NOT ALL FLIPS ARE CONFORMITY: DECOMPOSING STANCE CONVERGENCE IN MULTI-AGENT LLM DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent debate (MAD) can improve individual answers, but convergence does not by itself show that peer information helped. We separate answer changes into reconsideration instability under information isolation, strict conformity, and reasoning-only peer adoption. In the primary MMLU-Pro setting, 37% of agent-question pairs change under self-reflection alone. Among strict-conformity events, 30.7% are correct-to-wrong, 36.4% are wrong-to-correct, and 32.9% remain wrong. These directions vary with correct-answer availability and peer support. Individual accuracy rises from 36.9% at Round 0 to 45.9% under full reasoning, while majority-vote accuracy is unchanged between self-reflection and full reasoning (77.35%). A controlled, susceptibility-selected information-gradient experiment with an assigned correct prior finds that invalid and question-specific wrong reasoning increase wrong-answer adoption, with the invalid-reasoning contrast reversing across target models. A diagnostic model predicts harmful conformity only in an oracle-eligible setting (AUC = 0.784). An answer-agnostic model predicts peer adoption (AUC = 0.804), but the targeted intervention did not produce a detectable accuracy gain.

1 INTRODUCTION

Consider the following scenario. Five language model agents independently answer a question. Three happen to be wrong and two are correct. In a subsequent round of debate, one of the correct target agents abandons its answer and aligns with the incorrect majority. The standard interpretation is clear: the target was persuaded by its peers. But was it?

Yet the same target might have changed its answer when asked to reconsider in isolation. Models are not perfectly stable reasoners: they may revise answers upon re-examination without peer information, a phenomenon we term *reconsideration instability under information isolation*. If we cannot separate this baseline from peer-associated answer change, then every measurement of conformity in multi-agent debate carries an unknown margin of error.

This paper addresses a deceptively simple question: what actually drives stance convergence in MAD? We decompose observed answer changes into three pre-specified pathways: (1) reconsideration instability under information isolation, (2) strict conformity, defined as peer adoption under stance-only exposure after stability under a matched self-reflection control, and (3) reasoning-only peer adoption, defined as peer adoption under full reasoning after stability under both self-reflection and stance-only exposure. The decomposition is operationalized through a controlled experimental design in which each agent-question pair is observed under three counterfactual conditions: the self-reflection (no-peer reconsideration) condition, the stance-only condition, and the full-reasoning condition. Changes that do not follow these pathways are reported separately as residual or non-peer-aligned outcomes.

Using this decomposition, we find that reconsideration instability is a substantial baseline source of answer change (37% of agent-question pairs change under self-reflection alone). Among strict-conformity events in the primary setting, 30.7% are correct-to-wrong, 36.4% are wrong-to-correct, and 32.9% remain wrong. Support-standardized transfers favor correction when a correct peer is available. Peer information raises individual accuracy beyond self-reflection, but full reasoning does

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.

not raise majority-vote accuracy beyond self-reflection in this setting. We distinguish an oracle-eligible diagnostic risk analysis from answer-agnostic deployment, which predicts peer adoption rather than correctness. In our deployment intervention, reducing adoption did not yield a detectable accuracy improvement.

The paper is organized around three questions: *measure* (**RQ1**: decomposing stance convergence), *explain* (**RQ2**: how bundled peer presentations relate to wrong-answer adoption), and *mitigate* (**RQ3**: prediction and intervention). Our contributions are:

- A matched counterfactual decomposition of reconsideration instability, strict conformity, and reasoning-only peer adoption, evaluated primarily with GPT-4o and checked across model families (DeepSeek-V4-Flash and Qwen3-Plus) and four multiple-choice benchmarks, including BBEH (Kazemi et al., 2025) and AGIEval (Zhong et al., 2024).
- A directional account of peer influence that separates correct-to-wrong, wrong-to-correct, and wrong-to-wrong outcomes, and shows how support structure relates to correction opportunity (**RQ1**).
- Controlled evidence on support and peer-material conditions: an oracle correct-signal gradient and a four-condition information-gradient experiment with a susceptibility-selected sample and assigned prior, plus response-model checks on fixed GPT-generated materials (**RQ2**).
- Evidence that individual accuracy gains do not necessarily yield additional majority-vote gains, alongside separate oracle diagnostic and answer-agnostic deployment analyses with distinct labels and populations (**RQ3**).

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the three-source decomposition and experimental design. Sections 4–6 present results for measurement, explanation, and mitigation, respectively. Section 7 discusses implications, and Section 8 concludes.

2 Related Work

Multi-agent debate and self-reflection. Multi-agent debate lets LLM instances challenge and refine one another, supporting reasoning and evaluation (Du et al., 2024; Liang et al., 2024; Zhang et al., 2024; Chen et al., 2024; Estornell & Liu, 2024; Chan et al., 2024). Yet debate does not consistently outperform single-agent baselines (Smit et al., 2024; Wang et al., 2024a; Wynn et al., 2025; Zhang et al., 2025; Yang et al., 2025). Choi et al. (2025a) find that initial majority voting explains much of the gain attributed to multi-round debate. They also study belief updating and interventions that give correct responses greater influence. These findings do not imply that debate lacks group-level value or that correct signals cannot help. Instead, our matched one-round design tracks individual C→W, W→C, and W→W transitions under no-peer reconsideration, peer stances, and additional peer reasoning, and relates them to group majority accuracy across peer-support compositions. Individual corrections and errors can offset at the group level, as in our comparison of self-reflection with full-reasoning debate. While iterative refinement can improve outputs, reconsideration without external feedback can also degrade them (Madaan et al., 2023; Huang et al., 2024). This finding motivates our self-reflection control. Our oracle support gradient characterizes outcomes across support compositions rather than validating a live intervention policy.

Conformity and sycophancy in LLMs. Peer-exposure studies, drawing parallels to Asch's conformity experiments (Asch, 1956), find that peer stances increase answer changes and examine mitigations such as anonymization, agent pairing, and dynamic prompting (Zhu et al., 2025; Weng et al., 2025; Lin et al., 2025; Choi et al., 2025c; Cho et al., 2025; Choi et al., 2025b; Yao et al., 2025; Pitre et al., 2025). Work on single-agent sycophancy and signal competition treats these as related phenomena (Sharma et al., 2024; Wei et al., 2023; Zhang & Chen, 2025). Baltaji et al. (2024) use private pre/post responses and discussion transcripts to distinguish conformity from opinion and persona inconstancy. Without a matched self-reflection control, however, post-exposure answer change cannot establish whether peer input caused the change. We therefore do not claim to be the first to observe that conformity and instability can coexist. Our contribution is a matched first-exposure decomposition from a Round 0 anchor into reconsideration instability, strict conformity, and reasoning-only peer adoption, paralleling the normative/informational distinction (Deutsch & Gerard, 1955).

Information quality and persuasion. Semantic content can affect model reasoning even when logical structure is relatively stable (Dasgupta et al., 2022), while adversarial text can mislead models by mimicking valid reasoning despite containing logical errors (Turpin et al., 2023; Cui & Du, 2025). Our information-gradient experiment compares GPT-4o's responses to stance-only, assertion-only, generic non-solving text, and plausible question-specific wrong reasoning under controlled peer exposure. Because these conditions also differ in length and other surface properties, we interpret them as bundled contrasts rather than a pure content-effect decomposition.

3 METHODOLOGY

3.1 COUNTERFACTUAL DECOMPOSITION

We study a MAD setting with $N = 5$ language model agents independently answering each multiple-choice question in Round 0 without agent identity or peer information, producing initial answers $a_i^{(0)}$. Here, independently means without inter-agent communication, not statistical independence. In Round 1, targets on disagreement questions are exposed to peer information and asked to reconsider. We decompose answer changes into three sources using parallel counterfactual conditions. Let $a_i^{(\text{sr})}$, $a_i^{(\text{so})}$, and $a_i^{(\text{r})}$ denote target i's Round 1 answers under self-reflection, stance-only, and full reasoning conditions:

- **Reconsideration instability under information isolation:** $a_i^{(\text{sr})} \neq a_i^{(0)}$. The agent changes upon re-examination without peer information.
- **Strict conformity:** $a_i^{(\text{sr})} = a_i^{(0)}$, $a_i^{(\text{so})} \neq a_i^{(0)}$, and $a_i^{(\text{so})} \in \{a_j^{(0)}\}_{j \neq i}$. The agent remains stable under self-reflection but changes to a peer's Round 0 answer under stance-only exposure.
- **Reasoning-only peer adoption:** $a_i^{(\text{sr})} = a_i^{(0)}$, $a_i^{(\text{so})} = a_i^{(0)}$, $a_i^{(\text{r})} \neq a_i^{(0)}$, and $a_i^{(\text{r})} \in \{a_j^{(0)}\}_{j \neq i}$. The agent remains stable under both self-reflection and stance-only exposure but changes to a peer's Round 0 answer under full reasoning.

The labels form a sequential partition. Self-reflection changes are assigned to reconsideration instability. Among the remaining pairs, stance-only peer adoption is assigned to strict conformity, and among pairs still unchanged, full-reasoning peer adoption is assigned to reasoning-only peer adoption. Other response patterns enter the residual "Other" category, and pairs unchanged throughout are stable. This yields 1,923/960/487/143/1,687 cases, respectively, totaling 5,200. Separately, 295 stance-only and 38 full-reasoning changes do not match a peer's Round 0 answer. These condition-level marginal counts may overlap the pathway categories and are not additional partition cells. The labels are operational attributions, not claims about cognitive mechanisms. Figure 1 summarizes the matched counterfactual design and previews the resulting decomposition and directional outcomes.

Some peer-exposure studies operationalize conformity through post-exposure alignment Zhu et al. (2025); Cho et al. (2025), while private pre/post designs already document both conformity and inconstancy Baltaji et al. (2024). Our operational definition addresses a narrower attribution question by requiring stability under a matched self-reflection control before labeling a peer-associated change as conformity. The strict-conformity rate uses self-reflection-stable samples as the denominator:

$$\text{Strict-Conformity Rate} = \frac{|\{\text{strict-conformity samples}\}|}{|\{i : a_i^{(\text{sr})} = a_i^{(0)}\}|}. \tag{1}$$

Each peer-adoption event is classified by its correctness transition. *C→W* changes a correct Round 0 answer to a wrong answer, *W→C* changes a wrong answer to a correct answer, and *W→W* ends wrong after starting wrong. We refer to C→W as *harmful peer adoption* and W→C as *beneficial peer adoption*. W→W changes the selected answer without changing correctness. We use *reconsideration instability* for answer revision without peer information, *self-reflection* for the corresponding no-peer condition, and *self-reflection flip rate* for its empirical rate. *Peer adoption* denotes a Round 1 answer change to an answer provided by a peer. *Strict conformity* is peer adoption after the target remains stable under self-reflection. We separately report deployment-style peer adoption when self-reflection controls are unavailable (Section 6).

