

When Should Multi-Agent Debate Stop? A Comparison of Fixed-Round and Adaptive Stopping

Avik Shekhar¹

Received July 28, 2026

Accepted October 5, 2026

Electronic access October 31, 2026

In a multi-agent debate, language-model agents discuss their responses before the system chooses a final answer. Each additional round requires more inference-time computation. We examined whether setting a confidence threshold would reduce unneeded debate rounds while maintaining answer accuracy. We tested fixed-round and adaptive schedules in the publicly available Du and Liang frameworks. Each condition used gpt-4o-mini and the same 100 questions sampled from QuALITY. We recorded accuracy, completed rounds, model calls, judge or moderator confidence, and paired question-level changes. The single-agent baseline answered 77 of the 100 questions correctly. Varying the Du thresholds affected computation more consistently than accuracy. None of the Du adaptive conditions clearly improved on the fixed-one independent ensemble. For Liang, we examined fixed schedules of one, two, and three rounds, together with a prompt-matched fixed-two control. This control helped separate stopping behavior from changes to the scored-output prompt. Its performance was similar to strict adaptive stopping, indicating that stopping alone could not explain the original accuracy difference. Later rounds sometimes corrected earlier errors but also introduced regressions. In these experiments, confidence-threshold stopping offered no consistent improvement in accuracy or measured computational efficiency over simpler fixed schedules. Adaptive controllers therefore need comparisons with matched fixed baselines across multiple thresholds, with prompt changes assessed separately from the stopping rule.

Keywords: multi-agent debate; large language models; adaptive stopping; fixed-round stopping; QuALITY; inference-time compute

Introduction

Debate systems that involve multiple agents are being developed to help Large Language Models (LLMs) reason better through interactive dialogue. This is done by providing the LLMs with the opportunity to create responses based upon a prompt, then have them collaborate on reviewing the responses that they created, explain the reasons behind the responses that they gave, and engage in discussions and debates as to why they believed that their responses were correct or incorrect. A final answer is then selected according to the framework's scoring procedure.

Research in this area has shown that interactions among agents can improve reasoning or factuality in some scenarios. However, the degree to which these improvements occur depends heavily upon the methods employed. Du et al. employed a method in which agents provided independent responses to a question and exchanged information related to why each agent responded to the question in a certain manner.¹ Liang et al. developed a system that employed adversarial roles and a moderator to promote divergent opinions.² Other studies have examined specialized agents for software

development and cross-model communication.^{3,4} Research on debate has investigated the relationship between persuasiveness and truthful answers,⁵ consistency among collaborating models,⁶ confidence-weighted consensus,⁷ and sparse communication between agents.⁸ Other work has examined adversarial attacks on collaborating agents,⁹ debate for event extraction,¹⁰ and multi-agent debate for evaluating generated text.¹¹ AutoGen, MetaGPT, and AgentVerse provide additional frameworks for organizing multi-agent collaboration.^{12,13,14} Smit et al. discovered that multi-agent debate did not consistently outperform simple self-consistency or ensembling methods across their entire benchmark suite.¹⁵ Because of the varying degrees of success experienced in past research studies, it is important to evaluate competing debate protocols against simple baselines rather than assuming that additional interaction will always result in superior performance.

While prior research has improved reasoning through chain-of-thought prompting,¹⁶ and through STaR (which fine-tuned on generated rationales leading to correct answers),¹⁷ Huang et al. evaluated intrinsic self-correction without external feedback and found that it can fail or decrease accuracy on reasoning tasks.¹⁸ Although prior research has studied both multi-agent debate and inference-time reasoning, this paper

¹ Milpitas High School, California, USA

evaluates one adaptive stopping rule based on confidence levels vs. fixed-round limits on the exact same QuALITY subset, utilizing the same model ID and temperature in two existing debate architectures. The contribution is this controlled comparison of stopping schedules.

Two different types of methods were employed to schedule debates. One type is referred to as “fixed-round stopping,” wherein the same number of rounds are used per question. The other type is referred to as an “adaptive condition.” This adaptive condition uses the model’s degree of confidence as a factor when deciding whether to stop debating after a round has been completed. The primary adaptive conditions utilized a confidence threshold of 90 on a 0–100 scale. Additionally, there were four different Du thresholds (70, 80, 90 and 95) utilized during a sensitivity sweep. In the Du framework, a separate judge provides the confidence values. In Liang, the moderator supplied confidence as part of its output. A round represents an iteration of the debate process. The round limit concerns these iterations, rather than the length of an individual response.

There is a trade-off between answer quality and inference-time computation. Longer dialogues may revise or correct answers, but they require additional model calls and may also introduce new errors. An effective stopping rule should therefore identify whether further discussion is likely to improve the answer, rather than relying only on whether the current answer appears confident.

Were the adaptive stopping rules based upon confidence thresholds different in terms of answer accuracy and round completion relative to their respective fixed-round baselines?

The purpose of this research study was to compare whether the tested adaptive stopping rule resulted in changes either in answer accuracy or completed-round counts relative to fixed-schedule counterparts in controlled experiments. Answer accuracy and completed-round comparisons were made primarily within each individual framework. A single-agent baseline was used to assess whether incorporating debate produced a detectable improvement in accuracy.

In addition, adaptive control mechanisms impose their own overheads; therefore, we assessed the adaptive conditions against simple fixed schedules. Furthermore, we separated the effects of the modified Liang moderator prompt and scored-output pathway from stopping behavior and analyzed whether the Du results varied by the selected confidence threshold.

Both publicly available debate frameworks were tested utilizing a 100-question subset of the QuALITY v1.0.1 train split using the gpt-4o-mini model identifier at temperature = 0.0. Each experimental condition contained 100 total questions across 64 articles. Each condition included identical questions presented in identical orders, and no training or fine-tuning occurred.

Methods

Research Design

This study utilized an inference-only experimental design. No models were trained or fine-tuned nor were any model weights altered. The focal point of interest for this study was the debate process itself including: (i) the structure of the debate agents, (ii) termination criteria, (iii) how responses generated by the agents are aggregated to produce a final answer, and (iv) how that final answer is evaluated. Each experimental condition received identical sets of evaluation problems presented in identical orders. The final answer provided by each condition was then translated into one of four possible labels (A, B, C or D), and then compared against the gold-labeled correct answer. Additionally, the total number of completed rounds was tallied for each debate condition.

Dataset and Evaluation Sample

QuALITY is a long-context multiple-choice reading-comprehension dataset in which each question is associated with a source passage, four alternative answers, and a gold label indicating which answer is correct.¹⁹ Therefore, it provides an objective target that can be represented as A/B/C/D for evaluating each system’s final answer.

