CROSS-FAMILY EXCHANGE UNDER MATCHED CALL BUDGETS: SAMPLING, REVISION, AND TIE RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Does cross-family exchange improve a language-model panel beyond independent sampling or revisiting its own responses? We compare Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct on 200 GSM8K questions and a 100-question rational-answer MATH subset. For each question, 16 saved generations support four-call conditions that share initial answers, model composition, output ceilings, and aggregation. Both proposer orders are evaluated. Averaging the two orders within each question, exchange trails independent voting by 3.0 percentage points on GSM8K (descriptive 95% paired-bootstrap interval $[-5.5, -0.5]$) and 1.5 points on MATH ($[-6.0, 3.0]$). Relative to own-response revision, the differences are $+0.25$ and $+4.5$ points; neither interval excludes zero. Exchange adds 603 and 1,405 input tokens per question over independent voting. A post hoc order-invariant tie sensitivity reduces the MATH advantage over own-response revision to 1.1 points: answer order affects the fixed vote even without changing generated content. Shared valid initial agreement also fixes the final vote regardless of subsequent answers. These results identify restrictions of the decision procedure, not a general benefit of exchange or cognitive specialization.

1 INTRODUCTION

Multi-agent debate exchanges responses before a final decision, whereas voting aggregates independently generated answers (Du et al., 2024). Frameworks such as AutoGen and CAMEL make interaction easy to implement (Wu et al., 2024; Li et al., 2023), but an observed accuracy gain does not identify its cause. A compared system may use a stronger panel, more calls, another aggregation rule, or additional opportunities to revise. Even a fixed four-answer vote can change when ties are resolved in a different response order.

We ask how much exchanging responses adds for a fixed two-family panel with a four-call budget. The experiment shares initial answers between conditions and evaluates both proposer orders. Independent heterogeneous voting controls sampling; an own-response revision condition controls a related opportunity to revisit an answer without receiving another family's response. We retain realized input and output tokens rather than assuming equal calls mean equal computation. The 300 questions span two arithmetic subsets, and 4,800 generations preserve the complete evidence for every condition.

Choi et al. (2025) already separate majority voting from inter-agent debate across several benchmarks. Showing that voting is competitive is therefore not a new contribution by itself. Our contribution is the combined paired comparison of information flow, own-response revision, role order, and numerical aggregation. The controls remain imperfect for a pure communication effect: exchanged contexts differ in content and length, and the final call has a sequential dependency absent from independent sampling. The resulting contrast measures the complete exchange protocol, including these changes in context and dependency.

The results identify two restrictions of the final decision. First, when the two initial answers agree, the earliest-answer tie rule fixes the winner even if both later answers correct it. This occurs on 134 of 200 GSM8K questions. Second, permuting unchanged independent answers changes correctness on six of 100 MATH questions. Averaging uniformly over tied answers reduces the MATH exchange advantage

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.

over own-response revision from 4.5 to 1.125 points. These observations locate limits in the decision procedure itself, beyond the overall finding that exchange does not outperform independent voting.

2 Related work

Sampling and aggregation. Chain-of-thought prompting elicits intermediate reasoning (Wei et al., 2022), and self-consistency aggregates multiple sampled paths (Wang et al., 2023). These methods make independently sampled answers an essential baseline for any multi-call discussion protocol. Increasing the number of agents can also improve an ensemble through aggregation (Li et al., 2024a); the existence of such gains does not identify a benefit from exchanging messages. The fixed four-call comparison here asks about allocating calls between independent generation and review, not whether four samples outperform one.

Heterogeneous ensembles can rank and fuse candidate answers, as in LLM-Blender (Jiang et al., 2023), or pass previous-layer answers to later agents, as in Mixture-of-Agents (Wang et al., 2025). Ranking, fusion, and voting are different aggregation operations. We hold a simple numerical vote fixed to avoid conflating an exchange intervention with a learned or generative aggregator. This choice narrows the conclusion: the study does not test those systems or claim that a final synthesis model is interchangeable with a vote.

