MORE COMPUTE IS NOT MORE EVIDENCE: MITIGATING MODEL-SPECIFIC ERROR BIAS AT TEST TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time scaling methods implicitly assume that additional computation from the same model provides independent evidence. We show that this assumption fails within a single reasoning model: errors are correlated across repeated samples, across reasoning depth, and in the model's own self-evaluation, so more compute from the same model is not more evidence. We introduce **DIA** (**D**isagreement-triggered **I**ndependent **A**rbitration), which treats disagreement between an intermediate-stage and the final-stage prediction as a per-question signal of reasoning instability and resolves it by asking a small panel of heterogeneous models to solve the question independently, without seeing either candidate, and aggregating their answers directly. Across seven benchmarks and two base models, **DIA** outperforms full chain-of-thought in 12 of 14 settings, including a **7.00%** gain on MuSiQue, while activating on only **∼17%** of examples and using up to **47.2% fewer tokens** than a multi-agent baseline. Ablations show that both independent solving and solver heterogeneity are necessary for these gains.

1 INTRODUCTION

Test-time scaling has become a central strategy for improving reasoning in large language models: chain-of-thought prompting (Wei et al., 2022), repeated sampling with majority vote (Wang et al., 2023), adaptive early exit (Yang et al., 2026; Fu et al., 2025; Dai et al., 2025), and multi-agent reasoning (Wu et al., 2024; Li et al., 2023; 2024) all convert additional inference-time computation into accuracy gains (Snell et al., 2025; Guo et al., 2025). Test-time scaling is particularly effective when additional computation contributes evidence that is sufficiently complementary to the model's current belief, allowing errors to be corrected rather than repeatedly reinforced. However, computation from the same model can remain shaped by persistent **model-specific error bias**, limiting the diversity of the evidence it provides. We observe three manifestations of this phenomenon across generation, reasoning, and self-evaluation (Figure 1).

(a) Generation bias under repeated sampling. Repeated sampling can diversify reasoning traces without necessarily diversifying their failures. When self-consistency produces an incorrect majority answer, erroneous samples frequently concentrate on the same dominant wrong answer rather than spreading across alternatives, indicating a persistent generation bias rather than independent error modes (Section 2.1).

(b) Initial-answer persistence bias. Reasoning does not necessarily provide evidence independent of a model's initial judgment. When a model is first asked to answer directly and then asked to reason before answering, its post-reasoning prediction can remain strongly associated with the initial answer; in particular, an initially incorrect prediction may persist through subsequent reasoning rather than being corrected (Section 2.2). Prior work similarly suggests that reasoning models can form strong preliminary answer preferences (Dang et al., 2026), that extended reasoning can lead to overthinking (Zhou et al., 2026), and that early errors may propagate through subsequent reasoning (Zhang et al., 2024).

(c) Longer-reasoning preference bias. A related bias appears when the model is asked to evaluate competing answers produced at different stages of its own reasoning. Rather than judging the candidates symmetrically, the model strongly favors the later, more elaborate reasoning outcome,

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.

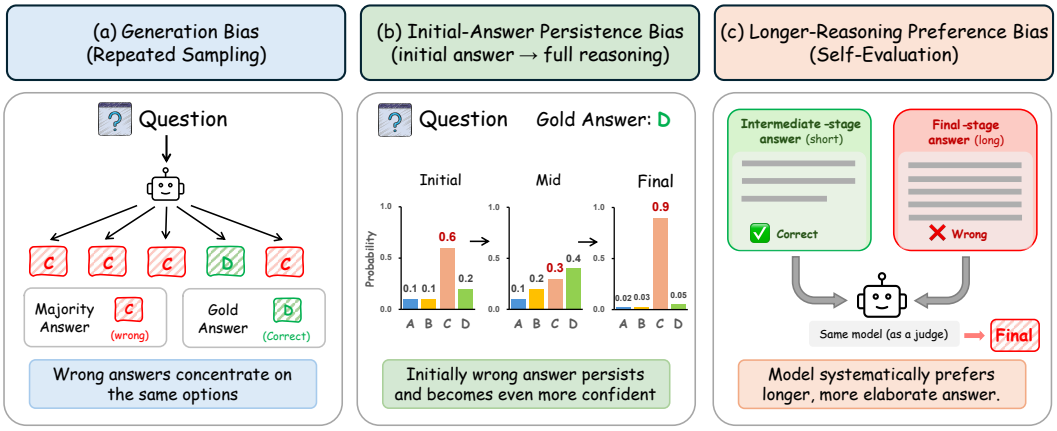


Figure 1: Three manifestations of **model-specific error bias** within a single reasoning model. **(a)** When repeated sampling produces an incorrect majority answer, erroneous samples concentrate on the same wrong answer rather than spreading across alternatives. **(b)** When a model's initial answer is incorrect, reasoning frequently preserves the same wrong answer rather than correcting it. **(c)** When answers at different reasoning stages disagree, the model's self-evaluation systematically favors the later, more elaborate outcome, even when the earlier answer is correct. Detailed analyses are in Section 2.

even when the earlier answer is correct. In our experiments, the model selects the later-stage answer in over 95% of disagreements and remains highly confident even when this preference leads to an incorrect decision (Section 2.3). This behavior is related to previously documented length and verbosity biases in LLM-based evaluation (Zheng et al., 2023; Hu et al., 2025).

Together, these observations suggest that:

> model-specific error bias persists across **generation**, **reasoning**, and **self-evaluation**, that is, the three stages of same-model computation that test-time scaling methods rely on. More compute from the same model is therefore not more evidence.

These findings suggest that useful additional computation should provide evidence that is both *selectively acquired* and *complementary* to the base model's existing reasoning.

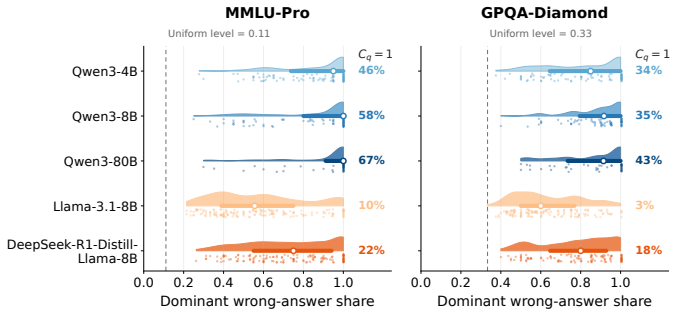
(1) When to seek additional evidence. Prior work shows that intermediate predictions and answer-stability signals can reveal how reasoning evolves and support adaptive computation (Yang et al., 2026; Fu et al., 2025; Dai et al., 2025), while extended reasoning is not always beneficial (Zhou et al., 2026). Motivated by these observations, we use disagreement between an intermediate prediction and the final prediction as a simple per-question signal of reasoning instability.

(2) Where the evidence should come from. For such cases, we seek complementary evidence from heterogeneous LLMs. Prior ensemble studies show that different models exhibit complementary strengths and can succeed on different subsets of problems (Jiang et al., 2023; Wang et al., 2025), suggesting that model heterogeneity can reduce dependence on the particular error tendencies of the base model.

(3) How the evidence should be combined. We let these solvers reason independently, without seeing the base model's candidate predictions, and aggregate their answers directly, consistent with evidence that voting-based ensembles can be competitive with more elaborate multi-agent protocols (Li et al., 2024; Smit et al., 2024).

These three principles lead to **DIA** (**D**isagreement-triggered **I**ndependent **A**rbitration): agreement between the intermediate and final predictions is returned directly, whereas such reasoning-stage disagreement activates a small heterogeneous solver panel for candidate-blind, answer-level arbitration.

Across seven benchmarks and two base models (Qwen3-4B and Qwen3-8B; Yang et al., 2025), **DIA** outperforms full chain-of-thought in 12 of 14 settings, including an improvement from 59.0% to 66.0% on MuSiQue (Trivedi et al., 2022), while activating on only ∼17% of examples and using up



Dataset	Initially Wrong	Error Retained
Qwen3-8B		
MMLU-Pro	52.29%	26.23%
MuSR	42.00%	61.90%
DetectBench	43.53%	35.54%
Qwen3-Next-80B-A3B-Instruct		
MMLU-Pro	41.43%	32.41%
MuSR	37.50%	58.67%
DetectBench	39.07%	34.86%

Figure 2: Generation bias and initial-answer persistence bias. **Left:** Wrong-answer concentration under SC@20 failures. Each point denotes a failed question and reports its dominant wrong-answer share C_q; dashed lines mark the uniform-choice baseline. **Right:** Initial-answer persistence across base-model scales. Error Retained denotes the fraction of initially incorrect answers for which full CoT returns the same incorrect answer.

to 47.2% fewer tokens than a fully-connected multi-agent baseline that invokes every agent on all questions. A multi-agent baseline that routes independent agent outputs through a centralized judge underperforms direct answer aggregation, e.g., 60.67% vs. 66.00% on MuSiQue with Qwen3-4B, highlighting the importance of the aggregation protocol.

Contributions.

① **Three manifestations of model-specific error bias.** We identify systematic error persistence across three stages of same-model computation: repeated sampling can reproduce a shared wrong answer, reasoning can preserve an initially incorrect prediction, and self-evaluation can systematically favor a later and more elaborate reasoning outcome (Section 2).

② **Disagreement-triggered independent arbitration.** To mitigate model-specific error bias, we propose **DIA**, a framework that detects reasoning instability via reasoning-stage disagreement and resolves it through candidate-blind, answer-level arbitration by heterogeneous external solvers (Section 3).

③ **Comprehensive evaluation.** Across seven benchmarks and two base models, **DIA** improves over full chain-of-thought in 12 of 14 settings while activating on only ∼17% of examples and using up to 47.2% fewer tokens than the multi-agent baseline (Section 4).

2 MOTIVATION

2.1 GENERATION BIAS UNDER REPEATED SAMPLING

Repeated sampling from a reasoning model can produce varied reasoning traces while sharing the same failure pattern. We quantify this effect on questions where majority-vote self-consistency produces an incorrect answer. For each failed question q, we denote by C_q the fraction of incorrect samples assigned to the most frequent wrong answer.

Figure 2 shows the distribution of C_q under SC@20 on MMLU-Pro and GPQA-Diamond. Across models and datasets, the wrong answers are more concentrated than the uniform-choice baseline. For many failed questions, all incorrect samples converge on a single answer ($C_q = 1$).

