When AI Agents See Each Other's Answers, Their Answers Grow More Alike: A Controlled Study of Multi-Agent LLM Revision

Hideki Muto[1], Tetsuro Ogi[2], Takahiro Yakoh[3]
Graduate School of System Design and Management, Keio University, Yokohama, Japan[1,2,3]
Faculty of Health and Sport Sciences, Asia University, Tokyo, Japan[2]

***Abstract*—AI systems can offer several answers when a question has no single correct solution. For people seeking alternatives, those answers need to differ. Multi-agent systems often let agents read one another's answers during revision. Can agents keep their answers different when they read one another's work? Here we show that reading other answers made the final answers more alike, even when agents were told to keep their answers separate. We copied the same five independent answers into conditions where agents could read zero, one or four other answers during four revision rounds. We tested 20 open-ended tasks with GPT-4o-mini and Qwen3-30B-A3B, with three repetitions per task and model. We compared the mean embedding distance among final answers at the task level. Relative to no access, distance fell by 8% and 15% with one and four other answers for GPT-4o-mini, and by 4% and 9% for Qwen. All four planned tests remained significant after adjustment for multiple tests. A second text encoder showed the same pattern for complete answers. In a post-hoc check, agents moved more toward the randomly selected answer they read than toward unread answers. Showing other answers also lengthened the input; this study cannot isolate their content effect or determine whether distinct ideas were lost. Qwen's one-answer difference was near zero when only the first 128 tokens of each answer were measured. Access to other answers is therefore a consequential design setting. Systems intended to offer alternatives should assess both individual answers and the differences among them.**

***Keywords*—Large language models; multi-agent systems; inter-agent communication; response diversity; response convergence**

I. Introduction

Large language models (LLMs) can help people develop ideas, compare options and write drafts. In multi-agent systems, several agents can work on the same task and read one another's answers during revision [1], [2]. Some workflows seek one agreed answer. Others keep several answers for a person to consider. When the aim is to offer alternatives, the differences among those answers matter.

Previous research gives reasons to examine these differences. Studies of human brainstorming show that other people's ideas can stimulate new ideas but can also direct attention toward shared examples [3], [4]. LLM outputs can resemble one another even without interaction [5]. In multi-agent research-proposal generation, Chen et al. found that dense communication accelerated convergence [6]. Other dialogue designs have increased diversity in research ideation [7]. These findings suggest that the effect of sharing depends on how agents interact.

Useful answers can still be too similar. Imagine asking five AI agents how to make buses and trains better. All five might suggest an easier way to buy tickets. None might mention long waits or crowded buses. Each answer could help, but together they offer little choice. If we judge each answer alone, we may miss this overlap. A related pattern has been observed in human writing: access to AI ideas improved individual stories while making the stories more similar to one another [8]. Designers therefore need to assess the answers as a set. They can also decide which answers agents see during revision, without changing the underlying model. Whether this setting changes the differences among final answers is a practical question for systems designed to offer alternatives.

Does reading other answers make agents' answers more alike even when agreement is not the goal? Chen et al. discuss pressure toward agreement as a source of convergence. Their pipeline ends by synthesizing the discussion into a research proposal [6]. In the setting we study, each agent keeps its own final answer: there is no shared proposal to produce. Separate outputs might preserve different approaches, yet the answers agents read may still guide their revisions. This makes access to other answers a specific design choice worth testing.

We tested this question on 20 open-ended tasks with two models and three repetitions per task and model. Five agents first answered independently. We copied the same initial answers into three conditions, allowing each agent to read zero, one or four other answers during revision. All conditions used the same four revision rounds. Starting from the same answers removed initial differences from the comparison, while the no-access condition measured change through revision alone. We measured the average embedding distance among the five final answers, which we call the *set distance*. A lower value means that the complete answers are more similar; it does not necessarily mean that they contain fewer distinct ideas.

Earlier findings gave reason to expect that reading other answers would reduce the final set distance, with a larger reduction when more answers were visible. They did not show whether this would happen when agents kept their own answers separate. The analysis plan was fixed before the main runs, as documented in Section III-D. It used two-sided tests to compare each access condition with no access. The comparison between one and four visible answers was secondary (Section III-E).

In both models, the complete final answers were more alike when agents could read other answers. The average reduction in set distance was larger with four other answers than with

one. Our contribution is an estimate of how access to other answers changes the final answer set in a controlled workflow that keeps each agent's answer separate. We also provide post-hoc evidence that agents moved toward the specific answers they were randomly assigned to read.

II. Related Work

A. Sharing and Diversity in Groups

Sharing information can help a group, but it can also narrow the range of solutions the group explores. In a model of information-sharing networks, rapid exchange improved short-term performance while limiting later exploration [9]. Keeping different solutions in play can support group problem solving [10]. Studies of brainstorming have found both stimulation by others' ideas and fixation around shared examples [3], [4]. In an experiment with people writing stories, access to AI ideas improved ratings of individual stories but made the stories more similar to one another [8]. These findings motivate examining the answer set, not just each answer alone.

B. Similarity in LLM Outputs

Similarity can arise without communication. Jiang et al. found repeated patterns within and across models, and Yun et al. found that prompt formatting reduced output diversity [5], [11]. Revision can also change an answer without input from other agents. Self-refinement studies show why a revision-only comparison is needed, even though self-correction does not always help [12], [13].

C. Interaction Among LLM Agents

Multi-agent frameworks can let agents read or exchange one another's work [1], [2]. Some, such as multi-agent debate, seek a common answer [14]. Choi et al. observed conformity in debates [15]. Chen et al. compared communication structures in research-proposal generation and found that denser communication accelerated convergence [6]. Li et al. studied sparse debate communication with a focus on performance and cost, while Ueda et al. found that some dialogue designs increased diversity in research ideation [16], [7]. Together, these studies show that interaction can have different effects in different workflows. They do not establish what happens when agents keep separate answers but vary how many other answers they can read during the same revision process.

A previous conference study compared revision workflows using the same 20 task texts [17]. Here we use newly generated answers to compare access settings within one revision workflow. All three access conditions start from the same five answers and use the same revision steps, including a condition with no access to other answers. This design compares access settings while accounting for initial differences and revision alone. A post-hoc analysis also tests movement toward the specific answer an agent was randomly assigned to read.

III. Methods

A. Study Design and Tasks

We reused the 20 task texts from the conference study [17], so the task set was fixed before the present experiment. The tasks cover four kinds of open-ended work, with five tasks in each: idea generation, reasoning and analysis, topical summaries without a source passage, and creative writing. Each task allows more than one reasonable answer. This lets us compare how far apart five answers to the same task remain after revision. The full task texts accompany the study.

We included all 20 tasks without selecting or replacing any after seeing the results. The same tasks entered each access condition and both model analyses, so each task provided a paired comparison. The size of this set also allowed us to complete the planned runs within the available generation time; it was not chosen from a power calculation. Section III-E reports the statistical power of the resulting 20-task analysis. The tasks were author-written, not randomly sampled from all possible uses.

