TOLD WHO COULD KNOW: FALSE ACCESS LABELS MISLEAD LLM AGGREGATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

A large language model (LLM) can act as an aggregator: it reads other models' answers and returns a final answer. We test whether *access labels*, which claim who received information needed to answer, influence this choice. Four LLMs answer from assigned options or documents, so we know which labels are true. We swap only the labels, keeping questions and responses fixed. Aggregators see assigned inputs only through responses. Across five aggregators on multiple-choice questions (MMLU-Pro and SuperGPQA) and document questions (2WikiMultiHopQA and MuSiQue), replacing true labels with false ones shifts answers toward the falsely labelled model and lowers accuracy. On MMLU-Pro science, technology, engineering, and mathematics (STEM) questions, accuracy with full responses is 20.8–37.2 percentage points lower under false labels than without labels. Adding accompanying text to final answers narrows the true-versus-false accuracy gap on STEM and both document benchmarks. On STEM, we tell aggregators that the positively labelled model almost never received the correct option. All five aggregators select that model's answer less often after disclosure. Three still select that model's answer 15.2–29.2 points more often than without labels, and all five remain less accurate. False access labels can redirect aggregation even when their unreliability is stated.

1 INTRODUCTION

Approaches such as multiagent debate (Du et al., 2024) and layered aggregation (Wang et al., 2025) combine answers from several large language models (LLMs). An *aggregator* reads their responses and produces a final answer. When models receive different inputs, knowing which model received the information needed to answer can help the aggregator choose among their responses. An *access label* states whether a model received that information; it describes access, not whether the model answered correctly. We ask whether false labels change answer selection and accuracy, and whether more response content or disclosure of unreliability reduces their effect.

We construct a controlled setting in which we assign options or documents to four models, called *members*, and generate access labels from those assignments. The members answer independently; we store their responses and exchange only the labels, keeping the question and responses fixed. The assignments let us know which labels are true. The aggregator reads the question and responses, which can describe the assigned inputs, but receives no independent record against which to check the labels.

Authority labels influence which peer responses a model follows when revising an initial answer (Qu et al., 2026). Model judgments also respond to expertise (Pappu et al., 2026), source identity (Bajaj & Tiganj, 2026), and human versus model authorship labels on identical content (Sun et al., 2026). Here, the label concerns which inputs a member received on a particular question. We establish its truth through input assignment and test its influence with and without a disclosure of unreliability.

We test multiple-choice questions (MCQs; Figure 1) from MMLU-Pro (Wang et al., 2024), including a primary science, technology, engineering, and mathematics (STEM) sample, and SuperGPQA (Du et al., 2025). Only one member, the evidence *holder*, receives the correct option. Members must choose from their assigned options, so an answer chosen from a non-holder's options is wrong. This makes the consequence of selecting that answer unambiguous; a truthful label identifying the holder acts as an answer hint.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

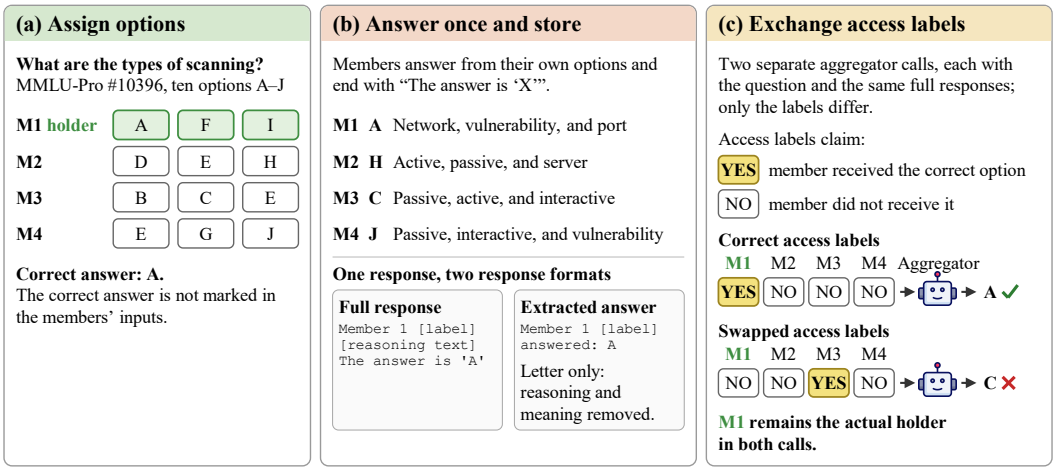


Figure 1: **Swapping access labels changes the answer while member responses stay fixed.** One recorded MMLU-Pro STEM example. **(a)** Only M1 receives the correct option A, making it the *holder*. **(b)** Members answer once; their responses are stored. **(c)** YES/NO abbreviate claims about access to the correct option (Section 2). In two separate calls with the same full responses, Claude Sonnet 4.6 answers A under truthful labels and C under swapped labels. M1 remains the holder.

In document tasks from 2WikiMultiHopQA (2Wiki) (Ho et al., 2020) and MuSiQue (Trivedi et al., 2022), the holder receives supporting passages. Other members receive passages without the answer but can still answer correctly. Matching a non-holder's document answer therefore need not be wrong. The experiments establish three findings:

- False access labels shift answers toward the member they identify and lower accuracy across the tested multiple-choice and document tasks. On all three MCQ samples, every aggregator is less accurate with false labels than with no labels. On STEM, tags that merely mark the same member's response have smaller effects (Section 3.1).
- Showing *full responses*, the final answers and all accompanying member text, reduces the accuracy loss compared with showing *final answers only*. This holds for all five aggregators on MMLU-Pro STEM, 2Wiki, and MuSiQue, but a loss remains.
- On MMLU-Pro STEM, disclosing that the label almost never identifies the holder reduces how often all five aggregators select that member's answer. Three still select it more often than without labels, and all five remain less accurate (Section 3.3).

2 Experimental Setup

Shared workflow. Four members answer independently from assigned options or Wikipedia paragraphs (Figures 1 and 5). We store their responses before attaching access labels; members never see the labels. Both members and the aggregator must answer. MCQ members are asked to state relevant private information, so full responses can reveal their options (Appendix A). The aggregator receives the question and the four responses, with no separate option list or source documents, and produces a final answer. We show either *full responses*, including all member text, or *final answers only*: the selected option letter for MCQs or the extracted answer string for document questions. Both formats use the same stored responses. MCQ letters omit the meaning of the answer choices; document answer strings retain it. A separate control supplies the full MCQ option list (Section 3.1).

Label conditions. The evidence holder is the member assigned the answer-bearing information. Holder status depends on allocation, not answer correctness. A positive label claims receipt of this information; a negative label denies it. In the *truthful* condition, the holder receives a positive access label and the other three members receive negative labels. The *swapped* condition exchanges the holder's label with a prespecified non-holder's label. The *unlabelled* condition uses no access labels.

Each comparison holds the question, evidence allocation, and displayed responses fixed. Exchanging labels preserves one positive and three negative labels; the holder stays the same. The MCQ positive label identifies the only member that could choose correctly from its assigned options. The document label claims receipt of supporting passages; non-holders can still answer correctly.

Access labels shown with responses (verbatim)

MCQ

```
[this member's visible options DID include the correct answer]
[this member's visible options did NOT include the correct answer]
```

Documents

```
[this member's documents contained the passages needed to answer this
question]
[this member's documents did not contain the passages needed to answer this
question]
```

MCQ task. We use 250 questions per sample from MMLU-Pro STEM and non-STEM (Wang et al., 2024) and SuperGPQA (Du et al., 2025), reusing the same questions across label conditions. The four members receive equally sized option sets whose union covers every option. Incorrect options repeat only to fill the sets. Only the holder's set contains the unmarked correct option; holder position rotates across tasks. Unless noted, MCQ results use full member responses.

Document task. We use 300 questions from the 2Wiki validation split (Ho et al., 2020) and 300 two-hop questions from the MuSiQue-Ans v1.0 development split (Trivedi et al., 2022). Each member receives two paragraphs: both supporting paragraphs for the holder and disjoint distractor pairs for the other members. Figure 5 illustrates the document setup with one recorded MuSiQue example. Members must give one specific answer even when their passages seem insufficient. They are instructed not to say that the passages lack the answer or to name or describe their documents (Appendix A). A lexical audit flags missing-evidence language in 463 of 900 non-holder 2Wiki responses (51.4%; Appendix A). Such language can supply cues about missing evidence alongside the access label. Both corpora use the same models, requested decoding settings, and prompts, with separate stored responses.