In total, 100 questions were selected from the QuALITY v1.0.1 train split. The original sample preparation used selection seed 42. In addition, all experimental conditions utilized the identical 100 rows of the data in exactly the same order. Each question came from one of 64 unique source documents, and no more than six questions were taken from any single source document. Errors were dispersed across source articles; none of the primary conditions had more than three errors from a single source article. The reproducibility package lists question IDs, source indices, answer choices, gold labels, checksums and sample-selection details.

For added background information, Pang et al. stated that they achieved a test-set accuracy rate of 55.4% on their best published baseline, 43.3% on a question-only baseline, and 93.5% for human annotators.¹⁹ Since this study used a different model, a 100 item sub-sample of the training split, and fully contextualized source passages to evaluate QuALITY, these results cannot be accurately compared to those previously obtained by Pang et al.

GSM8K was initially examined during pilot work; however, GSM8K was eventually excluded from further analysis because the selected model performed near ceiling level on the pilot questions.²⁰ Consequently, there was little room to examine accuracy differences among stopping conditions.

The 100 selected questions comprised a single evaluation sample. The gold labels associated with the questions were never seen by the agents prior to answering them. After each

system provided its final A/B/C/D answer for each question, that answer was then compared to the corresponding gold-labeled answer and classified as either “correct” or “incorrect”.

Framework Architectures and Stopping Point

Figure 1 depicts both architectures of debate (the Du framework and the Liang framework), as well as what information was present in each round and where each adaptive termination check occurred.

Debate Frameworks

To compare the effects of different stop-schedules on each of the two architectures, the researchers used the composable-models `lmmultiagentdebate` implementation associated with Du and colleagues,^{1,21} as well as the Multi-Agents-Debate repository associated with Liang and colleagues.^{2,22} Because the two architectures have differences in terms of the roles of the agents, the flow of information during debating, and the way that the final answer is chosen, cross-framework comparisons are descriptive rather than causal.

The Du framework has three agents that provide answers to questions. In Round 1 all of the three agents respond separately. However, starting in Round 2, the agents are given the other agents’ prior responses and may alter their reasoning process. Therefore, in each complete Du round, there will be one response from each of the three answer generating agents or three model calls. The final answer is decided based on a majority vote among the three A/B/C/D labels that were extracted from the last completed round. In the adaptive condition, after the three agent calls in each round, there is a separate judge call used to evaluate the stopping criterion. That judge call adds extra overhead and does not replace the agents’ majority vote answer.

The Liang framework has an affirmative debater, a negative debater, and a moderator. If a run reaches its fixed limit before an acceptable moderator answer is found, it uses a separate final judge. Each completed round consists of an affirmative response, a negative response, and a moderator evaluation. The full QuALITY prompt is embedded in each debater’s persistent meta-prompt at initialization, so that each debater retains the passage in their conversation context throughout the rest of the rounds. The moderator’s candidate answer is taken from `debateanswer`; when the repository invokes its separate final judge at the limit, the scored output follows that final-judge pathway.

Therefore, these frameworks do not utilize computationally equivalent rounds. Counts of completed rounds are thus only relevant within each framework. Round 1 includes three independent agent calls in Du; however, all rounds after Round

1 provide each agent with the other agents’ answers from the previous round and include a repetition of the full question block. Every normal round in Liang includes calls to both debaters (affirmative and negative), as well as to the moderator. The primary Du comparison is adaptive stopping versus fixed one round, because 95 of 100 adaptive questions concluded after Round 1.

Repository Updates

Both repositories were updated to support the common QuALITY-task format. Common updates involved formatting passages, questions and four possible answers; extracting final A/B/C/D labels; providing options for fixed-round and confidence-threshold adaptive stopping; requesting and parsing numeric-confidence values in adaptive conditions; enabling batch processing; and writing logs at question level. The adaptive enhancements differed between frameworks. Most notably, Liang’s strict adaptive condition added an extension to the prompt that generates `debateanswer` in the moderator prompt. As such, the original fixed versus adaptive comparison altered stopping behavior, the moderator prompt, and the scored-output path.

The initial role assignments and overall flow of argumentation (debate) were kept. The implementation of Du was done using commit `9846749350eb917ae5bfaaff4c645fc705b8d3af` of `composable-models/lmmultiagentdebate`. The implementation of Liang was done using commit `e58d146033568ccf737d8031a93748597907d7d3` of `Multi-Agents-Debate`. Documentation of source code has the prompt templates, the logic to extract information, temperature settings and the maximum token behaviors used in this research.

Model and Generation Parameters

All reported runs employed the OpenAI API model identifier `gpt-4o-mini` and temperature 0.0.²³ Python’s local random seed was set to 0 by the orchestration scripts. No model-generation seed parameters were provided in the API calls. In the Du implementation, `topp` and `maxtokens` were not explicitly set. In Liang, `topp` was also unspecified, but `maxtokens` was capped at the minimum of 2,048 tokens and the amount of context available, with a floor of 256.

Conditions Tested

Seven total conditions existed in the main matrix: Single-Agent Baseline; Du Fixed One Round; Du Fixed Two Rounds; Du Stopping Criteria (Adaptive), Confidence Threshold = 90, Six-Round Cap; Liang Fixed Two Rounds; Liang Strict Stopping Criteria (Adaptive), Confidence Threshold = 90, Six-Round Cap; Liang Prompt-Matched Fixed-Two Control. Ad-