Debate and communication design. Multiagent debate can exchange candidate reasoning before producing an answer (Du et al., 2024). Divergent-thinking debate (Liang et al., 2024) and sparse communication topologies (Li et al., 2024b) vary how agents interact. ChatEval instead studies multi-agent discussion as part of an evaluation procedure (Chan et al., 2024). The tasks, message graphs, and final decision rules differ across these settings. A negative or inconclusive result for our two-stage exchange therefore cannot establish a universal limitation of debate. Conversely, results from a richer interaction graph do not supply evidence for the particular four-call panel studied here.

Revision and feedback sources. Self-Refine uses a model's own feedback to revise an output (Madaan et al., 2023). Intrinsic self-correction can be unreliable on reasoning tasks without external feedback (Huang et al., 2024), whereas CRITIC uses tool-based feedback (Gou et al., 2024) and Reflexion maintains verbal feedback across attempts (Shinn et al., 2023). These distinctions motivate separating access to another model's answer from simply receiving another opportunity to answer. A communication comparison must specify the source of feedback and the information available at each call. Our exchange provides another family's response, without a trusted external verifier or a trained feedback policy.

Models, mathematical tasks, and scoring. Qwen2.5 (Qwen et al., 2024) and Llama 3 (Grattafiori et al., 2024) provide different model families, but family membership also changes pretraining, post-training, and individual task competence. It is not an independently randomized treatment. The MATH benchmark (Hendrycks et al., 2021) and quantitative-reasoning work such as Minerva (Lewkowycz et al., 2022) include mathematical outputs whose evaluation can require more than numerical string matching. GSM8K supplies numerical reference answers, and our MATH subset is restricted to rational answers. The checker normalizes these numbers; it does not assess proof validity or general symbolic equivalence. This restriction determines which tasks the reported accuracy can describe.

Statistical evidence. Paired comparisons preserve the common question and shared initial answers (Dror et al., 2018). Our bootstrap intervals (Efron, 1979) resample whole questions and are descriptive and unadjusted. Exact paired binary tests depend on discordant outcomes (McNemar, 1947), whereas our role-averaged scores may be fractional. Holm correction (Holm, 1979) addresses a prespecified family of tests; it is not applied to the descriptive intervals reported here. Uncertainty across questions remains distinct from replication across models and generation seeds.

3 A PAIRED INFORMATION-FLOW EXPERIMENT

3.1 MODELS, TASKS, AND SAMPLING

We use Qwen2.5-7B-Instruct (Qwen et al., 2024) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with bfloat16 weights, their native chat templates, and the same SDPA attention backend. All calls use temperature 0.7, top-p 0.95, top-k zero, and a 1,024-token output ceiling. Each question–call pair has a fixed seed. A paired exchange and own-response call use the same seed and model, but different contexts can still produce different samples. Parameter counts and pretraining are not matched between families, so family comparisons are descriptive.

We select 200 GSM8K test questions (Cobbe et al., 2021) by a fixed hash ordering after excluding the 100 questions in an earlier development comparison. No model outcomes enter that selection. The second set contains 100 previously selected MATH-Hard questions with rational final answers. Their inputs overlap an earlier cache study, but all generations here are fresh. The MATH set is a restricted task subset, not an independent replication on the full benchmark. GSM8K and MATH results are analyzed separately; pooling them would impose an arbitrary task mixture. All prompts request a boxed numerical answer and an exact fraction when appropriate.

3.2 SHARED STARTS AND THREE FOUR-CALL CONDITIONS

Let A be the first family and B the other family. We generate four independent initial answers from each family, denoted $a_1, \ldots, a_4$ and $b_1, \ldots, b_4$. The heterogeneous conditions share a_1, b_1. Writing $B[u]$ for a new response from B given the question and supplied text u, their four voting positions are

$$\mathcal{P}_{\text{ind}} = (a_1, b_1, a_2, b_2), \tag{1}$$

$$\mathcal{P}_{\text{exch}} = (a_1, b_1, B[a_1], A[B[a_1]]), \tag{2}$$

$$\mathcal{P}_{\text{own}} = (a_1, b_1, B[b_1], A[a_1]). \tag{3}$$

The supplied text appears inside fixed solution or review delimiters, and the original question is repeated in every call. The exchange asks B to check A's answer and then asks A to check that review. The final exchange call receives the review, not a concatenation of all previous answers. The own-response condition uses the corresponding templates and family call counts, but each family receives its own initial answer. Its two revisions are independent conditional on the shared starts: the last call receives a_1. The exchange instead passes $B[a_1]$ to that call, changing both the source and the processing history of the supplied text.

