MULTI-AGENT DEBATE AMPLIFIES BUT RARELY CREATES: A CALIBRATED DISSECTION AT THE FRONTIER OF MODEL CAPABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent debate is widely deployed in the belief that interaction lets large language models (LLMs) reason beyond their individual capabilities, yet many benchmark–model combinations used to justify it are now saturated, obscuring how much of the reported gain comes from interaction itself. We rebuild the evidence base on a pre-registered difficulty window that retains only items on which agents genuinely disagree, and evaluate homogeneous multi-instance debate, a heterogeneous five-model panel, Asch-style confederate manipulations, and equal-budget replications of three role- and aggregation-based designs. Here debate converges quickly but improves slowly, and its consensus is often confidently wrong: agreement forms by the second round, yet up to 72.7% of a model's contested items end in a unanimously wrong consensus. What drives convergence depends on where an item sits in the panel's capability distribution: where someone can solve an item, correct answers propagate; where no one can, wrong answers are adopted 108 times against 4, and the strongest model becomes the largest source of adopted errors. Exchanging reasoning traces from the first round doubles adoption of the observed majority and halves minority survival, so the channel that repairs weaker members also accelerates consensus collapse on unsolvable items. Neither scripted saboteurs nor dedicated critics shifted these outcomes, and no replicated design improved significantly over plain heterogeneous debate. Debate therefore acts more as a selector over answers its members can already produce than as a creator of new ones. Progress may depend less on richer protocols than on panel coverage, staged exposure of reasoning, and knowing when consensus deserves trust.

1 INTRODUCTION

Multi-agent debate has become a default recipe for improving the outputs of large language models (LLMs): several model instances answer a question independently, read one another's responses, and revise over a few rounds, with the final-round majority taken as the group's verdict (Du et al., 2024; Liang et al., 2024). The recipe shipped inside major agent frameworks from 2023 onward (Wu et al., 2024; Hong et al., 2024), and later variants add role prompts (Chen et al., 2024), dynamic topologies (Lin et al., 2025; Lu et al., 2026), critique–revision routing (Feng et al., 2026), anti-conformity aggregation (Cui et al., 2026), and confidence-aware orchestration (Wang et al., 2026). It is widely believed that, similar to how ensembles of independent estimators cancel out errors, exchanging responses lets models correct one another and outperform every member alone.

Three lines of evidence suggest this premise deserves a closer look. First, *measurement*: many of the benchmark–model combinations used in this literature are saturated for current models, since on MATH-500 the four strongest models in our panel answer 97.4–99.8% of items correctly in a single attempt, leaving little room for the answer flips and minority opinions that interaction effects live on. Second, a growing literature finds that debate's gains shrink under budget-matched baselines and that beliefs can behave like a martingale under homogeneous exchange (Smit et al., 2024; Choi et al., 2025a). Third, LLM agents adopt a stated majority even when it is wrong, so convergence is not by itself evidence of learning (Choi et al., 2025b; Zhu et al., 2025; Sharma et al., 2024); Section 5 reviews both strands of literature and discusses the conformity mechanism in detail.

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.

These observations raise a question that has not been settled under controlled measurement:

> *Does multi-agent debate genuinely enable LLMs to reason beyond their individual sampling distributions, or does it mainly select, sometimes poorly, among answers that the models can already produce?*

Average accuracy cannot answer this question, because on saturated benchmarks interaction has almost nothing left to move. We therefore pre-register a difficulty window admitting only items whose single-shot accuracy over the five models falls between 30% and 80%, where interaction effects can still appear and be measured. We decompose that question into three sub-questions. **RQ1**: On the contested items this window admits, how large are debate's gains, and does the configuration change what the group solves and adopts? We contrast different configurations on their contested pools, and attach an all-wrong pool, where no model initially holds the correct answer, as a probe of whether interaction can create solutions (Section 3.1). **RQ2**: Can external voices, adversarial or constructive, causally change group outcomes, or does convergence simply follow the panel's own dynamics? Following Asch's conformity paradigm (Asch, 1956), we insert scripted saboteurs or a dedicated critic into otherwise unmodified panels (Section 3.2). **RQ3**: Do role- and structure-based designs deliver real improvements on top of a heterogeneous panel? We replicate three recent designs (positional, structural, and prompting-based (Zhang et al., 2026; Feng et al., 2026; Cui et al., 2026)) against plain heterogeneous debate on the same pool, panel, and budget (Section 3.3).

Our measurements answer the question in the negative: on contested items, the group mostly ends with an answer some member could already produce, and beyond the panel's capability, it converges confidently on wrong ones.

- **The group agrees long before it improves, and reasoning traces make it agree faster rather than better.** Inside the window, the heterogeneous panel with traces is still our strongest condition, because its members make different mistakes and repair one another. However, exchanging traces from the first round doubles majority adoption and halves minority survival. Furthermore, on items that no member can solve, the group adopts incorrect answers 108 to 4, never producing a correct one. What makes the panel work is disagreement rather than agreement: members who err differently convert wrong answers into correct ones, while copies of one model mostly convert them into other wrong answers.
- **A persistent wrong voice cannot break a healthy consensus; it only freezes the wrong one the group was already forming.** Scripted saboteurs leave every stable-correct verdict intact yet hold the majority on most of the items the panel cannot settle, and a dedicated dissenter adds a voice without adding capability while losing accuracy itself.
- **Roles, gates, and anti-conformity prompts change where consensus lands, not what the panel can solve.** None passes the pre-registered improvement criterion on top of the heterogeneous panel, the best nominal gain being fully explained by a change in context mode, and the same five items defeat plain debate and all three designs alike.

Our key contributions are threefold: **(1) Measurable Evaluation Framework:** We formulate a controlled methodology that applies a pre-registered difficulty window over audited single-shot accuracy, filtering out saturated items to isolate cases where interaction dynamics are truly observable. **(2) Directional and Round-Level Attribution:** We attribute every answer change to a specific direction and round, cleanly separating *conformity* from *persuasion* and *amplification* from *creation*; this is a distinction unattainable by aggregate-accuracy studies. **(3) Empirical Bounds on Interventions:** We bound the efficacy of structural designs by combining Asch-style agent injections within real panels with matched-budget black-box replications of three published architectures.

2 Preliminaries and Experimental Setup

2.1 Calibration

We evaluate five publicly accessible frontier models ranging from 30B to 2.4T parameters (denoted as M1–M5 in Table 1), accessed via their public API endpoints under default sampling configurations. We refer to this fixed set of models as the panel, and designate settings in which all five models debate concurrently as five-model conditions. Across both calibration and main experiments,

Table 1: Single-shot accuracy of the five models on the different calibration domains, in which the bold cells fall inside the pre-registered difficulty window.

Alias	Model	MATH-500	Omni-MATH	Chess	Code	HLE-math
M1	Qwen3.8-Max	99.8	97.2	100.0	100.0	**51.3**
M2	DeepSeek-V4-Flash	99.4	89.7	92.0	100.0	**32.7**
M3	Qwen3.6-35B-A3B	99.6	90.3	97.0	97.3	24.0
M4	Gemma-4-31B-IT	98.6	86.9	98.0	99.7	26.5
M5	GLM-4.7-Flash	97.4	**66.2**	**43.0**	96.7	14.4

all responses are scored by a unified pipeline comprising rule-based extraction, a mathematical-equivalence layer for notation normalization, and a manual adjudication channel whose reliability is verified through an audit (Appendix A.1).

As a preliminary evaluation, we assess each model's single-shot accuracy across five task domains, namely MATH-500 (Hendrycks et al., 2021), Omni-MATH (Gao et al., 2025), checkmate-in-one chess positions, executable code problems from LiveCodeBench (Jain et al., 2025), and the mathematics subset of Humanity's Last Exam (Center for AI Safety et al., 2026). The audited results, summarized in Table 1, highlight two key observations that motivate our experimental design. First, standard benchmarks exhibit severe saturation, with MATH-500 accuracy spanning 97.4–99.8% and all five models exceeding 96% on code, rendering opinion shifts too rare and minority positions too sparse to reliably measure interaction effects. Second, model capabilities vary significantly by domain rather than uniformly, as single-shot accuracy ranges from 66.2% to 97.2% on Omni-MATH and from 43.0% to 100% on chess, which precludes a global ranking. While such performance gaps provide the foundation for heterogeneous collaboration, they offer leverage only where models remain fallible. Therefore, before conducting any debate experiments, we pre-registered a rule classifying an item as *contested* if its panel-wide single-shot accuracy lies within 30% to 80%, balancing the presence of model errors with the potential for collective resolution.