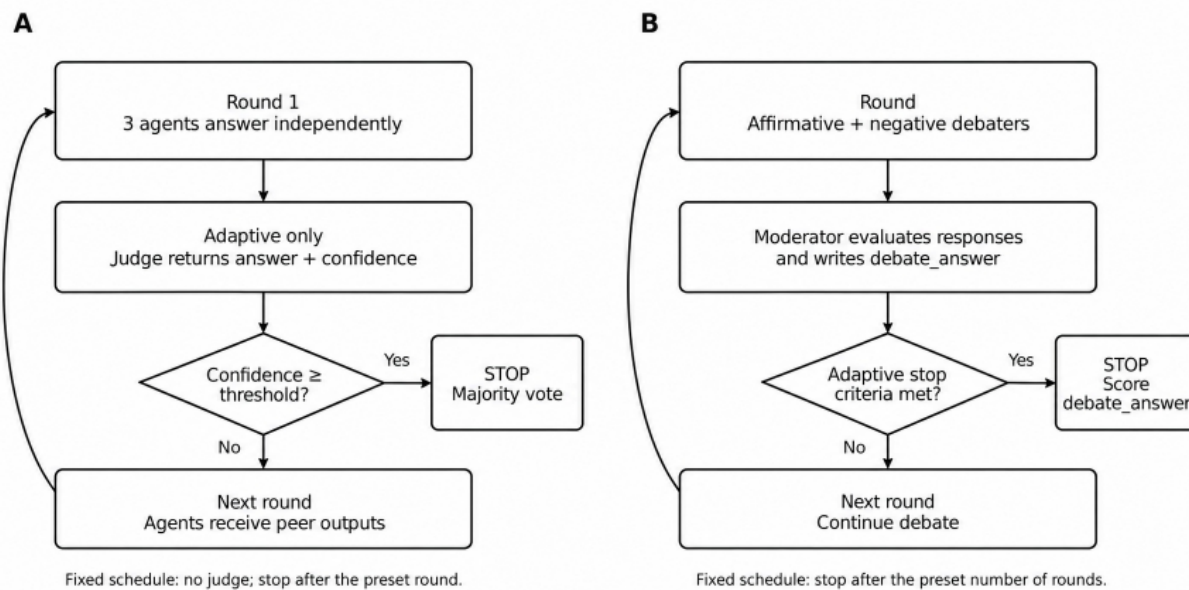


Figure. 1 Operational structures of (A) the Du framework and (B) the Liang framework. The diamonds mark the adaptive stopping checks; fixed schedules bypass the checks and stop at preset rounds.

ditionally, several sensitivity studies examined the following four Du thresholds (i.e., 70, 80, 90, and 95) as well as original-prompt Liang fixed-one and fixed-three schedules.

For all fixed-condition runs, the required number of rounds were completed. For adaptive runs, after each completed round, if their stopping criterion was met, the runs terminated. Otherwise, they ran until either their stopping criterion was met or the six-round cap was reached. The Liang prompt-matched fixed-two control enforced completion of exactly two rounds via use of the strict adaptive moderator prompt.

Sensitivity and Fixed-Round Bracketing

To assess how the stopping threshold influenced results, four Du adaptive runs were executed using the same 100 questions and three agents. Each of the four runs used the same model identifier, temperature, and six-round cap. Only the stopping threshold was varied across the four runs: 70, 80, 90, or 95. To provide a full range of sensitivity test data, the threshold-90 run was repeated in this batch.

Additionally, two other Liang fixed-schedule runs were executed to bracket the fixed-two schedule. Both fixed-one and fixed-three schedules used the original moderator prompt, disabled early-stopping, and retained the repository's typical final-judge processing. In addition to employing the same model identifier, temperature, and extraction process, these runs employed the same order of dataset presentation.

Single-Agent Baseline Condition

Each passage, question, and answer choice was presented to a single gpt-4o-mini model instance. Without any agent interaction or a stopping controller, this single-instance model was requested to select its best A/B/C/D answer choice. Based upon a uniform random distribution over the four possible choices, the expected accuracy is 0.25. A second simpler baseline exists within the Du conditions. Since the three Du agents never interact with one another during round one and since the final answer is selected based upon the majority vote of their responses in round one, Du fixed one round provides a three-sample ensemble of answers instead of providing interactive debate. This difference is important when evaluating whether debate provides added benefit over simply aggregating individual answers.^{15,24}

Stopping Criterion Used by the Judge in the Du Adaptive Condition

After each completed round in the Du adaptive condition, a separate language-model judge evaluated the three agent answers. The judge sent back two pieces of information: Answer and Confidence. Confidence was reported on a scale of 0 through 100. If the parsed confidence level met or exceeded 90, the process terminated. Below 90, the next round took place unless six rounds had already been completed. The judge's Answer did not become the system's final answer.

That answer was still decided by a majority vote among the three agents' answers from the last completed round.

Confidence parsing was successful for all 109 judge evaluations in the archived Du adaptive run. Those 109 evaluations equate to the 109 rounds that were finished. Therefore, the total amount of model calls in that run was the 109 judge calls added to the 327 agent calls (answer generating agents) which amounts to 436 total calls. That is compared to 300 agent calls in Du fixed one round and 600 in Du fixed two rounds.

Liang Strict Adaptive Stopping Criterion and Prompt-Matched Fixed-Two Control

Stopping early in the Liang strict adaptive condition was based upon three variables: there was a preferred option given by the moderator, debateanswer was non-empty, and the moderator's confidence was at least 90 on the 0–100 scale. Based upon pilot tests, the rule was tightened so that the debate would not finish prior to completing at least one full round.

The Liang strict adaptive moderator prompt was changed so that it asked for numerical confidence and left debateanswer empty if the moderator thought additional rounds may help improve the answer. Because debateanswer is evaluated for accuracy, this prompt change affects both the generation of the scored answer and how long a debate continues.

To separate the prompt change from stopping behavior, we created a Liang prompt matched fixed two control. In this control, we used the same modified moderator prompts as the strict adaptive condition but forced exactly two rounds to occur. Additionally, we directly scored the moderator's round-2 debateanswer. We were able to complete all 100 questions in this control with two rounds completed. There was one final debateanswer that was blank and therefore scored as incorrect.

There were 100 Liang adaptive questions answered. Eighty-one of them stopped prior to reaching the six-round cap. Nineteen of them reached the cap. Confidence parsing successfully parsed all 270 moderator evaluations.

Answer Extraction and Scoring

Answer extraction and scoring were done through scripted methods. The Du framework utilized a regular-expression-based technique to extract a single answer label from each of the agent responses. The predicted answer was the most frequently occurring label extracted from the agents' responses at the conclusion of the final round. The code additionally contained a deterministic fallback method. If all three labels were different then it selected the first answer in agent order. No such cases occurred in the Du adaptive evaluation.

Within the Liang framework, the scored output was reduced to a single A/B/C/D label; for moderator-scored paths, this was provided from debateanswer. For each condition, an ex-

act match between the extracted label and the QuALITY gold label was needed. Accuracy was defined as the number of correct answers divided by 100 questions.

Procedural Debate Length and Computational Usage

We calculated the mean and total number of completed rounds for each debate condition. Completed rounds are a procedural measure of debate length, not a direct measure of computational cost. Prompt tokens, completion tokens, and approximate API costs were reported in both Du and single-agent logs. Tokens consumed by Liang calls were not logged; however, we can reconstruct model call counts from the supplied code.

The single-agent baseline called for 100 model calls. Du fixed one used 300 calls, Du fixed two used 600 calls, and the original Du adaptive run used 327 calls to the agent plus 109 calls to the judge for a total of 436 calls. The original Liang fixed-two condition used 800 calls, Liang strict adaptive used 848, and prompt-matched fixed-two control used 600. Logged estimated costs were approximately \$0.106 for the single-agent baseline, \$0.317 for Du fixed one, \$0.945 for Du fixed two, and \$0.529 for Du adaptive.

Completed rounds are considered only in relation to their respective framework. Since Liang token totals were not logged, direct comparisons of token cost between the two frameworks were not possible.