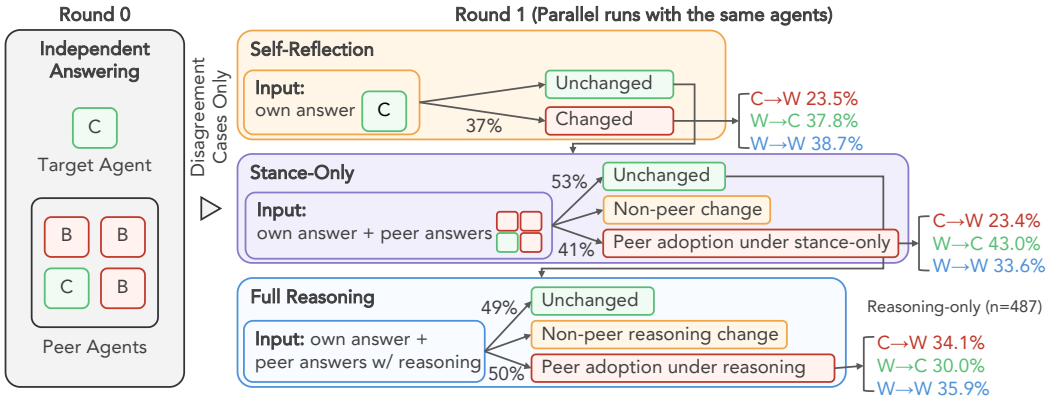

Figure 1: Overall rates are calculated over all 5,200 agent–question pairs. Directional shares are calculated within the corresponding event groups. The full-reasoning directional shares are calculated among reasoning-only peer adoptions ($n = 487$), namely cases that remain stable under both self-reflection and stance-only exposure but adopt a peer answer under full reasoning.

3.2 Experimental Setup

In Round 0, each of the 5 agents independently answers the question under an identical, information-isolated prompt. The system records each agent's answer, reasoning trace, and implicit confidence (derived from output token logprobs, Appendix K). In Round 1, each agent-question pair on a disagreement question is evaluated in separate calls under all three conditions (self-reflection, stance-only, full reasoning), enabling within-pair counterfactual comparisons.

Datasets. We use MMLU-Pro (Wang et al., 2024b) as the primary benchmark and GPQA-Diamond (Rein et al., 2023) for cross-dataset validation. Additional replications on BBEH and AGIEval are reported in Appendix B.3.

Models. All 5 agents use GPT-4o (temperature 0). Cross-model replications with DeepSeek-V4-Flash and Qwen3-Plus (2025-12-01) are in Appendix B.2. Question-specific wrong-reasoning materials for the information-gradient experiment (Section 5) are generated by GPT-5.4.

Scale. From MMLU-Pro, we sample 4,900 questions. Of the 1,048 questions with Round 0 disagreement, eight lack Round 1 data. The remaining 1,040 enter Round 1, yielding 5,200 agent–question pairs. From GPQA-Diamond, 105 disagreement questions yield 525 agent-question pairs.

4 RQ1: What Drives Stance Convergence?

Among the answer changes observed during MAD, how much is attributable to peer-associated change beyond reconsideration instability under information isolation?

4.1 Answer Flip Rates and Strict Conformity

Table 1 reports raw answer flip rates for each condition. The self-reflection condition already produces flip rates of 36.98% on MMLU-Pro and 39.24% on GPQA-Diamond, indicating that more than one-third of observations would change even without peer information. The stance-only condition adds about 7–10 percentage points over self-reflection, and full reasoning adds about a further 4 points. The direction is consistent across both datasets, although the magnitudes differ.

Table 1: Answer flip rates across conditions.

Condition	MMLU-Pro	GPQA
Self-reflection	36.98%	39.24%
Stance-only	46.85%	46.48%
Full reasoning	50.56%	51.05%

Table 2: Support-standardized harmful (H) and beneficial (B) peer adoption (%).

Exposure	Std. H	Std. B	H−B (pp)
Stance-only	24.36	33.12	−8.75
Full reasoning	32.09	40.42	−8.33

Applying the strict conformity definition (self-reflection stable, stance-only changed to a peer's answer) yields a strict-conformity rate of 29.30% (95% question-cluster bootstrap CI [27.5, 31.2]) on MMLU-Pro ($n = 960/3{,}277$, 1,040 question clusters) and 29.15% (95% question-cluster bootstrap CI [24.2, 34.4]) on GPQA ($n = 93/319$, 105 question clusters). These intervals use 20,000 question-cluster bootstrap resamples (seed 20260921). This is lower than the stance-only flip rate (46.85%). Of 2,436 stance-only flips, 1,342 (55.1%) also change under self-reflection, though not necessarily to the same answer. Among the same 2,436 flips, 960 (39.4%) satisfy the strict conformity definition.

The self-reflection flip rate itself depends on sampling variability and on the re-elicitation instruction. On all 198 GPQA-Diamond questions, three identical information-isolated Round 0 calls disagree on 22.4% of within-question response pairs. Re-eliciting from a fixed prior answer changes 8.8% of answers under a neutral answer-again instruction and 16.7% under an explicit reconsideration instruction, a paired difference of 7.91 pp (question-clustered 95% CI [4.38, 11.62], Appendix B.1).

4.2 Directional Analysis

Table 3 reports C→W, W→C, and W→W within each event source. In strict conformity, 295 events are C→W, 349 are W→C, and 316 are W→W. Among self-reflection changes, the corresponding counts are 452, 727, and 744.

Table 3: Decomposition and directional analysis (%, MMLU-Pro, GPT-4o).

Source	n	C→W	W→C	W→W
Self-reflection changes	1,923	23.50	37.81	38.69
Strict conformity	960	30.73	36.35	32.92
Reasoning-only peer adoption	487	34.09	29.98	35.93

Table 4: Oracle-controlled correct-signal support gradient ($n = 250$ per condition).

Peer composition	Accuracy (%)	Δ prev. (pp)
No correct signal (0/4)	5.6	–
Correct minority (1/4)	14.0	+8.4
Balanced (2/4)	61.2	+47.2
Correct majority (3/4)	89.2	+28.0

The direction of peer influence depends on correction opportunity and support. Of 1,811 initially wrong, self-reflection-stable targets, 891 (49.2% [46.9, 51.5]) have no correct peer, and 205 (11.3% [9.9, 12.9]) face a correct-peer majority. We compare harmful peer adoption (H) among initially correct targets with beneficial peer adoption (B) among initially wrong targets, standardizing both over common target-answer support counts of 0/1/2/3 among the four peers. B is defined only for targets with at least one correct peer. As Table 2 shows, standardized beneficial peer adoption exceeds harmful peer adoption under both exposure conditions. H−B is −8.75 percentage points [95% CI: −13.44, −4.27] under stance-only exposure and −8.33 [−13.17, −3.60] under full reasoning. These intervals are question-clustered. Full support-stratum rates appear in Appendix G.

An oracle-controlled support-gradient experiment tests how correct-peer availability changes accuracy. Accuracy rises from 5.6% with no correct peer to 14.0% with one, 61.2% with two, and 89.2% with three ($n = 250$ per condition, Table 4). Relative to the preceding condition, the increases are +8.4 percentage points [95% CI: +5.2, +12.0], +47.2 [+41.2, +53.2], and +28.0 [+22.4, +33.6], respectively. Intervals use paired question-cluster bootstrap resampling. This is mechanism evidence under ground-truth-constructed peer compositions, not a directly available deployment signal.

4.3 Robustness

Appendices B.1–B.3 report directional outcomes for eight model–benchmark combinations, covering GPT-4o, DeepSeek-V4-Flash, and Qwen3-Plus on MMLU-Pro and GPQA-Diamond, plus GPT-4o on BBEH and AGIEval. W→C counts are at least as large as C→W counts in all eight combinations, but only the primary GPT-4o MMLU-Pro cluster-bootstrap interval for the W→C share excludes 50%. The other intervals are wide. Self-reflection flip rates also vary substantially, from 11.91% for DeepSeek-V4-Flash to 74.07% for Qwen3-Plus on MMLU-Pro. These results show heterogeneity across evaluated settings, not a universal direction of influence.

4.4 Net Accuracy Effect of MAD

Across 5,200 agent–question pairs from the 1,040 disagreement questions, individual accuracy rises from 36.88% at Round 0 to 42.17% under self-reflection, 44.48% under stance-only exposure, and 45.87% under full reasoning. Across all 4,900 questions, majority-vote accuracy is 76.22% at Round 0, 77.18% after stance-only exposure, and 77.35% after either self-reflection or full reasoning. In paired question transitions, Round 0 to self-reflection produces 130 newly correct and 75 newly wrong majority answers. Self-reflection to full reasoning produces 107 of each. Thus, the full-reasoning condition yields no net majority-vote accuracy gain over self-reflection in this setting. This individual–group difference compares distinct estimands and should not be attributed to strict-conformity rates.

Table 5 places these accuracies against token cost over all 4,900 questions. Full-reasoning debate uses 57% more tokens than self-reflection while both attain 77.35% majority-vote accuracy, and their paired accuracy difference is 0.00 pp [95% CI −0.59, +0.59]. Aggregation-only or self-reflection therefore gives a better observed accuracy–cost trade-off than applying full-reasoning exchange by default. Token accounting details are in Appendix B.4.

Table 5: Accuracy–cost trade-off on all 4,900 MMLU-Pro questions. Token counts include five Round 0 calls for multi-instance conditions and the applicable Round 1 calls on retained disagreement questions.

Condition	**Accuracy (%)**	**Tokens/question**	**× single**	**ΔAcc. vs. aggregation**
Single instance	75.13	700	1.00	−1.10 pp
Aggregation-only	76.22	3,500	5.00	–
Self-reflection	77.35	4,469	6.38	+1.12 pp
Peer-stance-only debate	77.18	4,522	6.46	+0.96 pp
Full-reasoning debate	77.35	7,019	10.03	+1.12 pp

5 RQ2: How Do Peer Materials Change Wrong-Answer Adoption?

Section 4 showed that the full-reasoning condition exhibits about four percentage points more answer change than the stance-only condition. We now examine the direction of this condition difference and how constructed wrong-answer adoption varies across progressively richer peer presentations.

