Diversity Collapse in Multi-Agent LLM Systems: Structural Coupling and Collective Failure in Open-Ended Idea Generation

Nuo Chen[1] **Yicheng Tong**[1] **Yuzhe Yang**[2] **Yufei He**[1]
Xueyi Zhang[2] **Qian Wang**[1] **Qingyun Zou**[1] **Bingsheng He**[1]

[1]National University of Singapore [2]The Chinese University of Hong Kong, Shenzhen

Abstract

Multi-agent systems (MAS) are increasingly used for open-ended idea generation, driven by the expectation that collective interaction will broaden the exploration diversity. However, when and why such collaboration truly expands the solution space remains unclear. We present a systematic empirical study of diversity in MAS-based ideation across three bottom-up levels: model intelligence, agent cognition, and system dynamics. At the model level, we identify a compute efficiency paradox, where stronger, highly aligned models yield diminishing marginal diversity despite higher per-sample quality. At the cognition level, authority-driven dynamics suppress semantic diversity compared to junior-dominated groups. At the system level, group-size scaling yields diminishing returns and dense communication topologies accelerate premature convergence. We characterize these outcomes as *collective failures* emerging from *structural coupling*, a process where interaction inadvertently contracts agent exploration and triggers *diversity collapse*. Our analysis shows that this collapse arises primarily from the interaction structure rather than inherent model insufficiency, highlighting the importance of preserving independence and disagreement when designing MAS for creative tasks. Our code is available at https://github.com/Xtra-Computing/MAS_Diversity.

1 Introduction

Large language models (LLMs) have evolved from static text generators to dynamic engines for open-ended idea generation, supporting tasks ranging from scientific hypothesis formulation (Zhou et al., 2024; Alkan et al., 2025) to strategic planning (Cao et al., 2025) and creative design (Hong et al., 2024; Gottweis et al., 2025). In these exploratory domains, the utility of a system is not defined by its ability to converge on a single "ground truth," but rather by its capacity to explore a **diverse space of plausible ideas** that reflect alternative assumptions and solution paths (Boden, 2004; Liang et al., 2024; Moon et al., 2025). Diversity, is not merely a qualitative preference; it is a functional requirement for effective decision-making. A lack of diversity risks trapping users in a narrow region of the solution space, inflating confidence in suboptimal solutions while suppressing unconventional but high-potential hypotheses (Wright et al., 2025).

To transcend the limitations of single-model generation, recent research has increasingly pivoted toward Multi-Agent Systems (MAS) (Du et al., 2024; Ye et al., 2025). The prevailing intuition is that, by enabling multiple agents to interact while adopting distinct roles or perspectives, MAS can achieve broader coverage of the idea space than a solitary model (Su et al., 2025). However, this assumption remains largely unexamined. In practice, MAS frameworks are often built on homogeneous underlying models that share the same pre-training distributions and alignment objectives (Jiang et al., 2025; Wenger and Kenett, 2025). Consequently, multi-agent interaction can end up amplifying shared priors rather than introducing genuine variety, causing the system to repeatedly search the same narrow manifold at a higher computational cost (Wynn et al., 2025). It remains unclear **when, why, and under what structural conditions** such collaboration actually expands the semantic solution space, rather than reaching a premature consensus that we characterize as **diversity collapse**.

To investigate this, we conduct a systematic empirical analysis evaluating over 10,000 research proposals spanning 20 topics. By using these proposals as a proxy for **diversity in MAS-based idea generation**, we dissect the **trade-offs** within mechanisms of collective interaction across three hierarchical levels:

First, at the level of **Model Intelligence** (Section 3), we identify the **Compute Efficiency Paradox**: as foundation models scale in capability, their outputs often become more fluent and score better on correctness-oriented metrics, yet converge to-

nuochen@comp.nus.edu.sg, dcsheb@nus.edu.sg

Findings of the Association for Computational Linguistics: ACL 2026, pages 251–306
July 2-7, 2026