Both A =Qwen and A =Llama orders are evaluated for every question. Initial answers are shared between orders; only the voting order and the newly conditioned calls differ. Four independent answers from each single family provide additional panel baselines. Eight independent calls plus eight conditioned calls support all conditions, for 16 calls per question and 4,800 calls overall. The conditions overlap and are not independent experiments. The paired contrasts are computed within question before resampling.

3.3 EXTRACTION AND A FIXED DECISION RULE

A label-independent parser takes the last balanced boxed expression, falling back to an explicit final-answer field. Exact rational normalization merges supported fractions and decimals before counting votes. Invalid or absent answers abstain. The selected category is the most frequent nonmissing answer; if several categories tie, the earliest valid response among tied categories wins. An all-missing panel fails. This is a numerical-answer metric rather than a verifier of the full reasoning trace or arbitrary algebraic equivalence. Unparsed outputs remain part of the four-call outcome.

The fixed order matters. Reversing A, B in $\mathcal{P}_{\text{ind}}$ leaves its answer multiset unchanged, but can alter its tie winner. We retain this prespecified endpoint and report both role orders. A separate post hoc sensitivity averages uniformly over tied modal categories, isolating one aggregation choice without regenerating answers. This sensitivity is not a replacement endpoint selected to obtain a preferred contrast.

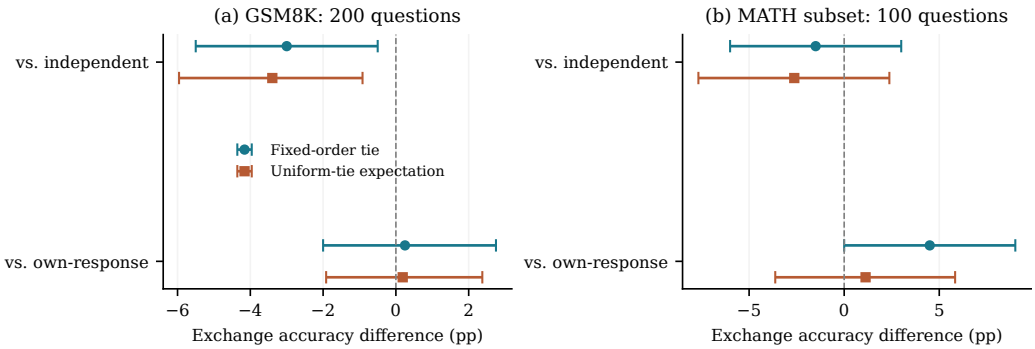

Figure 1: Exchange-minus-comparator accuracy, averaging both role orders within each question. Circles use the fixed earliest-answer tie rule; squares use the post hoc expected score under uniform tie resolution. Intervals resample whole questions and are unadjusted descriptive 95% intervals.

Table 1: Role-averaged fixed-vote accuracy and paired exchange contrasts. Intervals are in percentage points. Own-response denotes revisiting each family's initial answer under the corresponding prompts.

Quantity	GSM8K ($n = 200$)	MATH subset ($n = 100$)
Independent vote accuracy	92.0%	43.0%
Exchange accuracy	89.0%	41.5%
Own-response accuracy	88.75%	37.0%
Exchange − independent	-3.0 $[-5.5, -0.5]$	-1.5 $[-6.0, 3.0]$
Exchange − own-response	$+0.25$ $[-2.0, 2.75]$	$+4.5$ $[0.0, 9.0]$

3.4 Estimands and uncertainty

Let $Y^c_{i,k}$ be correctness on question i, role order $k \in \{0, 1\}$, and condition c. For comparator b, the primary task-wise contrast is

$$\widehat{\Delta}_b = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\sum_{k=0}^{1}(Y^{\text{exch}}_{i,k} - Y^{b}_{i,k}).$$