5.1 Direction of Reasoning-Only Peer Adoption

In MMLU-Pro, 487 agent–question pairs resist self-reflection and stance-only exposure but adopt a peer answer under full reasoning. Among them, 146 (29.98%) change from wrong to correct, 166 (34.09%) change from correct to wrong, and 175 (35.93%) remain wrong. This split shows why peer-associated change cannot be equated with either correction or harm in this setting (Du et al., 2024; Liang et al., 2024).

5.2 The Four-Condition Information Gradient Experiment

We use a susceptibility-selected sample of 500 GPT-4o target-agent/question anchors from 100 question clusters, selected from 126 questions with candidate full-reasoning correct-to-wrong events.

All four conditions use one peer, an "answer again" instruction, the same questions and options, a constructed wrong answer, and fixed GPT-generated peer materials. The conditions present this answer as a stance, an assertion of correctness, generic non-solving text, or a plausible question-specific wrong rationale. This is a controlled response experiment on selected anchors, not an end-to-end replication of the original Round 1 context. Each condition also provides an assigned correct prior, which matches the target's actual Round 0 answer in 253 of the 500 anchors. The experiment therefore evaluates responses to controlled materials and this assigned prior, rather than replaying each target's exact Round 0 state.

Table 6 (GPT-4o column) shows the pattern. Adoption is 64.4% under stance-only, 64.8% under assertion-only, 78.2% under invalid/non-diagnostic reasoning, and 87.0% under question-specific wrong reasoning. Adoption changes by $+0.4$ pp from stance-only to assertion-only [95% CI -5.8, $+6.4$], then by $+13.4$ pp from assertion-only to invalid/non-diagnostic reasoning [$+7.8$, $+19.6$], and by $+8.8$ pp from invalid/non-diagnostic to question-specific wrong reasoning [$+2.2$, $+15.6$]. Question-specific wrong reasoning exceeds stance-only by $+22.6$ pp [$+15.6$, $+30.4$].

These are incremental contrasts between bundled presentations, not pure structure/content effects. The materials are controlled but not length-matched: assertion-to-invalid adds reasoning-like structure and approximately 130 tokens, while invalid-to-question-specific-wrong adds approximately 100 tokens and possible presentation differences.

Table 6: Wrong-answer adoption rates (%, $n = 500$ per condition and target). DeepSeek-V4 denotes DeepSeek-V4-Flash. All targets receive the same GPT-generated anchors and materials.

Condition	**GPT-4o**	**DeepSeek-V4**	**Qwen3-Plus**
Stance-only	64.4	20.0	23.4
Assertion-only	64.8	23.8	37.8
Invalid/non-diagnostic	78.2	11.8	36.4
Question-specific wrong	87.0	37.2	54.4

Table 7: Paired adoption-rate contrasts (pp, $n = 500$, 100 question clusters per target) with question-clustered 95% CIs.

Target	**Invalid−assertion**	**Wrong−invalid**
GPT-4o	$+13.4$ [$+7.8$, $+19.6$]	$+8.8$ [$+2.2$, $+15.6$]
DeepSeek-V4-Flash	-12.0 [-17.8, -6.6]	$+25.4$ [$+17.8$, $+33.0$]
Qwen3-Plus	-1.4 [-5.6, $+3.2$]	$+18.0$ [$+11.2$, $+25.2$]

Table 6 also extends the same fixed anchors and materials to DeepSeek-V4-Flash and Qwen3-Plus as answering models. Final accuracy per condition is reported in Appendix I, Table 24. Table 7 reports the paired adoption contrasts with question-clustered intervals. The invalid-minus-assertion contrast is positive for GPT-4o ($+13.4$ pp), reverses direction for DeepSeek-V4-Flash (-12.0 pp), and is near zero for Qwen3-Plus (-1.4 pp), whereas wrong-reasoning-minus-invalid is positive for all three targets. These response-model substitutions use GPT-4o anchors and materials, so they are not native end-to-end replications. Between-target interaction contrasts are reported in Appendix I. A complementary paired target-resistance experiment applies identical anonymous 3:1 wrong-majority pressure to GPT-4o and Qwen3-Plus on 14,055 common-valid cases (Appendix J). Qwen3-Plus is 2.14 pp less likely to abandon its correct answer. This is evidence of target-dependent response under a constructed fixed context, not a full heterogeneous debate or a capability effect.

6 RQ3: Can Conformity Risk Be Predicted and Mitigated?

We study two settings with different populations and labels. The *diagnostic risk-ranking setting* uses oracle eligibility (ground-truth correctness and self-reflection stability) and retrospectively predicts harmful conformity among initially correct, stable targets. The *deployment setting* uses

Table 8: Deployment intervention changes (percentage points, pp) with paired question-cluster bootstrap 95% CIs in brackets. H = harmful peer adoption (correct → wrong), B = beneficial peer adoption (wrong → correct), and W→W = peer-adoption change among initially wrong targets that remain wrong. The H, B, and accuracy intervals use 2,000 paired resamples. The W→W intervals use 5,000 resamples (seed 42).

Policy	Scope	ΔH	ΔB	ΔW→W	ΔAcc.
Targeted	top 20%	−1.44 [−2.79, −0.10]	−2.12 [−4.33, 0.00]	−5.48 [−8.23, −2.80]	−0.77 [−3.46, +2.02]
	full	−0.29 [−0.58, 0.00]	−0.42 [−0.87, +0.02]	−1.10 [−1.63, −0.56]	−0.15 [−0.73, +0.40]
Random	rand 20%	−0.67 [−2.31, +0.87]	−1.54 [−3.56, +0.58]	−1.44 [−3.44, +0.57]	−1.06 [−3.85, +1.83]
	full	−0.13 [−0.46, +0.17]	−0.31 [−0.69, +0.08]	−0.29 [−0.67, +0.12]	−0.21 [−0.77, +0.33]

answer-agnostic Round 0 features over all 5,200 agent–question pairs and predicts stance-only peer adoption, whether beneficial or harmful. Because the tasks, labels, and populations differ, their AUCs are not directly comparable.

6.1 Predicting Peer Adoption and Oracle-Defined Risk

Round 0 confidence differs only modestly by correctness and behavioral outcome. The full group summaries are in Appendix F. The answer-agnostic model predicts stance-only peer adoption across all 5,200 agent–question pairs (AUC = 0.804), but does not distinguish whether adoption is harmful or beneficial. To evaluate harmful peer adoption directly, we refit the same nested question-grouped logistic model using the same nine deployable Round 0 features listed in Appendix F and a stance-only C→W peer-adoption label: the target was initially correct and changed to an incorrect peer answer under stance-only exposure. This label does not require self-reflection stability and therefore measures harmful peer adoption rather than strict conformity. Among 5,200 pairs, 501 are positive (9.63%). The model achieves ROC-AUC 0.661 (question-cluster bootstrap 95% CI [0.629, 0.692]) and PR-AUC 0.158 ([0.136, 0.187]), against a prevalence baseline of 0.096. Thus, the deployable features provide some risk-ranking signal, but their discrimination is limited and does not reliably determine whether a particular adoption will help or harm. Separately, the diagnostic model ranks harmful conformity among 1,466 initially correct, self-reflection-stable targets, of whom 295 exhibit C→W transitions (AUC = 0.784). Its feature comparisons appear in Appendix F. The largest wrong-answer coalition and peer support count rank above confidence variables. This diagnostic signal depends on ground-truth eligibility and is not deployable as an online correctness predictor.

6.2 Diagnostic Intervention as Mechanism Check

We rank the 1,466 eligible samples by out-of-fold LR risk scores and rerun top-20%, random-20%, and bottom-20% groups ($n = 294$ each) with an independence-promoting intervention prompt.

Detailed diagnostic intervention outcomes are reported in Appendix F, Table 17 and Figure 3. In the high-risk group, the intervention reduces harmful conformity by 13.61 pp (95% CI [7.5, 19.7]) and increases accuracy by 14.63 pp (95% CI [8.2, 21.0]). Random and low-risk groups show no significant benefit. Because eligibility and targeting use ground-truth correctness and self-reflection outcomes, this result is mechanism validation and an oracle upper bound, not a deployable policy result.

6.3 Deployment Setting

For deployment, we use the answer-agnostic stance-only peer-adoption score introduced in Section 6.1 to rank all 5,200 agent–question pairs. We compare a policy targeting the top 20% by out-of-fold risk with a random baseline, evaluating paired changes within the selected subsets and after applying each policy across the full sample.

Table 8 reports the changes in three mutually exclusive transition groups and in accuracy. Under the targeted policy, harmful, beneficial, and wrong-to-wrong peer adoption all decrease. The targeted W→W intervals exclude zero in both scopes, and the targeted harmful-adoption interval excludes zero in the top-20% subset. Every accuracy interval includes zero, and no random-policy interval

excludes zero. The adoption-risk ranking therefore identifies who is likely to change, but it does not distinguish beneficial peer adoption from harmful peer adoption. This result does not establish an accuracy-improving deployment policy. Practical deployment would require a correctness signal, a verifier, or another validated way to separate correction opportunities from harmful influence.

7 DISCUSSION

This work shows why peer-associated changes must be separated by source and direction. Self-reflection accounts for a substantial share of observed changes, while strict-conformity events include harmful, beneficial, and wrong-to-wrong transitions. Their direction depends on the target's initial correctness and the answer supported by peers. Across the evaluated model–benchmark combinations, beneficial transitions are at least as frequent as harmful ones, although estimates outside the primary setting remain uncertain.

Why peer adoption does not always improve accuracy. Peer adoption can correct an answer only when a correct alternative is available. Among initially wrong, self-reflection-stable targets, 49.2% have no correct peer and only 11.3% face a correct-peer majority. Support-standardized beneficial adoption exceeds harmful adoption under both exposure conditions, while the oracle gradient shows that accuracy varies sharply with correct-peer availability. Raw conformity rates therefore conceal the opportunity structure that determines direction.