2.2 Task pools

The audited calibration yields three pools. Each model debates its own contested items, namely the in-window items it answers incorrectly while at least one peer answers correctly, giving five per-model pools and the 133 valid homogeneous runs analyzed here. All five-model conditions share one merged pool of 45 items sampled from each model's contested items, so it reflects the panel's overall disagreement and keeps the three experiments comparable. A third pool of 40 items, mainly from HLE-math, on which no model answered correctly, even exhausting its reasoning budget at calibration, serves as a one-off probe of whether debate can create solutions that no individual holds.

2.3 Measurement and attribution

Our research questions concern mechanisms that endpoint accuracy cannot reveal. Raw flip counts conflate social influence with spontaneous drift, because re-asked models change answers without peers (Hao et al., 2026). An aggregate gain hides whether a group amplified an answer a member already held or created one nobody held. Accuracy averages also treat a unanimously wrong group the same as a unanimously right one. We therefore measure debate at three layers: its states, its transitions, and its endpoints.

At the level of states, let $a_{i,t}$ denote member i's adjudicated answer class at round t and m_t the plurality class. **Member accuracy** is the fraction of $a_{i,t}$ equal to the gold class, and **verdict accuracy** applies the same comparison to m_t. The **convergence index** $\mathrm{CI}(t) = \frac{2}{N(N-1)} \sum_{i<j} \mathbf{1}[a_{i,t} = a_{j,t}]$ measures agreement without reference to correctness, so it remains informative on items where accuracy deltas carry no signal. **Minority survival** $\mathrm{MSR}(t)$ is the fraction of $r{=}0$ minority classes still present at round t, and the **overconfident-consensus rate (OCS)** is the fraction of items whose final round ends in unanimous agreement on a wrong class.

At the level of transitions, the **flip matrix** classifies every answer change between consecutive rounds by its direction, by its effect on correctness, and by the class it adopts. Direction is what separates conformity, which concentrates changes toward m_{t-1}, from drift, which has no directional prefer-

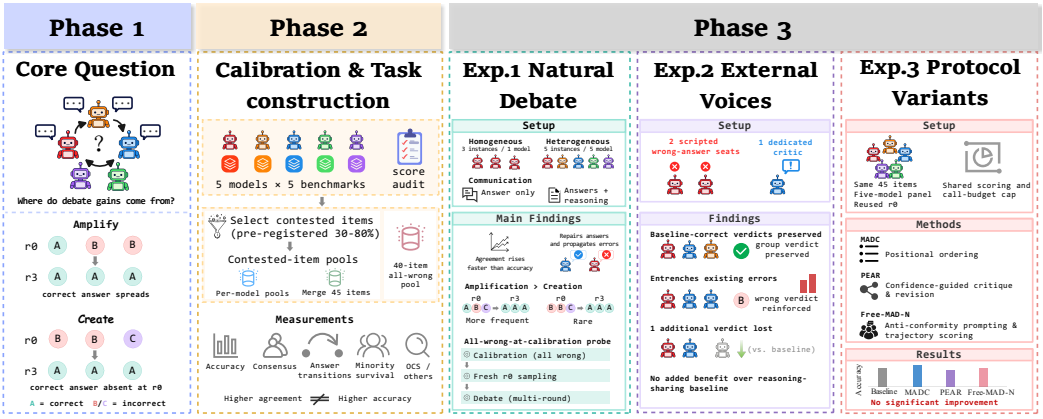

Figure 1: Overview of the study, in which single-shot calibration over five task domains is filtered through the pre-registered difficulty window into three task pools that feed the three experiments.

ence. Summarizing direction across rounds, **net conformity** is $\mathrm{NC} = \frac{\sum_t c_t}{\sum_t o_t} - \frac{\sum_t d_t}{\sum_t e_t}$, where c_t and d_t count, at round t, the minority holders who adopt m_{t-1} and the majority holders who leave it, while o_t and e_t count the members holding each position at round $t-1$.

At the level of endpoints, the **gain decomposition** compares the $r{=}3$ verdict with the $r{=}0$ verdict and with the member answers available at $r{=}0$. A run shows **amplification** when its verdict turns correct by $r{=}3$ and some member already held that answer at $r{=}0$, **creation** when no member held it, and **loss** when a correct verdict turns wrong. Round-wise states appear per model in Figure 3, pooled transitions in Figure 2, and endpoint quantities in Table A3, with the direction-by-effect matrices and the per-round generalization of overconfident consensus in Appendix A.5.

3 Experiments

All experiments share one protocol, summarized in Figure 1. A panel of N agents answers each item independently in round $r{=}0$. In each of $R{=}3$ consensus rounds, every agent receives the previous-round answers of all peers and is asked, with a single fixed prompt, to reconsider its answer. No roles, anti-conformity instructions, or early stopping are applied, and peer messages contain reasoning traces unless noted otherwise. The group verdict is the unique plurality equivalence class of the final round, and items without a unique plurality are counted as incorrect. Two configurations are used. The homogeneous configuration debates three instances of one model, and the heterogeneous configuration debates the panel of five distinct models with one instance each.

We design three experiments to systematically explore the mechanisms of multi-agent interaction. Section 3.1 addresses capability by contrasting the two configurations on contested pools and the all-wrong pool, where creation is the sole source of gain. Since observation alone cannot disentangle persuasion from compliance, Section 3.2 introduces scripted saboteurs and a dedicated critic into unmodified panels to test whether a persistent false voice alters the default consensus. To further validate our debate protocol, Section 3.3 replicates three designs that engineer role assignment and answer aggregation, under matched budgets and reused initial answers.

3.1 Amplification dominates; creation is rare

Setup. RQ1 evaluates whether collaboration enhances performance on contested items and identifies which configuration yields greater gains. To this end, Experiment 1 crosses the two configurations with two message formats, supplemented by the all-wrong pool as an exploratory probe, as shown in Table 2. Figures 2 and 3 present the four contested conditions, and Appendix A.5 provides the endpoint summary. Given that $r{=}0$ corresponds to independent sampling, any performance shifts in subsequent rounds are attributable to inter-agent interaction. Across all four conditions, performance gains concentrate heavily in the first exchange round, whereas later rounds largely drift.

Table 2: Round-wise and run-level quantities of the all-wrong probe, in which the five-model panel fails under the two message formats.

	Answer-only		With reasoning	
	r=0	r=3	r=0	r=3
Member accuracy (%)	3.5	5.5	4.0	9.0
Verdict accuracy (%)	5.0	5.0	2.5	7.5
Run-level quantities, pooled over the three consensus rounds				
Conform rate (%)	48.5		58.6	
Defect rate (%)	10.4		14.3	
Overconfident-consensus rate (%)	40.0		50.0	
Adoptions of a correct peer answer (n)	4		23	
Adoptions of a wrong peer answer (n)	108		140	

Homogeneous groups versus the heterogeneous panel. Since the two configurations evaluate different item pools, direct comparisons of absolute accuracy are invalid. A meaningful comparison lies in the intra-group dynamics induced by interaction. Both setups boost individual accuracy by 5–11 percentage points and achieve convergence, but they yield fundamentally different reliability guarantees. A homogeneous group reaches a unanimously incorrect consensus on roughly 25% of items, whereas the heterogeneous panel does so on only 2.2% of items. This disparity arises because identical model instances rapidly reinforce shared systematic errors, whereas a false consensus in the panel requires 5 distinct architectures to independently commit the same error. When homogeneous models fail to converge, minority answers simply persist.

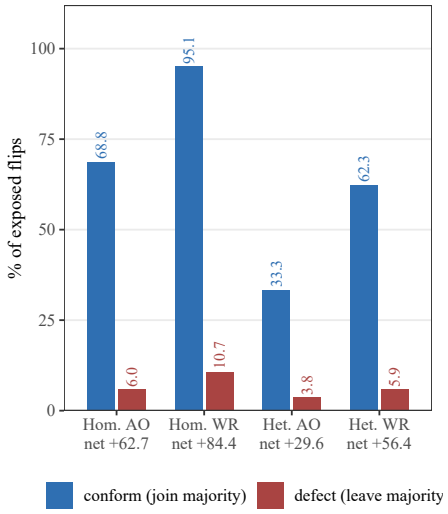

Figure 2: Exposure-normalized rates at which members join or leave the observed majority in the four contested conditions, with the net difference between the two rates printed under each condition.

