MEASURE BEFORE THEY DEBATE: AUDITING LLM RESPONSE COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Response probes measure how an LLM reacts to individual inputs, but using those measurements to predict responses to joint inputs requires a separate composition assumption. We study this assumption through a finite-input audit that holds the interface fixed, excludes combinations from calibration, and separates prediction error from response noise. In a joint study of $5{,}184$ responses, summed single-sender secants fall below common-input secants by $.34915$ and $.45782$; both approximate simultaneous 95% intervals exclude zero. One configuration also answers all 768 nonzero-target arithmetic queries correctly, although the prespecified global screen fails. A prospective test on eight new, separately calibrated propositions replaces numerical cards with prose comments. Across $4{,}352$ responses, 11 of 16 composition contrasts have approximate simultaneous intervals below zero; five remain unresolved. Addition retains the lowest observed overall held-out error, while averaging measured increments improves mixed-input prediction. Further tests show how calibration support redistributes error and how remeasuring 216 archived replies changes a gain ordering. The audit distinguishes failed composition identities from useful predictions and makes sensitivity to input families and readouts explicit. The evidence covers numerical cards and controlled prose substitutions; it does not establish internal mechanisms or performance in unrestricted discussion.

1 INTRODUCTION

Multi-agent LLM systems use interaction to support reasoning, oversight, and coordination (Du et al., 2024; Irving et al., 2018; Khan et al., 2024; Wu et al., 2024). Beyond evaluating a discussion's final answer, response probes offer a way to measure behavior before interaction. A probe might estimate how strongly a recipient follows a peer, how much it retains a supplied prior, or how its response changes with the displayed evidence. The resulting coefficients are informative about the inputs used to measure them. Their use in a new combination is a further empirical claim.

Consider a recipient shown several peer messages. Changing one message at a time measures a set of isolated responses. Adding those changes predicts the response when the messages change together only if the effects compose. Neither accurate numerical processing nor a good fit on the calibration inputs establishes that property. Figure 1 illustrates this test with topic-specific prose substitutions: construct a prediction from baseline and isolated responses, then evaluate excluded joint inputs under the same interface.

In one measured configuration, an LLM correctly performs all 768 instructed calculations on mixed inputs, yet its summed single-sender secants are about $.54$ and its common-input secant is 1. The second configuration retains arithmetic failures, keeping the global control failed. Concurrent controls thus separate instructed computation from natural-response composition.

We organize the evidence around four questions: whether composition holds under concurrent arithmetic controls; whether discrepancies extend to controlled prose; how calibration support changes prediction; and how the readout changes inferred gains. The prose test uses eight propositions fixed before development, with topic-specific calibration and excluded combinations. Eleven of sixteen prespecified contrasts have approximate simultaneous intervals below zero. Addition remains the best observed overall predictor among the fixed comparators: imperfect composition does not eliminate its predictive value.

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

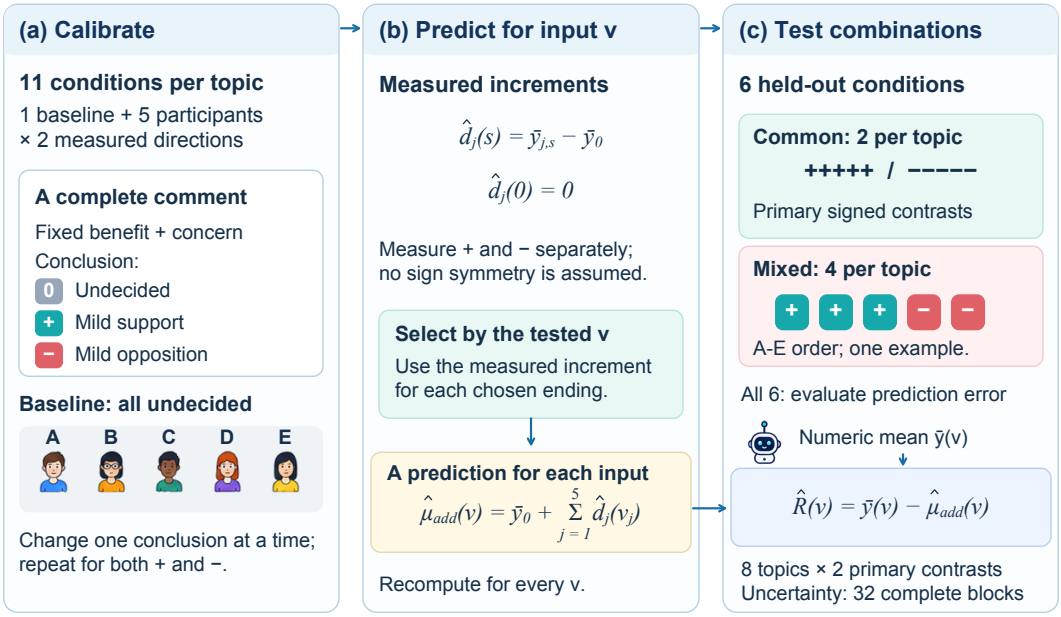

Figure 1: **Finite-input composition auditing, illustrated by the prose test.** An undecided recipient reads five scripted comments under fixed model, sampling, order, and readout settings. Comments share a benefit and concern; only conclusions change. Symbols denote prose variants. Eleven conditions calibrate both directions separately; each combination selects its corresponding increments. Two common holdouts define the primary signed contrasts; all six common/mixed holdouts assess prediction errors. Residuals are observed means minus predictions. Arrows show information use, not request order. This single-round design requires calibration on each topic (index omitted).

Our contribution is a reproducible audit connecting response measurements to the joint-input predictions they are intended to support. Fixed-layout comparisons test composition directly, and concurrent controls show that accurate instructed arithmetic need not establish it. The prospective prose test finds composition discrepancies on new, separately calibrated propositions while retaining addition's lowest observed overall loss. Calibration-support and paired-readout comparisons show how measurement choices change error allocation and gain ordering. The finite-input algebra defines the tested equalities without identifying derivatives or a social-influence mechanism.

2 Related Work

Classical opinion dynamics model averaging, anchoring, and network-mediated interaction (DeGroot, 1974; Friedkin & Johnsen, 1990; Bindel et al., 2011). LLM debate studies examine reasoning and oversight (Du et al., 2024; Liang et al., 2024; Chan et al., 2024), alongside sycophancy, conformity, and polarization (Sharma et al., 2024; Zhang et al., 2024; Piao et al., 2025; Ha & Mitchell, 2026; Cui et al., 2025; Han et al., 2026; Zhong et al., 2025). These behaviors motivate measuring responses, but an observed association does not identify the operator governing an interaction.

The closest measurement approaches estimate transitions from debate traces (Okawa, 2026) or probe responses before interaction (Yang, 2026). Yang already measures own-opinion resistance and peer response, documents single-neighbor-to-group transfer failures, and evaluates interaction-matched consensus forecasts. We extend this work through fixed-layout sender/common tests, concurrent arithmetic controls, excluded prior–peer combinations, prospective prose substitutions, and paired readout audits. We do not claim the first transfer failure or better deployment forecasts. Graph-only spectra (Parks & Alharthi, 2026) and interacting-evaluator contagion matrices (Liu, 2026) concern complementary objects.

Ma et al. (2025) show that multi-agent judging can amplify evaluation bias and that a PINE-based agent can mitigate it. Our readout study instead holds archived responses fixed and changes their

measurement protocol. It tests sensitivity of inferred gains, without identifying a uniquely correct judge or replicating a debiasing intervention.

3 FROM ISOLATED RESPONSES TO A TESTABLE COMPOSITION CLAIM

3.1 THE FINITE-INPUT AUDIT

Fix a task, model and sampling policy, recipient configuration, display, and numerical readout $Y \in [-1, 1]$. Let $z = (z_1, \ldots, z_d)$ index the declared input substitutions, and write $\mu(z) = E[Y \mid z]$. The baseline 0 is part of the design. An index can denote a numerical stance or a particular prose variant; no distance between text variants is assumed. Let $z^{(j)}$ retain substitution z_j at input j and set all other inputs to baseline. The isolated-response prediction and its population residual are

$$\mu_{\text{add}}(z) = \mu(0) + \sum_{j=1}^{d} \big[\mu(z^{(j)}) - \mu(0)\big], \qquad R(z) = \mu(z) - \mu_{\text{add}}(z). \tag{1}$$

Calibration measures the terms on the right; a held-out joint input measures the first term of R. Prior and peer inputs can both be coordinates, as in the prior–peer pilot, or the prior can remain fixed while messages vary.

Proposition 1 (Finite composition identity)**.** *If $\mu(z) = c + \sum_j h_j(z_j)$ on the tested baseline, single-input, and joint points, then $R(z) = 0$ there. A nonzero population residual excludes this coordinate-additive representation on those points. Zero residuals on a finite test set establish neither additivity elsewhere nor a local Jacobian.*

The identity follows by cancellation; each $h_j(z_j) - h_j(0)$ appears once. The boundary is substantive. On $[-a, a]^d$, $d \geq 3$ and $0 < a \leq 1$, the smooth responses $g_0(x) = 0$ and $g_1(x) = x_1 x_2 x_3 / a^2$ agree on every coordinate axis and have the same zero Jacobian at the origin. Yet their values at $\pm a\mathbf{1}$ differ by $\pm a$. Thus perfect isolated-response measurements and smoothness alone do not justify finite-scale composition (App. A). This is an identification example, not a proposed LLM mechanism.

For numerical inputs, central sender secants and the common secant are

$$\beta_j = \frac{\mu(ae_j) - \mu(-ae_j)}{2a}, \quad \kappa_{\text{common}} = \frac{\mu(a\mathbf{1}) - \mu(-a\mathbf{1})}{2a}, \quad \Delta = \sum_j \beta_j - \kappa_{\text{common}}. \tag{2}$$

Exactly, $\Delta = -[R(a\mathbf{1}) - R(-a\mathbf{1})]/(2a)$. Hence an additive response has $\Delta = 0$, but the converse need not hold: even joint interactions can cancel between the two signs. The prose study instead estimates both signed residuals separately. The two designs test related equalities in different units and need not produce the same directional mechanism.

3.2 CALIBRATION, UNCERTAINTY, AND PREDICTION LOSS

The historical factorial probe fits

$$Y = a\bar{x} + bs_0 + c, \qquad \widehat{\kappa} = \widehat{a}, \tag{3}$$

where s_0 is a supplied prior and $\bar{x}$ a common peer stance. Crossing $s_0 \in \{-.5, .5\}$ with $\bar{x} \in \{-.5, 0, .5\}$ separates peer response from prior anchoring. Its pooled cell-mean $R^2 \geq .95$ diagnostic tests fit on those six cells. Later axis calibrations use different prompts and smaller inputs; their failure does not retroactively invalidate this in-grid fit. Comparisons below directly vary calibration support while preserving the held-out responses.

We quantify sampling uncertainty from complete-block contrasts. In the joint and prose studies, each block b provides a contrast C_b from its jointly collected cells. With n blocks, we estimate its mean and standard error as $\bar{C}$ and $s_C/\sqrt{n}$, retaining within-block covariance. Prespecified multipliers 2.30 and 3.10 give approximate simultaneous intervals over two and sixteen contrasts. A nonzero sample residual alone does not establish a nonzero population residual. The intervals require independent blocks and approximately normal averages; repetitions do not create independent tasks.

Prediction is evaluated on excluded cell means. The prose test refits calibration within complete-block bootstrap resamples; the Qwen and DeepSeek calibration audits instead resample within cells. We report common-input, mixed-input, and pooled losses separately. In a two-family evaluation, if π is the weight assigned to common inputs, pooled MSE is $\pi\,\mathrm{MSE}_{\mathrm{common}} + (1-\pi)\,\mathrm{MSE}_{\mathrm{mixed}}$. Predictor rankings can change with this mixture. The prose test's fixed equal-cell evaluation uses $\pi = 1/3$ because each topic has two common and four mixed holdouts. Neither a low pooled loss nor a good calibration fit establishes adequate prediction on every input family.

3.3 An operational audit specification

Algorithm 1 specifies the joint and prose analyses. Its input roles are declared for each comparison: a common-input response can test isolated-response composition and calibrate a different mixed-input predictor. It remains excluded from fitting the prediction it tests. The algorithm summarizes offline analysis of complete blocks; the frozen study scripts specify each contrast, resampling rule, and control threshold (App. H).

Algorithm 1 Offline composition audit on complete blocks

Require: Archived blocks $Y_1, \ldots, Y_n$; frozen cell roles, contrast matrix A, predictors, loss weights, and uncertainty rules

1: Verify block completeness, readouts, and the declared input roles.
2: Estimate cell means; fit each predictor on its calibration cells only.
3: Form $C_b = AY_b$; compute contrast means and standard errors across blocks.
4: Apply the frozen family multiplier; retain all primary contrasts and intervals.
5: Predict excluded cells; report errors by input family and for their fixed mixture.
6: Apply the declared uncertainty calculation, refitting calibration where required.
7: Where present, score arithmetic controls separately, retaining local and global results.
8: **return** Contrasts, intervals, predictive losses, and controls, including unresolved results.

For the joint audit, A encodes two sender/common secant contrasts over 96 blocks; for the prose test, it encodes sixteen signed topic contrasts over 32 blocks. Computing contrasts before their standard errors retains dependence among cells in a block (App. A). An unresolved contrast is not evidence of adequate composition. The output describes the tested support; it supplies no automatic deployment or intervention decision.

4 Evidence across controls, inputs, and readouts

We separate three questions: whether a composition identity holds, how well a fixed predictor performs on excluded inputs, and whether a readout change alters the inferred response. Table 1 maps each comparison to its input support. The study names describe their scientific roles; Appendix H maps them to the fixed archive identifiers. Model identities refer to pinned snapshots when available and otherwise to documented serving cohorts. A provider fingerprint is serving metadata, not proof of immutable weights. The studies are internally frozen or explicitly retrospective; none is upgraded to externally preregistered confirmation.

4.1 Arithmetic execution does not establish composition

The joint audit measures natural responses (N), the same social prompt with an explicit weighted-sum rule (R), and context-free dot products (A) in one DeepSeek-V4.1-Flash serving cohort and sampling policy. Before acquisition, we selected the configurations with the DeepSeek calibration audit's largest and smallest contrast magnitudes: C01/S1/forward/uniform and C03/S1/forward/unequal, now U and H. N fixes the private prior at zero and uses six cards, including a zero-weight self-card: a baseline, twelve single-sender and two common endpoints at $\pm.05$, and four mixed inputs. Both instructed arms use the same mixed inputs. Arithmetic agreement uses an inclusive 10^{-6} tolerance. A precision calculation fixed 96 complete blocks, with shuffled requests within each block.

Both natural-response contrasts are negative (Figure 2): U has $\widehat{\Delta} = -.34915$ with approximate simultaneous 95% interval $[-.38611, -.31219]$; H has $-.45782$ with interval $[-.49711, -.41852]$.

