HIDE THE VERDICT, KEEP THE EVIDENCE: INQUIRY DIALOGUE FOR MULTI-AGENT LLM DEBATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-Agent Debate (MAD) typically requires agents to publicly disclose their initial answers to peers before any argument is exchanged, creating a persuasion-like setting vulnerable to conformity. We introduce INQUIRY, a verdict-blind, inquiry-inspired protocol in which agents still form a private answer each round, but explicit verdicts are withheld from peers until a final decision round. Across four multiple-choice QA benchmarks, we compare INQUIRY against Single-Agent, Independent (no-debate) aggregation, MAD-SoM, and MAD-Conformist baselines using both a frontier and a small open-weight (SLM) judge panel. On the SLM panel, INQUIRY achieves the highest multi-agent accuracy on every dataset, outperforming Independent by 5.7-20.6 percentage points, while SoM, the dominant early-commitment protocol, is often the weakest multi-agent method; only on CommonsenseQA does INQUIRY fall short of the single strongest individual judge's high solo accuracy. On the frontier panel, near-ceiling accuracies leave far less room for protocols to differ, suggesting verdict exposure matters most for weaker judges. Correct Reasoning Survival and Minority Truth Recovery diagnostics show that INQUIRY better preserves correct majorities and recovers substantially more minority-held correct answers than the other baselines. A further ablation finds that a mandatory argument-structure schema (Claim, Grounds, Warrant) harms small-panel accuracy on three of four datasets but has no detectable effect at frontier scale. These results identify verdict exposure as a consequential MAD design choice whose effects depend on model capability.

1 INTRODUCTION

Multi-Agent Debate (MAD) is increasingly used to improve LLM reasoning and evaluation: multiple agents independently generate initial responses, inspect their peers' reasoning, and revise their judgments before a final answer is selected (Du et al., 2024). Peer interaction, however, does not uniformly improve collective reasoning. Debate often fails to outperform simpler single-agent or ensembling baselines (Smit et al., 2024; Zhang et al., 2025), and its apparent gains are often attributable to majority voting rather than the exchange of arguments itself (Choi et al., 2026; Bertalanič & Fortuna, 2026). Although peer interaction can correct independent errors, it can also amplify confidently stated mistakes through sycophancy (tendency of LLMs to defer to confidently stated positions (Sharma et al., 2024)) and conformity to an incorrect majority, with agents abandoning correct answers to agree with persuasive but flawed peer reasoning (Pitre et al., 2025; Wynn et al., 2025; Weng et al., 2025). Identifying interaction designs that encourage evidence-based correction over conformity is key to building reliable MAD systems.

A structural choice shared by many MAD protocols which has received limited attention is that agents disclose an initial position before examining their peers' reasoning. In Walton's taxonomy of conversational contexts (Walton, 1998), this interaction resembles a Persuasion Dialogue, in which participants begin with committed positions and subsequently defend or revise them through argument. Many MAD systems inherit this early-commitment structure without explicitly considering whether it matches the epistemic objective of the task. If our aim is to identify the best-supported answer rather than the most successfully defended initial position, the objective is closer to Inquiry Dialogue, which seeks to establish whether a proposition is supported by the available evidence (Walton, 1998; Bezou-Vrakatseli et al., 2024).

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.

Early verdict exposure provides a direct channel for the failure modes above: a confidently stated peer verdict is the signal that sycophantic and conformist updating acts upon (Yao et al., 2025), and public commitment to an initial answer can suppress the subsequent generation of genuinely new reasoning, echoing the Degeneration-of-Thought effect observed in self-reflection (Liang et al., 2024). We therefore hypothesize that withholding explicit answer during evidence exchange removes the object of conformity while preserving the benefits of peer critique. We instantiate this principle in INQUIRY, a verdict-blind deliberation protocol in which judges exchange supporting evidence for up to three rounds while their explicit answer is withheld from peers until a final decision stage (Figure 1). Within this framework we also ask a secondary question: whether the withheld evidence is better exchanged as unconstrained free text or through a mandatory Toulmin-inspired Claim–Grounds–Warrant schema (Toulmin, 1958), a manipulation orthogonal to verdict-blindness that tests whether making claims more specific and falsifiable further improves peer critique.

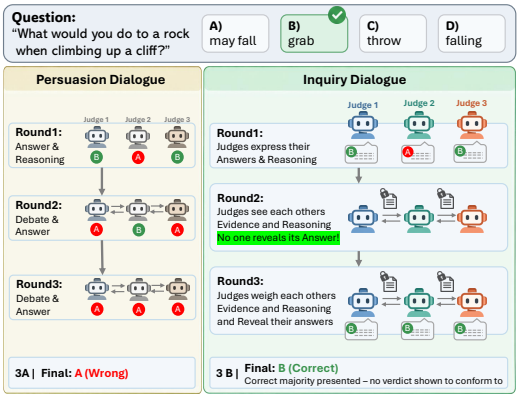

Figure 1: Persuasion versus Inquiry Dialogue. In persuasion protocols, judges commit to an answer before any evidence is exchanged, so a correct majority can erode as judges abandon their answer after observing a confidently stated but incorrect peer verdict. In Inquiry, judges exchange evidence while keeping their answers hidden until the final round, preserving the correct majority that visible-verdict exchange would otherwise erode.

We evaluate INQUIRY against Single-Agents, Independent aggregation and two multi agent debate protocol baseline: SoM (Du et al., 2024) and Conformist (Choi et al., 2026), using two heterogeneous three-judge panels (small open-weight models (SLMs) and frontier LLMs) across four benchmarks. On the SLM panel, INQUIRY achieves the highest accuracy among the evaluated multi-agent protocols on all four benchmarks, outperforming Independent aggregation by 5.7–20.6 percentage points (all $p < .01$), whereas SoM sometimes underperforms Independent aggregation. Deliberation diagnostics show that INQUIRY better preserves correct majorities than SoM and more frequently recovers minority-held correct answers than both MAD baseline, indicating that correct judges are less often pressured into abandoning their answers by confident but incorrect peers. On the frontier panel, accuracies cluster near ceiling and most protocol differences are not statistically detectable, limiting conclusions about the benefits of verdict-blindness at higher capability levels. Mandatory Toulmin-style argument structure reduces SLM accuracy on three of four benchmarks without detectable gains on the frontier panel, suggesting that withholding explicit verdicts and imposing structured argumentation have distinct, capability-dependent effects.

Our contributions are: **1) Verdict-blind framing.** We formulate explicit peer-verdict exposure, distinct from the reasoning exchanged, as a consequential MAD design variable and a potential conduit for conformity, inter-agent sycophancy, and Degeneration-of-Thought. Drawing on dialogue theory, we characterize common early-disclosure protocols as persuasion-like and propose INQUIRY, a verdict-blind, inquiry-inspired protocol for truth-seeking QA. **2) Controlled protocol evaluation.** We compare five protocol families on matched panels. We use independent aggregation to isolate the benefit of voting alone from the benefit of debate. Within INQUIRY, we factorially vary argument format and final decision rule and repeat the evaluation with both SLMs and frontier-model panels. **3) Capability-stratified findings.** We characterize when verdict-blind deliberation improves collective accuracy, distinguishing between the preservation of correct judgments and the recovery of minority-held correct answers to clarify the benefits and the limitations of INQUIRY.

2 RELATED WORK

MAD and Independent Aggregation. Du et al. (2024) popularized MAD protocols in which agents revise initial solutions after observing peers' answers and rationales. Variants employ role-diverse debate for evaluation (Chan et al., 2024), peer ranking and discussion (Li et al., 2024), heterogeneous

models with confidence-weighted consensus (Chen et al., 2024), and divergent debate to counter Degeneration-of-Thought (Liang et al., 2024). However, gains may arise from pooling independent predictions rather than interaction: Choi et al. (2026) find that voting explains much of MAD's benefit and that, under their assumptions, debate does not necessarily improve expected correctness. This motivates our INDEPENDENT baseline, which applies plurality voting to non-interacting panel members, separating aggregation from deliberation. Despite variations in roles and rounds, prior protocols generally expose peer answers during deliberation and thus do not isolate verdict channel.