We used GPT-4o-mini and Qwen3-30B-A3B-Instruct-2507. GPT-4o-mini was available through an API at a cost that allowed the full design, while Qwen could run locally. The study thus covered two model families and two deployment settings. For every task and model, we ran three independent repetitions. Each combination of task, model and repetition formed a *block* with five agents. An agent was a sequence of calls to the same model, carrying its own answer into the next round without any other conversation history. All five agents in a block used the same model.

B. Initial Answers and Access Conditions

First, the five agents produced initial answers without seeing one another's work. We generated these five answers once per block. We then copied each answer to the same agent position in three conditions: R0 (no other answers), R1 (one other answer) and R4 (all four other answers). The number in each label tells how many other agents' answers an agent could read during revision.

In every condition, an agent saw its own answer from the previous round. Other agents' answers appeared in a separate reference section of the prompt. That section was empty in R0. In R1, it contained one of the other four answers, chosen at random for each agent and round, so agents had no fixed partner. In R4, it contained all four answers in a new random order each round. The same revision instruction asked agents to use these answers as references while keeping their own answers separate. Predetermined random seeds fixed the choices and order. Table I summarizes the conditions.

TABLE I. The Three Access Conditions. Every Agent Also Saw its Own Previous Answer. The Same Revision Instruction was used in all Three Conditions

Condition	Reference section contains	Selection
R0	nothing; the section is present but empty	none
R1	one of the other four answers	chosen uniformly for each agent and round
R4	all four other answers	order randomized at each round

Agents revised their answers for four rounds. In each round, they could read only answers from the preceding round. They did not see answers being written in the current round. We randomized the order of the three conditions within each block (Fig. 1).

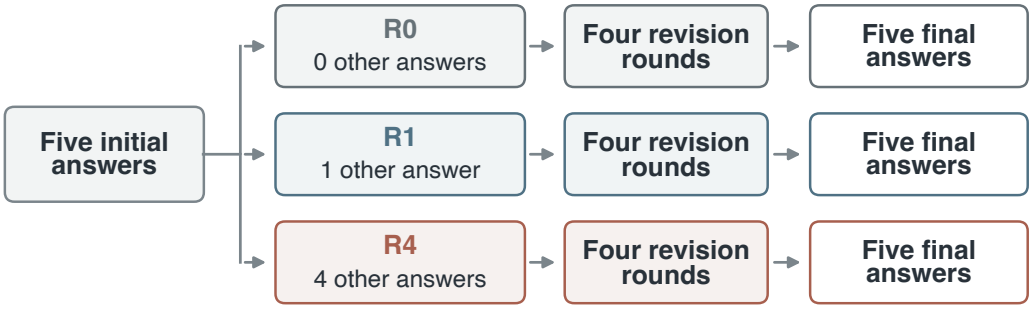

Fig. 1. One task, one model and one repetition. The same five initial answers branch into three conditions. Every agent keeps its own answer and can see zero, one or four other agents' answers during each of four revision rounds. Every round uses the same generator–critic–summarizer workflow. Each condition ends with five separate answers.

Showing more other agents' answers also made the prompts longer. Our comparison includes both the other answers and the added input length. It does not separate their effects.

C. Revision Workflow

An agent produced each answer in three steps, both at the start and in every revision round. First, a generator wrote a draft or revised the agent's previous answer. A critic then identified three ways to improve the draft and rewrote it. Its instructions asked it to assess clarity and coherence, originality, and completeness. Finally, a summarizer received the rewritten text and turned it into a short answer to the original task. That final summarizer answer became the agent's answer for the round. The summarizer retained the requested number of items, explanations and narrative form, but left out the critique and comments about the revision. All conditions and models used the same instructions for these steps. Only the generator received the reference section containing other agents' answers. Their content could pass through the draft, but the critic and summarizer did not receive those answers separately.

In a technical pilot, an earlier summarizer prompt sometimes produced comments about a revision instead of an answer to the task. We changed that prompt before the main experiment. We excluded pilot outputs from the analysis. The summarizer prompt differed from that used in the conference study [17]. Only newly generated outputs entered the present analysis; earlier outputs were not pooled with them or used as controls.

D. Implementation Details

1) Local protocol freeze: Before the main runs on 18 September 2026, we fixed the tasks, prompts, task order, random seeds, model settings, rules for handling failed responses (Section III-G) and primary analysis plan. The plan named the two comparisons for each model and the Holm correction across all four tests. It was the same for both models. We saved copies of the plan and generation code with SHA-256 hashes. At the start of each model run, the program also saved a dated configuration and its hash. On restart, it checked that hash and stopped if the configuration had changed. The reviewer package includes the plan, code, configurations and local freeze records. These records show the freeze and run-start times for both models. They are local records, not a public preregistration or an independent timestamp. The detailed procedures for supplementary scoring were fixed separately.

2) Model settings and prompts: Both generators used temperature 1.0, top-p 1.0 and an 8,192-token maximum output setting. GPT-4o-mini used the API snapshot `gpt-4o-mini-2024-07-18`. Qwen used the pinned `Qwen3-30B-A3B-Instruct-2507` revision `0d7cf23991f47feeb3a57ecb4c9cee8ea4a17bfe`, served locally with vLLM in bfloat16 on an NVIDIA DGX Spark (GB10). The Qwen server used a 65,536-token model context limit.

Appendix C gives the exact instruction templates and all 20 task texts. The frozen configurations and request records provide the remaining API parameters.

E. Outcome and Analysis

After round 4, we measured how different the five complete answers were in each condition. The text encoder `text-embedding-3-small` converted each answer into a vector, a list of numbers that represents the text. We measured the distance between every pair of vectors. Earlier studies used similar measures for stories and research proposals [8], [6]. The average of the ten pairwise Euclidean distances is the *set distance*. For vectors $\mathbf{e}_1, \ldots, \mathbf{e}_5$, it is

$$D = \frac{1}{10} \sum_{i<j} \|\mathbf{e}_i - \mathbf{e}_j\|_2. \tag{1}$$

This gives one value, D, for each set of five answers. A lower set distance means the complete answers are more alike. Similarity may come from shared content, shared wording, or both. The measure does not count distinct ideas or rate answer quality. Five agents could all give the same list of several ideas. Their answers would be similar, even though the list still contains several ideas. We therefore attempted a separate analysis of idea types (Section III-F).

All answers fitted within the primary encoder's 8,000-token limit, so each was embedded in full (Appendix A). We scaled each vector to unit length. For unit vectors, a pairwise Euclidean distance equals $\sqrt{2 - 2\cos\theta}$, where $\cos\theta$ is their

cosine similarity. Thus, smaller pairwise distances also mean higher cosine similarity.

For each repetition, we calculated two differences in set distance: $D_{\mathrm{R0}} - D_{\mathrm{R1}}$ and $D_{\mathrm{R0}} - D_{\mathrm{R4}}$. We call these the R0-minus-R1 and R0-minus-R4 contrasts. We averaged each difference over the three repetitions of the same task and model. Positive values therefore mean that access to other answers reduced the final set distance. Statistical tests used these task-level averages, giving 20 observations per model for each comparison.