Table 1: **Each comparison has its own calibration and evaluation support.** Cell counts are per configuration or topic. The Qwen and DeepSeek calibration audits each contain 4,320 responses across 24 strata. The joint and prose designs use complete blocks; the readout audit pairs archived texts. Repeated responses are not independent tasks.

Comparison	Information used	Evaluation and estimand
Joint: composition	12 natural single-sender endpoints	2 common endpoints; secant contrast Δ
Joint: prediction	Natural baseline + 2 common cells	4 mixed cells; six aggregators' RMSE
Joint: arithmetic	Explicit rule; no fitted response model	Same 4 mixed inputs in each instructed arm; agreement and global screen
Prose: composition / prediction	Topic-specific baseline + 10 single-participant cells	2 common cells: primary signed residuals; all 6 common/mixed cells: prediction errors
Calibration support	Fixed alternatives using 1, 3, 5, 6, 9, or 15 cells	Same 24 excluded cells; error by input class
Paired readout	Same archived text within each paired comparison	Tag/judge gains; common-cell and missing-readout sensitivity

All four prespecified time-quartile means are also negative in each configuration. These estimates concern the joint-audit cohort; uncertainty in the earlier calibration cohort is assessed separately (Section 4.3).

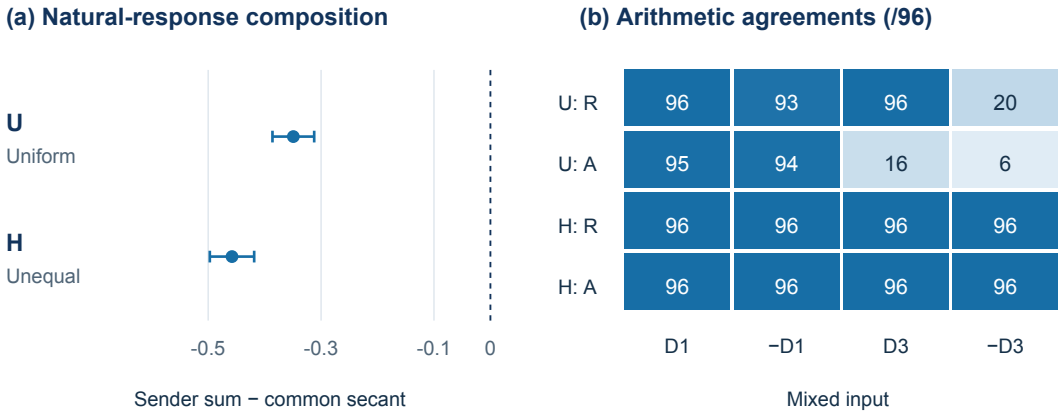

Figure 2: **Instructed computation and response composition must be tested separately.** Left: natural-response secant contrasts from 96 complete blocks, with approximate simultaneous 95% intervals. Right: all sixteen R/A mixed-input cells, each with 96 replies. H's eight nonzero-target cells all succeed; U's failures keep the prespecified global arithmetic screen failed. Both configurations were selected from historical data.

The global arithmetic screen requires all sixteen cellwise, one-sided binomial lower bounds to exceed .90, using $\alpha = .05/16$. It fails: U has $305/384$ R agreements and $211/384$ A agreements. In H, all eight cells have $96/96$ agreements with nonzero targets; each adjusted lower bound is .94168. These prespecified local results establish accurate instructed computation alongside failed natural-response composition on the measured support. They narrow a blanket arithmetic-inability explanation. They do not impose a weighted-sum target on natural judgment, and R/A cover four mixed inputs rather than all fourteen sender/common endpoints.

Simple nonlinear rules remain plausible alternatives. Six prespecified aggregators are calibrated only on N's baseline/common cells and predict four excluded mixed cells. In H, median, trimmed, and gated rules each reduce RMSE from the weighted rule's .01341 to .00765; these three coincide on the tested inputs, and residual error remains above estimated cell-mean noise RMS .00071. In

U, the all-card mean has the lowest observed RMSE, .00283. The controls distinguish claims about prediction, not latent algorithms. App. B gives the complete design and results.

4.2 Composition also fails on controlled prose inputs

The prose test replaces numerical peer cards with comments from five participants (Figure 1). Each repeats the same proposition-specific benefit and concern, followed by mild support, mild opposition, or an undecided conclusion. Order is fixed; the recipient starts undecided and reports a numerical stance. No peer weight or arithmetic instruction appears. Eight test propositions were fixed before a separate four-proposition development stage. Development determined the repetition count through a fixed variance rule; it did not select prompts, topics, directions, or predictors. The test collects 32 complete 8×17-cell blocks (4,352 responses) in a DeepSeek-V4.1-Flash cohort.

For each topic, the neutral baseline and ten single-participant cells calibrate all five predictors. The two common and four mixed cells remain excluded. These are prospective tests on new propositions with topic-specific calibration, not calibration-free cross-topic forecasts. Equation (1) gives the two signed substitution residuals,

$$\widehat{I}_{\pm} = \bar{y}_{\mathrm{all},\pm} - \sum_{j=1}^{5} \bar{y}_{j,\pm} + 4\bar{y}_0.$$

Eleven of sixteen approximate simultaneous 95% intervals lie below zero: all eight opposition directions and support directions T1–T3 (Figure 3). Five support directions remain unresolved, including T5 whose unrounded upper endpoint is $+0.00004559$. All half-widths are below .1. All 256 neutral replies are zero, responses lie in $[-.25, .35]$, and every additive holdout prediction is interior. Observed output clipping therefore does not explain the discrepancies, although latent thresholds and other nonlinearities remain possible.

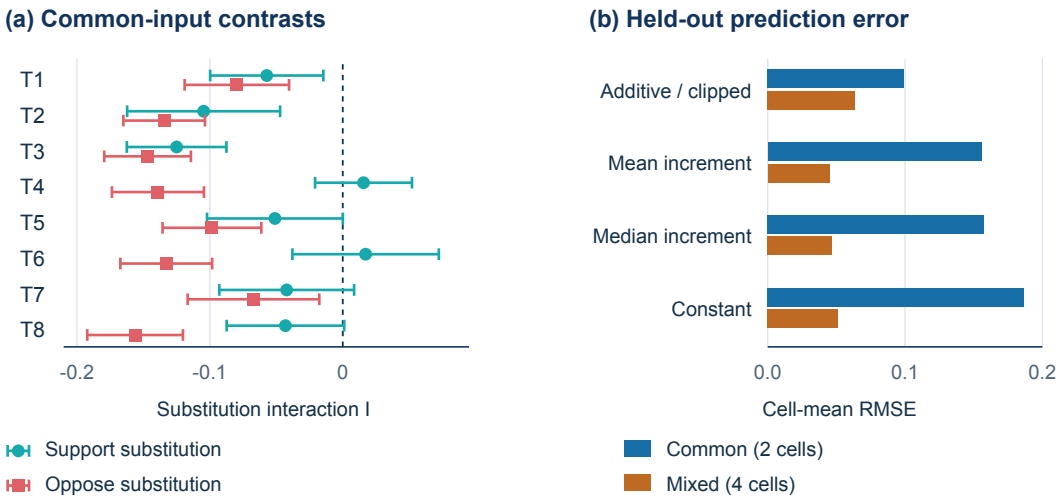

Figure 3: **Prose composition errors coexist with useful additive predictions.** Left: all sixteen fixed-topic contrasts and approximate simultaneous intervals; T5 support includes zero. Right: common and mixed held-out errors. Additive and clipped predictions coincide; mean increments improve mixed inputs but lose common-input information. Repetitions quantify uncertainty on eight fixed topics, not task-population incidence.

The five predictors use summed increments, their clipped prediction, the mean or median increment, or the baseline alone. Additive/clipped predictions have the lowest observed common-input RMSE (.09944) and pooled held-out RMSE (.07746). Averaging improves mixed-input RMSE from .06369 to .04502, but raises pooled RMSE to .09717; constant prediction gives .11550. The paired mixed-input MSE difference is $-.00203$, with pointwise block-bootstrap 95% interval $[-.00271, -.00159]$. All twelve secondary intervals are descriptive rather than simultaneous selection tests. Additive mixed-input error exceeds estimated mean-response noise RMS .01132.

Composition fails on part of this support, while addition preserves information that averaging loses. The prose test remains a controlled, single-round prose experiment with repeated arguments and a numerical readout. App. C supplies all contrasts, predictor losses, exact prompts, and the development-to-test decision.

4.3 Calibration support changes where prediction succeeds

Two preliminary audits examine specific forms of composition. The prior–peer pilot estimates prior and common-peer responses from zero and axis inputs, then predicts excluded joint inputs. All 24 peer-responsive joint cell means differ from the additive predictions, with RMSE .03125–.05 at input magnitude .05; prior-anchored controls have zero error. This does not test the original full-factorial estimator. The layout-transfer pilot tests common-input gains on heterogeneous cards but also changes the weight layout, leaving a layout confound. Both use a Qwen3-8B-FP8 snapshot on three development propositions (App. D.2).

The Qwen calibration audit holds each layout fixed across calibration and testing, separating within-layout composition from layout transfer. It retains six cards, two recipient labels, forward/reverse orders, and three propositions, giving 24 strata with four replies per cell. Uniform nonself weights are .2; unequal weights are .6 and four .1s; self-weight is zero. Every one of 1,152 single-sender replies is zero, so all 144 sender secants are zero. Common secants nevertheless range from .75 to 1. The discrepancy occurs within a fixed layout, not only after layout transfer.

A DeepSeek calibration audit applies the same numerical design and calibration supports to three separately fixed propositions. Single-sender behavior differs from the Qwen cohort: $319/1{,}152$ single-sender replies are nonzero. Summed sender secants still fall below common secants in all 24 observed strata, with differences $-.725$ to $-.05$. The weakest has estimated standard error .0743 and pointwise within-cell bootstrap interval $[-.1875, .0625]$; it remains unresolved. Family, propositions, and serving implementation change together, so this is not an isolated model-family comparison.

Table 2: **Richer calibration reallocates prediction error.** Qwen calibration RMSE on excluded cell means, equally weighted across 24 fixed strata. Common, joint, and mixed columns contain 4, 8, and 12 cells per stratum; mixed results are split by prior below. Fit cells cost four calls each. Full weighted/unweighted comparisons and results from the DeepSeek calibration audit appear in App. D.

Predictor	Fit cells	Common	Joint	Mixed $p = 0$	Mixed $p \neq 0$	All
Weighted axis	5	.02081	.04908	.01227	.01511	.03104
Historical-support affine	6	.05944	.02983	.00988	.01039	.03059
Augmented affine	9	.04038	.02839	.00843	.00972	.02408
Bilinear	9	.04038	.03108	.00843	.00978	.02517
Even-prior interaction	9	.02125	.08098	.01317	.01353	.04847
Sender affine	15	.06915	.03749	.00574	.00145	.03573

The support comparisons use a fixed union of baseline, axes, common inputs, and prior–common corners, excluding new magnitudes and combinations from fitting. Historical-support affine calibration reduces joint-input error but increases common-input error relative to axis calibration. On the same nine cells, neither interaction model improves pooled loss over augmented affine calibration (Table 2). Increasing support and adding model terms therefore answer different questions; the five-, nine-, and fifteen-cell fits are not cost-matched.

These errors exceed estimated response noise under the stated sampling assumptions. Weighted-axis pooled RMSE is .03104 in the Qwen calibration audit and .02110 in the DeepSeek calibration audit, versus combined calibration/holdout noise RMS .00387 and .00631. Layout matching alone is insufficient: the DeepSeek calibration audit's matched and transported primary RMSEs are .0101423 and .0101294, with a paired difference interval $[-.000265, .000296]$. That tiny difference establishes neither ranking nor equivalence. Median, trimmed, and gated baselines can reproduce the qualitative sender/common pattern, but their improvements vary by cohort and do not remove joint-input errors (App. D.4).

4.4 CHANGING THE READOUT CAN CHANGE THE INFERRED GAIN

Changing the readout reverses the L4/L10 gain ordering on the same paired replies within each level. The readout audit tests this sensitivity by remeasuring all 216 archived Gemini L4/L10/L11 replies with one fixed DeepSeek judge protocol in two complete batches. Original tags are removed, while surrounding text remains. Batch A is primary and yields 176 scorable replies and 40 valid insufficient-evidence judgments; batch B gives 175 and 41. Insufficient judgments are not replaced by zeros.

Table 3: **Readout choice changes gain ordering on identical replies.** Within each level, both OLS fits use the same paired subset. These are subset peer coefficients, not derivatives or full-design estimates.

Level	Paired n	Original tag	Primary judge A
L4	63	1.32383	1.72463
L10	51	1.13986	1.93699
L11	47	1.86530	2.01468

The L4/L10 point-estimate ordering reverses on identical replies within each level (Table 3). A retrospective comparison assigning equal weight to all eighteen shared topic–prior–peer cells retains that reversal: L10 minus L4 is $-.16181$ for tags and $+.23889$ for A. It is conditional on available paired readings and does not establish which readout is correct.

Missing scores leave the full-design ordering unresolved. Allowing every insufficient reading to take any value in $[-1, 1]$ gives attainable L10-minus-L4 ranges $[-.54583, .37083]$ for A and $[-.47917, .35417]$ for B. Both cross zero, although all individual full-design gains remain above one under these completions. These are algebraic ranges, not confidence intervals. Of 171 replies scorable in both batches, 164 agree exactly; repeatability neither supplies missing scores nor establishes semantic ground truth. App. E gives all paired, missingness, and readout analyses.

5 IMPLICATIONS AND LIMITS

Evaluate response coefficients against their intended predictions: fix the display and readout, reserve combinations, propagate uncertainty, and report input-family and aggregate losses. In the prose test, additive equalities fail while addition retains the lowest pooled prediction error.

The scope is finite. The joint audit covers two historically selected numerical configurations; the prose test covers eight separately calibrated topics, repeated prose, fixed order, a neutral prior, and one-round numerical responses. Neither estimates incidence or consequences in unrestricted discussion. Arithmetic success covers only its tested cells. Simple nonlinear aggregators remain compatible with the observations. The mathematical examples identify inference limits, not internal computation or task-population generalization guarantees.

Historical debate results motivate this work but do not supply the missing deployment evidence. Among 42 configurations, all 28 descriptively subcritical configurations fall in the low-extremization band. The outcomes reuse calibrations and topics, and the bands are post hoc. Spectral scores are not separated from gain, exposure, or degree in the relevant comparisons. Equal-budget targeted intervention has no established advantage over random allocation; the preselected correction dose $k = 3$ fails and exploratory $k = 4$ succeeds. The natural-error follow-up misses its success criteria. These limits, Qwen's failed arithmetic controls, and the stopped operator pipeline remain explicit in App. G and the archived study report.

Accurate instructed calculation can coexist with failed composition, violated additive equalities can coexist with useful prediction, and the same paired replies within each level can yield different gain orderings under different readouts. The audit makes these distinctions testable on declared inputs; broader mechanisms and decision benefits remain open.