Verdict Exposure, Conformity, and Sycophancy. LLMs may revise toward social signals rather than stronger evidence, aligning with an interlocutor's beliefs at the expense of accuracy (Sharma et al., 2024). In multi-agent settings, Weng et al. (2025) characterize the effects of majority size and interaction time and test persona and reflection-based mitigations, while Zhu et al. (2025) show that conformity to incorrect majorities increases with uncertainty. Yao et al. (2025) formalize inter-agent sycophancy and link it to premature disagreement collapse. These dynamics parallel *groupthink* in human committees, where pressure toward concurrence suppresses dissent and realistic appraisal of alternatives (Janis, 1972). Closest to our intervention, Choi et al. (2026) anonymize response sources, hiding *who* produced an answer while leaving the answer visible; we target the latter channel. Qu et al. (2026) further find that peer agreement more readily misleads initially correct models than corrects initially wrong ones, motivating our CRS and MTR analyses. Existing mitigations modify personas, reflection instructions, source identities, or aggregation rules; ours removes the explicit peer-verdict signal while preserving argument exchange.

Dialogue Type and Argument Structure. Persuasion dialogues begin from conflicting positions and resolve them by revising commitments; inquiry dialogues begin from a need for proof and seek to establish or refute a hypothesis through shared evidence (Walton, 1998; Black & Hunter, 2009), a recent work advocates for joint human–LLM reasoning (Bezou-Vrakatseli et al., 2024). We characterize both SoM (Du et al., 2024) and Conformist (Choi et al., 2026) as persuasion-like and INQUIRY as verdict-blind and inquiry-inspired: evidence is exchanged before explicit answers. Verdict concealment is our operational mechanism, not a claim to instantiate Inquiry Dialogue formally. Toulmin-inspired prompting makes an argument's claim, evidence, and inferential link explicit (Toulmin, 1958). Gupta et al. (2024) prompt LLMs to generate such components zero-shot, and Huang et al. (2026) uses Toulmin-structured arguments during multi-agent conflict resolution, combined with external evidence scoring and Bayesian confidence updates. Our structured condition instead requires a three-part *Claim–Grounds–Warrant* exchange within verdict-blind QA, varied factorially against free-text exchange while holding verdict exposure fixed.

3 PRELIMINARIES AND PROBLEM FORMULATION

Task setup. We consider a multiple-choice dataset $\mathcal{D} = \{(x_j, \mathcal{Y}_j, y_j^\star)\}_{j=1}^{J}$, where x_j is a question, $\mathcal{Y}_j$ is its candidate-answer set, and $y_j^\star \in \mathcal{Y}_j$ is the correct answer. A panel contains M language-model agents $\mathcal{A} = \{a_1, \ldots, a_M\}$. For question j, agent a_i produces at round t a private assessment

$$s_{ij}^{(t)} = \left(v_{ij}^{(t)}, e_{ij}^{(t)}\right), \tag{1}$$

where $v_{ij}^{(t)} \in \mathcal{Y}_j$ is its current verdict and $e_{ij}^{(t)}$ is its accompanying argument. We treat $e_{ij}^{(t)}$ as observable justification text, without assuming that it faithfully represents the model's latent reasoning or is necessarily correct.

Multi-agent deliberation. In MAD, agents initially answer independently and subsequently revise their assessments after observing peer messages (Du et al., 2024). Let $R_{\mathcal{P}}$ denote the disclosure rule of protocol $\mathcal{P}$: before round t, agent a_i conditions its updated assessment $s_{ij}^{(t)}$ on the question, its candidate answers, and the disclosed history $\{R_{\mathcal{P}}(s_{kj}^{(\tau)}) : k \neq i,\ \tau < t\}$ of its peers' prior assessments. The central distinction in our study concerns what $R_{\mathcal{P}}$ reveals:

$$R_{\mathcal{P}}(v, e) = \begin{cases} (v, \psi(e)), & \mathcal{P} = \text{PERSUASION}, \\ (\varnothing, \psi(e)), & \mathcal{P} = \text{INQUIRY}, \end{cases} \qquad t < T, \tag{2}$$

where T denotes the final decision stage and ψ specifies the argument format. Persuasion-style MAD exposes both the verdict and argument during deliberation. Inquiry withholds the explicit

verdict from peers while retaining it privately for analysis. Thus, Inquiry removes early *public* commitment without preventing agents from forming provisional private judgments.

We consider either unconstrained free-text arguments or Toulmin-structured arguments comprising a Claim, Grounds, and Warrant (Toulmin, 1958). Verdict-blindness is operational rather than information-theoretic: explicit answer labels are withheld, although the argument may still implicitly indicate an agent's preferred answer.

Collective decision. Let $v_{ij}^{(1)}$ and $v_{ij}^{(T)}$ denote the initial independent and final post-deliberation verdicts, respectively. The Independent baseline aggregates $\{v_{ij}^{(1)}\}_{i=1}^{M}$, whereas deliberative protocols aggregate $\{v_{ij}^{(T)}\}_{i=1}^{M}$. Under vote-based aggregation,

$$\widehat{y}_j = \arg\max_{y \in \mathcal{Y}_j} \sum_{i=1}^{M} \mathbb{1}\left[v_{ij}^{(T)} = y\right], \tag{3}$$

using a protocol-specified tie rule. We use *plurality* as the general term because agents may select three different options. Alternatively, when final verdicts are non-unanimous, a separate non-deliberating meta-judge selects the answer from the agents' final judgments and arguments.

4 Design and Experimental Setup

Protocols. We evaluate five conditions, spanning a spectrum from no interaction to fully verdict-blind deliberation. Only our own protocol, Inquiry, varies along the two design axes motivated in Section 1 (argument structure and decision rule); the remaining four are run as single fixed configurations for baseline, either because they have no peer interaction (Single-Agent, Independent) or because they are literature-defined baselines reproduced as originally specified (SoM, Conformist). The baselines are selected by explicit criteria rather than benchmark convention. SoM is the most widely adopted early-commitment MAD template, popularized by Du et al. (2024) and inherited by most subsequent debate protocols, so it directly instantiates the public commit-then-argue structure our study isolates. Conformist is a minimal, literature-specified modification of that same structure, proposed by the work that attributes much of MAD's benefit to voting rather than interaction (Choi et al., 2026). Single-Agent and Independent bracket the no-interaction end of the spectrum, separating individual judge capability and answer pooling from any effect of deliberation itself.

Panel size (three judges) and Inquiry's round budget (up to three rounds) follow Du et al. (2024)'s own three-agent, two-round configuration, adopted there as a computational-cost tradeoff rather than a point of saturation; we extend it to three rounds to give Inquiry's meta-judge escalation a longer window. We do not treat further rounds as a free source of gains: Choi et al. (2026) report that on SLM judges, unconstrained debate accuracy degrades over successive rounds rather than improving, the same premature-commitment concern motivating our verdict-blind design.

Formal setup. Building on our formulation, a decision rule D maps the panel's round-T verdicts (and, for the meta-judge, full transcript) to a final answer: D_{maj} takes the plurality of $v_{1j}^{(T)}, \ldots, v_{Mj}^{(T)}$ (a three-way split is scored incorrect, Appendix A.3); D_{meta} instead reads the transcript and returns a neutral arbitrator's answer. These five protocols differ only in M, T, $R_{\mathcal{P}}$, D, and two optional mechanisms, *freezing* (Conformist) and *short-circuiting* (Inquiry), will be defined later.

- **Single-Agent.** One answers each question once, with no peers and no deliberation. This is a reference point for what a single model can do unaided, not a multi-agent condition. Formally, $M = 1$, $T = 1$, no peer messages are exchanged, and D trivially returns $v_{1j}^{(1)}$.
- **Independent.** Three judges answer independently in a single call each, with no visibility into one another's reasoning or answers; the panel's answer is the plurality of the three responses. Independent isolates the benefit of pooling multiple opinions from the benefit of any actual peer interaction between them. Formally, $M = 3$, $T = 1$ and $D = D_{\text{maj}}$ over $v_{1j}^{(1)}, v_{2j}^{(1)}, v_{3j}^{(1)}$.
- **SoM** (Du et al.). Judges commit to a verdict in Round 1, visible to every peer, and then exchange two further rounds of free-text peer argument, framed cooperatively as "advice," before a final

majority vote (non-unanimous rounds are scored as incorrect). Formally, $M = 3$, $T = 3$, $R_{\mathcal{P}}$ exposes $(v_{ij}^{(t)}, e_{ij}^{(t)})$ every round, and $D = D_{\text{maj}}$.