Paired discordance rather than marginal accuracy totals determines a binary paired test (McNemar, 1947); role-averaged outcomes here can also be fractional. Each question receives equal weight, and its two orders are averaged rather than treated as independent samples. We use 20,000 whole-question bootstrap draws for descriptive percentile 95% intervals (Efron, 1979; Dror et al., 2018). They are unadjusted for multiple comparisons; a family-wise analysis would require an explicit procedure such as Holm correction (Holm, 1979). They are conditional on the chosen task subsets and generation procedure. They do not quantify variation across alternative model pairs or fully repeated generation seeds.

We also report the difference between the two role-specific exchange-minus- independent contrasts. This is an interaction summary for the implemented protocol, not a test of stable cognitive roles. All comparisons preserve the same questions, initial answers, four-call family composition, and aggregation rule. They do not match FLOPs, realized context lengths, or sequential latency.

4 Outcomes under the fixed vote

4.1 Independent sampling remains a strong control

On GSM8K, exchange is three points below independent heterogeneous voting. Its unadjusted descriptive interval excludes zero for this question set. For each role order, exchange loses eight or nine independent successes while rescuing two or three failures. The net balance is negative in both orders. Against own-response revision, the role-averaged difference is only 0.25 points with an

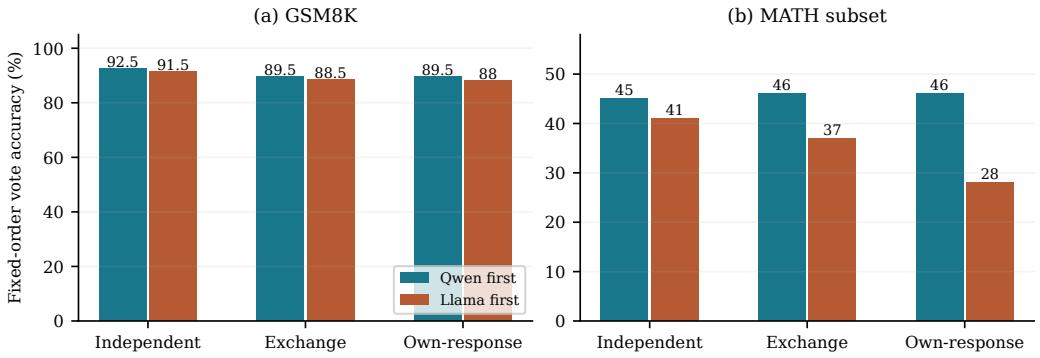

Figure 2: Both role orders under the fixed tie rule. The independent conditions contain the same answer multiset in different orders, so their accuracy difference is entirely due to tie resolution. Exchange and own-response conditions also contain different conditioned generations.

interval spanning zero. The observed calls therefore do not establish that another family's response adds useful information beyond the revision opportunity in this setting.

On the MATH subset, exchange trails independent voting by 1.5 points, with a wider interval. Its advantage over own-response revision is 4.5 points, with an interval whose lower endpoint is zero. These results do not establish superiority to either control. Qwen-only voting reaches 50% and Llama-only voting 21%, compared with 43% for the role-averaged heterogeneous vote. The asymmetry of constituent capabilities is substantial. A contrast against the weaker single-family baseline would change the panel and could not isolate an exchange effect. On GSM8K the single-family votes reach 92% and 89%.

4.2 Role order changes more than who reviews whom

On MATH, exchange accuracy is 46% with Qwen first and 37% with Llama first; own-response accuracy is 46% and 28%. The exchange-minus-own-response contrast is consequently zero in one order and nine points in the other. The two orders therefore give different empirical conclusions. Independent voting itself changes from 45% to 41% despite identical answer multisets. Six questions change correctness, with a four-point net difference, solely because the tie winner changes. GSM8K has two such questions and a one-point net difference.