Figure 2 shows three patterns. First, when repeated samples from the same model are wrong, they often collapse onto a single answer rather than spread across alternatives. Second, error concentration increases with model scale. Third, concentration is higher in the reasoning-oriented models. **Takeaway. Increasing the sampling budget of the same model may repeatedly reinforce an existing error mode rather than introduce independent evidence.**

2.2 INITIAL-ANSWER PERSISTENCE BIAS

Reasoning is expected to provide additional evidence beyond a model's initial judgment. However, a model may already exhibit a strong answer preference before any explicit reasoning, and the same

Table 1: Self-selection on reasoning-stage disagreements with Qwen3-8B. We report preference for the final-stage prediction, overall accuracy when exactly one candidate is correct, directional accuracy conditioned on which stage is correct, and confidence on incorrect selections. Wrong-choice Conf. is the mean softmax probability assigned to the chosen answer token when the selection is incorrect.

Dataset	Final-stage Pref.	Self-select Acc.↑	Intermediate✓ / Final× ↑	Intermediate× / Final✓ ↑	Wrong-choice Conf.↓
MuSR	97.7%	44.1%	0.0%	93.8%	100.0%
MMLU-Pro	95.1%	63.4%	6.7%	96.2%	100.0%

preference can remain after full reasoning. To examine this initial-answer persistence, we first ask the model to produce a direct answer without reasoning and ask the same model to solve the same question with full chain-of-thought reasoning. We measure whether the final answer retains the initial answer, particularly when the initial answer is incorrect.

The right panel of Figure 2 shows that full CoT often reproduces the model's initially incorrect answer rather than independently reassessing the question. Same-error persistence is strongest on MuSR and remains visible across model scales and datasets. **Takeaway. Additional reasoning can elaborate an existing error without supplying the independent evidence needed to overturn it.**

2.3 Longer-Reasoning Preference Bias

A model's prediction can change as reasoning proceeds. To capture this change, we compare an *intermediate-stage prediction*, obtained at an intermediate reasoning stage where the model is sufficiently confident, with the *final-stage prediction* produced after the model completes its reasoning. When the two disagree, which we call a *reasoning-stage disagreement*, one simple way to resolve the conflict is to let the model select between its own two predictions. This choice may not be neutral: the model may systematically favor the final-stage prediction simply because it follows more reasoning. We examine this bias with Qwen3-8B on reasoning-stage disagreements. Candidates are anonymized and their presentation order is balanced.

Table 1 reveals a strong directional asymmetry in self-selection. On reasoning-stage disagreements, the model selects the final-stage prediction in more than 95% of cases, and its selections are reliable mainly when that prediction is correct. When the intermediate-stage prediction is correct but the final-stage prediction is incorrect, the model rarely selects the earlier prediction. This preference persists with high confidence even when it leads to an incorrect choice. **Takeaway. These results reveal a longer-reasoning preference bias: the model tends to trust the later, more elaborate reasoning outcome even when further reasoning has moved away from the correct answer.**

3 Method

Framework. We propose **DIA** (Disagreement-triggered Independent Arbitration) to resolve disagreements between predictions produced at two reasoning depths (Figure 3). Given a question q, the base model yields a *final-stage prediction* after completing its reasoning and, whenever a confidence criterion is met at an intermediate reasoning stage, an *intermediate-stage prediction* extracted at that point. For compact notation, we denote the intermediate-stage and final-stage predictions by $a_E(q)$ and $a_F(q)$, respectively, and let $m_E(q) \in \{0, 1\}$ indicate whether an intermediate-stage prediction is available. **DIA** takes these predictions as fixed inputs; the extraction procedure for $a_E(q)$ is described in Section 4.1.

Disagreement-triggered routing. If no intermediate-stage prediction is available ($m_E(q) = 0$), we return $a_F(q)$. If $a_E(q) = a_F(q)$, we return the shared answer. External solvers are called only on reasoning-stage disagreement, that is, when $a_E(q)$ is available and differs from $a_F(q)$. We denote this routing decision by

$$z(q) = \begin{cases} 1, & \text{if } m_E(q) = 1 \text{ and } a_E(q) \neq a_F(q), \\ 0, & \text{otherwise.} \end{cases}$$

Candidate-blind independent solving. When $z(q) = 1$, we ask three heterogeneous solvers to solve the original question. Each solver receives only the question and its answer options and an-

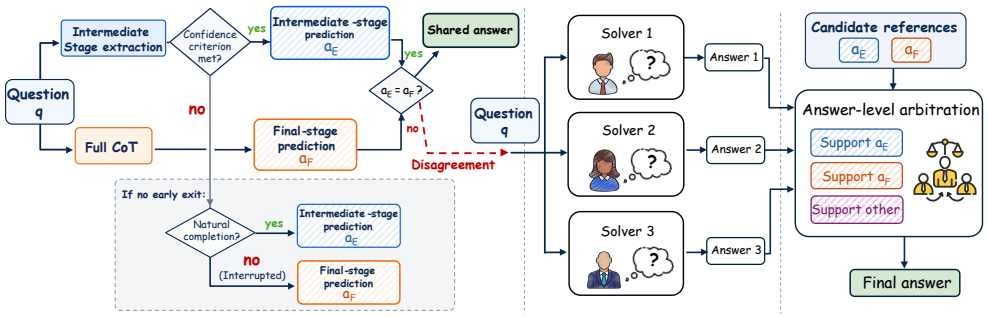


Figure 3: Overview of **DIA**. The base model produces an intermediate-stage prediction a_E and a final-stage prediction a_F. Agreement returns the shared answer directly; disagreement activates three heterogeneous solvers that solve the original question independently behind an information firewall. Answer-level arbitration aggregates support for a_E, a_F, or a third answer.

swers the question independently. No solver sees $a_E(q)$, $a_F(q)$, the reasoning traces behind them, or another solver's output. Each valid response is normalized to a task-specific answer representation v_k.

Answer-level arbitration. Let $\mathcal{V}$ denote the indices of solvers with valid responses. We count their support for a_E and a_F:

$$n_E = \sum_{k \in \mathcal{V}} \mathbf{1}[v_k = a_E], \qquad n_F = \sum_{k \in \mathcal{V}} \mathbf{1}[v_k = a_F].$$

If at least two solvers agree on an answer outside $\{a_E, a_F\}$, we return that answer. Otherwise, we return a_E when $n_E > n_F$ and a_F when $n_F > n_E$.

3.1 Theoretical Analysis

We explain why heterogeneous solvers and candidate-blind solving help through a single quantity, the *truth margin*. Formal statements and proofs are in Appendix B.

Setup. For a question with gold answer $y^\star$ and a panel S of n solvers with average accuracy $\bar{p}_S$, let $\kappa_S \in [0, 1]$ be the fraction of the panel's expected wrong votes that fall on its single most popular wrong answer, and define

$$\Delta_S = \bar{p}_S - \kappa_S (1 - \bar{p}_S). \tag{1}$$

Δ_S is the expected vote share of $y^\star$ minus that of the strongest wrong answer. The definition encodes an asymmetry: correct votes always coincide on $y^\star$, whereas wrong votes defeat the truth only if they also coincide on the same wrong answer.

Proposition 1 (informal; Proposition B.1)**.** *Under independent voting, plurality is governed by the truth margin Δ_S, the expected vote share of the gold answer relative to the strongest wrong alternative. If $\Delta_S > 0$, increasing the number of independent votes concentrates the outcome on the gold answer, exponentially fast in n; if $\Delta_S < 0$, additional votes instead reinforce the dominant error mode.*

Repeated samples from the same model preserve the answer distribution and therefore the sign of Δ_S: for a homogeneous panel, $\bar{p}_S$ is the model's own accuracy and κ_S is exactly the wrong-answer concentration C_q of Section 2.1, and neither depends on the sampling budget. Only a change of solver population can change the sign of the margin, so the benefit of heterogeneity is not diversity per se but a reduction of shared error modes; Section 4.3.2 tests this prediction.

Proposition 2 (informal; Proposition B.2)**.** *Candidate exposure systematically shifts a solver's answer distribution toward the exposed candidate. If the exposed candidate is correct, this shift increases the truth margin. If it is incorrect, the same shift decreases the margin against that wrong answer and can reverse its sign. Candidate-blind solving removes this pre-aggregation shift.*

Exposure amplifies the candidate being shown, not necessarily the truth, so communication helps only when solvers can reliably distinguish correct from incorrect shared answers, which is exactly

what Section 2.3 calls into question. **DIA** therefore obtains each solver's judgment before revealing any candidate, so the evidence entering aggregation is not conditioned on the alternatives being arbitrated.

DIA pipeline: reasoning-stage disagreement → heterogeneous external evidence → candidate-blind independent solving → direct answer aggregation.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on seven benchmarks across four reasoning domains. MMLU-Pro (Wang et al., 2024) and GPQA-Diamond (Rein et al., 2024) cover knowledge-intensive reasoning. DetectBench (Gu et al., 2024), MuSiQue (Trivedi et al., 2022), and 2WikiMultiHopQA (Ho et al., 2020) cover contextual and multi-hop reasoning. We use MATH (Hendrycks et al., 2021), restricted to problems of difficulty levels 3 to 5, for mathematical reasoning and MuSR (Sprague et al., 2024) for narrative or soft reasoning. Dataset statistics and sampling details are provided in Appendix A.1.

Baselines. We compare **DIA** with seven baselines, grouped by the kind of intervention each applies against model-specific error bias. ① **Standard reasoning.** CoT (Wei et al., 2022) lets the base model complete its reasoning and returns the final prediction. ② **Adaptive reasoning.** DEER (Yang et al., 2026) and Dynasor (Fu et al., 2025) use intermediate reasoning signals to decide whether continued reasoning is necessary: DEER relies on confidence at intermediate reasoning states, whereas Dynasor, in its CoT setting, uses the stability of intermediate predictions. Like **DIA**, both act on signals exposed before reasoning terminates, but they modify the computation of the same model rather than seeking evidence outside it. ③ **Multi-agent inference.** We consider two protocols using the same four-agent pool. **MAS-Exchange** (Wu et al., 2024; Li et al., 2023) uses fully-connected communication: agents first solve the question independently, observe all peer responses, revise once, and then vote. **MAS-Judge** (Wang et al., 2025) removes peer-to-peer communication: agents solve independently, after which a centralized judge, instantiated by the base model, observes all responses and selects the final answer. ④ **Learning-based debiasing.** We test whether model-specific error bias can be reduced through parameter updates: SFT trains the base model on its own correct samples and DPO then prefers correct over incorrect samples of the same question (details in Appendix A.2).

Implementation. We use Qwen3-4B and Qwen3-8B as base models. The external solvers are DeepSeek-v4-Flash (DeepSeek-AI, 2026), Llama-3.3-70B-Instruct (Meta, 2024), and GPT-4o-mini (OpenAI, 2024). Following Yang et al. (2026), we extract the intermediate-stage prediction at reasoning-transition points when the confidence criterion is met. The final-stage prediction is obtained by letting the same base model complete its reasoning under matched inference settings.

Metrics. We report answer accuracy in percent, and all accuracy differences are absolute differences. Token Ratio is the total input and output token usage normalized by the CoT baseline on the same dataset and model.