- **Conformist** (Choi et al.). Structurally identical to SoM, except that after each round, any judge whose answer already matches that round's plurality winner is *frozen*: it receives no further calls and its answer is carried forward unchanged for all remaining rounds. Only judges still in the minority continue to be re-prompted. This baseline tests whether simply freezing an already-formed majority, without any further deliberation on their part, recovers most of the benefit of full debate. Formally, M, T, $R_{\mathcal{P}}$, and D are identical to SoM's, plus a freezing rule: once $v_{ij}^{(t)}$ matches round t's plurality, $v_{ij}^{(t')} = v_{ij}^{(t)}$ for every $t' > t$ and a_i receives no further calls.
- **Inquiry (ours).** Judges remain verdict-blind for up to three rounds: each judge privately reasons toward a Claim every round, but that answer is stripped before being shown to peers, who see only the supporting evidence exchanged so far. The peer-review prompt is also framed differently from SoM/Conformist's cooperative "advice" exchange: judges are instructed to evaluate each peer's evidence and reasoning on their merits and give a specific factual counter-argument if they are wrong or irrelevant, rather than to weigh a peer's stated opinion.

Inquiry is the only condition in which we examine three separate ablations: 1.*argument structure* (free-text reasoning vs. a mandatory Toulmin-inspired Claim (the answer) / Grounds (The facts or evidence) / Warrant (The logical connection and reasoning) schema, Toulmin, 1958), 2.*decision rule* (plain majority vote vs. escalation to a neutral meta-judge on a non-unanimous final round), and 3.whether Round-1 short-circuiting is applied. Because SoM and Conformist are reproduced from prior work as single fixed configurations (free-text, majority vote), our main cross-protocol comparison is restricted to this shared configuration so it is not confounded by axes only Inquiry possesses; argument structure, decision rule, and short-circuiting are instead examined as three separate ablations internal to Inquiry.

Datasets. We evaluate four multiple-choice QA benchmarks (Table 1), selected to probe distinct failure modes in MAD, including peer pressure, sycophancy, and vulnerability to plausible but incorrect distractors. All four are standardized to a fixed four-option format, reducing/shuffling choices where the source format differs, with seeds fixed for reproducibility.

Table 1: Characteristics of used benchmarks in the study.

Dataset	N	Preprocessing and rationale
MedQA (Jin et al., 2021)	1,273	`GBaker/MedQA-USMLE-4-options` test split, natively four-option. Probes multi-step clinical/causal reasoning.
TruthfulQA (Lin et al., 2022)	684	MC1 validation split; reduced to 1 correct + 3 random incorrect options, items with $<$4 choices excluded (leaving 684 questions of 817). Probes resistance to confidently-worded common misconceptions.
CommonsenseQA (CSQA) (Talmor et al., 2019)	1,221	Validation split; native five-way, one of four incorrect options randomly dropped, remainder reshuffled into A–D. Probes plausible-but-subtly-wrong distractors of a different flavor than the above.
MMLU (Formal Logic) (Hendrycks et al., 2021)	126	Pure reshape (choices list to lettered dict); no exclusions. Probes symbolic/formal reasoning; matches Choi et al. (2026)'s subset.

Judge panels. We evaluate protocols on two heterogeneous judge panels of differing capability. The *frontier* panel comprises GPT-5-mini, Gemini 2.5 Flash, and Claude Haiku 4.5, with the full GPT-5 serving as meta-judge. The *small/open-weight* panel comprises Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Gemma-3n-E4B-it (each 7–8B parameters), and Phi-4 (14B) as meta-judge. In both panels the meta-judge is deliberately a larger model than the judges it arbitrates, so that escalation asks a strictly more capable arbitrator to resolve disagreement. Comparing the same protocols across both panels reveals whether their effects are consistent or depend on judges' competence.

Metrics and statistical testing. Our primary metric is accuracy. We compare protocols with McNemar's exact test, paired on the same set of questions, since every protocol is scored on identical items and the quantity of interest is discordant pairs (questions one protocol gets right and the other does not), not marginal accuracy considered in isolation. Each protocol is run once per question at

a fixed decoding configuration (greedy, temperature $= 0$; Appendix A.2) rather than averaged over repeated trials; statistical power comes from the sample size and is assessed via the paired exact test above. All significance tests are reported at $\alpha = 0.05$, computed separately within each dataset.

Beyond accuracy, we report a panel-level diagnostic that captures how a protocol reaches its final answer, rather than only whether it is correct. Let $y_j^\star$ be the gold answer and $g_i^t(x_j) = \mathbb{1}[v_{ij}^{(t)} = y_j^\star]$ indicate whether agent i's Round-t verdict is correct. **Correct Reasoning Survival (CRS)** measures the percentage of questions for which a correct two-of-three majority at Round 1 remains correct in the final decision, capturing whether deliberation preserves an initially correct majority.

$$\mathcal{M} = \left\{ j : \sum_{i=1}^{3} g_i^1(x_j) \geq 2 \right\}, \quad \mathrm{CRS} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \mathbb{1}\big[D(x_j) = y_j^\star\big]\,. \tag{4}$$

Minority Truth Recovery (MTR): Among questions where only one judge is initially correct and the other two are wrong, the percentage for which the panel ultimately reaches the correct answer.

$$\mathcal{N} = \left\{ j : \sum_{i=1}^{3} g_i^1(x_j) = 1 \right\}, \quad \mathrm{MTR} = \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} \mathbb{1}\big[D(x_j) = y_j^\star\big]\,. \tag{5}$$

Both metrics are computed from Round-1 judge outputs and the panel's final decision. They are undefined for Single-Agent and Independent, which have no post-Round-1 deliberation stage.

CRS and MTR can trade off against each other: a rule that never revisits an already-correct majority scores perfectly on CRS by construction while recovering almost no minority-held truths, which looks favorable on either statistic read alone. We summarize both modes using a harmonic mean:

$$H = \frac{2 \cdot \mathrm{CRS} \cdot \mathrm{MTR}}{\mathrm{CRS} + \mathrm{MTR}}, \tag{6}$$

5 Results

Main accuracy comparison. Table 2 and 3 report accuracy for all protocols on both judge panels. All debate protocols (SoM, Conformist, Inquiry) are restricted to free-text argument, majority-vote decision rule. Independent's row is the plurality of the three single agent responses without debate. At small/open-weight (SLM) panel (Table 2), the three Single-Agent rows are far from clustered:

Table 2: Accuracy (%), Small-Model panel, free text argument. $^*p < .05$ (paired McNemar) is significant difference from Independent.

	Protocol	MedQA	TruthfulQA	CSQA	MMLU
Single	Qwen2.5-7B	61.0	67.7	**82.6**	57.9
Single	Llama-3.1-8B	57.6	50.7	66.3	46.0
Single	Gemma-3n-E4B	50.7	48.4	70.3	45.2
	Independent (3x)	57.0	54.2	74.4	45.2
MAD	SoM	51.4*	49.9*	64.1*	50.8
MAD	Conformist	58.8*	56.6*	74.8	48.4
MAD	Inquiry (Ours)	**67.8***	**74.9***	80.1*	**58.7***

Table 3: Accuracy (%), Frontier panel, free-text argument. Same structure as Table 2. Accuracies sit close to ceiling across all protocols.

	Protocol	MedQA	TruthfulQA	CSQA	MMLU
Single	GPT-5-mini	94.0	89.0	90.0	95.0
Single	Gemini-2.5-Flash	91.0	89.0	89.0	95.0
Single	Claude-Haiku-4.5	84.0	85.0	90.0	84.0
	Independent (3x)	90.0	90.0	91.0	95.0
MAD	SoM	92.0	92.0	89.0	97.0
MAD	Conformist	93.0	91.0	91.0	96.0
MAD	Inquiry (Ours)	**95.0**	**99.0**	90.0	**98.0**

Qwen2.5-7B is the strongest individual judge on every dataset (57.9–82.6%), while Llama-3.1-8B and Gemma-3n-E4B trail by 3–19pp, so the panel's Independent/SoM/Conformist accuracy is effectively anchored by its two weaker members rather than reflecting Qwen's solo ceiling. On CommonsenseQA, Qwen alone (82.6%) already beats Independent's majority vote (74.4%). This mirrors findings in heterogeneous-capability multi-agent debate more broadly: pairing a strong model with weaker peers can pull the group's post-debate/post-vote accuracy down toward the weaker members' level rather than up toward the strongest member's ceiling (Wynn et al., 2025). SoM is frequently the weakest of the non-Inquiry multi-agent protocols, consistent with early verdict commitment actively harming small judges rather than helping them. Inquiry improves over every baseline, including the

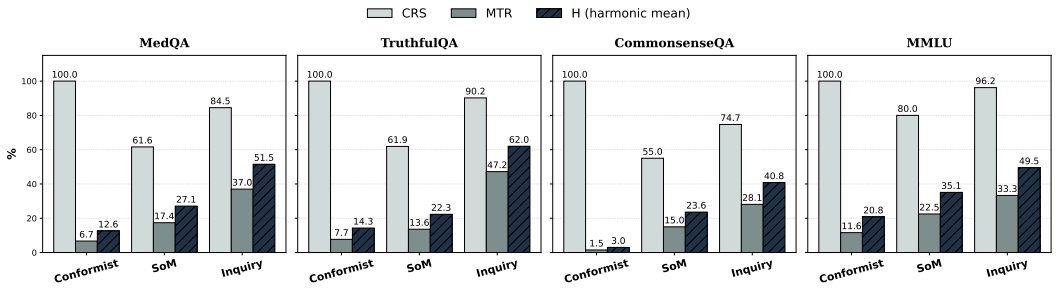

Figure 2: Panel-level CRS, MTR, and their harmonic mean H (%), small-model panel, free-text argument exchange. Conditioning-subset sizes are reported in Appendix A.4 (Table 6).

best individual Single-Agent model, on three of four datasets, reaching significance against Independent on MedQA (+10.8pp, $p < .001$), TruthfulQA (+20.6pp, $p < .001$), and MMLU (+13.5pp, $p < .01$); CommonsenseQA is the exception to beating the best individual model, where Qwen's strong solo performance (82.6%) leaves little headroom for any MAD to add value, though Inquiry still outperforms Independent there too (+5.7pp, $p < .001$), it does not surpass Qwen's ceiling.