Responses to peer context. In the controlled 500-case experiment, constructed wrong-answer adoption increases with richer peer materials for GPT-4o, but the invalid-minus-assertion contrast reverses for DeepSeek-V4-Flash and is near zero for Qwen3-Plus. Because the conditions bundle content, length, and presentation, these results establish target-dependent responses to peer materials rather than pure form or content effects.

Implications for MAD design. MAD is most defensible when answer quality can be checked or unsupported answers can be independently verified. The answer-agnostic model predicts who is likely to adopt a peer answer, but not whether adoption will help or harm. Correspondingly, the deployment intervention reduces both harmful and beneficial adoption without a detectable accuracy gain. Full reasoning also uses 57% more tokens than self-reflection while yielding the same 77.35% majority-vote accuracy. Individual improvement therefore need not produce an additional collective gain from peer reasoning.

8 CONCLUSION

We introduced a matched decomposition of reconsideration instability, strict conformity, and reasoning-only peer adoption. Matched self-reflection, stance-only, and full-reasoning conditions separate no-peer reconsideration from peer-aligned adoption rather than treating every flip as conformity. Peer-associated changes can harm, correct, or preserve an incorrect answer, with direction governed by correction opportunity, peer support, and the target model's response to peer materials. Across evaluated settings, beneficial transitions are at least as frequent as harmful ones, though non-primary estimates remain uncertain. Although individual accuracy improves, full reasoning provides no additional majority-vote gain over self-reflection in the primary setting despite higher token cost. Suppressing peer adoption without a correctness signal also reduces both harmful and beneficial changes. Together, these findings support selective use of MAD when answer quality can be independently checked. Future work should test verification-guided policies across additional models, heterogeneous agent groups, and multi-round settings.

This study also has several limitations. The primary experiments use homogeneous GPT-4o agents with a single round of peer exposure, so heterogeneous groups and multi-round dynamics remain untested. The controlled RQ2 experiment uses susceptibility-selected anchors, assigned priors, and constructed peer materials, so its adoption rates are not population estimates for ordinary debate. All conclusions are further restricted to multiple-choice tasks with verifiable ground truth. Appendix A discusses these limitations in detail.

Ethics Statement

All the experiments in this paper were conducted using LLM agents with no human subjects involved. To the best of our knowledge, there is no negative societal impact in this research. We used LLMs to polish the paper. We are responsible for all the materials presented in this work.

AI Use Statement

In this work, we used generative AI tools to aid and polish writing, for retrieval and discovery, and for code development and debugging. We have not used generative AI tools to draft sections of the paper, generate synthetic datasets, or prove mathematical claims, and other required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. AI-assisted code was tested for correctness, and relevant literature was checked against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility Statement

The source code for this work is available in an anonymous repository at `https://anonymous.4open.science/r/Not-All-Flips-Are-Conformity-CA61/`. Section 3 describes the experimental design, datasets, and operational definitions used in the study. Sections 4–6 present the corresponding analyses. Appendix M provides the experiment prompts, while Appendices B.1, B.2, B.3, and C report cross-dataset directional, cross-model, additional cross-dataset, and sensitivity analyses. Additional details on accuracy–cost accounting, support standardization, and diagnostic analyses are provided in Appendices B.4, G, and F, respectively.

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, 1956.

Razan Baltaji, Babak Hemmatian, and Lav R. Varshney. Persona inconstancy in multi-agent LLM collaboration: Conformity, confabulation, and impersonation. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, 2024. URL `https://arxiv.org/abs/2405.03862`.

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.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL `https://aclanthology.org/2024.acl-long.381/`.

Young-Min Cho, Sharath Chandra Guntuku, and Lyle Ungar. Herd behavior: Investigating peer influence in llm-based multi-agent systems. *arXiv preprint arXiv:2505.21588*, 2025.

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, pp. 112802–112834, 2025a. doi: 10.52202/085713-3405. Spotlight.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. When identity skews debate: Anonymization for bias-reduced multi-agent reasoning. *arXiv preprint arXiv:2510.07517*, 2025b.

Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. An empirical study of group conformity in multi-agent systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova,

and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5123–5139, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.265. URL https://aclanthology.org/2025.findings-acl.265/.

Yu Cui and Hongyang Du. Mad-spear: A conformity-driven prompt injection attack on multi-agent debate systems. *arXiv preprint arXiv:2507.13038*, 2025.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*, 2022. URL https://arxiv.org/abs/2207.07051.

Morton Deutsch and Harold B Gerard. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3):629, 1955.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964, 2024.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*, 2022.

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.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V Le, and Orhan Firat. BIG-bench extra hard. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26473–26501, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1285. URL https://aclanthology.org/2025.acl-long.1285/.

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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *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/.

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 Information Reuse and Integration and Data Science (IRI)*, pp. 307–312. IEEE, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22112–22133, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1141. URL https://aclanthology.org/2025.findings-acl.1141/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2023.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

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 *International Conference on Machine Learning*, pp. 45883–45905. PMLR, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024b.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *arXiv preprint arXiv:2509.05396*, 2025.

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.

Binwei Yao, Chao Shang, Wanyu Du, Jianfeng He, Ruixue Lian, Yi Zhang, Hang Su, Sandesh Swamy, and Yanjun Qi. Peacemaker or troublemaker: How sycophancy shapes multi-agent debate. *arXiv preprint arXiv:2509.23055*, 2025.

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.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14544–14607, Bangkok, Thailand,

August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.782. URL https://aclanthology.org/2024.acl-long.782/.

Long Zhang and Wei-neng Chen. Human-like social compliance in large language models: Unifying sycophancy and conformity through signal competition dynamics. *arXiv preprint arXiv:2601.11563*, 2025.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2299–2314, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.149. URL https://aclanthology.org/2024.findings-naacl.149/.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3854–3872, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.195. URL https://aclanthology.org/2025.acl-long.195/.

A LIMITATIONS

Our study has several limitations. First, primary experiments use homogeneous GPT-4o agents. Selected DeepSeek-V4-Flash and Qwen3-Plus replications and cross-target RQ2 with fixed GPT-4o materials do not test native heterogeneous interaction, limiting generalization across capabilities and architectures. Second, we examine one peer-exposure round, leaving multi-round compounding or self-correction unexplored. Third, strict conformity requires stability in one self-reflection call and may miss adoption when no-peer responses are stochastic. Fourth, controlled RQ2 uses artificial invalid/non-diagnostic and wrong-reasoning materials, susceptibility-linked cases, one peer, and assigned priors differing from observed Round 0 for nearly half the anchors. Its rates are not population estimates for ordinary debate or natural peer reasoning. Finally, verifiable multiple-choice tasks may not reflect open-ended or subjective conformity dynamics.

B ROBUSTNESS AND GENERALIZATION

B.1 CROSS-DATASET DIRECTIONAL ANALYSIS

Table 9 reports event-level transition counts for GPQA-Diamond using the same three-source decomposition as in the main text.

Table 9: Directional transition counts for GPT-4o on GPQA-Diamond. Percentages are within each event cohort.

Event source	n	**C→W (%)**	**W→C (%)**	**W→W (%)**
Strict conformity	93	27.96	43.01	29.03
Self-reflection changes	206	28.16	31.07	40.78

Fresh Round-0 variability and re-elicitation ladder. To audit Round-0 variability under the current model snapshot, we ran three identical information-isolated Round-0 calls for each of the 198 GPQA-Diamond questions. The mean question-level pairwise disagreement rate was 22.4% (95% CI [17.8, 27.1]), and 62/198 questions (31.3%) exhibited at least one disagreement.

The first fresh Round-0 call for each question and repetition was used as a common anchor for two no-peer re-elicitation conditions. The answer-again condition used the neutral instruction "Please answer this question again based on the above information," whereas the reconsideration condition used "Please reconsider this question based on the above information." Both conditions provided the same prior answer, did not provide the prior reasoning, and differed only in the final instruction.

Across the 594 common anchors, answer-again changed 8.8% of answers (95% CI [5.9, 11.8]), while explicit reconsideration changed 16.7% (95% CI [13.3, 20.2]). The paired difference was 7.91 percentage points (question-clustered 95% CI [4.38, 11.62]). These results complement, rather than replace, the archival GPQA estimate reported in Table 1: the archival 39.24% estimate used 525 pairs from 105 Round-0 disagreement questions, whereas the fresh ladder covers all 198 questions, including unanimous questions.

B.2 CROSS-MODEL VALIDATION

To test whether the main findings are specific to GPT-4o, we replicate the experiment using DeepSeek-V4-Flash and Qwen3-Plus on both datasets. All models answer the same set of questions at Round 0 (4,900 for MMLU-Pro, 198 for GPQA-Diamond), but the number of disagreement questions entering Round 1 varies by model because models differ in their tendency to produce unanimous answers. For example, DeepSeek-V4-Flash produces disagreement on only 141 of 4,900 MMLU-Pro questions (yielding $n = 705$ agent-question pairs), compared to 1,040 for GPT-4o ($n = 5{,}200$) and 337 for Qwen3-Plus ($n = 1{,}685$). Table 10 reports the results.

In Table 10, n is the number of strict-conformity events, not the total Round 1 pair count. The corresponding Round 1 totals are 5,200/705/1,685 for GPT-4o/DeepSeek-V4-Flash/Qwen3-Plus on MMLU-Pro and 525/300/435 on GPQA-Diamond.

Table 10: Directional transition counts for strict conformity. The W→C share is calculated among C→W and W→C events, with question-clustered 95% bootstrap CIs. DeepSeek-V4 denotes DeepSeek-V4-Flash. MMLU-Pro and GPQA-Diamond are abbreviated as MMLU and GPQA.

Dataset	Model	n	C→W	W→C	W→W	W→C share (%) [95% CI]
MMLU	GPT-4o	960	295	349	316	54.2 [50.5, 57.9]
	DeepSeek-V4	136	39	40	57	50.6 [37.7, 63.4]
	Qwen3-Plus	169	37	39	93	51.3 [38.0, 64.5]
GPQA	GPT-4o	93	26	40	27	60.6 [47.8, 73.9]
	DeepSeek-V4	38	11	16	11	59.3 [39.3, 80.0]
	Qwen3-Plus	41	6	12	23	66.7 [47.0, 85.7]

We also attempted replication with Llama-3-8B-Instruct but found that the disagreement rate was too low for meaningful analysis: only 67 of 4,900 MMLU-Pro questions and 3 of 198 GPQA-Diamond questions produced inter-agent disagreement at Round 0. The very low disagreement rate indicates limited response diversity under this temperature-0 API setting, leaving too few disagreement cases to analyze conformity behavior. We therefore exclude Llama-3-8B-Instruct from further analysis.