The role interaction for exchange versus independent voting is zero points on GSM8K, with interval $[-4, 4]$, and five points on MATH, with interval $[-4, 14]$. Neither identifies a stable proposer/reviewer specialization. The role-specific point estimates mix conditioned generation with an order-sensitive final selector. The following sensitivity separates the selector component.

5 Tie resolution as a measurement sensitivity

5.1 An order-invariant expected score

Let $\mathcal{T}_i^c$ be the set of tied modal nonmissing answer categories for one panel, and let g_i be the gold answer. Define

$$U_i^c = \begin{cases} |\mathcal{T}_i^c|^{-1}\,\mathbf{1}\{g_i \in \mathcal{T}_i^c\}, & |\mathcal{T}_i^c| > 0, \\ 0, & |\mathcal{T}_i^c| = 0. \end{cases}$$

This is the expected correctness of uniformly selecting a tied category. Selection assigns equal probability to each tied category without using gold. Gold is then used to evaluate the resulting expected correctness. The exact fractional score integrates tie randomness without Monte Carlo draws. Averaging and resampling these scores uses the same question-level units as the original analysis.

This transformation is invariant to permutations of a fixed answer multiset. Both independent role orders therefore have exactly the same transformed score for every question, not merely the

same aggregate mean. We verify that identity on all 300 questions. It removes deterministic order preference while retaining missingness, vote counts, and the generated content. It does not correct parsing errors or make exchange responses independent.

5.2 THE MATH REVISION CONTRAST IS TIE-SENSITIVE

Independent voting has multiple tied modal categories in 4% of GSM8K panels and 32% of MATH panels, averaging the two role orders. Exchange tie rates are 2% and 9.5%; own-response rates are 7.5% and 22.5%. Thus the same tie rule can affect conditions differently. Fewer ties do not by themselves imply better answers: repeated agreement on an incorrect category also avoids a tie.

Under uniform tie resolution, MATH expected accuracy is 43.0% for independent voting, 40.375% for exchange, and 39.25% for own-response revision. The exchange-minus-own-response difference falls from 4.5 to 1.125 points, with descriptive interval $[-3.625, 5.834]$. The exchange-minus-independent difference is -2.625 points, with interval $[-7.667, 2.376]$. These are sensitivity results on the same saved generations, not additional independent evidence.

On GSM8K, the corresponding exchange contrasts are -3.396 points versus independent voting, with interval $[-5.958, -0.917]$, and $+0.188$ versus own-response revision, with interval $[-1.917, 2.375]$. The qualitative conclusion is unchanged there. The comparison across tasks shows why reporting one aggregation rule alone can obscure how much a small observed effect depends on selecting among equally frequent answers.

5.3 A BOUND ON TIE-INDUCED SCORE CHANGES

For one panel with $m = |\mathcal{T}| > 1$ tied modal categories, let F be the fixed-order score and U the uniform-tie expectation. If the gold category is absent, both scores are zero. If it is present, $U = 1/m$ and F is either zero or one. Hence

$$|F - U| \leq \left(1 - \frac{1}{m}\right) \mathbf{1}\{m > 1\}.$$

The inequality is attained when the fixed order selects the gold category. When it selects another category, the difference is $1/m$, which is no larger than $1 - 1/m$ for $m \geq 2$. This elementary bound describes the selector alone; it makes no assumption about how the model produced its answers.

For a contrast between two conditions a, b, the triangle inequality gives

$$\left|(\overline{F}_a - \overline{F}_b) - (\overline{U}_a - \overline{U}_b)\right| \leq \overline{|F_a - U_a|} + \overline{|F_b - U_b|}.$$

The bars use the same question and role weights as the reported contrast. A coarse bound can replace each right-hand term by the corresponding tie frequency, but this discards the number of tied categories and whether gold is among them. The observed contrast shift is more informative than that worst-case bound when the full panels are available, as they are here.

This calculation explains why unequal tie rates deserve attention without turning them into an explanation of model competence. A condition can have many ties whose candidate answers are all wrong, producing no difference between selectors. Another can have a smaller number of ties that frequently contain gold, making its measured accuracy sensitive to response order. The MATH shift is therefore computed from complete candidate sets, not inferred solely from the 32% independent-panel tie rate.