At frontier scale (Table 3), the three Single-Agent rows are much closer together (84.0–95.0%), with Claude-Haiku-4.5 consistently the weakest judge (84.0–90.0%) and GPT-5-mini/Gemini-2.5-Flash close to each other and to the MAD; every protocol overall sits within 84.0–99.0% band and only TruthfulQA shows a significant Inquiry improvement over Independent. This is the expected contrast condition: the SLM-tier gains above are not a universal property of Inquiry, they specifically emerge once individual judges are weak enough, and heterogeneous enough in ability, for deliberation to have genuine errors to correct.

Short-circuiting stops deliberation when all judges privately agree after Round 1. It skips 22–48% of debate rounds while changing accuracy by at most 0.5 percentage points, with no significant differences. Continuing beyond unanimity also destabilizes correct consensus more often under SoM than Inquiry. Full results appear in Appendix A.6 .

Mechanism diagnostics. Figure 2 reports panel-level CRS and MTR (whether a correct Round-1 verdict survives to the final answer, or an incorrect one is corrected by it) for SLMs, free-text argument, restricted to the three protocols where such a Round-1-to-final transition exists (Independent is single-round, so Round 1 is the final answer). Conformist's CRS is 100.0% on every dataset by construction: an already-correct majority is frozen and can never be talked out of it, so this is a property of the freezing rule, not evidence of any deliberative benefit. Its MTR is correspondingly low (1.5–11.6%): a lone correct dissenter can rarely overturn a frozen two-judge majority under this rule, so Conformist can barely recover a minority-held truth by design. SoM shows real CRS erosion (55.0–80.0%), an already-correct majority is genuinely talked into the wrong answer some of the time under unconstrained early-commitment debate, and weak MTR (13.6–22.5%), well above Conformist's but still far short of Inquiry's. On the other side, Inquiry's CRS (74.7–96.2%) is consistently higher than SoM's on every dataset, and its MTR (28.1–47.2%) is consistently higher than both baselines. This indicates that Inquiry's accuracy advantage reflects a genuine mechanistic difference (better protection of correct majorities and substantially better recovery of correct minorities) rather than only a different aggregate outcome reached by an equally opaque process. Their harmonic mean H makes this explicit in a single number: Conformist's H (3.0–20.8) is the lowest of the three protocols on every dataset despite its perfect CRS, and Inquiry's H (40.8–62.0) is highest on every dataset, a consistent ranking that CRS or MTR alone does not always give.

Individual-judge transition dynamics. Figure 3 decomposes the panel-level CRS/MTR result down to Qwen2.5-7B, the strongest individual judge, splitting Round-1-to-final transitions into a correct-conditioned pair (C→C held correct vs. C→W flipped to wrong) and a wrong-conditioned pair (W→C flipped to correct vs. W→W held wrong), for every protocol×dataset combination; the full three-judge breakdown, including Llama-3.1-8B and Gemma-3n-E4B, is in Appendix A.7 (Figure 4). Two mechanistic patterns emerge that are invisible in the panel-level aggregates. First, Inquiry raises every judge's individual wrong-to-correct recovery rate (W→C) above both baselines on all 12 judge×dataset cells (Appendix A.7), confirming that the panel-level MTR advantage (Fig-

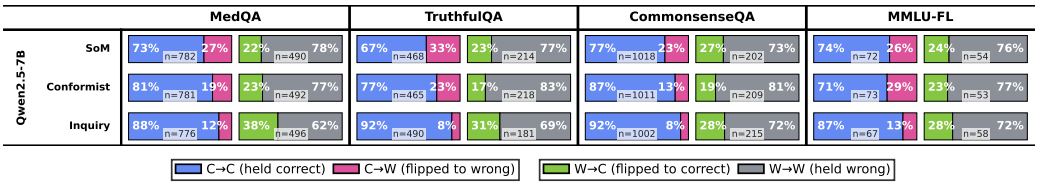

Figure 3: Qwen2.5-7B (strongest judge) Round-1→final-answer transitions, small-model panel, free-text argument exchange (Inquiry uses full deliberation, consistent with Table 2 and Figure 2); pair definitions as in the main text. n is the conditioning population size for each bar; Round-1 uncertain/unparsed claims are excluded. All three judges are reported in Appendix A.7.

ure 2) reflects a genuine, judge-general recovery effect rather than an artifact of majority-vote. The effect is largest for Llama-3.1-8B. Qwen2.5-7B, shows the same directional effect on every dataset but with a far more modest and uneven margin: as little as 1.04-fold (CommonsenseQA, where all three protocols cluster near 19–28%) up to 1.8-fold (TruthfulQA) over the stronger baseline.

Second, Inquiry's protection of already-correct judges (low C→W) is not judge-general in the same way: for Qwen2.5-7B it is a clear, consistent improvement over SoM (8–13% C→W under Inquiry vs. 23–33% under SoM on every dataset, matching or beating Conformist's mechanical freezing without relying on it). Llama and Gemma show a more mixed picture instead, detailed per-judge numbers are in Appendix A.7. Conformist's own low C→W is confirmed here to be a mechanical artifact of its freezing rule rather than a real deliberative benefit: it comes paired with only moderate W→C recovery for every judge, the same disconnect between protecting correct answers and creating new ones already noted for its panel-level CRS/MTR profile.

Argument structure: free-text vs. Toulmin. Table 4 isolates the effect of argument structure within Inquiry (SLMs panel, majority-vote decision rule). Free-text outperforms Toulmin on three of four datasets (MedQA, TruthfulQA, and MMLU). CSQA is the exception, where free-text and Toulmin are statistically indistinguishable (80.1% vs. 80.5%, $p = .746$). Table 5 repeats the same comparison on the frontier panel, with the majority-vote decision rule held constant (matching Table 3's main comparison). Here, free-text and Toulmin are statistically indistinguishable across all datasets. These results suggest that the format-following burden imposed by Toulmin's Claim/Grounds/Warrant schema is a judge-specific cost for smaller models rather than an inherent property: forcing small judges into Toulmin's structure either hurts or provides no benefit, whereas frontier judges applying the same schema to the same tasks exhibit no measurable loss in accuracy.

Table 4: Argument-structure ablation within Inquiry, small-model panel, majority-vote decision rule held constant. Exact paired McNemar test. Bold indicates a significant difference.

	MedQA	TruthfulQA	CSQA	MMLU
Free-text (%)	**67.8**	**74.9**	80.1	**58.7**
Toulmin (%)	63.5	70.9	80.5	46.8
p	.001	.014	.746(ns)	.009

Table 5: Argument-structure ablation within Inquiry, frontier panel, majority-vote decision rule held constant. Exact paired McNemar test. No dataset shows a significant difference.

	MedQA	TruthfulQA	CSQA	MMLU
Free-text (%)	95.0	99.0	90.0	98.0
Toulmin (%)	89.0	98.0	91.0	97.0
p	.070(ns)	1.000(ns)	1.000(ns)	1.000(ns)

Decision rule: majority vote vs. meta-judge. All main-text tables report Inquiry under the majority-vote rule, matching SoM/Conformist/Independent so the cross-protocol comparison is not confounded by an axis only Inquiry possesses; the meta-judge nonetheless already arbitrates every non-unanimous round regardless of which rule a table reports, so both answers come from the same saved transcripts (Appendix A.5). On SLM panel, escalating non-unanimous rounds to a meta-judge significantly improves accuracy over plain majority vote on three of four datasets (MedQA 67.8%→70.1%; TruthfulQA 74.9%→76.5%; CommonsenseQA 80.1%→81.7%: all $p = .05$; MMLU-FL n.s.). The reported majority-vote numbers are therefore a conservative lower bound on Inquiry's achievable accuracy at this panel scale. On the frontier panel the two rules are statistically indistinguishable on all datasets, consistent with far fewer non-unanimous rounds to arbitrate.