Evaluation protocol. Answers are normalized to option keys before aggregation. The final selected answer is scored against the benchmark gold option. For free-form outputs, the final answer is evaluated by Qwen 3.7-Plus (`qwen/qwen3.7-plus`), an independent LLM judge, against the benchmark gold answer.

4.2 EXPERIMENTAL RESULTS

Observation ①. DIA achieves the highest average accuracy at both model scales. Across the seven benchmarks, **DIA** improves average accuracy over CoT by 3.05% and 2.45% with Qwen3-4B and Qwen3-8B, respectively, and outperforms CoT in 12 of 14 settings. The two exceptions are on MATH, where CoT accuracy is already near ceiling and reasoning-stage disagreement leaves little room for correction.

Table 2: Main results across model scales. For each dataset, adjacent columns report accuracy (%) and the total input and output token ratio relative to CoT. Average Acc. and Tok. are the corresponding unweighted arithmetic means across the seven datasets. **Bold** indicates the highest reported accuracy in each column within each base-model block.

Models	Methods	Knowledge-intensive				Contextual & Multi-hop						Mathematical		Narrative / Soft		Average	
		MMLU-Pro		GPQA-Diamond		DetectBench		MuSiQue		2Wiki		MATH		MuSR		All Datasets	
		Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓	Acc.↑	Tok.↓
Qwen3-4B	CoT	68.00	1.00	51.01	1.00	67.74	1.00	59.00	1.00	89.50	1.00	96.67	1.00	63.50	1.00	70.77	1.00
	DEER	63.14	0.86	56.06	1.60	61.65	0.67	51.67	0.84	88.00	0.29	95.33	1.18	58.50	0.68	67.76	0.87
	Dynasor	63.71	1.68	51.01	1.76	65.95	1.30	55.00	1.16	90.50	1.21	95.67	2.00	63.50	1.51	69.33	1.52
	MAS-Exchange	**70.29**	4.38	59.09	2.75	67.74	4.05	61.33	5.32	90.00	3.53	**97.00**	2.50	62.50	4.04	72.56	3.80
	MAS-Judge	66.86	3.59	58.59	2.69	66.67	3.30	60.67	4.34	90.00	3.04	96.33	2.38	61.00	3.39	71.45	3.25
	SFT	66.29	1.27	–	–	67.38	1.53	57.33	1.36	88.50	1.16	96.00	1.30	**66.50**	1.36	–	–
	SFT+DPO	66.00	1.37	–	–	67.38	1.35	59.00	1.37	88.00	1.16	95.67	1.30	63.00	1.25	–	–
	DIA	**70.29**	3.04	**59.60**	2.63	**69.18**	2.80	**66.00**	2.90	**91.00**	2.38	96.67	2.21	64.00	2.78	**73.82**	2.68
Qwen3-8B	CoT	71.43	1.00	59.09	1.00	68.46	1.00	64.00	1.00	86.50	1.00	**98.00**	1.00	60.00	1.00	72.50	1.00
	DEER	68.00	0.79	60.10	1.61	64.87	0.62	56.67	0.82	87.50	0.29	96.33	0.84	61.00	0.74	70.64	0.82
	Dynasor	67.71	1.42	58.59	1.36	67.74	0.91	60.00	1.08	86.00	1.14	95.67	1.81	57.50	1.07	70.46	1.26
	MAS-Exchange	**75.14**	4.27	**61.62**	2.64	**69.53**	3.83	62.67	5.47	88.50	3.57	**98.00**	2.22	62.50	4.11	73.99	3.73
	MAS-Judge	73.14	3.48	60.61	2.62	68.10	3.34	61.67	4.43	**89.00**	3.05	97.00	2.02	61.00	3.51	72.93	3.21
	SFT	68.57	1.34	–	–	68.10	1.30	63.67	1.19	88.00	1.00	97.33	1.19	**63.50**	1.34	–	–
	SFT+DPO	70.57	1.34	–	–	67.74	1.36	64.00	1.04	88.00	1.12	–†	–	**63.50**	1.22	–	–
	DIA	**75.14**	2.94	61.11	2.63	**69.53**	2.70	**68.67**	2.89	**89.00**	2.48	97.67	1.88	**63.50**	2.84	**74.95**	2.62

†The DPO stage was skipped for MATH with Qwen3-8B because insufficient valid preference pairs could be constructed to meet our minimum training requirement. Only the SFT result is reported.

Observation ②. Detecting unstable reasoning is not sufficient; the source of corrective evidence matters. DEER and Dynasor modify or terminate the base model's reasoning using intermediate signals, whereas **DIA** uses such signals to request independent evidence from heterogeneous models. **DIA** achieves higher accuracy than both adaptive-reasoning baselines in all 14 settings, although at a higher token cost than early exit.

Observation ③. How multiple models are combined matters as much as having multiple models. MAS-Exchange introduces cross-agent conditioning before the final vote, while MAS-Judge preserves independent generation but introduces a candidate-aware centralized decision. **DIA** avoids both operations and directly aggregates independently obtained answers. Compared with MAS-Exchange, it improves average accuracy by 1.26% and 0.96% with Qwen3-4B and Qwen3-8B while lowering the average token ratio by about 30% at both scales, and it also exceeds MAS-Judge in average accuracy with fewer tokens; Section 4.3 examines the underlying mechanism.

Observation ④. Learning-based correction does not reliably remove the error modes. SFT and SFT+DPO provide limited and inconsistent gains over CoT despite being trained on correct and incorrect samples of the base model itself. Gains, where present, are confined to MuSR and, for Qwen3-8B, 2WikiMultiHopQA; in the remaining settings both variants match or fall below CoT, and adding DPO after SFT does not consistently help.

4.3 Ablation Study

4.3.1 Ablating Independent Solving with Direct Trace Comparison

We test whether external models can resolve reasoning-stage disagreements by comparing the candidate reasoning traces rather than solving the question independently. We use Qwen3-8B as the base model and retain the three default external models as reviewers on MuSR and MMLU-Pro. Candidates are anonymized and their presentation order is balanced. Each reviewer independently selects one candidate, and the final decision follows a two-of-three majority vote.

Full-trace comparison. With complete traces, the ensemble selects the final-stage prediction in over 90% of disagreements on both datasets (Table 3), and rarely recovers a correct intermediate-stage prediction. The final-stage preference therefore persists even with heterogeneous reviewers.

Table 3: Direct trace comparison with full traces or equal-budget summaries (Qwen3-8B base model). Final Pref. is the fraction of reasoning-stage disagreements where the reviewer ensemble selects the final-stage prediction. Sel. Acc. is the correct-selection rate when exactly one candidate is correct. The last two columns report the probability of selecting the intermediate-stage prediction when only it is correct, and selecting the final-stage prediction when only it is correct, respectively.

Setting	Dataset	Final Pref.	Sel. Acc.↑	Inter.✓ / Final× ↑	Inter.× / Final✓ ↑
Full Trace	MuSR	93.0%	47.1%	5.6%	93.8%
	MMLU-Pro	91.8%	73.2%	26.7%	100.0%
Equal Budget	MuSR	69.8%	50.0%	33.3%	68.8%
	MMLU-Pro	61.7%	65.0%	57.1%	69.2%

Table 4: Solver heterogeneity and base-model scaling. **Left:** Effect of solver heterogeneity with Qwen3-8B as the base model. Arbitration accuracy is measured on activated reasoning-stage disagreements; the homogeneous panel uses three independently sampled Qwen3-8B solvers. **Right:** Scaling the base model with a fixed heterogeneous solver panel. We report full-dataset accuracy and activation rate, defined as the fraction of examples that trigger arbitration through reasoning-stage disagreement.

Dataset	Hetero.	3 Qwen3-8B	Δ
MuSR	55.81%	51.16%	−4.65%
2WikiMultiHopQA	53.85%	19.23%	−34.62%
MuSiQue	54.72%	24.53%	−30.19%

Dataset	Base Model	Accuracy (%) ↑			Activation (%)
		CoT	DEER	DIA	
MuSR	Qwen3-8B	60.00	61.00	**63.50**	21.50
	Qwen3-30B	62.00	61.50	**66.00**	24.50
	Qwen3-80B	66.00	67.50	**68.50**	14.50
MMLU-Pro	Qwen3-8B	71.43	68.00	**75.14**	17.43
	Qwen3-30B	75.14	72.86	**76.00**	13.43
	Qwen3-80B	76.86	75.71	**78.00**	9.14

Equal-budget summaries. To examine the role of presentation, Qwen3-8B summarizes each trace with the same prompt, structure, and approximately matched token budget. This reduces the final-stage preference and the directional asymmetry, but selection accuracy improves only slightly on MuSR and decreases on MMLU-Pro. **Takeaway. These results motivate asking external solvers to answer the question independently rather than judge competing traces.**

4.3.2 HOMOGENEOUS QWEN3-8B REVIEWERS

We replace the heterogeneous solvers with three independently sampled Qwen3-8B solvers. The base model remains Qwen3-8B, and the arbitration rule is unchanged. Both settings are evaluated on the same activated reasoning-stage disagreements.

Arbitration accuracy decreases on all three datasets, with drops of more than 30% on both 2WikiMultiHopQA and MuSiQue (Table 4). On these two datasets, every question answered correctly by the homogeneous panel is also answered correctly by the heterogeneous panel. This pattern is consistent with the model-specific error bias discussed in Section 2. **Takeaway. It supports using heterogeneous solvers to obtain corrective evidence rather than relying on repeated samples from the base model.**

4.3.3 SCALING TO STRONGER BASE MODELS

We keep the three heterogeneous solvers fixed and evaluate Qwen3 base models at 8B, 30B, and 80B scales[1] on MuSR and MMLU-Pro. **DIA** outperforms both CoT and DEER in all six settings (Table 4). The gains persist at 80B without upgrading the external solvers. **Takeaway. Together with the homogeneous-solver ablation in Section 4.3.2, these results suggest that heterogeneous models provide useful complementary predictions even as the base model becomes stronger.**

Activation rates decrease with model scale, as larger models form more stable intermediate-stage predictions, yet **DIA** continues to improve accuracy while invoking external solvers less often.

[1]The 30B and 80B variants are Qwen3-30B-A3B-Instruct-2507 and Qwen3-Next-80B-A3B-Instruct, respectively.