Statistical Methods

Exact two-sided McNemar tests were applied to paired accuracy comparisons because the same 100 questions were evaluated in each condition.²⁵ In addition to comparing the main framework accuracies, the Liang prompt-matched control accuracies, the single agent baseline accuracies, and the original-prompt Liang fixed-round bracket accuracies, we also estimated 95 percent Wilson confidence intervals for each accuracy measure. Some of the McNemar comparisons are provided along with their respective discordant counts (questions answered correctly on only one of the paired comparisons), and matched pair odds ratios. Additionally, exact 95 percent confidence intervals for the odds ratios were estimated based on the binomial distribution of discordant pairs. Furthermore, as a means to avoid conducting multiple-comparison searches when testing different thresholds at varying levels, our primary focus was on presenting the descriptive findings from the four Du threshold sensitivity runs versus performing all six pair-wise threshold comparisons.

Exact two-sided paired sign tests were performed to determine whether or not the completed-round count differences for each individual question exhibited a consistent direction, while excluding tied round-count differences.²⁶ Mean and

median round difference values from the paired comparisons have been documented, as well as a bootstrap-based 95 percent confidence interval for the mean round difference. Given that a number of additional related paired comparisons were conducted subsequent to the initial analysis, it is important to recognize that the associated p-values should be treated as exploratory evidence only. As a result, even small one- to three-point differences in terms of accuracy should be approached with caution particularly given that no model-generation seed was passed into the API calls.

Reproducibility Materials

In order to support replication efforts, a comprehensive reproducibility package has been developed which includes the exact same 100-question CSV file, a question ID manifest, question level output values, a prompt-matched Liang control, all Du threshold-sensitivity and Liang fixed-round bracketing runs, orchestration and framework modification scripts, validation reports, checksums, prompt templates, extraction logic, and derived analysis table values. The package also documents the procedure used to select the 100-question sample from the QuALITY train split, as well as the repository commit IDs used for both frameworks.

Ethical Considerations

As this research utilizes publicly available datasets and source code repositories, no human participants were engaged in this study. No personal identifying information was used. Similarly, there was no intervention involving human subjects. Accordingly, institutional review board approval and/or informed consent were not applicable.

Results

Overall Performance

In the primary matrix, Du's fixed one round had the greatest observed accuracy, 0.80, whereas both the single-agent baseline and the original Du adaptive condition had 0.77. The original Liang fixed two had 0.73, and the prompt-matched fixed two and strict adaptive conditions had 0.52 and 0.53, respectively. Figure 2 shows the respective Wilson 95% confidence intervals.

Fixed-Round Versus Adaptive Stopping

Because 95 of the 100 questions in Du adaptive also ended after one round, Du fixed one is the most relevant Du baseline. Du fixed one had 0.80 accuracy with 100 completed rounds in total; Du adaptive had 0.77 with 109 completed rounds plus

109 additional judge calls. There was no statistically significant difference ($p = 0.508$) when comparing the pair-wise accuracy of these two conditions. Although adaptive stopping resulted in fewer rounds than fixed two, fixed two was not the best of the other simple baselines.

The original Liang fixed-two condition had 0.73 accuracy, and Liang's strict adaptive stopping condition had 0.53. However, the above comparison also changed the moderator prompt and the scored-output path. Liang's prompt-matched fixed-two control, which retained the modified moderator prompt and scored-output path, had 0.52 accuracy. Although there was a large statistically significant difference ($p < .001$) between Liang's original fixed-two condition and Liang's fixed two condition that matched prompts, there was not a statistically significant difference ($p = 1.00$) when comparing Liang's strict adaptive stopping condition to Liang's fixed two condition that matched prompts. Therefore, the original Liang gap cannot be due solely to the stopping behavior.

Paired Accuracy Tests

For example, in the comparison of Du fixed one and Du adaptive, there were only nine discordant question pairs, leaving little ability to distinguish small accuracy differences. For the comparison of Du fixed two and Du adaptive stopping, there were fourteen discordant questions. If we use an exact two-sided McNemar test to compare these two conditions at $\alpha = .05$, then if we have at least a 12 to 2 split among the fourteen discordant pairs, then we will be able to find a statistically significant difference (i.e., at least a ten point net difference in pair-wise accuracy in our hundred question sample).

When compared to the primary Du conditions, as well as to Liang's original fixed-two condition, there were no statistically significant differences found when comparing the single-agent baseline. When compared to Liang's prompt-matched fixed-two control and Liang's strict adaptive condition, however, the single agent baseline did have statistically significantly greater accuracy.

Stopping Distribution Analysis

Du's adaptive stopping distribution was highly skewed toward early stopping: ninety-five questions stopped after one round, four stopped after two rounds, and only one question reached six rounds. On the other hand, Liang's adaptive stopping distribution was less skewed: twenty-one stopped after round one, fifty-two after round two, four after round three, one after round four, three after five rounds, and nineteen reached round six. One bar for each stopping round is shown in Figure 3 for each framework.

Accuracy also varied with stopping round. Questions in Du's adaptive stopping that stopped after round one were cor-

Table 1 Performance Across Experimental Conditions

Condition	Accuracy	95% Wilson CI	Avg. rounds	Total rounds
Single-agent baseline	0.77	[0.678, 0.842]	n/a	n/a
Du fixed 1 ensemble	0.80	[0.711, 0.867]	1.00	100
Du fixed 2	0.79	[0.700, 0.858]	2.00	200
Du adaptive (T90)	0.77	[0.678, 0.842]	1.09	109
Liang fixed 2	0.73	[0.636, 0.807]	2.00	200
Liang fixed 2 prompt-matched	0.52	[0.423, 0.615]	2.00	200
Liang strict adaptive (T90)	0.53	[0.433, 0.625]	2.70	270

Note. Debate-round measures are not applicable to the single-agent baseline; it still required one model response per question. Du fixed 1 is labeled an independent ensemble because its three round-1 agents do not see one another before majority-vote aggregation.