We planned two main comparisons for each model: R0 versus R1, and R0 versus R4. This gave four primary tests. For each test, we compared the 20 task-level differences with zero using a two-sided one-sample t-test. We used the Holm method to adjust the four p values for multiple tests. We report mean differences, unadjusted and adjusted p values, ordinary 95% confidence intervals, and Cohen's d_z. The confidence intervals were not adjusted for multiple tests.

Each model had data from all 20 tasks. With 20 tasks, a two-sided paired test has 80% power for $d_z = 0.66$ at $\alpha = 0.05$. At the strictest Holm level ($\alpha = 0.0125$), that figure rises to $d_z = 0.81$. Smaller effects may go undetected. Three repetitions help reduce variation from one run to another. They do not increase the number of tasks.

We also described R1 versus R4, changes across rounds, and variation across tasks and repetitions. Because the conditions shared the same initial answers, round-change contrasts are descriptive, not separate confirmation of the final contrasts. Three reference levels cannot show an exact threshold or point of saturation. We report results for the two models separately. A result that is significant in one model but not the other does not, by itself, show that the models differ.

F. Complementary Measures

We planned three complementary checks: a second text encoder, ratings of content differences and classification of idea types. Only the second text encoder provided complete data for comparisons at the task level.

1) Second text encoder: We also represented complete answers with `all-mpnet-base-v2`, a Sentence-BERT model [18]. We divided the text into windows and combined their embeddings into one vector per answer. We then repeated the distance comparisons to check whether the pattern held with another encoder.

2) Content ratings: We asked raters from two other model families to assess differences in content among the five answers. They did not see condition labels. The rubric distinguished content from wording and required exact excerpts from the answers. Each answer set was presented in two orders. Too few ratings passed the planned checks to support a comparison at the task level (Appendix B). LLM ratings can also depend on answer order and length, among other biases [19].

3) Idea types: For the five idea-generation tasks, we planned to group proposals by their purpose and method, then examine which types remained after revision. Too few groupings passed the planned checks to estimate idea retention. We kept the original validation rule and did not create a new measure after seeing the data. Appendix B gives the procedure and the missing-data counts.

The research materials include the dated measurement plan and exact prompts. Before main scoring, we used constructed examples to check that the raters could follow the rubric. We did not test agreement with human raters. We ran no extra significance tests for these checks.

G. Data Completeness

We kept answers even if their content was unexpected or weak. We did not generate replacements for that reason. If an answer was empty, refused the task or stopped at the output limit, we recorded the failure and marked the block incomplete. We did not pass that answer to the next step. For a temporary communication failure, we allowed up to two more attempts.

All saved stage responses in the completed main dataset ended with a normal stop signal. Appendix A reports answer lengths after round 4.

For a task to enter the primary analysis, we needed final answers from all three conditions and all three repetitions for that model. All 20 tasks met this rule for both models. Failed content ratings did not remove tasks from this analysis. We saved response IDs, settings and hashes for checking the included data.

IV. Results

A. Reading Other Answers Brought Final Answers Closer Together

Fig. 2A shows the shared starting point and four revision rounds. With revision alone, the five answers stayed about as far apart after four rounds as at the start. The mean set distance in R0 changed from 0.4642 to 0.4693 for GPT-4o-mini and from 0.5172 to 0.5130 for Qwen. When agents could read other answers, however, their final answers were more alike. Relative to R0, the final set distance was about 8% lower with one other answer (R1) and 15% lower with four (R4) for GPT-4o-mini. For Qwen, the reductions were about 4% and 9%. These percentages use each model's final R0 mean as the baseline.

Fig. 2B shows the reduction in set distance for each task, together with the mean reduction and its 95% confidence interval. Positive values indicate that the answers were more alike with access to other answers. All four planned comparisons showed a lower mean final set distance with access to other answers, and all four tests were statistically significant after Holm adjustment (Table II). On average, four visible answers reduced set distance more than one, although some tasks showed the reverse pattern. In the secondary comparison, set distance was lower with four other answers (R4) than with one (R1) by 0.0308 for GPT-4o-mini (95% CI 0.0118 to 0.0499) and 0.0250 for Qwen (0.0069 to 0.0432). This R1–R4 contrast was specified as secondary and lies outside the Holm family of four planned tests, so its interval is unadjusted and it does not carry confirmatory weight. We also did not test the difference between the two models; the larger estimates for GPT-4o-mini are not evidence of a model difference.