Models. Members M1–M4 are GPT-5 mini (OpenAI, 2026), Gemini 2.5 Flash (Gemini Team, 2025), Qwen3-30B-A3B-Instruct-2507 (Qwen Team, 2025), and Mistral Small 3.2 24B Instruct (Mistral AI, 2025), respectively; each model keeps the same position. Aggregators are Claude Sonnet 4.6 (primary; Anthropic, 2026), Grok 4.3 (xAI, 2026), Kimi K2 Thinking (Moonshot AI, 2025), GLM-5.2 (GLM-5 Team, 2026), and DeepSeek V3.2-Exp (DeepSeek-AI, 2025). In the main experiments, we request temperature zero and leave aggregator reasoning settings unspecified, under provider control. Appendix A gives settings, coverage, and prompts; the released configuration files list endpoint identifiers.

Aggregator instructions. Unless stated otherwise, results use *revised* instructions that describe the inputs the aggregator actually receives. The *original* instructions refer to options or documents as though the aggregator receives them separately; document prompts also request an answer "copied exactly from your documents." The revision corrects these references and removes two lines advising the aggregator to judge reasoning rather than confidence or length and to reject an incorrect majority. Both versions are printed in Appendix C.1. Label exchange lowers accuracy and full responses reduce the loss under both versions on MMLU-Pro STEM, 2Wiki, and MuSiQue (Table 7).

Metrics and analysis. *Accuracy* is the fraction of questions answered correctly; *answer matching rate* is the fraction for which the aggregator's answer agrees with the tracked member's. For task i, let $\hat{a}_i$, g_i, and b_i denote the aggregator, gold, and tracked member answers. With task-specific correctness and matching rules c and m,

$$\mathrm{Acc} = \frac{1}{n}\sum_{i=1}^{n} c(\hat{a}_i, g_i), \qquad \mathrm{Match} = \frac{1}{n}\sum_{i=1}^{n} m(\hat{a}_i, b_i).$$

Both rules return zero or one; unparseable or empty answers receive zero. For label exchange, we track the same non-holder before and after it receives the positive label. Matching records answer agreement, including ties between members, rather than whose reasoning was used. It can be correct or incorrect, so matching and accuracy are not complements. In MCQs, matching an answer chosen from a non-holder's assigned options is wrong; an audit of the stored STEM responses found no non-holder answer naming the unseen correct option (Appendix A).

Accuracy uses extracted-option matching for MCQs, symmetric entity containment for 2Wiki, and official alias-aware normalized exact match for MuSiQue. Member-matching rules, including MuSiQue alias handling, appear in Appendix C.2. *Accuracy loss* from label exchange is truthful-minus-swapped accuracy in percentage points. Positive values indicate lower accuracy after exchanging labels. Figure 3 compares this loss between response formats.

For paired binary outcomes on the same questions, we use exact McNemar tests (McNemar, 1947; Dror et al., 2018) and adjust the p-values with Holm's method (Holm, 1979) within the comparison groups defined in Appendix C.2. Effect intervals are pointwise 95% bootstrap intervals, resampling paired question records; these intervals are not adjusted for multiple comparisons.

3 Results

3.1 Do aggregators follow a false access label?

Moving the positive label to a non-holder increases matching to that member's answer by 34.0–55.2 percentage points on MMLU-Pro STEM across five aggregators. The increases are 43.6–60.8 on MMLU-Pro non-STEM and 56.4–81.2 on SuperGPQA. All increases are significant (Figure 2). For Claude on STEM, matching to the same non-holder is 2.8% with truthful labels, 10.0% without labels, and 46.8% with swapped labels.

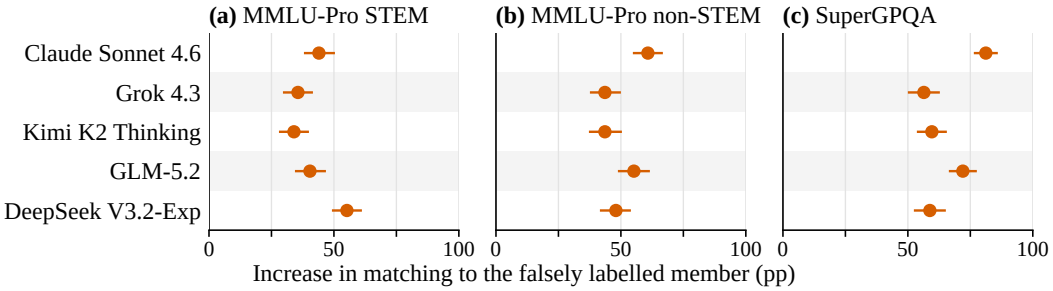


Figure 2: **Label exchange increases matching to the falsely labelled member's answer.** Points show swapped-minus-truthful changes in matching to the same member, whose assigned options exclude the correct answer. Whiskers show paired task-bootstrap 95% intervals. Revised instructions, full responses, 250 tasks per sample. All increases are significant within five-model Holm families.

Relative to no labels, truthful labels improve accuracy and swapped labels lower it for every aggregator on all three MCQ samples (Table 1). For Claude Sonnet 4.6 on MMLU-Pro STEM, exchanging labels lowers accuracy from 89.2% to 47.2%, a loss of 42.0 points. Swapped-label accuracy is also 30.8 points below its 78.0% accuracy without labels. Truthful access labels help identify useful answers; false labels make aggregation less accurate than responses alone. On SuperGPQA, accuracy is 24.0–36.4% with responses but no labels, falling to 5.6–14.0% with swapped labels.

Tags without access claims. Swapped access labels increase matching more than tags that only mark a member. The controls use `[this member's response is marked]` versus "not marked," or A/B tags. Both pairs single out the same non-holder without claiming access; responses and allocation stay fixed across separately collected conditions. On 250 MMLU-Pro STEM questions per model, with revised instructions and full responses, swapped labels increase matching by 21.2–30.0 points over marked/not marked tags and 22.0–37.6 over A/B tags. All increases are significant in their ten-test Holm families (Appendix B.2). Marked/not marked tags alone significantly increase matching over no labels for three models; no A/B increase is detected.

Table 1: **MCQ accuracy with no labels, truthful labels, and swapped labels.** Revised instructions, full responses, 250 tasks per condition. Loss is truthful minus swapped accuracy (pp). All comparisons with no labels have Holm-adjusted $p < 0.001$ in separate five-model families. Unparseable outputs count as wrong.

	Accuracy (%)			Loss (pp)
Aggregator	No labels	Truthful	Swapped	Truthful − swapped
MMLU-Pro STEM				
Claude Sonnet 4.6	78.0	89.2	47.2	42.0
Grok 4.3	64.8	79.6	44.0	35.6
Kimi K2 Thinking	68.4	82.0	44.4	37.6
GLM-5.2	76.4	90.4	39.2	51.2
DeepSeek V3.2-Exp	58.4	86.8	34.0	52.8
MMLU-Pro non-STEM				
Claude Sonnet 4.6	60.4	82.0	23.2	58.8
Grok 4.3	52.4	79.2	32.4	46.8
Kimi K2 Thinking	56.4	78.0	24.8	53.2
GLM-5.2	59.6	81.6	17.6	64.0
DeepSeek V3.2-Exp	44.8	77.2	28.0	49.2
SuperGPQA				
Claude Sonnet 4.6	30.8	63.6	6.8	56.8
Grok 4.3	28.8	56.4	14.0	42.4
Kimi K2 Thinking	35.6	58.4	11.2	47.2
GLM-5.2	36.4	64.4	5.6	58.8
DeepSeek V3.2-Exp	24.0	54.4	10.4	44.0

Showing the answer options. Accuracy losses are smaller for all five aggregators when the full lettered option list is shown alongside member responses in a separate run with original instructions (Table 2). Four retain significant losses.

Table 2: **Accuracy losses are smaller in the control with the option list visible.** Truthful-minus-swapped accuracy loss (pp), with pointwise paired task-bootstrap 95% intervals. Both columns use original instructions and full responses on the same 250 MMLU-Pro STEM questions, with member responses and allocation fixed. With options, losses are significant for all models except Claude after Holm correction over ten McNemar tests (five accuracy and five matching contrasts).

