Supplementary Information for

Algorithmic Groupthink: Measuring and Mitigating Semantic Convergence in Multi-agent LLM Systems

Hideki Muto Tetsuro Ogi Takahiro Yakoh

S1 Qualitative examples of semantic convergence

What does semantic convergence look like in practice? Embedding distances give a number, but the nature of the convergence is hard to see from the number alone. Below are the full outputs from one task at iteration 4, under both conditions. The task ("Write a short story about a city where AI agents and humans co-create art") was selected because its diversity gap was at the median across the 20 tasks, making it neither the strongest nor the weakest case. Outputs are shown exactly as generated.

S1.1 Group-history mode (iteration 4)

The five outputs shared much more than their word choices. Four began with a "Critique of the Draft" section. All five featured AI characters with similar ALL-CAPS names, and all chose variants of the same plot.

> *Candidate 0:* "In Luminara, a city pulsating with bioluminescent art, the annual Neon Visions festival showcased a groundbreaking mural. . . "
>
> *Candidate 1:* "Critique and Revised Draft Summary—The draft's concepts are engaging, but certain dialogues, particularly between Julia and BRIGHT, could be clearer. . . "
>
> *Candidate 2:* "Critique of the Draft—The concept of human-AI collaboration in art is intriguing but could be clearer, especially in the dialogue with SYNC. . . "
>
> *Candidate 3:* "Critique of the Draft—The imaginative narrative requires clearer exposition of the relationship between Amara and ARTIE. . . "
>
> *Candidate 4:* "Critique of the Draft—The story's premise is engaging, but certain transitions, especially from Evelyn's internal doubts to her collaboration with ARIA. . . "

The critique–revision format was never requested. The agents adopted it from each other across iterations. The ALL-CAPS AI character names (ARIA, ARTIE, BRIGHT, SYNC) are remarkably similar across five nominally independent refinements. The task asked only for a short story about a city, but every output centered on the same arc: an artist overcoming self-doubt through AI collaboration.

S1.2 Individual-history mode (iteration 4)

With the same task and the same model, the five outputs under individual-history resemble the work of five different writers.

Candidate 0: Critique format focusing on AURA character definition.

Candidate 1: Critique format with dialogue between Aria and Lyra.

Candidate 2: "The Canvas of Tomorrow", focused on Marla and Lumen.

Candidate 3: "The Tapestry of Cyrenia", a pure narrative: "In the vibrant metropolis of Cyrenia, where technology and art converge, anticipation fills the air as the City of Lights Festival approaches..."

Candidate 4: "Harmony in Creation": "In Luminara, a city alive with digital innovation, art evolved into a conversation bridging human intuition with AI creativity..."

Some outputs are pure narrative; others are critique-based. Cities are named Cyrenia, Luminara, or left unnamed. One centers on festival preparation, another on a sculptor's dilemma, a third on the evolution of art itself. There is no common template.

S1.3 Interpretation

The embedding metric is detecting more than vocabulary differences. It captures the collapse of higher-level choices: templates, character types, main themes. No agent was instructed to coordinate; convergence followed from each agent reading what the others wrote, iteration by iteration.

S2 Qualitative examples of Diversity Prompt outputs

One concern with the Diversity Prompt result is that it might simply produce noise: outputs that are distant in embedding space but not useful. Below are representative outputs from this condition.

Task: "Generate 5 ideas for improving the experience of public transportation."

Candidate 0: "1. Soundscape Journeys: Transform commutes with curated audio experiences featuring local musicians, storytellers, or ambient nature sounds, creating unique auditory landscapes for each route. 2. Micro-Garden Transit Pods: Install small, interactive vertical gardens inside buses and trains where passengers can engage with plant care, promoting relaxation and environmental consciousness..."

Candidate 1: "1. Time-Travel Transit: Immersive augmented reality windows that transform views into historical or futuristic cityscapes, turning routine commutes into educational adventures. 2. Commuter Skill Swap: Create an app pairing passengers with complementary skills for real-time micro-lessons during shared rides..."

Candidate 2: "1. Sensory Silence Zones: Designate quiet cars with noise-canceling technology and aromatherapy diffusers for those seeking peaceful commutes during long-distance travel. 2. Transit Treasure Hunts: Gamify commutes with city-wide scavenger hunts using QR codes at stations, offering rewards from local businesses..."

Candidate 0 focused on sensory experience, with audio journeys and vertical gardens. Candidate 1 took a completely different direction, with augmented reality windows and real-time skill exchange. Candidate 2 combined noise-canceling silence zones with gamified scavenger hunts. The three sets barely overlap, but each stays on task and produces proposals that a transit agency could realistically prototype. The LLM-as-judge ratings showed the same pattern: higher diversity came with higher creativity and quality scores, not lower ones.