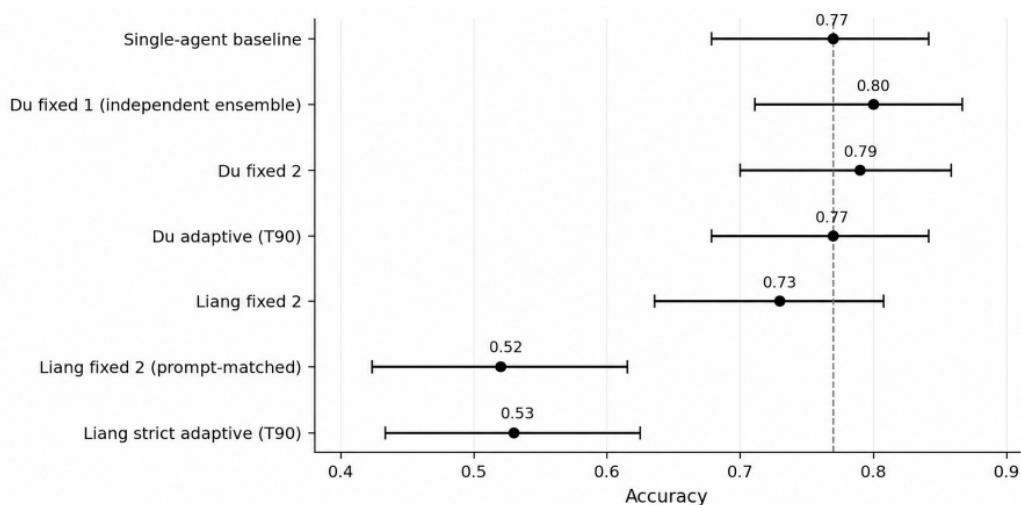


Figure. 2 Accuracy and 95% Wilson confidence intervals for the seven primary conditions. The vertical dashed line marks the single-agent baseline accuracy of 0.77.

Table 2 Accuracy Comparisons Between Conditions

Comparison (A vs. B)	A only	B only	McNemar p	Matched OR (95% CI)
Du fixed 1 vs. Du adaptive	6	3	0.508	2.00 [0.43, 12.36]
Du fixed 1 vs. Du fixed 2	6	5	1.000	1.20 [0.31, 4.97]
Du fixed 2 vs. Du adaptive	8	6	0.791	1.33 [0.41, 4.66]
Liang fixed 2 vs. Liang adaptive	30	10	0.002	3.00 [1.43, 6.88]
Single agent vs. Du fixed 1	2	5	0.453	0.40 [0.04, 2.44]
Single agent vs. Du fixed 2	5	7	0.774	0.71 [0.18, 2.61]
Single agent vs. Du adaptive	4	4	1.000	1.00 [0.19, 5.37]
Single agent vs. Liang fixed 2	11	7	0.481	1.57 [0.56, 4.78]
Single agent vs. Liang adaptive	35	11	<0.001	3.18 [1.58, 6.95]
Liang fixed 2 vs. prompt-matched fixed 2	29	8	<0.001	3.63 [1.62, 9.18]
Liang prompt-matched fixed 2 vs. adaptive	10	11	1.000	0.91 [0.35, 2.36]
Single agent vs. Liang prompt-matched fixed 2	36	11	<0.001	3.27 [1.63, 7.13]
Liang fixed 1 vs. fixed 2 (original prompt)	8	20	0.036	0.40 [0.15, 0.95]
Liang fixed 2 vs. fixed 3 (original prompt)	13	8	0.383	1.63 [0.62, 4.52]

Note. A only and B only are discordant question counts. The matched odds ratio is A-only/B-only. The prompt-matched Liang comparison isolates stopping while retaining the modified moderator prompt and scored-output path.

rect on seventy-five of ninety-five questions (0.789). Since only four questions stopped after two rounds in Du's adaptive stopping condition, only one of those four questions was answered correctly. The single question that stopped after six

rounds was also correct. In Liang's adaptive stopping condition, however, accuracy was: 0.619 for questions that stopped after round one; 0.481 for questions that stopped after round two; zero for both of the very small groups of questions that

Table 3 Debate Round Comparisons

Comparison	Adaptive fewer	Tied	Adaptive more	Mean diff. (95% CI)	Median	Sign-test p
Du fixed 1 vs. adaptive	0	95	5	+0.09 [0.01, 0.21]	0	0.063
Du fixed 2 vs. adaptive	95	4	1	-0.91 [-0.99, -0.79]	-1	<0.001
Liang fixed 2 vs. adaptive	21	52	27	+0.70 [0.37, 1.05]	0	0.471

Note. Sign tests exclude ties. Confidence intervals for paired mean differences are percentile bootstrap intervals with a fixed resampling seed.

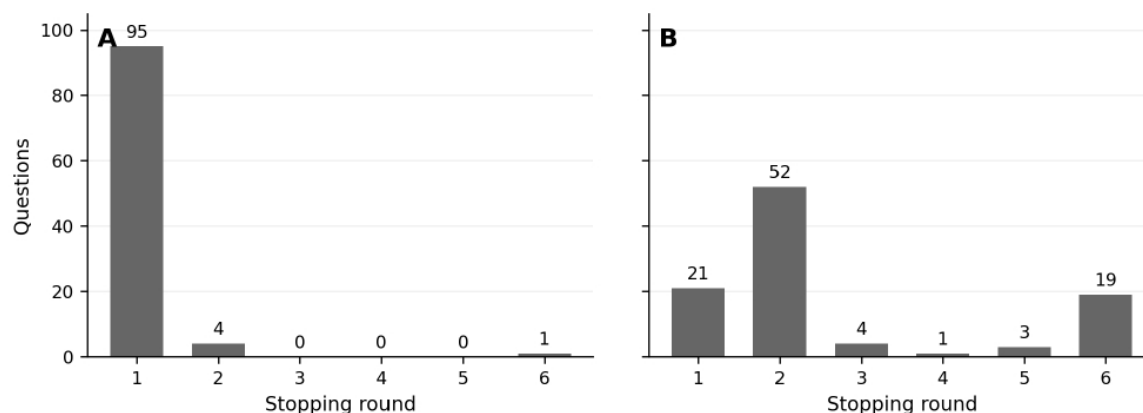


Figure. 3 Stopping-round distributions for (A) Du adaptive and (B) Liang strict adaptive. Each bar is the number of questions terminating at that round ($n = 100$ per condition).

stopped at rounds three and four; 0.333 for questions that stopped at round five; and 0.737 for the nineteen questions that made it to round six. These are simply descriptive results and not evidence that a specific stopping round caused either increased or decreased accuracy because question difficulty was not randomized across stopping rounds.

Threshold Sensitivity and Fixed-Round Bracketing

Additional studies were conducted to determine whether the Du results depended upon the confidence level, and/or whether the Liang results differed based upon the results of an adaptive stopping strategy versus a single fixed stopping schedule.

As well, the Du sweep showed an increase in the average number of completed rounds from 1.00 to 1.10 when the confidence limit rose from 70 to 95 on the 0–100 scale. There was no identifiable pattern of improvement in the accuracy measure of .77 to .80. Therefore, the sweep identified no confidence cutoff with a clear advantage over the simpler fixed schedules in both accuracy and computation.

Furthermore, a re-run of the threshold-90 condition yielded an accuracy of .78 and 104 completed rounds. These were very similar to the previous values (.77 and 109) obtained initially. Although the paired difference in accuracy between the two conditions was statistically insignificant ($p = 1.000$), the results indicate the need for caution when assessing small dif-

ferences in accuracy between experimental conditions.