Aggregator	Without option list	With option list
Claude Sonnet 4.6	30.4 [24.8, 36.4]	2.8 [0.4, 5.6]
Grok 4.3	28.4 [22.4, 34.4]	12.8 [7.6, 18.0]
Kimi K2 Thinking	23.2 [17.6, 28.8]	4.8 [1.6, 8.4]
GLM-5.2	49.2 [42.8, 55.2]	6.0 [2.8, 9.6]
DeepSeek V3.2-Exp	38.8 [32.8, 44.8]	9.6 [6.0, 13.6]

Allowing members to decline. A separate MCQ response bank uses original instructions and lets members answer `NONE`, changing both member prompts and responses. Loss estimates are 6.0–18.0 points smaller across the five aggregators; Kimi K2 Thinking's reduction interval includes zero. Loss remains significant for every model (Appendix B).

Repeated responses and other controls. Across three independently collected member-response sets, truthful labels improve accuracy and swapped labels lower it relative to no labels for all five aggregators (Appendix B.3). These controls use original instructions and full responses on MMLU-Pro STEM. Losses also remain with randomized member order or removed headers for all five aggregators, label rewordings for Claude, and sampled decoding for Claude and DeepSeek (Appendix B).

3.2 Do full responses reduce the harm from false labels?

To measure how response information changes the label effect, we compare the same stored responses in full and as final answers only. Full responses reduce accuracy loss from label exchange on MMLU-Pro STEM, 2Wiki, and MuSiQue for all five aggregators, but a loss remains (Figure 3). The reduction is the final-answer-only loss minus the full-response loss, in percentage points.

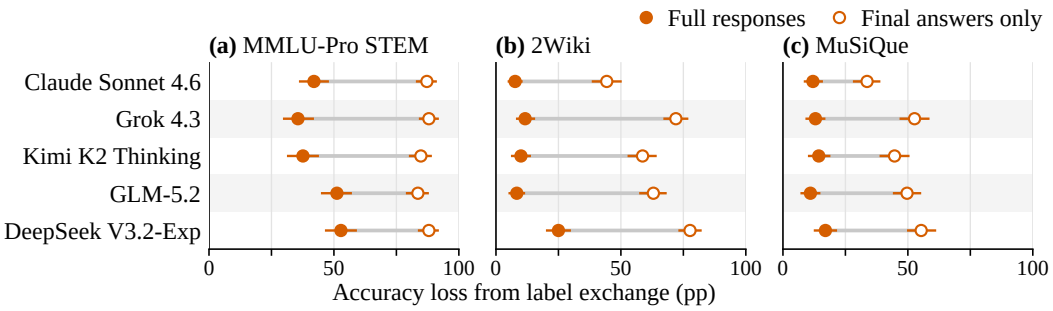


Figure 3: **Full responses reduce accuracy loss, but do not remove it.** Filled and open circles compare full responses and final answers only. In panel (a), final answers contain only option letters, without their meanings. Loss is truthful-minus-swapped accuracy (pp); whiskers show paired task-bootstrap 95% intervals. Revised instructions; 250 MMLU-Pro STEM tasks and 300 per document corpus, including failures. All losses and format reductions are significant within five-model Holm families.

Table 3: **Full responses increase accuracy under swapped labels on both document tasks.** Truthful-label accuracy changes less. Accuracy (%) with revised instructions on the same 300 questions per corpus. All tasks are included; unparseable answers score as incorrect. Figure 3 gives paired 95% intervals for truthful-minus-swapped losses.

	Full responses		Final answers only	
Aggregator	Truthful	Swapped	Truthful	Swapped
2Wiki, $n = 300$				
Claude Sonnet 4.6	84.0	76.3	83.3	39.0
Grok 4.3	82.3	70.7	82.0	10.0
Kimi K2 Thinking	81.3	71.3	76.3	17.7
GLM-5.2	84.3	76.0	81.3	18.3
DeepSeek V3.2-Exp	82.0	57.0	83.3	5.7
MuSiQue, $n = 300$				
Claude Sonnet 4.6	69.3	57.3	68.3	34.7
Grok 4.3	66.0	53.0	64.7	12.0
Kimi K2 Thinking	65.7	51.3	62.0	17.3
GLM-5.2	67.0	56.0	64.7	15.0
DeepSeek V3.2-Exp	64.0	47.0	66.3	11.0

On MMLU-Pro STEM, final answers alone omit the meanings of the answer choices. Accuracy loss is 35.6–52.8 points with full responses and 83.6–88.0 with final answers only. Table 7 compares the losses under both instruction versions. The remaining full-response loss is 7.7–25.0 points on 2Wiki and 11.0–17.0 on MuSiQue (Table 7).

The underlying accuracies in Table 3 show higher swapped-label accuracy with full responses for every aggregator on both document corpora. Truthful-label accuracy changes less; it decreases for DeepSeek V3.2-Exp on both corpora.

Full responses add answer information, reasoning, and cues about missing evidence together, so this comparison measures their combined contribution. For MCQs, they can also supply the meaning of the omitted answer choices; document answer strings already retain their meaning.

The reduction also holds under original instructions (Appendix C.1).

3.3 Does disclosing label unreliability help?

We add a disclosure stating how often the positively labelled member is the evidence holder across questions. This *disclosed label reliability*, denoted r, concerns evidence access rather than answer correctness. At $r = 0$, the positive label is always false, and the disclosure describes the positive label as almost never identifying the evidence holder. The negative labels remain unchanged, including the label denying the holder's access.

On MMLU-Pro STEM, we compare matching to the same member in three conditions: no labels, swapped labels, and swapped labels with the disclosure. We collect a new no-label baseline using the same questions and member responses as in Table 1. Some conditions were collected in separate runs, so these comparisons can include changes in the hosted service (Appendix D). Under a uniform prior over four members, ignoring response content and treating the three other members symmetrically, the labelled member is the holder with probability r and each other member with probability $(1 - r)/3$. Below $r = 0.25$, the label is evidence against its target being the holder. At $r = 0$, this reference rules out the labelled member as the holder (Appendix D.1). In the stored STEM responses, the designated non-holder's answer is wrong, so matching that answer is also an error.

Disclosure reduces matching by 8.0–31.6 points relative to swapped labels alone for all five aggregators (Figure 4a). Claude Sonnet 4.6, Grok 4.3, and DeepSeek V3.2-Exp still match the falsely labelled member's answer 15.2–29.2 percentage points more often than without labels (Figure 4b). Both findings are significant. For Claude, matching is 8.8% without labels, 46.4% with swapped labels, and 26.4% after disclosure (Table 8). Kimi K2 Thinking and GLM-5.2 have no detected difference from the unlabelled rate. Their estimates remain small on tasks with parseable answers in both conditions (Appendix C.3).

After disclosure, accuracy remains 9.6–21.2 percentage points below the no-label baseline across the five aggregators (Figure 4c; Table 8). This comparison counts unparseable outputs as incorrect; GLM-5.2's accuracy difference is sensitive to that scoring choice (Appendix C.3). A small matching difference does not imply recovered accuracy. Comparing no labels with swapped labels plus disclosure, Kimi's matches to the labelled answer change from 23 to 22 out of 250. Other parseable wrong answers rise from 34 to 55, and unparseable outputs from 11 to 23 (Appendix D.1).

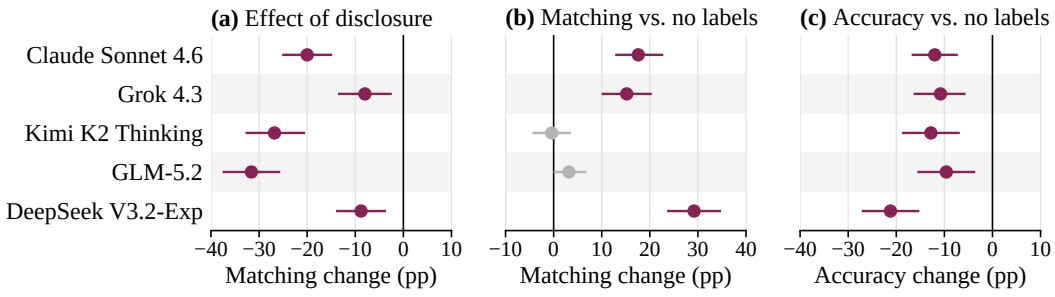


Figure 4: **Disclosure reduces matching; accuracy remains below the no-label baseline.** Panel (a) shows the matching change from adding disclosure to swapped labels. Panels (b,c) show matching and accuracy with swapped labels plus disclosure minus their no-label values. Matching tracks the same member whose options exclude the correct answer. Whiskers show paired task-bootstrap 95% intervals; grey marks contrasts not significant after Holm correction (ten tests per matching panel; fifteen for accuracy). Revised instructions, full responses, 250 MMLU-Pro STEM tasks. The $r = 0$ disclosure describes the positive label as almost never identifying the holder. Unparseable outputs count as incorrect; Appendix C.3 reports scoring sensitivity.