Behavior across configurations. Model dynamics depend heavily on ensemble context. Within its homogeneous group, M1 propagates incorrect consensus, reaching 72.7% unanimously wrong items, yet it experiences a performance decline in the heterogeneous panel, where its errors are no longer certified. The weakest model displays the mirror image: it benefits most from the panel, while deriving negligible gain from homogeneous copies that merely resample its own error distribution. Crucially, stance shifts are governed by procedural protocols rather than intrinsic model traits. In each condition, agents seldom desert a majority opinion and routinely adopt it; thus, models differ in the validity of the concepts they switch to, not in their willingness to adapt. Correct minority opinions are adopted across all conditions, whereas genuinely novel correct solutions arise only in the homogeneous, reasoning-augmented runs below.

Impact of reasoning granularity. Exchanging full reasoning traces rather than answers alone roughly doubles majority adoption and halves minority survival, indicating that richer context accelerates persuasion and convergence. Crucially, the net benefit of this heightened conformity depends on group diversity. In the heterogeneous panel, reasoning traces convey corrective signals that double accuracy gains from 5.3 to 10.6 percentage points. In homogeneous groups, however, they merely propagate shared errors with negligible effect. This increased exposure also incurs a calibration cost, elevating overconfident consensus in the panel from 2.2% to 11.1%. Overall, richer context intensifies the pull toward the majority, but its value critically hinges on whether group members exhibit distinct failure modes.

The all-wrong probe, creation, and single-model inference. On items no model solves, agreement still forms while correctness does not: the verdict accuracy never leaves 5.0% under answer-only

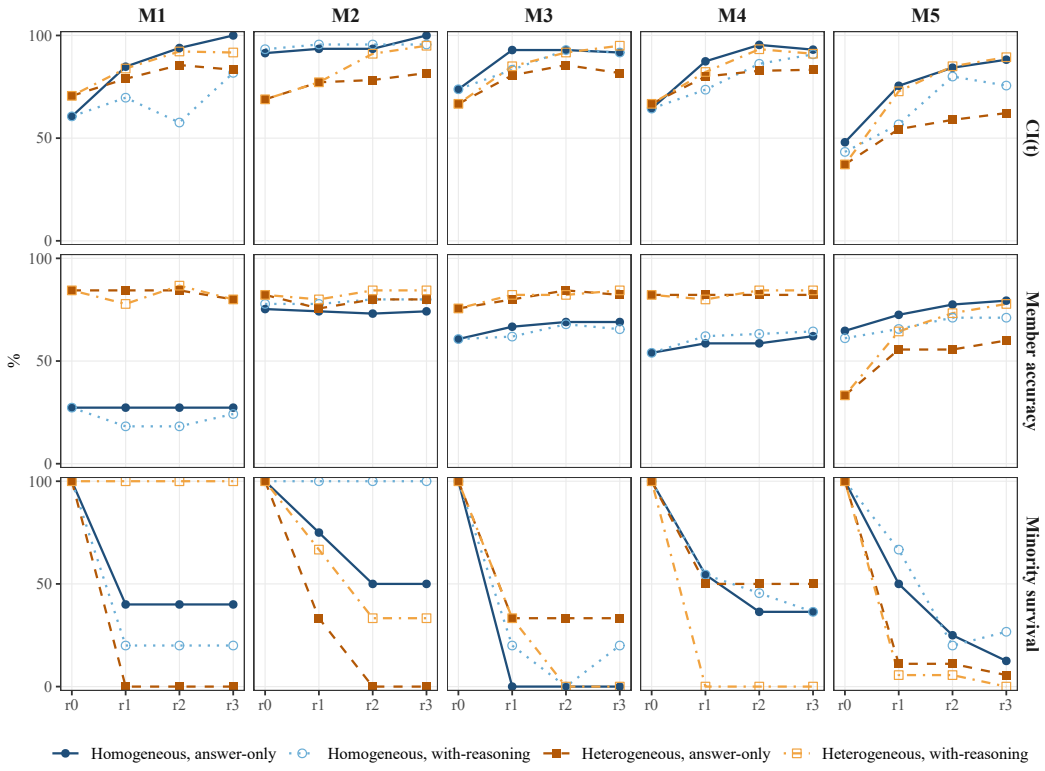

Figure 3: Round-wise trajectories of convergence, member accuracy, and minority survival in the four contested conditions.

exchange, and the correct answers that appear mid-debate are adopted only where a correct majority already existed at r=0. Adoptions are wrong 108 times against 4 without traces and 140 against 23 with them, so reasoning traces make propagation more efficient without making it more directional. Selection and amplification account for every gain in the contested pools. Creation occurs in only four homogeneous with-reasoning runs out of 128, and in each case one instance produced the new answer mid-debate and the group adopted it within a round. Three instances of one model are repeated sampling of a single distribution, so these four events are best read as the model's own flashes of insight that the richer context and the extra attempts happened to surface; debate's part was to notice and adopt the insight, not to produce it. The mechanism that debate does own lies in disagreement: in the panel, wrong answers are converted into correct ones far more often than into other wrong answers, whereas among copies of one model the two directions occur about equally often. Diversity of errors is therefore what lets a group escape its own blind spots, and it also hardens correct consensus, which in the panel survives even two seats repeating a verified wrong answer. Read against single-model inference, debate equalizes and selects rather than amplifies, lifting fallible reasoners toward the union of the panel's sampling distributions and leaving the strongest at or below their single-shot rates.

3.2 Saboteurs and a dedicated critic

Setup. RQ2 evaluates whether external interventions can causally shift group consensus or whether consensus follows the panel's own dynamics. Drawing on Asch's conformity paradigm (Asch, 1956), each condition introduces a distinct external voice to an otherwise unmodified five-model panel across 40 items, stratified by baseline trajectories: 20 stable-correct, 4 contested, and 16 non-convergent. All conditions exchange reasoning traces and are compared item-by-item against matched baseline trajectories from Experiment 1, ensuring that the external voice serves as the sole controlled variable. Specifically, under the saboteur condition, two scripted agents join from the initial consensus round and persistently repeat a pre-designated incorrect answer; under the critic

condition, a sixth seat instantiated by M1 is instructed to oppose the prevailing majority. Table 3 summarizes the results, while manipulation checks and member-level counter metrics are detailed in Appendix A.6.

Resilience of Correct Consensus. Where the baseline panel has already converged to a correct answer, neither intervention alters the outcome: every home item retains a correct majority across both conditions, and the scripted incorrect response never achieves a majority. Once a panel stabilizes on the correct answer, it remains resilient to external disruptions, regardless of whether the injected voice is adversarial or challenging. Furthermore, member alignment with saboteur outputs occurred almost exclusively on baseline-unstable items rather than stable ones.

Saboteurs reinforce pre-existing errors. At r=3, the scripted response holds a majority on 10 of the 16 non-convergent items. On every non-convergent item where the debate ends in a wrong majority, the matched baseline, which exchanges the same traces without the two scripted seats, already carries that same answer as its plurality, with a single exception. The primary impact of the two saboteur agents is injecting persistence: whereas member alignment with the incorrect response naturally decays across rounds in the baseline, the saboteurs sustain this alignment over time, leading to a slightly higher final adoption rate. The manipulation removes the drift that would otherwise have dissolved a wrong cluster, and the only correct verdict it destroys is one the baseline had barely rescued. Thus, the threat of a persistent adversarial voice lies not in fabricating a novel incorrect consensus, but rather in entrenching an emerging baseline error and occasionally disrupting a fragile correct one.

Table 3: Number of items whose final majority is correct, compared against the matched trajectories of Experiment 1.

Stratum (n)	Baseline	Saboteur	Critic
Stable (20)	20	20	20
Contested (4)	4	4	4
Non-conv. (16)	6	4	4

Compulsory dissent fails to enhance consensus and harms critic accuracy. The critic seat is played by the panel's strongest model and exchanges reasoning traces exactly as the other members do, so what the role contributes is a guaranteed opposing message rather than any new information. The panel's response follows from that asymmetry: it receptively incorporates the critic's feedback primarily during the initial consensus round, yet most non-convergent items still culminate in a unanimous incorrect consensus. The role does exact a measurable price on the seat that plays it, whose accuracy falls from 75% to 55% while the rest of the panel improves, since holding a fixed oppositional stance appears to pull a model away from answers it would otherwise have kept. Buying dissent with a dedicated seat therefore costs a call per round and returns a consensus that is no more often correct.