6 DISCUSSION

Deliberation depends on capability and protocol. For the SLM panel, INQUIRY significantly outperforms independent voting on all four benchmarks, whereas early-verdict SOM underperforms it on three. This supports findings that conventional MAD does not reliably improve over voting or ensembling (Smit et al., 2024; Choi et al., 2026), while showing that our verdict-blind interaction can provide additional gains. It also refines evidence that weaker models offer greater opportunities for deliberative improvement (Yang et al., 2025): they benefit only when the protocol limits harmful peer influence. At frontier scale, differences are mostly undetectable, likely reflecting stronger individual performance, limited headroom, and smaller samples.

Verdict exposure affects revision dynamics. INQUIRY better preserves correct majorities and recovers substantially more minority-held correct answers than SOM and CONFORMIST. This is consistent with research on conformity, uncertainty-sensitive revision, and premature consensus in multi-agent systems (Weng et al., 2025; Zhu et al., 2025; Yao et al., 2025; Qu et al., 2026). CONFORMIST's perfect CRS is produced mechanically by freezing the majority and is accompanied by weak minority recovery, showing that consensus stability alone is not evidence of productive deliberation. However, because reasoning may still imply a preferred answer, our findings concern the explicit verdict channel rather than complete stance concealment.

Structured arguments are not universally beneficial. Mandatory Claim–Grounds–Warrant exchange reduces small-panel accuracy on three benchmarks and provides no improvement on the fourth. Prior work shows that LLMs can generate Toulmin-style components (Gupta et al., 2024), but this does not guarantee better collective decisions. Our result instead aligns with evidence that restrictive formats can impair reasoning (Tam et al., 2024; Kaesberg et al., 2025). It also qualifies Huang et al. (2026)'s work, where argument structure is combined with evidence scoring and Bayesian aggregation rather than isolated experimentally. The absence of a frontier-scale effect suggests a capability-dependent formatting burden, although differences in decision rules and sample sizes prevent a formal cross-panel interaction claim.

Aggregation rule matters too. On SLMs, escalating non-unanimous rounds to a neutral meta-judge rather than taking a plain majority vote further improves accuracy on three of four datasets, without changing the underlying transcripts. This is a specific instance of a broader idea already present in the LLM-judge and evaluation literature (Li et al., 2024) and Choi et al. (2026)'s finding that voting alone explains much of MAD's benefit: even after debate, *how* a split panel's verdict is finally decided still matters. Our contribution here is narrower: confirming the effect on a fully open-weight small-model panel, where a plain majority vote would otherwise be the default choice.

7 CONCLUSION

This work examines how peer-verdict exposure, argument format, and decision rules shape collective accuracy, and revision dynamics across model-capability regimes. For SLM panels, early-verdict SOM can underperform INDEPENDENT aggregation, whereas verdict-blind, INQUIRY, achieves the highest accuracy on all benchmarks. CRS and MTR indicate that these gains reflect different revision dynamics, particularly improved recovery of correct minority answers, although preservation of initially correct reasoning varies across judges. CONFORMIST can prevent erosion of existing agreement through freezing, but provides less minority-truth recovery. At frontier scale, most protocol differences are not statistically detectable, with TruthfulQA as the principal exception. Requiring Claim–Grounds–Warrant structure reduces small-panel accuracy on three benchmarks but has no detectable frontier-scale effect. Since exchanged reasoning may still imply a preferred answer, our findings concern concealment of the explicit verdict channel rather than complete stance concealment. Overall, with weaker judges, effective deliberation relies more on controlling signals, preserving independent judgment, and minimizing interaction overhead.

Our evaluation is limited to four multiple-choice benchmarks, two three-model panels, and at most three rounds; generalization to open-ended tasks and larger-scale deliberation remains untested. Verdict blinding removes explicit answer labels but not positions implicitly revealed through reasoning. Moreover, the cross-panel structure ablations differ in sample size and decision rule, precluding a formal capability-by-format interaction claim. Future work should test broader tasks, adversarial misinformation, and fully local alternatives to costly or closed meta-judges

AI USE STATEMENT

In this work, we used generative AI tools (GitHub Copilot / Claude) for debugging the code that implements our experimental pipeline and for producing the statistical tests and figures reported in the paper. We have not used generative AI tools to formulate the core research question, hypotheses, or the protocol/decision-rule design presented in this paper, these were conceived and specified by the authors, with AI tools used only to refine and validate them in code. This work does not involve human subjects research, dataset collection from human participants, or theoretical proofs, so disclosure on those fronts is not applicable.

Additionally, we used AI tools for tasks with recommended disclosure: conducting literature search and summarizing related work as a starting point for our own reading, and editing the text of the paper for grammar, clarity, and structure. We have reviewed all AI-assisted work.

We note that the multi-agent debate protocols studied in this paper use large language models as judges, this is the object of study of the paper itself, not a tool used by the authors to produce the paper, and is unrelated to the disclosures above.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe our full experimental methodology in the main text (Section 4) and appendix. Every protocol's exact prompt template – Single-Agent/Independent, SoM, Conformist, the meta-judge, and all three rounds of Inquiry in both the Toulmin and free-text variants – is reproduced verbatim in Appendix A.1. Model versions, decoding parameters, completion-token budgets, the JSON-repair retry mechanism, and answer-parsing fallback are described in Appendix A.2. Dataset sources, preprocessing steps, and random seeds are given in Table 1. Our tie-breaking convention for non-unanimous outcomes and the short-circuit ablation methodology are detailed in Appendix A.3 and Appendix A.6, respectively. The statistical test used to compare protocols (McNemar's exact paired test) and the definitions of our CRS, MTR, and H metrics are given in Section 4. The anonymized code, prompts, and precomputed results needed to reproduce every reported number and figure are included as supplementary material with this submission, along with a no-API-key demo that exercises the full decision logic on scripted judge responses.

ETHICS STATEMENT

This work uses existing, publicly available multiple-choice QA benchmarks (TruthfulQA, MedQA, CommonsenseQA, MMLU) and does not collect any new data from human subjects; no personally identifiable information is involved. All judge panels are commercial or open-weight LLMs queried through their standard APIs, and no model was fine-tuned on any private or sensitive data. We do not foresee this work raising concerns under the ICLR Code of Ethics.

REFERENCES

Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. In *Proceedings of the ACM Conference on AI and Agentic Systems*, CAIS '26, pp. 311–329, New York, NY, USA, 2026. Association for Computing Machinery. doi: 10.1145/3786335.3813137.

Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. Towards dialogues for joint human-ai reasoning and value alignment. *arXiv preprint arXiv:2405.18073*, 2024.

Elizabeth Black and Anthony Hunter. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems*, 19(2):173–209, 2009.

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 Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085. Association for Computational Linguistics, 2024.

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? *Advances in Neural Information Processing Systems*, 38: 101732–101764, 2026.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235, pp. 11733–11763. PMLR, 2024.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. Harnessing Toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10259–10276. Association for Computational Linguistics, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Kenfeng Huang, Yi Cai, Xin Wu, Zikun Deng, and Li Yuan. SMADE-IE: Sparse multi-agent framework with evidence-driven debate for zero-shot information extraction, 2026.

Irving L. Janis. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin, Boston, 1972.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. doi: 10.3390/app11146421.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Voting or consensus? decision-making in multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11640–11671, 2025.

Ruosen Li, Teerth Patel, and Xinya Du. PRD: Peer rank and discussion improve large language model based evaluations. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904. Association for Computational Linguistics, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229.

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. doi: 10.18653/v1/2025.findings-acl.1141.

Jiaming Qu, Lucheng Fu, and Yibo Hu. Easier to mislead than to correct: Harmful and beneficial revision in llm conformity. *arXiv preprint arXiv:2606.01637*, 2026.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.

Andries P. Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? A look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235, pp. 45883–45905. PMLR, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1218–1236, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.91.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, Cambridge, UK, 1958.

Douglas Walton. *The New Dialectic: Conversational Contexts of Argument*. Toronto Studies in Philosophy. University of Toronto Press, Toronto, Canada, 1998. ISBN 9780802079879. doi: 10.3138/9781442681859.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. In *ICLR*, 2025.

