

Peer Review

Review of: "Diversity of Thought Elicits Stronger Reasoning Capabilities in Multi-Agent Debate Frameworks"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper extends Du et al.'s multi-agent debate protocol (ref. [1]) to heterogeneous model line-ups and argues that **who debates matters more than how large the debaters are**. Three debating LLMs answer a mathematical word problem; after every round, a fourth model — Gemini-Pro in every experiment (§3, "Response Summarization") — condenses the three answers and feeds the summary back as context for the next round. The central empirical claims are: (a) on GSM-8K, a diverse trio (Gemini-Pro, PaLM 2-M, Mixtral 8x7B, individually 78%, 64%, and 70% in §4.2) reaches **91%** after 4 rounds, while a homogeneous trio of three Gemini-Pro instances reaches only **80%** without CoT and **82%** with zero-shot CoT (§4.2, Figure 2), so the diverse arm beats GPT-4; (b) on ASDiv, a diverse trio (Gemini Flash 1.5, Gemini Pro, GPT 3.5, individually 89%, 86%, 81%) reaches **94%**, described as state of the art over ref. [20] (§4.2, Figure 1); (c) on MATH, the same trio (individually 55%, 32%, 33%) beats GPT-4 by 24 points and Gemini Ultra by 14 points (§4.2, Figure 8); (d) the benefit is scale-independent, holding from 0.75B to 100B+ (§4.1, Figure 5) and reproducing at 7B (+17% by round 4, Figure 6) and 2B (+10%, Figure 7); and (e) an appendix observation that mixing capacities produces "teacher-student" dynamics in which the strongest member *loses* accuracy (Appendix A.1, Figure 9a: 69% at round 0 degrading to 66%), leading the author to restate the requirement as diverse architectures *of similar size*. The framing conclusion is that diverse cooperating agents yield emergent capability beyond any single model.

2. Does the conclusion hold on the data?

The compute-budget confound is not addressed anywhere in the paper, and it is large enough to explain the headline result on its own. Per §3, one question consumes 3 debater calls plus 1 summarizer

call in every round. With round 0 plus 4 debate rounds (§4, “Baseline”; abstract, “after 4 rounds of debate”), that is at least $3 \times 5 = 15$ debater calls and 5 summarizer calls, i.e., **on the order of 16–20 generations per question**. The comparator in the abstract — GPT-4 on GSM-8K — is a single pass. The headline “91% vs GPT-4” is therefore a comparison at roughly 20:1 in inference cost, and the paper reports neither token counts, wall-clock, nor dollar cost anywhere. The claim as written is about an *ensemble-with-verification pipeline*, not about diversity.

Self-consistency at matched sample count is never used as a baseline. This is the decisive omission. The correct control for a 16–20-call system is majority vote over 16–20 samples from the single best member (Gemini-Pro, 78% in §4.2), and secondarily a summarizer-over-k-samples control that isolates the aggregator from the debate. Self-consistency on GSM-8K is well known to buy roughly the magnitude of gain claimed here, so without it, the reader cannot tell whether the 78%→91% movement (§4.2) is diversity, sampling, or verification-by-a-second-pass. The paper does cite chain-of-thought (ref. [8]) but no sampling-based aggregation baseline at all.

The baseline that is used is inconsistent with the paper’s own definition. §4 (“Baseline”) defines round 0 as each agent answering directly, i.e., a per-agent baseline. But the headline delta in §4.2 is stated as “rising from 78% to an impressive 91%,” where 78% is the accuracy of the *single strongest member* (Gemini-Pro, §4.2), not the round-0 output of the three-agent system. Appendix A.1 shows the author does compute a round-0 *aggregate* (“without any rounds of debate and simply by adopting the summary answer of the three models ... 69% accuracy,” Figure 9a). The two definitions of round 0 — per-agent in §4, aggregated in Appendix A.1 — are not the same object, and only the aggregated one is the honest debate-free control. The GSM-8K round-0 aggregate for the Gemini-Pro/PaLM 2-M/Mixtral trio is never reported, so **the actual effect attributable to debate (as opposed to aggregation) is not in the paper**.

The “diversity” contrast is not prompt-matched. In §4.2, the homogeneous arm is reported in two prompting conditions (80% without CoT, 82% with zero-shot CoT), while the diverse arm is reported as a single number (91%) with no CoT condition stated. The arms, therefore, differ in at least two variables simultaneously. Even taking the most favorable pairing for the author — 91% vs. 82% — the difference-of-differences is +13 points for the diverse arm against +4 points for the homogeneous arm, a 9-point gap that carries no uncertainty estimate.