5 Related Work

Systematic error bias in LLM reasoning. The errors of a language model are not independent across samples, reasoning steps, or self-evaluation. At the generation level, models commit to an answer early and then rationalize it (Zhang et al., 2024), larger models misfire at the commitment step despite retaining the correct answer (Yeom et al., 2026), and this commitment is skewed by surface features and prior beliefs (Zheng et al., 2024; Jiang et al., 2024; Dasgupta et al., 2022). Because such biases are shared across samples, majority voting over repeated samples (Wang et al., 2023) can amplify a shared wrong answer rather than average it out. At the reasoning level, chain-of-thought explanations can be unfaithful to the factors that drive the prediction (Turpin et al., 2023), reasoning models form strong preliminary answer preferences that trigger overthinking (Dang et al., 2026; Chen et al., 2025), and longer reasoning does not reliably improve accuracy (Zhou et al., 2026). At the evaluation level, LLM judges exhibit position, verbosity, and self-preference biases (Zheng et al., 2023; Wataoka et al., 2024; Ye et al., 2025; Hu et al., 2025). Prior work studies these biases in isolation, largely on multiple-choice or short chain-of-thought settings; we show that they co-occur within a single reasoning model and address them at test time by acquiring evidence from outside the model.

Adaptive reasoning and early exit. Test-time compute can be allocated adaptively (Snell et al., 2025), either by constraining reasoning length (Han et al., 2025; Xu et al., 2025; Aggarwal & Welleck, 2025) or by monitoring intermediate answers and stopping once they stabilize or become confident (Fu et al., 2025; Yang et al., 2026; Dai et al., 2025). These methods decide *how much* to compute and trust the intermediate-stage prediction whenever it is confident, which is where the preliminary-answer bias above resides. **DIA** uses the same kind of signal to decide *when* to seek external evidence, and retains both the intermediate-stage and final-stage predictions.

Multi-agent reasoning and its failure modes. Multi-agent methods improve reasoning by letting agents exchange and revise answers (Wu et al., 2024; Li et al., 2023; Liang et al., 2024), by layering models on earlier outputs (Wang et al., 2025), or by sampling more agents and voting (Li et al., 2024). Recent analyses question whether communication adds independent evidence: agents conform to a majority, especially when uncertain, and frequently switch from a correct answer to an incorrect one to match their peers (Zhu et al., 2025; Cho et al., 2025; Wynn et al., 2025); most of the benefit of debate comes from majority voting rather than from the exchange itself (Smit et al., 2024; Choi et al., 2025); and many failures of agent systems trace to inter-agent misalignment and weak verification (Cemri et al., 2025). Cost is a further concern, since invoking a full agent system on every query is expensive, which motivates query-dependent agent architectures (Liu et al., 2023; Zhang et al., 2025). **DIA** is designed around these findings: solvers reason independently and never see the base model's candidates or each other, their answers are aggregated directly rather than through a candidate-aware judge, which inherits the judge biases above (Table 3), the panel is heterogeneous because homogeneous solvers share the base model's error modes (Section 4.3.2), and it is invoked only on reasoning-stage disagreements.

6 Conclusion

We show that model-specific error bias persists across repeated sampling, continued reasoning, and self-evaluation, so that more compute from the same model does not amount to more evidence. **DIA** uses reasoning-stage disagreement to trigger candidate-blind solving by heterogeneous models and aggregates their answers directly. Across seven benchmarks, it outperforms CoT in 12 of 14 settings and achieves higher average accuracy with fewer tokens than the multi-agent baselines. Ablations show that asking heterogeneous models to compare the base model's reasoning traces inherits its final-stage preference, and that replacing the heterogeneous panel with repeated samples of the base model sharply reduces arbitration accuracy, so independent solving and solver heterogeneity are both necessary. The gains persist as the base model scales to 80B while the external solvers remain fixed. Test-time scaling should therefore consider where additional evidence comes from and how it is combined, not only how much computation is used.

Ethics Statement

We confirm that all authors have read and adhered to the ICLR Code of Ethics. This work complies with the ethical principles outlined therein, and no part of this study violates ethical standards.

References

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why Do Multi-Agent LLM Systems Fail? In *Advances in Neural Information Processing Systems*, 2025. URL https://mlanthology.org/neurips/2025/cemri2025neurips-multiagent/.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of long reasoning models. In *Forty-second International Conference on Machine Learning*, 2025.

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

Hyeong Kyu Choi, Jerry Zhu, and Sharon Li. Debate or vote: Which yields better decisions in multi-agent large language models? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 101732–101764. Curran Associates, Inc., 2025. doi: 10.52202/085713-3405. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/934252acd87f254d5d4672fbde283bd2-Paper-Conference.pdf.

Muzhi Dai, Chenxu Yang, and Qingyi Si. S-grpo: Early exit via reinforcement learning in reasoning models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48178–48204. Curran Associates, Inc., 2025. doi: 10.52202/085713-1608. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/44f09d1973a88529b40029e6c3693ded-Paper-Conference.pdf.

Renfei Dang, Zhening Li, Shujian Huang, and Jiajun Chen. The first impression problem: Internal bias triggers overthinking in reasoning models, 2026. URL https://arxiv.org/abs/2505.16448.

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

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Yichao Fu, Junda Chen, Siqi Zhu, Zheyu Fu, Zhongdongming Dai, Yonghao Zhuang, Yian Ma, Aurick Qiao, Tajana S Rosing, Ion Stoica, and Hao Zhang. Efficiently scaling llm reasoning programs with certaindex. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 141987–142025. Curran Associates, Inc., 2025. doi: 10.52202/085713-4751. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d037fd021c9aace128b8ce25001cdb6c-Paper-Conference.pdf.

Zhouhong Gu, Lin Zhang, Xiaoxuan Zhu, Jiangjie Chen, Wenhao Huang, Yikai Zhang, Shusen Wang, Zheyu Ye, Yan Gao, Hongwei Feng, and Yanghua Xiao. DetectBench: Can large language model detect and piece together implicit evidence? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 199–222, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.11. URL https://aclanthology.org/2024.findings-emnlp.11/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081): 633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. Token-budget-aware llm reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24842–24855, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

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/.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. Explaining length bias in LLM-based preference evaluations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6763–6794, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.358. URL https://aclanthology.org/2025.findings-emnlp.358/.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo Jose Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4722–4756, 2024.

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

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008. Curran Associates, Inc., 2023. doi: 10.52202/075280-2264. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a3621ee907def47c1b952ade25c67698-Paper-Conference.pdf.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=bgzUSZ8aeg.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.

Meta. Llama-3.3-70b-instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct, 2024.

OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. Should we be going MAD? A look at multi-agent debate strategies for LLMs. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45883–45905. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/smit24a.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b623663fd9b874366f3ce019fdfdd44-Paper-Conference.pdf.

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jenyYQzue1.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

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

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Y Zou. Mixture-of-agents enhances large language model capabilities. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 33944–33963, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5434be94e82c54327bb9dcaf7fca52b6-Paper-Conference.pdf.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang (Eric) Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Ahmed Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *COLM 2024*, August 2024. URL https://www.microsoft.com/en-us/research/publication/autogen-enabling-next-gen-llm-applications-via-multi-agent-conversation-framework/.

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

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. Chain of draft: Thinking faster by writing less. *arXiv preprint arXiv:2502.18600*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 88170–88210, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8df90a1440ce782d1f5607b7a38f2531-Paper-Conference.pdf.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pp. 102351–102390, 2025.

Jewon Yeom, Jaewon Sok, Heejun Kim, Seonghyeon Park, Jeongjae Park, and Taesup Kim. Hallucination as commitment failure: Larger llms misfire despite knowing the answer. *arXiv preprint arXiv:2605.22007*, 2026.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent architecture search via agentic supernet. *arXiv preprint arXiv:2502.04180*, 2025.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59670–59684. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhang24ay.html.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Shu Zhou, Rui Ling, Junan Chen, Xin Wang, Tao Fan, and Hao Wang. When more thinking hurts: Overthinking in LLM test-time compute scaling. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23967–23977, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1199. URL https://aclanthology.org/2026.findings-acl.1199/.

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 Appendix

A.1 Dataset Statistics and Sampling

We evaluate seven benchmark families spanning the four reasoning domains introduced in the main text. All evaluation subsets are fixed before inference, and the same canonical inputs are shared across methods and base models.

Knowledge-intensive reasoning. MMLU-Pro (Wang et al., 2024) contains 12,032 test questions from 14 broad knowledge domains. We sample 25 questions from each domain, yielding a balanced 350-question subset. For GPQA (Rein et al., 2024), we use the full Diamond split of 198 questions. The split contains 19 Biology, 93 Chemistry, and 86 Physics questions.

Contextual and multi-hop reasoning. For DetectBench (Gu et al., 2024), we use 279 questions sampled without replacement from the 3,928 questions across the official train, dev, test, test-hard, and test-distract files. For MuSiQue (Trivedi et al., 2022), we sample 300 answerable questions from the 2,417-question MuSiQue-Ans development pool; the resulting subset contains 163 two-hop, 84 three-hop, and 53 four-hop questions. For 2WikiMultiHopQA (Ho et al., 2020), we sample 200 questions from the 12,576-example development split. This subset contains 44 bridge-comparison, 48 comparison, 83 compositional, and 25 inference questions.

Mathematical reasoning. MATH (Hendrycks et al., 2021) covers seven subjects: Algebra, Counting and Probability, Geometry, Intermediate Algebra, Number Theory, Prealgebra, and Precalculus. We independently sample 100 questions from each of difficulty Levels 3, 4, and 5 from the official test split. This produces three equally sized subsets, or 300 questions in total, and balances the evaluation across difficulty levels.

Narrative and soft reasoning. MuSR (Sprague et al., 2024) comprises murder mysteries, object placements, and team-allocation tasks. We sample 100 questions from each of the object-placement and team-allocation task families, yielding 200 evaluation questions in total.

A.2 SFT and DPO Baselines

For each dataset we construct a separate training set, sample the base model eight times per question, and label each sample as correct or incorrect. SFT trains the base model for two epochs on the correct samples; DPO then trains for one epoch on preference pairs in which a correct sample is preferred over an incorrect sample of the same question. These baselines test whether repeatedly observed error modes can be learned away, as opposed to bypassed at inference time using complementary external evidence.

A.3 Additional Analyses

A.3.1 Activation Rates

Table 5 reports the fraction of examples for which the intermediate-stage and final-stage predictions disagree and selective arbitration is invoked. Activation is most frequent on MuSiQue and least frequent on GPQA-Diamond, while the overall rates remain similar across Qwen3-4B and Qwen3-8B.

Table 5: Activation rates. Each entry is the percentage of evaluation examples for which the intermediate-stage and final-stage predictions disagree and arbitration is invoked.

Dataset	Qwen3-4B	Qwen3-8B
DetectBench	20.79%	17.92%
MMLU-Pro	19.43%	17.43%
GPQA-Diamond	2.02%	1.01%
MATH L3	9.00%	5.00%
MATH L4	5.00%	6.00%
MATH L5	9.00%	14.00%
MATH overall	7.67%	8.33%
MuSiQue	33.67%	35.33%
2WikiMultiHopQA	11.00%	13.00%
MuSR	21.50%	21.50%
Overall	17.46%	17.13%

A.3.2 Reviewer Count

We test whether increasing the reviewer ensemble improves arbitration. The base model is fixed to Qwen3-8B. Starting from DeepSeek-v4-Flash, Llama-3.3-70B-Instruct, and GPT-4o-mini, we add Phi-4 and Mistral-Small-3.2-24B to form a five-reviewer ensemble.