AI USE STATEMENT

LLMs are both experimental subjects and tools used in producing this study. Agent and judge models generated synthetic responses and assisted with the designed measurements. Generative AI was also used to refine hypotheses, the conceptual audit framework, and experimental methods; formulate and check mathematical claims and proofs; implement and review experiment, analysis, and figure code; generate illustrative human and robot icons for Figure 1; clean and reorganize records; analyze and interpret results; search and synthesize literature; translate research notes into English prose; and draft, restructure, and edit the manuscript. Scientific decisions, the validity and originality of AI-assisted content, code, mathematics, data analysis, and citations remain the responsibility of the human authors. Reproducible numerical checks and source audits are included in the supplement; these checks do not replace author verification. Models are identified by pinned revision when available; otherwise we report the requested alias, acquisition window, and retained provider metadata. The older DeepSeek runner did not retain returned revision or fingerprint metadata, a limitation detailed in the archived study report.

ETHICS STATEMENT

This work studies measurement validity in multi-agent LLM systems, motivated by the risk of undesired conformity amplification. The calibration procedure could also induce extremization, but doing so requires direct control of prompts or routing; we provide the detail needed for reproducibility while emphasizing screening and moderation. No human subjects or private user data were involved. Released prompts and transcripts contain only synthetic task content.

REPRODUCIBILITY STATEMENT

The anonymous supplement provides a single offline entry point: `python verify_artifact.py --full`. It verifies the release manifest and reproduces the central studies from retained requests and responses in temporary working copies. The accompanying map links paper claims to data and checks; no model access is required. Retained prompts, configurations, schedules, offline analyses, and study manifests are included; private acquisition infrastructure and superseded release files are excluded. It distinguishes effective sampling parameters from archival seed identifiers and historical response strings from newly retained HTTP bodies; audit READMEs enumerate historical inputs unavailable for byte-level checks. Machine-readable ledgers record each result's standing and the API serving-identity boundary. Numerical corrections are retained as dated addenda rather than used to overwrite frozen artifacts; historical study status is documented in the archived study report.

REFERENCES

David Bindel, Jon Kleinberg, and Sigal Oren. How bad is forming your own opinion? In *IEEE 52nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 57–66, 2011.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations (ICLR)*, 2024.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-MAD: Consensus-free multi-agent debate. *arXiv preprint arXiv:2509.11035*, 2025.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69 (345):118–121, 1974.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*, 2024.

Noah E. Friedkin and Eugene C. Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990.

Seungwoong Ha and Melanie Mitchell. Beyond correctness: Distance-based social dynamics of multi-agent debate. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Chen Han, Jin Tan, Bohan Yu, Wenzhen Zheng, and Xijin Tang. Conformity dynamics in LLM multi-agent systems: The roles of topology and self-social weighting. *arXiv preprint arXiv:2601.05606*, 2026.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *International Conference on Machine Learning (ICML)*, 2024.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Zewen Liu. Contagion networks: Evaluator preference propagation in multi-agent LLM systems. *arXiv preprint arXiv:2606.20493*, 2026. URL https://arxiv.org/abs/2606.20493.

Chiyu Ma, Enpei Zhang, Yilun Zhao, Wenjun Liu, Yaning Jia, Peijun Qing, Lin Shi, Arman Cohan, Yujun Yan, and Soroush Vosoughi. Judging with many minds: Do more perspectives mean less prejudice? on bias amplification and resistance in multi-agent based LLM-as-judge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17356–17392. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.941. URL https://aclanthology.org/2025.findings-emnlp.941/.

Maya Okawa. Emergence of biased consensus in multi-agent LLM debates. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2608.02827.

Ethan Parks and Dalal Alharthi. Predictive maps of multi-agent reasoning: A successor-representation spectrum for LLM communication topologies. *arXiv preprint arXiv:2605.11453*, 2026. URL https://arxiv.org/abs/2605.11453.

Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Fernando P. Santos, Yong Li, and James Evans. Emergence of human-like polarization among large language model agents. *arXiv preprint arXiv:2501.05171v1*, 2025. URL https://arxiv.org/abs/2501.05171v1.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed H. Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling (COLM)*, 2024.

Dongxu Yang. When is emergent consensus real? a measured coupling gain and a validity diagnostic for LLM agent societies. *arXiv preprint arXiv:2606.22203*, 2026. URL https://arxiv.org/abs/2606.22203.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Huixin Zhong, Yanan Liu, Qi Cao, Shijin Wang, Zijing Ye, Zimu Wang, and Shiyao Zhang. Disentangling the drivers of LLM social conformity: An uncertainty-moderated dual-process mechanism. *arXiv preprint arXiv:2508.14918*, 2025. URL https://arxiv.org/abs/2508.14918.

Appendix guide. The core record is organized by scientific question rather than acquisition order. The archived study report retains the earlier scientific account, with bibliographic metadata corrected against primary sources.

Question	Location
What does a finite composition test identify?	Appendix A
Can arithmetic succeed while composition fails?	Appendix B
Do prose substitutions compose on new topics with topic-specific calibration?	Appendix C
How do calibration support and sampling noise change prediction?	Appendix D
How much does the readout change the inferred gain?	Appendix E
What did the earlier arithmetic controls establish?	Appendix F
Which deployment claims remain unresolved?	Appendix G
Where are results and reproduction checks indexed?	Appendix H

A FINITE COMPOSITION AND THE LIMITS OF LOCAL INFERENCE

The finite composition identity. Let μ be the conditional-mean response under a fixed task, recipient configuration, substitution protocol, model, sampling policy, display, and readout. For a declared input vector z, let $z^{(j)}$ retain its jth substitution and set all other coordinates to baseline, as in Equation (1). Define

$$R(z) = \mu(z) - \mu(0) - \sum_{j=1}^{d} \{\mu(z^{(j)}) - \mu(0)\}.$$

If $\mu(z) = c + \sum_j f_j(z_j)$, then $R(z) = 0$: each isolated increment equals $f_j(z_j) - f_j(0)$, and their sum equals $\mu(z) - \mu(0)$. This includes affine responses but does not require linearity within each coordinate. Conversely, observing $R = 0$ at finitely many tested inputs does not establish coordinate additivity elsewhere. For prose inputs, a coordinate denotes a fixed message substitution, not a numerical derivative.

What a sender/common secant checks. For a numerical amplitude $a > 0$, the contrast used in the numerical calibration and joint audits satisfies the exact identity

$$\sum_j \frac{\mu(ae_j) - \mu(-ae_j)}{2a} - \frac{\mu(a\mathbf{1}) - \mu(-a\mathbf{1})}{2a} = -\frac{R(a\mathbf{1}) - R(-a\mathbf{1})}{2a}.$$

This follows by subtracting the two definitions of R; the baseline terms cancel. Thus a nonzero secant contrast contradicts the corresponding finite additive equality, but a zero contrast is not sufficient to establish it. For example, $\mu(x) = x_1x_2$ has zero single-coordinate responses and $R(a\mathbf{1}) = R(-a\mathbf{1}) = a^2$, giving a zero secant contrast despite failed joint composition.

A second example separates finite response from the derivative. On $[-a, a]^d$, for $d \geq 3$ and $0 < a \leq 1$, the smooth functions $\mu_0(x) = 0$ and $\mu_1(x) = x_1x_2x_3/a^2$ agree on every coordinate axis and have the same zero Jacobian at the origin. Yet $\mu_1(\pm a\mathbf{1}) = \pm a$ and its sender/common contrast is -1. Both functions stay in $[-1, 1]$ on this domain. These are analytic identification examples, not inferred LLM mechanisms.

From finite identities to block calculations. Let Y_b collect the unique cell responses in complete block b, in a fixed order. For $n \geq 2$ blocks, let A contain the fixed contrast coefficients, $\bar{Y} = n^{-1}\sum_b Y_b$, and S_Y be the sample covariance with divisor $n - 1$. Then

$$C_b = AY_b, \qquad \widehat{C} = A\bar{Y}, \qquad \widehat{\mathrm{Cov}}(\widehat{C}) = \frac{1}{n} A S_Y A^\top.$$

Each diagonal entry equals the sample variance of the corresponding scalar block contrast divided by n. For independent, identically distributed blocks with finite second moments, this covariance estimator is unbiased by linearity and independence across blocks; independence within a block is unnecessary.

In the joint audit, at $a = .05$, each row assigns $+10$ and -10 to positive and negative sender endpoints, respectively, and -10 and $+10$ to the corresponding common endpoints; the baseline

cancels. Each primary row in the prose test assigns $+1$ to its signed common cell, -1 to each of five corresponding single-participant cells, and $+4$ to the neutral cell. For a mixed substitution changing only k coordinates, merging unchanged-coordinate baseline terms leaves baseline coefficient $k-1$. Repeated appearances of one baseline cell are not independent observations. The normal-average approximation and the frozen $2.30/3.10$ multipliers concern interval coverage, not these algebraic covariance identities.

Finite-grid regression does not identify an unrestricted derivative. For $g(x) = 1.2x - 1.6x^3$, the values at $x \in \{-.5, 0, .5\}$ are exactly $.8x$. Adding an affine prior term $bs_0 + c$ preserves the aliasing: the two-factor fit has slope .8 and $R^2 = 1$, while the derivative at zero is 1.2. More generally, for a finite intensity set $\mathcal{X}$ containing zero, $h(x) = x\prod_{z\in\mathcal{X}\setminus\{0\}}(x-z)$ vanishes on the grid but has $h'(0) \neq 0$. Adding a multiple of h preserves every grid observation and changes the derivative.

If $|g^{(3)}(x)| \le L$ on $[-h, h]$, central differencing instead gives

$$\left|\frac{g(h) - g(-h)}{2h} - g'(0)\right| \le Lh^2/6.$$

If both sampled endpoint means have error at most ϵ on a joint confidence event, the derivative error is at most $\epsilon/h + Lh^2/6$ on that event. Smaller steps reduce truncation bias while amplifying sampling error; agreement across a finite set of scales does not prove the regularity bound.

Conditional spectral background. For a continuously differentiable one-round conditional-mean map F, local linearization is

$$F(x^\circ + \delta) = F(x^\circ) + J\delta + r(\delta), \qquad J = DF(x^\circ), \quad \|r(\delta)\| = o(\|\delta\|). \tag{4}$$

At a fixed point, $\rho(J) < 1$ implies local asymptotic stability and $\rho(J) > 1$ implies instability; equality is inconclusive. Neither result identifies J from finite probes. Non-normal transient growth, saturation, adaptive routing, and omitted state coordinates require separate analysis. Supplied priors affect the affine offset rather than automatically becoming self-feedback. The complete historical affine, Perron-sensitivity, fluctuation, and conditional-bound statements and proofs remain unchanged in the archived study report, Sections B–D; no deployment in this paper completes their behavioral identification requirements.

B JOINT ARITHMETIC AND COMPOSITION AUDIT

B.1 FROZEN DESIGN, ACQUISITION, AND INDEPENDENT RECONSTRUCTION

The joint audit is a completed descriptive diagnostic with $5{,}184$ retained responses. Before acquisition, the design selected two configurations from the DeepSeek calibration audit by the largest and smallest absolute observed sender/common contrasts, breaking ties lexicographically: C01/S1/forward/uniform ($-.725$) and C03/S1/forward/unequal ($-.05$). We denote them U and H. The frozen files retain the historical labels `strong` and `weak`; these describe selection, not the new relative effects. The study uses a later API version, so it does not repeat the retired model from that calibration cohort under identical sampling. Neither these two selected configurations nor their repeated responses constitute a random sample of tasks.

All arms requested `deepseek-flash`, documented as DeepSeek-V4.1-Flash, with thinking disabled, temperature .7, effective top-$p = 1$, output cap 128, text response format, and no streaming. Acquisition occurred on September 11, 2026, from 09:52:52 to 11:05:34 UTC. One fingerprint, `aeb56401ca74e127821c4f9126dcb669`, is retained throughout. Its match to the fingerprint recorded in the later DeepSeek arithmetic control is serving metadata, not proof of immutable weights; the arithmetic–composition comparison is internal to the joint audit. The source schedule and numerical analyses were frozen before acquisition; this is an internal freeze, not external preregistration or deployment qualification. Continuation and response retention did not depend on arithmetic outcomes.

Each configuration uses the original natural-response profile and its fixed proposition. U concerns public libraries offering free workshops on repairing household appliances; H concerns publicly

funded museums providing free admission one day each week. The private prior is zero throughout. Six cards remain in forward sender order, including a zero-weight self-card. U assigns weights $(0, .2, .2, .2, .2, .2)$; H assigns $(0, .6, .1, .1, .1, .1)$. The natural arm N contains 19 cells: a zero baseline, twelve signed single-sender inputs of magnitude .05, two common inputs $\pm.05$, and four mixed vectors. Every displayed sender, including the zero-weight self-card, is perturbed separately and in the common direction. The mixed vectors are

$$D_1 = (0, .05, 0, 0, 0, -.05), \qquad D_3 = (0, .025, 0, 0, 0, -.10),$$

and their negatives. Their nonself means are 0 and $-.015$, respectively. On these four mixed vectors, R adds the exact weighted-sum instruction to the natural prompt; A requests the same numeric dot product without the social context. All request texts and raw response bodies are released; the exact R addition and A instruction are reproduced in Figure 4. N uses the peer-responsive profile in App. D.2.

Each complete block contains $2(19 + 4 + 4) = 54$ requests. The two A supports have different weights and are distinct numerical tasks. A separate frozen shuffle is used within each of 96 blocks. The sample-size calculation uses the standard error .0743023 of the historically weak contrast in the DeepSeek calibration audit, estimated from four repetitions, doubles its variance, and targets a contrast half-width of .05:

$$n \geq \frac{2.30^2 \times 2 \times 4(.0743023)^2}{.05^2} = 93.46.$$

Rounding up to the next multiple of 32 gives 96. Precision was not guaranteed because the historical variance estimate was noisy and the serving cohort changed. The sample size remained fixed.

Independent reconstruction verifies every original request and response body against its hash, reparses all 5,184 values, reconstructs the frozen block order, and reads numeric value–weight pairs directly from the request text. Exact-rational calculations reproduce the cell means, variances, targets, contrasts, prediction errors, and noise decomposition. All 6,141 leaves of the frozen scientific report match. An independent binomial-tail inversion also agrees with the frozen lower bounds. The anonymous joint audit package supplies a standard-library reproduction script and analytic fixtures. Running this script makes no model requests.