Temperature and decoding are never specified, and this is fatal to the homogeneous control. Neither §3, §4, nor the figure captions state a temperature, top-p, or sampling policy for any model, nor whether the setting is identical across arms. If the three Gemini-Pro instances in the homogeneous arm were run

at or near greedy decoding, they would produce near-identical round-0 responses, and the “homogeneous debate” arm collapses into one sample repeated three times — a control with no diversity of any kind, rather than a control with diversity of the cheap kind. The only way to show that *model-family* diversity is what matters is to beat two cheaper diversity sources at the same call budget: (i) one model at elevated temperature with k samples, and (ii) one model with k distinct personas/prompts. Neither is run. Consequently, the paper’s own thesis — that diversity of thought is the driver — is unsupported; what is shown is that *some* source of variation beats *near-zero* variation.

Diversity is not operationalized. “Diversity of thought” is instantiated throughout as “different model families” (§4.2: “three models of similar capacity but featuring diverse model families”), but no measurement of it is offered: no round-0 pairwise answer-disagreement rate, no output entropy, no coverage/oracle-ceiling statistic. Without such a quantity, there is no independent variable, only a two-level categorical label, and the paper cannot distinguish diversity from a simpler confound visible in its own numbers: the GSM-8K trio spans 64–78% (a 14-point spread, §4.2) while the ASDiv trio spans 81–89% (an 8-point spread, §4.2), and the reported gains over the best member are correspondingly +13 points on GSM-8K and +5 points on ASDiv. Gain tracking headroom is at least as good an explanation as gain tracking diversity.

Seeds, repeated runs, and variance: none reported. There is no statement anywhere of how many times any configuration was run, no seeds, no error bars mentioned in any figure caption (Figures 1, 2, 5–9), no confidence intervals, and no significance test. Every claimed effect is, therefore, a single point estimate of unknown stability. The homogeneous 78%→80% movement (§4.2, no-CoT) is 2 points and is not distinguishable from run-to-run noise on any plausible sample size.

Benchmarks and splits are not pinned down, and this changes the verdict. §3.1 describes GSM-8K as “8.5K ... problems” (the full corpus, not the 1,319-problem test split), ASDiv as 2,306 problems, and MATH as 12,500 problems, but no section states how many items were actually evaluated, or whether ASDiv means the full set or the commonly used ASDiv-A subset. This is not a formality. Taking the headline pair 91% (diverse) vs 82% (homogeneous, zero-shot CoT) from §4.2: on a 100-item subset, the two-proportion test gives $z \approx 1.9$ ($p \approx 0.06$) — not significant; on the full 1,319-item GSM-8K test split, the same two proportions give $z \approx 6.8$ — decisively significant. **The paper’s central claim is significant or not significant depending on a number it does not report.** Since the protocol in ref. [1] is customarily evaluated on small subsets, the reader must assume the unfavorable case.

Two results contradict the headline and are not reconciled. First, Appendix A.1/Figure 9a reports that debate *reduces* accuracy from 69% to 66% for the Gemma 7B / Qwen1.5 7B / Mixtral 8x7B trio — a diverse trio by the paper’s own definition — and Figure 9b reproduces the degradation. The author rescues the thesis by adding a similar-size requirement, but the headline GSM-8K trio (Gemini-Pro, PaLM 2-M, Mixtral 8x7B, §4.2) is itself not size-matched: Mixtral 8x7B has ~46.7B total parameters, while PaLM 2-M and Gemini-Pro sizes are undisclosed, so the revised criterion cannot be checked against the very configuration that produces the 91%. Second, and arithmetically: §4.2 puts Mixtral 8x7B at 70% on GSM-8K standalone, yet Appendix A.1 states that the round-0 summary of a trio containing that same Mixtral “achieve[s] the best performance compared to the individual models at 69% accuracy” on the same benchmark. 69% is below 70%, so either the two Mixtral evaluations use different subsets or prompts (in which case that must be stated) or the appendix claim of “best performance” is incorrect. I checked both figures twice against §4.2 and Appendix A.1.