As with the Du sweeps, the original-prompt Liang fixed-round bracket did not display a monotonically increasing trend in accuracy as the number of rounds increased. For example, for fixed rounds one through three the accuracy levels were .61, .73, and .68 respectively. Therefore, simply having more rounds does not guarantee better accuracy.

Error Analysis

This stopping decision analysis evaluates whether a better answer appeared in a later observed output or in a matched separate-run outcome for the same question. Each analysis uses paired question identification numbers. For Du, the fixed-one and fixed-two runs provide a cross-run two-round comparison. For Liang, the prompt-matched fixed-two control produced moderator outputs at rounds one and two. Cross-run comparisons are treated as empirical opportunities rather than deterministic alternative histories.

Of the ninety-five Du adaptive questions that stopped after round one, only five cases had an incorrect adaptive result and a correct result when examined separately in the fixed-two run. Conversely four of the ninety-five cases went in the opposite direction. Across all one hundred Du question IDs, comparing the separate fixed-one and fixed-two runs showed five corrections and six regressions. Thus, although the second-

Table 4 Sensitivity and Fixed-Round Results

Condition	Accuracy (95% Wilson CI)	Avg. rounds	Model calls	Stopping distribution / schedule
Du adaptive T70	0.79 [0.700, 0.858]	1.00	400	R1: 100
Du adaptive T80	0.77 [0.678, 0.842]	1.02	408	R1: 98; R2: 2
Du adaptive T90 (revision rerun)	0.78 [0.689, 0.850]	1.04	416	R1: 97; R2: 2; R3: 1
Du adaptive T95	0.80 [0.711, 0.867]	1.10	440	R1: 94; R2: 5; R6: 1
Liang fixed 1, original prompt	0.61 [0.512, 0.700]	1.00	500	Fixed 1
Liang fixed 2, original prompt (primary reference)	0.73 [0.636, 0.807]	2.00	800	Fixed 2
Liang fixed 3, original prompt	0.68 [0.583, 0.763]	3.00	1,100	Fixed 3

Note. Du model calls equal three answer-generating calls plus one judge call per completed round. Liang fixed-run model calls equal three calls per debate round plus the repository's two-call final judge. The Liang fixed-two row is the archived primary run and is included as the middle reference for the one/two/three-round bracket.

Table 5 Question-Level Error Analysis

Analysis	Observed transitions	Best-available/oracle result
Du: 95 adaptive questions stopping at round 1	71 correct in both; 15 wrong in both; 5 adaptive wrong / fixed-2 correct; 4 adaptive correct / fixed-2 wrong	80/95 (0.842) if an oracle could choose between the observed adaptive round-1 and fixed-2 outcomes; adaptive was 75/95 (0.789)
Du: fixed 1 to fixed 2, all 100 questions	5 wrong-to-correct; 6 correct-to-wrong; 89 unchanged in correctness	85/100 (0.85) if an oracle could choose the correct fixed-1 or fixed-2 outcome; best single fixed schedule was 80/100
Liang prompt-matched control: moderator round 1 to round 2	3 wrong-to-correct; 2 correct-to-wrong; 95 unchanged in correctness; only 5 answers changed	54/100 (0.54) if an oracle could choose the better moderator round; forced round 2 was 52/100
Liang adaptive: moderator round 1 to final scored output	10 wrong-to-correct; 5 correct-to-wrong; 85 unchanged in correctness	59/100 (0.59) if an oracle could choose among all observed moderator answers plus the final judge output; actual final was 53/100

round outcomes contained some possible correction opportunities, the fixed-two schedule did not improve accuracy overall. The two-round oracle accuracy of .85 represents the highest achievable accuracy by perfectly selecting between the observed fixed-one and fixed-two results.

Similarly, Liang's prompt matched controls demonstrated similar trends when analyzing moderator answers between rounds one and two. Of the 100 question-level comparisons performed from moderator round one to round two, only five answers changed: three from wrong-to-correct and two from correct-to-wrong. Additionally, there were twenty-one question IDs for which the original Liang adaptive run ended operation after completing round one. Only one answer in the corresponding prompt-matched forced-two control changed between its first and second rounds; however, that change was incorrect-to-correct. Because this control represented a distinct API run, it provided secondary evidence rather than being a direct counterfactual. In the original Liang adaptive trajectories an oracle selecting among all observed moderator answers and the final judge output would achieve an accuracy of 0.59, whereas the actual final accuracy achieved was 0.53. Thus, a six point recovery gap existed.

Most answer changes occurred among the 19 questions that reached the six-round cap. For nine of the 19 questions that reached the cap, there was a change from an incorrect mod-

Table 6 Liang Answer Transitions by Round

Rounds used	n	C→C	W→C	C→W	W→W	Answer changed
1	21	13	0	0	8	0
2	52	24	1	0	27	1
3	4	0	0	0	4	0
4	1	0	0	1	0	1
5	3	1	0	1	1	1
6	19	5	9	3	2	14
Total	100	43	10	5	42	17

erator's first round answer to a correct final answer. In contrast, three questions experienced a reversal of this type; they received a correct moderator answer in the first round, but an incorrect final answer. In addition, six adaptive questions ended with an incorrect final answer, although a correct moderator debateanswer had appeared at an earlier round for each of those questions. Conversely, eight of the 19 questions that reached the cap were corrected by the separate final judge after none of the previous moderator rounds had provided a correct answer. The additional computation allowed for multiple opportunities to correct answers; however, it also reversed some previously correct answers. Thus, whether continuation added value depended heavily on the question.

Table 7 lists confidence distributions and answer accuracy

in three confidence bins. There were 99 of 109 judge evaluations reported to be at or above 90 on the 0–100 confidence scale during the Du adaptive run. Yet, despite such high confidence ratings, incorrect judge answers were still present at these high-confidence evaluation levels. For the Liang moderator, accuracy increased from the lowest confidence bin to the middle confidence bin. Accuracy changed very little between the 80–89 and 90–100 bins.

In addition to reporting on accuracy rates for each of the three confidence bins, Table 7 reports on confidence distributions. The Liang Moderator produced a total of 270 confidence measures during its use in the Liang Strict Adaptive run. The median confidence measure was 85, and the interquartile range was 75–90. Across the 100 adaptive questions submitted during the Liang Strict Adaptive run, accuracy for the moderator’s first round of answers was .48. Accuracy for all of the adaptive questions’ final answers was .53. Eight of the twenty-one questions that stopped after Round 1 had incorrect answers. Five of nineteen questions reaching Round 6 were given an incorrect final answer. Results from a prompt-matched fixed two-control run indicated that having high confidence does not necessarily mean that the moderator will give a correct answer. Forty-four (or approximately 59%) of seventy-four final answers with confidence levels of at least 90 were accurate.