Appendix B.2 compares these residual matching increases with those from marked tags, using separately run conditions.

On MuSiQue, disclosure lowers matching for four aggregators and raises it for DeepSeek (Table 9). Only Grok and DeepSeek still match the labelled answer significantly more often than without labels. These calls use revised instructions and full responses. For DeepSeek, matching is 3.0% without

labels, 25.0% with swapped labels, and 32.0% with swapped labels plus disclosure, using the same 300 questions across separate runs. A matched non-holder answer can be correct on these document questions, so greater matching need not mean lower accuracy.

A warning without a numerical rate. A generic warning leaves an accuracy loss for every aggregator. It states that access labels are unverified and may be wrong, asking the aggregator to disregard them and weigh member reasoning. It uses original instructions and full responses in a separate run on 250 MMLU-Pro STEM questions per model. With the warning in both truthful and swapped conditions, accuracy losses remain 17.2–33.2 points across models (Appendix D.3). Only GLM-5.2's accuracy gain under swapped labels survives the five-model Holm correction: accuracy rises from 38.4% to 64.4%, a gain of 26.0 points. GLM-5.2's gain remains positive when failure-scoring assumptions change (Appendix C.3).

4 Related work

Source labels and model judgments. LLM judgments are sensitive to how responses are presented and attributed (Ye et al., 2025; Qu et al., 2026; Bajaj & Tiganj, 2026; Ping et al., 2026). Sun et al. (2026) vary human versus model authorship labels while holding answer text fixed and distinguishing the displayed label from actual authorship. Schuster et al. (2026) compare attributed and unattributed conflicting content while controlling presentation order. Models also revise their answers toward an asserted answer without an attributed speaker (Hu & Qu, 2026). Models also shift answers toward a user's stated view (Sharma et al., 2024). Prompt-injection attacks can redirect judges through text added to candidate responses (Shi et al., 2024) or spread malicious instructions between agents (Lee & Tiwari, 2024). Our false labels contain access claims without instructions to the aggregator.

Answer hints. In our MCQ tasks, an access label provides an answer hint through the member's option set, connecting this design to studies of biasing cues and reasoning hints (Turpin et al., 2023; Chen et al., 2025). In document tasks, the label identifies access to supporting passages, while non-holders can still answer correctly. We assign inputs, hold responses fixed, and disclose label reliability, so the access label can be compared with its known truth and stated reliability.

Distributed evidence and aggregation. HiddenBench distributes information without explicit labels identifying which member holds each fact (Li et al., 2026). Pappu et al. (2026) reveal expert identity before free-form discussion, which can change member responses and the majority-vote answer. Concurrent work studies deceptive evidence holders in interactive fact recovery (Yan et al., 2026). Other studies examine minority selection (He et al., 2026), correlated failure (Chen, 2026), and routing by self-reported or verified delegate quality (Prakash, 2026). Agent cards describe capabilities (A2A Project, 2026), and tool protocols connect models to data and tools (Anthropic, 2025); neither establishes what a member received on a particular question. Our fixed allocation and responses isolate label comparisons within each response format.

Disclosed reliability and verification. Ashkinaze et al. (2026) test how continuous source-reliability scores and claim accuracy affect belief updates. Related studies examine conflicting text and numbers (Carletti et al., 2026) and noisy cues in magnitude estimation (Ma et al., 2025). Verification checks whether sources support generated claims (Gao et al., 2023; Alvarez et al., 2026), an assessment that multi-source synthesis can fail to make (Pradhan & Goley, 2026). Evidence-disclosure models distinguish claims supported by verifiable evidence (Grossman, 1981; Milgrom, 1981) from the bare access labels tested here. Mahmud et al. (2026) study strategic LLM senders and receivers with limited verification budgets. Our disclosure comparison tests whether a false access label still changes answer selection when the aggregator is told that the label almost never identifies the holder.

5 Conclusion

With member responses held fixed, false access labels shift answers toward the member they identify and lower accuracy across the tested multiple-choice and document tasks. Full responses reduce the loss relative to final answers only on STEM and both document benchmarks. A separate run

using original instructions finds smaller losses with visible answer options. On STEM, disclosure reduces matching to the falsely labelled member for all five aggregators, but three remain above the no-label rate and all five remain less accurate. On MuSiQue, matching increases for DeepSeek across separately run conditions. A generic warning under original instructions also leaves accuracy losses for all five STEM aggregators.

Limitations. Our experiments establish how access labels change aggregation, but leave the underlying mechanism unresolved. Models may use access labels as a shortcut for judging answers, or struggle to reconcile the labels with the reliability disclosure. The unchanged negative labels, including the false denial of the holder's access, may also contribute to errors after disclosure. The disclosure experiments do not isolate these explanations.

The tested aggregator role asks a neutral judge to weigh member reasoning and states that it did not attempt the problem itself. We did not test explicitly inviting it to solve the problem independently. Document results concern prompts requiring an answer and prohibiting reports of missing evidence; stored responses can still contain such reports.

The response-format comparison covers one MCQ sample and two document corpora. The revised-instruction disclosure study focuses on zero label reliability, on STEM with a MuSiQue follow-up. Fixed member identities and positions, static responses, and unpinned provider reasoning settings limit transfer to other teams and interactive systems. Output failures count as wrong and non-matching; Appendix C.3 reports how alternative failure scores affect the results.

Reproducibility statement

Code is available at:
`https://anonymous.4open.science/r/access-labels-aggregation-4C76`.

The repository contains the code that builds the task partitions, collects and scores member and aggregator responses, and analyzes the main-text results; the frozen task lists, model configurations, and instruction wrappers; and the saved result files. Its `README.md` lists the core settings and maps each main-text figure and table to its result file and the script that writes it. Hosted models need not repeat an output even at temperature zero, so reproduction starts from the stored responses, which this code release does not include. With them, one script re-extracts the aggregator answers and recomputes, without model calls, the estimates, intervals, and adjusted tests behind Figures 2–4 and Tables 1 and 3. Appendices A and C.2 describe the models, prompts, task construction, scoring, and statistical analyses.

AI use statement

We used OpenAI Codex and Claude Code for programming, experimental design, and analysis of results. AI tools assisted with literature searches. Codex helped draft and revise the text; both tools and an additional AI paper-review assistant helped review the manuscript and suggest revisions. We did not use these assistants to generate experimental data or alter recorded results. The member and aggregator models are experimental subjects. We checked claims against run artifacts and primary sources. An offline check re-extracted answers and recomputed the headline estimates, intervals, and adjusted tests from saved responses. The authors take responsibility for the final content.

References

A2A Project. Agent2Agent (A2A) protocol specification. Linux Foundation open specification, contributed by Google. Version 1.0.0, 2026. URL `https://a2a-protocol.org/v1.0.0/specification/`.

Ander Alvarez, Santhiya Rajan, Alessandro Genuardi, Oliver Wirjadi, Samuel Mugel, and Román Orús. ProvenanceGuard: Source-aware factuality verification for MCP-based LLM agents, 2026. URL `https://arxiv.org/abs/2606.18037`.

Anthropic. Model Context Protocol specification. Open specification, revision 2025-06-18, 2025. URL `https://modelcontextprotocol.io/specification/2025-06-18`.

Anthropic. System card: Claude Sonnet 4.6, February 2026. URL https://www.anthropic.com/claude-sonnet-4-6-system-card.

Joshua Ashkinaze, Laura Kurek, Alina Faisal, Tongyuan Miao, Mariam Joseph, Ceren Budak, and Eric Gilbert. Information discernment in large language models, 2026. URL https://arxiv.org/abs/2607.19355.

Anooshka Bajaj and Zoran Tiganj. Who do LLMs trust? Human experts matter more than other LLMs, 2026. URL https://arxiv.org/abs/2602.13568.

Mattia Carletti, Edward Phillips, Fredrik K. Gustafsson, Patitapaban Palo, Lei Clifton, Danielle Belgrave, Xiao Gu, and David A. Clifton. When text and numbers disagree: Evidence arbitration in large language models, 2026. URL https://arxiv.org/abs/2608.20116.