Comparator numbers are asserted without coordinates. The abstract and §4.2 claim the framework “outperforms GPT-4” on GSM-8K without ever printing GPT-4’s GSM-8K accuracy or its prompting condition, and the supporting citation attached to that sentence in §4.2 is ref. [19] (Bubeck et al., “Sparks of AGI”) rather than the GPT-4 technical report, ref. [4]. The ASDiv “new state-of-the-art ... surpassing the previous record set by [20]” (§4.2) never prints ATHENA’s ASDiv score, so the margin is unverifiable. On MATH (§4.2, Figure 8), neither the framework’s own absolute accuracy nor GPT-4’s nor Gemini Ultra’s is given; only the margins “24%” and “14%” appear, which jointly imply Gemini Ultra outscoring GPT-4 on MATH by exactly 10 points — an implied gap far larger than the published figures for those two systems, and one that no cited source in the reference list supports. As it stands, the MATH section contains no absolute result at all.

Contamination is not discussed. ASDiv (2021) and GSM-8K (2021) are long-public sets, and the 94% ASDiv claim is a SOTA claim made with 2023–2024 proprietary models of undisclosed training data. Some statement of the risk, ideally with a held-out or perturbed variant, is necessary before the word “state-of-the-art” is used.

On the evidence presented, the defensible conclusion is the weak one: *a multi-model pipeline with an aggregating summarizer, at ~16–20× the single-pass budget, outperforms a single pass and outperforms three near-deterministic copies of one model.* The strong conclusions in the abstract and §5 — that diversity of thought specifically is the driver, that the system beats GPT-4, and that this constitutes emergent capability — are not established by the reported measurements.

3. Reproducibility

An independent replication cannot be attempted from this manuscript. Specifically:

Model identity. No snapshot identifiers or access dates appear anywhere. §4.2 names “GPT 3.5” (no dated snapshot), “Gemini Pro” (1.0 or 1.5 unclear, and it is used both as a debater in the GSM-8K trio and as the summarizer for every experiment), “Gemini Flash 1.5”, “PaLM 2-M”, “Mistral 7B” (no version), “Llama 2 7B” (base or chat unstated), “Gemma 7B” and “Gemma 2B” (base or instruction-tuned unstated), “Qwen 2B” (no such release; presumably Qwen1.5-1.8B, and §Appendix A.1 writes “Qwen1.5 7B” for the 7B case) and “Rho 1B” listed inside a group labelled “2B models”. Mixtral is written three different ways: “Mixtral 7B×8” (abstract), “Mixtral 7x8B” (§4.2) and “Mixtral 7Bx8” (Appendix A.1); the released model is Mixtral 8x7B. Figure 5 spans “0.75B to 100B+” without naming a single model on that axis.

Decoding. No temperature, top-p, top-k, max-token limit, or stop condition is given for any model or any arm.

Seeds and repetitions. None given; no indication that any configuration was run more than once.

Evaluation set sizes. Not stated for GSM-8K, ASDiv, or MATH; the split (test vs. train) is not stated either (§3.1).

Prompts. §4 (“Evaluation methods”) says prompts were configured to emit a boxed answer, and §4 (“Baseline”) says identical starting prompts were used, but no prompt text appears in the paper or appendix — not the debater prompt, not the summarizer prompt, not the zero-shot CoT variant used for the 82% homogeneous result in §4.2.

Answer extraction. §3 gives two mutually exclusive aggregation rules in the same bullet (“Final Summary”): the summarizer’s final output is the system output, and also “here we can extract what the mode of the answers of the 3 models was”. Which rule produced 91% (§4.2), 94% (§4.2/Figure 1), and 69% (Appendix A.1) is not stated. The boxed-answer parser and its handling of parse failures are not described.

Transcripts. The Qualitative Result paragraph in §4.2 promises “More of this can be found in the appendix,” but Appendix A.1 contains only the teacher-student analysis and Figure 9 — no debate transcripts appear anywhere in the paper. §4 (“Evaluation methods”) says a per-round JSON of all model outputs is generated for every question; that artifact is exactly what a reader needs, and it is not released.

Code and data. No repository URL, no data-availability statement, no license, no note on API access dates. Nothing in the manuscript is executable or checkable by a third party.

Presentation of results. The paper contains no tables. Every quantitative result is either embedded in prose or must be eyeballed off Figures 1, 2, 5, 6, 7, 8, and 9, none of whose captions state *n*, error bars, or the aggregation rule. Several per-round values discussed in the text (for example, the 7B “+17%” and 2B “+10%” of §4.2) are given only as deltas, with neither endpoint printed.