Causal Analysis. Across the 16 non-convergent items, the saboteur condition and its matched baseline share identical initial states and differ solely in the presence of two scripted agents. This controlled design isolates the causal impact of a persistent adversarial voice. Although the overall effect size is modest, its directional trajectory is consistent, showing that adoption of the scripted incorrect answer increases, and that the surrounding error cluster persists rather than dissipating. Conversely, the 24 baseline-correct items reveal a striking asymmetry: despite two scripted agents repeatedly injecting a verified incorrect answer across all rounds, no final verdict is altered. This serves as the causal complement to the observational findings in Section 3.1 and substantiates the same core insight: persistent dissenting voices systematically bias group consensus when belief certainty is fragile, yet exert negligible influence once a correct majority has stabilized.

3.3 Confidence roles and anti-conformity aggregation against plain debate

Setup. RQ3 asks whether engineering the two levers a debate protocol actually controls, who speaks in what order and how the group's answers are combined, improves on plain heterogeneous debate when the pool, the panel, and the budget are held fixed. We therefore replicate three designs that intervene at different points of the loop: MADC assigns speaking positions by confidence, so that roles follow evidence instead of turn order (Zhang et al., 2026), PEAR pairs a critic with a reviser and gates every revision by confidence together with a zero-call routing decision (Feng et al., 2026), and Free-MAD-N prompts members against conformity and aggregates whole trajectories rather than final answers (Cui et al., 2026). All three exchange reasoning traces at the plain panel's per-

Table 4: Black-box replications of three role- and aggregation-based designs against the plain panel on the merged pool under the same conditions, where net flips count items won minus items lost against the baseline and p is the paired exact-binomial probability.

Method	Axis	Calls per item	Final acc.	Net flips	p
Plain panel, ans-only	baseline	15	82.2	—	—
MADC	positional ordering	15	88.9	+3	0.25
PEAR	confidence, routing, critique	15	80.0	−1	1.00
Free-MAD-N	anti-conformity prompting	10	82.2	0	1.00

item call budget. The baseline is the answer-only panel verdict of Experiment 1 on the same items, and the improvement criterion was pre-registered: a paired exact-binomial $p < 0.05$ with at least three net item flips and mechanism indicators moving in the same direction. Table 4 summarizes the outcomes, and the replication caveats, deviations, and behavioral detail behind it are reported in Appendix A.6.

No design passes the criterion. The largest nominal gain, MADC's three items, is indistinguishable from chance at $p = 0.25$. Free-MAD-N ties the baseline exactly and PEAR ends below it. On a 45-item pool the whole spread between methods amounts to at most three item flips, which is below what a pool of this size can resolve.

MADC's nominal gain stems from a context-mode effect. The three items resolved by MADC are identical to those salvaged in Experiment 1 via reasoning-trace exchange upon transitioning from the answer-only format, two of which are likewise rescued by the dedicated critic in Experiment 2. Consequently, MADC's ordering mechanism yields no detectable benefit beyond the confounding context shift. Although MADC achieves the largest member-level accuracy gain among the three methods, this individual competence translates into correct panel verdicts only on items where reasoning traces already suffice.

Anti-conformity alters behavior, not outcomes; critique and revision degrade performance. Free-MAD-N's prompting functions as intended behaviorally, increasing answer switching and suppressing final consensus. However, its accuracy remains identical to the baseline item-by-item, demonstrating that avoiding lock-in cannot supply non-existent knowledge. PEAR deploys the heaviest mechanism, yet incurs the only net performance loss. On a pool where 93.3% of initial states $r{=}0$ contain a correct member answer, critique and revision offer no new knowledge, instead allowing flawed critiques to overwrite correct answers. Thus, PEAR maintains high post-revision confidence despite achieving the lowest accuracy across all four conditions. This echoes findings from Experiment 2, where a dedicated critic duplicated reasoning traces and suffered self-degradation.

Structural interventions fail to expand the panel's capability boundary. The same five items end up incorrect under the plain panel and under all three designs, the hard core the all-wrong probe exposed, and none of the 1,800 method calls solves them. Within this pool's resolution, engineering role assignment and answer aggregation redistributes where consensus lands inside the panel's sampling union rather than extending the union, which is the pattern every previous section reported.

4 Discussion: When Is Debate Worth It

Surprisingly, both failure modes manifest most clearly in our highest-performing setting: while exchanging reasoning traces elevates the heterogeneous panel's performance above all other configurations, the same traces double conformity and halve minority survival. Consequently, whenever an item exceeds the competence of all individual members, the panel overconfidently converges on an incorrect verdict. Incorrect verdicts originate from two distinct mechanisms with divergent remedies. *Selection failures* dominate: at least one member possesses the correct answer, yet the group fails to aggregate it into a consensus verdict—a failure exacerbated over time as consensus pressure suppresses the minority responses essential for solution coverage. *Capability failures* reflect intrinsic model limits: no member can solve the item, either because independent models converge on identical false attractors or because a model exhausts its compute budget on non-terminating

reasoning loops, rendering multi-instance replication ineffective. Together, these two mechanisms explain why multi-agent debate proves effective strictly within the pre-registered difficulty window.

These findings yield three major implications for deployment: **(1) Heterogeneity should be leveraged for coverage rather than consensus.** The panel's strength lies in diverse error modes: distinct mistakes translate into collective repairs, whereas replicated instances of a single model merely shuffle incorrect answers among themselves, with the benefit vanishing when the strongest member fails. **(2) Reasoning traces should be treated as a calibrated dosage rather than a default baseline.** While traces mediate error correction, they also accelerate conformity and collapse minority answers. Protocols that withhold traces until members commit to an independent answer, or reveal them strictly upon disagreement, can decouple the repair channel from the conformity channel, which is an avenue our findings motivate for future research. **(3) Consensus should be evaluated jointly with item difficulty.** Unanimous consensus below the difficulty threshold signals shared error attractors rather than collective knowledge, a failure unmitigated by any examined method. Because of limited endpoint access and benchmark coverage, the magnitudes we report may require further validation at larger scale. The judgment of this paper, however, still applies. The boundary most likely persists because consensus rounds expose answers without the evidence a member would need to verify them, and Appendix A.7 traces the item-level failure modes behind it.

5 Related Work

Multi-agent debate improves factuality and reasoning over single-model baselines (Du et al., 2024; Liang et al., 2024) and ships inside agent frameworks (Wu et al., 2024; Hong et al., 2024), while recent variants add roles, topologies, and critique–revision routing (Feng et al., 2026; Cui et al., 2026; Wang et al., 2026); its measured gains, however, shrink under budget-matched scrutiny: simple baselines are competitive (Smit et al., 2024), majority voting captures most of the gain and beliefs follow a martingale under homogeneous fully-connected exchange (Choi et al., 2025a), self-consistency is hard to beat at equal budget (Yang et al., 2025), heterogeneity is advocated as a remedy for degradation (Zhang et al., 2025), and prompt strength confounds comparisons (Wang et al., 2024). A parallel literature documents conformity and sycophancy as the mechanism behind such degradation (Sharma et al., 2024; Han et al., 2026; Pitre et al., 2025; Choi et al., 2025b), and trace taxonomies find that the dominant inter-agent failure is agreeing for the wrong reasons (Cemri et al., 2025), with Hao et al. (2026) closest to ours in decomposing single-round flips into spontaneous, conformist, and reasoning-induced sources. In contrast, our study pre-registers a difficulty window from calibration, tracks direction-aware flip dynamics across rounds in homogeneous and heterogeneous groups, manipulates social pressure causally with scripted seats, probes creation below the capability floor, and replicates three role- and aggregation-based designs at equal budget (Zhang et al., 2026; Feng et al., 2026; Cui et al., 2026); the closest single-model analogue is the pass@1-up, pass@k-down trade in RLVR (Yue et al., 2025), of which our finding that debate raises sampling efficiency inside the union of member distributions while shrinking the union is the multi-agent counterpart.