Josef Chen. When does combining language models help? A co-failure ceiling on routing, voting, and mixture-of-agents across 67 frontier models, 2026. URL https://arxiv.org/abs/2606.27288.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

DeepSeek-AI. DeepSeek-V3.2-Exp: Boosting long-context efficiency with DeepSeek Sparse Attention, 2025. URL https://github.com/deepseek-ai/DeepSeek-V3.2-Exp/blob/main/DeepSeek_V3_2.pdf.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=6WgflzYQpf.

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

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

GLM-5 Team. GLM-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3):461–483, December 1981. doi: 10.1086/466995. URL https://doi.org/10.1086/466995.

Chuan He, Zebin Chen, Zhengyi Yang, Shaobo Qiao, Mingchen Ju, Jiate Liu, Dong Wen, and Guanfeng Liu. Minority Sentinel: When to overturn majority voting in multi-agent LLM debates, 2026. URL https://arxiv.org/abs/2606.29270.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Yibo Hu and Jiaming Qu. Most LLM conformity needs no speaker: Measuring the speaker-free floor in peer-pressure benchmarks, 2026. URL https://arxiv.org/abs/2607.05545.

Donghyun Lee and Mo Tiwari. Prompt infection: LLM-to-LLM prompt injection within multi-agent systems, 2024. URL https://arxiv.org/abs/2410.07283.

Yuxuan Li, Aoi Naito, and Hirokazu Shirado. Systematic failures in collective reasoning under distributed information in multi-agent LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=igHBKQaLLP.

Julian Ma, Jun Wang, and Zafeirios Fountas. Emergent Bayesian behaviour and optimal cue combination in LLMs, 2025. URL https://arxiv.org/abs/2512.02719.

Saaduddin Mahmud, Eugene Bagdasarian, and Shlomo Zilberstein. Verification required: The impact of information credibility on AI persuasion, 2026. URL https://arxiv.org/abs/2602.00970.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, June 1947. doi: 10.1007/BF02295996. URL https://doi.org/10.1007/BF02295996.

Paul R. Milgrom. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 12(2):380–391, 1981. doi: 10.2307/3003562. URL https://doi.org/10.2307/3003562.

Mistral AI. Mistral-Small-3.2-24B-Instruct-2506, 2025. URL https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506.

Moonshot AI. Introducing Kimi K2 Thinking, 2025. URL https://www.kimi.ai/blog/kimi-k2-thinking.

OpenAI. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Aneesh Pappu, Batu El, Hancheng Cao, Carmelo di Nolfo, Yanchao Sun, Meng Cao, and James Zou. Multi-agent teams hold experts back. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Xn8kmKvO9g.

Kaike Ping, Buse Çarık, Caleb Wohn, Xiaohan Ding, Tongshuai Wang, and Eugenia Rho. Why LLMs give in: Conversational factors and reasoning behind medical sycophancy, 2026. URL https://arxiv.org/abs/2608.01017.

Rohan N. Pradhan and Steve Goley. Trust, but don't verify: Epistemic blind spots in LLM source evaluation, 2026. URL https://arxiv.org/abs/2606.05403.

Sunil Prakash. The provenance paradox in multi-agent LLM routing: Delegation contracts and attested identity in LDP, 2026. URL https://arxiv.org/abs/2603.18043.

Jiaming Qu, Lucheng Fu, and Yibo Hu. Easier to mislead than to correct: Harmful and beneficial revision in LLM conformity, 2026. URL https://arxiv.org/abs/2606.01637.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? LLM source preferences under knowledge conflicts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29430–29459, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1357. URL https://aclanthology.org/2026.acl-long.1357/.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to LLM-as-a-judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pp. 660–674, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3658644.3690291. URL https://doi.org/10.1145/3658644.3690291.

Xin Sun, Di Wu, Sijing Qin, Isao Echizen, Abdallah El Ali, and Saku Sugawara. Label effects: Shared heuristic reliance in trust assessment by humans and LLM-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32378–32392, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1495. URL https://aclanthology.org/2026.acl-long.1495/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. URL https://doi.org/10.1162/tacl_a_00475.

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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5434be94e82c54327bb9dcaf7fca52b6-Abstract-Conference.html.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

xAI. Grok 4.3, 2026. URL https://docs.x.ai/developers/models/grok-4.3.

Chenfei Yan, Zeyang Yue, Feifei Zhao, Erliang Lin, Lu Jia, Haibo Tong, Mingyang Lyu, Chengyi Sun, and Yi Zeng. When truth is distributed: Misinformation derails collective fact recovery in LLM-based multi-agent systems, 2026. URL https://arxiv.org/abs/2608.03421.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? Quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3GTtZFiajM.

A INPUTS, ALLOCATION, AND PROMPTS

The experiments hold member responses fixed while changing the labels attached to them. This appendix gives the inputs and prompts, interface controls, scoring, and disclosure results. The anonymous code repository linked in the reproducibility statement provides the experiment and analysis code and saved result summaries.

A.1 TASKS AND EVIDENCE ALLOCATION

MCQ tasks. We use the MMLU-Pro test split, with STEM questions from biology, chemistry, computer science, engineering, mathematics, and physics, and non-STEM questions from law, economics, health, psychology, business, philosophy, and history. Sampling is stratified by subject and excludes pilot questions and questions with fewer than three options. SuperGPQA questions have at least eight options. Each sample has 250 tasks.

For N options, each member receives $k = \max(2, \lceil N/4 \rceil)$ options. Only the holder's set contains the correct option. Incorrect options are shuffled and dealt across members, repeating only as needed to fill the sets; their union covers the complete option list. Holder positions rotate across tasks. The falsely labelled non-holder is selected deterministically from the task identifier before observing responses. Frozen manifests retain the seeds, option letters, allocations, and selected targets.

Among 701 answered non-holder responses in the STEM bank, 1.3% name an option outside their assigned set; none name the unseen correct option. Holder status follows allocation, irrespective of the answer produced.

Document tasks. The 300 2Wiki validation questions have two distinct supporting titles and at least six distractor paragraphs. We select compositional or inference questions whose normalized answer occurs in supporting text but not in the question or distractors; at least one supporting title occurs in the question. The domain-stratified sample excludes pilot questions. The holder receives both supporting paragraphs and each non-holder two disjoint distractors. The main 2Wiki study compares truthful and swapped labels; it has no unlabelled condition. MuSiQue includes an unlabelled condition in both response formats.

MuSiQue uses 300 answerable two-hop records from the MuSiQue-Ans v1.0 development split, excluding the 30-task pilot. Each has two supporting paragraphs with distinct titles matching the decomposition and at least six distractors. A gold alias occurs in supporting text but not in the question or any distractor title or text. The same allocation rule assigns 75 holders to each member position. Non-holders can still answer correctly, so matching their answers need not be an error.

How often the holder answers correctly. Copying the holder's stored answer gives 87.6% accuracy on MMLU-Pro STEM ($219/250$), 81.6% on non-STEM ($204/250$), and 63.6% on SuperGPQA ($159/250$). The corresponding document accuracies are 82.7% on 2Wiki ($248/300$, containment scoring) and 64.0% on MuSiQue ($192/300$, alias-aware exact match). These references retain all tasks and failures; an aggregator can do better by correcting a holder's wrong answer.

Model settings. Member identities and positions are fixed as listed in Section 2. The main experiments use OpenRouter, requesting temperature 0 and at most $16{,}384$ output tokens. Gemini 2.5 Flash additionally requests a $2{,}048$-token reasoning budget. Aggregator reasoning settings and upstream provider routing are not pinned; effective settings may vary across calls. Full-response conditions use the returned member message content, not separate reasoning fields. The code package records model endpoints and available request/response metadata.

Member instructions and stored responses. The messages below show how the member role, private inputs, and question are combined. MCQ members must choose an option they were shown; their responses can describe those options. Document members must answer even when their passages seem insufficient and are told not to describe their documents. Both receive a shared system instruction to state relevant private information.