The error taxonomy helps to clarify the observed failures. The original primary conditions had no blank final A/B/C/D answers, while the prompt-matched Liang control had one blank final debate answer, which was scored as incorrect. Confidence parsing succeeded in both archived adaptive runs. High reported confidence remained compatible with substantial error. Moreover, the prompt-matched Liang control reproduced almost exactly the same accuracy as strict adaptive stopping. Taken together, these results point to the modified Liang moderator/scored-output configuration, rather than adaptive termination alone, as the main explanation for the original Liang accuracy gap.

Discussion

In the Du framework, the most basic ensemble of three agents was difficult to improve upon. As agents do not communicate with each other in Round One, Du fixed one separates the contribution of ensembling from the effects of later debate. Neither adding rounds of debate nor using a confidence-based stopping judge provided a clear benefit over the less complex baseline. The threshold sweep also indicated there was no monotonic increase in accuracy based on how restrictive the stopping criteria were. These findings are consistent with earlier research showing that multi-agent debate does not reliably outperform ensembling or self-consistency across protocols and baselines.¹⁵

Accuracy from the single-agent baseline did not differ significantly from the primary conditions of Du and the original Liang fixed-two condition. The lower accuracy results from the Liang prompt-matched and strict adaptive conditions reflect the moderator/output configuration used in testing rather than a general effect of multi-agent interaction.

Diagnostics related to the stopping decision illustrated why adaptation had few opportunities to assist. Although subsequent rounds occasionally corrected inaccuracies, they also introduced new errors and only a minority of questions contained recoverable improvements. Thus, a controller must assess not only whether an existing answer appears confident, but also whether there is a good chance that another round would enhance that particular answer.

A similar conclusion can be drawn from examining the threshold sweep. More strictly defined confidence levels resulted in continued debating more consistently than improving the accuracy. Therefore, confidence served as an imperfect control signal within these experiments and could not be treated as a reliable indicator in these runs of whether another debate round would result in an enhanced answer.

Relation to Prior Work

These results need to be viewed alongside prior research instead of being seen as a contradiction of Liang et al. Du et al. demonstrated that multi-agent debate was beneficial on multiple tasks whereas Smit et al. showed that these same benefits are not always consistent across protocols and baselines.^{1,15} The previously published adaptive-break mechanism by Liang et al. was not reproduced unmodified here.² However, as shown in this study’s new control, the modified moderator-prompt and moderator-scored output-path can produce similarly low accuracy even when debate is constrained to exactly two rounds. Adaptive-Consistency provides a useful contrast since its stopping-signal is inter-sample agreement and not verbalized confidence.²⁷

Additionally, inference-time control mechanisms need to be evaluated as systems consisting of the control signal, threshold, prompt, aggregation rule and overhead. Prior work has examined confidence-weighted consensus and sparse communication between debating agents.^{7,8} Other approaches aggregate sampled reasoning paths through self-consistency or explore alternative reasoning paths through Tree of Thoughts.^{24,28} Self-Refine uses iterative feedback to revise outputs,²⁹ while Reflexion uses linguistic feedback to inform subsequent attempts.³⁰ Language Agent Tree Search combines search with model-generated evaluations and reflections.³¹ The current results demonstrate why measuring the value of the guidance supplied by the control mechanism is necessary and not simply assumed.

Table 7 Confidence and Accuracy

Framework	Confidence distribution	Accuracy, 50–79	Accuracy, 80–89	Accuracy, 90–100
Du adaptive	median 95; IQR 95–95; range 70–100	2/3 = 0.667	1/7 = 0.143	79/99 = 0.798
Liang strict adaptive	median 85; IQR 75–90; range 50–90	22/78 = 0.282	52/110 = 0.473	39/82 = 0.476

Practical Implications

For system designers, the most significant implication is methodological: Compare adaptive controllers with robust simple baselines; Separate ensembling from actual debate; Count overhead incurred by controller; Sweep or calibrate stopping thresholds; Use prompt-matched controls when the stopping intervention also modifies the prompt used in scoring output. Conducting a recoverability assessment before optimizing a controller is useful, as adaptation provides little upside when subsequent rounds rarely produce recoverable corrections.

Limitations

A primary limitation of this study’s results is that they may not generalize to the entire dataset or to other types of tasks, since only one 100-question QuALITY sample (spanning 64 articles) was used. Also, questions drawn from the same article are certainly not entirely independent. Most conditions were run only once, and some run-to-run variability is possible even at temperature 0.0. The threshold-90 rerun therefore provides only a very limited check on stability.

While the prompt-matched Liang control and the Liang fixed-one/fixed-three bracket were run independently of the archived primary conditions, they do serve to provide additional information regarding how to interpret the results; however, they must be viewed as sensitivity analyses rather than repeated replications. Similarly, the Du threshold sweep examined only four different cutoffs and did not calibrate those cutoffs on an independent development set.

Also, completed rounds cannot be considered as if each represents equal units of cost in terms of architecture across frameworks due to differing numbers and types of model calls made per round. Token and monetary-cost fields were retained for the Du and single-agent runs, but not for the Liang runs. Furthermore, this study utilized only one model family and one task, thereby limiting the ability to generalize to other models, datasets, and debate protocols.

Conclusion and Future Work

This study sought to determine whether confidence-threshold adaptive stopping could either improve the accuracy of answers provided by a debate protocol or simply decrease un-

necessary rounds of debate compared with fixed-schedule approaches. The adaptive settings within the Du framework did not demonstrate a clear advantage in accuracy or computational effort relative to the simpler fixed-one independent ensemble. Similarly, the threshold sweep had a greater effect upon computation than upon accuracy. Within the Liang framework, the prompt-matched fixed-two control demonstrated similar performance characteristics to strict adaptive stopping. Therefore, the original differences in accuracy between the fixed and adaptive conditions cannot be attributed to the stopping policy alone. Additionally, in Liang, the original-prompt fixed-one, fixed-two, and fixed-three schedules demonstrated non-monotonic behavior: the number of rounds increased, but did not increase answer quality consistently. Therefore, the study met its research objective through comparison of the confidence-threshold rules being tested against strong fixed baselines. These confidence-threshold rules did not consistently outperform their corresponding baselines in terms of accuracy or computation in these two testbeds. The most important lesson learned here is that a stopping policy must be evaluated as part of the overall debate system, which includes the signal, threshold, prompt, aggregation rule, and controller overhead.

Future studies should evaluate these stopping policies against fixed-schedules under several different sets of experimental conditions including multiple runs, models, and tasks. Future studies should also develop ways to use separate development data to calibrate stopping thresholds. Additionally, future studies should investigate alternative signals such as agent agreement, answer stability, and inter-sample consensus. More complete token and latency logging would enable direct efficiency comparisons.