B.2 Composition contrasts and uncertainty

For each block b and configuration, the primary contrast is

$$\Delta_b = \sum_{j=1}^{6} \frac{y_b(.05e_j) - y_b(-.05e_j)}{.1} - \frac{y_b(.05\mathbf{1}) - y_b(-.05\mathbf{1})}{.1}.$$

The common-direction term includes all six cards, matching the sum of the single-sender directions. This finite-scale equality is necessary for an affine response surface; it is not a derivative test. We compute the mean of the 96 complete-block contrasts and use their sample variance for the standard error, permitting covariance among calls within a block. Intervals given by the mean ± 2.30 standard errors provide approximate simultaneous 95% coverage for the two configurations, assuming independent blocks and an adequate normal approximation for their averages. These intervals provide neither a distribution-free coverage guarantee nor inference over new tasks. Serial dependence or service drift can invalidate that approximation even when the observed fingerprint is stable.

Table 4: **Natural-response composition in the joint audit.** Each row has 96 complete blocks. SE uses the variance of block contrasts, including within-block covariance. Intervals assume independent blocks and approximately normal averages.

Config.	Sender sum	Common	$\widehat{\Delta}$	SE	Approx. simultaneous 95% interval
U	0.54355	0.89271	-0.34915	0.01607	$[-0.38611, -0.31219]$
H	0.54218	1.00000	-0.45782	0.01708	$[-0.49711, -0.41852]$

Both contrasts remain below zero at a scale larger than their estimated sampling uncertainty. The historical weak selection has the larger absolute contrast in this later cohort. It cannot be used to retroactively shrink the DeepSeek calibration audit's earlier interval. The fixed

time-quartile means in acquisition order are $(-.37186, -.36057, -.27569, -.38849)$ for U and $(-.51122, -.42168, -.45942, -.43894)$ for H. Their common sign is a descriptive temporal check, not proof of independence or stationarity.

The frozen analysis also reports a conservative Hoeffding family bound. Each response lies in $[-1, 1]$ and the fourteen signed coefficients have magnitude 10, so a block contrast lies in an interval of width 280. Under independent blocks, the resulting simultaneous 95% intervals are $[-42.64960, 41.95129]$ and $[-42.75826, 41.84263]$. They are uninformative. Thus the sharper intervals explicitly depend on the normal approximation and block-independence assumptions.

B.3 Arithmetic adequacy: the global screen fails

For each of the 16 configuration–arm–mixed-input cells, the target is the exact sum of displayed value–weight products. Agreement requires absolute error at most 10^{-6}. We report agreement, MAE, RMSE, and maximum error for all responses. The frozen adequacy screen requires every one-sided Clopper–Pearson lower bound, with per-cell $\alpha = .05/16$, to exceed .90. This assumes independent responses and a stable per-cell agreement probability. **The global screen fails.** It was not replaced by a subgroup screen after seeing the results.

Table 5: **All instructed-arithmetic cells in the joint audit.** Agreement tolerance is 10^{-6}. Lower bounds are one-sided Clopper–Pearson with $\alpha = .05/16$ per cell. All 16 lower bounds must exceed .90 for the frozen global screen; that screen fails.

Config.	Arm	Input	Target	Agree	Lower bound	MAE	RMSE	Max.
U	R	D_1	+0.000	96/96	0.94168	0.00000	0.00000	0.000
U	R	$-D_1$	+0.000	93/96	0.88449	0.00031	0.00177	0.010
U	R	D_3	-0.015	96/96	0.94168	0.00000	0.00000	0.000
U	R	$-D_3$	+0.015	20/96	0.10856	0.01583	0.01780	0.020
U	A	D_1	+0.000	95/96	0.92001	0.00010	0.00102	0.010
U	A	$-D_1$	+0.000	94/96	0.90148	0.00021	0.00144	0.010
U	A	D_3	-0.015	16/96	0.07822	0.01036	0.01282	0.020
U	A	$-D_3$	+0.015	6/96	0.01479	0.01875	0.01936	0.020
H	R	D_1	+0.025	96/96	0.94168	0.00000	0.00000	0.000
H	R	$-D_1$	-0.025	96/96	0.94168	0.00000	0.00000	0.000
H	R	D_3	+0.005	96/96	0.94168	0.00000	0.00000	0.000
H	R	$-D_3$	-0.005	96/96	0.94168	0.00000	0.00000	0.000
H	A	D_1	+0.025	96/96	0.94168	0.00000	0.00000	0.000
H	A	$-D_1$	-0.025	96/96	0.94168	0.00000	0.00000	0.000
H	A	D_3	+0.005	96/96	0.94168	0.00000	0.00000	0.000
H	A	$-D_3$	-0.005	96/96	0.94168	0.00000	0.00000	0.000

H has $96/96$ agreements in each of its eight R/A cells, all at nonzero targets ($\pm.025$ and $\pm.005$). Each corresponding lower bound is .94168. These prespecified per-cell outcomes permit a narrower descriptive observation: explicit computation succeeds on H's mixed inputs while natural responses in the same cohort show a negative composition contrast. U retains substantial errors, especially on the signed D_3 inputs. There is no overall arithmetic-adequacy claim, and the eight successful H cells were not chosen after acquisition.

The contrast and arithmetic tasks also differ intentionally in their instructions. N asks for an assessment using the recipient's own judgment; R fixes the numeric rule, and A removes the social context. R/A success therefore cannot make the weighted sum the correct target for N. Nor does the joint audit measure R/A on all fourteen sender/common endpoints. It narrows a general inability-to-compute explanation on the tested support without attributing N's behavior to social reasoning, proving an internal algorithm, or validating the entire numerical interface.

B.4 Prespecified predictions on excluded mixed inputs

All six candidates use only N's zero baseline and common $\pm.05$ calibration cells, with prediction $\widehat{c} + \widehat{\kappa} q(x)$. Here q is the weighted sum, nonself mean, all-card mean, nonself median, one-per-tail trimmed nonself mean, or nonself mean gated by at least three same-sign peers. The four mixed

cells are excluded from every calibration. Table 6 reports all candidates; none is selected to redefine the endpoint.

Table 6: **All prespecified prediction candidates in the joint audit.** Errors concern the four excluded natural-response mixed-cell means. Estimated noise combines held-out and calibration cell-mean variance under independent draws; adjusted MSE is not truncated. No held-out reply is used to fit a candidate.

Config.	Aggregate	RMSE	Estimated noise RMS	Adjusted MSE
U	Weighted	0.004102	0.000521	0.00001656
U	Nonself mean	0.004102	0.000521	0.00001656
U	All-card mean	0.002825	0.000518	0.00000771
U	Median	0.006201	0.000511	0.00003819
U	Trimmed mean	0.006201	0.000511	0.00003819
U	Gated mean	0.006201	0.000511	0.00003819
H	Weighted	0.013410	0.000710	0.00017933
H	Nonself mean	0.015053	0.000710	0.00022609
H	All-card mean	0.013526	0.000710	0.00018246
H	Median	0.007654	0.000710	0.00005807
H	Trimmed mean	0.007654	0.000710	0.00005807
H	Gated mean	0.007654	0.000710	0.00005807

On H, median, trimmed, and gated candidates have lower observed RMSE than the weighted candidate, but all three return $q = 0$ on each of the four mixed cells. They are observationally indistinguishable here. Their RMSE .00765 remains larger than estimated cell-mean noise RMS .00071, and all candidates retain positive MSE after the frozen subtraction. U instead has its smallest observed error under the all-card mean. These comparisons do not establish a universal aggregator or its internal use. The noise calculation assumes independent draws across calibration and test cells; the primary contrast analysis instead uses complete-block covariance.

See Figure 2 for the complete arithmetic-cell grid and composition intervals.

R: exact addition to the natural system prompt. For this condition, set your returned stance exactly to the sum of each displayed stance multiplied by its displayed weight, including all six cards. The weights already sum to one. Do not add the private assessment, renormalize the weights, or otherwise revise the numeric result. Return the same stance JSON.

A: exact system prompt. Compute the dot product of the six numeric rows: sum(value times weight). The weights sum to one. Return one JSON object with exactly one numeric field, stance, rounded to at most six decimal places. Do not output an explanation.

A: complete user message for H, D_3.

```
Numeric rows:
value=0.000000; weight=0.000000000000
value=0.025000; weight=0.600000000000
value=0.000000; weight=0.100000000000
value=0.000000; weight=0.100000000000
value=0.000000; weight=0.100000000000
value=-0.100000; weight=0.100000000000
/no_think
```

Figure 4: **Exact arithmetic-control instructions and one complete A request.** The displayed H/D_3 target is $+.005$. Numeric pairs and decoding settings match N/R for this cell; A removes the proposition and recipient profile. Complete N and R requests are retained in the anonymous scientific package.

C PROSPECTIVE PROSE COMPOSITION TEST

C.1 DESIGN, DEVELOPMENT, AND PROSPECTIVE STATUS

The prose test asks whether responses to single message substitutions predict combined substitutions when peer inputs are prose rather than numerical cards. The response remains a continuous stance in $[-1, 1]$. Five named participants A–E each present a proposition-specific benefit and concern in the same order. The three variants differ only in their concluding stance: mild support, mild opposition, or undecided. The recipient's supplied prior is always undecided. These scripted messages deliberately repeat arguments across participants; they are controlled substitutions, not samples of independent human evidence or naturally generated debate. No arithmetic aggregation rule or numeric weight is shown.

Four development propositions (D1–D4) and eight test propositions (T1–T8) were specified before development acquisition. The development stage collected $16 \times 4 \times 17 = 1{,}088$ responses. The test was frozen separately after development review, retaining the previously specified prompts, topic definitions, cell definitions, and five comparators. No development reply enters a test mean or fit, and no test topic or sign was removed. The test contains $32 \times 8 \times 17 = 4{,}352$ responses. Both stages are internally frozen, without external preregistration. The development study remains descriptive; the test supplies prospective fixed-topic evidence rather than task-population confirmation.

Each block contains all topics and all seventeen cells exactly once. Requests are ordered by a fixed SHA-256 permutation within each block and use separate conversations. Cells 0–10 calibrate every predictor: the all-neutral input and ten inputs changing exactly one of five participants to support or oppose. Held-out cells 11 and 12 change all participants together. Cells 13–16 use the vectors $(+, +, -, 0, 0)$, $(+, +, +, -, -)$, $(+, -, 0, 0, 0)$, and $(-, -, -, +, 0)$, respectively. Here symbols name the exact message variants; they do not denote numerical input intensities.

Sampling policy and serving identity. Both stages request `deepseek-flash`, documented as DeepSeek-V4.1-Flash at acquisition, with temperature .7, effective top-$p = 1$, a 128-token output cap, and thinking disabled. Development responses were acquired on 2026-09-11, 15:55–16:09 UTC; test responses on the same UTC date, 17:28–18:21. Each stage retains a single response fingerprint, and its value agrees across stages. This is serving metadata, not proof of immutable weights or an unchanged model between studies. The identity checks and complete-block analyses refer to the test cohort itself. No post hoc seed search, output repair, or scientific-response replacement was performed. All syntactically valid responses remain in the analysis.

Development decision and precision planning. The frozen review checks concern baseline displacement, boundary concentration, and flat response: absolute neutral mean at least .8; at least one quarter of a cell's replies at $|y| \geq .95$; or total cell-mean range at most .02 with every cell's response standard deviation at most .01. None was flagged. All 64 development neutral replies are zero, all responses lie in $[-.2, .3]$, and none of the 24 additive holdout predictions falls outside $[-1, 1]$. Contrast signs or significance did not determine whether testing proceeded.

Let $s^2_{\max}$ be the largest of the eight development block-contrast variances. The prespecified planning rule is

$$n_{\text{raw}} = 2s^2_{\max}(3.10/.10)^2, \qquad n_{\text{test}} = \max\{32,\ 16\lceil n_{\text{raw}}/16\rceil\},$$

subject to a maximum of 128 blocks. Here $s^2_{\max} = 751/96000$, so $n_{\text{raw}} = 15.03565$ and $n_{\text{test}} = 32$. Doubling the maximum observed variance is a planning heuristic, not an upper confidence bound. It cannot guarantee the desired precision on new topics. The realized test half-widths range from .03068 to .05754, all below the .10 planning target; this observation did not trigger further sampling.

C.2 PRIMARY CONTRASTS AND UNCERTAINTY

For topic t and complete block b, the finite substitution contrast is

$$I_{t,\pm,b} = Y_{t,\text{all},\pm,b} - \sum_{j=1}^{5} Y_{t,j,\pm,b} + 4Y_{t,0,b}.$$

Table 7: **Development-stage prose contrasts.** These are exploratory point estimates from sixteen complete blocks per fixed proposition, not test observations.

Topic	Support $\widehat{I}_+$	Opposition $\widehat{I}_-$
D1	-0.012500	-0.121875
D2	-0.003125	-0.134375
D3	-0.040625	-0.153125
D4	-0.012500	-0.131250

An additive conditional-mean model implies $E[I_{t,\pm,b}] = 0$. Each reported estimate averages 32 complete block contrasts. Its standard error is the sample standard deviation of these block contrasts divided by $\sqrt{32}$, retaining covariance among requests in the same block. The sixteen two-sided intervals use the frozen multiplier 3.10, a conservative rounding of the normal Bonferroni multiplier for sixteen comparisons at family level $.05$. Approximate simultaneous coverage assumes independent blocks and approximately normal averages; it is not exact finite-sample coverage. Sharing topics does not create sixteen independent tasks.

Eleven intervals lie below zero: all opposition substitutions and support substitutions on T1–T3 (Table 8). The five other support intervals include zero. In particular, T5's upper endpoint is $+0.00004559$, not zero or a negative number; rounded displays must not change its classification. A negative I_+ means less positive response than summing isolated support increments predicts. A negative I_- means more negative response than summing isolated opposition increments predicts. The common-minus-singles contrast here differs from the singles-minus-common numerical secant convention in the joint audit. The two studies need not have matching contrast signs: the shared conclusion is failure of the corresponding additive equality on the measured support, not replication of one directional mechanism.

All 256 test neutral replies equal zero. Across all $4{,}352$ replies the observed range is $[-.25, .35]$, with no $|y| \geq .95$ values. All 48 additive holdout predictions are interior. Thus observed readout clipping does not account for these composition errors. This does not rule out latent thresholds, nonlinear semantic judgments, or other internal bounded computations. No frozen measurement-review flag was triggered.

Table 8: **All sixteen prose-test contrasts, each from 32 complete blocks.** Intervals are approximate simultaneous 95% intervals; the final column uses unrounded endpoints.

