# How Much Independent Information Does an LLM Committee Provide?

**Anonymous authors**
Paper under double-blind review

## Abstract

Adding agents to an LLM committee can improve accuracy without providing proportionally more independent information. Member accuracy, error dependence, and aggregation quality change together, obscuring what a larger or more communicative committee contributes. We study this problem with a committee information audit. End-task accuracy alone cannot tell whether a gain comes from better individual answers, less redundant errors, or better selection of existing answers. We therefore separate three quantities on frozen output records within each protocol arm: dependence among member outputs, complementarity available in that arm's candidate answers, and the gain recovered by an aggregator. This separation distinguishes information that is present in the committee from information that a deployable policy can actually use. The audit combines explicit evidence allocation, matched communication contrasts, and held-out evaluation of fixed-candidate selectors. Across controlled domains, held-out template families, and two external benchmarks, variance-equivalent effective size shows saturation. In a paired Qwen3-4B full-topology control, the observed decay is steeper than a fixed-correlation null. Five related diagnostics show the same directional pattern on the shared output matrix. Conditioning on input structure explains much of the unconditional dependence, while quality-matched model diversity does not consistently reduce it. In the original output-budget-matched Qwen comparisons, peer discussion increased plurality accuracy over self-revision by 5.4 percentage points on HotpotQA and 7.4 on MuSiQue, while increasing error correlation and reducing candidate-oracle coverage. In the stricter input-token-matched Qwen validation, the peer–self plurality changes were $+0.4$ pp (95% CI $[-2.4,\ 3.0]$) on HotpotQA and $+4.0$ pp (95% CI $[2.0,\ 6.0]$) on MuSiQue; fixed- candidate oracle coverage fell by 15.2 and 10.6 pp, respectively. Grouped validation further shows that weighting and learned selection recover useful but uneven fractions of the available complementarity. Thus, communication can improve an endpoint score while increasing error correlation and reducing the fraction of questions with at least one correct candidate. Committee evaluation should therefore report dependence, candidate coverage, and realized gain together rather than use accuracy as a proxy for all three.

## 1 Introduction

Multi-agent LLM systems use debate, verification, and voting to improve decisions by combining several model outputs (Du et al., 2024; Liang et al., 2024; Li et al., 2024a). Increasing the number of agents or mixing model families offers more candidate answers, but the value of those answers depends on how their errors overlap. Classical ensemble analysis relates aggregation gains to member accuracy and joint errors (de Condorcet, 1785; Krogh & Vedelsby, 1994; Kuncheva & Whitaker, 2003); N-version programming similarly studies the benefits and limits of software diversity (Avizienis, 1985). For LLM teams, Bertalanič & Fortuna (2026) connect correlation, effective team size, communication, and saturation.

An LLM committee also changes its own dependence structure. Members receive overlapping evidence, exchange answers and justifications, and revise their outputs. Communication can help agents find a correct answer while making their remaining errors more alike. Accuracy alone therefore cannot reveal whether a protocol adds independent information, concentrates existing information, or



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.

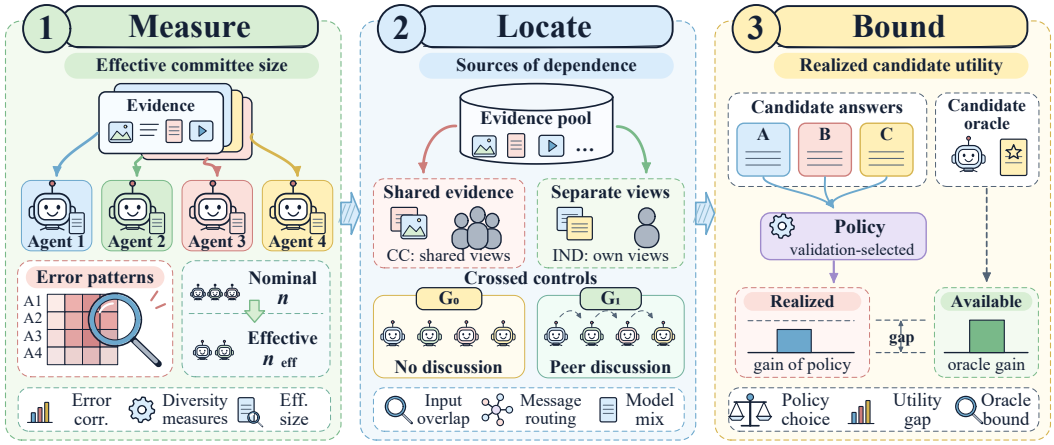



Figure 1: **Measure independence, locate dependence, assess utilization.** The audit links effective size, dependence controls, and realized gain. CC/IND controls and the fixed-candidate oracle are schematic; bars are illustrative.

improves its aggregation. A higher score may come from stronger individual members, less overlapping errors, or a better rule for choosing among unchanged candidate answers; these explanations imply different conclusions about whether adding members or changing the protocol will help. We ask: *how much independent and usable information does an LLM committee provide, and how does its protocol change the answer?*

We address this question with a *Committee Information Audit* (Figure 1). On each frozen item record it asks three distinct questions: (i) are members repeating the same errors, (ii) when the baseline is wrong, is another fixed candidate correct, and (iii) how much of that opportunity does a deployable policy recover? The first is summarized by variance-equivalent effective size and complementary dependence diagnostics; the second by the candidate-oracle gap; and the third by held-out policy gain and utilization. Separating them identifies whether the bottleneck is redundancy, candidate coverage, or selection. Our contribution is a matched, item-level audit that makes these measurements comparable while evidence allocation and communication vary.

Three findings organize the evidence. First, variance-equivalent effective size shows a saturation pattern across the tested controlled and external settings. A paired Qwen3-4B full-topology control shows decay beyond a fixed-$\bar{\rho}$ null. Agreement among five related diagnostics is treated as a convergent sensitivity check on the same output matrix, rather than as five independent tests. Second, dependence has multiple sources. Input-structural conditioning explains much of the unconditional association, and quality-matched family diversity provides no consistent reduction. In the original output-budget-matched Qwen comparisons, peer discussion improves plurality accuracy over self-revision by 5.4 percentage points on HotpotQA and 7.4 on MuSiQue, while increasing error correlation and reducing candidate-oracle coverage. Under strict input-token matching, the corresponding Qwen changes are $+0.4$ pp (HotpotQA; 95% CI $[-2.4, 3.0]$) and $+4.0$ pp (MuSiQue; 95% CI $[2.0, 6.0]$), while fixed-candidate oracle coverage falls by 15.2 and 10.6 pp. Third, weighting and learned selection yield positive average gains under grouped validation, yet their recovery of available complementarity varies across domains and protocols.

The audit applies to fixed member outputs. Its oracle $o = \mathbb{E}_t[\max_i c_{it}]$ bounds selection among those answers; new-answer synthesis and further interaction define different candidate sets. Source contrasts require paired condition IDs and evidence mappings. The study distinguishes the original pre-committed experiments, subsequent post-hoc analyses, and a follow-up that excludes recorded historical questions. It reports failed predictions and provenance boundaries alongside the results (§4.6, Appendix K).

## 2 RELATED WORK

**Multi-agent LLM systems.** Debate, role specialization, and orchestrated conversation aim to improve end-task accuracy (Du et al., 2024; Liang et al., 2024; Smit et al., 2024; Chan et al., 2023; Wu et al., 2023; Li et al., 2023; Hong et al., 2024). AgentForest studies scaling through sampling and voting (Li et al., 2024a); evaluation frameworks coordinate LLM judges (Chan et al., 2023; Zheng et al., 2023); orchestration systems optimize communication (Zhuge et al., 2024; Zhang et al., 2025). Sparse debate topology is a direct precedent for changing interaction structure (Li et al., 2024b). We examine these design choices through their effects on dependence and on available versus realized gain.

**Ensemble diversity and correlated errors.** The relationship between correlation and effective ensemble size is classical (de Condorcet, 1785; Krogh & Vedelsby, 1994; Kuncheva & Whitaker, 2003; Dietterich, 2000). Diversity measures, accuracy tradeoffs, and software diversity are also well studied (Kuncheva & Whitaker, 2003; Margineantu & Dietterich, 1997; Avizienis, 1985). The closest LLM-team analysis is Bertalanič & Fortuna (2026), which studies effective team size at both the answer and correctness levels, communication-induced dependence, heterogeneity, and alternative aggregation rules. We extend this empirical perspective with explicit evidence-allocation controls, input-structural permutation nulls, quality-matched composition contrasts, and oracle-normalized gains under held-out policy evaluation. Their heterogeneity results concern fitted scaling ceilings; ours concern error correlation in selected committees. Likewise, our one-round communication contrasts estimate a finite-stage change rather than an asymptotic ceiling parameter. In concurrent work, Li et al. (2026) analyze evidence partitioning under independence assumptions on private error components and evaluate deliberation for forecasting. We examine evidence allocation empirically in controlled tasks and QA, together with effective-size scaling and fixed-candidate utilization.

The supplement places this paper alongside the two closest LLM-team analyses by measurement scope; moving that scope comparison out of the main text lets the paper foreground measured outcomes. Our contribution is a matched, item-level audit that jointly evaluates dependence, candidate coverage, and held-out policy recovery as evidence allocation and communication vary. The final outputs of each condition are frozen before these quantities are computed, so the audit can distinguish a change in the available correct alternatives from a change in dependence or in the gain recovered by selection.

**Communication and network dependence.** Social-learning theory studies herding, cascades, and convergence when agents observe one another (Banerjee, 1992; Bikhchandani et al., 1992; DeGroot, 1974; Golub & Jackson, 2010). Network structure affects whether dispersed information can be aggregated (Golub & Jackson, 2010; Hong & Page, 2004). These models motivate treating communication as a source of dependence. Our matched-input LLM comparisons measure changes in error correlation and accuracy after one full-mesh discussion round (§4.3).

**Aggregation and expert combination.** Weighted voting, online expert aggregation, stacking, ensemble selection, and mixture models provide established ways to combine predictions (Littlestone & Warmuth, 1994; Freund & Schapire, 1997; Cesa-Bianchi & Lugosi, 2006; Wolpert, 1992; Margineantu & Dietterich, 1997; Jacobs et al., 1991). LLM ensembles use learned routers (Tekin et al., 2025) and stronger models as judges (Zheng et al., 2023; Chan et al., 2023); debate diagnostics also distinguish judge scores from correctness (Keramati et al., 2026). Our main comparison evaluates four fixed-candidate selection methods using grouped nested validation and frozen-policy external follow-up. The supplement also reports broader historical aggregator and judge evaluations, with each result tied to its available provenance.

## 3 MEASURING INFORMATION IN AN LLM COMMITTEE

The audit takes a frozen item-by-member record: item and condition IDs, member answers and correctness, model family and composition, evidence-view hashes, policy outputs, and training, validation, and test group IDs. Message traces are optional. Measure computes dependence summaries and nulls; Locate reports matched or conditional contrasts; Bound compares oracle coverage, baseline accuracy, policy gain, and utilization. Unavailable or degenerate inputs are marked not estimable.

All candidate-selection comparisons are conditional on the recorded answers. For each communication arm, we freeze the final candidate answers before computing dependence, oracle coverage, and policy gain. Arms may produce different candidates; all selectors within an arm receive the same fixed set.

**Committee testbed.** Our testbed fixes a question set and an evidence pool per question, then instantiates committees by (i) allocating each agent a subset of the evidence, and (ii) running a communication protocol in which agents may exchange messages before answering. Two protocol components matter throughout: the *evidence allocation*, which determines what each agent sees, and the *communication structure*, which determines whose answers and justifications each agent observes before committing. Both are experimental controls in our setting, and both are design choices in deployed systems. Models are drawn from four families (Qwen3, Mistral, Gemma, Phi) at 4B–14B scale (Yang et al., 2025; Jiang et al., 2023; Gemma Team, 2025), plus 32B and 8B judges in the historical supplementary evaluation; committees are homogeneous or heterogeneous, with $n \in \{1, 2, 4, 8\}$.

**Effective independent information.** The per-item correctness of agent $i$ on question $t$ is a binary variable; the committee's joint correctness distribution is what matters. Following the equicorrelated special case of standard effective-sample-size analysis (Kish, 1965), we summarize pairwise error dependence by the mean error correlation $\bar{\rho}$ and report the effective committee size $n_{\text{eff}} = n/(1 + (n-1)\bar{\rho})$. This is a variance-equivalent size for standardized errors, not a Shannon-information estimate or a count of independent reasoners; zero pairwise correlation does not imply joint independence. If agents fail independently, $\bar{\rho} = 0$ and $n_{\text{eff}} = n$; if they fail together, $\bar{\rho} \to 1$ and $n_{\text{eff}} \to 1$. Negative error correlation can yield $n_{\text{eff}} > n$: under this variance-based ESS summary, anti-correlated members can be more efficient in variance terms than $n$ independent copies. To examine features beyond mean pairwise correlation, §4.2 also reports a cofailure-based effective number, the participation ratio of the error covariance, its spectral effective rank, and an answer-entropy effective number. Each is calibrated against within-committee permutation nulls that preserve each agent's marginal accuracy while breaking cross-agent item alignment. These original nulls do not condition on item difficulty. We additionally permute within input-defined structural strata (Appendix R); the five summaries are complementary diagnostics, not identical estimands. Because effective size summarizes redundancy rather than utility, it must be read alongside candidate coverage and policy gain below.

**Available and realized gain.** Let $c_{it}$ indicate a correct final answer from seat $i$ on item $t$. On a fixed evaluation set, $b = \max_i \mathbb{E}_t[c_{it}]$ is the retrospectively best seat, not a deployable selection rule, and $o = \mathbb{E}_t[\max_i c_{it}]$ is the label-reading oracle. Available complementarity is $g = o - b$, realized gain is $a_\pi - b$, and utilization is $u_\pi = (a_\pi - b)/g$ for $g > 0$. This separates two bottlenecks: a small $g$ means that the committee offers few recoverable alternatives, whereas a large $g$ with a small realized gain means that the alternatives exist but the selector fails to find them. For any selector restricted to these fixed candidate answers, $c_{\pi(t),t} \leq \max_i c_{it}$ on every item, so $a_\pi \leq o$ and $u_\pi \leq 1$; negative utilization means worse accuracy than $b$. This bound applies to selectors over fixed candidate answers; synthesis of new answers or evidence, or further interaction, defines a different candidate set. For held-out policy evaluation, we additionally define $b_{\text{val}}$ as the held-out accuracy of a seat selected on validation data, not on test labels. Internally this is an out-of-fold policy: the selected seat may differ across outer folds. We report $\Delta_{\text{val}} = a_\pi - b_{\text{val}}$ and $U_{\text{val}} = (a_\pi - b_{\text{val}})/(o - b_{\text{val}})$ where the denominator is positive. The retrospective $b$ and its utilization remain diagnostic references, not competing deployable policies. Main summaries retain all 66 internal evaluation cells for accuracy and gain; thresholded utilization is a sensitivity analysis rather than a filter defining the main accuracy comparison. Means of ratios and ratios of pooled gains are distinct quantities and are not interchanged.

Figure 2 separates generation from audit. The original experiments freeze candidates after the assigned protocol; the external follow-up freezes each communication arm generated from a shared initial state. Aggregation is evaluated within each resulting candidate set.

**Pre-commitment.** The original hypotheses, metrics, and success rules were frozen in version-controlled documents before the corresponding evaluation; the results report both confirmations and failures (§4.6). The subsequent audit is post hoc (Appendix Q).