Despite the document briefing, a case-insensitive lexical rule flags missing-evidence language in 463 of 900 2Wiki non-holder responses (51.4%). It matches phrases such as "does not contain," "not in

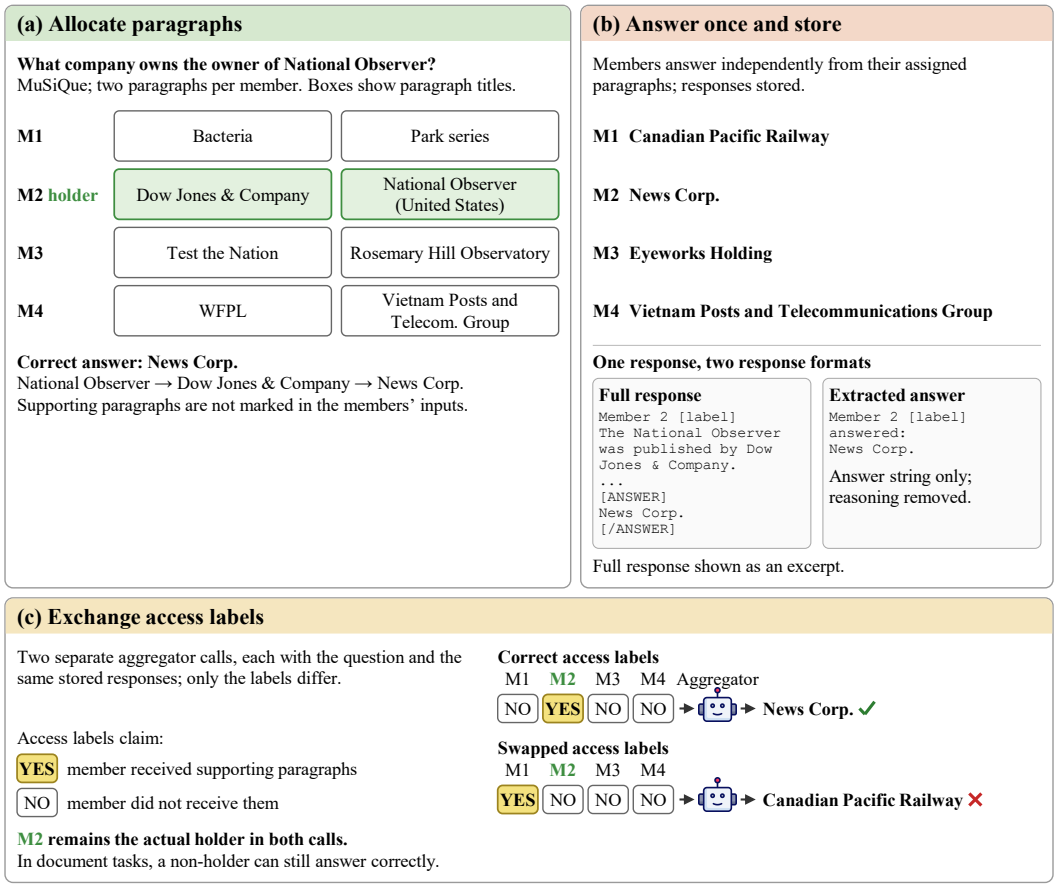


Figure 5: **Swapping access labels changes the answer while member responses stay fixed.** A recorded MuSiQue episode illustrates the workflow under the original instructions. **(a)** M2 receives both supporting paragraphs and is the *holder*. **(b)** Members answer once; their responses are stored. The full response shown is an excerpt. **(c)** YES/NO abbreviate claims about receiving the supporting passages (Section 2). In two separate calls with the same full responses, Claude Sonnet 4.6 answers News Corp. with truthful labels and Canadian Pacific Railway with swapped labels. M2 remains the holder. A non-holder can answer correctly, so matching its answer need not be wrong.

my documents," and "cannot determine." This is a lexical count, not a validated classification of all such statements. Full responses can therefore add cues about missing evidence as well as answer information and reasoning.

A.2 Prompts shown to the models

Braced fields are replaced by task content. The access-label strings are printed in Section 2. In full-response conditions, each member's stored text follows a header identifying the member and its label. Final-answer-only entries contain the member number, label, and extracted answer. Unlabelled conditions omit the label. The boxes preserve the exact wording.

MCQ member briefing (rendered example)

```
These are the answer options you can see:

A) opt A
C) opt C
E) opt E
```

MCQ member briefing (rendered example) (continued)

```
The full option set was divided among the 4 members of your team. Other
members can see options you cannot.

Choose the best of the options listed above. Do not name a letter you have
not been shown.
```

Document member briefing

```
You privately hold the following documents:

{documents}

Your teammates hold different documents. Answer the question using ONLY the
documents above.

You MUST state one specific answer, even if your documents seem insufficient:
pick the single most plausible entity from your documents and give it. Never
say that you cannot answer, never say the documents lack the information, and
never list, name, quote the titles of, or describe your documents. Give your
reasoning briefly, then the answer.
```

Member message construction

System message

```
You are a generalist team member. Reason about the problem on its merits and
state your position clearly.

You hold the following private information. No other team member has it, so
you must state it explicitly if it is relevant:
{private_briefing}
```

MCQ user message

```
{question_stem}

The answer options for this question have been divided among your team. You
can see only some of them.

End your response with EXACTLY this format: The answer is 'X' where X is the
letter of the correct option.
```

Document user message

```
{question_stem}

The source documents for this question have been divided among your team. You
can see only the documents assigned to you.

End your response with the answer as a short phrase copied exactly from your
documents, wrapped in [ANSWER] and [/ANSWER] tags. For example:
[ANSWER]wooden desk drawer[/ANSWER].
```

Aggregator: system message and revised MCQ user message

System message (all conditions)

```
You are a neutral aggregator. You did not attempt the problem yourself. Weigh
the submitted answers on the quality of their reasoning and report a single
final answer.
```

Aggregator: system message and revised MCQ user message (continued)

User message (revised MCQ)

```
{question_stem}

The answer options for this question were divided among four team members.
You receive their responses below; their option lists are not shown to you.

Independent team members submitted these responses:

{peer_block}

Give the single best answer to the question using the information shown.

End your response with EXACTLY this format: The answer is 'X' where X is the
letter of the correct option.
```

For revised document tasks, substitute the following allocation note and final-answer paragraph; keep the rest of the user message unchanged.

Revised document substitutions

Allocation note

```
The source documents for this question were divided among four team members.
You receive their responses below; the source documents are not shown
separately.
```

Final-answer paragraph

```
End your response with the answer as a short phrase, wrapped in [ANSWER] and
[/ANSWER] tags. For example: [ANSWER]wooden desk drawer[/ANSWER].
```

B How the interface affects label influence

The controls below examine omitted options, labels that single out a member, instructions requiring members to answer, aggregator sampling, and member count. They use full responses on the same MMLU-Pro STEM questions, with the study-specific coverage stated below.

B.1 Showing the full option list

Adding the full lettered option list gives the aggregator answer information beyond member reports. With original instructions and the same tasks, responses, and targets, truthful-minus-swapped accuracy losses are 2.8–12.8 points, compared with 23.2–49.2 in the earlier run without options. The visible-option losses survive correction for four aggregators; Claude Sonnet 4.6 is the exception.

B.2 Marking a response without claiming access

The marked-tag control assigns `[this member's response is marked]` to the designated non-holder and `[this member's response is not marked]` to the others. A secondary version uses `[member tag: A]` and `[member tag: B]`. Under revised instructions, access labels increase matching more than either tag pair for every aggregator (Table 4).

For Claude, Grok, and DeepSeek, the residual matching increases after disclosure are 17.6, 15.2, and 29.2 points, compared with marked-tag increases of 10.4, 2.0, and 8.0 points. These descriptive comparisons use no labels as the baseline and draw on separate runs.

Table 4: Matching changes for tags without access claims on MMLU-Pro STEM, using revised instructions and full responses. Both columns measure matching to the same designated non-holder in percentage points, retaining all 250 tasks and failures. Brackets give pointwise paired-bootstrap 95% intervals. Asterisks mark $p_H < 0.05$ for exact McNemar tests, with a separate ten-comparison Holm family in each panel. Tag conditions were collected separately from the reused no-label and swapped-label conditions.