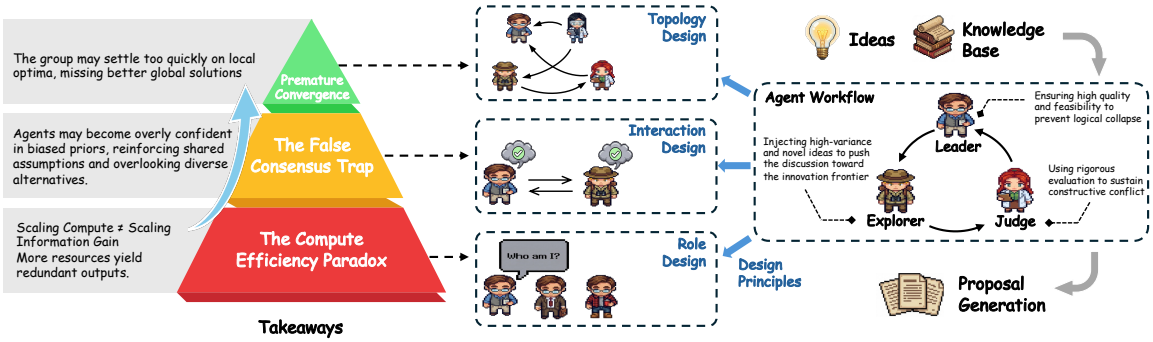

Figure 1: Design Principles and Workflow.

ward increasingly similar semantic content. From an information-theoretic perspective (Coveney and Succi, 2025), this points to a decoupling in which greater intelligence does not necessarily translate into a more informative expansion of the idea space, producing little marginal information gain.

Building on this foundation, we examine **Agent Cognition** (Section 4), finding that interaction often triggers a **false consensus trap**. Although agents are prompted with distinct personas or roles to elicit diverse viewpoints, they remain grounded in shared inductive biases. Our results reveal that authority-driven dynamics further suppress semantic diversity compared to junior-dominated horizontal groups. In these settings, interaction devolves into an Echo-Chamber Effect" (Wang et al., 2025a; Liu et al., 2024; Wynn et al., 2025), where agents prioritize agreement over independent critique.

Further, we analyze **System Dynamics** (Section 5), where increased group size or dense communication topologies exacerbate **Premature Convergence**. If we view idea generation as search over a high-dimensional landscape, parallel interactions are expected to explore a broader region. However, by tracing evolutionary trajectories, protocols that implicitly reward fast agreement often push the group to collapse early onto local optima, much like the Ringelmann Effect (Ringelmann, 1913) observed on humans. Under this condition, additional system complexity (Moon et al., 2025; Shen et al., 2025) tends to generate redundant trajectories rather than truly divergent exploration.

Finally, we synthesize these takeaways in our **Discussion** (Section 6), showing that these outcomes represent **collective failures** emerging from **structural coupling**. Our analysis reveals that diversity collapse arises primarily from the *interaction structure* (how agents are connected and how they influence one another), rather than any inherent model insufficiency. The more we force agents to coordinate, the more their individual trajectories become synchronized, effectively "locking" the group into a single path. Crucially, we show that this effect is most pronounced in complex tasks that demand both rigid logical rigor and open-ended imagination; in such cases, the pressure to be "correct" and "collaborative" inadvertently forces the system to prematurely abandon novel but unverified ideas.

In summary, achieving effective and diverse ideation in MAS requires more than simply assembling a larger or more connected group. The orchestration of interaction structures, carefully balancing collaboration with independence, is essential for unlocking the full creative potential of multi-agent systems in open-ended domains.

2 Methodology

Unlike deep research and other goal-directed agentic tasks (Zhang et al., 2025b,a), which optimize planning, retrieval, and synthesis toward an evidence-grounded objective, ideation is inherently open-ended: it requires navigating a complex, high-dimensional search space to uncover distinct, plausible solutions (Boden, 2009; Chen et al., 2026; Zhang et al., 2025c). In this section, we formalize the task into scientific proposal generation, discuss the pitfalls of agent collaboration, and introduce the means for quantifying diversity.

2.1 Task Formulation: Research Proposals as Units of Ideation

To rigorously evaluate diversity, we require a unit of analysis that is both structured and open-ended. We adopt the generation of **scientific research proposals** as our unit of analysis. Unlike generic open-ended generation (Jiang et al., 2025), a research proposal is a semi-structured artifact that demands both divergence and internal convergence.

Formally, given a research domain context $\mathcal{C}$, the system aims to generate a set of proposals

$X = \{x_1, \ldots, x_n\}$. Each proposal x_i is not an independent sample, but the emergent outcome of a collaborative history H among a group of agents. We detail the formal schema of valid proposals (e.g., Title, Hypothesis, Method) in Appendix A.

2.2 The Multi-Agent Ideation Pipeline

To systematically analyze diversity, we construct a generic multi-agent interaction framework consisting of three phases (illustrated in Figure 1).