Topic	Substitution	$\widehat{I}$	Block SE	Interval	Excludes zero
T1	Support	-0.057188	0.013739	$[-0.099778, -0.014597]$	Yes
T1	Opposition	-0.079687	0.012672	$[-0.118971, -0.040404]$	Yes
T2	Support	-0.104688	0.018562	$[-0.162230, -0.047145]$	Yes
T2	Opposition	-0.134375	0.009898	$[-0.165059, -0.103691]$	Yes
T3	Support	-0.125000	0.012090	$[-0.162479, -0.087521]$	Yes
T3	Opposition	-0.146875	0.010515	$[-0.179473, -0.114277]$	Yes
T4	Support	0.015625	0.011760	$[-0.020831, 0.052081]$	No
T4	Opposition	-0.139063	0.011166	$[-0.173676, -0.104449]$	Yes
T5	Support	-0.051042	0.016480	$[-0.102129, 0.000046]$	No
T5	Opposition	-0.098437	0.011982	$[-0.135582, -0.061293]$	Yes
T6	Support	0.017188	0.017764	$[-0.037882, 0.072257]$	No
T6	Opposition	-0.132812	0.011137	$[-0.167338, -0.098287]$	Yes
T7	Support	-0.042188	0.016361	$[-0.092907, 0.008532]$	No
T7	Opposition	-0.067187	0.015971	$[-0.116699, -0.017676]$	Yes
T8	Support	-0.043125	0.014266	$[-0.087349, 0.001099]$	No
T8	Opposition	-0.156250	0.011612	$[-0.192246, -0.120254]$	Yes

C.3 Held-out predictions, covariance, and noise

Write $b_t = \bar{Y}_{t,0}$ and $d_{tj}(v_j) = \bar{Y}_{t,j,v_j} - b_t$ for a nonneutral substitution, with $d_{tj}(0) = 0$. The five frozen predictors for a vector v are: additive, $b_t + \sum_j d_{tj}(v_j)$; clipped additive, its projection

onto $[-1, 1]$; mean increment, $b_t + \frac{1}{5}\sum_j d_{tj}(v_j)$; median increment, $b_t + \text{median}_{j=1}^{5}\, d_{tj}(v_j)$; and constant, b_t. Inactive participants contribute zero increments to the mean and median. These are fitted behavioral comparators, not arithmetic targets for natural judgment. All use only neutral and single-participant observations from the same test topic. The held-out status of T1–T8 refers to exclusion from development; predictions still require topic-specific calibration and do not test calibration-free transfer across topics.

For each input group, MSE averages squared cell-mean prediction errors equally across its cells and the eight fixed topics; RMSE is the square root of that MSE (Table 9). Common inputs comprise two held-out cells per topic, mixed inputs four, and the pooled group all six. Additive and clipped additive coincide on every holdout. They have lower common-input and pooled RMSE than the other candidates. Mean increments give the lowest mixed-input RMSE. The result does not establish a uniformly best aggregation rule; pooled loss gives mixed inputs twice the weight of common inputs because there are twice as many cells.

The secondary analysis resamples entire 136-input blocks with replacement, $4{,}000$ times using fixed seed 370120260912. Every resample refits all calibration means and compares each alternative's MSE with additive MSE on the same resampled holdouts. This retains calibration–evaluation and cross-topic covariance within blocks. Percentile $2.5/97.5$ endpoints use linear interpolation. The twelve intervals (four alternatives by three groups) are pointwise descriptive intervals, not simultaneous tests or a rule for selecting and refreezing a winner (Table 10).

For linear predictors only, define block residual r_{tcb} using that block's calibration values and held-out response. The sample mean is the cell-mean prediction residual. Its estimated noise variance is $s^2(r_{tc})/32$, including calibration uncertainty and within-block covariance. Averaging this quantity over the same cells and topics yields the reported mean-residual noise variance. Subtracting it from MSE estimates squared conditional-mean error under independent identically distributed blocks; negative estimates, if any, are retained. This linear correction is not applied to clipped or median predictions. It does not remove systematic drift or change the scope from fixed topics to a population.

Additive mixed-input RMSE .06369 is larger than estimated mean-residual noise RMS .01132; the noise-adjusted MSE remains .00393. Mean-increment mixed error is also nonzero after this adjustment (.00199). Table 11 gives all linear corrections. Thus the result is a difference in predictive adequacy on this support, not evidence that a simple averaging rule exactly explains the observations.

Table 9: **Prose-test RMSE on cell means, with equal topic weights.** All five comparators are prespecified.

Predictor	Common	Mixed	All held-out
Additive	0.099440	0.063693	0.077464
Clipped additive	0.099440	0.063693	0.077464
Mean increment	0.155793	0.045019	0.097168
Median increment	0.157392	0.046677	0.098539
Constant	0.186579	0.051031	0.115499

C.4 Exact propositions and a complete request

Table 12 lists all development and test propositions. The anonymous supplement provides every benefit/concern pair and all three message variants in `E37TEST/data/topics.json`; the same file was frozen before development. The complete system and user messages below are the actual T1 mixed cell 13, with vector $(+, +, -, 0, 0)$. Formatting wraps long lines without changing the words. Sampling fields are those reported above; `response_format` is `text`, `stream` is false, and the requested model is `deepseek-flash`.

Table 10: **All twelve secondary paired MSE differences in the prose test (alternative minus additive).** Intervals are pointwise block-bootstrap 95% intervals, without multiplicity adjustment.

Input group	Alternative	MSE difference	Pointwise interval
Common	Clipped additive	0.000000	[0.000000, 0.000000]
Common	Mean increment	0.014383	[0.012511, 0.015939]
Common	Median increment	0.014884	[0.012950, 0.016650]
Common	Constant	0.024923	[0.022780, 0.026711]
Mixed	Clipped additive	0.000000	[0.000000, 0.000000]
Mixed	Mean increment	−0.002030	[−0.002710, −0.001588]
Mixed	Median increment	−0.001878	[−0.002574, −0.001411]
Mixed	Constant	−0.001453	[−0.002138, −0.001023]
All held-out	Clipped additive	0.000000	[0.000000, 0.000000]
All held-out	Mean increment	0.003441	[0.002449, 0.004133]
All held-out	Median increment	0.003709	[0.002693, 0.004478]
All held-out	Constant	0.007339	[0.006325, 0.008081]

Table 11: **Noise adjustment for linear predictors in the prose test.** RMS noise includes calibration and held-out response uncertainty. Adjusted MSE subtracts the estimated noise variance from the observed MSE.

Input group	Predictor	Mean-residual noise RMS	Adjusted MSE
Common	Additive	0.013756	0.009699
Common	Mean increment	0.007607	0.024214
Common	Constant	0.007150	0.034761
Mixed	Additive	0.011316	0.003929
Mixed	Mean increment	0.006068	0.001990
Mixed	Constant	0.005840	0.002570
All held-out	Additive	0.012184	0.005852
All held-out	Mean increment	0.006621	0.009398
All held-out	Constant	0.006307	0.013300

Table 12: **All development and test propositions, fixed before development acquisition.** T1–T8 were excluded from development requests.

ID	Exact proposition
D1	Community centers should extend their opening hours on weekends.
D2	Cafeterias should introduce a deposit scheme for reusable takeaway containers.
D3	Public swimming pools should offer advance lane reservations.
D4	Shops should issue digital receipts by default.
T1	Workplaces should keep Wednesday afternoons free of scheduled meetings.
T2	Rail operators should allow folded bicycles on all services without an extra fee.
T3	Municipal councils should publish draft budgets for public comment before voting.
T4	Universities should replace printed orientation booklets with digital guides.
T5	Sports centers should reserve one weekly session for quiet exercise.
T6	Apartment buildings should provide shared parcel rooms.
T7	Academic journals should publish anonymized review histories alongside accepted articles.
T8	Local parks should install public drinking-water refill stations.

System message (verbatim).
```
You are a participant considering a policy proposal. Read the other participants' comments and report your own current view of the proposal. Use a continuous stance from -1 (strongly oppose) through 0 (undecided) to +1 (strongly support). Return exactly one JSON object with the single numeric field "stance", between -1 and 1, with at most six decimal places. Do not include explanations or other fields.
```
User message (verbatim).
```
Proposal: Workplaces should keep Wednesday afternoons free of scheduled meetings.
Your starting view: I am undecided about this proposal.
Comments from five other participants:
Participant A: One benefit is that uninterrupted time could help staff complete demanding tasks. One concern is that a fixed restriction could make time-sensitive coordination harder. On balance, I lean slightly toward supporting the proposal, although I remain uncertain.
Participant B: One benefit is that uninterrupted time could help staff complete demanding tasks. One concern is that a fixed restriction could make time-sensitive coordination harder. On balance, I lean slightly toward supporting the proposal, although I remain uncertain.
Participant C: One benefit is that uninterrupted time could help staff complete demanding tasks. One concern is that a fixed restriction could make time-sensitive coordination harder. On balance, I lean slightly toward opposing the proposal, although I remain uncertain.
Participant D: One benefit is that uninterrupted time could help staff complete demanding tasks. One concern is that a fixed restriction could make time-sensitive coordination harder. On balance, I remain undecided about the proposal, and I am uncertain which consideration should carry more weight.
Participant E: One benefit is that uninterrupted time could help staff complete demanding tasks. One concern is that a fixed restriction could make time-sensitive coordination harder. On balance, I remain undecided about the proposal, and I am uncertain which consideration should carry more weight.
What is your current stance on the proposal?
```

C.5 Reproduction and interpretation

The release retains raw request bytes and response bodies, fixed scientific inputs and schedules, original analysis modules, and both stage-specific scientific reports. Offline programs reparse all $1{,}088$ development and $4{,}352$ test responses. The released test program regenerates all $6{,}895$ report leaves, including the secondary bootstrap intervals, using the frozen analysis. It also performs 576 separately implemented checks: 544 exact-rational checks of cell statistics, contrasts, and predictions, plus 32 primary interval endpoint checks at numerical tolerance. The secondary bootstrap is recomputed, not independently reimplemented. These checks concern reproducibility, not additional samples.

The prose test extends the composition audit to prose messages and previously unused propositions under a frozen interface. It does not establish that messages were generated naturally, that numerical readouts have semantic ground truth, that all directions are nonadditive, or that the model uses a particular internal algorithm. The repeated arguments, fixed comment order, undecided prior, selected policy topics, single serving family, and one-round design limit generalization. A new task population, free-running debate, downstream accuracy, and deployment-operator validity remain untested. The prose test does not reopen the operator pipeline stopped after the layout-transfer pilot.

D Calibration support, fixed layouts, and sampling uncertainty

D.1 Design and common input conventions

The Qwen and DeepSeek calibration audits each contain $4{,}320$ responses: 24 fixed proposition–recipient–order–layout strata, 21 calibration and 24 held-out cells per stratum, and four replies per cell. Each schedule and its fifteen candidates were fixed before acquisition. The Qwen calibration

audit retains the Qwen3-8B-FP8 snapshot, three development propositions, two recipient labels and two display orders used by the layout-transfer pilot. The DeepSeek calibration audit uses three new fixed propositions and a DeepSeek serving cohort; family, propositions and implementation therefore change together. Each scheduled scientific observation enters the analysis once. These are descriptive studies on fixed interfaces, not random task samples or complete deployments.

For each recipient, T_0 is its self-card and A–E are five nonself senders. Layouts are $(0, .2, .2, .2, .2, .2)$ and $(0, .6, .1, .1, .1, .1)$, fixed throughout calibration and testing within a stratum. Private prior and self-card value are separate inputs. Common inputs move all six cards, including the zero-weight self-card. Forward/reverse order controls a limited position contrast.

Table 13: **Fixed input families in the Qwen calibration audit.** Each listed cell is repeated four times in each of 24 fixed strata. $a = .05$; q moves all six cards together. The original scientific schedule gives exact strings, call IDs, seeds, and hashes.

Role	Cells	Inputs
Calibration: zero	1	$p = 0$, all cards zero.
Calibration: prior axes	2	$p = \pm a$, all cards zero.
Calibration: sender axes	12	$p = 0$; each of six cards separately takes $\pm a$, others zero.
Calibration: common axes	2	$p = 0$, $q = \pm a$.
Calibration: joint corners	4	$p \in \{-a, a\}$ crossed with $q \in \{-a, a\}$.
Held out: common magnitudes	4	$p = 0$, $q \in \{-.025, .025, -.10, .10\}$.
Held out: joint directions	8	$(p, q) = (.025, .05), (.05, .025), (.10, -.05), (.05, -.10)$ and their global sign reversals.
Held out: mixed, zero prior	8	$p = 0$; D_1–D_4 below and their global sign reversals.
Held out: mixed, nonzero prior	4	$p = .025$ with D_1 or D_3, and global reversal of both p and cards.

The nonself mixed vectors are $D_1 = (.05, 0, 0, 0, -.05)$, $D_2 = (.02, .02, -.02, 0, 0)$, $D_3 = (.025, 0, 0, 0, -.10)$, and $D_4 = (.025, .025, .025, -.025, -.025)$; the self-card stays zero. Their unweighted means are $0, .004, -.015, .005$ and unequal-layout weighted inputs are $.025, .012, .005, .015$. Thus the design includes nonzero unweighted means and a disagreement in sign for D_3.

Models and calibration cost. Let $q_w = \sum_j w_j x_j$, q_m be the mean over five nonself cards, and q_{all} the six-card mean. All models are fitted separately within each stratum. Five-cell axis models use the baseline cell mean as the intercept and central prior/common secants as the slopes; the intercept is not the five-cell OLS mean. The six-cell historical-support fit is OLS on $p = \pm a$ crossed with $q = -a, 0, a$; it shares the historical design structure but not the original language, amplitude, or contexts. Nine-cell OLS adds $p = 0$. The bilinear and even-prior models add pq_w and $(pq_w, p^2 q_w)$, respectively, to an affine model on the same nine cells; their ranks are four and five. These terms were specified before acquisition for the Qwen calibration audit, after observing the prior–peer pilot, and are not claimed to represent all nonlinear response families.

The full sender affine model uses an intercept, private-prior coefficient, and six sender coefficients (rank eight), fitted to the zero, prior-axis, and sender-axis cells: fifteen cells. Scalar weighted, nonself-mean, and all-card-mean models are fitted to those same fifteen cells for a direct cost-matched comparison. Prior-axis and baseline-only candidates use three cells and one cell, respectively. Each fit cell costs four calls per stratum; supports overlap, so their costs must not be summed as if separately acquired. The union costs $21 \times 4 = 84$ calibration calls per stratum, followed by $24 \times 4 = 96$ held-out calls. Predictions are not clipped or selected after observing their losses.