5.4 SEPARATING ROLE ASSIGNMENT FROM SELECTION

Independent voting permutes the same four answers, so its uniform-tie scores agree for every question. Its fixed-order difference therefore has an exact selector explanation. Exchange and own-response revision generate different third and fourth answers in the two orders; their residual differences can reflect contexts, model roles, and sampling.

On MATH, the nine-point exchange advantage over own-response revision with Llama first becomes 4.5 points after balancing proposer orders and 1.125 points after uniform tie averaging. The two operations address different features of the same observations: proposer assignment and preference among equally frequent answers. They are not independent replications.

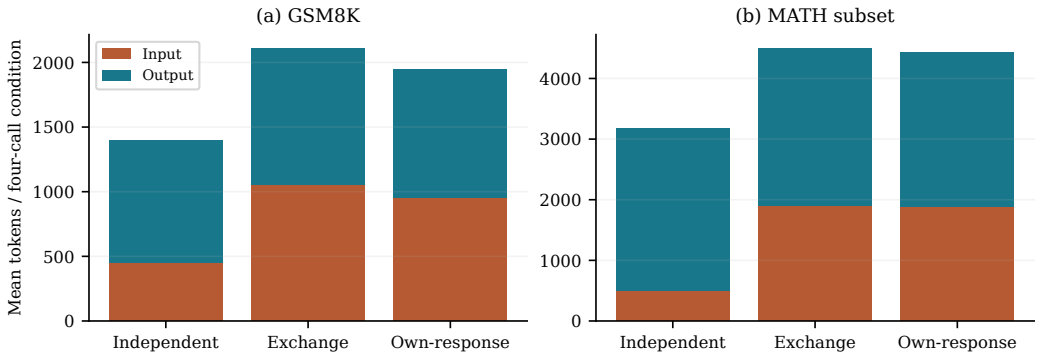

Figure 3: Input and output tokens per four-call condition, averaged over questions and both role orders. Stacking the counts displays the accounted text volume; it is not a FLOP, memory, or latency measurement.

Table 2: Role-averaged realized costs. Missing is the mean number of unparsed answers among the four calls. All calls have the same 1,024-token ceiling.

Task	Condition	Input tokens	Output tokens	Missing
GSM8K	Independent	447.2	948.4	0.385
	Exchange	1050.4	1059.3	0.353
	Own-response	950.5	995.7	0.420
MATH	Independent	494.9	2675.1	1.250
	Exchange	1899.7	2599.0	1.055
	Own-response	1890.8	2548.4	1.190

Missing answers still abstain. Uniform tie resolution uses only valid modal categories; an all-missing panel still fails. It neither imputes answers nor credits unparsed reasoning. Changing extraction would alter the candidate sets and requires a separate analysis of the saved text.

5.5 Attribution within the saved panels

For fixed generated answers, the change from fixed-order to uniform-tie scoring is attributable to selection by construction. It does not explain why models produced those answers or predict a new adaptive aggregator. Exchange-induced changes in dependence, missingness, and agreement remain part of the protocol effect.

This post hoc diagnostic retains the primary endpoint and needs no new model calls or fitted selector. For independent voting, it attributes the entire order effect to selection; for revised panels, it separates that component from the changed answers. The calculation is one sensitivity analysis of the original observations, not an additional replication.

6 Realized costs and missing answers

Exchange adds 603.2 input tokens per question over independent voting on GSM8K and 1,404.8 on MATH. Its input use is about 2.35 and 3.84 times the respective independent baseline. Generated-token differences are $+110.9$ on GSM8K and -76.0 on MATH. A shorter continuation can therefore coexist with substantially more input processing. Four matched calls and equal output ceilings do not establish equal realized compute.

Compared with own-response revision, exchange adds 99.9 input tokens on GSM8K and 8.9 on MATH. Generated-token differences are $+63.6$ and $+50.7$. The conditions have the same family call counts, but messages differ in length and the exchange introduces a sequential review dependency. These dependencies can affect serving latency, which is not measured in this experiment.