6 Conclusion

In this paper, we subject the generative reading of multi-agent debate to controlled measurement: a pre-registered difficulty window over audited single-shot accuracy, direction-aware attribution of every answer change, causal insertions of scripted voices, and equal-budget replications of three designs that engineer role assignment and answer aggregation across five models and four reasoning domains. We show that debate amplifies but rarely creates: gains come from selecting and repairing answers already present in the panel's sampling union, that union contracts as consensus absorbs minority answers, and no manipulation or protocol design we tested moved the boundary. The measurements also expose a dosage effect that practitioners can act on: reasoning traces exchanged from the first round double conformity and halve minority survival, accelerating consensus collapse on items nobody can solve, whereas panels composed for complementary failure modes supply the repairs and the coverage behind every gain we observed. We therefore recommend calibrating difficulty before debating and skipping saturated items, composing panels for different failure modes rather than adding roles, and withholding reasoning traces until members have committed to independent answers. The value of multi-agent debate lies less in richer interaction protocols than in who is in the room, what they are asked to share, and when consensus deserves trust.

AI USE STATEMENT

Generative AI tools were used in this work for literature searching and summarization, drafting and polishing manuscript text, and assisting with analysis code. We have not used generative AI tools for study design decisions, data adjudication, or the construction of the task pools beyond the documented scoring pipeline. No generative AI tool was used to run the experiments, which the authors executed on their own infrastructure, and the studied language models are the objects of the experiments rather than authoring tools. Additionally, we used generative AI tools for wording suggestions and consistency checks. We have reviewed all AI-assisted work: AI-assisted text was checked against the cited sources by all authors, AI-assisted code was verified and tested against the frozen experimental records, and every reported statistic was re-derived from the raw logs by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study involves no human subjects: all data come from public benchmarks and from API calls to publicly served models, and no private or personally identifiable data were used. The saboteur manipulation is a defensive analysis of a known adversarial risk in multi-agent systems rather than a new attack, and the scripted wrong answers were confined to the authors' own experiments and are not transferable to deployed systems without access to the panel. Model identities are disclosed in Table 1 so that the capability claims can be checked against the released models. The authors declare no conflicts of interest and no sponsorship from any model provider.

REPRODUCIBILITY STATEMENT

All conditions share one documented protocol at the start of Section 3, one scoring pipeline with a manual-adjudication channel in Section A.1, and a pre-registered difficulty window fixed before any debate run. Every reported statistic is recomputed from frozen run logs by the analysis scripts. Condition-level counters, manipulation checks, and replication deviations are listed in Appendix A.6, and per-model indicator trajectories in Appendix A.5. Prompts, run-level settings, seeds, and call budgets are provided in the supplementary material, together with the adjudicated answer keys and the item-level outcome tables. Known reproducibility limits, including single-seed point estimates and the deviations introduced when replicating the published designs, are stated in Section 5 and Appendix A.6.

REFERENCES

Solomon E. Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9):1–70, 1956. doi: 10.1037/h0093718. URL https://doi.org/10.1037/h0093718.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent llm systems fail? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4082. URL https://doi.org/10.52202/085713-4082.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649:1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://doi.org/10.1038/s41586-025-09962-4.

Pei Chen, Shuai Zhang, and Boran Han. CoMM: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1720–1738. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.112. URL https://aclanthology.org/2024.findings-naacl.112/.

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*, vol-

ume 38, 2025a. doi: 10.52202/085713-3405. URL https://openreview.net/forum?id=iUjGNJzrF1.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. An empirical study of group conformity in multi-agent systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5123–5139. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.findings-acl.265. URL https://aclanthology.org/2025.findings-acl.265/.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-MAD: Consensus-free multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31977–31997. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1600. URL https://aclanthology.org/2026.findings-acl.1600/.

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. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Yang Feng, Ziwei Xu, Xia Hu, and Fengxiang He. Pear: Permutation-equivariant adaptive routing multi-agent debate. *arXiv preprint arXiv:2606.20621*, 2026.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yaqPf0KAlN.

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

Xiqi Hao, Zengqing Wu, Yu-Xuan Qiu, Chuan Xiao, Ruiqi Xu, Shuyuan Zheng, and Jianbin Qin. Not all flips are conformity: Decomposing stance convergence in multi-agent llm debate. *arXiv preprint arXiv:2606.00820*, 2026.

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*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, 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. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VtmBAGCN7o.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Hsien-Tsung Lin, Pei-Cing Huang, Chan-Tung Ku, Chan Hsu, Pei-Xuan Shieh, and Yihuang Kang. Towards simulating social influence dynamics with llm-based multi-agents. In *2025 IEEE International Conference on Intelligence and Informatics (IRI)*, pp. 307–312. IEEE, 2025. doi:

10.1109/IRI66576.2025.00064. URL https://doi.org/10.1109/IRI66576.2025.00064.

Yuxing Lu, Yucheng Hu, Xukai Zhao, and Jiuxin Cao. Dytopo: Dynamic topology routing for multi-agent reasoning via semantic matching. *arXiv preprint arXiv:2602.06039*, 2026.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22112–22133. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1141. URL https://aclanthology.org/2025.findings-acl.1141/.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 45883–45905. PMLR, 2024. URL https://proceedings.mlr.press/v235/smit24a.html.

Jingbo Wang, Sendong Zhao, Jiatong Liu, Haochun Wang, Wanting Li, Bing Qin, and Ting Liu. Orchestrating intelligence: Confidence-aware routing for efficient multi-agent collaboration across multi-scale models. *arXiv preprint arXiv:2601.04861*, 2026.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6106–6131. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.331. URL https://aclanthology.org/2024.acl-long.331/.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) 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. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1933. URL https://openreview.net/forum?id=4OsgYD7em5.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. Stop overvaluing multi-agent debate – we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*, 2025.

Qian Zhang, Jinyi Liu, Yan Zheng, Hebin Liang, and Lanjun Wang. Key decision-makers in multi-agent debates: Who holds the power? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29883–29891, 2026. doi: 10.1609/aaai.v40i35.40235. URL https://ojs.aaai.org/index.php/AAAI/article/view/40235.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3854–3872. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.195. URL https://aclanthology.org/2025.acl-long.195/.

A APPENDIX

Notations. Task items are named by source dataset: `omni-N` and `math500-N` denote item N of the Omni-MATH and MATH-500 test splits, `hle-*` a Humanity's Last Exam (mathematics) item, `chess-N` a checkmate-in-one position, and `lcb-[a, b, c]` a LiveCodeBench execution item, whose identifier carries the bracketed index triple of the source record.

Contents. The subsections below provide details on the following components:

- the calibration audit, saturation evidence, and per-model pool sizes (A.1);
- the construction and governance of the three task pools (A.2);
- the design and replication specifications of Experiments 2 and 3 (A.3);
- the run-level settings, seed policy, and call budgets (A.4);
- the endpoint results and per-model analysis of Experiment 1 (A.5);
- the counters and manipulation checks of all three experiments (A.6);
- the item-level failure modes underlying the discussion (A.7);
- four worked trajectories illustrating representative cases (A.8).

A.1 CALIBRATION, SATURATION, AND SCORING AUDIT

Saturation. Table 1 reports the audited single-shot accuracy of the five models over the five domains. The established benchmarks are saturated for these models: MATH-500 scores span 97.4% to 99.8%, and all five models exceed 96% on code, leaving answer flips too rare and minority opinions too empty for interaction effects to be measured on those domains. The domains within the pre-registered difficulty window are predominantly Omni-MATH (66.2%–97.2%) and HLE, with chess (43.0%) being the exception for the weakest model. Table A1 lists the per-model contested pools that this calibration yields and that feed Experiment 1.

Table A1: Per-model contested pool sizes, in items, after the three-wave scoring audit. The capped pool is what Experiment 1 runs; the valid set excludes escape and invalid-retained items.

	M1	M2	M3	M4	M5
Contested pool, capped at 50	11	38	28	29	50
Valid items analyzed	11	31	28	29	34

Scoring audit. All answers are scored by one pipeline: rule-based extraction, a mathematical-equivalence layer, and a manual-adjudication channel. Three audit waves corrected 201, 159, and 44 calibration records respectively. The first wave, manual adjudication over the rule-based judgments, corrected 201 records, revised five gold answers (math500-176, omni-0, omni-8, omni-112, and omni-148), and removed five defective items (omni-2, omni-47, omni-90, omni-96, and omni-143). The second wave repaired the choice-answer extractor, whose word-boundary, markdown-decoration, and \boxed handling defects misread models' answer styles; its 159 corrections were all wrong-to-correct, and the repair exposed the single largest error, in which 40 of the 50 code errors initially attributed to one model were extractor misreads, moving that model's code profile from the panel's weakest to mid-tier. The third wave upgraded the equivalence layer and added a semantic-adjudication channel for free-text answers, contributing 44 corrections. Because every correction reclassifies an item from wrong to correct, each wave shrank the contested pools rather than enlarging them.