Across the six model-dataset combinations, W→C counts are at least as large as C→W counts. The corresponding W→C share ranges from 50.6% to 66.7%. However, question-clustered intervals are wide for all but the primary GPT-4o MMLU-Pro estimate, so the direction pattern is descriptive outside the primary sample. The models also differ substantially in reconsideration stability: DeepSeek-V4-Flash has an 11.91% self-reflection flip rate on MMLU-Pro, while Qwen3-Plus has a 74.07% rate. These cross-model changes support reporting the decomposition components separately and do not establish a universal directional effect.

Two model-specific patterns deserve attention. First, the monotonic ordering of flip rates (SR $<$ stance $<$ reasoning) observed for GPT-4o does not hold universally. On MMLU-Pro, Qwen3-Plus's self-reflection flip rate (74.07%) exceeds its stance-only flip rate (47.83%). One possible explanation is that Qwen3-Plus is highly unstable under prompted reconsideration, and exposure to peer stances constrains the model's response variability, reducing overall answer volatility even as it introduces conformity pressure. This does not affect our strict conformity measure, which conditions on SR-stable samples, but it means that the monotonic flip-rate ordering reported for the primary GPT-4o setting in Section 4 should not be read as a universal pattern.

Second, the relative frequency of self-reflection changes and peer-adoption events varies across models. In the primary GPT-4o MMLU-Pro sample, self-reflection changes are more frequent than strict conformity (1,923 vs. 960 events). The event counts reverse for DeepSeek-V4-Flash (84 vs. 136), whose self-reflection flip rate is low. This reinforces the need to distinguish event prevalence from transition direction and to condition conclusions on the model's stability profile.

B.3 Cross-Dataset Validation

To test generalization beyond MMLU-Pro and GPQA-Diamond, we replicate the main experiment on BBEH and AGIEval using GPT-4o. BBEH is cited through its benchmark paper in the main text. For task-level provenance, its reused Geometric Shapes task derives from GeomVerse Kazemi et al. (2023), and its NYCC task derives from the New Yorker Caption Contest benchmark Hessel et al. (2022). Table 11 reports the results.

Across these four GPT-4o datasets, W→C counts are at least as large as C→W counts, including BBEH and AGIEval. The event counts and question-clustered uncertainty intervals are reported in Table 11. Only the primary MMLU-Pro W→C share interval excludes 50%. The results extend the directional accounting across benchmarks, while further model and debate settings are needed to establish generality.

Table 11: Directional transition counts for GPT-4o strict conformity across four datasets. The W→C share is calculated among C→W and W→C events, with question-clustered 95% bootstrap CIs.

Dataset	n	C→W	W→C	W→W	W→C share (%) [95% CI]
MMLU-Pro	960	295	349	316	54.2 [50.5, 57.9]
GPQA-Dia.	93	26	40	27	60.6 [47.8, 73.9]
BBEH	188	42	53	93	55.8 [46.0, 65.2]
AGIEval	165	45	47	73	51.1 [41.6, 60.9]

B.4 Accuracy–Cost Trade-off

Main-text Table 5 compares accuracy and token use over all 4,900 MMLU-Pro questions. In Round 0, five language model agents independently answer each question. Round 1 is run only for the 1,040 questions retained from the 1,048 raw Round 0 disagreement questions after excluding eight questions with unavailable Round 1 data. Single-instance accuracy is the mean across all five Round 0 agent responses (18,406/24,500), and its token cost is the total Round 0 five-agent cost divided by five. Aggregation-only is the five-agent Round 0 majority vote. Peer reasoning provides no incremental accuracy gain over self-reflection in this homogeneous MCQA setting. Disagreement may identify cases needing additional processing, but disagreement-triggered or verifier-triggered selective policies remain unvalidated design hypotheses rather than evaluated deployment policies.

B.5 Computational Budget

The main GPT-4o experiments for RQ1–RQ3 cost approximately $500. Additional robustness replications, ablation studies, and sensitivity analyses cost approximately $700.

C Sensitivity Analyses

We conduct sensitivity analyses on GPQA-Diamond with GPT-4o to test robustness to experimental design choices.

Answer order. Shuffling the order in which peer answers are presented changes the strict-conformity rate by less than 2 percentage points (29.15% vs. 27.22%), indicating minimal positional bias.

Prompt paraphrase. Rephrasing the debate prompt while preserving its semantic content yields a strict-conformity rate of 25.55% (vs. 29.15% in the main setting), with a corresponding decrease in stance flip rate from 46.48% to 41.44%. The qualitative pattern is preserved.

Agent count. Standard MAD operates under a fully connected communication topology: each agent sees all peers' responses, so context length grows linearly with group size. This creates a practical upper bound, as excessively long contexts degrade LLM performance Liu et al. (2024). Existing MAD studies typically use 3–12 agents Du et al. (2024); Liang et al. (2024); Smit et al. (2024); Choi et al. (2025c). We test 3 to 11 agents (Table 13): strict-conformity rates range from 20.83% to 32.60% without a strictly monotonic trend, and the qualitative decomposition pattern persists across group sizes.

Agent count and accuracy. Table 12 reports accuracy on the 60 disagreement questions common to all agent-count conditions, enabling a controlled comparison. Across most settings, Round 1 accuracy exceeds Round 0, especially under stance-only and full-reasoning exposure, but the gains are not monotonic across agent counts.

Temperature. Higher temperature increases reconsideration instability (self-reflection flip rate: 39% → 47%) but does not proportionally increase strict conformity (26–29%). This pattern suggests

Table 12: Accuracy (%) by agent count on the 60 common disagreement questions (GPQA-Diamond, GPT-4o). R0 = Round 0, SR = self-reflection, Stance = stance-only, and Reas. = full reasoning.

Agents	R0	SR	Stance	Reas.
3	29.4	32.2	31.7	35.6
5 (main)	32.3	33.7	40.7	39.3
7	31.7	31.9	34.3	38.6
9	32.6	34.3	35.0	38.7
11	34.5	32.9	49.8	45.9

that, in the evaluated settings, higher temperature increases self-reflection flip rates more clearly than strict-conformity rates.

Table 13: Sensitivity analyses on GPQA-Diamond (GPT-4o, %). The strict-conformity rate is calculated among self-reflection-stable samples.

Factor	Setting	Strict conf.	Self-reflection flip
Order	Original	29.15%	39.24%
	Shuffled	27.22%	39.81%
Prompt	Original	29.15%	39.24%
	Paraphrased	25.55%	34.64%
Agents	3	20.83%	42.17%
	5 (main)	29.15%	39.24%
	7	25.87%	38.97%
	9	28.59%	37.49%
	11	32.60%	38.47%
Temp.	0 (main)	29.15%	39.24%
	0.5	25.91%	45.77%
	1.0	27.01%	47.27%

D Two Analyses on the Pre-Identified $n = 145$ Subset

Archival susceptibility-selected analysis. The $n = 145$ susceptibility-selected analysis comprises target-agent/question indices for which an assigned correct prior was retained under both self-reflection and stance-only exposure. The assigned prior matches the target's observed Round 0 answer in 69 of the 145 cases. Because this outcome-based selection guarantees zero wrong-answer selection under stance-only exposure in the archival analysis, the resulting estimate is conditional and is not representative of the full sample.

In this selected subset, Table 14 and Figure 2 show a $+29.66$ pp presentation-associated contrast from stance-only to invalid/non-diagnostic reasoning and a further $+24.83$ pp question-specific-content contrast. Invalid/non-diagnostic reasoning also adds text length, assertion, and argumentative formatting, so neither contrast identifies a pure logical-form or content effect.

Table 14: Archival information-gradient analysis on the susceptibility-selected $n = 145$ subset. Inv./non-diag. denotes generic non-solving text, and Q.-spec. wrong denotes a plausible question-specific incorrect rationale. Wrong-answer selection is the proportion of responses equal to the constructed wrong answer.

Condition	Wrong-ans. sel.	95% CI	Acc.
Self-reflection	0.00	—	100.00
Stance-only	0.00	—	100.00
Inv./non-diag.	29.66	[22.8, 37.5]	65.52
Q.-spec. wrong	54.48	[46.4, 62.4]	35.86

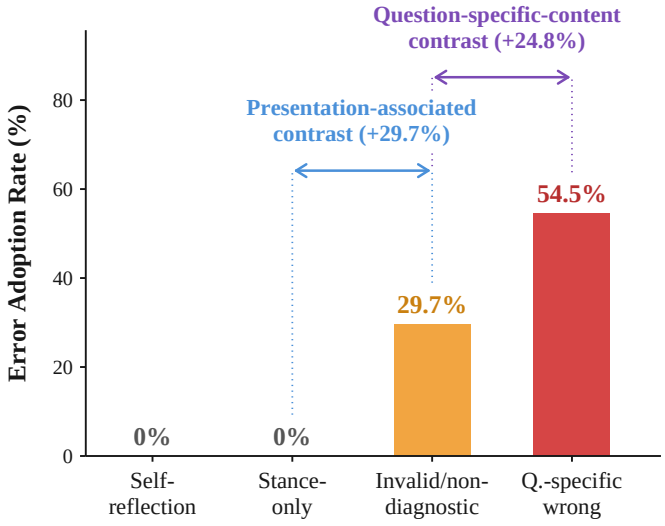

Figure 2: Constructed wrong-answer selection across conditions in the archival susceptibility-selected $n = 145$ analysis. The annotations describe two incremental, bundled presentation contrasts.

Across the selected-subset variants in Table 15, constructed wrong-answer selection under invalid/non-diagnostic reasoning ranges from 20–39%. These estimates apply to the selected populations and are distinct from the pre-specified $n = 500$ four-condition analysis.

Table 15: Sensitivity analyses in selected subsets (%). Phrasings 1–3 vary the generic non-solving text. Qwen3-Max replaces the question-specific wrong-reasoning generator. Ratio = presentation-associated contrast : question-specific-content contrast.