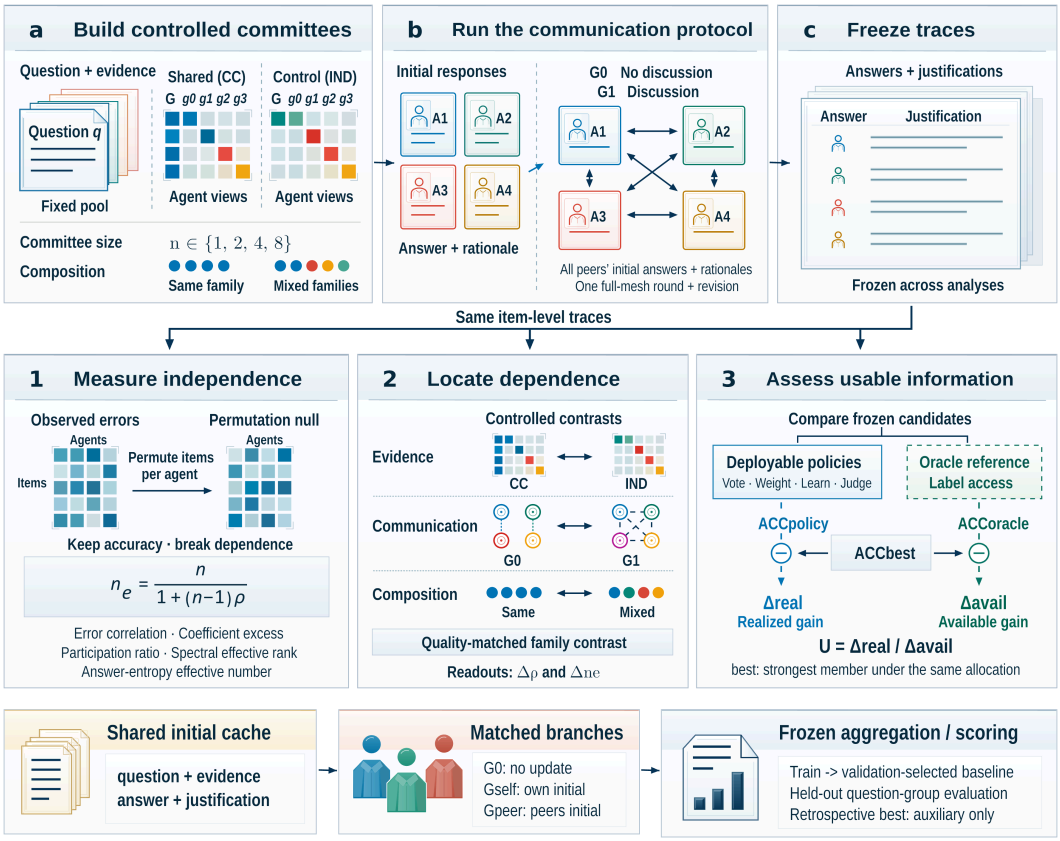



Figure 2: **An audit of fixed committee outputs.** Top: allocate evidence and composition, run the protocol, and freeze traces. Middle: reuse the same records to measure independence, dependence, and usable information. Bottom: matched no-update, self-revision, and peer arms. The oracle applies only to the fixed candidate set.

# 4 Results

## 4.1 Experimental setup

**Tasks.** Three controlled domains instantiate the mechanisms by which committees share information: direct evidence integration (A), two-hop joins (B), and conflicting partially shared reports (C); each item carries an explicit evidence allocation across 8 seats with shared-evidence (CC) and designed-independent (IND) variants. Domain language is realized by template families; three held-out families (F2–F4), generated under latent-structure-equivalence audits, test transfer to unseen surface forms. For external validity we use HotpotQA (Yang et al., 2018) (distractor setting) and MuSiQue (Trivedi et al., 2022) (2–4-hop): each question's own gold and distractor paragraphs are partitioned across agents. The committee uses only benchmark-provided text, and the single-agent condition retains the original full context.

**Committees.** $n \in \{1, 2, 4, 8\}$; homogeneous (one model) or heterogeneous (four families, seat order fixed); one justification-plus-answer round for all agents, an optional full-mesh discussion round, greedy decoding, batch 48 (24 for heterogeneous long-context external cells). 148 internal cells over 1,800 items; 500-item external cells for three model families per benchmark; the grouped internal aggregation audit covers 66 cells. The follow-up compares four-seat Qwen committees on 500 questions per external benchmark, with no discussion ($G_0$), self-revision ($G_{\text{self}}$), and peer discussion ($G_{\text{peer}}$). The latter two share the initial answers and evidence allocation, distinguishing peer information from another revision opportunity. The follow-up excludes questions found in the available historical ID and content-hash records. These records are incomplete, so exclusion does not establish global non-exposure or pretraining independence. Parsing and invalid-output diagnostics are reported in the supplement.

**Statistics.** Original trajectory intervals use paired item bootstrap (2,000 replicates). New matched communication intervals use 2,000 paired-item draws; the stronger null uses 1,999 within-stratum permutations. External communication intervals use 2,000 paired native-question draws within benchmark, shared across arms; these are pointwise percentile 95% intervals, without multiplicity adjustment. Internal aggregation fitting uses five outer and three inner group folds with three declared seeds. Grouping by domain and latent item suffix keeps repeated realizations together: 375 groups underlie 8,250 cell–item rows across 66 cells, not 8,250 independent questions. Model selection and baseline seat selection exclude outer-test labels. The main internal summary is domain-balanced over all cells, with seed repeats not treated as new observations. External policies and validation-selected seats are frozen using historical train/validation records before follow-up scoring, without test refitting. Their uncertainty is conditional on those frozen policies. Historical S1 folds remain unavailable, and S3 reuses cells; those evaluations are not pooled with the grouped validation results (Appendix S).

### 4.2 Effective information saturates with committee size

At a fixed positive error correlation, $n_{\text{eff}}/n$ decreases with $n$ by construction. The empirical question is whether the observed trajectories also show changing error dependence after this mechanical component is removed. We therefore compare the observed curve with a paired fixed-$\bar{\rho}$ null anchored at the $n = 2$ correlation (Appendix A.1). Across the three controlled domains, per-agent effective size ($n_{\text{eff}}/n$) falls as $n$ goes from 2 to 8 in all six domain-by-composition trajectories. Paired item-bootstrap intervals exclude zero in five, and the pooled slope is $-0.071$ with interval $[-0.080, -0.061]$. In the Qwen3-4B full-topology trace, the observed $n_{\text{eff}}/n$ is $0.666, 0.442, 0.332, 0.243, 0.183$ for $n = 2, 3, 4, 6, 8$; the fixed-$\bar{\rho}_2 = 0.5004$ null predicts $0.666, 0.500, 0.400, 0.286, 0.222$. Observed-minus-null intervals are $[-0.082, -0.035]$, $[-0.090, -0.048]$, $[-0.062, -0.026]$, and $[-0.055, -0.025]$ for $n = 3, 4, 6, 8$, respectively. Thus the saturation pattern is not explained by the fixed-correlation formula alone; removing this mechanical $n$-dependence prevents the ESS formula itself from being mistaken for empirical saturation. At $n$=8, effective committee size is typically between 2 and 3, so adding seats supplies diminishing variance-equivalent information.

The pattern extends beyond the correlation summary and the original template language. All five per-agent diagnostics decline in every shared-evidence track (30/30 metric–track comparisons; these are convergent sensitivity checks on one output matrix; Figure 3). The original unconditional permutation nulls yield dependence-excess $z > 1.96$ in 25 of 27 cells. Three held-out template families re-express the same latent items: all 18 family-by-domain-by-composition trajectories decline. One family-domain case lies exactly at the pre-committed 0.75 threshold for a separate level-based check (Appendix H).

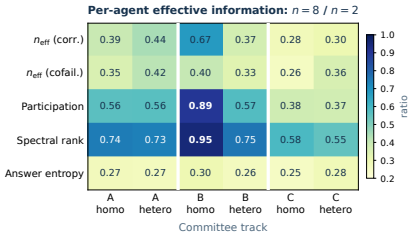
Saturation coexists with positive candidate-oracle gains: committees can add correct alternatives while increasing redundancy. A single accuracy curve would conflate these effects, whereas the audit keeps candidate coverage separate from the per-agent information ratio.

### 4.3 Where the dependence comes from

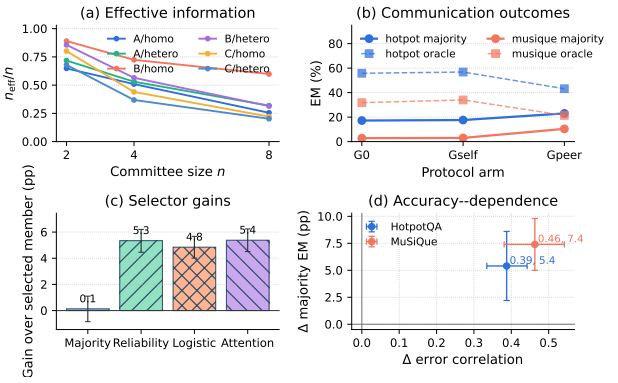
Model similarity and shared protocols can both contribute to dependence; family diversity alone does not separate them.

**Model composition.** Family diversity shows no consistent reduction. For heterogeneous committees drawn from four families, the pooled paired contrast gives $\Delta\bar{\rho} = +0.056$ $[0.012, 0.103]$ and $\Delta n_{\text{eff}} = -0.25$ $[-0.44, -0.08]$. Quality matching limits observed differences in single-agent accuracy: the maximum absolute deviation from the selected models' mean accuracy is 0.023 (Appendix L). Against the homogeneous committee with the lowest observed error correlation, the domain-B interval includes zero ($[-0.004, 0.271]$), while domain C is higher ($[0.019, 0.258]$). Matching and comparator selection use evaluation items, and these intervals condition on that selection; they establish neither equivalence in B nor the absence of residual capability or model-similarity effects.

**Input structure.** Shared evidence and common item difficulty can both correlate errors. Across 66 cells, a post-hoc null conditioned on template, contamination, and input-evidence support re-




**(a) Metric robustness.** Ratios compare $n = 8$ with $n = 2$ across six frozen committee tracks.




**(b) Four diagnostic views.** Effective information, communication-arm outcomes, held-out selector gains, and paired accuracy–dependence changes.

Figure 3: **Robustness and communication effects on frozen committee outputs.** Panel (a) retains the original diagnostic heatmap. Panel (b) replaces the previous communication-only panel with four complementary views. Axes denote committee size or protocol arm, not training steps.

duces mean excess correlation from 0.276 to 0.065. Residual excess is 0.070/0.006/0.118 in A/B/C; domain B provides little residual evidence. Finer strata are too sparse for strong absence claims (Appendix R). The remaining excess is conditional on the observed input structure; unmeasured difficulty may still contribute.

**Communication.** Matched-input contrasts reveal simultaneous changes in accuracy and dependence. Matched initial answers and evidence remove changes in the starting committee as an explanation for the paired differences. In domain C (CC), homogeneous Qwen3-4B committees with $n = 4$, the two base-evidence comparisons have identical initial answers and justifications on all 125 items. With no bulletin intervention, discussion changes $\bar{\rho}$ from $-0.032$ to $0.425$ ($\Delta = 0.457$, paired 95% interval $[0.358, 0.556]$) while majority accuracy rises from 44.0% to 66.4% ($\Delta = +22.4$ pp, 95% CI $[15.2, 30.4]$). With the bulletin intervention, $\Delta\bar{\rho} = 0.422$ ($[0.302, 0.536]$) and accuracy improves by 18.4 pp ($[11.2, 27.2]$). These controlled frozen-trace comparisons do not randomize treatment assignment. $G$-dependent supplement seeds also change the realized inputs, so the refined-evidence average is excluded from the mechanism estimate (Appendix M).

**Peer discussion versus self-revision.** The follow-up reuses each seat's initial answer and justification across no-update ($G_0$), self-revision ($G_{\text{self}}$), and peer-update ($G_{\text{peer}}$) arms. The original updates match model, decoding, evidence, and output allowance but not input-token length, so they are secondary. The strict input-token-matched Qwen result is primary; Mistral's format failures preclude reliable accuracy estimates (Appendix A).

On identical initial traces, peer discussion improves plurality EM but lowers candidate-oracle coverage (Table 1). The natural-text placebo matches final token length and remains positive, but its shuffled, cross-item construction makes it a sensitivity control. The refined-evidence cells use $G$-dependent supplement seeds and are not mechanism estimates (Appendix M). The accuracy gain therefore comes with a smaller fraction of items having a correct fixed candidate, rather than demonstrating a richer candidate pool.

### 4.4 Tested aggregators do not consistently exploit complementarity

Positive oracle complementarity quantifies the selection opportunity in the candidate set. To measure how much a deployable policy recovers, we compare its accuracy with a member selected on validation data and report utilization as a normalized diagnostic (Table 2).

We evaluate majority voting, training-reliability-weighted voting, a logistic seat scorer, and A7 attention-weighted voting. A7 uses learned seat scores as vote weights; it is not a separately trained DeepSets or set-attention network. Correctness is a training target only. Reliability, agreement, and missingness features follow the declared splits, and all four methods select among fixed mem-

Table 1: **Matched communication outcomes.** Peer minus self-revision (placebo: peer minus shuffled peer); all entries are percentage points. The strict input-token row is primary. Accuracy estimates and their paired-item 95% intervals are separate columns; oracle columns are descriptive point estimates because interval estimates were not retained for those contrasts.

| Contrast | Accuracy Δ (pp) | | | | Oracle Δ (pp) | |
|---|---|---|---|---|---|---|
| | HotpotQA | | MuSiQue | | HotpotQA | MuSiQue |
| | Est. | 95% CI | Est. | 95% CI | Δ | Δ |
| Output-budget matched | +5.4 | – | +7.4 | – | −13.6 | −12.6 |
| Strict input-token matched | +0.4 | [−2.4, 3.0] | +4.0 | [2.0, 6.0] | −15.2 | −10.6 |
| Natural-text placebo | +5.6 | [2.6, 8.6] | +4.8 | [2.8, 7.0] | −2.0 | 0.0 |

Table 2: **Held-out aggregation gains.** Grouped nested evaluation over 66 domain-balanced cells. Accuracy is percent; gains are percentage points. Utilization is cell-wise $U_{\text{val}}$ over positive-gap cells. Test labels are excluded from fitting and selection; intervals use 2,000 paired draws.

| Method | Accuracy (%) | Gain (pp) | 95% CI | Utilization |
|---|---|---|---|---|
| *Deployable selectors* | | | | |
| Majority voting | 78.42 | +0.11 | [−0.85, 1.10] | -0.045 |
| Reliability weighting | 83.65 | +5.33 | [4.46, 6.21] | 0.384 |
| Logistic aggregation | 83.15 | +4.84 | [4.01, 5.67] | 0.349 |
| Attention-weighted vote | 83.68 | +5.37 | [4.51, 6.24] | 0.383 |
| *Reference bounds* | | | | |
| Validation-selected member | 78.32 | 0.00 | – | 0.000 |
| Candidate oracle (label-reading) | 93.27 | +14.96 | – | 1.000 |

ber answers. The internal procedure uses five outer and three inner group folds; external policies and baseline seats are frozen before follow-up scoring. Test labels enter neither fitting nor baseline selection.