A.2 TASK-POOL CONSTRUCTION AND DATA GOVERNANCE

The difficulty window is fixed before any debate run: an item is *contested* when its panel-wide single-shot accuracy falls between 30% and 80%. Three pools follow from this rule. Each model's *contested pool* contains the items it answers incorrectly while at least one peer answers correctly, giving the per-model pools of Table A1 and the 133 valid homogeneous runs. The *merged pool* samples 45 items from the focal models' contested pools (9, 8, 8, and 15 from M2, M3, M4, and

M5, plus 5 shared-frontier items added for paired comparison with Experiment 1), and is shared by every five-model condition so the experiments remain comparable. The *all-wrong pool* holds 40 items, 38 from the HLE-math subset and omni-108 and omni-45, on which no model answered correctly at calibration; it serves as a one-off probe of whether debate can create a solution nobody holds.

Task lists were frozen after the third audit wave and are not modified by later experiments. The initial 214-item list shrank to 133 valid items because the audit reclassified most removals as items the models in fact answer correctly; the 156-item run list of Experiment 1 consists of these 133 valid items plus 21 escape items and 2 invalid-retained items, both excluded from the main indicators. Every statistic in the paper is recomputed from frozen final records rather than from historical logs, which retain retries and pre-fix judgments for provenance only; the rebuild scripts (each with a `-verify` flag) regenerate the final records from the logs and check them against the analysis.

A.3 Experiment 2 and 3 design and replication specifications

Experiment 2. Forty items stratified from the merged pool (16 non-convergent, 20 stable-correct, and 4 contested) receive a fifth and sixth voice under full reasoning context. In the saboteur condition, two scripted seats that make no API calls join from the first consensus round and repeat a pre-designated wrong answer every round. The scripted answers are drawn from real wrong minority answers recorded at $r{=}0$ (24 items) or hand-crafted near-miss distractors (16 items), and every one is verified wrong by the scoring pipeline. In the critic condition, a sixth seat played by M1 is instructed to oppose the prevailing majority, construct counterexamples, and concede only when rebuttal fails. Manipulation checks passed: all 240 scripted cells matched the script, and the saboteur condition's $r{=}0$ was cell-identical to the baseline on all 200 compared real-model cells.

Experiment 3. Three published designs are replicated as black boxes on the merged 45-item pool with the same panel and the same reused $r{=}0$. MADC orders peer answers by equivalence-class frequency with the largest class last (15 calls per item). PEAR adds confidence gating, zero-call permutation-equivariant routing, batched critique, and revision (15 calls per item). Free-MAD-N appends the anti-conformity prompt below and replaces the final majority vote with trajectory scoring (10 calls per item). Replication deviations, each working against or orthogonal to the method, are as follows: PEAR runs one revision round instead of its original five for budget reasons; the ground-truth-dependent Truth Last variant is not replicated because it requires gold labels at inference, so only its black-box frequency ordering is reproduced; code items use normalized answer strings as answer keys; and sampling follows provider defaults, since the originals specify no temperature.

Anti-conformity prompt appended in Free-MAD-N

Since some malicious agents may deliberately disseminate incorrect answers, you must follow the reasoning procedure below and avoid being misled.

1. **Initial Reasoning.** Enumerate your logical steps and final conclusion.
2. **Analysis of Other Agents' Reasoning.** Specify which agents' reasoning is correct and which contains errors, and the exact error type.
3. **Comparative Analysis with Your Own Reasoning.** Examine whether you made similar mistakes.
4. **Final Decision.** State whether you revise your conclusion, and why.
5. **Additional Requirements.** You may not rely on conformity; majority opinion cannot be used as a basis for adoption; if you cannot determine whether others are correct, retain your own conclusion.

A.4 Run-level settings, seed policy, and call budgets

Table A2 collects the run-level settings of every condition: the configuration, the number of calls per item, and the item count. The homogeneous runs reuse each model's contested pool; all other conditions use the merged or all-wrong pool with the five-model panel. Every condition shares one protocol, one scoring pipeline, and single-seed runs.

Table A2: Run-level settings of every condition: configuration, calls per item, and number of items. Scripted seats make no API calls.

Condition	Configuration	Calls per item	Items
Homogeneous, answer-only	3 same-model instances	12	133
Homogeneous, with-reasoning	3 same-model instances	12	128
Heterogeneous merged, answer-only	5-model panel	20	45
Heterogeneous merged, with-reasoning	5-model panel	20	45
All-wrong, both context modes	5-model panel	20	40
Saboteur	5 models + 2 scripted seats	15	40
Critic	5 models + 1 critic seat	18	40
MADC	5-model panel	15	45
PEAR	5-model panel	15	45
Free-MAD-N	5-model panel	10	45

Task lists are sampled once with a fixed seed; the debate runs use a single seed throughout, so all results are point estimates. The initial answers of the merged and all-wrong pools are reused across experiments through a reuse chain that prefers a prior experiment's full answer body, then a calibration answer, and samples only where neither exists. Two execution-level changes are recorded as protocol caveats: the reasoning effort of M1 and M2 was reduced from its highest to a medium setting partway through the heterogeneous runs, with the affected cells timestamped for stratification; and two items with free-text gold answers are scored through the semantic-adjudication channel rather than the equivalence layer. The complete study used roughly 12,300 API calls, under 5% of the total budget.

A.5 Experiment 1 detail

Table A3 collects the endpoint quantities that Section 3.1 reports in prose, and Figure A1 gives the direction-by-effect flip matrices behind the pooled rates of Figure 2. Toward-majority moves are mostly wrong-to-correct repairs in the panel conditions and mostly wrong-to-wrong in the homogeneous ones, and correct-to-correct changes are absent by construction, because adjudicated correct answers form a single equivalence class. The all-wrong probe records CI rising from 42.5% to 67.3% under answer-only exchange and from 44.3% to 77.8% with reasoning. Correct answers appear at some round on twelve and ten items respectively and persist to $r{=}3$ on five and six. Minority survival at $r{=}3$ is 14.3% and 52.9%, and per-model member accuracy at $r{=}3$ is 5.0/5.0/7.5/7.5/2.5 without traces and 10.0/10.0/7.5/7.5/10.0 with them. Cells without exposure of a given kind are left empty in Figure A2.

Figure A2 completes the per-model view of the Section 2.3 toolkit with two indicators that the main text reports only in pooled form. The top row generalizes OCS to every round: the fraction of items on which all members agree on a wrong answer at round t; at $t{=}3$ this equals the OCS column of Table A3. Overconfident consensus forms within a single consensus round wherever failure modes are shared: M1's homogeneous OCS rises from 27.3% of its items at $r{=}0$ to 72.7% at $r{=}3$, while panel unanimity on the merged pool stays at or near zero in every round. The bottom row reports net conformity per transition, the conform rate minus the defect rate of Section 2.3 computed for each model separately. The gap is positive for nearly every model and condition, confirming that the pooled asymmetry in Figure 2 is not driven by a single model; it widens rather than narrows when messages carry reasoning traces, since conformity accelerates faster than defection in both configurations.

A.6 Counters and manipulation checks

Experiment 1. The homogeneous runs reuse each model's own contested pool, so per-model accuracy levels are not cross-comparable and only within-model changes are. The 133 valid runs feed the main indicators, and the 21 escape items and 2 invalid-retained items are excluded from them. The all-wrong probe reuses calibration answers for M1 and M2 and freshly samples M3 through M5, so its $r{=}0$ is anchored to the calibration records rather than resampled. Two items (omni-60 and omni-100) carry free-text gold answers that the equivalence layer cannot match and are scored through

Table A3: Endpoint summary of the four contested conditions of Experiment 1, in which the homogeneous groups are three instances of one model over its own contested pool and the heterogeneous conditions are the five-model panel over the merged pool. Values with arrows are $r{=}0{\rightarrow}r{=}3$; minority survival and overconfident consensus are reported at $r{=}3$.