Table 3: Initial-answer agreement and final role-averaged correct totals. Every fixed-vote condition preserves a shared valid initial agreement. Half counts arise from averaging the two role orders, not from partial credit for an individual fixed-order answer. This is a post hoc descriptive partition.

Task	Initial state	Questions	Independent	Exchange	Own-response
GSM8K	Agree, correct	133	133	133	133
	Agree, wrong	1	0	0	0
	Differ or missing	66	51	45	44.5
MATH	Agree, correct	13	13	13	13
	Agree, wrong	5	0	0	0
	Differ or missing	82	30	28.5	24

Missing-answer counts are also protocol-dependent. Exchange has fewer unparsed answers than independent voting on MATH, yet does not achieve a higher role-averaged accuracy. Reducing missingness can add either correct or incorrect votes. Conditioning on panels with four parsed answers would change the target population and is not the primary analysis. The fixed outcome retains every question, including all-missing panels, under the declared extraction rule.

7 Initial agreement limits where exchange can matter

A structural restriction of the four-answer vote. Suppose the two shared initial answers are the same valid category v. Under the fixed earliest-answer tie rule, the four-answer vote must select v regardless of the two new answers. The initial category already has two votes. A different category can receive at most two votes, and if it does, the earliest response among the tied categories is still one of the initial answers. Otherwise v has a strict plurality. This argument also covers abstentions in either new answer. It does not cover two missing initial extractions, which supply no valid votes.

This restriction applies equally to independent voting, exchange, and own-response revision. Initial agreement therefore fixes their correctness: correct agreement cannot be overturned, and incorrect agreement cannot be rescued by the two additional calls. This is a property of the specified aggregator and call allocation, not evidence that a model never generates a useful correction. Even two correct new answers cannot overturn two identical wrong starts under this tie rule.

An exact decomposition of the contrast. Let D_i indicate that the shared starts differ or include a missing answer, and let δ_i be the role-averaged exchange-minus-control score. Under the fixed vote, $\delta_i = 0$ whenever $D_i = 0$. Thus, with $n_D = \sum_i D_i > 0$,

$$\widehat{\Delta} = \frac{n_D}{n}\left(\frac{1}{n_D}\sum_{i:D_i=1}\delta_i\right), \qquad |\widehat{\Delta}| \le \frac{n_D}{n}.$$

The fraction n_D/n measures where this decision rule allows an effect; the conditional mean measures its direction and size on those questions. Valid initial agreement occurs on 134/200 GSM8K and 18/100 MATH questions, leaving $n_D = 66$ and 82, respectively. We verify the zero-contrast restriction on every saved panel. On that subset, the GSM8K exchange-minus-independent difference is $-6/66$, or -9.09 points; multiplying by its $66/200$ frequency recovers the overall -3-point contrast. The analogous MATH difference is $-1.5/82$, or -1.83 points, yielding the overall -1.5 points after weighting. These conditional numbers are descriptive, not separately selected confirmatory tests, and we retain all questions in the primary endpoint.

Why a selector sensitivity is informative here. Under uniformly random tie resolution, two new matching answers can tie the shared starts and receive half the selection probability. Initial agreement then no longer fixes the final score in those cases. This explains one way that the uniform-tie sensitivity changes the decision procedure without changing any generated text. It does not make uniform tie resolution a recommended replacement: allowing a correction also allows a correct initial agreement to be overturned. The desirable rule depends on the quality of the competing answers, which frequency alone does not certify.

Dependence on the allocation of calls. The restriction extends beyond four answers. Suppose the earliest valid initial answer is v, its initial count is c_v, the largest competing initial count is $c_{\max}$ (zero if none exists), and r calls remain. If $c_v \geq c_{\max} + r$, no competitor can exceed v's count, even if every remaining answer supports that competitor. Any tie still favors v. Thus the number of retained initial votes can restrict correction before any review is generated. With two agreeing starts, three later matching answers can outvote them; two cannot under the fixed rule. This identifies an opportunity allowed by the selector, not a benefit guaranteed by more calls. Changing which answers enter the final vote defines a different protocol and must be evaluated at its corresponding cost.