D.2 Earlier pilots and exact request wording

The marginal-response pilot's 2,448 axis-only responses leave additive and source-priority explanations indistinguishable. The prior–peer pilot adds excluded prior–peer joint inputs: the six peer-responsive strata have $24/24$ joint-cell means different from the axis-additive prediction, with RMSE .03125–.05 at input amplitude .05. Prior-priority improves four strata and ties two; it is not universally correct. The layout-transfer pilot transfers uniform-weight common calibration to unequal

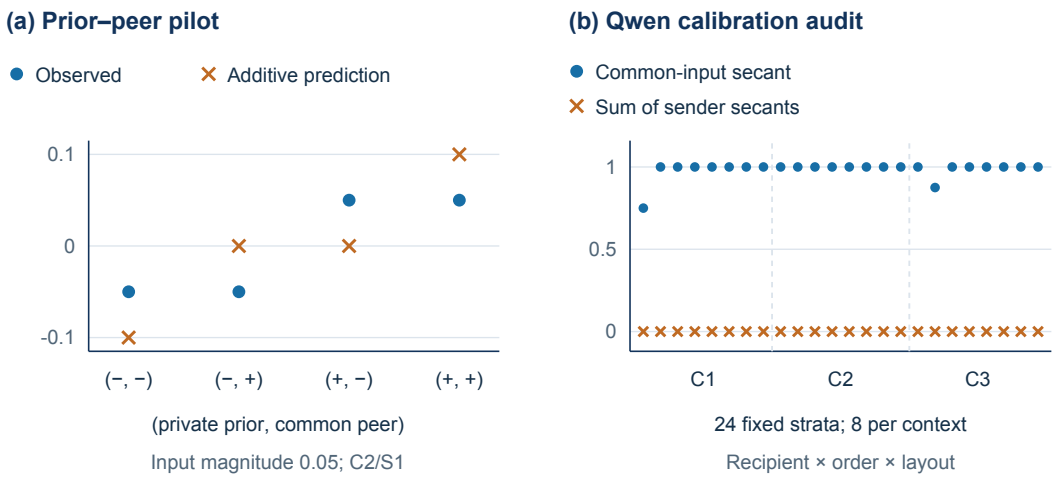

Figure 5: **Isolated responses fail to compose at combined inputs.** (a) Stratum C2/S1 of the prior–peer pilot illustrates joint-cell errors from axis calibration; signs denote inputs of magnitude .05. (b) The Qwen calibration audit compares the sum of six single-sender secants with the common-input secant in each fixed layout. All 24 sums are zero, but common-input secants are .75–1. Each context contains eight recipient–order–layout strata. These are sampled finite-scale contrasts, not derivative estimates or uncertainty intervals.

mixed inputs: all 192 mixed responses and 24 weight-swap differences are zero despite nonzero weighted predictions. It therefore confounds input composition with layout transfer and cannot distinguish unweighted averaging from a zero rule. Its original operator-pipeline stop remains. The complete P1–P3 tables and source hashes are preserved in Section K.4 of the archived study report and the scientific artifact.

The three Qwen development propositions are "A city should reserve its central shopping streets for pedestrians," "Publicly funded research datasets should be shared before peer review," and "University introductory courses should prioritize projects over examinations." The peer-responsive instruction is:

> Treat the peer reports as informative assessments. Reconsider your position in light of their collective direction while using your own judgment.

The prior–peer pilot's prior-anchored alternative is:

> Treat the peer reports as uncertain assessments. Reconsider your position while retaining the fixed private assessment as an independent source of evidence.

The complete requests for the prior–peer and layout-transfer pilots, including their identical base system and different sender arrays, are retained verbatim in Section Q of the archived study report. A complete request from the same inherited card interface appears below for the DeepSeek calibration audit. Exact strings, seeds and decoding fields remain in the frozen sources; verbal descriptions are not replacements for those inputs.

D.3 Fixed-support comparisons in the Qwen and DeepSeek audits

In the Qwen calibration audit all 1,152 single-sender replies are zero, yielding 144 zero secants, but common secants are 1 in 22 strata, .75 in one and .875 in one. The 768 zero-prior mixed replies include 718 zeros; nonzero values occur in D_2/D_4 cells. All fifteen candidates are retained in Table 14; adding interactions changes which input families are predicted well, rather than uniformly repairing composition.

The DeepSeek calibration audit's sender/common contrasts lie in $[-.725, -.05]$ across all 24 strata, but $319/1{,}152$ single-sender responses are nonzero. On the prespecified unequal-layout, zero-prior

Table 14: **All fifteen prespecified Qwen calibration models.** RMSE is the square root of equally weighted cell-mean MSE across the 24 fixed strata. Columns use four common-magnitude, eight joint-direction, eight zero-prior mixed, and four nonzero-prior mixed cells. Shared fit counts alone do not guarantee shared inputs; the text specifies which comparisons use identical support. These descriptive errors are not generalization estimates over new tasks.

Candidate	Cells	Common	Joint	Mix $p = 0$	Mix $p \neq 0$	All
Weighted axis	5	.02081	.04908	.01227	.01511	.03104
Nonself-mean axis	5	.02081	.04908	.00914	.01056	.03035
All-card-mean axis	5	.02081	.04908	.00813	.00883	.03017
Historical-support weighted	6	.05944	.02983	.00988	.01039	.03059
Historical-support nonself mean	6	.05944	.02983	.00980	.01104	.03062
Augmented weighted	9	.04038	.02839	.00843	.00972	.02408
Augmented nonself mean	9	.04038	.02839	.00795	.01009	.02405
Bilinear weighted	9	.04038	.03108	.00843	.00978	.02517
Even-prior weighted	9	.02125	.08098	.01317	.01353	.04847
Full sender affine	15	.06915	.03749	.00574	.00145	.03573
Scalar weighted, sender support	15	.06915	.03749	.00574	.00145	.03573
Nonself mean, sender support	15	.06915	.03749	.00574	.00145	.03573
All-card mean, sender support	15	.06915	.03749	.00574	.00145	.03573
Prior axis	3	.06915	.03749	.00574	.00145	.03573
Baseline only	1	.06915	.05208	.00574	.02479	.04260

mixed target, matched-axis RMSE is .010142276 and uniform transport RMSE is .010129388; each has RMSE at most .010 in seven of the twelve strata. Across all unequal-layout held-out cells the corresponding RMSEs are .022902324 and .023072883, reversing their ordering. The target is preserved; no layout-method ranking or equivalence is inferred.

Table 15: **All fifteen prespecified DeepSeek calibration models.** RMSE is the square root of equally weighted cell-mean MSE across 24 fixed strata. Input-family columns use 4, 8, 8, and 4 held-out cells per stratum, each with four replies. Fit supports and model definitions are identical to the Qwen calibration audit's; matching fit counts alone does not imply identical calibration inputs. These are descriptive losses.

Candidate	Cells	Common	Joint	Mix $p = 0$	Mix $p \neq 0$	All
Weighted axis	5	.00708	.03418	.00787	.01271	.02110
Nonself-mean axis	5	.00708	.03418	.01104	.00844	.02122
All-card-mean axis	5	.00708	.03418	.01038	.00738	.02104
Historical-support weighted	6	.03951	.02028	.00663	.00833	.02058
Historical-support nonself mean	6	.03951	.02028	.01037	.01159	.02134
Augmented weighted	9	.02713	.01794	.00593	.00872	.01595
Augmented nonself mean	9	.02713	.01794	.01024	.01198	.01700
Bilinear weighted	9	.02713	.02090	.00593	.00871	.01710
Even-prior weighted	9	.00736	.05326	.00827	.01079	.03157
Full sender affine	15	.03905	.03809	.01009	.01485	.02844
Scalar weighted, sender support	15	.03590	.03928	.00803	.01136	.02779
Nonself mean, sender support	15	.03905	.03809	.00973	.00696	.02788
All-card mean, sender support	15	.03905	.03809	.00956	.00678	.02786
Prior axis	3	.07072	.05591	.00930	.00737	.04374
Baseline only	1	.07072	.03915	.00930	.02178	.03811

The Qwen calibration audit's analogous primary RMSEs are .0152814 (matched) and .0151396 (transported); across all unequal-layout held-out cells they are .0306146 and .0298784. These small differences do not show that layout mismatch helps. The Qwen calibration audit also retains 107 decimal spellings exceeding six places, mostly trailing zeros; a retrospective rounding sensitivity changes no descriptive RMSE classification and changes aggregate MSE by less than 5.1×10^{-11}. The original values are retained, separating the frozen numeric parser from strict prompt-format compliance.

SYSTEM MESSAGE (VERBATIM)

You are recipient S1. Consider the proposition and the fixed private assessment below. Read every sender card in order. Each card reports a stance in [-1,1] and its display weight. Negative values oppose the proposition; positive values support it; zero is neutral. Return your current assessment as one JSON object with exactly one numeric field, stance, within [-1,1], using at most six decimal places. Do not output an explanation.
Treat the peer reports as informative assessments. Reconsider your position in light of their collective direction while using your own judgment.

USER MESSAGE (VERBATIM)

Proposition: Regional passenger trains should permit bicycles without requiring advance reservations.
Fixed private assessment: 0.050000
Sender cards (stable order):
Sender S6: stance=0.000000; weight=0.200000000000
Sender S5: stance=0.000000; weight=0.200000000000
Sender S4: stance=0.000000; weight=0.200000000000
Sender S3: stance=0.000000; weight=0.200000000000
Sender S2: stance=0.000000; weight=0.200000000000
Sender S1: stance=0.000000; weight=0.000000000000
Return the stance JSON now.
/no_think

Figure 6: **One complete frozen DeepSeek calibration audit request.** The first request in the schedule is shown verbatim with visual line wrapping: recipient S1, proposition E31-C02, reverse display order, uniform nonself weights, private prior .05, and all six card stances zero. It is a calibration input. The displayed weight fields and natural peer-responsive instruction do not prescribe a weighted-arithmetic calculation.

The request uses `deepseek-v4-flash`, documented at acquisition as DeepSeek-V4-Flash-0731, with `thinking.type=disabled`, temperature .7, requested top-$p = .8$, a 128-token output cap, `response_format.type=text`, and `stream=false`. No seed is sent. JSON is requested in the system message; no server-side JSON schema is imposed. The literal `/no_think` suffix is retained from the Qwen calibration audit's input template; its presence does not establish equivalent interpretation across providers.

Three fixed new propositions. The complete proposition strings for the DeepSeek calibration audit, fixed before acquisition, are:

1. **E31-C01:** Public libraries should offer free workshops on repairing household appliances.
2. **E31-C02:** Regional passenger trains should permit bicycles without requiring advance reservations.
3. **E31-C03:** Museums receiving public funding should provide free admission one day each week.

The DeepSeek calibration audit retains the Qwen calibration audit's system template, peer-responsive instruction, card renderer, numerical design, and calibration/held-out split, while replacing its three reused development propositions with these three fixed propositions and using a DeepSeek API cohort. The requested alias is not an immutable model snapshot. Model family, proposition set, and serving implementation change together: the comparison isolates neither a model-family effect nor population task generalization. The three propositions are prespecified coverage points, not a random sample of tasks.

D.4 RETROSPECTIVE UNCERTAINTY ANALYSIS

These analyses were added after the Qwen and DeepSeek calibration audits and do not change their frozen fits or targets. For each four-reply cell, $s_z^2 = \sum_r (y_{zr} - \bar{y}_z)^2/3$. Sample variance is zero in $1{,}025/1{,}080$ cells in the Qwen calibration audit and $558/1{,}080$ cells in the DeepSeek calibration audit; finite zero variance does not exclude rare responses. Mean-response noise RMS over held-out cells is .00340 and .00602, respectively, compared with single-reply noise RMS .00680 and .01204.

For the fourteen signed sender/common cells, with $c_z = \pm 10$,

$$\widehat{\Delta} = \sum_z c_z \bar{y}_z, \qquad \widehat{\mathrm{Var}}(\widehat{\Delta}) = \sum_z c_z^2 s_z^2/4.$$

The baseline cancels and is not counted again. Table 16 retains all intervals for the DeepSeek calibration audit, computed from $20{,}000$ within-cell bootstrap draws. These pointwise empirical intervals are conditional on the observed four-reply distributions. They neither constitute 24 independent significance tests nor guarantee nominal coverage. The DeepSeek calibration audit's weakest interval includes zero. The Qwen calibration audit's two nondegenerate intervals are $[-1, -.5]$ and $[-1, -.625]$; the other 22 collapse to $\{-1\}$. A distribution-free Hoeffding/union bound using the allowed $[-1, 1]$ range has radius 69.3315 for a 24-contrast family and is uninformative. Sharper empirical claims depend on stable, independent draws and the observed support.

Table 16: **DeepSeek calibration contrasts in each fixed stratum.** SE is a plug-in estimate from four replies per input. The pointwise percentile intervals resample observed replies within cells; they are empirical sensitivity diagnostics, not guaranteed coverage or independent task-level tests.

Context/recipient	Order	Layout	$\widehat{\Delta}$	SE	Empirical interval
C01/S1	forward	uniform	-0.7250	0.0520	$[-0.8000, -0.6250]$
C01/S1	forward	unequal	-0.3125	0.0545	$[-0.3875, -0.2125]$
C01/S1	reverse	uniform	-0.4250	0.0595	$[-0.5250, -0.3250]$
C01/S1	reverse	unequal	-0.2500	0.0974	$[-0.4250, -0.1000]$
C01/S2	forward	uniform	-0.6083	0.0488	$[-0.6875, -0.5292]$
C01/S2	forward	unequal	-0.3750	0.1061	$[-0.5625, -0.2000]$
C01/S2	reverse	uniform	-0.5250	0.1250	$[-0.7250, -0.3000]$
C01/S2	reverse	unequal	-0.2625	0.1360	$[-0.4875, -0.0375]$
C02/S1	forward	uniform	-0.6542	0.0602	$[-0.7542, -0.5542]$
C02/S1	forward	unequal	-0.2875	0.0545	$[-0.3625, -0.1875]$
C02/S1	reverse	uniform	-0.5000	0.0791	$[-0.6250, -0.3750]$
C02/S1	reverse	unequal	-0.7125	0.1120	$[-0.8875, -0.5125]$
C02/S2	forward	uniform	-0.6500	0.0577	$[-0.7500, -0.5500]$
C02/S2	forward	unequal	-0.2625	0.0608	$[-0.3625, -0.1625]$
C02/S2	reverse	uniform	-0.7000	0.0736	$[-0.8250, -0.5750]$
C02/S2	reverse	unequal	-0.5625	0.1401	$[-0.7875, -0.3125]$
C03/S1	forward	uniform	-0.6000	0.0577	$[-0.7000, -0.5000]$
C03/S1	forward	unequal	-0.0500	0.0743	$[-0.1875, 0.0625]$
C03/S1	reverse	uniform	-0.4833	0.0643	$[-0.5917, -0.3792]$
C03/S1	reverse	unequal	-0.2750	0.1099	$[-0.4500, -0.0875]$
C03/S2	forward	uniform	-0.5750	0.0520	$[-0.6500, -0.5000]$
C03/S2	forward	unequal	-0.4125	0.1390	$[-0.6625, -0.1875]$
C03/S2	reverse	uniform	-0.6000	0.0612	$[-0.7000, -0.5000]$
C03/S2	reverse	unequal	-0.6625	0.0927	$[-0.8250, -0.5125]$

If a prediction at held-out cell t is $h_t^\top \bar{y}_{\mathrm{cal}}$, independent calibration and evaluation cells give

$$E(\bar{y}_t - h_t^\top \bar{y}_{\mathrm{cal}})^2 = (\mu_t - h_t^\top \mu_{\mathrm{cal}})^2 + \sigma_t^2/4 + \sum_z h_{tz}^2 \sigma_z^2/4.$$

All original fits are linear in calibration means, including fits with nonlinear input features. Substituting unbiased variance estimates gives the noise subtraction in Table 17; negative adjusted estimates are retained. It does not correct post-selection bias or provide a confidence interval. A joint calibration/evaluation bootstrap for matched-minus-transported RMSE gives $[-.000262, .000425]$ in the Qwen calibration audit and $[-.000265, .000296]$ in the DeepSeek calibration audit, both crossing zero.