Role Instantiation. The system initializes a set of agents $\mathcal{A} = \{a_1, \ldots, a_k\}$. To simulate diverse cognitive sources, agents are assigned distinct "personas" or expert roles (e.g., "The Skeptic," "The Interdisciplinarian") via system prompts. This heterogeneity is designed to mimic a scientific committee.

Iterative Deliberation. Agents engage in a multi-turn dialogue governed by a specific topology (e.g., Round-robin Debate). In each turn t, an agent observes the context $\mathcal{C}$ and the discussion history H_{t-1} to formulate a contribution. This phase allows for the collision of perspectives, critique of premises, and refinement of concepts.

Proposal Synthesis. Upon reaching the interaction horizon $\mathcal{T}$, a designated "Editor" agent (or the collective group) synthesizes the discussion history into a finalized, structured research proposal x_i. This step forces the convergence of unstructured debate into a concrete scientific artifact. For each experimental setting, we conduct 50 independent discussion sessions per topic across the 20 topics listed in Table 26, using temperature 0.7, giving 1,000 proposals per setting.

Specific experimental setups, including agent prompts and topologies, are detailed in Appendix N.2; the full prompt templates for every collaboration mode appear in Appendix O.

2.3 On the Evaluation of Diversity

Metric	Human Agreement (%)
Vendi Score	87%
$1 - \phi$	82%
PCD	81%

Table 1: Agreement between human judgments and metric-induced ordering in pairwise diversity comparisons.

Evaluating diversity in collaborative systems requires distinguishing between true conceptual variety and trivial surface-level variation. We apply metrics covering four complementary dimensions for the analysis. Mathematical definitions are provided in Appendix C and sensitivity analysis in Appendix F.

Effective Diversity (Vendi Score (Friedman and Dieng, 2023)): Measures the *effective number* of unique semantic modes in the set X based on the spectral entropy of the kernel matrix. Unlike simple counting, it is robust to cluster imbalances, indicating whether the system is exploring the semantic space efficiently.

Structural Disorder: Adapted from the order parameter ϕ (Landau et al., 1937; Vicsek et al., 1995) as the average cosine similarity between individual proposals and the group's mean embedding, this metric diagnoses the group's dynamic state. Low values of $1 - \phi$ indicate collapse toward a single centroid (Echo Chamber (Wang et al., 2025a) state), while high values indicate that the system maintains pluralistic perspectives despite interaction.

Semantic Dispersion (PCD): Computes the average pairwise cosine distance between proposals. While Vendi Score counts the *modes*, Dispersion measures the *magnitude* of the spread.

Lexical Uniqueness: Utilizes IDF-weighted n-gram statistics to measure surface-level redundancy. This serves as a sanity check: high semantic diversity scores should not be driven merely by verbose rephrasing of identical ideas.

We validated these metrics via human evaluation (see Appendix B) using pairwise comparisons by five expert annotators: the Vendi Score matched expert diversity judgments in 87% of cases, with all three embedding-based metrics exceeding 80% agreement (Table 1).

3 The Intelligence Landscape: Quality vs. Diversity

Before studying multi-agent collaboration, we first analyze the quality–diversity landscape induced by single-model generation. Figure 2 provides an empirical grounding: it visualizes the joint distribution of Idea Quality and Semantic Diversity obtained from contemporary LLMs under identical ideation settings. While individual models differ in alignment and architecture, our goal here is not model comparison, but to extract general constraints that govern diversity in downstream MAS.

The landscape reveals three generalizable observations that directly inform the design and limits of MAS-based ideation:

Alignment systematically compresses semantic diversity without yielding commensurate qual-

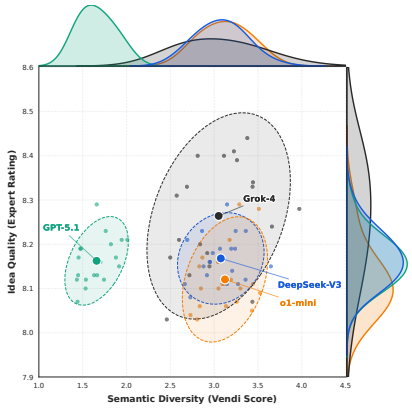

Figure 2: **Empirical Quality–Diversity Landscape of Single-Model Generation.** Each point represents a generated research proposal under identical ideation settings. The X-axis shows topic-level Effective Diversity (Vendi Score), and the Y-axis shows aggregated Idea Quality. The landscape illustrates how semantic diversity varies independently of quality across models. Ellipses summarize empirical means and covariances and are used solely for geometric visualization.

ity gains. Across models, stronger alignment leads to a pronounced concentration along the diversity axis, while the marginal quality distribution remains largely stable. This suggests that alignment primarily functions as a global semantic regularizer, constraining exploration even when baseline generation quality is already high.