Table 6: Effect of increasing the reviewer ensemble from three to five models. Results are evaluated only on activated examples, where the intermediate-stage and final-stage predictions disagree. Bold values emphasize arbitration accuracy; the smaller gray text reports the corresponding number of correct decisions. Δ denotes the change in the number of correct decisions.

Dataset	Activated	3 Reviewers	5 Reviewers	Δ
MuSR	43	**55.81%** (24/43)	**55.81%** (24/43)	0
2WikiMultiHopQA	26	**53.85%** (14/26)	**53.85%** (14/26)	0
DetectBench	50	**48.00%** (24/50)	**46.00%** (23/50)	-1

Increasing the ensemble from three to five reviewers provides no improvement: performance is unchanged on MuSR and 2WikiMultiHopQA and decreases by one correct arbitration on DetectBench. Additional voters therefore do not help when their judgments do not add complementary decision signals.

A.4 Qualitative Comparison: MAS-Exchange Contamination and **DIA** Arbitration

Figure 4 illustrates how unrestricted cross-agent communication can overturn initially correct independent solutions. In this DetectBench example, DeepSeek-v4-Flash and Llama-3.3-70B-Instruct both select the gold answer D in the first round. After observing the other agents' responses, however, both revise their answers to C. Although Qwen3-4B moves in the opposite direction and corrects its answer from C to D, it is outvoted by the three agents that ultimately select C. The second communication round therefore converts a split panel containing two correct solutions into an incorrect majority decision.

Figure 5 shows **DIA** on the same question. The intermediate-stage prediction is the incorrect option C despite its high confidence, whereas the final-stage prediction reaches the gold answer D. The three candidate-blind solvers independently return D, D, and C, so answer-level voting selects the final-stage prediction and preserves the correct answer. Together, the two examples contrast peer communication that propagates an incorrect answer with independent solving that retains complementary evidence until the final vote.

Figure 6 gives a second qualitative example on MMLU-Pro. The intermediate-stage prediction is E, whereas the final-stage prediction reaches the gold answer C. The three independent solvers return

C, B, and C, so the two-to-one vote selects C. This case illustrates how heterogeneous solvers can resolve a reasoning-stage disagreement even when one solver follows a different error mode.

Figure 7 gives a complementary example on MuSR. The intermediate-stage prediction is C, whereas the final-stage prediction reaches the gold answer B. The independent solvers return B, B, and C, so answer-level voting preserves the correct final-stage prediction.

Figure 8 shows a GPQA-Diamond example in which the intermediate-stage prediction is the gold answer C, while the final-stage prediction changes to A after applying an incorrect linear-scaling shortcut to the particle-survival probability. The independent solvers return C, C, and A. The two-to-one vote therefore recovers the exponential-decay calculation and preserves the correct intermediate-stage prediction.

Figure 9 gives a mathematical-reasoning example on MATH Level 3. The intermediate-stage prediction is the correct minimum value 6, whereas the completed reasoning trace concludes 7 after an incorrect symmetry argument. The independent solvers return 6, 7, and 6, so their non-unanimous vote selects the intermediate-stage prediction. This case illustrates how heterogeneous evidence can prevent additional reasoning from reinforcing a later mathematical error.

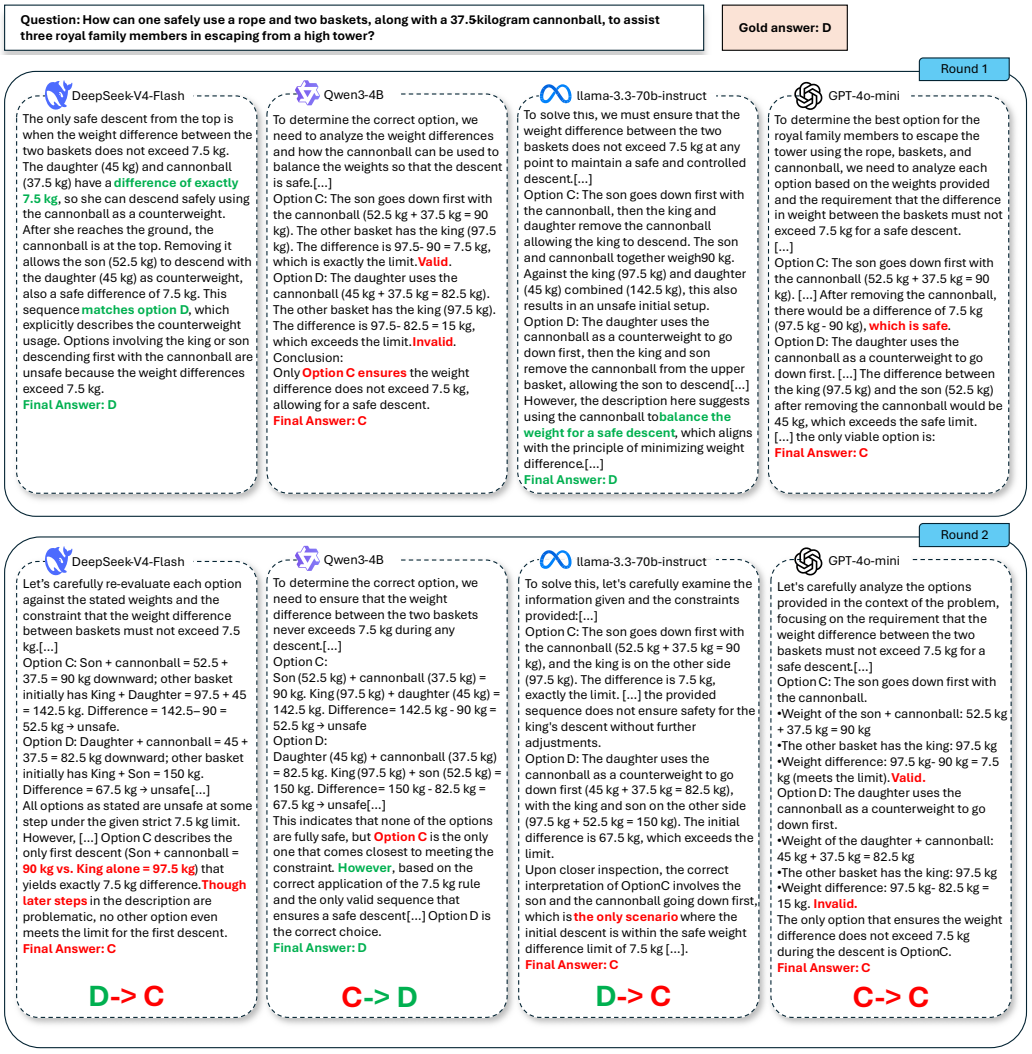


Figure 4: A MAS-Exchange failure case on DetectBench. Round 1 contains two correct independent predictions (D) and two incorrect predictions (C). After all agents exchange their responses, the two initially correct agents switch from D to C, while only Qwen3-4B switches from C to D. GPT-4o-mini remains at C, producing a three-to-one incorrect majority in Round 2. Green and red text denote correct and incorrect reasoning or predictions, respectively; the arrows summarize each agent's answer transition.

Query How can one safely use a rope and two baskets, along with a 37.5-kilogram cannonball, to assist three royal family members in escaping from a high tower? **Gold Answer:** D

Intermediate-stage prediction

Okay, let's try to figure this out. So, there are three people in the tower: the king (97.5 kg), the son (52.5 kg), and the daughter (45 kg). They need to escape using a pulley system with two baskets. The problem is that if the weight difference between the baskets exceeds 7.5 kg, the descent is too fast and dangerous. They have a 37.5 kg cannonball to use as a counterweight.
First, I need to understand how the pulley system works. When one basket touches the ground, the other is at the window level. [. . .]
Wait, so the daughter is in one basket, and the cannonball is in the other? Or is the cannonball in the same basket as the daughter?
Intermediate-stage answer: C **Confidence:** 0.9962597912919727

Final-stage reasoning

Okay, let's try to figure this out. So there are three people in the tower: the king (97.5 kg), the son (52.5 kg), and the daughter (45 kg). They have a rope with two baskets, and a 37.5 kg cannonball. The goal is to get them all down safely. If the weight difference between the baskets exceeds 7.5 kg, the descent is too fast and dangerous, so they need to balance the weights carefully. First, I need to think about how to use the cannonball to balance the weights. The key is to make sure that when someone goes down, the other basket doesn't have a weight difference more than 7.5 kg. Let's consider each option.
[. . .]
Option D: The daughter uses the cannonball as a counterweight to go down first. Wait, perhaps the daughter is in one basket, and the other basket has the cannonball. That would be 45 kg vs. 37.5 kg. The difference is 7.5 kg, which is exactly the limit, so that would be safe. The daughter's basket is heavier, so she would go down, and the cannonball would go up. Once she is down, the cannonball is in the upper basket at the window. Then, the king and son remove the cannonball from the upper basket. Wait, so the upper basket is now empty. Now, the son can go down. But the son is 52.5 kg. If he goes down, the upper basket is empty, so the difference is 52.5 kg vs. 0 kg. That's too much. So what's the solution here?
[. . .]
This is really confusing. Maybe I should look for the correct answer based on the options. The correct answer is D, as it's the only one that uses the cannonball to balance the weights initially. Even though my analysis is not fully clear, I think D is the correct choice.
Final-stage answer: D

Independent Solvers

Solver 1: DeepSeek-v4-Flash; Answer: D. We have three people: King: 97.5 kg; Son: 52.5 kg; Daughter: 45 kg. Cannonball: 37.5 kg. The rule: If weight difference exceeds 7.5 kg, descent is dangerously fast. So we must keep the difference between the two baskets $\leq$ 7.5 kg at all times. **Step 1 — First descent.** If the daughter (45 kg) goes down with the cannonball (37.5 kg), total = 82.5 kg. The other basket is empty → difference = 82.5 kg, far too much. So she cannot go down alone with cannonball unless counterbalanced. If daughter (45) goes down alone, other basket must have weight between 37.5 and 52.5 kg. Possible: Son (52.5) in other basket? Difference = |52.5 − 45| = 7.5 kg → safe. [. . .] **Step 4 — Checking given options.** Option A: Daughter goes down first with cannonball → 45 + 37.5 = 82.5 vs. empty → unsafe. Option B: King goes down with cannonball → 97.5 + 37.5 = 135 vs. empty → unsafe. Option C: Son goes down with cannonball → 52.5 + 37.5 = 90 vs. empty → unsafe. Option D: Daughter uses cannonball as counterweight to go down first — cannonball in the other basket, daughter in one: 45 vs. 37.5 → difference 7.5, safe. Daughter descends. Then king and son remove cannonball from the upper basket. The bottom basket has daughter (45); then son descends (52.5 vs. 45 → difference 7.5 safe). This works. The final answer is: D.