4. What specifically to fix

1. **Add a budget-matched self-consistency baseline (§4, “Baseline”; new table).** For every debate configuration, report majority-vote accuracy over *k* samples of the single strongest member, with *k* equal to the debate’s total generation count (~16–20 per §3). Until 91% on GSM-8K (§4.2) is compared against Gemini-Pro self-consistency at matched *k*, the diversity claim has no control.
2. **Report the call budget explicitly (§3 and §4).** State generations per question, prompt and completion tokens, and cost/latency for every arm including the GPT-4 and Gemini Ultra comparators in §4.2. Remove or heavily qualify the abstract’s “outperforms GPT-4” unless the comparison is budget-matched.
3. **Add the two cheap-diversity controls (§4.2).** One model at elevated temperature with *k* samples, and one model with *k* distinct role/persona prompts, both at the debate’s call budget. The paper’s thesis requires model-family diversity to beat both; at present neither is run.
4. **Fix the baseline definition mismatch (§4 “Baseline” vs §4.2 vs Appendix A.1).** Report the round-0 *aggregated* accuracy for the Gemini-Pro / PaLM 2-M / Mixtral 8x7B trio on GSM-8K, in the same manner as the 69% round-0 aggregate of Figure 9a, and restate the headline delta against that number instead of against the 78% single-model Gemini-Pro figure.
5. **State decoding parameters and prove they are identical across arms (new §4 subsection).** Temperature, top-*p*, max tokens, stop conditions, per model and per arm. If the homogeneous three-Gemini-Pro arm of §4.2 was run near-greedy, say so and re-run it with sampling; as reported, it is not a fair control.
6. **Report seeds, repetitions, and uncertainty (all of §4, Figures 1, 2, 5–9).** At least 3 seeds per configuration, with means, standard deviations, and error bars on every figure, plus a paired significance test for the two comparisons the paper leans on: 91% vs 82% on GSM-8K (§4.2) and 94% vs 89% on ASDiv (§4.2).

7. **State evaluation set sizes and splits (§3.1).** Exact n for GSM-8K (test split vs the 8.5K full corpus as currently written), ASDiv (full 2,306 vs ASDiv-A), and MATH (which subjects/levels), plus the sampling method if subsets were used. As shown in section 2 above, the significance of the headline result flips between n=100 and n=1319.
8. **Prompt-match the CoT condition (§4.2).** The homogeneous arm is reported at 80% (no CoT) and 82% (zero-shot CoT) while the diverse 91% carries no stated CoT condition. Report both arms in both conditions in one table.
9. **Operationalize “diversity of thought” (§4.2).** Add a measured diversity statistic per configuration — round-0 pairwise answer-disagreement rate, and the oracle ceiling (fraction of questions at least one member answers correctly at round 0) — and plot final accuracy against it. This also tests the competing headroom explanation: the GSM-8K trio spans 64–78% and gains 13 points over its best member, while the ASDiv trio spans 81–89% and gains 5 points.
10. **Disambiguate the answer-extraction rule (§3, bullet “Final Summary”).** The bullet offers both “the final summarized response ... is considered as the output” and mode-of-three-answers. Pick one, state it, and report the other as an ablation; note which rule produced 91%, 94%, and 69%.
11. **Isolate the summarizer (§3, “Response Summarization”).** Gemini-Pro is the summarizer in all experiments and simultaneously a debater in the GSM-8K trio and the ASDiv/MATH trio, so the “diverse” arm contains four model families in the loop while the homogeneous arm contains one. Vary the summarizer (at minimum: a non-Gemini summarizer, and a rule-based majority vote with no summarizer model) so the aggregator’s contribution is separable from the debaters’.
12. **Supply absolute MATH numbers (§4.2, Figure 8).** Report the framework’s own MATH accuracy per round, and print GPT-4’s and Gemini Ultra’s MATH scores with citations and prompting conditions. The current text gives only margins of “24%” and “14%”, which imply a 10-point Gemini-Ultra-over-GPT-4 gap that no reference in the list supports.
13. **Substantiate the ASDiv SOTA claim (§4.2, Figure 1).** Print ATHENA’s (ref. [20]) ASDiv accuracy and the evaluation subset it used, and confirm the two numbers are computed on the same split; otherwise, drop “state-of-the-art”.
14. **Correct the MATH dataset citation (§3.1).** MATH is cited as ref. [16], which the reference list gives as Enderton, *A Mathematical Introduction to Logic* (2001). The intended citation is Hendrycks et al., “Measuring Mathematical Problem Solving with the MATH Dataset” (2021).
15. **Correct the GPT-4 citation in §4.2.** “This substantial improvement outperforms GPT-4[19]” cites Bubeck et al., “Sparks of AGI”; the GPT-4 technical report is ref. [4], and in any case, a specific GSM-

8K number and prompting condition must be quoted.