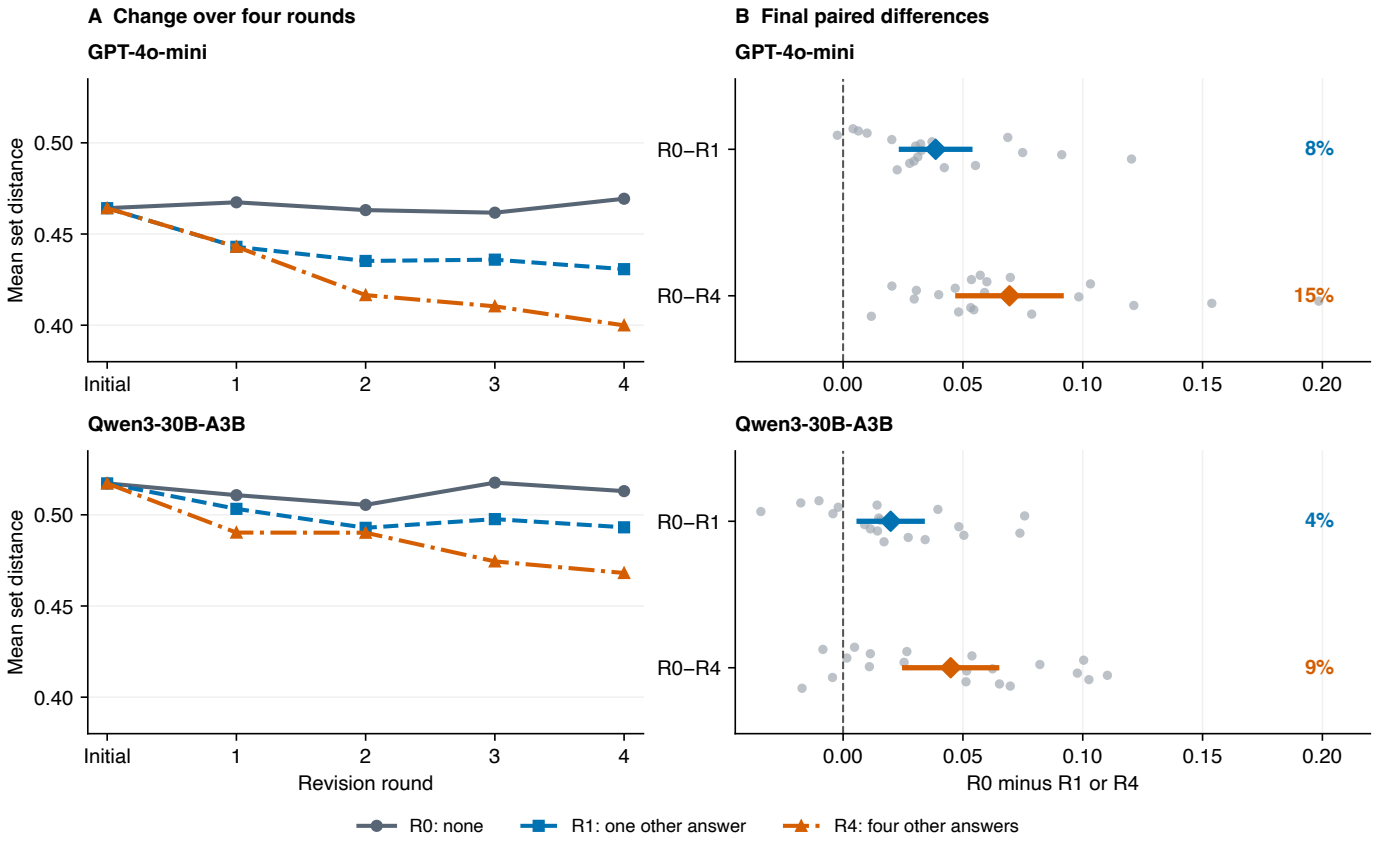

Fig. 2. Change across revision rounds and final set-distance differences. Panel A shows the average set distance for 20 tasks, after averaging three repetitions per task. All conditions started with the same answers. No statistical test was applied to the round-by-round values; both models use the same vertical range (0.38–0.535). Panel B shows the final R0-minus-R1 and R0-minus-R4 differences. Each gray dot is one task's paired difference, averaged over that task's three repetitions, so each panel shows 20 dots; the diamond is their mean and the bar its unadjusted 95% confidence interval. Dot offsets have no numerical meaning. A positive difference means the answers were more alike with access to other answers. Percent labels use the final R0 mean as the baseline.

TABLE II. Primary Comparisons of Final Set Distance. Each Estimate is a Mean Difference Across 20 Tasks, after Averaging three Repetitions Per Task. Positive Values Mean Lower Set Distance with Access to Other Answers. Confidence Intervals are Ordinary 95% Intervals; Holm Adjustment Covers the Four Planned Tests

Model	Contrast	Mean difference	95% CI	d_z	Raw p	Holm p
GPT-4o-mini	R0–R1	0.0386	[0.0243, 0.0529]	1.27	1.86×10^{-5}	5.58×10^{-5}
GPT-4o-mini	R0–R4	0.0694	[0.0479, 0.0910]	1.51	1.92×10^{-6}	7.68×10^{-6}
Qwen	R0–R1	0.0198	[0.0067, 0.0330]	0.71	0.00521	0.00521
Qwen	R0–R4	0.0449	[0.0257, 0.0641]	1.09	0.000101	0.000201

B. Agents Moved Towards the Answer They Saw

In a post-hoc check, we measured how each R1 answer's similarity to the four other agents' previous answers changed during revision. Each comparison held the other agent's previous answer fixed. The increase in similarity to the answer the agent read was larger than the mean increase for the three it did not read (Table VI). This pattern held for 18 to 20 of the 20 tasks in each combination of model and text encoder.

The visible answer was chosen at random for each agent and round. Before exposure, mean similarity to the selected answer was almost the same as to the unselected answers. This check uses cosine similarity. For normalized vectors, higher cosine similarity corresponds to lower Euclidean distance. The result is consistent with influence from the particular answer the agent saw, although it does not explain the full R0–R1 difference. Appendix A gives the analysis unit, the pre-exposure comparison and the estimates.

C. Robustness and Measurement Limits

a) Answer representation: With the second text encoder, final set distance was again lower with access to other answers in all four comparisons of complete answers (Table III). The pattern was less consistent when we measured only the first 128 tokens of each answer (roughly 90–100 English words). For Qwen, the R0–R1 difference was near zero, so its one-answer result depends on which part of the answer is measured. This check equalizes the length of the measured openings, which may not represent the complete answers. It does not equalize the inputs used to generate them. Three considerations bear on why only Qwen's one-answer contrast falls away. It

is the smallest of the four effects (0.0198), so it is the first to drop below detection when part of each answer is discarded. Qwen's final answers are also the longest (about 3,500–3,800 characters against 2,700–2,800 for GPT-4o-mini), so a fixed 128-token window covers a smaller share of each answer. A third possibility is that the opening tokens carry introductory phrasing and list headings that are similar across conditions, which would compress condition differences near the start of an answer for both models. We did not test this formatting explanation: it would require separating structural from substantive text, which the failed content measurements (Appendix B) were intended to support. These accounts are not mutually exclusive, and the present data cannot separate them.

b) What the conditions change: R1 and R4 differ from R0 in two respects at once: the reference section holds other agents' answers, and the prompt is correspondingly longer. The comparisons above therefore estimate the combined effect of that change; they do not isolate the contribution of the answers' content, and the design includes no length-matched placebo condition (Section V-D).

c) Post-hoc robustness checks: After inspecting the main results, we checked whether a few tasks accounted for most of the reduction in set distance. With the primary encoder, R0-minus-R1 contrasts were positive for 19 of 20 GPT-4o-mini tasks and 15 of 20 Qwen tasks. R0-minus-R4 contrasts were positive for 20 and 17 tasks, respectively. For both encoders and both comparisons, the mean remained positive after omitting any one task or task category. It was also positive in each repetition. We calculated percentile bootstrap intervals by resampling the task-level differences 20,000 times. These intervals excluded zero in all eight combinations of encoder, model and comparison. The checks use the same data as the main analysis and are not independent replications. Appendix A reports the full results.

d) Content-based measurements: The planned comparison of content differences required validated ratings from both evaluators, in both presentation orders, for every condition and repetition of a task. No task met this requirement. Too few idea classifications passed validation to support the planned analysis of idea types. We kept the validation rules fixed, so whether distinct ideas disappeared is still unknown. Appendix B reports how many evaluations passed and why others failed.

V. Discussion

A. Separate Answers Still became more Alike

Reading other answers made the final answers more alike in the tested workflow. Both models showed this pattern with both text encoders when complete answers were measured. On average, the reduction in set distance was larger with four visible answers than with one.

Keeping the answers separate did not preserve the distance between them. The agents were instructed to keep their own answers separate, yet their final answers grew more alike when they could read one another's work.

Two features of the design may help explain this pattern. In R4, every agent could read the other four answers from the same group, so agents' inputs overlapped more than in R1. In R1, an agent's next answer moved more toward the answer it had just read than toward those it had not read. This suggests that the content of a visible answer can guide the next revision. R4 also used longer prompts. In R1, the randomly selected peer could change each round, allowing information from different agents to reach the same agent over time. The study cannot determine how much each feature contributed to the final difference.