	Homogeneous		Heterogeneous	
	answer-only	with reasoning	answer-only	with reasoning
Items (n)	133	128	45	45
States (%)				
Convergence index	68.2→93.7	68.0→87.8	62.0→78.4	62.0→92.4
Member accuracy	60.9→67.9	60.4→66.4	71.6→76.9	71.6→82.2
Verdict accuracy	62.4→68.4	61.7→68.8	80.0→82.2	80.0→84.4
Oracle (any member correct)	75.9→70.7	75.8→72.7	93.3→86.7	93.3→86.7
Minority survival	24.4	33.3	8.0	4.0
Overconfident consensus	26.3	24.2	2.2	11.1
Transitions and endpoints				
Conform rate (%)	68.8	95.1	33.3	62.3
Defect rate (%)	6.0	10.7	3.8	5.9
Adoptions of a correct peer answer (n)	62	101	31	48
Adoptions of a wrong peer answer (n)	72	126	17	39
Amplified (n)	11	9	1	3
Created (n)	0	4	0	0
Lost (n)	3	4	0	1

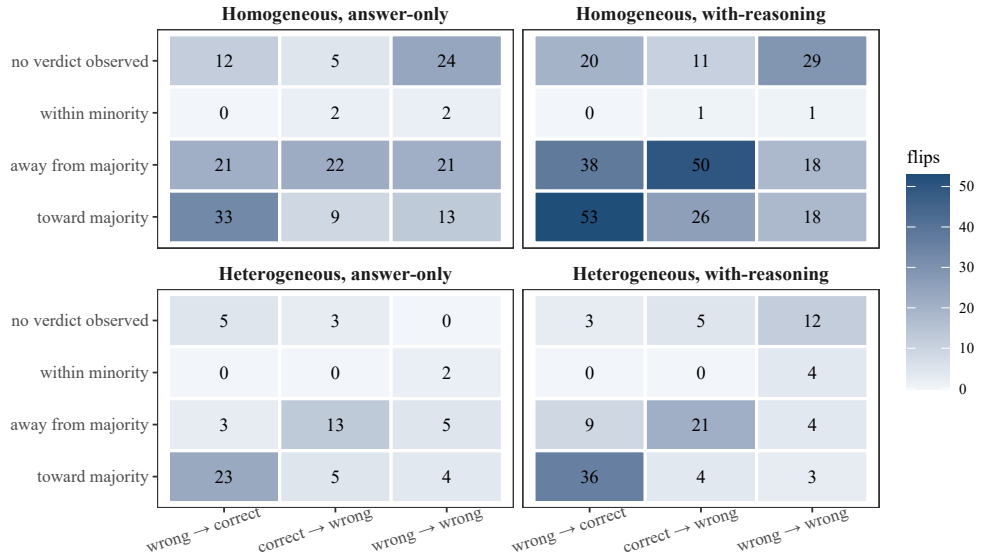

Figure A1: Counts of answer changes in the four contested conditions of Experiment 1, cross-tabulated by direction relative to the previous round's verdict and by effect on correctness.

the semantic-adjudication channel with feature-guarded accept patterns; this adjudication changes the all-wrong gain decomposition from zero to one generated item and shrinks its always-wrong set from five to three. The 2×2 context-mode comparison anchors each full-context run to its answer-only pair by direct $r{=}0$ injection, verified cell-for-cell on 384 and 225 cells for the homogeneous and heterogeneous pools.

Experiment 2. Baselines in Table 3 are the matched Experiment 1 trajectories: answer-only for the stable and contested strata, with-reasoning for the non-convergent stratum. Manipulation checks

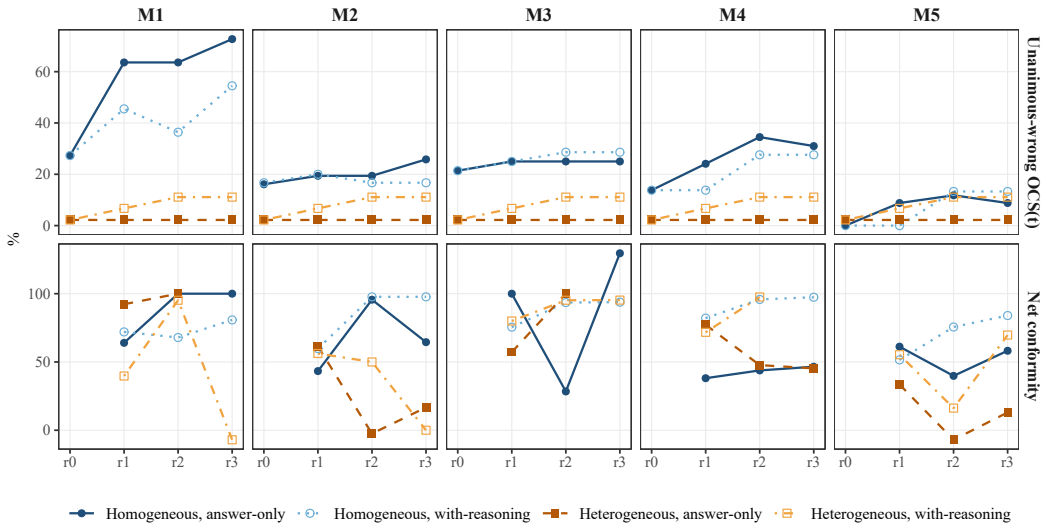

Figure A2: Per-model overconfident-consensus rate by round and per-model net conformity in the contested conditions of Experiment 1, which generalize to each model the pooled quantities reported in the main text.

passed throughout: every scripted answer was adjudicated wrong, all 240 scripted cells matched the script, and the saboteur condition's $r{=}0$ was cell-identical to the baseline on all 200 compared real-model cells. Member-level counters behind Section 3.2: only eight of the 160 member answers matching a scripted answer occurred on home items; on the non-convergent items, such answers stayed at 50 to 52 cells per round under the saboteurs against a decay from 50 to 36 in the matched baseline, with 40 adoptions against 30; the single dragged-down item is omni-50 (Example 2); the critic's answer was adopted in 36 transitions, 26 of them in the first consensus round, and the critic seat fell from 75% accuracy to 55% while the panel mean rose from 64% to 68%.

Experiment 3. Caveats behind Table 4: the methods exchange reasoning traces while the baseline exchanges answers only, which fully explains MADC's three wins and has unknown direction for the other two; each condition uses a single seed; and on 45 items fewer than ten discordant pairs cannot reach significance, so the verdict is a null within this pool's resolution rather than evidence that the designs cannot help in other regimes. Behavioral detail: MADC's wins are omni-147, omni-50 and omni-97 and it lifts mean cell accuracy by 12.8 points; Free-MAD-N records 112 answer switches against 322 holds with 64.4% final consensus; PEAR's post-revision confidence averages 4.56 out of 5 against 80% accuracy; the items no condition solves are omni-18, omni-20, omni-40, omni-71 and omni-132.

A.7 Failure modes

The three experiments establish what debate does. This subsection evaluates why groups fail when they fail. Tracing every failed item separates two failure modes with different remedies: items where no member ever holds the correct answer, and items where the group holds it and does not assemble it into a verdict.

Most verdict failures are selection failures. On the merged pool under answer-only debate, three of the 45 items never see a correct member answer in any round, but five of the eight wrong final verdicts fall on items where some member held the correct answer at some point. Interaction also erodes the raw material: the share of items on which any member holds a correct answer falls from 93.3% at $r{=}0$ to 86.7% at $r{=}3$, because consensus pressure removes exactly the minority answers on which coverage depends. Debate thus raises sampling efficiency inside the union of member distributions while shrinking the union itself, the item-level face of the amplification-without-creation pattern of Section 3.1.

Some failures are shared traps. A second class resists from every direction: on one shared item four of the five homogeneous pools independently converge to the same wrong value, and on such items every model pays a full reasoning cost without approaching the gold answer. These are the hard core of Section 3.3 and the unanimously wrong consensuses of Section 3.1; when failure modes are shared, replication adds no information and consensus carries none.