The internal domain-balanced results show useful gains from aggregation. Reliability weighting improves on the validation-selected member by 5.33 percentage points, logistic aggregation by 4.84, and attention-weighted voting by 5.37; all three paired intervals exclude zero. Majority voting gains 0.11 points with an interval spanning zero. The candidate oracle gains 14.96 points, leaving a gap between available selection opportunities and the gains recovered by these methods. Comparing policy gain with this oracle gap distinguishes a committee bottleneck from a selector bottleneck. Mean cell-wise utilization is 0.384, 0.349, and 0.383 for the three weighted or learned methods. These averages support partial recovery, with variation across domains and communication arms, rather than a common utilization ceiling.

Accuracy and $\Delta_{\text{val}}$ across all evaluation cells are the primary outcomes. Small oracle gaps make utilization unstable, so threshold sensitivity and domain breakdowns are reported separately. Mean cell-wise utilization is not the ratio of mean gains: majority can have a small positive average gain and negative average utilization because cells have different oracle gaps. The retrospectively best seat answers a separate diagnostic question and is not a deployable baseline. Conditional rescue and harm likewise have different denominators; their rate difference is not a net accuracy change (Appendix A).

The Qwen external cells and internal summaries are reported separately. The supplement also reports broader historical aggregator and unrestricted-judge results with their evaluation scope (Appendices A and S). The finite selector comparison leaves open the performance of improved selection, answer synthesis, and additional interaction.

### 4.5 External validation of the scaling pattern

HotpotQA and MuSiQue test whether declining per-agent effective size extends beyond the controlled task generator. These experiments use benchmark-authored questions and each question's own documents, partitioned across committee members.

Table 3: **Cross-setting scaling summary.** Values summarize frozen committee outputs; $U_{\text{val}}$ is the domain-balanced majority utilization. External intervals are paired item-bootstrap intervals for $U_{\text{val}}$; controlled and held-out rows have no retained intervals. Estimates and intervals are shown in separate columns.

| Setting | $n_{\text{eff}}/n$ | Hetero. | $U_{\text{val}}$ | 95% CI | Key pattern |
|---|---|---|---|---|---|
| Controlled A | $0.51 \to 0.26$ | 0.32 | $-0.19$ | – | all shared-evidence tracks decline |
| Controlled B | $0.72 \to 0.60$ | 0.31 | $-0.02$ | – | decline; residual excess is small |
| Controlled C | $0.44 \to 0.22$ | 0.20 | $-0.67$ | – | decline with strongest utilization loss |
| Held-out templates | $18/18$ decline | $< 0.75$ in $5/6$ | $-0.21$–$-0.50$ | – | transfer across three families |
| HotpotQA (500) | $0.30 \to 0.14^{\dagger}$ | 0.37 | $-0.12$ | $[-0.25, -0.01]$ | external scaling confirmed |
| MuSiQue (500) | $0.33 \to 0.16^{\dagger}$ | 0.42 | $-0.10$ | $[-0.19, -0.01]$ | external scaling confirmed |

On 500 HotpotQA and 500 MuSiQue items, per-agent effective information declines from $n{=}4$ to $n{=}8$ across three model families (for example, Qwen3-4B HotpotQA $0.301 \to 0.143$ and Gemma-3-4B MuSiQue $0.473 \to 0.197$, with paired intervals excluding zero). Heterogeneous committees remain far from independent ($n_{\text{eff}}/n = 0.374, 0.415$), while the surface-only SHAM control changes little. Because benchmark construction changes committee size together with evidence partitioning, the matched contrasts in §4.3 are needed to interpret this transfer. These benchmarks therefore strengthen external validity, while the matched controlled contrasts carry the mechanism claim.

### 4.6 Robustness and falsification

Several pre-committed hypotheses were falsified: heterogeneous committees did not consistently increase independence, DEC was not distinguishable from RANDI, and a cross-model policy proxy failed on a second family. The full falsification table and negative controls are in Appendix K; the post-hoc audit separates reconstructed from unreproduced aggregates (Appendix Q).

## 5 Limitations

Members are open-weight models of 4B–14B; generalization to frontier proprietary systems is untested. External benchmarks partition documents across agents while the single-agent condition retains full context, so evidence partitioning and committee size are not separately identified.

Communication covers one round, not long-horizon tool use. Follow-up reliability is limited to Qwen because of Mistral format failures. Original cells differ in input length. Strict validation and the placebo match lengths, but cross-item neutral padding leaves the placebo as a sensitivity control. Historical exposure records are incomplete.

The effective-size summary captures variance-equivalent redundancy, while structural strata leave difficulty unresolved and composition contrasts condition on selected evaluation sets. Utilization covers four selectors; the candidate oracle does not bound unrestricted synthesis. Experiments concern English reasoning, not adversarial safety.

## 6 Conclusion

An LLM committee can improve accuracy while adding less independent information. Effective size saturates in the tested settings; a paired Qwen3-4B control shows decay beyond a fixed-correlation null. Input structure explains much of the dependence, while peer discussion can raise plurality accuracy and error correlation together and reduce fixed-candidate coverage. Held-out selectors recover only part of available complementarity; report dependence, coverage, and realized gain together when choosing interventions.

### Reproducibility statement

The supplement gives formal definitions, protocols, model settings, and the separation of original experiments from post-hoc audits and follow-up validation (Appendix A). The companion evidence package now includes the strict matched and natural-text placebo per-item traces, input-length assertions, frozen scorer, fixed-$\bar{\rho}$ analysis, manifests, and one-command verification scripts. The Overleaf upload remains an aggregate manuscript package rather than a complete generation pipeline. Historical folds remain unavailable, and S3 is not strict family-disjoint transfer. Reported bootstrap intervals condition on fitted policies and observed traces; they do not include all training variation. Upstream revision and historical exposure limitations are documented in the supplement.

### Ethics statement

This study uses public benchmarks and synthetic controlled tasks. It includes no human participants, personal data, or intervention; benchmark use follows the stated dataset terms.

### AI use statement

AI tools assisted with literature retrieval and metadata checks, experiment and protocol design, implementation of analysis and experiment scripts, statistical reconstruction, interpretation checks, manuscript revision, template integration, figure preparation, and language, consistency, and build checks. Human authors independently reviewed the outputs and remain responsible for the submitted content.

## References


Algirdas Avizienis. The N-version approach to fault-tolerant software. *IEEE Transactions on Software Engineering*, SE-11(12):1491–1501, 1985.

Abhijit V. Banerjee. A simple model of herd behavior. *Quarterly Journal of Economics*, 107(3): 797–817, 1992.

Blaž Bertalanič and Carolina Fortuna. The Ringelmann effect in multi-agent LLM systems: A scaling law for effective team size. *arXiv preprint arXiv:2606.02646v1*, 2026. URL `https://arxiv.org/abs/2606.02646v1`.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

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. *arXiv preprint arXiv:2308.07201*, 2023.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale, 1785.

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

Thomas G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pp. 1–15, 2000.

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. arXiv:2305.14325.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352v7*, 2024. URL https://arxiv.org/abs/2308.00352v7.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Ali Keramati, Justin Cheok, Jacob Horne, and Mark Warschauer. The confident liar: Diagnosing multi-agent debate with log-probabilities and LLM-as-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1361–1375, 2026. doi: 10.18653/v1/2026.acl-srw.121. URL https://aclanthology.org/2026.acl-srw.121/.

Leslie Kish. *Survey Sampling*. Wiley, 1965.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.17760.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024a. arXiv:2402.05120.

Yuante Li, Yicheng Tao, Kate Zhang, Taozhi Wang, Gefei Gu, and Yaxin Zhou. Diverse evidence, better forecasts: Multi-agent deliberation under information asymmetry. *arXiv preprint arXiv:2607.01661*, 2026.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7281–7294, 2024b. doi: 10.18653/v1/2024.findings-emnlp.427. URL https://aclanthology.org/2024.findings-emnlp.427/.

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2305.19118.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Dragos D. Margineantu and Thomas G. Dietterich. Pruning adaptive boosting. In *International Conference on Machine Learning (ICML)*, pp. 211–218, 1997.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? A look at multi-agent debate strategies for LLMs. *arXiv preprint arXiv:2311.17371*, 2024.

Selim Furkan Tekin, Fatih Ilhan, Gaowen Liu, Ramana Rao Kompella, and Ling Liu. Dynamic optimizations of LLM ensembles with two-stage reinforcement learning agents. *arXiv preprint arXiv:2502.04492*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. arXiv:2108.00573.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. arXiv:1809.09600.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02506.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.16823.

Table 4: Scope comparison with the closest related work. "Matched" uses paired item records; "fixed-candidate" forbids answer synthesis.

| Study | Evidence allocation | Communication | Dependence audit | Candidate utility |
|---|---|---|---|---|
| Ringelmann | composition / scaling | protocol contrasts | answer–correctness correlation | aggregation |
| InfoDelphi | partitioned evidence | forecasting deliberation | independence assumptions | forecasting |
| **This paper** | explicit CC/IND controls | matched no-update, self, and peer protocols | five diagnostics plus structural nulls | held-out fixed-candidate selection |

# A FOLLOW-UP VALIDATION: NEW AND HISTORICAL EVIDENCE BOUNDARIES

This follow-up was designed after examining historical results; it is separate from the original pre-committed analyses. All four external configurations are retained, although Qwen is emphasized in the main text because Mistral frequently violates the fixed answer format. All invalid outputs remain incorrect; these estimates jointly reflect reasoning and format adherence. Recorded-history exclusion is not proof of complete non-exposure or pretraining independence.

**Grouped held-out evaluation.** Five outer folds and three inner group folds yield OOF predictions on 66 internal cells. Conservative domain–item-index groups keep template, model, mechanism, and committee variants together. There are 375 such groups, not 8,250 independent trials. Validation-selected seats are fixed per outer fold, not one globally fixed test-selected seat. Reliability and logistic features use training data only. A7 is attention-weighted voting from logistic scores, not a separately trained DeepSets model. The main comparison covers these four fixed-candidate selection rules.

Table 5: **Domain breakdown of held-out selectors.** Entries are domain-balanced means over seed-42 frozen cells and report $\Delta_{\text{val}}$ in percentage points / $U_{\text{val}}$. They are descriptive decompositions; pooled paired intervals are reported in the main table. Light row tints are visual guides only and do not encode significance or a rank order.

| Domain | Majority | Reliability | Logistic | Attention |
|---|---|---|---|---|
| A | $+0.95/-0.027$ | $+4.69/0.491$ | $+4.36/0.463$ | $+4.73/0.488$ |
| B | $+3.38/0.176$ | $+8.62/0.478$ | $+8.11/0.452$ | $+8.51/0.468$ |
| C | $-4.00/-0.284$ | $+2.69/0.185$ | $+2.04/0.134$ | $+2.87/0.193$ |

**External transfer.** Each benchmark/model uses 400 historical training and 100 historical validation items. Frozen policies are applied to all three new arms with no test calibration. This is old-to-new same-benchmark transfer under changed evidence allocation, not strict family-disjoint transfer. Three deterministic solver seeds agree; they do not increase test sample size. Local dataset and model bytes are hashed; upstream immutable revision provenance remains incomplete. The placebo run contains 8,000 generated traces (1,000 questions, four seats, two

Table 6: Communication trade-off: peer minus self. EM gains and correlation changes are descriptive point estimates from the frozen traces; no interval estimates were retained for this table. Light row tints identify model family and are visual grouping aids.

| Dataset | Model | $\Delta$EM | $\Delta\rho$ |
|---|---|---|---|
| HotpotQA | Qwen3-4B | +5.4 | +0.388 |
| MuSiQue | Qwen3-4B | +7.4 | +0.463 |
| HotpotQA | Mistral-7B | +1.2 | +0.221 |
| MuSiQue | Mistral-7B | +2.2 | +0.161 |

Table 7: Qwen external communication decomposition on 500 questions per benchmark. EM is in percent; member EM is the seat mean. The oracle-to-plurality headroom is oracle EM minus plurality EM in percentage points, distinct from $g = o - b$. These are descriptive point estimates from the frozen traces, without new interval estimates.

| Benchmark | Arm | Member EM | Plurality EM | Oracle EM | Headroom | $\rho$ / $n_{\text{eff}}$ |
|---|---|---|---|---|---|---|
| HotpotQA | $G_{\text{self}}$ | 22.15 | 17.60 | 56.80 | 39.20 | 0.174 / 2.627 |
| HotpotQA | $G_{\text{peer}}$ | 25.80 | 23.00 | 43.20 | 20.20 | 0.562 / 1.489 |
| MuSiQue | $G_{\text{self}}$ | 9.65 | 3.00 | 34.00 | 31.00 | 0.016 / 3.818 |
| MuSiQue | $G_{\text{peer}}$ | 10.90 | 10.40 | 21.40 | 11.00 | 0.479 / 1.641 |

Table 8: Strict input-budget-matched robustness validation on the full 1,000-item external set. Each entry is peer minus self; EM and invalid changes are in percentage points, with paired 95% bootstrap intervals for EM. The final tokenizer input length is exactly equal per item and seat; plurality ties and invalid parses are errors.

| Dataset | Model | $\Delta$EM | 95% CI | $\Delta\rho$ | $\Delta$ invalid |
|---|---|---|---|---|---|
| HotpotQA | Qwen3-4B | $+0.4$ | $[-2.4, 3.0]$ | $+0.305$ | $+0.7$ |
| MuSiQue | Qwen3-4B | $+4.0$ | $[2.0, 6.0]$ | $+0.352$ | $-0.5$ |
| HotpotQA | Mistral-7B | $+4.2$ | $[2.2, 6.2]$ | $+0.316$ | $+22.4$ |
| MuSiQue | Mistral-7B | $+1.0$ | $[0.0, 2.0]$ | $+0.259$ | $+15.2$ |

arms). Generation did not load gold labels; the separate scorer uses the frozen official parser and `SCORING_ONLY_gold.json`. Both arms have identical final tokenizer lengths for every item and seat. Because exact matching may require inserting a neutral padding marker on the shorter source, the placebo is an answer-content sensitivity control rather than a semantically neutral causal intervention.

### A.1 Fixed-correlation null for committee-size saturation

The correlation-based $n_{\text{eff}}/n$ ratio decreases with $n$ even when $\bar{\rho}$ is held fixed. We therefore use the paired Qwen3-4B full-topology traces to estimate a descriptive null anchored at the observed $n = 2$ correlation. For each bootstrap draw, the same item indices are used at all committee sizes and the null curve is $n_{\text{eff}}/n = 1/[1+(n-1)\bar{\rho}_2]$. The observed-minus-null intervals remain negative at every larger size, showing that the observed trajectory is steeper than the mechanical fixed-correlation curve. This is a null for the variance-equivalent summary, not a test of mutual information or joint independence.

### A.2 Strict matched candidate coverage

The strict external traces also permit a fixed-candidate coverage check. Using the same frozen parser and gold-separated scorer, Qwen candidate-oracle EM falls from 56.4% to 41.2% on HotpotQA and from 34.0% to 23.4% on MuSiQue (self to peer), changes of $-15.2$ and $-10.6$ percentage points. Together with the strict plurality results in Table 8, this preserves the dependence–coverage trade-off while avoiding a claim of universal accuracy improvement. Mistral coverage is not interpreted because its invalid-output rate rises by 22.4/15.2 percentage points.

**Mistral failure boundary.** Invalid-format rates are 54.55/31.95/41.25% on HotpotQA and 46.40/31.70/42.80% on MuSiQue ($G_0/G_{\text{self}}/G_{\text{peer}}$). All invalids are scored wrong. MuSiQue peer–self correlation is undefined in 320/2,000 bootstrap draws because a member has zero correctness variance; its conditional-on-estimability interval is unstable and not evidence of robust significance. No format-based deletion or selective rerun was performed.