Andrea Wynn, Harsh Satija, and Gillian Hadfield. Talk isn't always cheap: Understanding failure modes in multi-agent debate. In *Proceedings of Machine Learning Research*, volume 267, 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*, abs/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, 2025.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3854–3872, 2025.

A Additional Analyses and Implementation Details

A.1 Prompt templates

For exact reproducibility, this section reproduces every judge-facing prompt template verbatim, with the runtime-substituted fields (`question_context`, `peer_arguments`, `peer_answers`) replaced by a bracketed description of their content. `question_context` is always the question stem followed by its four lettered options (A–D). All prompts require a JSON response matching the schema shown; no free-form prose responses are accepted.

Single-Agent / Independent Multi-Agent (shared baseline template). Also used, unchanged, as Round 1 of SoM and Conformist below.

```
[question_context]
Answer this question.
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "reason": "<concise reasoning that weighs the options and identifies
             the correct one, max 50 words>",
  "answer": "A" | "B" | "C" | "D"
}
```

Inquiry (ours), Toulmin variant – Round 1: independent evaluation.

```
[question_context]
---
ROUND 1 -- INDEPENDENT EVALUATION
Think through the evidence for and against each option before answering.
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "option_arguments": [
    {
      "option": "A" | "B" | "C" | "D",
      "stance": "supports" | "refutes",
      "grounds": "<specific fact or evidence relevant to this option,
                  max 25 words>",
      "warrant": "<why this fact supports or refutes the option,
                  max 25 words>",
      "backing": "<short label for a source or established fact,
                  max 12 words>"
    }
  ],
  "claim": "A" | "B" | "C" | "D",
  "confidence": "high" | "medium" | "low"
}

IMPORTANT:
- Do NOT state your chosen answer anywhere except the private "claim" field.
- List option_arguments for AT MOST 2 options -- whichever ones you find
  evidence-worthy; you are not required to fill all 2 slots.
- Keep every field within its stated word limit; write short factual
  phrases, not full sentences.
```

Inquiry (ours), Toulmin variant – Round 2: peer review (verdicts hidden).

```
[question_context]
---
ROUND 2 -- PEER REVIEW (evidence-first critique; verdicts hidden)
Your peers' Round 1 evidence (their private claim has been removed):
[peer_arguments: each peer's Round-1 option_arguments, with "claim" stripped]
Evaluate each peer's Grounds and Warrant BEFORE deciding whether to update
your own view:
- If a peer's evidence is valid and relevant, say so and reinforce it
  with additional support.
- If a peer's evidence is wrong or irrelevant, give a specific factual
  counter-argument (not just disagreement).
- Reconsider your own Grounds and Warrant in light of this critique, then
  decide whether to keep or revise your Claim.
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "option_arguments": [
    {
      "option": "A" | "B" | "C" | "D",
      "stance": "supports" | "refutes",
      "grounds": "<max 25 words>",
      "warrant": "<max 25 words>",
      "backing": "<short label, max 12 words>"
    }
  ],
  "peer_responses": [
    {"to": "<judge id>", "point": "<your specific factual response to
                                   this judge's evidence, max 40 words>"}
  ],
  "claim": "A" | "B" | "C" | "D",
  "confidence": "high" | "medium" | "low"
}
IMPORTANT:
- Do NOT state your chosen answer anywhere except the private "claim" field.
- List option_arguments for AT MOST 2 options.
```

Inquiry (ours), Toulmin variant – Round 3: final answer (simultaneous reveal).

```
[question_context]
---
ROUND 3 -- FINAL ANSWER (simultaneous reveal)
Your peers' Round 2 evidence and rebuttals (verdicts still hidden):
[peer_arguments: each peer's Round-2 option_arguments and peer_responses]
This is the final round. Reveal your answer now. No peer has seen your
answer yet -- this is a simultaneous reveal, not a response to anyone
else's final choice.
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "option_arguments": [
    {
      "option": "A" | "B" | "C" | "D",
      "stance": "supports" | "refutes",
      "grounds": "<max 25 words>",
      "warrant": "<max 25 words>",
      "backing": "<short label, max 12 words>"
    }
  ],
  "peer_responses": [
    {"to": "<judge id>", "point": "<max 40 words>"}
  ],
  "claim": "A" | "B" | "C" | "D",
  "confidence": "high" | "medium" | "low",
  "answer": "A" | "B" | "C" | "D",
  "answer_reason": "<one decisive sentence citing the key reason,
                   max 40 words>"
}
```

The free-text variant below reuses this exact three-round structure and verdict-blindness mechanism unchanged; only the schema differs, replacing the four Toulmin fields (`option_arguments`: `grounds`/`warrant`/`stance`) with a single free-text `reasoning` field. Free-text is the argument style used for all main-text tables and figures; Toulmin above is the structured-argument ablation reported in Table 4 (and Table 5 for the frontier panel).

Inquiry (ours), free-text variant – Round 1: independent evaluation.

```
[question_context]
---
ROUND 1 -- INDEPENDENT EVALUATION
Think through the evidence for and against each option before answering.
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "reasoning": "<free-text reasoning about the evidence for/against the
               most relevant options, max 100 words>",
  "claim": "A" | "B" | "C" | "D",
  "confidence": "high" | "medium" | "low"
}
IMPORTANT:
- Do NOT state your chosen answer anywhere except the private "claim" field.
```

Inquiry (ours), free-text variant – Round 2: peer review (verdicts hidden).

```
[question_context]
---
ROUND 2 -- PEER REVIEW (evidence-first critique; verdicts hidden)
Your peers' Round 1 reasoning (their private claim has been removed):
[peer_arguments: each peer's Round-1 reasoning, with "claim" stripped]
Evaluate each peer's reasoning BEFORE deciding whether to update your own
view:
- If a peer's reasoning is valid and relevant, say so and reinforce it
  with additional support.
- If a peer's reasoning is wrong or irrelevant, give a specific factual
  counter-argument (not just disagreement).
- Reconsider your own reasoning in light of this critique, then decide
  whether to keep or revise your Claim.
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "reasoning": "<free-text reasoning about the evidence for/against the
               most relevant options, incorporating your peers' critique,
               max 100 words>",
  "claim": "A" | "B" | "C" | "D",
```

```
  "confidence": "high" | "medium" | "low"
}
IMPORTANT:
- Do NOT state your chosen answer anywhere except the private "claim" field.
```

Inquiry (ours), free-text variant – Round 3: final answer (simultaneous reveal).

```
[question_context]
---
ROUND 3 -- FINAL ANSWER (simultaneous reveal)
Your peers' Round 2 reasoning (verdicts still hidden):
[peer_arguments: each peer's Round-2 reasoning]
This is the final round. Reveal your answer now. No peer has seen your
answer yet -- this is a simultaneous reveal, not a response to anyone
else's final choice.
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "reasoning": "<free-text reasoning about the evidence for/against the
               most relevant options, max 100 words>",
  "claim": "A" | "B" | "C" | "D",
  "confidence": "high" | "medium" | "low",
  "answer": "A" | "B" | "C" | "D",
  "answer_reason": "<one decisive sentence citing the key reason,
                   max 40 words>"
}
```

SoM (Du et al., 2024) and Conformist (Choi et al., 2026) – Rounds 2–3: peer exchange (verdicts visible). Round 1 reuses the Single-Agent/Independent template above unchanged. The same peer-exchange template below is reused for both Round 2 and Round 3 (no separate final-reveal framing, since answers are already visible from Round 2 onward). Conformist uses this identical template but freezes a judge's answer (stops calling it) once that judge matches the previous round's plurality winner; the prompt text itself does not change.

```
[question_context]
These are the answers to the question from other agents:
[peer_answers: each other judge's "answer" and "reason" from the previous round]
Using the opinion of other agents as additional advice, can you give an
updated response?
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "reason": "<concise reasoning that weighs the options and identifies
             the correct one, max 50 words>",
  "answer": "A" | "B" | "C" | "D"
}
```

Meta-judge (invoked only when Inquiry's Round 3 is non-unanimous).

```
[question_context]
Three judges independently debated this question across multiple rounds
and did NOT reach a unanimous final answer. Below is their full
deliberation transcript (each round's evidence, critiques, and final
answers):
[transcript: the full round-by-round record for all three judges]
You are an independent meta-judge. Read the transcript and decide the
correct answer using your own judgment -- do not simply defer to the
majority if you find their reasoning unconvincing.
You MUST respond in valid JSON matching EXACTLY this schema:
{
  "reason": "<concise reasoning that weighs the judges' arguments and
             identifies the correct answer, max 50 words>",
  "answer": "A" | "B" | "C" | "D"
}
```

A.2 Decoding parameters and API implementation details

All judge and meta-judge calls, across every protocol and both panels, use greedy decoding (temperature $= 0$) for determinism; the only other sampling parameter set is a per-panel completion-token budget. The frontier panel (GPT-5-mini, Gemini 2.5 Flash, Claude Haiku 4.5) is allowed 2,000 completion tokens per call, since GPT-5-mini and Gemini 2.5 Flash consume part of this budget on hidden reasoning/thinking tokens before producing visible output. The SLM panel (Qwen2.5-

7B-Instruct, Llama-3.1-8B-Instruct, Gemma-3n-E4B-it, self-hosted via vLLM) uses a smaller 600-token budget: these models have no hidden reasoning tokens competing for the budget, and the self-hosted deployment has a hard 4,096-token combined prompt+completion context window, so a smaller completion allowance leaves adequate room for the prompt once later rounds accumulate peer-argument text.