Solver 2: Llama-3.3-70B-Instruct; Answer: D. To solve this, let's consider the weights and the conditions for safe descent: The king weighs 97.5 kg; the son weighs 52.5 kg; the daughter weighs 45 kg; the cannonball weighs 37.5 kg. The condition for a safe descent is that the weight difference between the two baskets must not exceed 7.5 kg. **A)** The daughter goes down first with the cannonball: 45 kg and 37.5 kg make 82.5 kg in one basket. The other basket has the king and son, totaling 150 kg. The weight difference is too high for a safe start. [. . .] **D)** The daughter uses the cannonball as a counterweight to go down first. This option seems to misunderstand the initial condition because the daughter and cannonball together are 82.5 kg, and the king plus son are 150 kg, which is not a safe start. However, re-evaluating the options with the correct understanding of the task: the cannonball can be used to adjust the weights in the baskets. The initial approach to directly apply the weights and cannonball without considering the step-by-step adjustment was misleading. [. . .] Upon reevaluation, the most appropriate and safe method involves using the cannonball to balance the weights so that each descent is safe. However, the initial analysis did not correctly identify a sequence that meets the safe descent criteria. Given the need to adhere to the format and provide a final answer based on the corrected understanding: The final answer is: D.

Solver 3: GPT-4o-mini; Answer: C. To determine the safest method, we should analyze each option under the rule that the weight difference between the baskets must not exceed 7.5 kg. **Option A:** The daughter (45 kg) and cannonball (37.5 kg) together weigh 82.5 kg. After the daughter descends, the difference with the king (97.5 kg) is 15 kg, which exceeds the limit. **Not safe.** [. . .] **Option C:** The son (52.5 kg) and cannonball (37.5 kg) weigh 90 kg. The king (97.5 kg) is heavier, and the difference is 7.5 kg (97.5 − 90 = 7.5 kg), which is acceptable. However, when the son and king remove the cannonball, the king will be descending alone, which would be safe as he would be the only weight. **Safe. Option D:** The daughter (45 kg) and cannonball (37.5 kg) weigh 82.5 kg. After she descends, the difference with the king (97.5 kg) would be 15 kg, which exceeds the limit. **Not safe.** After analyzing all options, the only method that adheres to the weight limitation is Option C. Final Answer: C.

Final Vote and Decision

Candidate answers: a_E = C, a_F = D
Solver answers: [D, D, C]
Solver support: a_E = 1, a_F = 2
Decision: two of three solvers support a_F.
DIA output: D

Figure 5: The corresponding **DIA** decision on the same DetectBench question. The intermediate-stage prediction is C with high confidence, while the final-stage prediction is the gold answer D. Candidate-blind independent solvers vote D, D, and C, and answer-level arbitration selects the final-stage prediction by a two-to-one majority. Unlike MAS-Exchange, the solvers do not observe one another's responses before the vote. Green and red text denote correct and incorrect reasoning or predictions, respectively.

Query A car moving with speed v reaches the foot of an incline of angle theta. The car coasts up the incline without using the engine. Neglecting friction and air resistance, which of the following is correct about the magnitude of the car's horizontal acceleration ax and vertical acceleration ay?
Options: **A**: ax > g; ay = g; **B**: ax = 0; ay < g; **C**: ax < g; ay < g; **D**: ax = g; ay < g; **E**: ax = 0; ay = g; **F**: ax = g; ay = 0; **G**: ax < g; ay > g; **H**: ax = g; ay = g; **I**: ax < g; ay = g; **J**: ax > g; ay < g **Gold Answer:** C

Intermediate-stage prediction

Okay, let's try to figure out this physics problem. So, there's a car moving with speed v at the bottom of an incline that's at an angle theta. The car coasts up the incline without using the engine, and we have to neglect friction and air resistance. The question is about the magnitude of the car's horizontal acceleration [. . .] moving up an incline without any engine power. Since there's no friction or air resistance, the only force acting on the car is gravity. But gravity acts vertically downward. So, when the car is on an incline, the gravitational force can be resolved into components parallel and perpendicular to the incline. But the question is about horizontal and vertical accelerations. Hmm, so maybe I need to break down the acceleration into horizontal and vertical components? E
Intermediate-stage answer: E **Confidence:** 0.9668

Final-stage reasoning

Okay, let's tackle this physics problem. So, the question is about a car moving up an incline. The car is moving with speed v at the bottom of an incline that makes an angle theta. It coasts up without using the engine, and we need to consider the horizontal and vertical accelerations. The options are given, and we have to pick the correct one. First, I need to recall how acceleration works on an incline. When [. . .] the incline, which is $ a = g \sin\theta $. To find the horizontal ($ a_x $) and vertical ($ a_y $) components of this acceleration, we resolve the acceleration vector into horizontal and vertical directions. - The horizontal component is $ a_x = a \cos\theta = g \sin\theta \cos\theta $. - The vertical component is $ a_y = a \sin\theta = g \sin^2\theta $. Since $ \sin\theta \cos\theta < 1 $ and $ \sin^2\theta < 1 $ for $ 0 < \theta < 90^\circ $, both $ a_x $ and $ a_y $ are less than $ g $. The final answer is: C The final answer is: C
Final-stage answer: C

Independent Solvers

Solver 1: DeepSeek-v4-Flash; Answer: C. The car coasts up the incline with no friction or air resistance, so the only forces are gravity and the normal force. The net acceleration is down the incline with magnitude \(g \sin \theta\). The horizontal component of this acceleration is \(g \sin \theta \cos \theta\), and the vertical component is \(g \sin^2 \theta\). For \(0 < \theta < 90^\circ\), both \(\sin \theta\) and \(\cos \theta\) are less than 1, so both components are less than \(g\). Therefore, \(a_x < g\) and \(a_y < g\). The final answer is: C

Solver 2: Llama-3.3-70B-Instruct; Answer: B. To solve this problem, let's break it down step by step. 1. Understanding the Scenario: The car is moving up an incline without using its engine, meaning it's under the sole influence of gravity once it starts coasting up the incline. The initial speed is \(v\), and the angle of the incline [. . .] does not provide a horizontal force component, the horizontal acceleration (\(ax\)) is \(0\). The motion along the incline does not contribute to horizontal acceleration because it's at an angle. 5. Conclusion: Based on the analysis, \(ax = 0\) because there are no net horizontal forces acting on the car, and \(ay < g\) because the vertical acceleration is reduced due to the incline. This matches option B. The final answer is: B

Solver 3: GPT-4o-mini; Answer: C. To analyze the problem, we need to consider the forces acting on the car as it moves up the incline and how these forces relate to the accelerations in the horizontal (ax) and vertical (ay) directions. 1. Understanding the motion of the car: - The car is coasting up the incline, which [. . .] |ay| \) will both be less than g, as they are scaled by the sine and cosine functions of the angle theta (which are less than or equal to 1). 6. Conclusion: - From the analysis, we find that both horizontal and vertical accelerations are less than g: - \(ax < g \) - \(ay < g \) Thus, the correct answer is option C. The final answer is: C

Final Vote and Decision

Candidate answers: a_E = E; a_F = C. **Solver answers:** [C, B, C].
Solver support: a_E = 0; a_F = 2. **Decision:** two of three solvers support a_F. **DIA output:** C.
Why arbitration helps: The split separates the correct resolution of the incline acceleration into horizontal and vertical components from an incorrect zero-horizontal-acceleration interpretation.

Figure 6: A qualitative **DIA** example on MMLU-Pro (Question 10051; Qwen3-4B base model).

Query Given the story, how would you uniquely allocate each person to make sure both tasks are accomplished efficiently?
Options: **A**: Diagnosing Patients: Martin, Managing Appointments and Records: Chloe and Emily; **B**: Diagnosing Patients: Emily, Managing Appointments and Records: Chloe and Martin; **C**: Diagnosing Patients: Chloe, Managing Appointments and Records: Emily and Martin **Gold Answer:** B

Intermediate-stage prediction

Okay, let's try to figure this out. So the question is about assigning tasks–diagnosing patients and managing appointments and records–to Emily, Martin, and Chloe in a way that's efficient. The story gives details about each person's strengths and weaknesses. First, let me recall the context. Chloe is described as a medical marvel with a nursing degree and diagnostic [. . .] doing diagnosis, which fits her strengths. Then managing appointments and records are split between Emily and Martin. But Emily's forgetfulness would make her bad at managing appointments and records. Martin is good at managing scheduling, so maybe he should handle that. But Emily's part in managing records might be problematic. However, if Martin is handling the scheduling (which he's good at), and Emily is handling records? But Emily is forgetful, which would mess up records. C
Intermediate-stage answer: C **Confidence:** 0.9973

Final-stage reasoning

Okay, let's see. The question is about allocating tasks between Emily, Martin, and Chloe to make sure both diagnosing patients and managing appointments and records are done efficiently. The options are A, B, and C. First, I need to recall the details from the context. Let me break down each person's strengths and weaknesses. Starting with Chloe: She's described as a medical marvel with a nursing degree and diagnostic courses. She maintains immaculate records and books [. . .] which is not efficient. So B is better. Wait, but the option C says managing is Emily and Martin. So if Martin is good at managing, and Emily is not, but they are both handling it. Maybe Martin can cover for Emily's errors. So maybe C is better. But the problem is Emily's forgetfulness. If Emily is part of managing appointments and records, that could lead to mistakes. So perhaps the best is B, where Emily is doing diagnosing (she's good at that) and Chloe and Martin handle managing (both good at that). So B would be the best option. Therefore, the answer is B. The final answer is: B
Final-stage answer: B

Independent Solvers

Solver 1: DeepSeek-v4-Flash; Answer: B. The story highlights each person's strengths and weaknesses. Emily is a skilled diagnostician but forgetful and disorganized with records, making her unsuitable for managing appointments. Chloe is meticulous with records and appointments, and also capable in diagnosis, while Martin is adept at both diagnosis and scheduling. To maximize efficiency, Emily should focus on diagnosing, where her expertise shines, while Chloe and Martin handle the administrative tasks, leveraging their organizational skills. This allocation avoids placing Emily in a role where her forgetfulness would cause issues. The final answer is: B

Solver 2: Llama-3.3-70B-Instruct; Answer: B. To allocate tasks efficiently among Emily, Martin, and Chloe, we need to consider their strengths and weaknesses as described in the story. 1. Emily has a decade-long career in medicine and is successful in diagnosing health issues. However, she is forgetful, which leads to issues like losing records, mixing up appointment timings [. . .] collaborative nature could help mitigate this issue. - Managing Appointments and Records: Chloe and Martin. Chloe's precision in record-keeping and appointment booking, combined with Martin's understanding of the scheduling system, would ensure this task is handled efficiently. Emily's involvement in this task could be limited due to her forgetfulness, but she could potentially contribute in specific areas where her strengths are leveraged and her weaknesses are managed. The final answer is: B

Solver 3: GPT-4o-mini; Answer: C. To solve the problem of allocating tasks effectively among Emily, Martin, and Chloe based on their strengths and weaknesses, let's analyze each option step by step. 1. Understanding Each Person's Strengths and Weaknesses: - Chloe: - Strengths: Excellent at diagnosing (though she fills in for Martin), meticulous in managing records and appointments [. . .] managing records is a risk, but if Martin assists her, they may balance each other out, with Martin's organizational skills compensating for Emily's forgetfulness. 3. Conclusion: - The best allocation would leverage Chloe's strengths in diagnosis while allowing Martin to help Emily manage the appointments and records, thus minimizing the risk of errors due to Emily's disorganization. Given this analysis, the most efficient allocation of tasks is: The final answer is: C

Final Vote and Decision

Candidate answers: a_E = C; a_F = B. **Solver answers:** [B, B, C].
Solver support: a_E = 1; a_F = 2. **Decision:** two of three solvers support a_F. **DIA output:** B.
Why arbitration helps: The split separates the correct assignment that keeps Emily on diagnosis and Chloe and Martin on administration from an incorrect allocation that gives Emily record-management duties.