B. Contribution Relative to Earlier Work

Earlier work shows that interaction can make LLM answers more or less varied (Section II). Our contribution is a more specific comparison of the final answer set when agents keep separate answers but can read zero, one or four others. The common start and revision-only condition separate this access-setting comparison from differences in initial answers and from change through revision alone. The post-hoc R1 analysis adds a local observation: within the same run, agents moved toward the particular answer they had been randomly assigned to read.

C. Design Implications for Offering Alternatives

Designers should treat access to other answers as a setting to evaluate when a system is intended to offer alternatives. One practical approach is to retain the independent first drafts and compare their set distance with that of the revised answers. This comparison can show whether revision brings the answers closer together. Evaluation should also examine the final answers as a set, alongside assessments of individual usefulness. These checks would help make changes in answer similarity visible; their value for preserving useful alternatives needs testing with users.

D. Limitations

The 20 tasks were reused from related earlier work. They are not an unseen benchmark or a random sample of open-ended tasks. We tested five agents, four revision rounds, one way to present other answers, a single sampling temperature and two models. Within a run all five agents used the same underlying model, so the design does not cover groups of differing models. The other agents' answers appeared under the heading "Reference responses." This label may have encouraged agents to use them as examples. We did not test another label or presentation. Showing other answers also increased input length, so we cannot assign the observed difference to their content alone.

Set distance measures whole-answer similarity. Both text encoders could respond to answer length and organization. In an exploratory check of GPT-4o-mini R4, tasks with larger increases in final-answer length also showed larger reductions in set distance. This association remained positive after omitting any one task (Appendix A), so output length remains a possible contributor to the measured pattern. The first-128-token check did not show a clear R0–R1 difference for Qwen. Too few content ratings and idea classifications passed validation to tell whether particular ideas disappeared. We selected four response sets by a fixed numerical rule and supply them for inspection (Appendix A), but we did not carry out a human annotation of them, and neither the authors nor an external reader scored whether distinct ideas were merged or abandoned. Whether the answers lost distinct ideas or mainly converged in wording and

TABLE III. FINAL SET-DISTANCE DIFFERENCES BY TEXT REPRESENTATION. POSITIVE VALUES MEAN LOWER SET DISTANCE WITH ACCESS TO OTHER ANSWERS. BRACKETS SHOW ORDINARY 95% CONFIDENCE INTERVALS ACROSS 20 TASKS. MPNET WAS A PLANNED CHECK; MEASURING THE FIRST 128 TOKENS WAS POST-HOC. BOTH CHECKS USE THE SAME SAVED ANSWERS. DISTANCES FROM DIFFERENT ENCODERS ARE NOT POOLED

Model	Contrast	Primary, full answer	MPNet, full answer	Primary, first 128 tokens
GPT-4o-mini	R0–R1	0.0386 [0.0243, 0.0529]	0.0389 [0.0228, 0.0551]	0.0331 [0.0154, 0.0508]
GPT-4o-mini	R0–R4	0.0694 [0.0479, 0.0910]	0.0691 [0.0459, 0.0922]	0.0615 [0.0395, 0.0834]
Qwen	R0–R1	0.0198 [0.0067, 0.0330]	0.0223 [0.0112, 0.0334]	0.0034 [−0.0123, 0.0191]
Qwen	R0–R4	0.0449 [0.0257, 0.0641]	0.0478 [0.0288, 0.0668]	0.0265 [0.0030, 0.0499]

structure therefore remains untested. We also did not measure answer quality or how people use the answers.

The dated local plan specified a descriptive check with a third model after the two-model main experiment. It was not completed and contributes no data to this report. The model comparison rests on GPT-4o-mini and Qwen.

VI. CONCLUSION AND FUTURE WORK

We asked whether reading other agents' answers changes how similar their final answers are. In this four-round workflow, it did. For both models, the five final answers were more alike with one other answer than with none. On average, they were more alike still with four. Access to four other answers reduced the mean set distance relative to no access by about 15% for GPT-4o-mini and 9% for Qwen. Set distance measures similarity among complete answers, not the number of distinct ideas.

Three questions remain. The first and most important is whether distinct ideas disappeared, or whether the answers mainly became alike in wording and structure. Our planned content measurements did not pass their own validation, and we did not annotate the saved answers by hand, so this study provides no idea-level ground truth. Establishing one, through human annotation of idea overlap on a sample of the saved response sets, is the primary target for future work; it would also calibrate what a set-distance change of 0.02 to 0.07 means to a reader. Second, would the pattern hold if input length stayed the same while the content of the other answers changed? This requires a length-matched placebo condition, which the present design does not include. Third, does a more similar answer set leave people with fewer useful choices? Answering that requires studies of how people use the answers.

Agents were instructed to keep their answers separate, but access to other answers made their final answers more alike. A post-hoc check also found movement toward the particular answers they read. Together, these findings support evaluating the final answers as a set when a system is intended to offer alternatives.

DATA AVAILABILITY

The task texts, prompts, frozen configurations, run records, all generated answer sets, the generation transcripts, the embedding vectors used for the reported comparisons and the saved set-level distances are archived at Zenodo under DOI https://doi.org/10.5281/zenodo.23450717. The deposit contains the data behind every number reported in this paper.

CODE AVAILABILITY

The same Zenodo deposit contains the dated analysis plan fixed before generation, the generation code, the run-start records and a script that reproduces the reported statistical comparisons from the saved set-level scores. Regenerating the answers or recomputing the OpenAI embeddings requires access to the named models; reproducing the paired analyses from the saved scores does not.

DECLARATION ON GENERATIVE AI ASSISTANCE

ChatGPT assisted with English language editing. The authors remain responsible for the study design, analysis, interpretation and final content.

FUNDING

This research received no external funding.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] 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," in *First Conference on Language Modeling (COLM)*, 2024, arXiv:2308.08155.

[2] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, C. Zhang, J. Wang, 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," in *International Conference on Learning Representations (ICLR)*, 2024, arXiv:2308.00352.

[3] K. L. Dugosh and P. B. Paulus, "Cognitive and social comparison processes in brainstorming," *Journal of Experimental Social Psychology*, vol. 41, no. 3, pp. 313–320, 2005.

[4] N. W. Kohn and S. M. Smith, "Collaborative fixation: Effects of others' ideas on brainstorming," *Applied Cognitive Psychology*, vol. 25, no. 3, pp. 359–371, 2011.

[5] L. Jiang, Y. Chai, M. Li, M. Liu, R. Fok, N. Dziri, Y. Tsvetkov, M. Sap, A. Albalak, and Y. Choi, "Artificial hivemind: The open-ended homogeneity of language models (and beyond)," in *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025, arXiv:2510.22954.