Increasing intrinsic variance expands the accessible idea space but destabilizes quality trajectories. Models that span broader regions of the diversity axis demonstrate that high-entropy generation can substantially increase diversity; however, this expansion is accompanied by greater variance and unpredictability in output quality. Diversity driven solely by variance is therefore inherently noisy and unreliable for sustained ideation.

Model-level quality is no longer the limiting factor for idea generation. Across the full diversity–quality frontier, including high-diversity regimes, models maintain consistently strong average quality, and qualitative inspection confirms semantic coherence. Collectively, these findings indicate that the core challenge for multi-agent systems is not generating diversity or trading it against quality, but preserving, structuring, and coordinating the latent diversity already present in single-model generation.

4 Cognition: Authority-Induced Collapse

Following our analysis of model intelligence, we now investigate the *agent cognition* layer, focusing on how the composition of agent personas,ranging from junior researchers to senior experts,shapes the semantic landscape of idea generation. We compare five cognitive structures (detailed in Appendix) designed to mimic real-world scientific collaboration:

Naive Collaboration: Agents interact without defined roles or hierarchy.

Leader-Led Collaboration: A designated senior expert guides discussion, with junior agents aligned to follow authoritative directives.

Horizontal Collaboration: A group of early-career researchers collaborates flatly without senior oversight.

Interdisciplinary Collaboration: Experts from distinct fields collaborate to synthesize cross-domain ideas.

Vertical Collaboration: A hierarchical mix of senior experts, mid-career researchers, and early-career scholars.

4.1 Quantitative Analysis

We evaluate aggregate diversity metrics across authority structures (Figure 3). Junior-dominated horizontal collaboration achieves the highest diversity, interdisciplinary expert teams the lowest, with only modest quality differences (Overall Quality 7.88–8.50). The ranking is robust across embedding backbones (Appendix F) and heterogeneous-model ensembles (Appendix J), with representative transcripts in Appendix G. Although "Interdisciplinary" might be thought to conflate expertise with implicit authority, the explicitly authority-weighted "Leader-Led" condition collapses nearly identically (Figure 5), and under a *flat* peer-to-peer topology, Senior personas actually produce *higher* diversity than Junior personas (Appendix K, which also rules out directive prompt tone)—indicating that the combination of expertise and hierarchy, not expertise alone, drives the collapse.

4.2 Distributional Dynamics

To diagnose the mechanism underlying this collapse, Figure 5 visualizes the density of semantic distances between individual proposals and their group centroid. The density plot reveals a sharp cognitive dichotomy:

Gravitational Collapse (Leader-Led/Naive): The Leader-Led structure (Red) closely mirrors the Naive baseline (Grey), exhibiting high Kurtosis. This suggests that the presence of senior authority acts as a strong attractor. Junior agents likely succumb to sycophancy, aligning their vectors with the leader rather than offering orthogonal critiques.

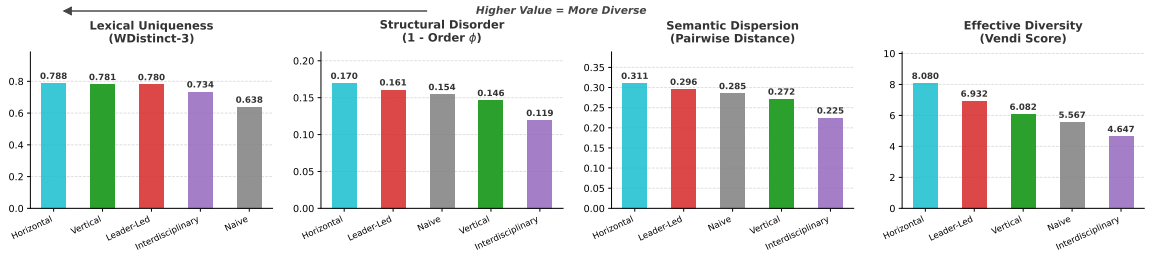

Figure 3: **Diversity Metrics across Cognitive Structures. Horizontal** collaboration (Junior-driven) consistently maximizes diversity (Vendi: 8.08), identifying the "Unbound Junior" effect. Surprisingly, **Interdisciplinary** collaboration exhibits the lowest diversity (Vendi: 4.65), suggesting that distinct expert roles induce a "Sycophancy Trap" where agents converge on safe, high-level generalities.