**Rescue and harm.** For the validation-selected baseline $B$, candidate oracle $O$, and selector $M$, rescue is $P(M = 1 \mid B = 0, O = 1)$; harm is $P(M = 0 \mid B = 1)$. The denominators differ. The net accuracy gain is $(N_{\text{rescue}} - N_{\text{harm}})/N$, not rescue rate minus harm rate. The identity is checked for every method, arm and seed. Counts below use Qwen peer-arm predictions and the previously frozen validation-selected seat; the historical reconstruction counts use a different evaluation population.

Table 9: Natural-text placebo validation under exact input-token matching. Each entry is real peer minus shuffled peer; intervals are paired item-bootstrap 95% intervals ($B = 2{,}000$). The shuffled arm replaces each peer initial with a response from a different item in the same benchmark, preserving the own initial, evidence, and final input length. Estimates and intervals are separate columns.

| Dataset | Δ plurality | 95% CI | Δ oracle | 95% CI | $\Delta\rho$ | 95% CI | Δ invalid |
|---|---|---|---|---|---|---|---|
| HotpotQA | +5.6 | [2.6, 8.6] | −2.0 | [−5.8, 1.6] | +0.274 | [0.203, 0.342] | +0.8 pp |
| MuSiQue | +4.8 | [2.8, 7.0] | 0.0 | [−3.2, 3.2] | +0.312 | [0.221, 0.399] | 0.0 pp |

Table 10: Fixed-$\bar{\rho}$ null for the full-topology Qwen3-4B committee traces. Intervals are paired item-bootstrap 95% intervals ($B = 2{,}000$).

| $n$ | observed $\bar{\rho}$ | observed $n_{\text{eff}}/n$ | fixed-$\bar{\rho}_2$ null | observed − null |
|---|---|---|---|---|
| 2 | 0.5004 | 0.666 | 0.666 | – |
| 3 | 0.6312 | 0.442 | 0.500 | $[-0.082, -0.035]$ |
| 4 | 0.6711 | 0.332 | 0.400 | $[-0.090, -0.048]$ |
| 6 | 0.6247 | 0.243 | 0.286 | $[-0.062, -0.026]$ |
| 8 | 0.6390 | 0.183 | 0.222 | $[-0.055, -0.025]$ |

**Historical reconstruction analysis.** The following original table is retained only as a post-hoc audit. Historical folds are unavailable. Repeated cells require clustering; its original pooled intervals and retrospective test-best seats are not interchangeable with the new question-group evaluation. S3 was not a strict family-disjoint test. They are retained as historical diagnostics, with these limitations.

## B Formal definitions and estimands

This appendix fixes the formal objects used in the main text and distinguishes the historical and held-out evaluation baselines.

A *committee* is a set of $n$ agents (language models with assigned roles), an item set $\mathcal{T}$, and a protocol that assigns each agent an evidence view $x_i(t)$ and a message set $m_i(t)$ for each item $t$. Agent $i$ produces a final answer $a_i(t)$; correctness is $c_i(t) = \mathbf{1}[a_i(t) = y(t)]$ with $y(t)$ the ground-truth answer.

**Effective independent information.** With $\bar{\rho}$ the mean over the $n(n-1)/2$ pairwise error correlations, $n_{\text{eff}} = n/(1 + (n-1)\bar{\rho})$ is the equicorrelated effective-sample-size identity. Four further operationalizations are defined in Appendix G: cofailure-based effective number, covariance participation ratio, spectral effective rank, and answer-entropy effective number.

**Available and realized gain.** Let $\text{best} = \max_i \mathbb{E}[c_i]$ over committee members under the allocation, and $\text{oracle} = \mathbb{E}[\max_i c_i(t)]$ (a selector that returns any correct member answer). Available complementarity is $\text{oracle} - \text{best}$; realized gain of policy $\pi$ is $\text{acc}(\pi) - \text{best}$; utilization is their ratio. The historical aggregation summaries use the threshold $0.02$ for available complementarity. The main grouped evaluation instead uses $b_{\text{val}}$, the held-out accuracy of a seat selected on validation data, and reports $\Delta_{\text{val}} = a_\pi - b_{\text{val}}$ over all cells. Its normalized diagnostic is $U_{\text{val}} = (a_\pi - b_{\text{val}})/(o - b_{\text{val}})$ on positive-gap cells. Seats may differ across outer folds. Thresholded historical utilization and the main validation-based quantity therefore use different baselines and evaluation populations.

## C Identifiability perspective

This appendix records the conceptual framework that motivated the design, kept out of the main text because the paper's claims are empirical.

We index a committee system by $(O, I, \pi)$: an observation operator $O$ (what each agent sees), an intervention family $I$, and a policy class $\pi$. The communication structure $G$ is absorbed into $O$ as its message-routing block, $O = (E, G)$, where $E$ is evidence assignment: an internal reduction audit rejected the alternatives (communication as a fourth independent axis; as a policy relabeling; as an intervention), because no single policy-indexed statistic reproduces the topology-axis outcomes (the

Table 11: Strict input-matched fixed-candidate coverage. Oracle EM is the fraction of items for which at least one of four frozen candidate answers is correct; invalid answers count as incorrect.

| Dataset | Model | Self oracle EM | Peer oracle EM | Peer − self |
|---|---|---|---|---|
| HotpotQA | Qwen3-4B | 56.4 | 41.2 | −15.2 pp |
| MuSiQue | Qwen3-4B | 34.0 | 23.4 | −10.6 pp |
| HotpotQA | Mistral-7B | 10.2 | 11.0 | +0.8 pp |
| MuSiQue | Mistral-7B | 2.6 | 2.8 | +0.2 pp |

Table 12: New-test paired change counts; logistic selector, seed 42 (other deterministic seeds agree).

| Dataset | Rescued/eligible | Harmed/baseline correct | Total |
|---|---|---|---|
| HotpotQA | 30/88 | 26/128 | 500 |
| MuSiQue | 16/52 | 13/55 | 500 |

audit, with the ordering breaks that forced this conclusion, is described in the study documentation). In idealized bit-world batteries we derived separability conditions for mechanism identification under $(O, I, \pi)$; the LLM experiments instantiate the same mechanisms (shared evidence, injected bulletins, message routing), and we use the framework only to choose which ablations to run. No theorem is claimed for the LLM setting.

## D Experimental protocol

This appendix specifies the protocol and estimands; complete re-execution additionally requires the generation artifacts that are outside the submission package. Further detail (prompts, template text) is in Appendices E and F.

**Controlled tasks.** Three domains instantiate the mechanisms by which committees share information: **A** (direct evidence integration: each seat holds two signal documents and one distractor, with per-document noise rate $\varepsilon$), **B** (two-hop joins: a link document and a property document must be composed; a decoy link is always present), and **C** (conflicting partially shared reports: a majority of seats receive reports naming the true item, a minority receive reports naming a shared wrong item). Each item carries an explicit allocation across 8 seats with mechanism variants: IND (no shared corruption), CC (shared bulletin asserting a wrong answer), MED (bulletin to one seat), BOTH (bulletin plus higher $\varepsilon$). Items are generated from fixed seeds (1500 main + 300 cross-seed); latent structure (truth, decoys, noise draws, flags) is identical across all language realizations, verified item-by-item by an automated audit.

**Held-out language realizations.** Template families F2 (restructured syntax), F3 (natural verbose with code-free filler sentences), and F4 (discourse reordering) realize the same latent items in unseen language. The latent-equivalence audit requires zero mismatches on every latent field before any run.

**Committee protocol.** Round 1: every agent answers with a one-sentence justification plus a code answer (few-shot prompt; greedy decoding). Round 2 (unless ablated): full mesh — every agent sees all other agents' round-1 justification and answer, then revises. The G0 ablation sets round-2 answers equal to round-1 answers.

**Statistics.** Item-level paired bootstrap (2,000 replicates) within cells, paired across $n$ within a track; cluster bootstrap over cells for aggregator utilization; within-cell permutation nulls (200 replicates) for dependence excess (permuting each agent's correctness vector across items preserves marginals and destroys dependence). The intended audit gate requires exact endpoint reproduction from per-item outputs; historical reproduction remains incomplete (Appendix Q).

## E Prompts and template families

This appendix makes the surface forms auditable; the saturation result must not depend on them, which is what the held-out families test.

The round-1 prompt is a wrapper with role ("You are analyst Desk A reviewing internal records"), the agent's source documents, the question, an instruction not to mention source names, and two few-shot examples ending in "Answer: CODE". Round 2 appends other agents' round-1 outputs as messages. External benchmarks replace the code few-shot with two span-answer examples and ask for "the exact short answer". Full wrapper text and one complete item per family (F1–F4, all three domains) are in the study code; the few-shot content is frozen in the runners and identical across all conditions.

## F Models and decoding

This appendix records the compute configuration.

Committee members: Qwen3-4B, Mistral-7B-v0.3, Gemma-3-4B-IT, Qwen3-14B (heterogeneous committees seat one of each family in fixed order) (Yang et al., 2025; Jiang et al., 2023; Gemma Team, 2025); screening adds Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Phi-4-mini-instruct; judges add Qwen3-32B and Llama-3.1-8B-Instruct. All are local checkpoints of public models; HF revision hashes are part of the planned release metadata. Decoding is greedy, `max_new_tokens`=110, batch 48 (24 for heterogeneous cells with long external contexts; disclosed per-cell in manifests). Hardware: 2× NVIDIA A100-SXM4-80GB; PyTorch 2.9.0, Transformers 4.57.6, Python 3.12. No quantization.

## G Dependence metrics and null calibration

This appendix verifies that the saturation result does not depend on the correlation-based definition of $n_{\text{eff}}$.

For each metric we give the definition, intuition, null behavior, and limitation. M1–M4 use the per-item seat-correctness matrix $A \in \{0,1\}^{|\mathcal{T}|\times n}$; M5 uses the corresponding answer identities and does not require correctness.

**M1 Correlation-based $n_{\text{eff}}$.** $n/(1+(n-1)\bar{\rho})$. Intuition: the number of independent standardized errors with the same variance of their mean. Null: $\bar{\rho}=0 \Rightarrow n_{\text{eff}}=n$. Limitation: reduces the full dependence structure to its mean; inapplicable at $n$=1.

**M2 Cofailure-based effective number.** With $q_{\text{obs}}$ the observed cofailure rate and $q_{\text{ind}}=\prod_i(1-\text{acc}_i)$ its prediction under independence, $n_{\text{eff,cof}}=n\log q_{\text{obs}}/\log q_{\text{ind}}$ (guard: if $q_{\text{obs}}\le q_{\text{ind}}$, set $n$). Intuition: how many independent copies reproduce the observed tail of common failures. Limitation: tail-focused; unstable when cofailures are very rare.

**M3 Participation ratio.** Eigenvalues $\lambda$ of the seat-error covariance; $(\sum\lambda)^2/\sum\lambda^2$. Intuition: the effective dimension of the error space. Limitation: second-moment only.

**M4 Spectral effective rank.** $\exp(-\sum_j \hat{\lambda}_j \log\hat{\lambda}_j)$ with $\hat{\lambda}$ normalized. Intuition: entropy-weighted dimension; closer to 1 than M3 when one direction dominates. Limitation: same as M3.

**M5 Answer-entropy effective number.** Per item, the exponential of the entropy of the empirical distribution of final answers across seats; averaged over items. Intuition: how many distinct answers the committee actually produces. Limitation: ignores correctness; a committee agreeing on a wrong answer scores low.

**Null calibration.** Within each cell we permute each seat's correctness (and answer) vector independently across items (200 replicates). This preserves every marginal while destroying between-seat dependence; the observed-minus-null mean, standardized by the null standard deviation, gives the dependence-excess $z$ reported in the main text (exceeds 1.96 in 25 of 27 shared-evidence cells). All five per-agent normalized metrics decline in every shared-evidence track (Appendix H).

## H Full scaling results

This appendix reports the per-track numbers behind Figure 3 and the saturation claim.

Internal tracks (mechanism CC, anchor model Qwen3-4B and heterogeneous; per-agent $n_{\text{eff}}/n$, item-level paired bootstrap): pooled slope $-0.071$ $[-0.080, -0.061]$; per-track slopes A/homo $-0.065$ $[-0.082, -0.050]$, A/hetero $-0.065$ $[-0.084, -0.043]$, B/homo $-0.046$ $[-0.079, 0.004]$ (the one directional track, $P(\text{neg}) = 0.965$), B/hetero $-0.086$ $[-0.110, -0.066]$, C/homo $-0.091$ $[-0.111, -0.076]$, C/hetero $-0.074$ $[-0.087, -0.063]$. Held-out families: all 18 family-domain-composition trajectories decline (F2 4/4 frozen checks, F4 4/4, F3 3/4 with one boundary case at exactly the pre-committed 0.75 threshold). External benchmarks: every model family on both benchmarks declines from $n{=}4$ to $n{=}8$ with intervals excluding zero (HotpotQA: Qwen3-4B $0.301 \to 0.143$ $[-0.169, -0.148]$; Mistral-7B $0.350 \to 0.155$ $[-0.213, -0.178]$; Gemma-3-4B $0.355 \to 0.167$ $[-0.206, -0.172]$. MuSiQue: Qwen3-4B $0.327 \to 0.155$ $[-0.187, -0.158]$; Mistral-7B $0.364 \to 0.167$ $[-0.220, -0.175]$; Gemma-3-4B $0.473 \to 0.197$ $[-0.325, -0.236]$.)

## I AGGREGATOR RED-TEAM DETAILS

This appendix reports the historical aggregator comparison. The main question-group evaluation is described in Appendix A.

Deployable aggregators use final answers and, where specified, training-estimated reliability or justifications; oracle-information controls are exceptions. $S_1$ denotes historical item CV, whose folds are unavailable, not a newly verified strict holdout. $S_2$ holds out domains; $S_3$ is a record-weighted model protocol with repeated heterogeneous cells and records lacking the nominal held-out model. Utilization is relative to candidate-oracle gain. Historical means use 65 eligible cells (available gain $> 0.02$), with record weighting in S3. The following intervals are historical; the revised cell-cluster intervals are in Table 13. A1 majority $-0.344$ $[-0.552, -0.140]$; A2 reliability weighting $0.174$ $[0.089, 0.264]$; A3 calibrated reliability $0.174$; A4 logistic stacking $0.110$; A5 MLP agent-choice $-0.325$; A6 DeepSets-style scorer $-0.319$; A7 attention-weighted vote $0.175$ $[0.078, 0.260]$; A8 dependence-aware weighting $0.166$; A9t candidate-selector oracle bound $1.000$ by construction; A9l *oracle-feature regressor* (cheating reference, not deployable: features include each member's correctness) $0.07$–$0.15$ under distribution shift; A11 greedy $\kappa$-diversity selection $-0.227$ (max cell $0.000$). Deployable families are A1–A8, A11, and the judges below; A9l is marked as an oracle-information control.

## J LLM JUDGES

This appendix fully discloses the judge configuration and its failures, including the cases where judges succeed.