8 Interpretation and limitations

The control identifies a protocol contrast. Sharing initial answers removes one source of sampling variation. Matching family composition and calls avoids crediting a stronger or larger panel to communication. Own-response revision supplies an additional comparison beyond independent sampling. Nevertheless, exchange changes context content, length, and dependency structure together. Its last response sees the other family's review; the own-response control sees its own initial answer. Their difference is not a perfectly isolated causal effect of cross-family information under otherwise identical contexts. The supported estimand is the specified protocol contrast, with those changed inputs made explicit.

Statistical precision and scope. An interval spanning zero does not establish equivalence. In particular, the 100-question MATH subset leaves appreciable positive and negative contrasts unresolved. These two open-weight families and one exchange pattern do not represent all multi-agent methods, tool use, longer discussions, or final-model synthesis. Selecting the stronger single-family panel after inspecting these results would also require separate evaluation. We do not infer an optimal deployment policy from this small set of operating points.

Costs and quality are separate outcomes. The exchange uses more input tokens without demonstrating an accuracy advantage over independent voting. This is evidence about the measured operating point, not a theorem that exchange cannot help at another budget. Calls, output ceilings, total tokens, and sequential latency constrain different resources. A quality– latency frontier requires a serving benchmark and multiple operating points; neither follows from a single call-matched comparison.

Task and metric scope. GSM8K questions are fresh relative to the development comparison; the MATH inputs are a reused, rational-answer subset. A restricted numerical checker can miss equivalent symbolic answers and does not verify proof correctness. The families differ in initial capability and may differ in formatting behavior. These factors limit generalization without invalidating the paired comparison under its declared endpoint. New models, tasks, and generation seeds would supply additional replication rather than being implied by question bootstrap intervals on these runs.

Boundaries of a mechanism claim. The experiment records each call's inputs and outputs, not independent labels of cognitive roles. A changed answer need not be a correction, and agreement need not imply successful information integration. The paired outcomes and selector sensitivities identify effects of the implemented decision procedure; claims about cognitive specialization require separately validated behavioral measurements and interventions.

9 Conclusion

With initial answers, panel composition, calls, and the vote held fixed, cross-family exchange does not establish an advantage over independent sampling on either arithmetic subset. The MATH advantage over own-response revision shrinks under order-invariant tie resolution, and valid initial agreement fixes the winner on many questions. More input processing therefore accompanies an unresolved or negative accuracy contrast at the tested budget. Evaluating exchange requires both generation controls and a decision-rule audit: compare against sampling and revision, evaluate both proposer orders, and identify which questions the final vote allows later responses to change.

Reproducibility Statement

The supplement provides the data and code needed to reproduce the reported analyses, following Pineau et al. (2021). The matched-panel section specifies the shared initial responses, call composition, generation settings, answer extraction, tie-breaking rule, and problem-level resampling. The supplement contains every generated path and a portable analysis that reconstructs condition scores, paired contrasts, token costs, and the post hoc uniform-tie sensitivity. Both role orders remain paired within question.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *International Conference on Learning Representations*, volume 2024, pp. 9079–9093, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25cc3adf8c85f7c70989cb8a97a691a7-Abstract-Conference.html.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/934252acd87f254d5d4672fbde283bd2-Abstract-Conference.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations*, volume 2024, pp. 32808–32824, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving Quantitative Reasoning Problems with Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857. Curran Associates, Inc., 2022. doi: 10.52202/068431-0278. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract.html.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://proceedings.neurips.cc/paper/2023/hash/a3621ee907def47c1b952ade25c67698-Abstract-Conference.html.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More Agents Is All You Need. *Transactions on Machine Learning Research*, 2024a. URL https://openreview.net/forum?id=bgzUSZ8aeg.

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, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. 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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996. URL https://doi.org/10.1007/BF02295996.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL https://jmlr.org/papers/v22/20-303.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-Agents Enhances Large Language Model Capabilities. In *International Conference on Learning Representations*, volume 2025, pp. 33944–33963, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5434be94e82c54327bb9dcaf7fca52b6-Abstract-Conference.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

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 conversations. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.