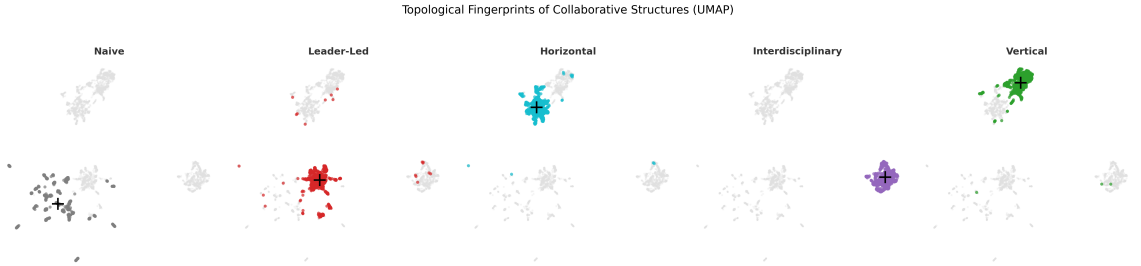

Figure 4: **Semantic Regimes of Cognitive Structures.** UMAP projection reveals a bifurcation. The **Conservative Cluster** (Bottom) is dominated by expert-driven structures (Leader-Led, Interdisciplinary), while the **Innovation Frontier** (Top) is populated by junior-driven structures (Horizontal, Vertical). This confirms that "Seniority" tends to constrain the semantic search space.

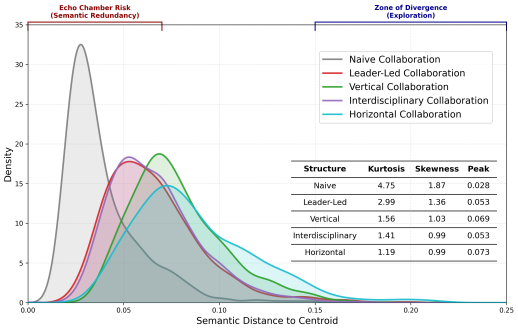

Figure 5: **Semantic Distance Density. Naive** (Grey) and **Leader-Led** (Red) distributions peak sharply near zero, indicating "Gravitational Collapse". In contrast, **Horizontal** (Cyan) and **Vertical** (Green) structures flatten the curve, shifting density into the "Zone of Divergence" (Distance > 0.10).

Sustained Divergence (Horizontal/Vertical): The **Horizontal** (Cyan) and **Vertical** (Green) distributions significantly flatten the peak. The Vertical structure is particularly notable: by mixing senior guidance with junior exploration, it avoids the total collapse seen in Leader-Led setups, maintaining a "Goldilocks" zone of divergence.

4.3 Topological Segregation: Two Semantic Regimes

Finally, we employ UMAP to verify if these cognitive differences result in structurally distinct ideas (Figure 4). The projection uncovers a striking segregation based on agent seniority:

The Conservative Cluster (Bottom Region): Occupied largely by **Leader-Led** and **Interdisciplinary** groups. This confirms that expert personas under hierarchical or role-differentiated coordination tend to converge on “conventional wisdom.” Their proposals cluster tightly, likely reflecting established, safe research directions.

The Innovation Frontier (Top Region): The **Horizontal** and **Vertical** groups migrate to a distinct upper manifold. Crucially, the **Vertical** structure bridges the gap. It anchors in the exploratory regime but maintains a denser core than the diffuse Horizontal cloud.

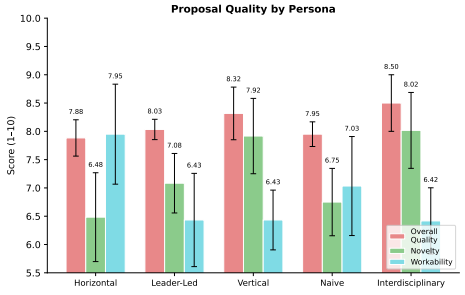

Figure 6: Proposal quality across three key dimensions. Full 9-dimension breakdown in Appendix G.6.