The judge produces an answer after reading the question and every agent's round-2 justification and final answer (few-shot prompt with two examples; greedy decoding). It may generate a new answer code: the candidate oracle is a reference, not an upper bound on unrestricted judge synthesis. **Qwen3-32B judge**: mean utilization $0.162$ $[0.027, 0.284]$; domain means A $0.037$, B $0.390$, C $0.054$; *best cell* $1.000$*, on a committee whose members share the judge's model family* (a disclosed self-preference risk), and the elevated domain-B mean does not meet the pre-committed success threshold ($> 0.50$ stable in two or more domains). **Llama-3.1-8B judge** (family-disjoint from all seat models): mean $-0.028$ $[-0.119, 0.060]$; domains A $-0.133$, B $0.060$, C $-0.015$. These historical, fixed-prompt evaluations show heterogeneous judge performance. Their utilization is relative to a retrospectively selected seat, and the candidate oracle is a comparison reference for unrestricted judges. These results are separate from the main grouped selector evaluation.

## K PRE-COMMITTED FALSIFICATION TABLE

This appendix records predictions fixed before their corresponding evaluations and the resulting limits on the claims. Historical results are retained; the verdicts below are interpreted against the stated success rules and subsequent provenance checks.

Table 13: Historical reconstruction analysis. Post-hoc audit; original folds unavailable, retrospective test-best reference, repeated cells clustered in subsequent audits. Estimates and historical pooled 95% intervals are separate columns; these are not the main paired-item intervals. Historical S3 is not strict family-disjoint transfer.

| Aggregator | $S_1$ (historical CV) | | $S_2$ (domain holdout) | | $S_3$ (model protocol) | |
|---|---|---|---|---|---|---|
| | Est. | 95% CI | Est. | 95% CI | Est. | 95% CI |
| Majority vote | -0.344 | $[-0.552, -0.140]$ | -0.344 | $[-0.555, -0.149]$ | -0.451 | $[-0.612, -0.295]$ |
| Reliability wt. | 0.174 | $[0.089, 0.264]$ | 0.162 | $[0.068, 0.252]$ | 0.103 | $[0.014, 0.189]$ |
| Calibrated rel. | 0.174 | $[0.087, 0.263]$ | 0.160 | $[0.062, 0.253]$ | 0.105 | $[0.012, 0.186]$ |
| Logistic stack | 0.110 | $[0.020, 0.208]$ | 0.043 | $[-0.136, 0.172]$ | -0.061 | $[-0.189, 0.053]$ |
| MLP agent-choice | -0.325 | $[-0.457, -0.213]$ | -0.001 | $[-0.166, 0.120]$ | -0.075 | $[-0.188, 0.024]$ |
| DeepSets | -0.319 | $[-0.438, -0.217]$ | -0.041 | $[-0.204, 0.078]$ | -0.030 | $[-0.112, 0.040]$ |
| Attention vote | 0.175 | $[0.078, 0.260]$ | 0.052 | $[-0.115, 0.183]$ | -0.048 | $[-0.182, 0.074]$ |
| Dependence-aware | 0.166 | $[0.078, 0.261]$ | 0.082 | $[-0.086, 0.218]$ | 0.103 | $[0.002, 0.184]$ |
| Candidate oracle | 1.000 | $[1.000, 1.000]$ | 1.000 | $[1.000, 1.000]$ | 1.000 | $[1.000, 1.000]$ |
| LLM judge (Qwen3-32B) | 0.162 | $[0.023, 0.281]$ | – | – | – | – |
| LLM judge (Llama-3.1-8B) | -0.028 | $[-0.117, 0.062]$ | – | – | – | – |
| Greedy $\kappa$-selection | -0.227 | $[-0.299, -0.165]$ | – | – | – | – |

Table 14: Pre-committed failures and their effects. Full version-control provenance in the study ledger.

| **Prediction (frozen)** | **Observed** | **Verdict** | **Effect on claim** |
|---|---|---|---|
| Injection signature separates IND (D4) | identical on 2000/2000 items | FAIL | signatures scoped to CC family |
| Cross-model separator transfers (C8) | lost all power on second family | FALSIFIED | per-model reporting |
| DEC structurally distinct from RANDI | indistinguishable in bit battery | PARTIAL | correction claim scoped |
| FP2 strong force-privacy effect | direction held, magnitude below threshold | DOWNGRADED | secondary evidence |
| Task E pilot V1 | key predictions missed | FAIL | redesigned |
| Task E pilot V2 | measurement bug invalidated endpoints | FAIL | V3 correction; V3 used throughout |
| Heterogeneity raises effective information | B: CI includes zero; C: correlation increase CI $[0.019, 0.258]$ | FALSIFIED | diversity = unreliable proxy |
| MLP aggregator closes gap (B11) | utilization $-0.333$ | FAIL | red-team extended |
| SHAM $\ll$ RANDI universally | domain A unstable ($P$=0.25) | DOWNGRADE | domain-conditional nullity |
| Observation refinement restores independence (OIPG F1) | $\rho$ effect CI includes zero | NULL | O is an accuracy lever, not independence lever |
| High-authority intervention increases separation (PI1) | direction reversed | FALSIFIED | authority claim inverted |
| Hetero $n_{\text{eff}}/n < 0.75$ all domains (F3) | domain A exactly 0.750 | BOUNDARY | threshold sensitivity disclosed |
| Multi-master debate as core primitive | confounded measurement | DESIGN KILL | committee protocol instead |
| LLM judge recovers $> 50\%$ | 0.162 mean, one cell 1.000 | THRESHOLD NOT MET | historical judge result only |
| Greedy selection beats zero | max cell 0.000 | FAIL | tested rule gives no gain |

## L QUALITY-MATCHED HETEROGENEITY

This appendix reduces observed single-agent accuracy differences in the diversity comparison. Matching and comparator selection use the evaluation items, so the estimates remain conditional on that selection.

Candidates were screened on the same items as the main matrix (pooled single-agent accuracy): Gemma-3-4B $0.628$, Mistral-7B $0.669$, Phi-4-mini $0.644$, Qwen3-14B $0.663$ (maximum absolute deviation from their mean $0.023$; Llama-3.1-8B and Qwen3-4B screened but not selected). The quality-matched heterogeneous committee is compared against each member's homogeneous committee on the same items with paired item-bootstrap conditional on selection. Error correlation against the homogeneous committee with the lowest observed correlation: domain B $+0.130$ $[-0.004, 0.271]$ (interval includes zero); domain C $+0.141$ $[0.019, 0.258]$ (the conditional pointwise interval excludes zero). Majority utilization of the matched heterogeneous committee: $-0.361$; reliability-weighted $+0.013$. Under this matching rule, family diversity does not consistently improve effective independence. Residual capability differences and selection effects remain possible.

## M COMMUNICATION AND FACTORIAL ABLATIONS

This appendix retains the original factorial as a historical descriptive analysis. It is superseded for mechanism attribution by the matched-input post-hoc contrasts in Appendix R.

Domain C, mechanism CC, $n{=}4$, homogeneous anchor model; $2 \times 2 \times 2$ over evidence refinement (O: base vs. one extra signal document per seat), bulletin intervention (I: none vs. correction), and communication (G: round-1 only vs. full mesh); aggregation policy ($\pi$: majority vs. reliability weighting) is evaluated post hoc. Main effects on $\bar{\rho}$: O $+0.030$ $[-0.038, 0.097]$ (null); I $-0.076$ $[-0.144, -0.012]$; G $+0.455$ $[0.379, 0.532]$. On majority accuracy: O $+0.108$, I $+0.168$, G $+0.176$. Cell endpoints ($\bar{\rho}$ / $n_{\text{eff}}$): G0 cells $-0.03/4.4$, $-0.06/4.9$, $-0.03/4.3$, $-0.04/4.5$; G1 cells $0.43/1.8$, $0.36/1.9$, $0.54/1.5$, $0.34/2.0$. Two-way interactions on $\bar{\rho}$: O×I $-0.062$ $[-0.162, 0.036]$; O×G $+0.032$ $[-0.086, 0.148]$; I×G $-0.111$ $[-0.236, 0.007]$ — all intervals include zero. The O1 supplement generator includes G in its seed, so realized evidence differs across those G contrasts. These historical main effects are confounded and are not evidence that communication dominates other mechanisms. The main text uses the two O0 matched comparisons instead.

## N EXTERNAL BENCHMARKS: CONSTRUCTION AND FULL RESULTS

This appendix documents exactly what is and is not author-controlled in the external evaluation — the position a reviewer will most want to check.

**What is unchanged.** Questions, gold answers, and the document set are exactly the benchmark's own (HotpotQA dev, distractor setting: 10 paragraphs, 2 gold, 8 distractors; MuSiQue-Ans dev: 20 paragraphs, 2–4 supporting). We modify no question, answer, or document. The single-agent condition sees all paragraphs — the benchmark's original setting — which anchors the protocol (Qwen3-4B single-agent EM 0.552/0.602 at 250/500 items, within the published range for models of this size).

**What is author-controlled.** The partition of the question's own paragraphs across committee seats (shared-gold "CC" partition vs. random "IND" partition), and the two-round protocol. These are the treatments under study.

**Item selection.** First 500 eligible items in file order (HotpotQA: exactly-2-gold, non-yes/no, 10 paragraphs; MuSiQue: answerable, $\geq$2 supporting). The 250-item pilot uses the first 250 of the same scan.

**Full results.** See Table 15. Heterogeneous committees (four families): $n_{\text{eff}}/n$ 0.374 (HotpotQA), 0.415 (MuSiQue). Majority utilization pooled over committee cells: HotpotQA $-0.115$ $[-0.247, -0.011]$; MuSiQue $-0.101$ $[-0.194, -0.005]$. SHAM surface-shuffle control (250-item pilot): $\bar{\rho}$ 0.782 vs. 0.784 baseline.

Table 15: External scaling, per model family: per-agent $n_{\text{eff}}/n$ at $n{=}4 \rightarrow 8$. The estimate and paired item-bootstrap 95% CI for each benchmark are shown in separate columns.

| Model | HotpotQA | | MuSiQue | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| Qwen3-4B | $0.301 \rightarrow 0.143$ | $[-0.169, -0.148]$ | $0.327 \rightarrow 0.155$ | $[-0.187, -0.158]$ |
| Mistral-7B | $0.350 \rightarrow 0.155$ | $[-0.213, -0.178]$ | $0.364 \rightarrow 0.167$ | $[-0.220, -0.175]$ |
| Gemma-3-4B | $0.355 \rightarrow 0.167$ | $[-0.206, -0.172]$ | $0.473 \rightarrow 0.197$ | $[-0.325, -0.236]$ |

## O Reproducibility and audit pipeline

This appendix records the intended audit pipeline. The post-hoc audit in Appendix Q identifies unresolved historical fold and confidence-interval provenance; not every number has been reproduced.

Generation seeds: 20260901 (main), 20260902 (cross-seed); analysis RNG seeds are fixed per script. Historical S1 specified item-grouped 5-fold CV by $\text{hash}(\text{item_id}) \bmod 5$, but its hash seed and folds were not retained. S2 uses domain-held-out rotations; S3 uses the record-weighted model protocol, not a verified strict model holdout. Bootstrap: item-level, paired across committee sizes within a track, 2,000 replicates; cluster bootstrap over cells for aggregator utilization; permutation nulls 200 replicates. **Audit gates**: (i) endpoints recompute from per-item outputs to four decimal places before any interval is read; (ii) latent-equivalence audit (zero mismatches) before any held-out-template run; (iii) allocation determinism by per-cell hash. **Provenance**: the original pre-committed hypotheses and success rules were fixed before their corresponding evaluations. Post-hoc audits and the subsequent follow-up are identified separately. The submission package contains frozen aggregate evidence, figure sources, LaTeX sources, and build instructions. It does not contain raw model checkpoints, per-item prediction matrices, the private infrastructure ledger, or the complete generation pipeline. Accordingly, the package supports inspection of the reported aggregates and figures; it does not support end-to-end regeneration of model outputs.

## P Artifact release notes and additional limitations

This appendix states release scope and residual caveats.

Future artifact release: testbed generators (all domains and template families), committee runners, external-benchmark builders and runners, analysis scripts, frozen prediction matrices, and per-cell manifests. Excluded from that release: raw model checkpoints (public; identified by revision hash) and the private ledger with infrastructure paths. Until that release is public, the submission package supports audit of the reported aggregates and figures, rather than full rerunning of model generation. Additional limitations beyond the main text: results are for English tasks; initial generation followed by one optional discussion round is one point in the space of communication schedules (the G0 comparison omits discussion; a sweep over multiple rounds was not performed); and the falsification table, while complete for pre-committed predictions, cannot cover unconceived analyses.

## Q Post-hoc reconstruction and interval repair

We distinguish historical summary reproduction from reconstruction of policies and from statistical validity. The saved report's A1–A8 intervals can be reproduced from its aggregate records, whereas the general paper-data generator uses a different RNG stream and percentile procedure. Matching one generator does not certify another. More importantly, the historical S3 report resampled 137 evaluation records separately, despite their belonging to only 65 unique cells.

The revised aggregation table therefore keeps the original point-estimate estimand but replaces its uncertainty with an explicitly post-hoc cell bootstrap. For every one of 2,000 draws (NumPy default RNG, seed 20260907, reset per series), sample 65 cell IDs with replacement and carry every evaluation record of each selected cell. Compute the record-weighted mean of the resampled values; take the 2.5th and 97.5th percentiles. S1/S2, Greedy, and fixed-prompt judges have one aggregate per cell. In S3, 41 homogeneous cells appear once and 24 heterogeneous cells each contribute four model-test records; their within-cell dependence is preserved. Twelve records do not contain the

nominal held-out model. Keeping record weighting is not equivalent to cell-equal sensitivity, which can reverse some method signs.

These intervals are conditional on saved policies and a best seat selected using the evaluation labels. They do not include training, seat-selection, or cross-cell shared-item uncertainty and must not be read as population-level guarantees. The new paired rescue/harm results below use reconstructed policies, not recovered historical predictions. The original S1 and Greedy split rule used Python hash; its seed and fold assignments were not retained. CRC32 gives a new deterministic post-hoc split, not historical recovery or verification of the historical S1 holdout. S2/S3 aggregate accuracy matches do not prove itemwise identity; S3 is a record-weighted model protocol, not a verified strict model holdout. Judges retain aggregate outputs only; their prompt and parser permit a newly generated answer code, so the candidate oracle bounds selectors, not arbitrary judge synthesis. Raw old tables remain archived and are not silently overwritten.

## R STRUCTURAL-DIFFICULTY-CONDITIONED DEPENDENCE AUDIT

We reanalyzed all 66 base E-FULL s21 cells (125 items each), without selecting on oracle gap or fitting or querying any model. The unit of analysis is a cell; summaries below weight cells equally. Difficulty covariates were recovered from the preserved input generator and item files, independently of all tested outputs: template, pre-generated contamination, and input evidence support. Within each domain/mechanism/model/size cell, evidence support counts truthful signal statements in A, truthful link and property support in B, and truthful source reports in C. We divide by the maximum support (respectively $2n, 2n, 3n$) and use fixed bins $[0, 1/3], (1/3, 2/3], (2/3, 1]$. Ground-truth labels here characterize generated inputs, not model-derived difficulty.