Aggregator	Tags minus no labels	Swapped access labels minus tags
Marked/not marked (primary)		
Claude Sonnet 4.6	$+10.4$ $[6.8, 14.4]^*$	$+27.2$ $[21.6, 32.8]^*$
Grok 4.3	$+2.0$ $[-1.6, 5.6]$	$+21.2$ $[15.2, 27.2]^*$
Kimi K2 Thinking	$+4.4$ $[0.8, 8.0]$	$+22.0$ $[16.0, 28.0]^*$
GLM-5.2	$+4.8$ $[1.6, 8.0]^*$	$+30.0$ $[24.4, 36.0]^*$
DeepSeek V3.2-Exp	$+8.0$ $[4.4, 12.0]^*$	$+30.0$ $[24.0, 36.0]^*$
A/B tags (secondary)		
Claude Sonnet 4.6	$+2.4$ $[0.0, 4.8]$	$+35.2$ $[29.6, 41.2]^*$
Grok 4.3	$+1.2$ $[-2.4, 4.8]$	$+22.0$ $[16.0, 28.0]^*$
Kimi K2 Thinking	$+2.4$ $[-0.8, 5.6]$	$+24.0$ $[18.4, 29.6]^*$
GLM-5.2	$+1.2$ $[-1.6, 4.0]$	$+33.6$ $[28.0, 39.2]^*$
DeepSeek V3.2-Exp	$+0.4$ $[-3.2, 4.0]$	$+37.6$ $[31.6, 43.6]^*$

B.3 Allowing members to decline

A separate response bank changes the member prompts to permit `NONE`, while still requiring the aggregator to answer. With these new responses, accuracy losses are 6.0–18.0 points smaller than with required-answer members, but remain significant for all five aggregators. Kimi K2 Thinking's reduction interval includes zero. This control covers MCQs only.

Repeated responses and presentation. Three independently collected member-response sets preserve the direction of the label comparisons for all five aggregators. Losses also remain with randomized order or removed member headers for all five and label rewording for Claude. The corresponding saved result summaries report each design and its uncertainty separately.

B.4 Sampling the aggregator's answer

Claude Sonnet 4.6 and DeepSeek V3.2-Exp were also tested at aggregator temperature 0.7, with three seeds and the same stored member responses. Accuracy loss remains positive in all six model–seed runs (Table 5). This check concerns direction, not equivalence with temperature 0.

Table 5: **Accuracy loss persists under sampled aggregator decoding.** Truthful-minus-swapped accuracy (pp) on the same 250 MMLU-Pro STEM questions, with original instructions and full responses. Only the aggregator is sampled; stored member responses stay fixed. The temperature-0 reference comes from an earlier collection. Sampled columns are individual runs, not uncertainty intervals.

	Temperature 0	Temperature 0.7		
Aggregator	Reference	Seed 101	Seed 102	Seed 103
Claude Sonnet 4.6	30.4	28.8	32.0	29.2
DeepSeek V3.2-Exp	38.8	37.2	36.0	39.6

B.5 Adding or removing a member response

Swapping access labels reduces accuracy with three, four, and five member responses (Table 6). We retain the same evidence holder and falsely labelled member throughout. The three-member condition removes another non-holder; the five-member condition adds a Llama 3.3 70B response generated

from a non-holder's option set, changing both member count and composition. Four-member calls were collected alongside these conditions.

Table 6: **Accuracy loss persists with three and five member responses.** Truthful-minus-swapped accuracy (pp) on 250 MMLU-Pro STEM questions, using full responses, temperature 0, and revised instructions adjusted to state the member count. Brackets show paired task-bootstrap 95% intervals. All six losses remain significant after Holm correction of exact paired McNemar tests.

Aggregator	3 members	4 members	5 members
Claude Sonnet 4.6	46.8 [40.8, 53.2]	42.4 [36.4, 48.4]	42.4 [36.4, 48.8]
DeepSeek V3.2-Exp	48.8 [42.8, 54.8]	52.0 [46.0, 58.0]	52.0 [45.6, 58.4]

C Instruction comparisons and scoring

C.1 Original and revised instructions

The original instructions refer to options or documents the aggregator does not receive separately, including a request to copy an answer from "your documents." The revised instructions describe the actual inputs. They also remove user-message advice to judge reasoning rather than confidence or length and not follow an incorrect majority; the system message stays fixed.

The paired study changes instruction version while retaining the same tasks, member responses, labels, and models, with counterbalanced call order. It covers 250 MMLU-Pro STEM tasks and 300 tasks per document corpus, using both response formats. The design and analysis were specified before calls; the technical pilot remains included and continuation did not depend on its effect direction or significance. Table 7 shows that label exchange lowers accuracy and full responses reduce that loss under both versions. The revised instructions yield larger losses on STEM; several clauses differ, so the comparison measures their combined effect. Neither version was designated as primary in advance, and the timing of the presentation choice is undocumented.

Original instruction substitutions

MCQ allocation note

```
The answer options for this question have been divided among your team. You
can see only some of them.
```

Document allocation note

```
The source documents for this question have been divided among your team. You
can see only the documents assigned to you.
```

Decision paragraph (both task types)

```
Decide which response is correct. Judge the reasoning, not the confidence or
the length. If several agree but their shared reasoning is wrong, do not
follow the majority.
```

The original document answer-format instruction also asks for the answer to be "copied exactly from your documents." The revised version removes those words. All these substitutions change together; the system message stays fixed.

Other MCQ samples. The non-STEM and SuperGPQA extension likewise pairs original and revised instructions on fixed 250-task samples and stored responses, with counterbalanced call order. It tests full responses only. Matching increases and accuracy losses survive correction for every aggregator under both versions; the size of the instruction effect varies by model and benchmark.

C.2 Scoring and paired comparisons

All primary comparisons retain every selected task, including output failures. Unparseable or empty answers receive zero accuracy and matching credit. MCQ accuracy uses the extracted option letter;

Table 7: Accuracy loss after label exchange (percentage points), using the same tasks and member responses. Full denotes full responses; final denotes final answers only. Both instruction versions yield a loss and a reduction in that loss from full responses for all five aggregators on all three corpora, after correction within each five-model family. These are paired comparisons within each instruction version.

	Original instructions		Revised instructions	
Aggregator	Full	Final	Full	Final
MMLU-Pro STEM				
Claude Sonnet 4.6	34.0	86.8	42.0	87.2
Grok 4.3	27.2	87.2	35.6	88.0
Kimi K2 Thinking	27.2	86.8	37.6	84.8
GLM-5.2	48.8	78.4	51.2	83.6
DeepSeek V3.2-Exp	35.2	87.2	52.8	88.0
2Wiki				
Claude Sonnet 4.6	6.0	40.3	7.7	44.3
Grok 4.3	19.3	71.3	11.7	72.0
Kimi K2 Thinking	9.7	55.0	10.0	58.7
GLM-5.2	8.3	57.3	8.3	63.0
DeepSeek V3.2-Exp	34.7	72.7	25.0	77.7
MuSiQue				
Claude Sonnet 4.6	6.3	27.3	12.0	33.7
Grok 4.3	16.0	48.7	13.0	52.7
Kimi K2 Thinking	10.7	42.7	14.3	44.7
GLM-5.2	9.3	40.3	11.0	49.7
DeepSeek V3.2-Exp	27.0	48.0	17.0	55.3

matching trims whitespace and ignores case. Every matching member receives full credit in a tie. Matching records answer agreement, not whose reasoning was used.

For 2Wiki, symmetric entity containment removes surrounding quotes, a leading article, and trailing punctuation, then lowercases and collapses whitespace. Both strings must be nonempty and either equal or the shorter must contain at least four characters and occur as a whole phrase in the longer. MuSiQue uses official normalized exact match over the canonical answer and gold aliases; member matching uses the same normalization and merged alias groups.

For instruction version w and response format r, let $P_{w,r} = 100(\mathrm{Acc}_{\mathrm{truthful},w,r} - \mathrm{Acc}_{\mathrm{swapped},w,r})$ on the same tasks. The reduction from full responses is $R_w = P_{w,\mathrm{final}} - P_{w,\mathrm{full}}$, where final denotes final answers only. Positive P means a loss after label exchange; positive R means a smaller loss with full responses.

Binary contrasts use exact two-sided McNemar tests. Effect intervals use $5{,}000$ percentile bootstrap resamples of complete paired task records, unless stated otherwise. They are pointwise 95% intervals and describe task variability, not repeated-generation variability. Response-format reductions and instruction interactions use saturated cell-means linear probability models with task-clustered, finite-sample-corrected covariance (CR1) and normal Wald tests. No equivalence margin is tested.

Core contrasts have separate five-model Holm families for each benchmark, instruction version, response format, population, and named contrast. Each STEM disclosure contrast jointly corrects five revised effects and five instruction interactions. The two MuSiQue revised matching contrasts share ten tests across the five models. Each placebo pair and version uses ten tests: tag versus unlabelled and access label versus tag. Visible-options accuracy losses and matching increases share ten primary tests. Member-declining losses form a five-model family. Captions identify the family for each retained display; the saved JSON result summaries record secondary families.