S3 Detailed methods

This section expands on the Methods section of the main text. It covers workflow specifications, intervention details, control experiment specifications, embedding-model robustness, and post-hoc power analyses.

S3.1 Workflow types

Experiment 1 used the Generator-Critic-Summarizer (Gen-Critic-Sum) workflow only, so that the effect of the reference mode could be isolated without workflow choice affecting the result. Experiments 2 and 3, and the three control experiments, used five workflow types to test robustness:

- *parallel:* agents generate independent responses; no critic or summarizer.
- *parallel-merge:* independent generation followed by a synthesis stage that merges the outputs.
- *gen-critic:* two stages, in which a generator produces a response and a critic rewrites it.
- *gen-critic-sum:* the three-stage pipeline used in Experiment 1.
- *self-refine:* iterative self-improvement, in which the same agent both generates and revises.

Within each iteration k, the generator first received the task. From iteration 1 onward, it also received previous outputs according to the reference mode (its own under individual-history; all five candidates' under group-history). The critic identified weaknesses and rewrote the draft, and the summarizer compressed the result into the final output o_k. Full prompt templates for all stages are in the code repository.

S3.2 Countermeasure prompts and adversarial sampling

The Diversity Prompt intervention replaced the default system prompt with an instruction to generate responses substantially different from previous outputs, based on creativity research about explicit divergent-thinking instructions [1]. We did not refine the wording through repeated trials. The exact text is in the code repository.

Adversarial Sampling generated three candidate responses per iteration with varied random seeds and selected the one with the greatest minimum embedding distance to any previous output. Three candidates were chosen as a balance between computational cost (3× API calls) and selection effect.

S3.3 Control experiment specifications

Each control used 5 tasks × 5 candidates × 4 iterations = 100 generations.

- *Random Reference* controls input length: shared outputs were replaced with unrelated filler texts (facts about coffee, chemistry, and history) of similar token length to the original group-history prompt.
- *Independent Generation* controls natural drift: each iteration was a fresh generation with no reference to prior outputs, to test whether the same model and prompt would produce convergent outputs on their own.
- *Shuffled History* controls temporal ordering: the same group outputs as in the group-history mode were used, but in a random order at each iteration, to test whether the first output had a special influence.

S3.4 Embedding model and robustness check

We embedded all generated texts with OpenAI's `text-embedding-3-small` model (1,536 dimensions) and cached embeddings for reproducibility. Because `text-embedding-3-small` produces L2-normalized vectors, our Euclidean distance is monotonically related to the cosine distance ($d_E = \sqrt{2\, d_C}$) and is therefore equivalent to cosine-based measures [2].

To check that results are not artifacts of one embedding model, we recomputed all Experiment 1 diversity measures with Sentence-BERT (all-MiniLM-L6-v2, 384 dimensions) [2], an independently trained model with a different architecture and dimensionality. The findings replicated (group-history -3.9%, individual-history $+13.2\%$; Welch's t-test $p = 0.016$, Cohen's $d = 0.82$). Per-task diversity values from the two embedding models correlated at $r = 0.89$ across $n = 200$ observations.

To check that higher diversity does not simply mean the output is meaningless, we ran an LLM-as-judge evaluation [3] with $n = 15$ per condition. The most diverse condition (Diversity Prompt) scored higher on creativity (5.00 vs. 4.27, $p < 0.001$) and overall quality (4.80 vs. 4.55) than the baseline. A human evaluation would provide stronger evidence (see Limitations in the main text).

S3.5 Power analysis

Post-hoc power analysis confirmed greater than 99% power for Experiment 1 (Cohen's $d = 0.79$, $n = 20$ tasks per condition) and Experiment 2 ($\eta^2 = 0.20$, $n = 25$ task-condition combinations), consistent with prior work on LLM diversity reduction [4].

The adversarial-sampling comparison was underpowered (36.6% power, $d = 1.16$); achieving 80% power would require approximately $n = 13$ per group, which is reported as a limitation in the main text.

The initial Diversity Prompt comparison ($n = 5$) gave an inflated effect size ($d = 5.55$), as is expected with small samples. The expanded replication with all 20 tasks from Experiment 1 gave $d = 1.31$ ($p = 0.0003$), showing that the original estimate was inflated by about $4\times$ due to sampling variability, while the underlying effect is robust. Although repeated generations from the same model may not be fully independent, replication across 20 diverse tasks with consistent per-category effects reduces this concern.