For each of 1,999 replicates (seed 20260906), each member's binary error vector is independently permuted within its input-defined stratum, preserving its stratum-specific error count exactly. The statistic is mean pairwise error correlation. One-sided Monte Carlo $p$ values use $(1 + \#\{T^* \geq T\})/2000$. The observed cell-equal mean correlation is 0.2759. The table gives conditional-null means, observed-minus-null excesses, counts with $p < .05$, descriptive BH-adjusted counts, and the percentage of items in nonsingleton strata. Overlapping cells are not independent replications; BH counts are descriptive, not a guaranteed error-control claim under arbitrary cross-cell dependence.

| Stratification | Null $\rho$ | Excess | Raw | BH | Movable (%) |
|---|---|---|---|---|---|
| Unconditional | 0.0000 | 0.2759 | 58/66 | 58/66 | 100.0 |
| Template and contamination | 0.0894 | 0.1865 | 57/66 | 57/66 | 100.0 |
| + structural evidence bins | 0.2111 | 0.0648 | 30/66 | 28/66 | 98.0 |
| Exact seat-evidence vector | 0.2649 | 0.0110 | 7/66 | 2/66 | 47.6 |

The structural-bin null leaves mean excess correlations of 0.0697, 0.0064, and 0.1183 in domains A, B, and C (raw significant cells: 10/22, 3/22, and 17/22). Its mean excess all-seat cofailure probability is 0.01346, versus 0.05059 under unconditional permutations. Thus much of the unconditional signal is explainable by observed input heterogeneity, and domain B provides little residual evidence under this adjustment. The exact seat-evidence sensitivity further reduces excess, but only 47.6% of items belong to nonsingleton strata (22.4% in C); this severe sparsity restricts its power. Neither coarse strata nor exact evidence-count vectors establish exchangeability conditional on all latent difficulty. These results do not identify shared failure mechanisms independently of unmeasured item properties.

**Matched communication contrasts, not a conditional null.** For the domain-C, CC, homogeneous Qwen3-4B, $n = 4$ factorial, the two $O = 0$ comparisons have identical round-1 answers and raw texts in all 125 paired items. With $I = 0$, majority accuracy rises from 0.440 to 0.664: 29 gains and one harm, $\Delta = 0.224$ (paired-item percentile 95% interval $[0.152, 0.304]$; exact two-sided McNemar $p = 5.77 \times 10^{-8}$). Mean error correlation rises by 0.4569, interval $[0.3577, 0.5564]$. With $I = 1$, accuracy rises from 0.600 to 0.784: 26 gains and three harms, $\Delta = 0.184$ ($[0.112, 0.272]$; $p = 1.52 \times 10^{-5}$), while correlation rises by 0.4219 ($[0.3020, 0.5357]$). Intervals use 2,000 paired-item bootstrap replicates (seed 20260907, reset for each contrast) and are conditional on these frozen traces; they do not cover model-training uncertainty.

Importantly, the $O = 1$ supplement generator seeds its random draw using the full cell identifier, which includes $G$. Hence those between-cell $G$ contrasts do not hold realized evidence fixed: round-1 answers differ in 38/125 and 47/125 items for $I = 0, 1$ (raw texts differ in 123/125 and 124/125). A four-cell average must not be described as a clean matched-input communication effect. As a separate fixed-input transition analysis, comparing round 2 to round 1 within the $O = 1, G = 1$ traces gives majority gains of 0.152 and 0.176 and correlation increases of 0.5973 and 0.4608. These paired transitions support communication-associated dependence under the frozen protocol but are neither conditional-independence tests nor evidence of randomized treatment assignment.

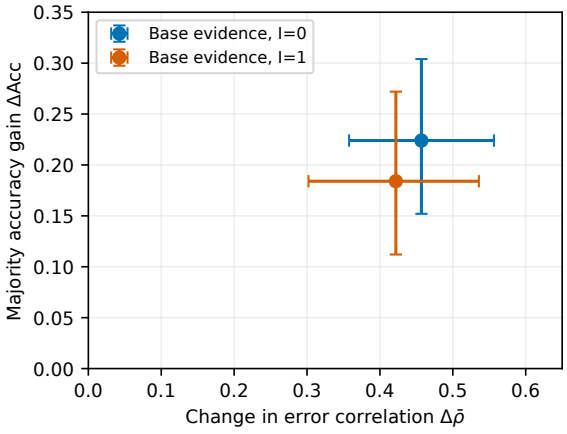



Figure 4: **Accuracy–dependence trade-off on matched inputs.** Two base-evidence communication contrasts, each with 125 paired domain-C items. Horizontal and vertical bars are marginal paired-item 95% bootstrap intervals, not a joint confidence region. Both contrasts improve accuracy while increasing error correlation. The refined-evidence factorial contrasts are excluded because their generated inputs differ across communication conditions. This is a zero-versus-one discussion comparison, not a multi-round dose-response curve.

## S UTILIZATION SENSITIVITY AND RESCUE/HARM

Across nine strict gap thresholds, pooled cell-equal utilization point estimates for all reconstructed non-oracle policies remain below 0.30. The maximum is 0.218 (S2 reliability/calibrated reliability at $t = 0.30$), based on only two cells. The statement is not universal across domains: at $t = 0.02$, domain-B S1-new attention reaches 0.405, S2 reliability 0.381, and S3 reliability 0.366 (22 cells each). These counterexamples are retained. Raising the threshold changes the population of cells, not just a denominator in a fixed population. S3 averages its model-test records within each cell before averaging cells, unlike the historical table's record weighting.

Table 16: Unique eligible cells by strict gap threshold. Empty domain subsets have undefined estimates.

| Domain | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| ALL | 66 | 66 | 65 | 63 | 55 | 39 | 17 | 5 | 2 |
| A | 22 | 22 | 21 | 19 | 11 | 4 | 3 | 1 | 0 |
| B | 22 | 22 | 22 | 22 | 22 | 21 | 7 | 4 | 2 |
| C | 22 | 22 | 22 | 22 | 22 | 14 | 7 | 0 | 0 |

For best-seat correctness $B$, oracle correctness $O$, and reconstructed selector correctness $P$, define rescue as $R = \Pr(P = 1 \mid B = 0, O = 1)$ and harm as $H = \Pr(P = 0 \mid B = 1)$. S3 counts average folds within cell, so fractional counts are valid. Their denominators differ: $R - H$ is not

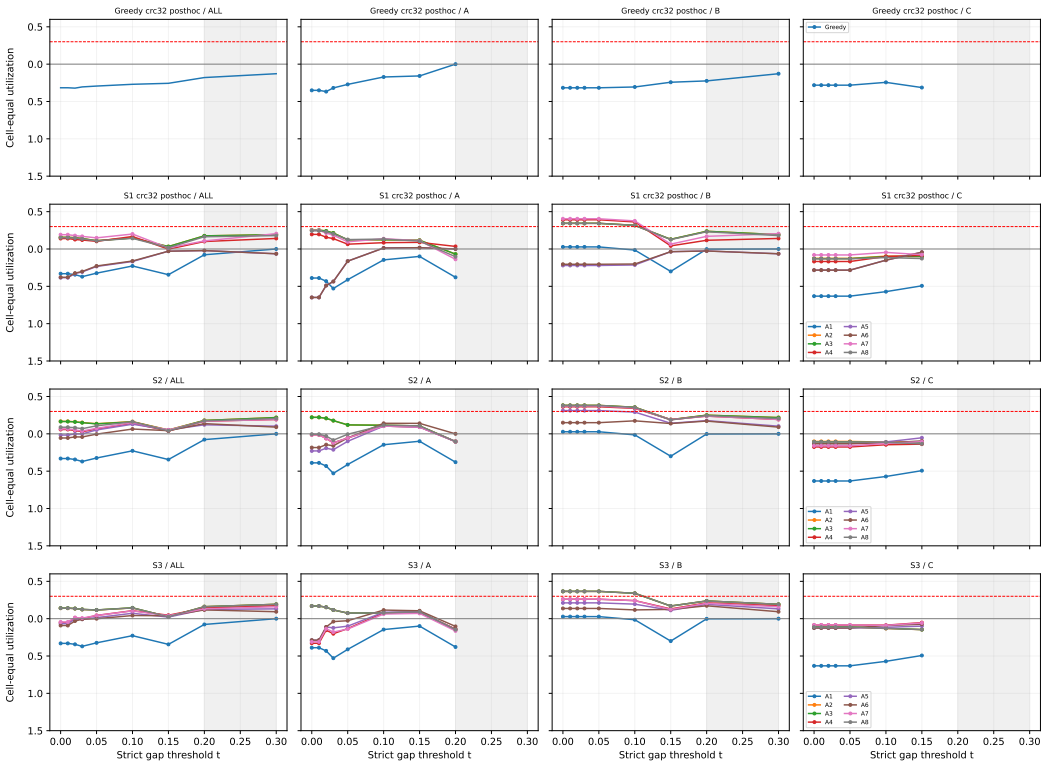



Figure 5: **Full reconstructed-policy threshold sensitivity.** Rows distinguish Greedy-new, S1-new, S2 and S3; columns distinguish pooled domains and A/B/C. A1 majority, A2 reliability, A3 calibrated reliability, A4 logistic stack, A5 MLP chooser, A6 DeepSets, A7 attention, A8 dependence-aware. Each available method is shown; these are point estimates, not simultaneous confidence bands. The 0.30 reference and high-threshold shading do not establish significance. Empty subsets have no estimate.

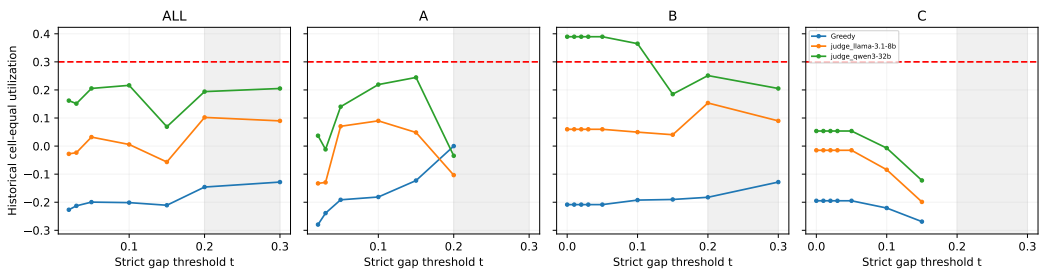



Figure 6: Historical fixed-prompt judges and historical Greedy, kept separate from reconstructed-policy sensitivity. Incomplete historical coverage is omitted rather than filled with reconstructed predictions.

an accuracy change. For selectors the matched accuracy change is $(N_R - N_H)/(125n_{\text{cells}})$, not the mean of cell-specific utilization ratios.

At $g > 0.02$, majority rescues 344 of 952 eligible cell-items but harms 616 of 6618 best-correct cases (net loss 272); Greedy-new rescues 121 and harms 389. S2 dependence-aware weighting rescues 508 and harms 397. All reconstructed policies at the default threshold are shown next, with full domain/threshold counts delivered as CSV rather than selectively reported. Original policy predictions and judge itemwise outputs are unavailable; these are conditional post-hoc reconstructions, not recovered historical error transitions. The best seat is chosen on each full evaluation set and ties use the smallest seat index. No uncertainty from fitting, baseline selection, or tie choice is included; this section deliberately reports verified point estimates and counts only.

Table 17: All reconstructed policies at gap $> 0.02$, pooled domains, 65 equally weighted cells. These are new paired outcomes, not recovered historical item predictions.

| Policy | Rescue count | $R$ | Harm count | $H$ | $\Delta$acc. |
|---|---|---|---|---|---|
| Greedy-new Greedy | 121/952 | 0.127 | 389/6618 | 0.059 | -0.033 |
| S1-new A1 | 344/952 | 0.361 | 616/6618 | 0.093 | -0.033 |
| S1-new A2 | 490/952 | 0.515 | 362/6618 | 0.055 | 0.016 |
| S1-new A3 | 490/952 | 0.515 | 363/6618 | 0.055 | 0.016 |
| S1-new A4 | 374/952 | 0.393 | 259/6618 | 0.039 | 0.014 |
| S1-new A5 | 474/952 | 0.498 | 674/6618 | 0.102 | -0.025 |
| S1-new A6 | 473/952 | 0.497 | 668/6618 | 0.101 | -0.024 |
| S1-new A7 | 413/952 | 0.434 | 260/6618 | 0.039 | 0.019 |
| S1-new A8 | 496/952 | 0.521 | 374/6618 | 0.057 | 0.015 |
| S2 A1 | 344/952 | 0.361 | 616/6618 | 0.093 | -0.033 |
| S2 A2 | 505/952 | 0.530 | 361/6618 | 0.055 | 0.018 |
| S2 A3 | 505/952 | 0.530 | 363/6618 | 0.055 | 0.017 |
| S2 A4 | 521/952 | 0.547 | 439/6618 | 0.066 | 0.010 |
| S2 A5 | 532/952 | 0.559 | 470/6618 | 0.071 | 0.008 |
| S2 A6 | 512/952 | 0.538 | 491/6618 | 0.074 | 0.003 |
| S2 A7 | 524/952 | 0.550 | 433/6618 | 0.065 | 0.011 |
| S2 A8 | 508/952 | 0.534 | 397/6618 | 0.060 | 0.014 |
| S3 A1 | 344/952 | 0.361 | 616/6618 | 0.093 | -0.033 |
| S3 A2 | 520.5/952 | 0.547 | 397.75/6618 | 0.060 | 0.015 |
| S3 A3 | 521.25/952 | 0.548 | 397.75/6618 | 0.060 | 0.015 |
| S3 A4 | 549/952 | 0.577 | 487.5/6618 | 0.074 | 0.008 |
| S3 A5 | 545.25/952 | 0.573 | 506.25/6618 | 0.076 | 0.005 |
| S3 A6 | 509.75/952 | 0.535 | 489.75/6618 | 0.074 | 0.002 |
| S3 A7 | 550/952 | 0.578 | 488/6618 | 0.074 | 0.008 |
| S3 A8 | 520.5/952 | 0.547 | 397.75/6618 | 0.060 | 0.015 |

## S.1 COMPLETE UTILIZATION THRESHOLD POINT TABLES

Table 18: Reconstructed policies, domain ALL. Cell-equal utilization for strict gap thresholds. An asterisk marks incomplete historical coverage, not a full eligible-set estimate. No confidence intervals.