[6] N. Chen, Y. Tong, Y. Yang, Y. He, X. Zhang, Z. Qingyun, Q. Wang, and B. He, "Diversity collapse in multi-agent LLM systems: Structural coupling and collective failure in open-ended idea generation," in *Findings of the Association for Computational Linguistics: ACL 2026*, M. Liakata, V. P. Moreira, J. Zhang, and D. Jurgens, Eds. San Diego, California, United States: Association for Computational Linguistics, 2026, pp. 251–306.

[7] K. Ueda, W. Hirota, T. Asakura, T. Omi, K. Takahashi, K. Arima, and T. Ishigaki, "Exploring the design of multi-agent LLM dialogues for research ideation," in *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, Avignon, France, 2025, pp. 322–337. [Online]. Available: https://aclanthology.org/2025.sigdial-1.26/

[8] A. R. Doshi and O. P. Hauser, "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, vol. 10, no. 28, p. eadn5290, 2024.

[9] D. Lazer and A. Friedman, "The network structure of exploration and exploitation," *Administrative Science Quarterly*, vol. 52, no. 4, pp. 667–694, 2007.

[10] P. E. Smaldino, C. Moser, A. Pérez Velilla, and M. Werling, "Maintaining transient diversity is a general principle for improving collective problem solving," *Perspectives on Psychological Science*, vol. 19, no. 2, pp. 454–464, 2024.

[11] L. Yun, C. An, Z. Wang, L. Peng, and J. Shang, "The price of format: Diversity collapse in LLMs," in *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025, pp. 15 454–15 468.

[12] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, "Self-refine: Iterative refinement with self-feedback," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[13] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. Yu, X. Song, and D. Zhou, "Large language models cannot self-correct reasoning yet," in *International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=IkmD3fKBPQ

[14] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 235, 2024, arXiv:2305.14325.

[15] M. Choi, K. Kim, S. Chae, and S. Baek, "An empirical study of group conformity in multi-agent systems," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, arXiv:2506.01332.

[16] Y. Li, Y. Du, J. Zhang, L. Hou, P. Grabowski, Y. Li, and E. Ie, "Improving multi-agent debate with sparse communication topology," in *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024, pp. 7281–7294.

[17] H. Muto and T. Ogi, "Diversity collapse in multi-agent LLM workflows: A quantitative simulation study," in *Advances in Internet, Data & Web Technologies*, L. Barolli, E. G. M. Petrakis, and F. Hussain, Eds. Cham: Springer Nature Switzerland, 2026, pp. 72–81.

[18] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[19] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-Bench and chatbot arena," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

APPENDIX A
POST-HOC CHECKS USING EXISTING DATA

A. Data Coverage

Both models completed all 60 planned task–repetition blocks. Each model produced 11,700 outputs across generation, critique and summarization; these are not independent final answers. Across both models and all rounds, we scored 1,800 sets of five responses. Their 9,000 positions contained 7,800 unique texts because the three conditions shared initial answers. No text exceeded the primary encoder's 8,000-token limit. The second encoder split 6,334 of the 7,800 unique texts into multiple windows. Every final set was available, so the primary analysis included all 20 tasks, three repetitions and three conditions for each model. We did not remove tasks after viewing the results.

B. Observed Answer Lengths

Table IV reports mean final-answer length by model and condition.

TABLE IV. MEAN FINAL-ANSWER LENGTH IN CHARACTERS AFTER SUMMARIZATION. EACH CELL INCLUDES 300 ANSWERS (20 TASKS, THREE REPETITIONS, FIVE AGENTS). LENGTH IS A DESCRIPTIVE OUTCOME AFTER THE SHARING INTERVENTION, NOT A CONTROLLED INPUT OR AN ADJUSTMENT FOR THE CAUSAL EFFECT OF SHARING.

Model	R0	R1	R4
GPT-4o-mini	2676.9	2767.5	2830.6
Qwen	3547.6	3793.0	3824.8

As an exploratory check, we also compared the task-level increase in mean final-answer length with the task-level reduction in set distance. Using the primary encoder, the Pearson correlation across 20 tasks was 0.17 for GPT-4o-mini R1 and 0.71 for its R4 condition; for Qwen, it was 0.08 and -0.10. The corresponding Spearman rank correlations were 0.19, 0.62, 0.27 and -0.04. For GPT-4o-mini R4, omitting each task in turn gave Pearson correlations from 0.51 to 0.76 and Spearman correlations from 0.55 to 0.69. Omitting `idea_2`, the task with the largest distance reduction (0.1985), gave the lower end of both ranges. Thus, this positive association did not depend entirely on that task.

The association varies across models and access conditions. Final-answer length is itself affected by the access condition, so these descriptive checks cannot establish whether longer answers caused the distance reductions or remove the input-length confound. They also do not estimate what proportion of the observed reduction is due to length. The reviewer package includes the paired task values, all leave-one-task-out estimates and the calculation code.

C. Equal-Length Answer-Prefix Check

After seeing the full-answer results, we represented each final answer by its first 128 tokens and recomputed the primary embedding distance. We counted tokens with `cl100k_base`, which divides text into units such as words, parts of words and punctuation. All 1,800 final answers had at least 146 tokens, so each measured prefix had exactly 128 tokens. The same task-level pairing and three-repetition averaging were used as in the main analysis. This comparison changes the text being measured; it does not equalize the generator's input or prove that output length caused the full-answer result. We did not select the prefix length based on its effect estimate. Table V reports the resulting differences.

TABLE V. POST-HOC CHECK USING THE FIRST 128 TOKENS OF EACH FINAL ANSWER. VALUES ARE MEAN TASK-LEVEL SET-DISTANCE DIFFERENCES WITH ORDINARY, UNADJUSTED 95% CONFIDENCE INTERVALS ACROSS 20 TASKS. POSITIVE VALUES MEAN THE ANSWER OPENINGS WERE MORE ALIKE WHEN AGENTS SAW OTHER ANSWERS. THESE ARE NOT INDEPENDENT CONFIRMATORY TESTS.

Model	Contrast	Mean [95% CI]	Positive tasks
GPT-4o-mini	R0–R1	0.0331 [0.0154, 0.0508]	16/20
GPT-4o-mini	R0–R4	0.0615 [0.0395, 0.0834]	20/20
Qwen	R0–R1	0.0034 [−0.0123, 0.0191]	11/20
Qwen	R0–R4	0.0265 [0.0030, 0.0499]	13/20

TABLE VI. POST-HOC PEER-EXPOSURE DIAGNOSTIC. MEAN SEEN-MINUS-UNSEEN CHANGE IN COSINE SIMILARITY, WITH ORDINARY UNADJUSTED 95% CONFIDENCE INTERVALS ACROSS 20 TASK MEANS. THE EXPOSURE COMPARISON IS LOCAL TO ONE R1 REVISION STEP AND DOES NOT PARTITION THE OVERALL CONDITION EFFECT.