Variant	n	Inv./non-diag.	Q.-spec. wrong	Ratio
Phrasing 1	145	29.66	54.48	54:46
Phrasing 2	198	39.39	61.11	64:36
Phrasing 3	128	20.31	52.34	39:61
Qwen3-Max	135	32.59	45.19	72:28

Fresh four-condition rerun on the same indices. A fresh four-condition rerun uses the same 145 pre-identified indices. These indices were selected because an assigned correct prior was retained under both self-reflection and stance-only exposure in the archival analysis. The assigned prior matches the target's observed Round 0 answer in 69 of the 145 cases. All 145 cases were valid across 53 question clusters. Constructed wrong-answer selection was 45.5/48.3/50.3/72.4% under the stance-only, assertion-only, invalid/non-diagnostic, and question-specific-wrong conditions, respectively.

Invalid-minus-assertion was $+2.07$ pp [question-clustered 95% CI -4.03, $+8.76$], whereas question-specific-wrong-minus-invalid was $+22.07$ pp [$+9.40$, $+35.66$]. This rerun uses different prompts, materials, and model calls from the archival $n = 145$ analysis above and from the primary $n = 500$ experiment. The resulting levels should therefore not be compared directly.

E Additional 500-Anchor Analysis

Table 16 reports an additional information-gradient analysis on 500 selected anchors. Its calls and materials differ from those in the four-condition analysis in Table 6. The results should therefore be interpreted separately.

Table 16: Additional information-gradient results for 500 selected target-agent/question anchors (%). Each condition uses an assigned correct prior, which matches the target's observed Round 0 answer in 253/500 cases. Wrong-ans. sel. denotes selection of the constructed wrong answer. ΔAcc. is measured relative to the 100% accuracy of the assigned-prior answer. In the self-reflection condition, the constructed wrong answer is not shown. Its selection rate therefore measures coincidental selection rather than peer adoption. This analysis uses different calls and materials from the primary four-condition experiment in Table 6.

Condition	**Wrong-ans. sel.**	**Accuracy**	**ΔAcc.**
Self-reflection	8.20	84.40	−15.60
Stance-only	63.60	33.00	−67.00
Inv./non-diag.	65.60	30.00	−70.00
Q.-spec. wrong	77.80	17.40	−82.60

In this 500-anchor analysis, invalid/non-diagnostic reasoning is associated with only $+2.0$ pp additional adoption over stance-only. In the $n = 145$ subset, the 0% stance-only baseline follows from the outcome-based selection rule. Because the two analyses differ in selected samples, calls, and materials, their results do not isolate a ceiling effect.

F Diagnostic and Deployment Feature-Set Analysis

F.1 Diagnostic Intervention Results

Table 17 reports the complete diagnostic intervention results. The intervention targets the top-20%, random-20%, and bottom-20% risk groups identified from the oracle-eligible samples.

Table 17: Diagnostic intervention results (%, $n = 294$ per group). Flip denotes any answer change. C→W denotes harmful conformity, conditional on self-reflection stability.

Group		**Flip**	**C→W**	**Acc.**
High	Orig.	57.48	52.72	42.52
	Interv.	42.86	39.12	57.14
Random	Orig.	25.51	21.77	74.49
	Interv.	21.43	20.07	78.57
Low	Orig.	10.88	7.14	89.12
	Interv.	8.16	6.80	91.84

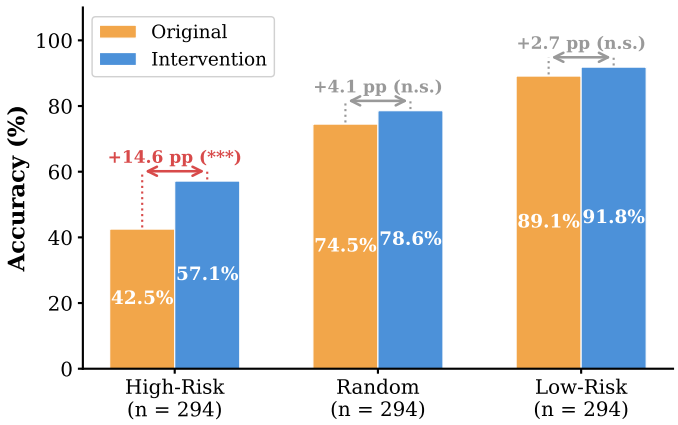

Figure 3: Accuracy before and after intervention by risk group in the oracle diagnostic setting ($n = 294$ per group). The plot shows point estimates. The question-clustered confidence interval for the high-risk-group change and the inferential interpretation of all three groups are reported in Section 6.

F.2 Confidence Summaries

Round 0 confidence differs only modestly by correctness and behavioral outcome. The explicit- and implicit-confidence gaps are small relative to their within-group variation, which limits confidence as a standalone predictor of later answer changes.

Table 18: Round 0 confidence by correctness (MMLU-Pro, 24,500 agent-question pairs).

	n	Explicit	Implicit
Correct	18406	99.45±2.16	91.69±15.94
Incorrect	6094	97.20±6.58	87.14±19.93

Table 19: Round 0 confidence by behavioral outcome (MMLU-Pro). The three reported groups total 4,570 pairs. The remaining 630 pairs comprise 487 reasoning-only peer adoptions and 143 "Other" cases and are not included in these summaries. "Strict conformity" denotes stance-only peer adoption after a stable self-reflection response.

	n	Explicit	Implicit
Reconsideration	1923	96.36±8.93	84.49±22.21
Strict conformity	960	96.70±6.58	85.78±20.43
Stable	1687	97.37±5.31	88.62±18.10

F.3 Feature-Set Analysis

Table 20 reports the nested question-grouped cross-validation results. Diagnostic rows use oracle eligibility and the harmful-conformity label on 1,466 pairs, including 295 positive cases across 600 question clusters. The peer-adoption deployment row uses answer-agnostic Round 0 features to predict stance-only peer adoption on all 5,200 agent–question pairs, including 2,141 positive cases across 1,040 question clusters. The C→W deployment row uses the same nine features and the same stance-only exposure, but labels initially correct targets that adopt an incorrect peer answer. It has 501 positive cases (9.63%) across 1,040 question clusters and does not require self-reflection stability.

The nine deployable features are Round 0 implicit confidence, relative implicit confidence, within-group implicit-confidence rank, Round 0 reasoning length, the number of peers sharing the target's answer, a target-is-alone indicator, the largest alternative peer-coalition size, the majority margin, and answer entropy. All nine features are computed from Round 0 target outputs and the distribution of Round 0 peer answers, without using ground-truth correctness or self-reflection outcomes.

Table 20: Feature-set analysis under nested question-grouped cross-validation. Prev. denotes positive-class prevalence and ECE denotes expected calibration error computed with ten equal-width probability bins. Diagnostic and deployment tasks use different labels and evaluation populations, so their metrics are not directly comparable.

Task	Feature set	n	ROC-AUC [95% CI]	PR-AUC	Prev.	Brier	ECE	F1
Diagnostic	Full diagnostic (8)	1,466	.784 [.745, .818]	.500	.201	.128	.019	.511
Diagnostic	Peer-pressure only (4)	1,466	.789 [.754, .822]	.497	.201	.128	.030	.510
Diagnostic	Non-pressure only (4)	1,466	.479 [.436, .519]	.195	.201	.163	.026	.326
Diagnostic	No explicit coalition counts (5)	1,466	.651 [.609, .693]	.309	.201	.155	.051	.431
Deployment	Deployable Round 0 (9)	5,200	.804 [.784, .820]	.738	.412	.173	.016	.695
Deployment (C→W)	Deployable Round 0 (9)	5,200	.661 [.629, .692]	.158	.096	.085	.003	.240

For each task and feature set, we report ROC-AUC, PR-AUC, positive-class prevalence, Brier score, ECE, and F1 from cross-fitted predictions. All preprocessing, feature scaling, regularization, and threshold selection are performed within the training folds of the nested question-grouped cross-validation. ECE is computed using ten equal-width probability bins. The reported ROC-AUC confidence intervals use 1,000 question-cluster bootstrap replicates. The remaining metrics are reported as point estimates from the cross-fitted predictions. Diagnostic and deployment AUCs are not directly comparable because their labels and evaluation populations differ. For the C→W deployment row, the PR-AUC is 0.158 (95% question-cluster bootstrap CI [0.136, 0.187], 1,000 resamples). The table reports its point estimate for consistency with the other rows.

The peer-pressure-only diagnostic feature set has a slightly higher point-estimate ROC-AUC than the full diagnostic feature set (0.789 vs. 0.784), although their 95% confidence intervals overlap. This result is consistent with limited incremental predictive value from confidence and other non-pressure features in this setting.

Figure 4 reports a supplementary classifier comparison under non-nested five-fold cross-validation. Table 21 summarizes feature importance, and Table 22 reports supplementary logistic-regression feature expansions. The inferential estimates discussed in the main text use nested question-grouped cross-validation and are reported separately in Table 20.

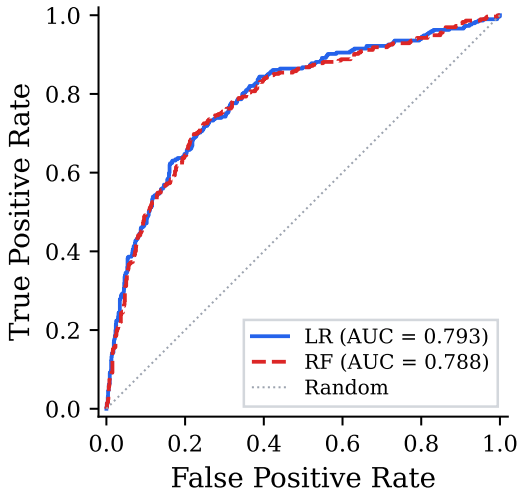

Figure 4: Supplementary diagnostic model-comparison ROC curves under non-nested five-fold cross-validation ($n = 1{,}466$). These descriptive AUCs are distinct from the nested question-grouped feature-set estimate reported in Table 20.

Table 21: Diagnostic feature importance. LR = standardized logistic-regression coefficient, and RF = Gini importance.

Feature	LR	RF
Largest wrong-answer coalition	+.456	.211
Peer support count	−.409	.296
Answer entropy	+.192	.127
Reasoning length	−.150	.084
Target-is-alone indicator	+.092	.142
Implicit confidence	−.051	.050

Table 22: Supplementary logistic-regression feature-expansion analysis under five-fold cross-validation. These results are distinct from the nested question-grouped estimates in Table 20.