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| Greedy-new Greedy | -0.316 | -0.316 | -0.321 | -0.305 | -0.294 | -0.270 | -0.257 | -0.179 | -0.128 |
| S1-new A1 | -0.331 | -0.331 | -0.344 | -0.370 | -0.324 | -0.228 | -0.344 | -0.077 | 0.000 |
| S1-new A2 | 0.157 | 0.157 | 0.152 | 0.141 | 0.112 | 0.149 | 0.033 | 0.176 | 0.192 |
| S1-new A3 | 0.156 | 0.156 | 0.151 | 0.140 | 0.110 | 0.147 | 0.033 | 0.176 | 0.192 |
| S1-new A4 | 0.140 | 0.140 | 0.127 | 0.120 | 0.102 | 0.168 | -0.002 | 0.100 | 0.141 |
| S1-new A5 | -0.384 | -0.384 | -0.329 | -0.307 | -0.234 | -0.168 | -0.029 | -0.021 | -0.064 |
| S1-new A6 | -0.379 | -0.379 | -0.323 | -0.302 | -0.228 | -0.161 | -0.029 | -0.021 | -0.064 |
| S1-new A7 | 0.191 | 0.191 | 0.179 | 0.169 | 0.149 | 0.200 | 0.015 | 0.108 | 0.205 |
| S1-new A8 | 0.152 | 0.152 | 0.146 | 0.135 | 0.110 | 0.142 | 0.018 | 0.164 | 0.179 |
| S2 A1 | -0.331 | -0.331 | -0.344 | -0.370 | -0.324 | -0.228 | -0.344 | -0.077 | 0.000 |
| S2 A2 | 0.167 | 0.167 | 0.162 | 0.151 | 0.136 | 0.164 | 0.053 | 0.179 | 0.218 |
| S2 A3 | 0.165 | 0.165 | 0.160 | 0.149 | 0.133 | 0.161 | 0.050 | 0.179 | 0.218 |
| S2 A4 | 0.058 | 0.058 | 0.043 | 0.029 | 0.064 | 0.142 | 0.037 | 0.169 | 0.192 |
| S2 A5 | -0.016 | -0.016 | -0.001 | -0.001 | 0.052 | 0.128 | 0.053 | 0.121 | 0.103 |
| S2 A6 | -0.055 | -0.055 | -0.041 | -0.042 | -0.004 | 0.064 | 0.043 | 0.137 | 0.090 |
| S2 A7 | 0.066 | 0.066 | 0.052 | 0.038 | 0.074 | 0.151 | 0.051 | 0.169 | 0.192 |
| S2 A8 | 0.089 | 0.089 | 0.082 | 0.069 | 0.107 | 0.162 | 0.040 | 0.174 | 0.205 |
| S3 A1 | -0.331 | -0.331 | -0.344 | -0.370 | -0.324 | -0.228 | -0.344 | -0.077 | 0.000 |
| S3 A2 | 0.141 | 0.141 | 0.135 | 0.124 | 0.116 | 0.144 | 0.026 | 0.160 | 0.192 |
| S3 A3 | 0.142 | 0.142 | 0.136 | 0.124 | 0.117 | 0.145 | 0.026 | 0.160 | 0.192 |
| S3 A4 | -0.052 | -0.052 | 0.009 | 0.000 | 0.043 | 0.107 | 0.046 | 0.142 | 0.173 |
| S3 A5 | -0.067 | -0.067 | -0.007 | -0.007 | 0.014 | 0.072 | 0.021 | 0.124 | 0.131 |
| S3 A6 | -0.092 | -0.092 | -0.032 | -0.007 | -0.001 | 0.042 | 0.039 | 0.117 | 0.093 |
| S3 A7 | -0.045 | -0.045 | 0.016 | 0.008 | 0.043 | 0.104 | 0.040 | 0.132 | 0.160 |
| S3 A8 | 0.141 | 0.141 | 0.135 | 0.124 | 0.116 | 0.144 | 0.026 | 0.160 | 0.192 |

Table 19: Reconstructed policies, domain A. Cell-equal utilization for strict gap thresholds. An asterisk marks incomplete historical coverage, not a full eligible-set estimate. No confidence intervals.

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| Greedy-new Greedy | -0.351 | -0.351 | -0.367 | -0.318 | -0.271 | -0.172 | -0.158 | 0.000 | – |
| S1-new A1 | -0.389 | -0.389 | -0.431 | -0.529 | -0.411 | -0.146 | -0.099 | -0.379 | – |
| S1-new A2 | 0.253 | 0.253 | 0.241 | 0.214 | 0.123 | 0.124 | 0.117 | -0.069 | – |
| S1-new A3 | 0.253 | 0.253 | 0.241 | 0.214 | 0.123 | 0.124 | 0.117 | -0.069 | – |
| S1-new A4 | 0.196 | 0.196 | 0.158 | 0.139 | 0.064 | 0.084 | 0.089 | 0.034 | – |
| S1-new A5 | -0.649 | -0.649 | -0.489 | -0.436 | -0.163 | 0.013 | 0.018 | 0.000 | – |
| S1-new A6 | -0.649 | -0.649 | -0.489 | -0.436 | -0.163 | 0.013 | 0.018 | 0.000 | – |
| S1-new A7 | 0.250 | 0.250 | 0.214 | 0.184 | 0.096 | 0.138 | 0.112 | -0.138 | – |
| S1-new A8 | 0.243 | 0.243 | 0.231 | 0.203 | 0.126 | 0.133 | 0.106 | -0.103 | – |
| S2 A1 | -0.389 | -0.389 | -0.431 | -0.529 | -0.411 | -0.146 | -0.099 | -0.379 | – |
| S2 A2 | 0.221 | 0.221 | 0.208 | 0.177 | 0.119 | 0.115 | 0.106 | -0.103 | – |
| S2 A3 | 0.221 | 0.221 | 0.208 | 0.177 | 0.119 | 0.115 | 0.106 | -0.103 | – |
| S2 A4 | -0.013 | -0.013 | -0.062 | -0.121 | -0.056 | 0.104 | 0.091 | -0.103 | – |
| S2 A5 | -0.229 | -0.229 | -0.192 | -0.212 | -0.100 | 0.104 | 0.091 | -0.103 | – |
| S2 A6 | -0.184 | -0.184 | -0.145 | -0.160 | -0.055 | 0.141 | 0.140 | 0.000 | – |
| S2 A7 | -0.013 | -0.013 | -0.062 | -0.121 | -0.056 | 0.104 | 0.091 | -0.103 | – |
| S2 A8 | -0.005 | -0.005 | -0.029 | -0.085 | -0.006 | 0.115 | 0.106 | -0.103 | – |
| S3 A1 | -0.389 | -0.389 | -0.431 | -0.529 | -0.411 | -0.146 | -0.099 | -0.379 | – |
| S3 A2 | 0.169 | 0.169 | 0.153 | 0.117 | 0.075 | 0.078 | 0.092 | -0.147 | – |
| S3 A3 | 0.169 | 0.169 | 0.153 | 0.117 | 0.075 | 0.078 | 0.092 | -0.147 | – |
| S3 A4 | -0.329 | -0.329 | -0.154 | -0.201 | -0.133 | 0.077 | 0.079 | -0.138 | – |
| S3 A5 | -0.288 | -0.288 | -0.111 | -0.123 | -0.100 | 0.086 | 0.092 | -0.147 | – |
| S3 A6 | -0.288 | -0.288 | -0.111 | -0.039 | -0.027 | 0.115 | 0.106 | -0.103 | – |
| S3 A7 | -0.310 | -0.310 | -0.135 | -0.179 | -0.139 | 0.062 | 0.071 | -0.164 | – |
| S3 A8 | 0.169 | 0.169 | 0.153 | 0.117 | 0.075 | 0.078 | 0.092 | -0.147 | – |

Table 20: Reconstructed policies, domain B. Cell-equal utilization for strict gap thresholds. An asterisk marks incomplete historical coverage, not a full eligible-set estimate. No confidence intervals.

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| Greedy-new Greedy | -0.317 | -0.317 | -0.317 | -0.317 | -0.317 | -0.306 | -0.242 | -0.224 | -0.128 |
| S1-new A1 | 0.028 | 0.028 | 0.028 | 0.028 | 0.028 | -0.014 | -0.300 | -0.002 | 0.000 |
| S1-new A2 | 0.344 | 0.344 | 0.344 | 0.344 | 0.344 | 0.317 | 0.130 | 0.237 | 0.192 |
| S1-new A3 | 0.344 | 0.344 | 0.344 | 0.344 | 0.344 | 0.317 | 0.130 | 0.237 | 0.192 |
| S1-new A4 | 0.391 | 0.391 | 0.391 | 0.391 | 0.391 | 0.362 | 0.042 | 0.117 | 0.141 |
| S1-new A5 | -0.221 | -0.221 | -0.221 | -0.221 | -0.221 | -0.214 | -0.036 | -0.027 | -0.064 |
| S1-new A6 | -0.205 | -0.205 | -0.205 | -0.205 | -0.205 | -0.201 | -0.036 | -0.027 | -0.064 |
| S1-new A7 | 0.405 | 0.405 | 0.405 | 0.405 | 0.405 | 0.376 | 0.065 | 0.169 | 0.205 |
| S1-new A8 | 0.343 | 0.343 | 0.343 | 0.343 | 0.343 | 0.316 | 0.126 | 0.231 | 0.179 |
| S2 A1 | 0.028 | 0.028 | 0.028 | 0.028 | 0.028 | -0.014 | -0.300 | -0.002 | 0.000 |
| S2 A2 | 0.381 | 0.381 | 0.381 | 0.381 | 0.381 | 0.356 | 0.190 | 0.250 | 0.218 |
| S2 A3 | 0.381 | 0.381 | 0.381 | 0.381 | 0.381 | 0.356 | 0.190 | 0.250 | 0.218 |

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| S2 A4 | 0.363 | 0.363 | 0.363 | 0.363 | 0.363 | 0.341 | 0.190 | 0.237 | 0.192 |
| S2 A5 | 0.311 | 0.311 | 0.311 | 0.311 | 0.311 | 0.291 | 0.142 | 0.177 | 0.103 |
| S2 A6 | 0.150 | 0.150 | 0.150 | 0.150 | 0.150 | 0.174 | 0.138 | 0.171 | 0.090 |
| S2 A7 | 0.373 | 0.373 | 0.373 | 0.373 | 0.373 | 0.347 | 0.190 | 0.237 | 0.192 |
| S2 A8 | 0.379 | 0.379 | 0.379 | 0.379 | 0.379 | 0.354 | 0.187 | 0.244 | 0.205 |
| S3 A1 | 0.028 | 0.028 | 0.028 | 0.028 | 0.028 | -0.014 | -0.300 | -0.002 | 0.000 |
| S3 A2 | 0.366 | 0.366 | 0.366 | 0.366 | 0.366 | 0.340 | 0.169 | 0.237 | 0.192 |
| S3 A3 | 0.366 | 0.366 | 0.366 | 0.366 | 0.366 | 0.340 | 0.169 | 0.237 | 0.192 |
| S3 A4 | 0.260 | 0.260 | 0.260 | 0.260 | 0.260 | 0.242 | 0.131 | 0.212 | 0.173 |
| S3 A5 | 0.212 | 0.212 | 0.212 | 0.212 | 0.212 | 0.195 | 0.112 | 0.192 | 0.131 |
| S3 A6 | 0.137 | 0.137 | 0.137 | 0.137 | 0.137 | 0.117 | 0.124 | 0.172 | 0.093 |
| S3 A7 | 0.261 | 0.261 | 0.261 | 0.261 | 0.261 | 0.241 | 0.124 | 0.206 | 0.160 |
| S3 A8 | 0.366 | 0.366 | 0.366 | 0.366 | 0.366 | 0.340 | 0.169 | 0.237 | 0.192 |

Table 21: Reconstructed policies, domain C. Cell-equal utilization for strict gap thresholds. An asterisk marks incomplete historical coverage, not a full eligible-set estimate. No confidence intervals.

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| Greedy-new Greedy | -0.282 | -0.282 | -0.282 | -0.282 | -0.282 | -0.244 | -0.313 | – | – |
| S1-new A1 | -0.632 | -0.632 | -0.632 | -0.632 | -0.632 | -0.571 | -0.493 | – | – |
| S1-new A2 | -0.126 | -0.126 | -0.126 | -0.126 | -0.126 | -0.096 | -0.101 | – | – |
| S1-new A3 | -0.129 | -0.129 | -0.129 | -0.129 | -0.129 | -0.102 | -0.101 | – | – |
| S1-new A4 | -0.168 | -0.168 | -0.168 | -0.168 | -0.168 | -0.101 | -0.086 | – | – |
| S1-new A5 | -0.283 | -0.283 | -0.283 | -0.283 | -0.283 | -0.150 | -0.043 | – | – |
| S1-new A6 | -0.283 | -0.283 | -0.283 | -0.283 | -0.283 | -0.150 | -0.043 | – | – |
| S1-new A7 | -0.080 | -0.080 | -0.080 | -0.080 | -0.080 | -0.046 | -0.077 | – | – |
| S1-new A8 | -0.132 | -0.132 | -0.132 | -0.132 | -0.132 | -0.115 | -0.128 | – | – |
| S2 A1 | -0.632 | -0.632 | -0.632 | -0.632 | -0.632 | -0.571 | -0.493 | – | – |
| S2 A2 | -0.102 | -0.102 | -0.102 | -0.102 | -0.102 | -0.111 | -0.108 | – | – |
| S2 A3 | -0.107 | -0.107 | -0.107 | -0.107 | -0.107 | -0.120 | -0.115 | – | – |
| S2 A4 | -0.176 | -0.176 | -0.176 | -0.176 | -0.176 | -0.146 | -0.138 | – | – |
| S2 A5 | -0.131 | -0.131 | -0.131 | -0.131 | -0.131 | -0.109 | -0.053 | – | – |
| S2 A6 | -0.132 | -0.132 | -0.132 | -0.132 | -0.132 | -0.123 | -0.095 | – | – |
| S2 A7 | -0.160 | -0.160 | -0.160 | -0.160 | -0.160 | -0.129 | -0.104 | – | – |
| S2 A8 | -0.108 | -0.108 | -0.108 | -0.108 | -0.108 | -0.113 | -0.135 | – | – |
| S3 A1 | -0.632 | -0.632 | -0.632 | -0.632 | -0.632 | -0.571 | -0.493 | – | – |
| S3 A2 | -0.113 | -0.113 | -0.113 | -0.113 | -0.113 | -0.131 | -0.146 | – | – |
| S3 A3 | -0.110 | -0.110 | -0.110 | -0.110 | -0.110 | -0.127 | -0.146 | – | – |
| S3 A4 | -0.086 | -0.086 | -0.086 | -0.086 | -0.086 | -0.087 | -0.053 | – | – |
| S3 A5 | -0.126 | -0.126 | -0.126 | -0.126 | -0.126 | -0.115 | -0.100 | – | – |
| S3 A6 | -0.125 | -0.125 | -0.125 | -0.125 | -0.125 | -0.092 | -0.074 | – | – |
| S3 A7 | -0.085 | -0.085 | -0.085 | -0.085 | -0.085 | -0.090 | -0.057 | – | – |
| S3 A8 | -0.113 | -0.113 | -0.113 | -0.113 | -0.113 | -0.131 | -0.146 | – | – |

Table 22: Historical aggregate records, domain ALL. Cell-equal utilization for strict gap thresholds. An asterisk marks incomplete historical coverage, not a full eligible-set estimate. No confidence intervals.