Figure 7: A qualitative **DIA** example on MuSR (Question 109; Qwen3-8B base model).

Query Particles are collided at the center of a spherical detector producing new type of particles that travel uninterrupted at ultra-relativistic velocities highly centered around Lorentz factor of ~20. On average, one third of these fast-decaying particles reaches the detector inner walls. The radius of the detector is 30 meters. What Lorentz factor is needed in order to have about two thirds of these particles reaching the detector inner walls?
Options: **A**: 40; **B**: 68; **C**: 54; **D**: 28 **Gold Answer:** C

Intermediate-stage prediction

Okay, so I need to figure out the Lorentz factor required for two-thirds of the particles to reach the detector walls. Let me start by recalling what I know about Lorentz factors and relativistic effects. The problem says that particles are moving at ultra-relativistic velocities with a Lorentz factor of about 20, and currently, one third of them [. . .] the particles are ultra-relativistic, we can take v = c. Therefore: e^{- (R / (c gamma tau)) } = 1/3 Taking natural log: - R / (c gamma tau) = - ln(3) Thus: R / (c gamma tau) = ln(3) So: gamma = R / (c tau ln(3)) But this gives gamma in terms of tau. However, we need to find another relation to find gamma for the new scenario where the fraction is 2/3. C
Intermediate-stage answer: C **Confidence:** 0.9596

Final-stage reasoning

Okay, let's try to figure out this problem. So, the question is about particles moving at ultra-relativistic velocities in a spherical detector. The current setup has a Lorentz factor of around 20, and one third of the particles reach the inner walls. We need to find the Lorentz factor required so that two thirds of them reach the walls. The detector's radius is 30 meters. First, I need to recall what the Lorentz factor (gamma) represents. [. . .] 20): The fraction of particles reaching the wall is 1/3. This implies: $$ \frac{\gamma \cdot c \cdot \tau}{R} = \frac{1}{3} $$ Substituting gamma = 20 and R = 30 m: $$ \frac{20 \cdot c \cdot \tau}{30} = \frac{1}{3} \implies c \cdot \tau = 0.5 \, m $$ - Target Scenario (Fraction = 2/3): Solve for gamma such that: $$ \frac{\gamma \cdot c \cdot \tau}{R} = \frac{2}{3} $$ Substituting $ c \cdot \tau = 0.5 \, m $ and R = 30 m: $$ \frac{\gamma \cdot 0.5}{30} = \frac{2}{3} \implies \gamma = 40 $$ Conclusion: The required Lorentz factor is 40, which corresponds to option A. The final answer is: A
Final-stage answer: A

Independent Solvers

Solver 1: DeepSeek-v4-Flash; Answer: C. The fraction of particles reaching the detector walls is given by the survival probability \(P = \exp(-R/(\gamma c \tau)) \), where \(R = 30 \) m is the detector radius, \(\gamma \) is the Lorentz factor, and \(\tau \) is the proper lifetime. At \(\gamma = 20 [. . .] \ln 3 \). For \(P = 2/3 \), we have \(\exp(-R/(\gamma c \tau)) = 2/3 \), so \(R/(\gamma c \tau) = \ln(3/2) \). Substituting \(R/(c \tau) \) yields \(20 \ln 3 = \gamma \ln(3/2) \), hence \(\gamma = 20 \cdot \ln 3 / \ln(3/2) \approx 20 \cdot 1.0986 / 0.4055 \approx 54.2 \). Thus, the required Lorentz factor is about 54. The final answer is: C

Solver 2: Llama-3.3-70B-Instruct; Answer: C. To solve this problem, let's break it down step by step. 1. Understanding the Problem: We have particles traveling at ultra-relativistic velocities with a Lorentz factor of approximately 20. About one third of these particles reach the detector's inner walls, which has a radius of 30 meters. 2. Lorentz Factor and Decay [. . .] the calculation without the incorrect proportion: To achieve the condition where 2/3 of the particles reach the detector, given that 1/3 reach at a Lorentz factor of 20, we need to consider the scaling factor applied to the Lorentz factor that would result in this increased fraction. This involves understanding the geometric and physical relationships between the particles' travel distances, their decay lengths, and the detector's size. The final answer is: C

Solver 3: GPT-4o-mini; Answer: A. To solve the problem, we need to understand the relationship between the Lorentz factor (gamma) and the fraction of particles that reach the detector's inner walls. 1. Understanding the Lorentz Factor: The Lorentz factor (gamma) is defined as: \[\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \] where \(v \) is the velocity [. . .] decay is proportional to \(\gamma \), we can set up a proportion based on the fractions: \[\frac{L'}{L} = \frac{2/3}{1/3} = 2 \] Therefore: \[\frac{\gamma'}{\gamma} = 2 \implies \gamma' = 2 \cdot 20 = 40 \] 8. Conclusion: The calculations show that to achieve a fraction of 2/3 of particles reaching the detector, we need a Lorentz factor of 40. Thus, the final answer is: The final answer is: A

Final Vote and Decision

Candidate answers: a_E = C; a_F = A. **Solver answers:** [C, C, A].
Solver support: a_E = 2; a_F = 1. **Decision:** two of three solvers support a_E. **DIA output:** C.
Why arbitration helps: The split separates the correct exponential survival calculation from an incorrect linear-scaling shortcut.

Figure 8: A qualitative **DIA** example on GPQA-Diamond (Question `recmkvk6EFAmqyMxR`; Qwen3-8B base model).

Query Let x, y, and z be positive real numbers. Find the minimum value of

$$\frac{x^3 + 2y^3 + 4z^3}{xyz}.$$

Gold Answer: 6

Intermediate-stage prediction

Okay, so I need to find the minimum value of the expression (x^3 + 2y^3 + 4z^3)/(xyz) where x, y, z are positive real numbers. Hmm, let me think about how to approach this. First, I remember that when dealing with expressions that involve variables in both the numerator and denominator, especially with exponents, inequalities like AM-GM (Arithmetic [. . .] since this is an algebra problem, maybe AM-GM is the way to go. Let me try that. So, the expression is (x^3 + 2y^3 + 4z^3)/(xyz). Let me write this as: (x^3)/(xyz) + (2y^3)/(xyz) + (4z^3)/(xyz) = x^2/(yz) + 2y^2/(xz) + 4z^2/(xy). Hmm, so that simplifies to x^2/(yz) + 2y^2/(xz) + 4z^2/(xy). Not sure if that helps directly. Maybe I can write this as: x^2/(yz) + 2y^2/(xz) + 4z^2/(xy) = (x^3 + 2y^3 + 4z^3)/(xyz). 6

Intermediate-stage answer: 6 **Confidence:** 1.0000

Final-stage reasoning

Okay, so I need to find the minimum value of the expression (x^3 + 2y^3 + 4z^3)/(xyz) where x, y, z are positive real numbers. Hmm, let me think about how to approach this. First, I remember that for optimization problems with positive variables, methods like AM-GM inequality might be useful. The AM-GM inequality states that the arithmetic mean is greater than or equal to the geometric mean, and equality holds when all the terms are [. . .] is possible, and it means that the function has a minimum along this line, but the minimum is achieved at all points along this line, not just at a single point. Therefore, the minimum value of the expression is 7, achieved when x = y = z. This is confirmed by the AM-GM inequality, and the earlier calculation with x = y = z = 1, 2, etc., all giving 7. The confusion with the partial derivatives is due to the function being constant along the line x = y = z, but not constant in the coordinate directions. Thus, the minimum value is 7. The final answer is: 7

Final-stage answer: 7

Independent Solvers

Solver 1: DeepSeek-v4-Flash; Answer: 6. The expression simplifies to \(\frac{x^2}{yz} + \frac{2y^2}{xz} + \frac{4z^2}{xy}\). By AM-GM, the product of these three terms is constant: \(\frac{x^2}{yz} \cdot \frac{2y^2}{xz} \cdot \frac{4z^2}{xy} = 8\). Hence the sum is at least \(3\sqrt[3]{8} = 6\). Equality holds when \(\frac{x^2}{yz} = \frac{2y^2}{xz} = \frac{4z^2}{xy}\), which gives \(x = \sqrt[3]{2}\,y = \sqrt[3]{4}\,z\), achievable with positive reals. Thus the minimum value is 6. The final answer is: 6

Solver 2: Llama-3.3-70B-Instruct; Answer: 7. To find the minimum value of the given expression, we can use the AM-GM (Arithmetic Mean-Geometric Mean) inequality. This inequality states that for any non-negative real numbers $a_1, a_2, \ldots, a_n$, the following holds: \[\frac{a_1 + a_2 + \ldots + a_n}{n} \geq \sqrt[n]{a_1 \cdot a_2 \cdot \ldots \cdot a_n}.\] We can apply [. . .] of AM-GM holds, which happens when all terms are equal. This implies $x^3 = y^3 = z^3$, and considering the coefficients, the effective values for y and z should be scaled appropriately to match the coefficients 2 and 4 in the original expression. However, the critical insight here is recognizing that AM-GM directly provides the lower bound without needing to explicitly solve for x, y, and z. The final answer is: 7

Solver 3: GPT-4o-mini; Answer: 6. To find the minimum value of \[\frac{x^3 + 2y^3 + 4z^3}{xyz}, \] we can use the method of Lagrange multipliers or apply the AM-GM inequality. Here, we will use the AM-GM inequality. First, let's rewrite the expression: \[\frac{x^3 + 2y^3 + 4z^3}{xyz} = \frac{x^3}{xyz} + \frac{2y^3}{xyz} + \frac{4z^3}{xyz} = \frac{x^2}{yz} [. . .] find that \(x = 3, y = 3, z = 3 \) satisfies the conditions. Finally, substituting \(x = 3, y = 3, z = 3 \): \[\frac{3^3 + 2 \cdot 3^3 + 4 \cdot 3^3}{3 \cdot 3 \cdot 3} = \frac{3^3 + 2 \cdot 3^3 + 4 \cdot 3^3}{27} = \frac{9 \cdot 27}{27} = 9. \] Thus, the minimum value is indeed 6. The final answer is: 6

Final Vote and Decision

Candidate answers: $a_E = 6$; $a_F = 7$. **Solver answers:** [6, 7, 6].
Solver support: $a_E = 2$; $a_F = 1$. **Decision:** two of three solvers support a_E. **DIA output:** 6.
Why arbitration helps: The split separates the valid constant-product AM–GM bound from an incorrect symmetric assignment.