Acknowledgments

I would like to thank my research mentor, Clayton Greenberg, a former professor at the University of Pennsylvania, for his guidance throughout this research project. He helped me clarify my research question, refine my experimental methodology, and better understand the findings of the study.

References

1 Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through

- multiagent debate. Proceedings of the 41st International Conference on Machine Learning. Vol. 235, pg. 11733–11763, 2024. <https://proceedings.mlr.press/v235/du24e.html>.
- 2 T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu. Encouraging divergent thinking in large language models through multi-agent debate. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pg. 17889–17904, 2024. <https://doi.org/10.18653/v1/2024.emnlp-main.992>.
- 3 C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun. ChatDev: Communicative agents for software development. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pg. 15174–15186, 2024. <https://doi.org/10.18653/v1/2024.acl-long.810>.
- 4 Z. Yin, Q. Sun, C. Chang, Q. Guo, J. Dai, X. Huang, and X. Qiu. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pg. 15135–15153, 2023. <https://doi.org/10.18653/v1/2023.emnlp-main.936>.
- 5 A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, and E. Perez. Debating with more persuasive LLMs leads to more truthful answers. Proceedings of the 41st International Conference on Machine Learning. Vol. 235, pg. 23662–23733, 2024. <https://proceedings.mlr.press/v235/khan24a.html>.
- 6 K. Xiong, X. Ding, Y. Cao, T. Liu, and B. Qin. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. Findings of the Association for Computational Linguistics: EMNLP 2023, pg. 7572–7590, 2023. <https://doi.org/10.18653/v1/2023.findings-emnlp.508>.
- 7 J. Chen, S. Saha, and M. Bansal. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pg. 7066–7085, 2024. <https://doi.org/10.18653/v1/2024.acl-long.381>.
- 8 Y. Li, Y. Du, J. Zhang, L. Hou, P. Grabowski, Y. Li, and E. Ie. Improving Multi-Agent Debate with Sparse Communication Topology. Findings of the Association for Computational Linguistics: EMNLP, pg. 7281–7294, 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.427>.
- 9 A. Amayuelas, X. Yang, A. Antoniadis, W. Hua, L. Pan, and W. Y. Wang. MultiAgent Collaboration Attack: Investigating Adversarial Attacks in Large Language Model Collaborations via Debate. Findings of the Association for Computational Linguistics: EMNLP, pg. 6929–6948, 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.407>.
- 10 S. Wang and L. Huang. Debate as Optimization: Adaptive Conformal Prediction and Diverse Retrieval for Event Extraction. Findings of the Association for Computational Linguistics: EMNLP, pg. 16422–16435, 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.958>.
- 11 C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. ChatEval: Towards Better LLM-Based Evaluators through Multi-Agent Debate. International Conference on Learning Representations, 2024. <https://openreview.net/forum?id=FQepisCUWu>.
- 12 Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations. Conference on Language Modeling, 2024.
- 13 S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, and J. Schmidhuber. MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework. International Conference on Learning Representations, 2024. <https://openreview.net/forum?id=VtmBAGCN7o>.
- 14 W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung, C. Qian, Y. Qin, X. Cong, R. Xie, Z. Liu, M. Sun, and J. Zhou. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. International Conference on Learning Representations, 2024.
- 15 A. P. Smit, N. Grinsztajn, P. Duckworth, T. D. Barrett, and A. Pretorius. Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs. Proceedings of the 41st International Conference on Machine Learning. Vol. 235, pg. 45883–45905, 2024. <https://proceedings.mlr.press/v235/smit24a.html>.
- 16 J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems 35, 2022. <https://doi.org/10.52202/068431-1800>.
- 17 E. Zelikman, Y. Wu, J. Mu, and N. Goodman. STaR: Bootstrapping Reasoning With Reasoning. Advances in Neural Information Processing Systems 35, 2022. <https://doi.org/10.52202/068431-1126>.
- 18 J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. International Conference on Learning Representations, 2024. <https://openreview.net/forum?id=Ikmd3fKBPQ>.
- 19 R. Y. Pang, A. Parrish, N. Joshi, N. Nangia, J. Phang, A. Chen, V. Padmakumar, J. Ma, J. Thompson, H. He, and S. R. Bowman. QuALITY: Question answering with long input texts, yes! Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pg. 5336–5358, 2022. <https://doi.org/10.18653/v1/2022.naacl-main.391>.
- 20 K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. arXiv, 2021. <https://doi.org/10.48550/arXiv.2110.14168>.
- 21 composable-models.llm_multiagent_debate[computer software]. GitHub, commit 9846749350eb917ae5bfaaff4c645fc705b8d3af, 2023. Repository version verified July 27, 2026. https://github.com/composable-models/llm_multiagent_debate.
- 22 T. Liang. Multi-Agents-Debate [computer software]. GitHub, commit e58d146033568ccf737d8031a93748597907d7d3, 2023. Repository version verified July 27, 2026. <https://github.com/Skytliang/Multi-Agents-Debate>.
- 23 OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. 2024. Accessed September 2026. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- 24 X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. International Conference on Learning Representations, 2023. <https://openreview.net/forum?id=1PL1NIMMrw>.
- 25 Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. Vol. 12, pg. 153–157, 1947. <https://doi.org/10.1007/BF02295996>.
- 26 W. J. Dixon and A. M. Mood. The statistical sign test. Journal of the American Statistical Association. Vol. 41, pg. 557–566, 1946. <https://doi.org/10.1080/01621459.1946.10501898>.
- 27 P. Aggarwal, A. Madaan, Y. Yang, and Mausam. Let’s Sample Step by Step: Adaptive-Consistency for Efficient Reasoning and Coding with LLMs. Proceedings of EMNLP, pg. 12375–12396, 2023. <https://doi.org/10.18653/v1/2023.emnlp-main.761>.
- 28 S. Yao, D. Yu, J. Zhao, I. Shafra, T. Griffiths, Y. Cao, and K. Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. Advances in Neural Information Processing Systems 36, 2023. <https://doi.org/10.52202/075280-0517>.
- 29 A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhume, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-Refine: Iterative Refinement with Self-Feedback. Advances in Neural Information Processing Systems 36, 2023. <https://doi.org/10.52202/075280-2019>.
- 30 N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. Advances in Neural Information Processing Systems 36, 2023.

<https://doi.org/10.52202/075280-0377>.

- 31 A. Zhou, K. Yan, M. Shlapentokh-Rothman, H. Wang, and Y.-X. Wang. Language Agent Tree Search Unifies Reasoning, Acting, and Planning in Language Models. Proceedings of the 41st International Conference on Machine Learning. Vol. 235, pg. 62138–62160, 2024. <https://proceedings.mlr.press/v235/zhou24r.html>.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-67623-6>.