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| Greedy-old Greedy | -0.227* | -0.227* | -0.227 | -0.213 | -0.200 | -0.201 | -0.211 | -0.146 | -0.128 |
| Judge-fixed judge_llama-3.1-8b | -0.028* | -0.028* | -0.028 | -0.023 | 0.032 | 0.006 | -0.057 | 0.102 | 0.090 |
| Judge-fixed judge_qwen3-32b | 0.162* | 0.162* | 0.162 | 0.151 | 0.205 | 0.216 | 0.069 | 0.194 | 0.205 |
| S1-old A1 | -0.344* | -0.344* | -0.344 | -0.370 | -0.324 | -0.228 | -0.344 | -0.077 | 0.000 |
| S1-old A2 | 0.174* | 0.174* | 0.174 | 0.164 | 0.139 | 0.157 | 0.048 | 0.164 | 0.179 |
| S1-old A3 | 0.174* | 0.174* | 0.174 | 0.164 | 0.138 | 0.157 | 0.051 | 0.164 | 0.179 |
| S1-old A4 | 0.110* | 0.110* | 0.110 | 0.103 | 0.090 | 0.113 | -0.029 | 0.019 | 0.128 |
| S1-old A5 | -0.325* | -0.325* | -0.325 | -0.304 | -0.230 | -0.169 | -0.029 | -0.021 | -0.064 |
| S1-old A6 | -0.319* | -0.319* | -0.319 | -0.297 | -0.223 | -0.154 | -0.029 | -0.021 | -0.064 |
| S1-old A7 | 0.175* | 0.175* | 0.175 | 0.165 | 0.144 | 0.169 | 0.021 | 0.144 | 0.218 |
| S1-old A8 | 0.166* | 0.166* | 0.166 | 0.156 | 0.134 | 0.150 | 0.032 | 0.169 | 0.192 |
| S2 A1 | -0.344* | -0.344* | -0.344 | -0.370 | -0.324 | -0.228 | -0.344 | -0.077 | 0.000 |
| S2 A2 | 0.162* | 0.162* | 0.162 | 0.151 | 0.136 | 0.164 | 0.053 | 0.179 | 0.218 |
| S2 A3 | 0.160* | 0.160* | 0.160 | 0.149 | 0.133 | 0.161 | 0.050 | 0.179 | 0.218 |
| S2 A4 | 0.043* | 0.043* | 0.043 | 0.029 | 0.064 | 0.142 | 0.037 | 0.169 | 0.192 |
| S2 A5 | -0.001* | -0.001* | -0.001 | -0.001 | 0.052 | 0.128 | 0.053 | 0.121 | 0.103 |
| S2 A6 | -0.041* | -0.041* | -0.041 | -0.042 | -0.004 | 0.064 | 0.043 | 0.137 | 0.090 |
| S2 A7 | 0.052* | 0.052* | 0.052 | 0.038 | 0.074 | 0.151 | 0.051 | 0.169 | 0.192 |
| S2 A8 | 0.082* | 0.082* | 0.082 | 0.069 | 0.107 | 0.162 | 0.040 | 0.174 | 0.205 |
| S3 A1 | -0.344* | -0.344* | -0.344 | -0.370 | -0.324 | -0.228 | -0.344 | -0.077 | 0.000 |
| S3 A2 | 0.135* | 0.135* | 0.135 | 0.124 | 0.116 | 0.144 | 0.026 | 0.160 | 0.192 |
| S3 A3 | 0.136* | 0.136* | 0.136 | 0.124 | 0.117 | 0.145 | 0.026 | 0.160 | 0.192 |
| S3 A4 | 0.009* | 0.009* | 0.009 | 0.000 | 0.043 | 0.107 | 0.046 | 0.142 | 0.173 |
| S3 A5 | -0.007* | -0.007* | -0.007 | -0.007 | 0.014 | 0.072 | 0.021 | 0.124 | 0.131 |
| S3 A6 | -0.032* | -0.032* | -0.032 | -0.007 | -0.001 | 0.042 | 0.039 | 0.117 | 0.093 |
| S3 A7 | 0.016* | 0.016* | 0.016 | 0.008 | 0.043 | 0.104 | 0.040 | 0.132 | 0.160 |
| S3 A8 | 0.135* | 0.135* | 0.135 | 0.124 | 0.116 | 0.144 | 0.026 | 0.160 | 0.192 |

Table 23: Historical aggregate records, domain A. Cell-equal utilization for strict gap thresholds. An asterisk marks incomplete historical coverage, not a full eligible-set estimate. No confidence intervals.

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| Greedy-old Greedy | -0.280* | -0.280* | -0.280 | -0.239 | -0.191 | -0.181 | -0.123 | 0.000 | – |
| Judge-fixed judge_llama-3.1-8b | -0.133* | -0.133* | -0.133 | -0.129 | 0.070 | 0.090 | 0.048 | -0.103 | – |
| Judge-fixed judge_qwen3-32b | 0.037* | 0.037* | 0.037 | -0.012 | 0.140 | 0.219 | 0.244 | -0.035 | – |
| S1-old A1 | -0.431* | -0.431* | -0.431 | -0.529 | -0.411 | -0.146 | -0.099 | -0.379 | – |
| S1-old A2 | 0.236* | 0.236* | 0.236 | 0.208 | 0.113 | 0.097 | 0.106 | -0.103 | – |
| S1-old A3 | 0.233* | 0.233* | 0.233 | 0.205 | 0.106 | 0.079 | 0.106 | -0.103 | – |
| S1-old A4 | 0.132* | 0.132* | 0.132 | 0.111 | 0.051 | 0.050 | 0.114 | -0.034 | – |
| S1-old A5 | -0.489* | -0.489* | -0.489 | -0.436 | -0.163 | 0.013 | 0.018 | 0.000 | – |
| S1-old A6 | -0.489* | -0.489* | -0.489 | -0.436 | -0.163 | 0.013 | 0.018 | 0.000 | – |
| S1-old A7 | 0.198* | 0.198* | 0.198 | 0.167 | 0.064 | 0.106 | 0.117 | -0.069 | – |
| S1-old A8 | 0.228* | 0.228* | 0.228 | 0.199 | 0.119 | 0.115 | 0.106 | -0.103 | – |
| S2 A1 | -0.431* | -0.431* | -0.431 | -0.529 | -0.411 | -0.146 | -0.099 | -0.379 | – |
| S2 A2 | 0.208* | 0.208* | 0.208 | 0.177 | 0.119 | 0.115 | 0.106 | -0.103 | – |
| S2 A3 | 0.208* | 0.208* | 0.208 | 0.177 | 0.119 | 0.115 | 0.106 | -0.103 | – |
| S2 A4 | -0.062* | -0.062* | -0.062 | -0.121 | -0.056 | 0.104 | 0.091 | -0.103 | – |
| S2 A5 | -0.192* | -0.192* | -0.192 | -0.212 | -0.100 | 0.104 | 0.091 | -0.103 | – |
| S2 A6 | -0.145* | -0.145* | -0.145 | -0.160 | -0.055 | 0.141 | 0.140 | 0.000 | – |
| S2 A7 | -0.062* | -0.062* | -0.062 | -0.121 | -0.056 | 0.104 | 0.091 | -0.103 | – |
| S2 A8 | -0.029* | -0.029* | -0.029 | -0.085 | -0.006 | 0.115 | 0.106 | -0.103 | – |
| S3 A1 | -0.431* | -0.431* | -0.431 | -0.529 | -0.411 | -0.146 | -0.099 | -0.379 | – |
| S3 A2 | 0.153* | 0.153* | 0.153 | 0.117 | 0.075 | 0.078 | 0.092 | -0.147 | – |
| S3 A3 | 0.153* | 0.153* | 0.153 | 0.117 | 0.075 | 0.078 | 0.092 | -0.147 | – |
| S3 A4 | -0.154* | -0.154* | -0.154 | -0.201 | -0.133 | 0.077 | 0.079 | -0.138 | – |
| S3 A5 | -0.111* | -0.111* | -0.111 | -0.123 | -0.100 | 0.086 | 0.092 | -0.147 | – |
| S3 A6 | -0.111* | -0.111* | -0.111 | -0.039 | -0.027 | 0.115 | 0.106 | -0.103 | – |
| S3 A7 | -0.135* | -0.135* | -0.135 | -0.179 | -0.139 | 0.062 | 0.071 | -0.164 | – |
| S3 A8 | 0.153* | 0.153* | 0.153 | 0.117 | 0.075 | 0.078 | 0.092 | -0.147 | – |

Table 24: Historical aggregate records, domain B. Cell-equal utilization for strict gap thresholds. An asterisk marks incomplete historical coverage, not a full eligible-set estimate. No confidence intervals.

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| Greedy-old Greedy | -0.208 | -0.208 | -0.208 | -0.208 | -0.208 | -0.192 | -0.190 | -0.182 | -0.128 |
| Judge-fixed judge_llama-3.1-8b | 0.060 | 0.060 | 0.060 | 0.060 | 0.060 | 0.050 | 0.040 | 0.154 | 0.090 |
| Judge-fixed judge_qwen3-32b | 0.390 | 0.390 | 0.390 | 0.390 | 0.390 | 0.365 | 0.185 | 0.251 | 0.205 |
| S1-old A1 | 0.028 | 0.028 | 0.028 | 0.028 | 0.028 | -0.014 | -0.300 | -0.002 | 0.000 |
| S1-old A2 | 0.364 | 0.364 | 0.364 | 0.364 | 0.364 | 0.338 | 0.186 | 0.231 | 0.179 |
| S1-old A3 | 0.364 | 0.364 | 0.364 | 0.364 | 0.364 | 0.338 | 0.186 | 0.231 | 0.179 |
| S1-old A4 | 0.363 | 0.363 | 0.363 | 0.363 | 0.363 | 0.333 | 0.038 | 0.032 | 0.128 |
| S1-old A5 | -0.211 | -0.211 | -0.211 | -0.211 | -0.211 | -0.217 | -0.036 | -0.027 | -0.064 |
| S1-old A6 | -0.192 | -0.192 | -0.192 | -0.192 | -0.192 | -0.188 | -0.036 | -0.027 | -0.064 |
| S1-old A7 | 0.409 | 0.409 | 0.409 | 0.409 | 0.409 | 0.381 | 0.126 | 0.197 | 0.218 |
| S1-old A8 | 0.369 | 0.369 | 0.369 | 0.369 | 0.369 | 0.343 | 0.190 | 0.237 | 0.192 |
| S2 A1 | 0.028 | 0.028 | 0.028 | 0.028 | 0.028 | -0.014 | -0.300 | -0.002 | 0.000 |
| S2 A2 | 0.381 | 0.381 | 0.381 | 0.381 | 0.381 | 0.356 | 0.190 | 0.250 | 0.218 |
| S2 A3 | 0.381 | 0.381 | 0.381 | 0.381 | 0.381 | 0.356 | 0.190 | 0.250 | 0.218 |
| S2 A4 | 0.363 | 0.363 | 0.363 | 0.363 | 0.363 | 0.341 | 0.190 | 0.237 | 0.192 |
| S2 A5 | 0.311 | 0.311 | 0.311 | 0.311 | 0.311 | 0.291 | 0.142 | 0.177 | 0.103 |
| S2 A6 | 0.150 | 0.150 | 0.150 | 0.150 | 0.150 | 0.174 | 0.138 | 0.171 | 0.090 |
| S2 A7 | 0.373 | 0.373 | 0.373 | 0.373 | 0.373 | 0.347 | 0.190 | 0.237 | 0.192 |
| S2 A8 | 0.379 | 0.379 | 0.379 | 0.379 | 0.379 | 0.354 | 0.187 | 0.244 | 0.205 |
| S3 A1 | 0.028 | 0.028 | 0.028 | 0.028 | 0.028 | -0.014 | -0.300 | -0.002 | 0.000 |
| S3 A2 | 0.366 | 0.366 | 0.366 | 0.366 | 0.366 | 0.340 | 0.169 | 0.237 | 0.192 |
| S3 A3 | 0.366 | 0.366 | 0.366 | 0.366 | 0.366 | 0.340 | 0.169 | 0.237 | 0.192 |
| S3 A4 | 0.260 | 0.260 | 0.260 | 0.260 | 0.260 | 0.242 | 0.131 | 0.212 | 0.173 |
| S3 A5 | 0.212 | 0.212 | 0.212 | 0.212 | 0.212 | 0.195 | 0.112 | 0.192 | 0.131 |
| S3 A6 | 0.137 | 0.137 | 0.137 | 0.137 | 0.137 | 0.117 | 0.124 | 0.172 | 0.093 |
| S3 A7 | 0.261 | 0.261 | 0.261 | 0.261 | 0.261 | 0.241 | 0.124 | 0.206 | 0.160 |
| S3 A8 | 0.366 | 0.366 | 0.366 | 0.366 | 0.366 | 0.340 | 0.169 | 0.237 | 0.192 |

Table 25: Historical aggregate records, domain C. Cell-equal utilization for strict gap thresholds. An asterisk marks incomplete historical coverage, not a full eligible-set estimate. No confidence intervals.

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| Greedy-old Greedy | -0.195 | -0.195 | -0.195 | -0.195 | -0.195 | -0.221 | -0.269 | – | – |
| Judge-fixed judge_llama-3.1-8b | -0.015 | -0.015 | -0.015 | -0.015 | -0.015 | -0.084 | -0.199 | – | – |
| Judge-fixed judge_qwen3-32b | 0.054 | 0.054 | 0.054 | 0.054 | 0.054 | -0.007 | -0.122 | – | – |
| S1-old A1 | -0.632 | -0.632 | -0.632 | -0.632 | -0.632 | -0.571 | -0.493 | – | – |
| S1-old A2 | -0.074 | -0.074 | -0.074 | -0.074 | -0.074 | -0.097 | -0.114 | – | – |
| S1-old A3 | -0.072 | -0.072 | -0.072 | -0.072 | -0.072 | -0.094 | -0.107 | – | – |
| S1-old A4 | -0.164 | -0.164 | -0.164 | -0.164 | -0.164 | -0.198 | -0.157 | – | – |

| Policy | 0 | 0.01 | 0.02 | 0.03 | 0.05 | 0.1 | 0.15 | 0.2 | 0.3 |
|---|---|---|---|---|---|---|---|---|---|
| S1-old A5 | -0.283 | -0.283 | -0.283 | -0.283 | -0.283 | -0.150 | -0.043 | – | – |
| S1-old A6 | -0.283 | -0.283 | -0.283 | -0.283 | -0.283 | -0.150 | -0.043 | – | – |
| S1-old A7 | -0.082 | -0.082 | -0.082 | -0.082 | -0.082 | -0.130 | -0.124 | – | – |
| S1-old A8 | -0.095 | -0.095 | -0.095 | -0.095 | -0.095 | -0.130 | -0.157 | – | – |
| S2 A1 | -0.632 | -0.632 | -0.632 | -0.632 | -0.632 | -0.571 | -0.493 | – | – |
| S2 A2 | -0.102 | -0.102 | -0.102 | -0.102 | -0.102 | -0.111 | -0.108 | – | – |
| S2 A3 | -0.107 | -0.107 | -0.107 | -0.107 | -0.107 | -0.120 | -0.115 | – | – |
| S2 A4 | -0.176 | -0.176 | -0.176 | -0.176 | -0.176 | -0.146 | -0.138 | – | – |
| S2 A5 | -0.131 | -0.131 | -0.131 | -0.131 | -0.131 | -0.109 | -0.053 | – | – |
| S2 A6 | -0.132 | -0.132 | -0.132 | -0.132 | -0.132 | -0.123 | -0.095 | – | – |
| S2 A7 | -0.160 | -0.160 | -0.160 | -0.160 | -0.160 | -0.129 | -0.104 | – | – |
| S2 A8 | -0.108 | -0.108 | -0.108 | -0.108 | -0.108 | -0.113 | -0.135 | – | – |
| S3 A1 | -0.632 | -0.632 | -0.632 | -0.632 | -0.632 | -0.571 | -0.493 | – | – |
| S3 A2 | -0.113 | -0.113 | -0.113 | -0.113 | -0.113 | -0.131 | -0.146 | – | – |
| S3 A3 | -0.110 | -0.110 | -0.110 | -0.110 | -0.110 | -0.127 | -0.146 | – | – |
| S3 A4 | -0.086 | -0.086 | -0.086 | -0.086 | -0.086 | -0.087 | -0.053 | – | – |
| S3 A5 | -0.126 | -0.126 | -0.126 | -0.126 | -0.126 | -0.115 | -0.100 | – | – |
| S3 A6 | -0.125 | -0.125 | -0.125 | -0.125 | -0.125 | -0.092 | -0.074 | – | – |
| S3 A7 | -0.085 | -0.085 | -0.085 | -0.085 | -0.085 | -0.090 | -0.057 | – | – |
| S3 A8 | -0.113 | -0.113 | -0.113 | -0.113 | -0.113 | -0.131 | -0.146 | – | – |