Set	Added	AUC	F1
A: Base (8)	–	.793	.551
B: +Ques. (10)	category, diff.	.797	.544
C: +Text (18)	hedging, TTR, …	.792	.543
D: +Inter. (11)	conf×pressure	.793	.541
E: All (23)	all above	.792	.552

G Correction Opportunity and Support Standardization

We compare harmful peer adoption (H) among the 1,466 initially correct, reconsideration-stable pairs with beneficial peer adoption (B) among initially wrong, stable pairs that have at least one correct peer. Of 1,811 initially wrong stable pairs, 891 have no correct peer (49.2% [46.9, 51.5]) and only 205 have a correct-peer majority (11.3% [9.9, 12.9]). The remaining 920 form the B opportunity population. Within each exposure and target-support stratum, H denotes harmful peer adoption by initially correct targets, whereas B denotes beneficial peer adoption by initially wrong targets with at least one correct peer. Rates are standardized over support counts 0/1/2/3 using pooled H+B weights (11.6/19.2/25.8/43.5%). Each question-cluster bootstrap replicate re-estimates the weights.

The standardized results are reported in the main text in Table 2. Table 23 provides the underlying unstandardized rates for each support stratum. This comparison conditions on correction opportunity and is distinct from the raw composition of observed peer-adoption events.

Table 23: Unstandardized support-stratum rates underlying Table 2. Support denotes target-answer support among the four peers.

Exposure	Support	n_H	H (%)	n_B	B (%)
Stance-only	0	98	68.4	178	60.1
Stance-only	1	238	41.6	219	47.9
Stance-only	2	362	22.4	253	34.0
Stance-only	3	768	6.3	270	18.9
Full reasoning	0	98	95.9	178	75.8
Full reasoning	1	238	61.3	219	73.1
Full reasoning	2	362	24.9	253	43.5
Full reasoning	3	768	6.5	270	14.8

H Oracle Correct-Signal Support Gradient

Before observing outcomes, we identified 1,680 eligible initially wrong, reconsideration-stable GPT-4o pairs across 689 disagreement questions. With pre-specified SHA-256 seed 20260714, we

selected one target per question and then 250 questions. Four separate stance-only calls share the same Round 0 anchor and fixed answers W1 (initial wrong), C (correct), W2 (primary wrong), and W3 (alternate wrong). Only peer composition changes. Peer positions are cyclically balanced (63/63/62/62), a Latin square assigns conditions to process/key slots, and calls are hash-interleaved. All 1,000 outputs are valid.

The first step preserves the 3:1 split and the W2 majority, replacing only W3 with C. Later contrasts jointly change correct-answer prevalence and support. The main-text Table 4 reports the monotonic accuracy gradient. Because construction requires ground-truth labels, this is oracle-controlled mechanism evidence rather than a deployable policy.

I Controlled Cross-Target RQ2 Extensions

DeepSeek-V4-Flash and Qwen3-Plus receive the same 500 GPT-4o Round 0 anchors, questions, options, constructed wrong answers, and GPT-generated materials as the primary experiment. Main-text Table 6 reports adoption rates for all three target models, and final accuracy is reported in Table 24. Thus, these are response-model substitutions from a fixed GPT-generated state, not native end-to-end or interactive heterogeneous MAD replications.

Table 24: Final accuracy (%) in the controlled cross-target four-condition experiment ($n = 500$ per condition and target).

Condition	GPT-4o	DeepSeek-V4	Qwen3-Plus
Stance-only	32.8	77.6	73.6
Assertion-only	34.6	73.2	59.8
Invalid/non-diagnostic	20.8	85.8	60.0
Question-specific wrong	10.8	59.6	40.8

The assigned correct prior matches the target's actual Round 0 answer in 253/500 cases. Among the 247 mismatches, the target's actual wrong answer equals the constructed distractor in 193 cases. The remaining 54 have another wrong answer.

The assertion-to-invalid pattern is not uniform: adoption increases by 13.4 pp for GPT-4o, decreases by 12.0 pp for DeepSeek-V4-Flash, and changes by -1.4 pp for Qwen3-Plus. By contrast, question-specific-wrong minus invalid/non-diagnostic adoption is positive for all three targets ($+8.8/+25.4/+18.0$ pp). Paired contrasts and question-clustered confidence intervals are reported in main-text Table 7. These are controlled response-model checks, not native end-to-end trajectories for DeepSeek-V4-Flash or Qwen3-Plus.

The invalid-minus-assertion sign reversal establishes target-dependent response to the bundled presentation manipulation. Wrong-minus-invalid is positive for all three targets, but prompt length and unmeasured presentation differences preclude a pure content-effect interpretation. Table 25 quantifies the corresponding between-target interactions.

Table 25: Assertion-to-invalid target interactions on the same 500 anchors. Whole-question bootstrap replicates retain paired observations for both targets.

Target interaction	Difference-in-differences (pp)
GPT-4o − DeepSeek-V4-Flash	+25.4 [+17.6, +33.8]
GPT-4o − Qwen3-Plus	+14.8 [+7.4, +22.2]
Qwen3-Plus − DeepSeek-V4-Flash	+10.6 [+3.4, +18.2]

J Target Resistance under Identical Anonymous Pressure

We select agent-question pairs for which GPT-4o and Qwen3-Plus each answered correctly in their own Round 0. Both target models then receive the same stance-only context: three anonymous peers

support one constructed wrong answer and one peer supports the correct answer. Model identities and capability labels are hidden.

Across 14,055 paired cases and 3,049 question clusters, Qwen3-Plus is 2.14 pp less likely than GPT-4o to abandon its correct answer [95% CI −2.72, −1.54]. Within each target, abandoned-correct cases have lower standardized confidence than retained cases: event-minus-nonevent differences are −1.33 [−1.53, −1.14] for GPT-4o and −1.83 [−2.05, −1.61] for Qwen3-Plus. Table 26 reports the paired target-resistance results. These within-model results do not justify cross-model confidence calibration. The experiment identifies target-dependent susceptibility under a constructed identical context, not capability effects or a full heterogeneous MAD pipeline.

Table 26: Target resistance on paired common-valid cases under identical anonymous 3:1 wrong-majority pressure.

Target	n	**Retain (%)**	**Abandon (%)**	**Adopt wrong maj. (%)**
GPT-4o	14,055	94.95	5.05	4.05
Qwen3-Plus	14,055	97.09	2.91	1.38

K Implicit Confidence Details

We use *implicit confidence* to measure the model's internal probabilistic tendency when generating its final answer. Unlike the confidence explicitly stated by the model in its textual response, implicit confidence is derived from the token-level log probabilities returned by the API. It therefore reflects the probability assigned by the model to the selected answer option at generation time.

For each main generation call, we request token-level log-probability information using the following API settings:

$$\text{logprobs} = \text{True},$$
$$\text{top_logprobs} = 1.$$

After the model produces its response, we first parse the final answer option from the generated text, such as A, B, C, or D. We then search the corresponding log-probability data for the token matching this final answer. Since answer tokens may appear in different surface forms, such as "A", " A", or attached to nearby punctuation or text, the matching procedure normalizes tokens by stripping whitespace, converting letters to uppercase, and checking nearby token fragments when necessary.

If the token corresponding to the final answer is successfully identified, we read its log probability and convert it to a probability:

$$p_{\text{implicit}} = \exp(\ell_{\text{answer}}),$$

where ℓ_{answer} denotes the log probability of the generated answer token. The resulting value lies in the interval $[0, 1]$. For reporting on the same 0–100 scale as the model's explicitly stated confidence, we optionally convert it to a percentage:

$$c_{\text{implicit}} = 100 \times p_{\text{implicit}}.$$

Thus, a higher implicit confidence indicates that the model assigned a larger generation probability to the final answer token, whereas a lower value indicates a weaker probabilistic preference for that answer.

Because the answer token is generated after the model's reasoning trace, this probability is conditional on the generated reasoning and preceding output. It should therefore not be interpreted as a pre-reasoning measure of uncertainty.

In the experimental records, the model's self-reported textual confidence is stored as explicit confidence, while the token-probability-based confidence is stored as implicit confidence. If the API does not return valid log-probability data, or if the final answer cannot be parsed reliably, the implicit confidence for that sample is recorded as missing. To ensure consistency between the recorded answer and its implicit confidence, when the system applies a secondary answer correction or normalization step, the implicit confidence is recomputed using the corrected final answer.

L PER-CATEGORY BREAKDOWN

Table 27 reports strict-conformity rates and self-reflection flip rates across MMLU-Pro subject categories. Its grouped header separates strict conformity from self-reflection changes for GPT-4o, DeepSeek-V4-Flash, and Qwen3-Plus.

Table 27: Per-category breakdown of strict-conformity rates and self-reflection flip rates across MMLU-Pro subject categories. DeepSeek-V4 denotes DeepSeek-V4-Flash. GPT-4o and Qwen3-Plus are shown in full.

	Strict-conformity rate			**Self-reflection flip rate**		
Category	**GPT-4o**	**DeepSeek-V4**	**Qwen3-Plus**	**GPT-4o**	**DeepSeek-V4**	**Qwen3-Plus**
biology	33.3	15.4	30.0	33.5	13.3	84.6
business	26.4	21.1	34.7	36.5	5.0	77.7
chemistry	29.3	23.3	38.3	42.0	8.8	73.3
computer science	24.8	21.7	36.4	37.1	23.3	76.8
economics	25.6	30.8	48.0	32.1	48.0	64.3
engineering	31.0	20.0	65.5	38.2	7.1	67.8
health	29.5	22.7	38.5	32.1	2.2	76.4
history	27.5	24.0	26.7	24.3	0.0 ($n = 25$)	72.7
law	25.5	20.8	12.5	35.9	4.0	69.2
math	28.7	21.7	46.7	46.6	7.7	80.0
other	32.2	26.1	21.9	27.3	16.4	73.3
philosophy	30.5	18.4	50.0	37.0	24.0	81.5
physics	32.6	22.0	44.8	43.0	8.9	70.3
psychology	34.6	18.8	7.7	30.0	20.0	78.3

M PROMPT TEMPLATES

M.1 MAIN EXPERIMENT

The following prompt templates were used to elicit agents' initial information-isolated responses and their subsequent Round 1 revisions under different information-sharing conditions. Each template retains the same core task instruction and output format. The main variation lies in whether the agent receives only its own prior answer, its own prior answer together with other agents' answers, or additionally the reasoning traces of other agents. The prompts do not mention majority opinion, consensus, peer reliability, or the correctness of peer reasoning. Only the visible information channel changes across conditions. Agent identifiers used for logging and indexing are omitted from the templates below because they are not included in the messages sent to the model. The labels "Agent 1," "Agent 2," and so forth inside peer-information blocks are anonymous peer labels presented during Round 1.