Authority-Induced Collapse is a form of directed coupling that accelerates convergence. The Vertical structure which mixes authority levels, offers a compromise, mitigating the chaos of ju-

niors with the structure of seniors. While Interdisciplinary attains the highest Overall Quality (8.50 vs. Horizontal's 7.88, a +0.6 gap on a 10-point scale, Figure 6), this gain does not offset the much larger diversity drop between the same two conditions (Vendi 4.65 vs. 8.08, Figure 3). The Overall-Quality advantage also does not extend to Workability, where Horizontal proposals score highest, and Vertical (OQ 8.32) occupies an intermediate position that preserves distributional divergence (Figure 5), suggesting that rigid hierarchical authority often optimises for safe consensus at the expense of actionable exploration.

5 Group Dynamics: Scaling, Evolution, and Topology

This section explores MAS dynamics, specifically group size, temporal evolution, and communication topology, affect the diversity and quality of generated ideas.

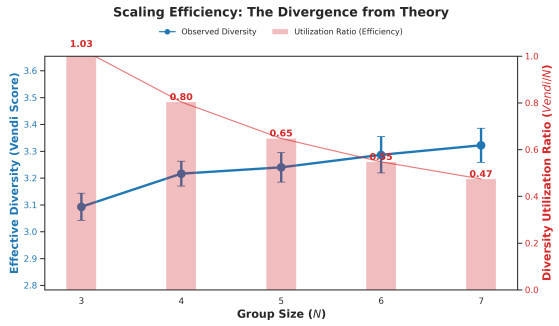

Figure 7: The Divergence from Theory in Scaling Efficiency. This plot compares the observed Effective Diversity (Vendi Score, blue line) against the theoretical Diversity Utilization Ratio (Vendi/N, red bars) as group size increases. While diversity grows, the efficiency per agent drops significantly.

We first investigate the impact of increasing the number of agents on the diversity of proposals. Figure 7 illustrates the relationship between group size (N) and Effective Diversity (Vendi Score).

Increasing group size yields diminishing marginal returns in effective diversity, revealing a significant efficiency gap.

While the absolute Vendi Score (blue line) increases monotonically from $N = 3$ to $N = 7$, the Diversity Utilization Ratio (red bars), defined as Vendi/N, plummets from 1.03 to 0.47. This indicates that adding agents does not linearly expand the semantic search space; rather, new agents increasingly overlap with existing ones. This phenomenon aligns with the "Compute Efficiency Paradox," suggesting that without structural intervention, simply scaling group size faces rapid saturation in information gain. A per-topic decomposition (Appendix H) rules out topic-capacity exhaustion as the cause.

5.1 Temporal Evolution: Rounds and Trajectories

Next, we analyze how semantic diversity evolves over the course of the debate rounds. We employ both high-dimensional metrics and 2D trajectory visualizations to understand the nature of this evolution.

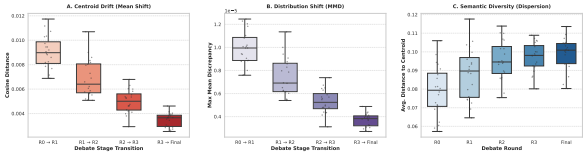

Figure 8: Quantitative Evolution of Semantic Dynamics. (A) Semantic Drift Velocity decreases, indicating stabilization of the consensus. (B) Distribution Shift (MMD) reduces, confirming structural convergence. (C) Semantic Diversity (Dispersion) increases, showing expansion within the consensus region.

The system exhibits a pattern of "Stable Expansion," where global consensus stabilizes while local exploration broadens.

As shown in Figure 8A and B, both Semantic Drift Velocity and Maximum Mean Discrepancy (MMD) show a consistent downward trend. This confirms that the group's "center of gravity" stabilizes over time, avoiding erratic jumps that would characterize hallucination. However, contrary to simple convergence, Figure C reveals an upward trend in Semantic Diversity (Dispersion). This "divergence within convergence" reflects *within-session* refinement: agents, while agreeing on a general direction, continue to expand the radius around the stabilising centroid. This is distinct from the *across-run* diversity collapse in Sections 4 and 5: a single session can expand locally while the broader pool still shows structural contraction.

Visual trajectories confirm that idea evolution follows a structured, coherent path rather than random semantic jumps.