16. **Resolve the 69% vs 70% inconsistency (Appendix A.1 vs §4.2).** Appendix A.1 calls the 69% round-0 aggregate “the best performance compared to the individual models” on GSM-8K, while §4.2 reports Mixtral 8x7B alone at 70% on GSM-8K. Either report Mixtral’s standalone score inside the Appendix A.1 experiment, or correct the claim.
17. **Reconcile Figure 9a/9b with the headline configuration (Appendix A.1 and §4.2).** The revised criterion — “diverse architectures of similar sizes” — is not satisfied by the §4.2 headline trio, in which Mixtral 8x7B (~46.7B total parameters) sits beside two models of undisclosed size. Either report parameter counts for PaLM 2-M and Gemini-Pro or weaken the criterion, and state explicitly in §5 that debate degraded performance in two of the reported configurations.
18. **Quantify §4.1 (Figure 5).** The section asserts “similar performance gains across all model scales” from 0.75B to 100B+ without naming a single model, giving a single accuracy, stating the number of agents and rounds, or defining “similar”. Add the underlying table with model names, parameter counts, per-round accuracies, and the CoT/no-CoT condition. Note also that this claim sits uneasily with Figure 9a, where the largest member loses more than 20 points.
19. **Use percentage points consistently (§4.2, Appendix A.1).** “a performance increase of 17%” (7B, Figure 6), “a 10% improvement” (2B, Figure 7), “margins of 24% and 14%” (MATH), “decreased it’s performance by more than 20%” (Figure 9a) are all ambiguous between relative and absolute change; print both endpoints in each case.
20. **Add a master results table.** The manuscript currently has none. One table with rows = configuration (models, sizes, diverse/homogeneous, CoT condition), columns = benchmark × round 0–4 accuracy ± s.d., plus n, generations per question, and cost, would make every claim in §4.2 checkable and would remove the need to read values off Figures 1, 2, and 5–9.
21. **Fix figure/text cross-references.** (a) §4.2 attributes the homogeneous three-Gemini-Pro result (80%/82%) to Figure 2, whose caption reads “Diverse Model Debate Performance ... on the GSM-8K benchmark”; a homogeneous arm needs its own panel or an amended caption. (b) Figure 6 (“7B Diverse Models ...”) is placed inside the medium-capacity discussion but is first referenced several paragraphs later alongside Figure 7. (c) The Qualitative Result paragraph in §4.2 says Figure 4 shows Mixtral changing its answer at round 3, but Figure 4 is captioned “Illustration of the debate procedure”; the transcript figure is missing.
22. **Define the round indexing (§4 and all figure captions).** The captions say “Across 4 rounds” while the text describes round 0 plus 4 debate rounds; state whether the x-axis has 5 points and whether

“after 4 rounds” in the abstract includes round 0.

23. **Release artefacts.** Publish the per-round JSON described in §4 (“Evaluation methods”), the debater and summarizer prompts, the boxed-answer parser, and the evaluation script, with a data-availability statement and API access dates. Deliver the debate examples promised by “More of this can be found in the appendix” in §4.2.
24. **Add a contamination statement (§3.1).** GSM-8K and ASDiv predate all evaluated models; a SOTA claim of 94% on ASDiv (§4.2) should be accompanied by a perturbed or held-out check, or by an explicit acknowledgement of the limitation.
25. **Soften the framing in the abstract and §5.** “Emergent capabilities beyond even the most powerful individual models” and “revolutionize the field of AI” are not supported by single-run, non-budget-matched point estimates on undisclosed subset sizes. A copy-edit pass is also needed: “mutliagent” and “it’s” for “its” (Appendix A.1), “agenitc” (§5), “different architectures” agreement in §3.

5. Recommendation

Major revisions. The question is well chosen, and the teacher-student observation in Appendix A.1 (Figures 9a, 9b) is a genuinely interesting negative result that the field should see. But the central claim — that *diversity* drives the gain — currently rests on a comparison between a ~16–20-generation pipeline and both a single pass and a homogeneous arm whose decoding temperature is unstated and may be near-deterministic, with no self-consistency control at matched budget, no seeds, no variance, and no reported evaluation-set size. Items 1, 3, 4, 5, 6, and 7 of the list above are the load-bearing ones: with a budget-matched self-consistency baseline, a temperature-matched homogeneous arm, and *n* and error bars reported, this could become a solid contribution even if the effect shrinks. Without them, the reported numbers cannot distinguish the paper’s hypothesis from the cheaper alternatives. I would be glad to review a revised version.

Declarations

Potential competing interests: No potential competing interests to declare.