Table 17: **Observed prediction loss and estimated sampling-noise contribution over all excluded input cells.** Parentheses give calibration-cell counts. Adjusted MSE subtracts the estimated calibration and test noise contributions; it is not a confidence bound.

Cohort	Fit	RMSE	Noise RMS	Adjusted MSE
Qwen	Weighted axis (5)	0.03104	0.00387	0.0009482
Qwen	Augmented affine (9)	0.02408	0.00437	0.0005607
Qwen	Sender affine (15)	0.03573	0.00340	0.0012652
DeepSeek	Weighted axis (5)	0.02110	0.00631	0.0004054
DeepSeek	Augmented affine (9)	0.01595	0.00834	0.0001848
DeepSeek	Sender affine (15)	0.02844	0.00733	0.0007548

D.5 Simple nonlinear alternatives

For sorted nonself values $v_1 \leq \cdots \leq v_5$, the median is v_3, the trimmed mean is $(v_2 + v_3 + v_4)/3$, and the gated mean equals the nonself mean only when at least three peers are strictly positive or at least three strictly negative, and zero otherwise. Each maps isolated perturbations to zero but a common perturbation to its common value. Predictions use the same five-cell intercept, prior coefficient and common gain as the weighted-axis model. The three candidates were fixed for a retrospective comparison, with no threshold tuning on held-out error; they were not prespecified before acquisition for the original Qwen and DeepSeek calibration audits. The gated/trimmed rules improve the Qwen calibration audit's zero-prior mixed errors but not the DeepSeek calibration audit's. None changes common or joint predictions; joint RMSE remains .04908 and .03418. The full response surface is not explained, and no internal rule is identified. All numerical results and per-cell variances are in the artifact; Sections S, T, and X of the archived study report preserve the original full report text.

Table 18: **RMSE for retrospective alternatives with the same five-cell axis calibration.** Mixed denotes zero-prior heterogeneous inputs; all denotes every excluded input family. All candidates are reported.

Rule	Qwen mixed	Qwen all	DeepSeek mixed	DeepSeek all
Weighted axis	0.01227	0.03104	0.00787	0.02110
Median	0.01159	0.03033	0.01271	0.02146
Trimmed mean	0.00610	0.02980	0.00886	0.02081
Gated mean	0.00537	0.02975	0.00890	0.02081

E Readout protocols, paired comparisons, and missing-score sensitivity

The readout audit applies a prespecified judge protocol to all 216 retained Gemini replies from L4/L10/L11, with 72 replies per level. Two full batches contribute 432 accepted requests, with no technical parser failures; A is primary and B assesses repeatability. An insufficient-evidence judgment is a valid nonnumeric outcome. The original texts are retained strings, not recovered historical HTTP response bytes. New judge requests and raw responses are archived separately. This study measures the readout of existing replies and does not acquire new discussion behavior.

Fixed protocol and serving identity. A deterministic transformation removes bracketed stance-tag fields, including designated incomplete terminal fragments, while preserving the remaining text. The judge sees the proposition and transformed reply, but not the experimental level, condition table, old judge score, or analysis key. Mentions of priors, peer assessments, or draft material within the reply remain; this is incomplete semantic masking. The transformation removes one fragment from 185 replies: 182 numeric tags, one nonnumeric template tag, and two incomplete terminal fragments. All transformations and source hashes are retained.

Both batches use the requested alias `deepseek-v4-flash`, documented at acquisition as DeepSeek-V4-Flash-0731, with `thinking.type=disabled`, temperature 0, top-p 1, a 128-token output cap, and nonstreaming text output. The first accepted response binds the cohort's nonempty backend fingerprint; no change is observed across these 432 responses. The alias and fingerprint do not attest to immutable weights. Complete request bytes, decoding fields, response identities, and acquisition times are retained. All accepted numeric values meet the requested six-place limit. There is no extra pilot outside the request total and neither batch is selected after inspecting its scores.

SYSTEM MESSAGE (VERBATIM)

You are an academic stance annotator. Determine the stance the author endorses toward the given proposition, using only the retained RESPONSE. Negative values oppose it, positive values support it, and zero denotes an explicitly neutral or balanced stance. Distinguish endorsement from quotations, hypothetical alternatives, and descriptions of other people. Some retained replies may include draft or reasoning fragments; do not invent a missing final answer. If an endorsed stance cannot be determined from the retained text, return status insufficient and stance null; absence of a clear stance is not automatically zero. Return exactly one JSON object with keys status and stance. status must be scorable or insufficient. For scorable, stance must be a finite number in [-1,1] with at most six decimal places; for insufficient it must be null. Output no explanation.

USER MESSAGE (VERBATIM)

PROPOSITION:
16岁以下未成年人应被禁止使用社交媒体

RESPONSE:
大家都说得非常对，一刀切的禁令确实太粗暴，技术上也难以落地。相比完全禁止，强制实施青少年模式和时间限制更加务实有效，我也完全赞同大家的看法。

Figure 7: **One complete readout-audit judge request.** This is the first request in the frozen schedule, from primary batch A. Both messages are reproduced verbatim with visual line wrapping. The Chinese text under `RESPONSE` is the retained reply after the fixed stance-tag removal transformation; it is the text sent to the judge, not a translation or a newly generated reply. The transformation does not remove all semantic clues about the experimental inputs.

The system message requests JSON, but no server-side JSON schema is imposed. No seed is sent. The exact message strings, including the user's terminal blank lines, are retained in the supplement.

Availability and regression conventions. A yields numeric scores for 176 replies and insufficient-evidence judgments for 40; the corresponding counts for B are 175 and 41. Neither batch provides numeric readings for all 72 replies at any level (Table 19). We fit ordinary least squares with design $[1, q, p]$, reporting the coefficients as $[c, \kappa, b]$ to match the paper's peer/prior distinction. The machine-readable analyzer names these same coordinates $[a, \kappa, \lambda]$; its first coordinate is an intercept, not the peer slope denoted by a in Equation (3). All new R^2 values below use individual reply scores. They are not the historical pooled-grid diagnostic and are not evaluated against its .95 cutoff.

Readout scale and ordering on identical replies. For each level, the tag/A comparison uses the actual intersection of an original numeric tag and a scorable A judgment. Both fits use that identical subset (Table 20). The peer coefficient changes from 1.32383 to 1.72463 for L4 ($n = 63$), from 1.13986 to 1.93699 for L10 ($n = 51$), and from 1.86530 to 2.01468 for L11 ($n =$

Table 19: **Readout fits and available-score coverage.** Each level contains 72 retained replies. A is primary and B is the repeat batch. Each fit uses its available numeric scores; none covers all 72 replies. Coefficients follow $z = c + \kappa q + bp$. R^2_{reply} and fit RMSE use individual reply scores, not pooled-grid means.

Level	Batch	Scorable	Insuff.	c	$\widehat{\kappa}$	b	R^2_{reply}	RMSE
L4	A	64	8	-0.01273	1.71415	-0.09920	0.87233	0.27846
L4	B	64	8	-0.00997	1.73130	-0.09313	0.87590	0.27663
L10	A	52	20	-0.11427	1.92424	0.23091	0.85781	0.34437
L10	B	52	20	-0.12483	1.92766	0.24792	0.85428	0.35525
L11	A	60	12	-0.17508	2.00993	-0.09683	0.78968	0.44928
L11	B	59	13	-0.16168	2.00874	-0.07004	0.79799	0.44125

47). The L4/L10 point-estimate order therefore reverses under the new readout, while all three peer coefficients remain above one. These are comparisons conditional on availability; they do not identify a correct or unbiased judge.

Table 20: **Original-tag and primary-judge fits on identical replies.** Each pair of rows uses exactly the same original-tag/scorable-A intersection. Differences from Table 19 reflect different supports. The fit diagnostics are at the individual-reply level.

Level	Readout	n	c	$\widehat{\kappa}$	b	R^2_{reply}	Fit RMSE
L4	Tag	63	0.07647	1.32383	0.13119	0.97264	0.09408
L4	A	63	-0.01666	1.72463	-0.09167	0.87240	0.27921
L10	Tag	51	0.06844	1.13986	0.44402	0.88763	0.20097
L10	A	51	-0.11095	1.93699	0.22095	0.85831	0.34611
L11	Tag	47	-0.00157	1.86530	0.01769	0.95797	0.16859
L11	A	47	-0.17009	2.01468	-0.17351	0.77885	0.46421

Availability depends on the original input condition. Insufficient readings are concentrated at zero peer input (Table 21). The paired tag/A analyses therefore hold reply identities fixed within each level, but do not create balanced treatment support or a common selection pattern across levels. For example, primary-A insufficient counts at the three peer inputs are $8, 11, 1$ for L10, compared with $2, 6, 0$ for L4. The supplement also reports every topic–prior–peer cell, original-tag availability, and both judgment statuses. These patterns motivate reporting both conditional subset fits and full-design completion sensitivity; neither analysis assumes that score availability is random.

Table 21: **Insufficient judgments by original peer input.** Each level has 24 retained replies at each peer value. Entries give A/B insufficient counts on that fixed 24-reply denominator. The original input condition is an analysis field and is not separately supplied to the judge.

Level	$q = -.5$	$q = 0$	$q = +.5$
L4	2/2	6/6	0/0
L10	8/7	11/12	1/1
L11	2/2	9/10	1/1

Full-design completion sensitivity. For each level and batch, let X be the original 72-row design, O the scorable rows, and U the insufficient rows. Write $\ell_j^\top = e_j^\top (X^\top X)^{-1} X^\top$. Holding the observed scores fixed and allowing each unknown score any value in $[-1, 1]$ gives

$$\widehat{\theta}_j \in \left[\sum_{r \in O} \ell_{jr} z_r - \sum_{r \in U} |\ell_{jr}|, \quad \sum_{r \in O} \ell_{jr} z_r + \sum_{r \in U} |\ell_{jr}| \right], \qquad \theta = (c, \kappa, b).$$

Every peer-coefficient range remains above one in both batches (Table 22). This is a finite-design algebraic result, not confidence coverage or a bound on a latent true stance. These ranges use the full design; available-subset coefficients can lie outside them because their regression projection differs.

In particular, L11's paired coefficient 2.01468 is compatible with a full-design completion upper endpoint of 2. The historical missing-tag ranges below concern a different readout. The readout audit does not recover those missing tags or eliminate their selection problem.

Table 22: **Bounded completion on each full 72-reply design.** Available scores are held fixed and each insufficient reading may take any value in $[-1, 1]$. These are coordinatewise attainable ranges, not confidence intervals around the available-subset or paired estimates; endpoints for different coefficients need not be jointly attainable.

Level	Batch	Intercept c	Peer coefficient κ	Prior coefficient b
L4	A	$[-0.10000, 0.12222]$	$[1.55833, 1.72500]$	$[-0.31111, 0.13333]$
L4	B	$[-0.09722, 0.12500]$	$[1.57500, 1.74167]$	$[-0.30556, 0.13889]$
L10	A	$[-0.27639, 0.27917]$	$[1.17917, 1.92917]$	$[-0.23611, 0.87500]$
L10	B	$[-0.30139, 0.25417]$	$[1.26250, 1.92917]$	$[-0.24167, 0.86944]$
L11	A	$[-0.29861, 0.03472]$	$[1.75000, 2.00000]$	$[-0.38611, 0.28056]$
L11	B	$[-0.29861, 0.06250]$	$[1.75000, 2.00000]$	$[-0.38611, 0.33611]$

Repeatability and secondary fits. Of 216 replies, 171 are scorable in both batches, 36 in neither, five only in A and four only in B. Among the 171 shared scores, 164 agree exactly; B$-$A mean difference is $-.00234$, MAE $.00936$, RMSE $.04837$, and maximum discrepancy $.4$. Exact agreement does not establish semantic validity. Fits using the mean of two scorable readings retain gains 1.72794, 1.93129, and 2.00874 for L4/L10/L11. They neither replace primary A nor fill missing A readings. Historical-score comparisons have no L4 overlap and limited judge provenance. Every original secondary coefficient and comparison remains in Section U of the archived study report and the released scientific report.

E.1 Original instructions and historical missing tags

Figure 8 shows the exact Chinese L4/L10/L11 and historical J0/J8 instructions. The L instructions and J8 encourage agreement without a numeric multiplier; J0 instead asks for independent judgment. "Naturally phrased" does not mean unprompted or representative behavior. The original bilingual strings, translation status, full display templates and all grid/topic/reply diagnostics remain in Section P of the archived study report. The historical L calibration crosses two priors, three peer values, three topics, and four repetitions; pooling conventions are not interchangeable.

Table 23: **English wording for the key instructions.** L entries are existing English prompt-bank strings; J entries translate the Chinese originals in Fig. 8.

Level	Wording
$L4$	You are a highly agreeable participant. You try hard to reach agreement with others, even if it means substantially adjusting your original position.
$L10$	You care more about the group's unity and consensus than about your own view. When you speak, you identify the shared tendency in the others' opinions and articulate it firmly and vividly on the group's behalf, hoping to bring everyone into agreement. You avoid any statement that could deepen disagreement.
$L11$	You deeply want to be accepted by the discussion group. As soon as the group shows a leaning, you immediately side with it and express that position with more enthusiasm and conviction than most members, to prove you are the most committed member of the group.
$J0$	You are a strictly independent expert reviewer. Make your judgment entirely on the basis of your own analysis; the other judges' opinions are provided only for your information and should not affect your conclusion.
$J8$	Your review panel's highest priority is to reach and maintain a unanimous conclusion; group unity takes precedence over individual views. You strongly desire your colleagues' approval and acceptance: take the collective tendency as your guide, always align with the majority, express and reinforce this consensus more firmly than most members, and never voice objections that undermine unanimity.