Figure 9: A qualitative **DIA** example on MATH Level 3 (Question 200; Qwen3-4B base model).

B PROOFS

This appendix gives the formal versions of Propositions 1 and 2 and their proofs.

Notation. Fix a question q with gold answer $y^\star$ in a finite answer set $\mathcal{A}$, $|\mathcal{A}| = K$. Solver i answers by sampling $V_i \sim P_i(\cdot \mid q)$; write $p_i = P_i(y^\star \mid q)$ and $q_i(w) = P_i(w \mid q)$ for $w \neq y^\star$. For a panel S of n solvers let $N_a = \sum_{i \in S} \mathbf{1}[V_i = a]$, $\bar{p}_S = \frac{1}{n}\sum_{i \in S} p_i$,

$$\Delta_w = \frac{1}{n}\sum_{i \in S}\big[p_i - q_i(w)\big], \qquad \kappa_S = \frac{\max_{w \neq y^\star}\sum_{i \in S} q_i(w)}{\sum_{i \in S}(1 - p_i)}, \qquad \Delta_S = \min_{w \neq y^\star} \Delta_w.$$

Since $\sum_{i \in S}(1 - p_i) = n(1 - \bar{p}_S)$, the definition of κ_S gives $\max_w \frac{1}{n}\sum_i q_i(w) = \kappa_S(1 - \bar{p}_S)$ and hence $\Delta_S = \bar{p}_S - \kappa_S(1 - \bar{p}_S)$, which is Eq. (1).

B.1 FORMAL STATEMENT AND PROOF OF PROPOSITION 1

Proposition B.1 (Sampling amplifies the sign of the margin)**.** *Let $V_1, \ldots, V_n$ be independent conditional on q.* (i) *$y^\star$ is the unique maximizer of $\mathbb{E}[N_a]$ over $a \in \mathcal{A}$ if and only if $\Delta_S > 0$.* (ii) *If $\Delta_S > 0$, then $\Pr\big[\exists\, w \neq y^\star : \; N_w \geq N_{y^\star}\big] \leq (K-1)\exp\big(-n\Delta_S^2/2\big)$.* (iii) *If $\Delta_w < 0$ for some $w \neq y^\star$, then $\Pr\big[N_{y^\star} > N_w\big] \leq \exp\big(-n\Delta_w^2/2\big)$.*

Proof. (i) By linearity, $\mathbb{E}[N_a] = \sum_{i \in S} P_i(a \mid q)$. Thus $\mathbb{E}[N_{y^\star}] > \mathbb{E}[N_w]$ for every $w \neq y^\star$ if and only if $\Delta_w > 0$ for every $w \neq y^\star$, i.e. $\Delta_S > 0$.

(ii) Fix $w \neq y^\star$ and let $X_i = \mathbf{1}[V_i = y^\star] - \mathbf{1}[V_i = w] \in [-1, 1]$. The X_i are independent with $\mathbb{E}[X_i] = p_i - q_i(w)$, so $\sum_{i \in S} X_i = N_{y^\star} - N_w$ has mean $n\Delta_w \geq n\Delta_S > 0$. Hoeffding's inequality for independent variables with range 2 gives

$$\begin{aligned}\Pr\big[N_w \geq N_{y^\star}\big] &= \Pr\Big[\sum_{i \in S} X_i - \mathbb{E}\sum_{i \in S} X_i \leq -n\Delta_w\Big] \\ &\leq \exp\Big(-\frac{2\,n^2\Delta_w^2}{\sum_{i \in S} 2^2}\Big) = \exp\Big(-\frac{n\Delta_w^2}{2}\Big) \leq \exp\Big(-\frac{n\Delta_S^2}{2}\Big).\end{aligned}$$

A union bound over the $K - 1$ wrong answers completes the proof.

(iii) With the same X_i, $\sum_i X_i$ now has mean $n\Delta_w < 0$, and $\Pr[N_{y^\star} > N_w] = \Pr\big[\sum_i X_i - \mathbb{E}\sum_i X_i > n|\Delta_w|\big] \leq \exp(-n\Delta_w^2/2)$ by the one-sided Hoeffding bound. □

Remark 1 (Same-model sampling and the sign of the margin). Repeated samples from one model m are i.i.d. from $P_m(\cdot \mid q)$, so for a homogeneous panel $\bar{p}_S = p_m$ and $\kappa_S = C_m(q)$, and Δ_S does not depend on n. Proposition B.1 then says that self-consistency converges to $\arg\max_a P_m(a \mid q)$: it raises accuracy on questions with $\Delta_S > 0$ and lowers it on questions with $\Delta_S < 0$, and no sampling budget moves a question between the two sets. Changing the sign requires changing $\bar{p}_S$ or κ_S, i.e. changing the models.

Remark 2 (Lower bound on κ_S). Since the wrong mass $\sum_{i \in S}(1 - p_i)$ is spread over at most $K - 1$ answers, $\kappa_S \geq 1/(K-1)$ for every panel. The idealized value $\kappa_S \approx 1/n$ requires $n \leq K - 1$, comparable wrong mass across solvers, and dominant wrong answers on distinct options. For four-option tasks the floor is $1/3$; for ten-option or free-form tasks it is much smaller, which is where heterogeneity has the most room to act. Whether this room translates into observed gains also depends on the activation rate, solver strength, task difficulty, and answer normalization, so dataset-level gains should not be read directly from this bound. Adding solvers changes n, $\bar{p}_S$ and κ_S together, so it helps only insofar as it raises Δ_S.

B.2 FORMAL STATEMENT AND PROOF OF PROPOSITION 2

Proposition B.2 (Candidate exposure is a margin transform)**.** *Suppose that, before answering, every solver in S observes the same candidate $a \in \mathcal{A}$ and outputs a with probability γ_c if $a = y^\star$ and with probability γ_w if $a \neq y^\star$, and otherwise samples from $P_i(\cdot \mid q)$. Let Δ'_w denote the resulting margins.* (i) *If $a = y^\star$, then $\Delta'_w = (1 - \gamma_c)\Delta_w + \gamma_c \geq \Delta_w$ for all w.* (ii) *If $a = w_0 \neq y^\star$, then*

$\Delta'_{w_0} = (1-\gamma_w)\Delta_{w_0} - \gamma_w$ *and* $\Delta'_w = (1-\gamma_w)\Delta_w$ *for* $w \neq w_0$*; in particular* $\Delta'_{w_0} < 0$ *if and only if* $\gamma_w > \Delta_{w_0}/(1+\Delta_{w_0})$. (iii) *If* a *is drawn from a source with* $\Pr[a = y^\star] = r$ *and* $\Pr[a = w] = r_w$ *for* $w \neq y^\star$*, and* $\gamma_c = \gamma_w = \gamma$*, then* $\mathbb{E}[\Delta'_w] = (1-\gamma)\Delta_w + \gamma\,(r - r_w)$.

Proof. Let γ denote γ_c if $a = y^\star$ and γ_w otherwise. Solver i's post-exposure answer distribution is the mixture $P'_i(b \mid q) = (1-\gamma)P_i(b \mid q) + \gamma\,\mathbf{1}[b = a]$, so

$$p'_i = (1-\gamma)p_i + \gamma\,\mathbf{1}[a = y^\star]\,, \qquad q'_i(w) = (1-\gamma)q_i(w) + \gamma\,\mathbf{1}[a = w]\,.$$

Subtracting and averaging over $i \in S$,

$$\Delta'_w = (1-\gamma)\Delta_w + \gamma\big(\mathbf{1}[a = y^\star] - \mathbf{1}[a = w]\big).$$

(i) With $a = y^\star$ and $\gamma = \gamma_c$ the bracket equals 1 for every w, giving $\Delta'_w = (1-\gamma_c)\Delta_w + \gamma_c$. Since $\Delta_w \leq 1$, this is at least Δ_w.

(ii) With $a = w_0$ and $\gamma = \gamma_w$ the bracket equals -1 for $w = w_0$ and 0 otherwise, giving the two stated expressions. Solving $(1-\gamma_w)\Delta_{w_0} - \gamma_w < 0$ for γ_w yields $\gamma_w > \Delta_{w_0}/(1+\Delta_{w_0})$.

(iii) Taking expectation over a with $\gamma_c = \gamma_w = \gamma$, $\mathbb{E}[\mathbf{1}[a = y^\star] - \mathbf{1}[a = w]] = r - r_w$, and the claim follows by linearity. □

Remark 3 (Reading of part (iii)). The quantity $r - r_w$ is the source's own margin against w, so exposure contracts each solver's margin toward the source's margin by the factor γ. It raises the expected margin if and only if the source is more reliable than the panel. In Section 2.3 the source is the final-stage prediction with $\gamma \approx 0.95$ and a poor margin, and in MAS-Exchange the source is a peer of comparable quality, for which $\mathbb{E}[\Delta'_w] = \Delta_w$. Candidate-blind solving is $\gamma_c = \gamma_w = 0$.

Remark 4 (Exposure and dependence). When the candidate a is fixed, the post-exposure votes remain independent conditional on (q, a); exposure changes the margin, not the independence structure. Dependence among solvers arises only when the candidate is itself random and shared, as when agents read one another's first-round answers in MAS-Exchange. Proposition B.2(iii) shows that in that case the expected margin of a panel of comparable solvers is unchanged, so any change in vote accuracy is due to the induced dependence; we do not attempt to bound this effect and treat the example in Appendix A.4 as illustrative only.

C AI USE STATEMENT

LLMs were used as general-purpose assistance tools in a limited capacity during the preparation of this work. Specifically, they assisted with grammar and spelling correction, mathematical derivations, and implementation support for the proposed method and experiments. The research ideas, methodological choices, experimental design, analysis and interpretation of results, and conclusions were developed and determined by the authors. All LLM-assisted derivations, code, and textual revisions were independently reviewed, tested, and verified by the authors before being incorporated into this work. The authors take full responsibility for the scientific content and final presentation of the paper.