Every call is wrapped in a single bounded repair retry: if a judge's raw reply cannot be parsed as a JSON object matching the expected schema, one follow-up turn is sent asking the model to resend only the JSON object, with no surrounding text or markdown code fence; whatever comes back (repaired or not) is then used as-is. If the response still cannot be parsed as JSON after this retry, a regex-based fallback extracts a claim or answer letter directly from the raw text (matching against the literal schema key, e.g. `"claim": "A"`) rather than discarding the response outright, so a single malformed turn does not need to be dropped from the dataset. All judge, meta-judge, and dataset-preprocessing random seeds are fixed.

A.3 Tie handling for non-unanimous outcomes

Neither Du et al.'s SoM (Du et al., 2024) nor Choi et al.'s Conformist (Choi et al., 2026) specify a tie-break rule for a non-unanimous final round, and Independent has no such precedent to draw on either. We do not introduce one: a genuine three-way split (no unique plurality answer) is scored as incorrect for all three protocols, the same convention used for every other non-unanimous outcome under the majority-vote rule. In the SLM-panel data used in this paper, such splits are infrequent but non-negligible for Independent (2.2–15.1% of questions across the four datasets, highest on MMLU, the smallest dataset) and rarer for SoM and Conformist (0.0–3.2%), consistent with Independent being the condition most likely to produce a genuine disagreement since its judges never see each other's reasoning.

A.4 Mechanism diagnostics: CRS/MTR conditioning subsets

Figure 2 in the main text reports CRS, MTR, and their harmonic mean H for the small-model panel, free-text argument exchange. Table 6 below reports the same values together with the size n of each conditioning subset (the number of questions with a correct two-of-three Round-1 majority, for CRS, or a correct one-of-three Round-1 minority, for MTR), which the figure omits for readability.

Table 6: Panel-level CRS, MTR, and their harmonic mean H, small-model panel, free-text argument. n is the size of conditioning subset.

Dataset	Protocol	CRS % (n)	MTR % (n)	H %
MedQA	SoM	61.6 (284)	17.4 (265)	27.1
	Conformist	100.0 (286)	6.7 (268)	12.6
	Inquiry (Maj.)	84.5 (310)	37.0 (238)	51.5
TruthfulQA	SoM	61.9 (155)	13.6 (162)	22.3
	Conformist	100.0 (149)	7.7 (168)	14.3
	Inquiry (Maj.)	90.2 (173)	47.2 (106)	62.0
CommonsenseQA	SoM	55.0 (238)	15.0 (200)	23.6
	Conformist	100.0 (236)	1.5 (194)	3.0
	Inquiry (Maj.)	74.7 (190)	28.1 (128)	40.8
MMLU-FL	SoM	80.0 (25)	22.5 (40)	35.1
	Conformist	100.0 (24)	11.6 (43)	20.8
	Inquiry (Maj.)	96.2 (26)	33.3 (33)	49.5

A.5 Decision rule: majority vote vs. meta-judge

Every main-text table reports Inquiry under the majority-vote rule, matching SoM/Conformist/Independent, so that the cross-protocol comparison is not confounded by an axis only Inquiry possesses. Because the meta-judge is invoked on every question whose final round is non-unanimous, regardless of which decision rule a table reports, both the plain majority-vote

Table 7: Decision-rule ablation within Inquiry, small-model panel, free-text argument, full deliberation held constant. Exact paired McNemar test. Bold indicates a significant difference.

	MedQA	TruthfulQA	CSQA	MMLU
Majority (%)	67.8	74.9	80.1	58.7
Meta-judge (%)	**70.1**	**76.5**	**81.7**	61.9
p	$<.001$	.019	.001	.424 (n.s.)

Table 8: Decision-rule ablation within Inquiry, frontier panel, free-text argument exchange. Exact paired McNemar test, same 100 questions per dataset as Table 3.

	MedQA	TruthfulQA	CSQA	MMLU
Majority (%)	95.0	99.0	90.0	98.0
Meta-judge (%)	96.0	97.0	90.0	98.0
p	1.000 (n.s.)	0.500 (n.s.)	1.000 (n.s.)	1.000 (n.s.)

answer and the meta-judge's answer are already recorded for every question, letting us isolate the decision rule's own effect directly from these same saved transcripts, holding panel and argument structure (free-text) fixed. Tables 7 and 8 report this same-panel, same-argument-structure ablation for the SLM and frontier panels respectively.

On the SLM panel, escalating non-unanimous rounds to a meta-judge significantly improves accuracy over plain majority vote on three of four datasets: MedQA (67.8% vs. 70.1%, $p < .001$), TruthfulQA (74.9% vs. 76.5%, $p = .019$), and CommonsenseQA (80.1% vs. 81.7%, $p = .001$); MMLU is the exception (58.7% vs. 61.9%, $p = .424$, n.s.), where the non-unanimous subset is smallest (48 of 126 questions). The meta-judge is therefore not a neutral bookkeeping choice at this panel scale: reading the transcript and arbitrating recovers real accuracy that plain majority vote leaves on the table whenever the panel is split. On the frontier panel, by contrast, majority vote and meta-judge are statistically indistinguishable on all four datasets, consistent with the frontier panel's small number of non-unanimous rounds (2–6 out of 100 per dataset) leaving little room for either rule to differ.

A.6 Short-circuit ablation

Table 9 isolates the round-1-unanimous subset (questions where all three judges agree in Round 1) and compares that Round-1 consensus accuracy to the accuracy after every question is forced through all 3 rounds (SoM/Conformist already run full rounds by default, no short-circuit mechanism exists for them; Inquiry re-run with `--no-short-circuit`). Conformist never erodes (C→W = 0 on every dataset): its freezing rule mechanically guarantees that an already-unanimous verdict cannot be talked out of by further rounds. SoM shows real, sometimes substantial erosion: 10 flips on MedQA, 5 on TruthfulQA, and a striking 49 correct-to-wrong flips on CommonsenseQA (out of 748 unanimous cases, a 6.6% erosion rate), continued unconstrained debate under early verdict commitment can talk an already-correct majority out of the right answer. Inquiry's net effect from continued deliberation (free-text) is small but mostly the same direction as SoM's: a net loss of 3 questions on MedQA (5 C→W vs. 2 W→C) and 6 on CommonsenseQA (7 C→W vs. 1 W→C), exactly neutral on MMLU (1 C→W vs. 1 W→C), and the sole net gain on TruthfulQA (0 C→W vs. 1 W→C) – i.e., empirically, continued deliberation under Inquiry's private-reasoning protocol causes markedly less erosion than SoM's, which is both more consistent in direction and far larger in magnitude where it occurs (up to 49 flips on CommonsenseQA vs. Inquiry's 6). This is the basis for reporting all main-comparison tables (Tables 2 and 3) under the no-short-circuit condition uniformly: it isolates the pure effect of argument structure and protocol design from the confound of when the panel is allowed to stop deliberating.

A.6.1 Short-circuiting in practice: accuracy vs. deliberation cost

Within Inquiry, if all three judges' private Round-1 claims already agree, there is no disagreement left to resolve, and the run can stop after Round 1 rather than continuing through Rounds 2–3, a pure cost optimization that changes nothing about the decision logic itself. This raises a deployment-relevant question distinct from the erosion result above: not "does continuing to deliberate erode an already-unanimous verdict" (answered above, on the unanimous subset only), but "if short-circuiting is switched on, does overall dataset accuracy change, and how much deliberation does it save." We answer this by reconstructing the short-circuit-ON condition directly from the existing no-short-circuit data, rather than by comparing against a separately-executed short-circuit-enabled

Table 9: Round-1-unanimous subset (see text): Round-1 accuracy vs. final accuracy after every question is forced through all 3 rounds. Final accuracy uses the majority-vote decision rule throughout. C→W / W→C count individual questions whose answer flips from correct to wrong / wrong to correct between Round 1 and the final round. Small-model (SLM) panel, free-text argument exchange for SoM/Conformist.