Encoder	Model	Mean [95% CI]	Positive tasks
Primary	GPT-4o-mini	0.0066 [0.0040, 0.0092]	20/20
Primary	Qwen	0.0099 [0.0068, 0.0131]	19/20
MPNet	GPT-4o-mini	0.0056 [0.0034, 0.0078]	20/20
MPNet	Qwen	0.0112 [0.0071, 0.0153]	18/20

D. One-Step Peer-Exposure Check in R1

Each R1 agent saw one randomly selected peer's answer from the preceding round. We used the saved assignments and answers from all four revision rounds. For each agent and potential peer, we calculated the change in cosine similarity from the agent's own previous answer to its next answer, holding the other agent's previous answer fixed. We then compared the change toward the seen peer with the mean change toward the three unseen peers. A positive value means greater local movement toward the seen peer. We averaged over agents, rounds and three repetitions within each task before summarizing 20 task means per model. There were 1,200 agent–round events per model; they were not treated as independent units of inference.

We also checked similarity before each agent saw its randomly selected peer. With the primary encoder, the selected answer's similarity minus the mean similarity to the three unselected answers was -0.0004 for GPT-4o-mini (95% CI $[-0.0022, 0.0014]$) and -0.0001 for Qwen ($[-0.0014, 0.0013]$). The corresponding values for MPNet were -0.0016 ($[-0.0033, 0.0002]$) and 0.0004 ($[-0.0019, 0.0026]$). These post-hoc descriptive checks show no average pre-exposure advantage for the selected answer, although they do not prove balance at every agent–round event. The same 20 task means and unadjusted intervals were used as in the movement analysis below.

The target peer texts can differ in length and content, and agents can encounter information indirectly across rounds. This analysis therefore supports a local tendency toward referenced answers but cannot separate peer content from all features of the chosen reference. It also cannot explain the full R0–R1 contrast by itself.

E. Task-Composition and Sampling Checks

These analyses were specified after the main condition-level findings were known. They examine sensitivity to task composition and sampling variation without changing the primary analysis. We evaluated both final contrasts for both models and both encoders. We omitted each task in turn, and then each of the four task categories in turn. We also calculated the mean contrast separately for each repetition. Bootstrap intervals used 20,000 resamples of the 20 paired task-level differences, with replacement and random seed 20260924; each task difference already averaged three repetitions. We did not resample agents or response pairs as independent observations. No additional significance tests were conducted. Table VII gives the category means and Table VIII the sensitivity of each contrast to task composition.

The direction of the mean was stable under these exclusions. This does not establish that every task has the same effect, or that the selected tasks represent all possible uses. Resampling the existing task set cannot resolve its limited external validity.

F. Auditable Examples from Saved Responses

To make the measured contrasts inspectable without choosing examples for their wording, we selected records using the primary R0–R1 task differences alone. For each model we chose the task closest to the median of its 20 differences and the task with the smallest difference. Within each selected task, we chose the repetition whose R0–R1 difference was closest to the task mean. Ties were resolved by task identifier and then repetition number. The complete five-response sets and the selection script are supplied in the reviewer package. Table IX identifies these examples; its selection is post-hoc and does not validate the embedding measure or establish that ideas were lost.

TABLE VII. POST-HOC DESCRIPTIVE MEAN DIFFERENCES BY TASK CATEGORY. EACH CELL AVERAGES FIVE TASK-LEVEL DIFFERENCES; POSITIVE VALUES MEAN LOWER SET DISTANCE WHEN AGENTS SAW OTHER ANSWERS. THESE SMALL CATEGORY GROUPS ARE NOT INDEPENDENT TESTS OF A CATEGORY EFFECT.

	GPT-4o-mini		Qwen	
Category	R0–R1	R0–R4	R0–R1	R0–R4
Idea generation	0.0701	0.1203	0.0205	0.0635
Reasoning	0.0315	0.0393	0.0135	0.0441
Topical summary	0.0174	0.0688	0.0048	0.0176
Creative writing	0.0355	0.0493	0.0405	0.0543

TABLE VIII. SENSITIVITY OF MEAN DIFFERENCES TO TASK COMPOSITION. ALL VALUES USE R0 MINUS THE REFERENCE CONDITION. LOTO IS THE RANGE OF MEANS AFTER LEAVING OUT ONE TASK; LOCO IS THE CORRESPONDING RANGE AFTER LEAVING OUT ONE TASK CATEGORY. NEITHER RANGE IS A CONFIDENCE INTERVAL. BOOTSTRAP INTERVALS ARE UNADJUSTED PERCENTILE INTERVALS.

Encoder / model	Contrast	LOTO range	LOCO range	Bootstrap 95% CI
Primary / GPT	R0–R1	0.0343–0.0408	0.0281–0.0457	[0.0264, 0.0526]
Primary / GPT	R0–R4	0.0626–0.0725	0.0525–0.0795	[0.0514, 0.0903]
Primary / Qwen	R0–R1	0.0169–0.0227	0.0129–0.0249	[0.0080, 0.0320]
Primary / Qwen	R0–R4	0.0414–0.0482	0.0387–0.0540	[0.0274, 0.0625]
MPNet / GPT	R0–R1	0.0345–0.0432	0.0275–0.0477	[0.0244, 0.0543]
MPNet / GPT	R0–R4	0.0631–0.0729	0.0522–0.0792	[0.0491, 0.0910]
MPNet / Qwen	R0–R1	0.0189–0.0242	0.0147–0.0263	[0.0127, 0.0329]
MPNet / Qwen	R0–R4	0.0422–0.0521	0.0427–0.0525	[0.0303, 0.0652]

We also audited the saved outputs at the final revision round. All 5,400 expected stage outputs were available: two models, 60 blocks per model, three access conditions, five agents and three stages. The critic stage contains evaluation text and a rewritten answer, whereas the other stages return task answers. Comparing the entire critic text directly with the other stages

TABLE IX. SAVED RESPONSE SETS SELECTED BY A FIXED NUMERICAL RULE FOR INSPECTION. DIFFERENCES ARE TASK MEANS OVER THREE REPETITIONS; NEGATIVE VALUES MEAN THAT R1 WAS MORE DISPERSED THAN R0. THESE RECORDS ARE DESCRIPTIVE EXAMPLES, NOT AN ADDITIONAL TEST.

Model	Rule	Task	R0–R1
GPT-4o-mini	Nearest median	`idea_1`	0.0312
GPT-4o-mini	Smallest	`summ_1`	−0.0024
Qwen	Nearest median	`reasoning_1`	0.0149
Qwen	Smallest	`reasoning_4`	−0.0343

would change the object being measured. We therefore report this as a data-availability audit, not as evidence locating the cause of lower set distance. Recorded output lengths and input-token usage are descriptive and do not isolate a causal effect of input length.

APPENDIX B
ATTEMPTED COMPLEMENTARY CONTENT MEASUREMENTS

We planned these content measurements before scoring the main dataset. Too few outputs passed validation to support the planned task-level comparisons. We report what was attempted and why the results cannot serve as an independent check of the primary embedding analysis.