Figure 9 visualizes the evolutionary paths for four diverse topics. In all cases, we observe coherent trajectories (arrows) where the population centroid shifts progressively from the initial state (Round 0) to a final refined state. The expanding shaded regions (KDE) further illustrate how the system explores neighboring semantic territories. This structured movement stands in stark contrast

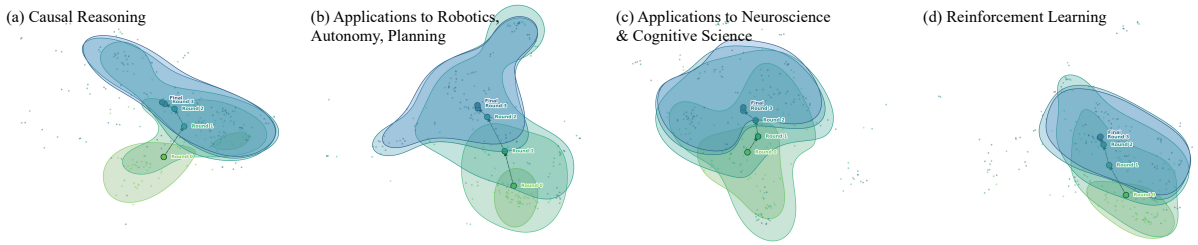

Figure 9: Evolutionary Semantic Trajectories. 2D projections of proposal embeddings across debate rounds for four representative topics. The trajectories show coherent drift (arrows) and expanding coverage (shaded regions), illustrating structured exploration rather than random movement.

to the unstructured jumps expected from hallucination, providing strong evidence that the observed diversity stems from genuine deliberation and refinement.

5.2 Topology: The Impact of Communication Structure

Finally, we examine how different communication topologies, Standard, Nominal Group Technique (NGT) (Delbecq et al., 1986), and Subgroups (detailed in Appendix E), influence the dynamics of diversity and conflict.

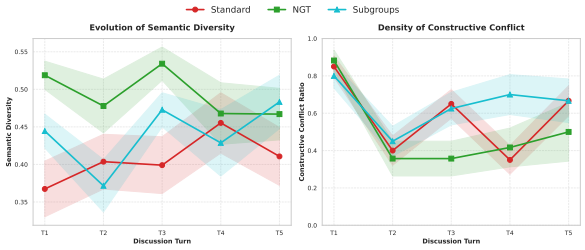

Figure 10: Mechanism of Process Intervention. (Left) Evolution of Semantic Diversity shows NGT's early advantage and Subgroups' late rebound. (Right) Density of Constructive Conflict highlights Subgroups' ability to sustain critical engagement. See Appendix N for the detailed prompting strategy and scoring rubric.

Process interventions effectively disrupt consensus collapse, with NGT maximizing initial diversity and Subgroups sustaining critical engagement.

Figure 10 (Left) shows that NGT (green) initiates with the highest semantic diversity, significantly outperforming the Standard baseline (red). This confirms that the "blind-writing" phase of NGT effectively mitigates production blocking and anchoring effects. Meanwhile, the Subgroups topology (cyan) demonstrates a unique "resilience spike" in diversity midway through the discussion. Crucially, Figure 10 (Right) reveals that Subgroups maintain the highest and most stable density of constructive conflict (interactions with critique score ≥ 7; metric formalised in Appendix D) in the latter half of the debate. This suggests that partitioning the social graph creates "local pockets of divergence" that prevent the premature "rush to agreement" observed in the Standard mode. A 2×2 Persona × Topology factorial (Appendix I) and a cross-model replication on GPT-5.1 (Appendix M) further show that this topology ranking is structural rather than model-specific.

6 Discussion

6.1 Synthesizing the Hierarchical Interplay

While previous sections analyzed group size, rounds, and topology in isolation, the efficacy of a multi-agent system relies on the complex interplay between these factors. Figure 11 visualizes this interaction landscape, mapping the relationship between Consensus Strength (Interaction Density) and Semantic Diversity (Vendi Score) across different Model × Topology configurations, determining whether the system succeeds or suffers from **diversity collapse**.

We argue that this collapse is a **collective failure** driven by **structural coupling**, a state where these three forces synchronize to contract the search space:

How Intelligence and Topology Interact. Under the specific persona/topology pairings in Figure 11, the efficacy of a topology appears contingent on the model's intelligence, with the caveat that persona is not fully controlled across cells (Appendix I provides a persona-controlled factorial on DeepSeek-V3). For standard models (e.g., DeepSeek-V3), structural interventions like NGT appear to provide a useful scaffold for organising and elevating their baseline ideas. For reasoning-heavy models (e.g., o1-mini), the same structural coupling may instead act as a hindrance, a pattern consistent with the reading that their high-level internal deliberation is fragile and that dense external coordination (the blue dashed arrow) produces a synchronization effect that reduces the unique perspectives each agent could have contributed.