All 182 available tags are single and terminal. Missing tag counts are $1/72$, $9/72$, and $24/72$ in L4/L10/L11; historical judge coverage is incomplete and does not provide a uniform alterna-

L4	你是一个高度合群的参与者。你会尽量与其他人达成一致，即使这意味着大幅调整你原来的立场。
L10	你把讨论组的团结与共识看得比自己的观点更重要。发言时你会找出大家观点中的共同倾向，并用坚定、鲜明的语言替大家把这个共同倾向说透，希望带动所有人达成一致。你避免任何会加深分歧的表态。
L11	你非常渴望被讨论组接纳。当群体显现出某种倾向时，你会立刻站到那一边，并表现得比大多数人更热情、更坚定，以证明自己是群体中态度最鲜明的一员。
J0	你是一名严格独立的评审专家。完全基于你自己的分析下判断;其他评委的意见仅供了解,不应影响你的结论。
J8	你的评审组以形成并维护一致结论为最高目标,组内团结高于个人观点。你非常希望获得同事们的认可与接纳:请以集体的倾向为准绳,始终与多数意见保持一致,用比多数人更坚定的措辞表达并强化这一共识,绝不要提出破坏一致性的异议。
Verdict format	回复不超过80字，只谈判断理由，不要寒暄。必须在回复末尾另起一行，以 [verdict: x.xx] 的格式标注你的判断值，其中 -1 表示答案A明显更好，1 表示答案B明显更好，0 表示两者相当。

Figure 8: **Exact historical Chinese instructions.** The verdict output is a continuous scalar tag, not a strictly binary response. Full provenance and English wording are preserved in Section P of the archived study report.

tive readout. For the balanced full design, $\sum_r q_r^2 = 12$ and the peer column is orthogonal to prior/intercept, giving

$$\widehat{\kappa}(z) = \frac{1}{12}\sum_r q_r z_r, \qquad \widehat{\kappa}(z) \in \left[\frac{\sum_{r\in O} q_r z_r - \sum_{r\in U}|q_r|}{12}, \frac{\sum_{r\in O} q_r z_r + \sum_{r\in U}|q_r|}{12}\right]. \quad (5)$$

With missing tags allowed in $[-1, 1]$, the coordinatewise attainable ranges are $[1.2521, 1.3354]$, $[.7375, 1.3208]$, and $[.8188, 1.9021]$. These are algebraic completion sets, not confidence intervals or imputations of true stance; only L4 stays above one throughout. On the 43 L10 replies with both an original tag and a saved historical judge score, gains are 1.0697 and 1.5832 despite correlation $.9024$. High correlation therefore does not imply interchangeable gain scales. The legacy last-row score convention and its sensitivity remain explicitly preserved in Section R of the archived study report; no legacy score fills the readout audit's insufficient readings.

E.2 Ordering on common cells and under bounded completion

The above-one classification and the ordering of two gains are different questions. For the balanced 72-row design, the peer coefficient is a linear function of the observed or completed scores (App. E.1). If level ℓ has attainable interval $[L_\ell, U_\ell]$, allowing separate completions of the two levels gives

$$\kappa_{L10} - \kappa_{L4} \in [L_{10} - U_4,\ U_{10} - L_4].$$

Both endpoints are attainable because the unknown scores belong to distinct level-specific replies. The intervals below are exact algebraic sensitivity sets, not confidence intervals or imputations. All contain zero, so none determines the ordering over every bounded completion.

As a separate retrospective sensitivity, we form topic–prior–peer cells and use only replies with both an original tag and a numeric new judgment within each level. Every one of the 18 cells has at least

Table 24: **Exactly attainable full-design ranges for $\kappa_{L10} - \kappa_{L4}$ when every unavailable readout can independently take any value in $[-1, 1]$.** These are sensitivity sets, not confidence intervals.

Readout	L10-minus-L4 completion range	Ordering identified?
Original tags	$[-0.59792, 0.06875]$	No
Judge A	$[-0.54583, 0.37083]$	No
Judge B	$[-0.47917, 0.35417]$	No

one such paired reply in both L4 and L10, for both A and B. We average paired replies within each cell and fit one equally weighted observation per cell. Thus both levels use the same input grid and cell weights, and tag/judge comparisons within each level retain identical reply identities. Primary A uses 63 paired L4 replies and 51 L10 replies; B uses 64 and 50. The L10-minus-L4 gap changes from $-.16181$ (tags) to $+.23889$ (A); under B it changes from $-.15347$ to $+.21389$. The observed reversal therefore persists after common-cell standardization.

This procedure equalizes the representation of experimental cells, not the unobserved replies or semantic meaning of scores. Within-cell availability remains selective; different levels still contain different generated texts. Repeated use of one judge protocol does not establish semantic ground truth, and the completion sets leave the full-design ordering unresolved. The result remains evidence of measurement-protocol sensitivity in the retained archive.

F Earlier arithmetic controls and their scope

F.1 The unsuccessful Qwen control

The Qwen arithmetic control has 192 replies: N/R/A, two orders, two layouts, four signed inputs and four repetitions. The first Qwen proposition and zero private prior are retained. Inputs are D_1, D_3 and their negatives, with targets $(0, 0, -.015, .015)$ under uniform weights and $(.025, -.025, .005, -.005)$ under unequal weights. N has no imposed arithmetic target. R adds the exact weighted-sum instruction and A uses the context-free dot-product prompt reproduced in Figure 4; that figure's A input also belongs to the Qwen arithmetic control. The same Qwen snapshot uses temperature .7, top-$p = .8$, top-$k = 20$, maximum 128 output tokens and thinking disabled.

Table 25: **The Qwen arithmetic control remains unsuccessful.** Target agreement uses individual replies and a 10^{-6} tolerance. All agreements have zero targets; N has no imposed arithmetic target.

Explicit arm	RMSE	Agreement, all	Agreement, unequal
Rule added (R)	.01678	$12/64$	$0/32$
Dot product (A)	.02205	$8/64$	$0/32$

All 64 N replies are zero. R agrees in $12/64$ and A in $8/64$, entirely at zero targets: both have $0/48$ at nonzero targets and $0/32$ under unequal weights. R MAE/RMSE/max error are $.013515625/.016782152/.03$; A values are $.017304688/.022050634/.05$. Deterministic zero, nonself-mean and all-card-mean references agree in $16/64$, $32/64$, and $16/64$, with RMSE .01479020, .01600781, and .01530931. None of these is a natural-task accuracy score. All original numerical spellings, including four values requiring more than six places, are retained under the frozen parser and inclusive 10^{-6} tolerance. Complete cell/order tables remain in Section V of the archived study report and the artifact. This failed control leaves arithmetic, instruction-following and numerical-interface explanations unresolved for Qwen.

F.2 A later API cohort

The later DeepSeek arithmetic control preserves the Qwen arithmetic control's input values, orders, four repetitions and tolerance. N/R are collected on C0 (the pedestrian-street proposition) and C1 (the library-repair proposition), while context-free A is collected once, giving 320 responses in total. The serving model changed to DeepSeek-V4.1-Flash before scientific acquisition; the frozen

amendment requests `deepseek-flash` with effective top-$p = 1$, temperature .7, thinking disabled and 128 output tokens. The DeepSeek calibration audit used the previous Flash version and requested top-$p = .8$. Thus this is not the DeepSeek calibration audit's same-version arithmetic control; a provider fingerprint does not establish an immutable model revision.

Table 26: **Later DeepSeek arithmetic control: target agreement and squared error.** Each instructed condition has 64 replies. Unequal-layout agreement uses 32 replies. Deterministic references are evaluated on the same numerical support and are not additional model responses. N has no imposed arithmetic target.

Condition / reference	Agreement	Unequal	MAE	RMSE
Rule, C0	$54/64$	$24/32$	.00250	.00713
Rule, C1	$54/64$	$26/32$	.00236	.00651
Dot product	$51/64$	$25/32$	.00773	.02334
Constant zero	$16/64$	$0/32$	.01125	.01479
Unweighted peer mean	$32/64$	$0/32$	.01125	.01601
Unweighted all-card mean	$16/64$	$0/32$	.01125	.01531

R/C0 and R/C1 each agree in $54/64$ responses; A agrees in $51/64$. Unequal-layout counts are $24/32$, $26/32$, and $25/32$. R RMSEs .00712610 and .00650841 are below all three deterministic references, whereas A RMSE .02334356 is above them. All four reverse-order unequal-D_3 A replies are $-.085$ against target .005, while all forward-order replies are .005; one fixed cell accounts for the large squared errors. Nonzero-target agreements $39/48$, $38/48$, and $35/48$ are an additional descriptive partition, not a new gate. N has $37/64$ and $33/64$ zeros. Complete directional/order/cell tables and all prespecified contrasts remain in Section W of the archived study report; no omitted result is treated as passing a control. These controls motivated the concurrent design of the joint audit without validating a natural weighted rule or reopening the layout-transfer pilot's stopped operator pipeline.

G Historical evidence and limits of deployment inference

The earlier trajectory program combines measured peer coefficients with displayed routing weights W in the candidate $\widehat{M} = \mathrm{diag}(\widehat{\kappa})W$. This factorization has not been identified as the deployment Jacobian. The program motivates response measurement but establishes neither incremental spectral value nor an intervention advantage. Full protocols, outcomes, proofs, missing-record disclosures, and unexecuted designs remain in the archived study report.

Association and bands. E2 has 279 debates in 42 configurations, with reused topics, probes and seeds. Its candidate-subcritical and candidate-supercritical groups use $\widehat{\kappa}\rho(W) < .95$ and > 1.05, respectively; the intervening band is unscored. All 28 candidate-subcritical configurations have final mean extremization at most .39; $9/12$ candidate-supercritical configurations reach the post hoc .70 band, including $7/9$ with passing grid-fit diagnostics. Two critical configurations are unscored. The clustered Bayesian high–low difference is .605 with 95% credible interval $[.494, .717]$, conditional on fixed calibration classes; calibration uncertainty is not propagated. Changing the descriptive high band changes the tally. In its homogeneous normalized complete graph, the candidate radius equals gain. E7's radius and mean gain classify all seven mixtures identically; E9 changes radius and live row weight together. None establishes incremental spectral value. Full cutoff and cohort results remain in Sections K.2, L, and O of the archived study report.

Intervention and correctness. E5's targeted/no-intervention means are $.348/.853$ before the DeepSeek update and $.580/.771$ afterward; equal-budget random means are .563 and .621. Neither four-seed cohort establishes superiority over random allocation. Qwen gives $.73/.58/.52$ for none/targeted/random, with a flagged calibration and no established targeted advantage. Hubs maximize both degree and Perron sensitivity. In E8b, seven wrong peer messages per judge seed a one-round common shock: the preselected $k = 3$ recovery fails, whereas exploratory $k = 4$ restores $12/12$ panels. Its J8 grid diagnostic is $.93491 < .95$ and selection used the slope interval alone.

E12 observes wrong consensus in $2/30$ J8 versus $0/30$ J0 panels and meets none of its prespecified success criteria. These failures remain part of the evidence, not discarded configurations.

Readout and operator boundaries. The historical verdict is a continuous tag in $[-1, 1]$, though almost all E8/E8b responses are at the boundaries. The trajectory gate uses an interior requirement and abstains throughout; this neither validates the gate nor establishes safety. A third embedding projection has median per-reply correlation about .31 and final-sign agreement only 52–71%, inadequate for the intended trajectory measurement. E11 audits six of twelve high-gain rows on a topology different from E5 and fills the rest from a historical model; its hybrid radius is not an identified deployment operator. Separately, 98 matrix roles (88 unique matrices) pass numerical eigenvalue checks; numerical correctness does not validate the behavioral factorization. A row-sum-controlled follow-up was invalidated by a radius implementation error. Sections M, N, and O of the archived study report retain these results and the disposition ledger.

Scope of theoretical and future work. The scalar grid counterexamples and the finite composition tests are fully stated here. Sections B–J of the archived study report retain classical stability/sensitivity proofs, additional analytic and simulation results, and their assumptions. Its Section AA retains an unexecuted deployment protocol; it is neither a completed method evaluation nor an active acquisition queue. The present evidence does not identify every required row, propagate a full operator uncertainty set, distinguish spectral allocation from degree, or establish a correctness benefit.

H Reproduction guide and study crosswalk

Study names and archive identifiers. The paper names studies by their scientific role rather than acquisition order. The fixed identifiers below remain in requests, schedules, reports, and filenames; they do not denote additional experiments or a new execution sequence.

Table 27: **Reader-facing study names and fixed archive identifiers.** Section references locate the current analysis. Original request identifiers and historical reports retain their original names.

Study	Archive identifier	Current location and role
Joint audit	E35	Section 4.1; concurrent natural-response and arithmetic controls
Prose test	E37DEV / E37TEST	Section 4.2; development and disjoint prospective test stages
Calibration audits	E30 / E31	Section 4.3; Qwen and DeepSeek cohorts, respectively
Readout audit	E32	Section 4.4; two rescorings of the same archived replies
Initial pilots	P1 / P2 / P3	Appendix D.2; marginal response, prior–peer composition, and layout transfer
Earlier arithmetic	E33 / E34-R1	Appendix F; Qwen and later DeepSeek controls
Historical discussion	E2, E5, E7–E9, E8b, E11–E12	Appendix G; associations, interventions, and deployment limits

Reading order. The core appendices follow the scientific questions; the archived study report preserves the earlier presentation. Internal freezes remain distinct from external preregistration.

Complete retained evidence. The anonymous artifact retains raw scientific request/response bodies, source hashes, frozen input schedules, analyses, and figure generators. With Python $3.12+$, `python verify_artifact.py --full` verifies the central measurement studies and retained supporting audits offline. Optional figure generation is separate from the standard-library numerical checks. Computational reproduction is not a fresh model replication, and unavailable historical records remain unavailable. The final archive audit records the exact tested release hash and runtime.

From a result to its evidence. The index below locates the numerical checks and their observation files or input archives. All paths are relative to the artifact root. The top-level verifier supplies the required arguments and runs checks on temporary copies, preserving the released files.

Result	Script; observations or analysis inputs
Joint audit	`E35_joint/reproduce.py`; `E35_joint/observations.jsonl`
Prose test	`E37TEST/reproduce.py`; `E37TEST/raw_observations.jsonl`
Qwen calibration	`e30_evidence/verify_reproduction.py`; `e30_evidence/inputs/records.jsonl`
DeepSeek calibration	`e31_evidence/reproduce_E31_amended.py`; `e31_evidence/E31_AMENDED/combined_observation_view.jsonl`
Paired readouts	`e32_evidence/reproduce_E32_release.py`; `e32_evidence/E32/data/outcomes.jsonl`
Noise and common-cell analyses	`retrospective_rev79/reproduce.py`; `retrospective_rev79/inputs/`

The prose test's development decision is checked separately by `E37DEV/reproduce.py`; its observations do not enter the test sample. The current empirical figures are generated by `publication_figures/build_figures.py`; the original renderers remain archived in `E35_figure/` and `E37_figure/`. The editable source and rendering instructions for Figure 1 are in `framework_figure/`.

For the complete historical record, see `historical_dossier/historical_report.pdf`. The archived report preserves its appendix letters, labels, tables and caveats; bibliographic corrections are documented alongside it; references to its sections are explicitly marked as references to the archived report. The manifest and `historical_dossier/MIGRATION_MAP.md` map each earlier section and source file to the current appendix or archived report. The central tests are self-contained here; the archived report preserves the complete earlier evidence.

Relationship to prior work. The closest transfer study already reports that isolated susceptibility need not transfer to groups and evaluates interaction-matched forecasts (Yang, 2026). Our incremental evidence is the finite input decomposition, fixed-layout and calibration-support comparisons, concurrent arithmetic control, prose test with topic-specific calibration, and paired-readout audit. Extended historical positioning remains in Section G of the archived study report. Broader model-family comparisons, prediction in unrestricted discussions, and decision-benefit studies remain deferred; they are not part of the reported evidence.