Capability failures are model-specific. Calibration exposes a third pattern: on particular items particular models never answer at all, burning their entire token budget on thinking that does not terminate and returning empty content. A two-feature rule, exhausted budget and empty answer, identifies all 149 such cells among 2,833 with zero false positives. Escape thinking is not shallow: it averages 4.7 times the length of normal traces, but explicit closure markers appear at the end of only 6.7% of escape traces against 62.1% of normal ones. In homogeneous debate, escape is rescued only when it is isolated; an instance that escapes while its peers converge reads their answers in the next round and recovers, its thinking length collapsing to about a fifth, whereas when all instances share the divergence every member sees only silence and twelve calls produce nothing. A homogeneous group therefore equals twelve independent failures exactly where its members are alike. On the all-wrong pool, 38 items drawn from HLE-math and 2 from Omni-MATH, the same exhaustion is systematic rather than incidental: at $r{=}0$ the weakest member produces no answer on 37.5–40% of the forty items depending on message format, and once debate begins the no-answer share rises to 22.5% for that member and to 20% for a second one, so the group is often choosing among missing answers rather than among wrong ones.

Cost and capability come apart. The sharpest asymmetry: a 284B model burned its full 65,536-token budget in each of twelve calls on one item without producing an answer, while a 31B model solved the same item with a converging trace at about a fifth of that per-call cost. A contamination check rules out memorization for the efficient solver, which answered the item only partially at calibration and derived the full classification under a longer budget, a partial-to-complete pattern typical of derivation rather than recall. Capability is also domain-relative: the weakest member overall sits near the ceiling on code at 96.7% and at the bottom on chess at 43.0%, while a math and code leader is the second-weakest chess model. Consistently, escape events barely overlap across models, with a single item out of forty shared ones drawing escapes from two models, so the panel's coverage advantage is structural rather than a margin of safety.

The non-termination is visible in the trace itself. On omni-60, the number-theory item reproduced in full as Example 3 below, the 284B model opens a new branch to ask "now can we lift to mod 9?" and is cut off mid-sentence at the limit after twelve attempts, whereas the 31B model closes with a counterexample and answers.

Divergent trace on omni-60 (truncated)

```
...  there is a polynomial mod 3 realizing the identity function
with derivative delta at 0, so the derivative delta is possible.
Now can we lift to mod 9?  We have a polynomial P0 over F3 with P0
the identity and derivative delta; the desired values P(x) mod 9 are
P(0)=0, P(1)=1, P [truncated at the 65,536-token limit]
```

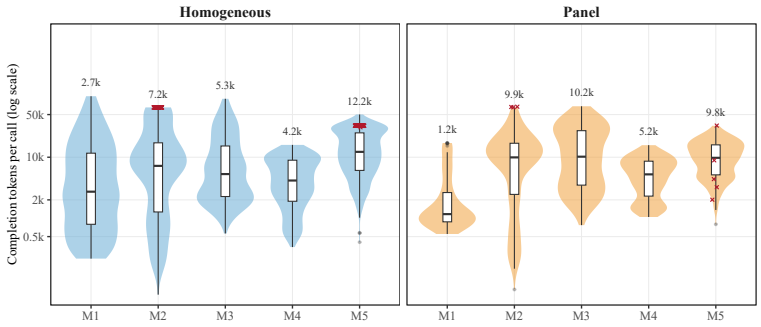

Figure A3: Distribution of per-call output tokens by model and configuration (log scale); red crosses mark non-terminating calls that exhausted their token budget and returned an empty answer.

Groups therefore fail in two ways that call for different remedies: selection failures call for better assembly of answers members already hold, and capability failures call for members with different failure modes. No communication protocol supplies either; Section 4 draws out the deployment consequences.

A.8 WORKED EXAMPLES

This subsection exhibits four representative trajectories, one per mechanism the paper measures. Each box states the item's prompt and gold answer, then tabulates the answer of every model slot in each debate round. A ✓ marks an answer the scoring pipeline adjudicates correct, a × an incorrect one, and an em dash a non-terminating response (an empty output at the model's output-token limit). Slots map to models as M1=Qwen3.8-Max, M2=DeepSeek-V4-Flash, M3=Qwen3.6-35B-A3B, M4=Gemma-4-31B-IT, and M5=GLM-4.7-Flash (Table 1).

Example 1— Repair (amplification): item math500-42

Question. A gecko is in a room that is 12 feet long, 10 feet wide and 8 feet tall. The gecko is currently on a side wall ($10'$ by $8'$), one foot from the ceiling and one foot from the back wall ($12'$ by $8'$). The gecko spots a fly on the opposite side wall, one foot from the floor and one foot from the front wall. What is the length of the shortest path the gecko can take to reach the fly, assuming it does not jump and can only walk across the ceiling and the walls? Express your answer in simplest radical form.
Gold. $2\sqrt{113}$. Four models answer correctly from the start; M3 opens with the wrong $2\sqrt{106}$ and is repaired to $2\sqrt{113}$ after one exchange.

	r=0	r=1	r=2	r=3
M1	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓
M2	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓	—
M3	$2\sqrt{106}$ ×	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓
M4	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓	$2\sqrt{113}$ ✓
M5	$2\sqrt{113}$ ✓	—	—	—

Example 2— Conformity (overconfident consensus): item omni-50

Question. There is a grid of height 2 stretching infinitely in one direction. Between any two edge-adjacent cells there is a door that is locked with probability $\frac{1}{2}$, independently of all other doors. Philip starts in a corner cell. Compute the expected number of cells Philip can reach, assuming he can only move between cells when the door between them is unlocked.
Gold. $\frac{32}{7}$. M1 and M2 open with the correct $\frac{32}{7}$ but abandon it for the incorrect $\frac{22}{5}$, and the panel ends unanimous on $\frac{22}{5}$.

	r=0	r=1	r=2	r=3
M1	$\frac{32}{7}$ ✓	$\frac{22}{5}$ ×	$\frac{22}{5}$ ×	$\frac{22}{5}$ ×
M2	$\frac{32}{7}$ ✓	$\frac{22}{5}$ ×	—	$\frac{22}{5}$ ×
M3	$\frac{22}{5}$ ×	$\frac{22}{5}$ ×	$\frac{22}{5}$ ×	$\frac{13}{2}$ ×
M4	$\frac{22}{5}$ ×	$\frac{22}{5}$ ×	$\frac{22}{5}$ ×	$\frac{22}{5}$ ×
M5	$\frac{22}{5}$ ×	—	$\frac{22}{5}$ ×	$\frac{22}{5}$ ×

Example 3— Escape (non-terminating reasoning): item omni-60

Question. Find all positive integers n for which there exists a polynomial $P(x) \in \mathbb{Z}[x]$ such that for every positive integer $m \geq 1$, the numbers $P^m(1), \ldots, P^m(n)$ leave exactly $\lceil n/2^m \rceil$ distinct remainders when divided by n. (Here P^m means P applied m times.)
Gold. n is prime or $n = 2^k$ (free text). Correct answers take phrasings such as "n is a power of 2 or an odd prime" or "$n = p$ or $n = 2^k$"; the semantic-adjudication channel accepts any equivalent form. M5 returns an empty answer in every round and M2 in two rounds, each exhausting its output budget (32,768 and 65,536 tokens per call), while M1 and M3 repair toward the classification M4 holds throughout. The r=0 answers are reused from prior runs and retain their inherited verdicts.

	r=0	r=1	r=2	r=3
M1	×	✓	✓	✓
M2	×	—	—	✓
M3	×	✓	✓	✓
M4	✓	✓	✓	✓
M5	—	—	—	—

Example 4— No creation (all-wrong pool): item hle-66f3f373

Question. Heads-up in a poker game, it's your turn; you may check or bet. Canonical reasons for betting are (1) getting calls from weaker hands and (2) folding out better hands. Subtler reasons may include (3) denying equity to drawing hands, (4) gaining information about your opponent's hand, and (5) avoiding revealing your own hand in a showdown. If we add the assumption that both you and your opponent play a game-theoretically optimal strategy (and give away no tells), which of these subtler reasons always disappear? Choices: A. 3; B. 4; C. 5; D. 4 and 5; E. all three disappear.
Gold. C. No model answers correctly in any round, and the panel ends unanimous on the wrong choice D. This is the all-wrong pool's modal outcome: debate cannot create an answer that no member holds.

	r=0	r=1	r=2	r=3
M1	D ×	D ×	D ×	D ×
M2	D ×	D ×	D ×	D ×
M3	D ×	D ×	D ×	D ×
M4	D ×	D ×	D ×	D ×
M5	D ×	D ×	D ×	D ×