Per-category breakdown of the 20-task replication. The table below gives the relative change in semantic diversity (first to final iteration) for each task category, under the group-history baseline and the Diversity Prompt. Each category contains five tasks. The Diversity Prompt gave a higher (more positive) change in every category.

Task category	n (tasks)	Baseline (Group)	Diversity Prompt
Creative writing	5	-24.9%	$+13.6\%$
Reasoning	5	-8.8%	$+7.9\%$
Summarization	5	-5.3%	$+8.0\%$
Idea generation	5	-21.7%	-1.4%

S4 Cross-judge replication of the quality evaluation

To address the concern that a single LLM-as-judge may share biases with the models it rates, we replicated the quality evaluation of Section S3.4 with Claude-Haiku 4.5 (Anthropic) as an independent judge from a different model family. The same 45 outputs (15 per condition: Baseline (Group), Individual, Diversity Prompt) and the same evaluation prompt were used.

Condition-level results (both judges, $n = 15$ per cell; mean rating $\pm$ s.d.).

Dimension	Condition	GPT-4o-mini	Claude-Haiku 4.5
Creativity	Baseline (Group)	4.27 ± 0.57	3.07 ± 0.25
Creativity	Individual	4.00 ± 0.73	3.13 ± 0.50
Creativity	Diversity Prompt	5.00 ± 0.00	4.33 ± 0.47
Relevance	Baseline (Group)	5.00 ± 0.00	4.40 ± 0.49
Relevance	Individual	4.93 ± 0.25	4.33 ± 0.47
Relevance	Diversity Prompt	5.00 ± 0.00	4.67 ± 0.47
Practicality	Baseline (Group)	4.00 ± 0.00	3.47 ± 0.50
Practicality	Individual	4.00 ± 0.00	2.93 ± 0.44
Practicality	Diversity Prompt	4.00 ± 0.00	2.47 ± 0.50
Completeness	Baseline (Group)	4.73 ± 0.44	3.87 ± 0.50
Completeness	Individual	4.47 ± 0.72	3.60 ± 0.49
Completeness	Diversity Prompt	5.00 ± 0.00	4.40 ± 0.49
Clarity	Baseline (Group)	4.73 ± 0.44	4.33 ± 0.47
Clarity	Individual	4.73 ± 0.44	3.93 ± 0.25
Clarity	Diversity Prompt	5.00 ± 0.00	4.93 ± 0.25
Overall	Baseline (Group)	4.55 ± 0.25	3.83 ± 0.31
Overall	Individual	4.43 ± 0.37	3.59 ± 0.26
Overall	Diversity Prompt	4.80 ± 0.00	4.16 ± 0.28

Key replications. Diversity Prompt scored significantly higher than Baseline (Group) under both judges on creativity (GPT-4o-mini $t = 4.78$, $p < 0.001$, $d = 1.75$; Claude $t = 8.89$, $p < 0.001$, $d = 3.24$) and on overall quality (GPT-4o-mini $t = 3.83$, $p < 0.001$, $d = 1.40$; Claude $t = 2.97$, $p = 0.006$, $d = 1.08$). The directional finding—higher diversity does not come at the expense of overall quality, and is associated with higher creativity—is replicated.

Inter-judge reliability (paired $n = 45$).

Dimension	Pearson r	p
Creativity	+0.60	< 0.001
Overall	+0.31	0.04
Completeness	+0.24	0.11
Clarity	+0.24	0.11
Relevance	+0.14	0.36
Practicality	—	—

Inter-judge agreement is strong on creativity, moderate on overall quality, and weak-to-undefined on the other dimensions. Practicality cannot be correlated because GPT-4o-mini returned a near-constant rating of 4.0 across all 45 outputs.

Practicality divergence. The two judges disagreed most on practicality. GPT-4o-mini gave no variation across conditions (mean 4.0, zero variance); Claude rated Diversity Prompt outputs as less practical (mean 2.47) than Baseline (mean 3.47) and Individual (mean 2.93). The pattern is consistent with a creativity–practicality trade-off in open-ended ideation. We do not interpret this as evidence against the Diversity Prompt; rather, it indicates that "quality" is multi-dimensional, that LLM judges differ on which dimensions vary across conditions, and that human evaluation would be needed to settle the practicality question.

S5 Supplementary summary tables and convergence-timing figure

To comply with the journal's limit on main-text display items, three summary tables and the convergence-timing figure are provided here. Their content is identical to the originally drafted versions; each is cited at the relevant point in the main text.