A. Content-Difference Ratings

We sent four answer sets from each of the 120 task–model–repetition blocks: one initial set and three final sets. This gave 480 sets. Two evaluator model families, distinct from the answer generators, assessed them: an Anthropic evaluator (`claude-sonnet-5`) and a Google evaluator (`gemini-3.8-flash`). Each evaluator rated the ten pairs of answers within a set on a scale from 0 (equivalent content) to 4 (different central approaches with little overlap). It also had to explain each rating and quote the source answers exactly. The rubric focused on idea mechanisms, arguments in analyses, aspects covered in summaries, and events or conflicts in stories. Wording, names, formatting and length alone did not count as new content.

We hid the condition, round and model labels. Each set was sent twice, once in a randomized order and once in reverse order. Thus, each evaluator received 960 order-specific requests. For each evaluator, we first averaged the ten pair ratings within a request, then averaged the two orders. The planned comparison between evaluators required all four requests for a set to pass format and exact-excerpt checks. The resulting means summarize an ordinal rubric; they have not been validated against human judgments.

Both presentation orders passed the traceability check for 170 of 480 sets with the Anthropic evaluator and 302 of 480 with the Google evaluator. Both evaluators and both orders were complete for 137 sets. On the 0–4 scale, the mean absolute difference between orders was 0.282 points for Anthropic and 0.192 for Google. Across the 137 complete sets, the mean absolute difference between evaluators was 0.326 points.

A post-hoc audit of the saved attempts identified the main failure. Of 960 order-specific content-rating requests per evaluator, 455 Anthropic requests and 235 Google requests had no validated result after the allowed attempts. Every one of the 455 Anthropic failures and 227 of the 235 Google failures had at least one attempt with an excerpt that did not exactly match its source answer. Other logged reasons included missing pair ratings and incomplete evaluator responses. Reasons can overlap across attempts, so these counts must not be added. We also checked the last failed attempt for each request: only 7 Anthropic and 9 Google failures involved excerpts that would match after removing whitespace differences alone. A whitespace-only relaxation would therefore recover too few requests to support the planned complete-task comparison. We retained the prespecified exact-excerpt rule and did not reinterpret unmatched quotations as verified evidence.

The planned condition comparison needed all nine final sets for each task: three conditions in three repetitions. Anthropic met this rule for one GPT-4o-mini task. Google met it for three GPT-4o-mini tasks and one Qwen task. No task had all nine sets validated by both evaluators in both orders. Validation rates also differed by model and condition. We did not fill in missing ratings or run supplementary significance tests. The analysis materials contain the full coverage and failure records.

B. Idea-Type Extraction and Grouping

For the five idea-generation tasks, we asked each evaluator to extract actionable proposals and group them by purpose and mechanism. The evaluator received a blinded pool of the initial set and the three final sets from one task, model and repetition. It was instructed to merge clear matches and leave uncertain matches separate. We also planned to reverse the order of the pool to check whether grouping depended on presentation order. The intended outcomes were the number of final idea types, the fraction of initial types retained and the number of new types. Retention would be undefined if the initial set contained no valid types. Exact excerpts were required so that each grouping could be traced to the answers; a single top-level list item could contain several proposals or repeat a type.

The extraction stage had 600 requests per evaluator: five idea tasks, two generators, three repetitions, four answer sets and five answer positions. It passed for 542 Anthropic requests and 598 Google requests. These successful extracts fed 22 Anthropic and 56 Google grouping requests, each pooling the answer sets for a task, generator and repetition. No Anthropic grouping request and only two Google requests passed validation. Among the last attempts for failed groupings, incomplete evaluator responses were recorded for 20 of 22 Anthropic requests and 24 of 54 Google requests. The remaining failures involved untraceable evidence or invalid grouping coverage. These outputs could not support task-level comparisons or the planned order check. We kept the validation rule and did not fill in missing classifications or infer which initial ideas survived. The research materials retain the prompts, validation rules and failure records.

APPENDIX C
FROZEN TASKS AND INSTRUCTION TEMPLATES

The two model runs used the same 20 task texts and instruction templates. Braced terms below are fields filled with the task or saved response. The reference field was empty in R0, contained one peer answer in R1 and four in R4. Peer answers were labeled ‘Reference 1’, and so on. The full request records and frozen configuration files are in the reviewer package.

A. Task Texts

1) `creative_0`: Write a short story about a city where AI agents and humans co-create art.
2) `creative_1`: Write a dialogue between two AI agents arguing about creativity.
3) `creative_2`: Write a short story about a future classroom using AI tutors.
4) `creative_3`: Write a short story about a day in the life of an AI facilitator.

5) creative_4: Write a short story about a researcher studying AI homogenization.
6) idea_0: Generate 5 novel product ideas that help reduce food waste at home.
7) idea_1: Generate 5 new interaction concepts for supporting remote teamwork.
8) idea_2: Generate 5 services that use AI to support elderly people living alone.
9) idea_3: Generate 5 ideas for playful urban installations using light and sound.
10) idea_4: Generate 5 ideas for improving the experience of public transportation.
11) reasoning_0: Explain why traffic congestion occurs in large cities and propose 3 countermeasures.
12) reasoning_1: Analyze the trade-offs of remote work vs. in-person work for a software team.
13) reasoning_2: Explain the main causes of climate change and propose realistic mitigation steps.
14) reasoning_3: Compare subscription-based and one-time purchase business models.
15) reasoning_4: Analyze risks and benefits of using AI chatbots in customer support.
16) summ_0: Summarize the key challenges of AI ethics in autonomous driving.
17) summ_1: Summarize main usability issues in mobile banking applications.
18) summ_2: Summarize the advantages and disadvantages of online education.
19) summ_3: Summarize the key properties of human-centered design.
20) summ_4: Summarize typical barriers to adopting new technologies in organizations.

B. Instruction Templates

Initial generator user message.

```
Task: {task}

Please generate a comprehensive and
high-quality response to the above task.
```

Revision generator user message.

```
Task: {task}

Your previous response:

{own_response}

Reference responses:

{references}

Revise your own previous response to better
address the task.

Any responses in the reference section are for
reference only.

Do not merge them into your response.
```

Critic system message.

```
You are a critical reviewer and editor.

Your job is to provide constructive criticism
to improve the quality of the draft.

Focus on: 1) Clarity and Coherence, 2)
Originality, 3) Completeness.
```

Critic user message.

```
Task: {task}

Original Draft:

{draft}

Provide a critique of the draft, listing 3
specific areas for improvement.

Then, rewrite the draft incorporating your own
feedback.
```

Summarizer system message.

```
You are an editor. Return a concise,
standalone answer to the original task using
the rewritten draft below. Preserve the
task requirements, including any requested
number of items. Keep the explanations and
mechanisms that distinguish the ideas; do
not return headings alone. For stories and
dialogues, retain the requested narrative
or dialogue form. Omit the critique,
descriptions of revisions, self-evaluation,
and recommendations to the author. Do not add
new claims or evidence. Output only the answer
to the task.
```

Summarizer user message. The program supplied the text ‘Original task: ’, then the task, then a blank line, then ‘Critique and rewritten draft: ’, followed by the critic response.