STEM disclosure accuracy uses a separate fifteen-test Holm family: five effects under each instruction version and five instruction interactions.

C.3 SENSITIVITY TO OUTPUT FAILURES

To test how failure scoring affects accuracy loss, let f_T and f_S count unparseable truthful and swapped outputs among n tasks. Allowing each failure either correctness score, while fixing parsed scores, bounds the loss by $[P - 100f_S/n,\ P + 100f_T/n]$. Subtracting opposite endpoints gives bounds on the response-format reduction. These are bounds over failure scores, not confidence intervals.

Every accuracy loss and response-format reduction in the paired study of Table 7 retains a positive lower bound under both instruction versions. Uneven failures still limit interpreting differences between versions as changes in answer selection alone. Empty content, parsing failure, and output-limit finishes can overlap.

For disclosure, a post-hoc check retains tasks with parseable answers in both the disclosed and unlabelled conditions. Claude, Grok, and DeepSeek retain positive residual matching after correction on STEM; Grok and DeepSeek do so on MuSiQue. Kimi and GLM have smaller positive estimates that do not survive correction in either corpus. These subsets depend on the outputs, so they do not replace the primary all-task comparison. Their intervals use $10{,}000$ paired resamples and each corpus has a separate five-model Holm family.

For STEM disclosure accuracy, allowing each unparseable output either correctness score leaves an accuracy deficit for Claude, Grok, Kimi, and DeepSeek relative to no labels. GLM's difference has bounds $[-18.0, +2.0]$ percentage points; it has 21 unparseable answers without labels and 29 with disclosed false labels. These bounds describe sensitivity to failure scores, not sampling uncertainty.

In the warning control, GLM's swapped-label parsing failures fall from 51 to 32. Its 26.0-point accuracy gain has failure-score bounds $[5.6, 38.8]$, retaining a positive gain under every assignment to unparsed outputs.

MuSiQue includes one duplicate question pair, retained in the paired study.

D WHAT DISCLOSURE CHANGES AND WHERE IT FALLS SHORT

D.1 FALSE LABELS WITH AND WITHOUT DISCLOSURE

The comparison tracks the same non-holder under no labels, swapped labels, and swapped labels with disclosed zero reliability. Actual allocation and full member responses stay fixed.

Zero label reliability (STEM)

```
Exactly one member below is labeled as having seen the correct answer. Across
questions, that positively labeled member is the true holder approximately 0%
of the time.
```

The disclosure is prepended to the member-response block, after the question and allocation note.

Zero label reliability (MuSiQue)

```
Exactly one member below is labeled as having received the passages needed to
answer this question. Across questions, that positively labeled member
actually received them approximately 0% of the time.
```

This document-task disclosure is inserted immediately before the first member entry, after the sentence introducing the independent member responses. The same disclosure is used with original and revised instructions.

The uniform-prior reference in Section 3.3 describes holder probabilities using only the disclosed rate. Matching also depends on response content, so this reference does not specify a required matching rate relative to the unlabelled condition.

On STEM, disclosure lowers matching for all five aggregators, while Claude, Grok, and DeepSeek still match the labelled member more often than without labels. Table 8 gives the matching rates and accuracy comparison; Figure 4 shows the matching and accuracy contrasts. The two matching analyses share the unlabelled and disclosed calls. A *run* is a separately executed set of aggregator

calls. Disclosed and unlabelled calls mostly share a collection day, while many disclosed/swapped pairs come from different days and provider routing varies. These comparisons therefore include run differences despite fixed tasks and requested settings.

Table 8: **Accuracy remains below the no-label baseline after disclosure.** Matching tracks the same non-holder; contrasts appear in Figure 4. Accuracy differences are swapped labels with disclosure minus no labels (pp), with paired task-bootstrap 95% intervals. All five survive the 15-test Holm correction. Revised instructions, full responses, 250 MMLU-Pro STEM tasks. Unparseable outputs count as incorrect (Appendix C.3).

Matching (%)			
Aggregator	No labels	Swapped	Swapped with disclosure
Claude Sonnet 4.6	8.8	46.4	26.4
Grok 4.3	14.8	38.0	30.0
Kimi K2 Thinking	9.2	35.6	8.8
GLM-5.2	6.8	41.6	10.0
DeepSeek V3.2-Exp	16.4	54.4	45.6

Accuracy (%)			
Aggregator	No labels	Swapped with disclosure	Difference (pp) [95% CI]
Claude Sonnet 4.6	78.4	66.4	-12.0 $[-16.8, -7.2]$
Grok 4.3	61.6	50.8	-10.8 $[-16.4, -5.6]$
Kimi K2 Thinking	72.8	60.0	-12.8 $[-18.8, -6.8]$
GLM-5.2	76.0	66.4	-9.6 $[-15.6, -3.6]$
DeepSeek V3.2-Exp	62.8	41.6	-21.2 $[-27.2, -15.2]$

For Kimi, moving from no labels to swapped labels with disclosure reduces correct answers from 182 to 150 out of 250. Matches to the labelled member fall from 23 to 22, while other wrong answers rise from 34 to 55 and unparseable outputs from 11 to 23. Similar matching rates can therefore accompany different accuracies.

D.2 Disclosure on document questions

The MuSiQue comparison uses 300 tasks with fixed full responses and target selection. Under revised instructions, residual matching over no labels survives correction for Grok and DeepSeek. Adding the disclosure lowers matching for four models but raises it for DeepSeek (Table 9). Thus the reduction observed for all five models on STEM does not carry over uniformly to document questions. Non-holder document answers can be correct, so greater matching need not mean lower accuracy. The conditions were collected in separate runs.

Table 9: Disclosure on MuSiQue does not reproduce the STEM pattern for every model. Revised instructions, full responses, all 300 tasks and failures. Values are changes in matching to the same non-holder (percentage points), with pointwise paired 95% intervals and exact McNemar p_H. The two contrasts share a ten-test Holm family across five aggregators.

Aggregator	Disclosed minus swapped Change [95% CI], p_H	Disclosed minus unlabelled Change [95% CI], p_H
Claude Sonnet 4.6	-4.3 $[-7.3, -1.7]$, 0.027	$+1.3$ $[0.3, 2.7]$, 0.375
Grok 4.3	-5.3 $[-9.0, -1.7]$, 0.043	$+9.3$ $[6.3, 12.7]$, < 0.001
Kimi K2 Thinking	-13.0 $[-17.0, -9.3]$, < 0.001	$+1.7$ $[0.0, 3.3]$, 0.375
GLM-5.2	-7.0 $[-10.3, -4.0]$, < 0.001	$+1.3$ $[-0.7, 3.7]$, 0.375
DeepSeek V3.2-Exp	$+7.0$ $[2.0, 12.0]$, 0.043	$+29.0$ $[24.0, 34.0]$, < 0.001

D.3 A WARNING WITHOUT A NUMERICAL RATE

The warning below was tested in a separate run on STEM with original instructions and full responses. It precedes the member-response block in both label conditions. Table 10 gives the accuracy gain and remaining loss for each aggregator.

Warning without a numerical rate

```
Caution: the bracketed access annotations attached to the responses below are
unverified and may be wrong. Do not rely on them; weigh each member's own
reasoning instead.
```

Table 10: Warning without a numerical rate on MMLU-Pro STEM, using original instructions and full responses. Warning conditions were collected separately from the reused no-warning conditions. All 250 tasks and failures are retained. The gain compares accuracy with and without the warning under swapped labels, in percentage points. Brackets give pointwise paired-bootstrap 95% intervals from 5,000 resamples; p_H uses exact McNemar tests with a five-model Holm correction. Remaining loss is truthful-minus-swapped accuracy with the warning in both conditions, reported as a point estimate in percentage points. Failure-scoring sensitivity is reported in Appendix C.3.

Aggregator	Accuracy gain from warning under swapped labels	p_H	Remaining loss with warning
Claude Sonnet 4.6	$+4.8$ $[0.8, 8.8]$	0.143	24.4
Grok 4.3	$+0.0$ $[-4.4, 4.4]$	1.000	27.2
Kimi K2 Thinking	$+5.6$ $[0.0, 11.2]$	0.195	17.2
GLM-5.2	$+26.0$ $[20.0, 32.0]$	< 0.001	22.0
DeepSeek V3.2-Exp	$+4.0$ $[-1.6, 9.6]$	0.405	33.2