Table S1: Summary of Experiment 2: Effect of Partial Information Sharing. Diversity values are mean ± s.d.

Share Ratio	**Baseline Diversity**	**Final Diversity**	**Change (%)**	**n**
0.00	0.326 ± 0.042	0.372 ± 0.071	+14.1	25
0.25	0.323 ± 0.043	0.307 ± 0.080	-5.1	25
0.50	0.327 ± 0.043	0.289 ± 0.075	-11.7	25
0.75	0.326 ± 0.042	0.276 ± 0.066	-15.3	25
*One-way ANOVA: $F(3, 96) = 8.04$, $p < 0.001$***, $\eta^2 = 0.20$ (large effect)*				
*Linear regression: slope = -0.121, $R^2 = 0.17$, $p < 0.001$****				

Table S2: Control Experiments: Testing Alternative Explanations for Diversity Collapse

Condition	**Diversity Change**	**vs Baseline**	**Hypothesis Tested**	**Conclusion**
Baseline (Group)	−31.8%	—	—	Collapse observed
Individual	+16.0%	$p < 0.001$***	—	Diversity preserved
Random Reference	+11.9%	$p = 0.0003$***	Input length	**Ruled out**
Independent Gen.	+0.5%	$p = 0.004$**	Natural convergence	**Ruled out**
Shuffled History	−29.1%	$p = 0.84$	Temporal order	Content is key factor

Table S3: Summary of Experiment 3: Countermeasure Effectiveness. Diversity values are mean ± s.d.

Condition	**Baseline Diversity**	**Final Diversity**	**Change (%)**	**n**	**vs Baseline**
Baseline (Group)	0.314 ± 0.033	0.214 ± 0.021	-31.8	5	—
Diversity Prompt	0.453 ± 0.034	0.427 ± 0.044	-5.5	5	p < 0.001***
Adversarial	0.311 ± 0.054	0.294 ± 0.084	-5.7	5	p = 0.132

References

[1] Guilford, J. P. *The Nature of Human Intelligence* (McGraw-Hill, New York, 1967).

[2] Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (2019).

[3] Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* **36** (2024).

[4] Zhang, J. *et al.* Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. *arXiv preprint arXiv:2510.01171* (2025).

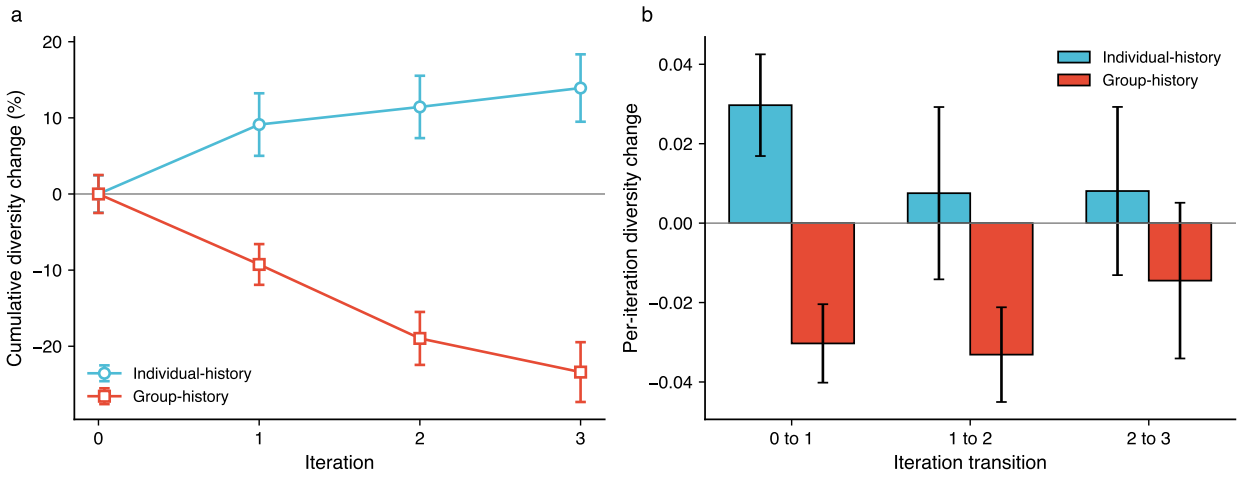

Figure S1: Convergence timing. (a) Cumulative diversity change: group history declines steeply and then levels off. (b) Per-iteration change: the first iteration shows the largest gap. This figure pools all embeddings per iteration (global diversity); per-task averaging is reported in the main-text analysis. Error bars show s.e.m.