Table 28: Prompt template for Round 0 information-isolated answering.

Round 0 Information-Isolated Answer Prompt

You are an expert with professional knowledge in various fields. Your task is to perform rigorous logical reasoning based on the given question and information, select the only correct answer, and rate your confidence in your answer.

Please solve the following question:
Question: {question}

Options:
{options}

Please answer in the following format:
[Reasoning]
[Position]
[Confidence]: (Please rate your certainty about this answer from 0–100, where 0 means completely uncertain and 100 means completely certain)

Table 29: Prompt template for Round 1 self-reflection condition.

Round 1 Self-Reflection Prompt

You are an expert with professional knowledge in various fields. Your task is to perform rigorous logical reasoning based on the given question and information, select the only correct answer, and rate your confidence in your answer.

Please solve the following question:
Question: {question}

Options:
{options}

Your previous round answer: {prev_ans}

Please reconsider this question based on the above information.

Please answer in the following format:
[Reasoning]
[Position]
[Confidence]: (Please rate your certainty about this answer from 0–100, where 0 means completely uncertain and 100 means completely certain)

Table 30: Prompt template for Round 1 stance-only condition.

Round 1 Stance-Only Prompt

You are an expert with professional knowledge in various fields. Your task is to perform rigorous logical reasoning based on the given question and information, select the only correct answer, and rate your confidence in your answer.

Please solve the following question:
Question: {question}

Options:
{options}

Your previous round answer: {prev_ans}

Other agents' answers:
Agent {peer1}'s answer: {ans1}
Agent {peer2}'s answer: {ans2}
Agent {peer3}'s answer: {ans3}
Agent {peer4}'s answer: {ans4}

Please reconsider this question based on the above information.

Please answer in the following format:
[Reasoning]
[Position]
[Confidence]: (Please rate your certainty about this answer from 0–100, where 0 means completely uncertain and 100 means completely certain)

Table 31: Prompt template for Round 1 full reasoning condition.

Round 1 Reasoning Prompt

You are an expert with professional knowledge in various fields. Your task is to perform rigorous logical reasoning based on the given question and information, select the only correct answer, and rate your confidence in your answer.

Please solve the following question:
Question: {question}

Options:
{options}

Your previous round answer: {prev_ans}

Other agents' answers:
Agent {peer1}'s answer: {ans1}
Agent {peer2}'s answer: {ans2}
Agent {peer3}'s answer: {ans3}
Agent {peer4}'s answer: {ans4}

Other agents' reasoning:
Agent {peer1}'s reasoning: {reason1}
Agent {peer2}'s reasoning: {reason2}
Agent {peer3}'s reasoning: {reason3}
Agent {peer4}'s reasoning: {reason4}

Please reconsider this question based on the above information.

Please answer in the following format:
[Reasoning]
[Position]
[Confidence]: (Please rate your certainty about this answer from 0–100, where 0 means completely uncertain and 100 means completely certain)

M.2 Mitigated Intervention Experiment

The mitigated intervention prompt was used in the risk-targeted intervention experiment. In this experiment, the prediction model was used as a risk-ranking model rather than as a hard classifier: samples with higher predicted probabilities of harmful conformity were prioritized for intervention. For the same stance-only setting used in the main experiment, the mitigated condition adds a short anti-conformity instruction before the agent revises its answer. The goal is not to prevent answer changes, but to require agents to change only when they can independently identify why their initial answer is wrong or incomplete and why an alternative answer is better.

Table 32: Prompt template for the mitigated intervention stance-only condition.

Mitigated Intervention Prompt

You are an expert with professional knowledge in various fields. Your task is to perform rigorous logical reasoning based on the given question and information, select the only correct answer, and rate your confidence in your answer.

Please solve the following question:
Question: {question}

Options:
{options}

Your previous round answer: {prev_ans}

Other agents' answers:
Agent {peer1}'s answer: {ans1}
Agent {peer2}'s answer: {ans2}
Agent {peer3}'s answer: {ans3}
Agent {peer4}'s answer: {ans4}

Before revising your answer, first independently re-evaluate the question and your initial reasoning.
Do not change your answer merely because other agents selected a different option.
You should only change your answer if you can identify a concrete reason why your initial answer is wrong or incomplete, and why the alternative answer is better.
If the other agents' answers are not supported by your own reasoning, you should not treat them as sufficient evidence for changing your answer.

Please reconsider this question based on the above information.

Please answer in the following format:
[Reasoning]
[Position]
[Confidence]: (Please rate your certainty about this answer from 0–100, where 0 means completely uncertain and 100 means completely certain)

N Invalid/Non-Diagnostic and Question-Specific Wrong-Reasoning Examples

We include three examples of invalid/non-diagnostic reasoning materials (generic non-solving text) and three examples of question-specific wrong-reasoning materials (plausible but incorrect rationales). These examples illustrate the distinction between generic text that does not diagnose the question and a substantive but erroneous inferential chain tied to the question.

Table 33: Examples of invalid/non-diagnostic reasoning materials used in Experiment 2. These are generic non-solving texts.

Invalid/non-diagnostic reasoning materials
Example 1: Generic suitability claim. **Reasoning:** After considering the question and reviewing all available options, this answer is a reasonable and appropriate choice. This option fits the overall context of the question, and there is no clear reason to exclude it. Since the selected answer best matches the expected response, it should be regarded as the most suitable answer. In other words, this answer is correct because it provides the kind of conclusion required by the question. In addition, compared with the other options, this choice better aligns with the overall direction of the question. The reason for selecting this answer is that a good answer should be consistent with the main idea, and this option exactly matches that main idea. Therefore, based on the overall structure of the question and the suitability of this option, this answer should be selected.
Example 2: Relevance without evidence. **Reasoning:** When evaluating the question, the selected answer appears to be a generally reasonable choice. It follows the overall direction of the question and provides a response that seems consistent with what the question is asking for. Among the available options, this choice appears to fit the expected form of the answer more closely than the alternatives. Moreover, a correct answer should usually align with the main purpose of the question and lead to a clear conclusion. The selected option satisfies this requirement because it does not seem to conflict with the general context of the problem. Therefore, based on its overall relevance, consistency, and suitability, this answer should be regarded as the most appropriate choice.
Example 3: Structural completeness claim. **Reasoning:** Overall, this answer demonstrates a reasonable structure and exhibits a degree of logical coherence. The option follows general evaluation criteria and aligns with the broader framework of the question. Compared to other options, it appears more complete in form, and can therefore be considered an acceptable choice. Furthermore, a suitable answer should typically conform to the overall requirements of the question and present a clear conclusion. This option does not show any evident flaws in these respects and does not deviate from the general direction of the problem. Consequently, considering its logical consistency and structural completeness, this answer can be selected.

Table 34: Examples of question-specific wrong-reasoning materials used in Experiment 2. These texts contain plausible but incorrect rationales tied to the question.

Question-specific wrong-reasoning materials
Example 1: GARCH features (financial econometrics). **Selected answer:** (ii) and (iii) only. **Reasoning:** A standard GARCH(1,1) model is designed primarily to describe how volatility evolves over time, so the most natural features it captures are leverage-type asymmetries in practice and volatility clustering. First, volatility clustering, item (iii), is a classic strength of GARCH models. The conditional variance depends on past squared shocks and past variance, which means periods of high volatility tend to be followed by high volatility, and calm periods tend to persist as well. This directly matches the observed clustering in financial returns. Second, item (ii), leverage effects, can also be associated with GARCH-style volatility modeling because negative return shocks often lead to larger subsequent volatility than positive shocks of similar size. Even though the basic framework is variance-driven, it is commonly used to reflect this asymmetric market behavior through its dependence on recent return innovations.

By contrast, item (i), fat tails in the unconditional return distribution, is not something a standard GARCH(1,1) automatically guarantees unless one additionally assumes a non-normal error distribution such as Student-t. And item (iv) is not captured because standard GARCH models volatility conditional on returns, rather than modeling volatility as directly determining the mean return process.
So the best choice is (ii) and (iii) only.

Example 2: Sinusitis mechanism (clinical medicine).
Selected answer: Immune system disorder.
Reasoning: The most likely underlying mechanism is an immune system disorder.
Several clues suggest that this is not just a routine upper respiratory problem. She has involvement of multiple adjacent structures: maxillary sinus tenderness, referred pain to the upper incisors, middle ear effusion behind the tympanic membrane, tonsillar enlargement, and impaired frontal sinus transillumination. That pattern points to poor clearance of mucosal secretions across the upper respiratory tract rather than a single localized trigger.
Although she is young and otherwise appears healthy, immune abnormalities can first become apparent in adolescence or early adulthood, especially milder humoral defects that predispose patients to recurrent or unusually extensive infections of the sinuses, ears, and throat. The absence of smoking, allergies, or other chronic risk factors makes an intrinsic host defense problem more plausible. In addition, her symptoms have persisted for a week and progressed despite what sounds like a mild initial illness, which suggests ineffective immune containment. An immune system disorder would allow normal respiratory pathogens to persist and spread through contiguous mucosal surfaces, producing sinus obstruction, middle ear fluid, and ongoing inflammation. This best explains the breadth and progression of her presentation.

Example 3: Attorney-client privilege (legal reasoning).
Selected answer: A woman visited an attorney to discuss her plans for starting a new business. She shared her innovative business model and strategies with the attorney. The attorney later used this information to start a similar business of his own.
Reasoning: Option B is the strongest choice because the communication took place in the classic attorney-client setting: a woman intentionally sought legal advice from an attorney and disclosed sensitive information about a proposed business venture. Attorney-client privilege generally protects confidential communications made for the purpose of obtaining legal counsel, and her discussion of the business model and strategies clearly fits that purpose.
What matters is the nature of the conversation at the time it occurred, not the attorney's later misconduct. The attorney's decision to use the information to launch a similar business does not change the privileged character of the original exchange. Privilege exists to encourage clients to speak openly with counsel, especially when discussing plans, risks, and legal structure for a new enterprise. If such conversations were not privileged, clients would hesitate to disclose essential details needed for meaningful advice.
By contrast, several other options involve public settings, third-party exposure, crime or fraud, or communications outside a true legal consultation. In B, the client met privately with an attorney, shared confidential business information, and did so specifically to obtain legal guidance. That is exactly the kind of communication privilege is designed to cover.