Protocol	n uni. (%)	R1 acc	Final acc	C→W	W→C
MedQA					
SoM	582 (45.7)	75.95	74.23	10	0
Conformist	574 (45.1)	77.35	77.35	0	0
Inquiry (free-text)	596 (46.8)	85.07	84.56	5	2
Inquiry (Toulmin)	534 (42.0)	80.52	80.34	2	1
TruthfulQA					
SoM	281 (41.1)	80.43	79.36	5	2
Conformist	281 (41.1)	80.07	80.07	0	0
Inquiry (free-text)	345 (50.4)	86.38	86.67	0	1
Inquiry (Toulmin)	293 (42.8)	82.94	83.62	2	4
CommonsenseQA					
SoM	748 (61.3)	89.71	83.16	49	0
Conformist	755 (61.8)	89.27	89.27	0	0
Inquiry (free-text)	874 (71.6)	92.11	91.42	7	1
Inquiry (Toulmin)	816 (66.8)	93.63	93.50	2	1
MMLU-FL					
SoM	43 (34.1)	74.42	74.42	0	0
Conformist	43 (34.1)	74.42	74.42	0	0
Inquiry (free-text)	42 (33.3)	78.57	78.57	1	1
Inquiry (Toulmin)	60 (47.6)	53.33	51.67	1	0

run: round1-unanimous questions never diverge from their Round-1 answer whether or not later rounds run, so Table 9's own C→W/W→C counts convert directly into an overall accuracy delta, at full dataset scale, with no additional judge calls. This avoids conflating the effect of short-circuiting with the sampling noise that two independently-executed API runs would introduce (an earlier, discarded pass comparing genuinely separate short-circuit-ON and no-short-circuit runs produced inconsistent, noisier deltas for exactly this reason).

Table 10 shows the result: overall accuracy changes by at most half a percentage point in either direction (–0.15pp to +0.49pp across all four datasets, free-text), none significant (largest effect, CommonsenseQA, $p = .070$), while short-circuiting skips 22–48% of rounds depending on how often the panel already agrees at Round 1. Measuring directly from each question's recorded per-call token/latency stats, rather than assuming savings scale one-to-one with the round reduction (judges are called concurrently within a round, so wall-clock time is not simply proportional to token count), total judge-API tokens fall by 25.9–59.9% and wall-clock latency by 15.3–47.7%. Token savings exceed the round reduction on every dataset (later rounds carry more accumulated context, so skipping them saves more than an equal share of rounds would suggest), while latency savings track the round reduction more closely, since wall-clock time per round is dominated by generation length rather than prompt size. Toulmin shows the same accuracy pattern at a similar or slightly smaller magnitude (–0.29pp to +0.79pp, all non-significant; not tabulated here for space). The largest savings occur on the dataset where judges agree most often at Round 1, CommonsenseQA (71.6% short-circuited, 59.9% tokens / 47.7% latency saved) – which is also the dataset where Inquiry's absolute accuracy gain over Independent is smallest, since a strong individual judge already leaves little for deliberation to correct. With these two considerations together, negligible, directionally inconsistent accuracy cost against a substantial, measured reduction in compute and latency, short-circuiting is a favorable default for deployment: it recovers a real fraction of Inquiry's compute cost with no detectable accuracy cost at this sample size. We note the discordant-pair counts underlying the significance tests are small (1–8 flips across datasets), so "not significant" here should be read as "no detectable difference at this sample size," not as proof of exact equivalence; the inconsistent sign of the delta across datasets, however, is itself evidence against a systematic accuracy cost in either direction.

Table 10: Inquiry, free-text, small-model panel: accuracy under full deliberation ("full") vs. short-circuiting switched on ("SC-ON," reconstructed as described in the text). Δ is SC-ON minus full, in percentage points; p is the exact paired McNemar test (same discordant pairs as Table 9). % SC is the fraction of questions resolved by short-circuiting at Round 1. Avg rounds is the mean number of debate rounds run per question under SC-ON. Tok./Lat. saved (%) is the reduction in total judge-API tokens and wall-clock latency under SC-ON relative to full deliberation.

Dataset	Acc(%) full	Acc(%) SC-ON	Δ	p	% SC	Avg rounds	Tok./Lat. saved (%)
MedQA	67.79	68.03	+0.24	.453	46.8	2.06	38.3 / 27.9
TruthfulQA	74.85	74.71	–0.15	1.000	50.4	1.99	41.1 / 27.1
CommonsenseQA	80.10	80.59	+0.49	.070	71.6	1.57	59.9 / 47.7
MMLU-FL	58.73	58.73	0.00	1.000	33.3	2.33	25.9 / 15.3

A.7 Individual-judge transition dynamics: full three-judge breakdown

Figure 4 reports the same Round-1-to-final transition decomposition as Figure 3 in the main text, for all SLM judges (Qwen2.5-7B, Llama-3.1-8B, Gemma-3n-E4B) rather than Qwen2.5-7B alone.

Llama-3.1-8B shows the largest wrong-to-correct recovery effect of any judge: on TruthfulQA, Llama recovers from an incorrect Round-1 claim to a correct final answer 49% of the time under Inquiry, versus only 10% under SoM and 13% under Conformist, a roughly four-fold improvement and the single largest per-judge recovery effect observed in this study. Gemma-3n-E4B shows the same directional effect (Inquiry $\geq$ both baselines) on every dataset, with margins ranging from 1.1-fold (CommonsenseQA) up to 1.8-fold (TruthfulQA) over the stronger baseline.

Inquiry's protection of already-correct judges (low C→W), by contrast, is not judge-general: Llama shows a partial improvement over SoM (lower C→W on three of four datasets, higher on MMLU-FL), while Gemma is slightly worse on three of four datasets (Inquiry C→W 1–4pp higher than SoM on MedQA, TruthfulQA, and MMLU-FL, improving only on CommonsenseQA). Gemma's transition rates are nonetheless the most stable across protocols of the three judges: its C→C spans only 83–98% over all 12 protocol×dataset cells, versus 67–92% for Qwen and 78–97% for Llama, and it never shows a swing as large as Llama's Inquiry-specific jump in W→C – i.e., Gemma's answers appear comparatively insensitive to which debate protocol it is placed in, while Qwen and especially Llama shift substantially once Inquiry replaces SoM/Conformist.

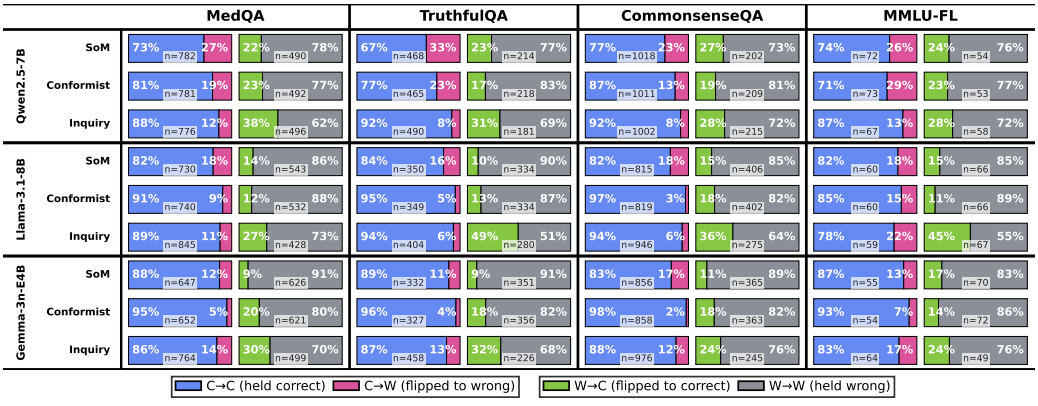

Figure 4: Individual-judge Round-1→final-answer transitions, all three small-model judges, free-text argument exchange (Inquiry uses full deliberation, consistent with Table 2 and Figure 2). For each judge/protocol/dataset cell, the left bar conditions on a correct Round-1 claim (C→C held correct vs. C→W flipped to wrong) and the right bar conditions on a wrong Round-1 claim (W→C flipped to correct vs. W→W held wrong); n is the conditioning population size for that bar. Round-1 uncertain/unparsed claims are excluded from both populations.