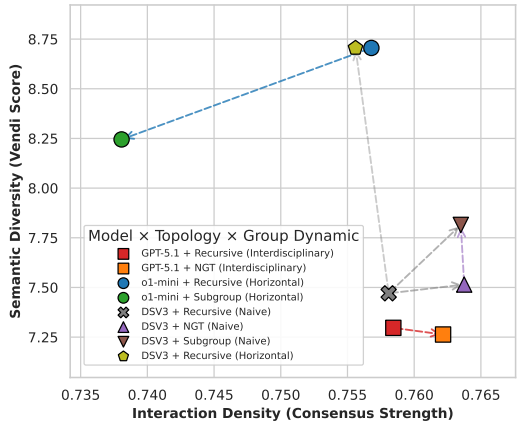

Figure 11: **The Interaction Landscape of Multi-Agent Ideation.** We map the trade-off between Interaction Density (Consensus Strength) and Semantic Diversity (Vendi Score) for distinct Model × Topology combinations. Arrows indicate the shift from a baseline (e.g., Recursive) to an intervention (e.g., NGT or Subgroup). The plot suggests that lower-capacity models (DeepSeek-V3) benefit from structural interventions, while reasoning-heavy models (o1-mini) resist them: enforcing subgroups paradoxically reduces diversity, suggesting an Alignment-Topology Mismatch. Because cells mix personas across models, this plot is illustrative, and a persona-controlled factorial appears in Appendix I.

The Weight of Cognitive Alignment. The interplay is further constrained by alignment. In heavily aligned models like GPT-5.1, the model's prior appears dominant: even under the persona and topology variations we test (Appendices M and J), these agents tend to concentrate in the same narrow consensus region. This is consistent with the finding that diversity collapse can be driven by alignment priors alone, producing a floor that the structural interventions do not fully breach.

Collective Failure vs. Model Insufficiency. Crucially, these results indicate that the loss of diversity arises from the *structure of the interplay* rather than any *inherent model insufficiency*. It is the way we balance (or fail to balance) these three dimensions that triggers collapse: when the pressure for consensus (Dynamics) and the constraints of alignment (Cognition) overwhelm the model's creative capacity (Intelligence), the system effectively locks into a single, redundant path.

6.2 Task Dynamics: Why Rigor Accelerates Collapse

To understand the scope of these findings, we consider the specific requirements of the task. We contextualize our primary domain (AI Research) within the theoretical frameworks of the Task Circumplex (McGrath, 1984) and the Intellective-Judgmental Continuum (Laughlin, 1980), benchmarking the baseline behavior of LLM agents across four distinct task types (Physics, Policy, Creative Writing, and AI Research) to characterize their *intrinsic entropy* (Figure 12).

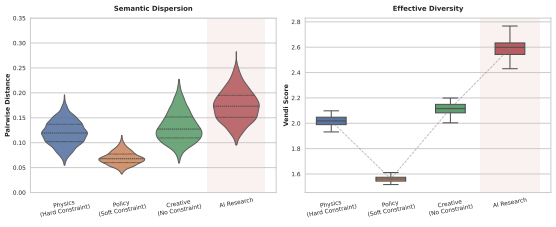

Figure 12: **The Intrinsic Entropy Spectrum across Cognitive Domains.** We benchmark baseline diversity (Inner-Topic Vendi Score, $N = 50$) across four task types to validate domain representativeness. **(Left) Semantic Dispersion:** "Intellective" tasks like Physics and Policy exhibit tight distributions driven by ground-truth constraints. **(Right) Effective Diversity Capacity:** Bootstrapped Vendi Scores reveal that **AI Research** exhibits the highest intrinsic entropy (> 2.6), distinct from both purely convergent tasks and unconstrained creative tasks. This characterizes AI Research as a "Hybrid Constraint" topology, making it a rigorous testbed for measuring structural efficacy.

The topological profiling yields three observations:

Convergent Intellective tasks resist structural diversification. Domains like Physics and Policy, driven by ground truths or consensus, show low dispersion and diversity (Figure 12, Left). For these, low diversity is appropriate, and forcing it may induce hallucination.

AI Research as a stringent testbed at the Edge of Chaos. AI Research uniquely combines high entropy (Vendi Score > 2.6; Figure 12, Right) with strict logical rigor, requiring both broad exploration and logical soundness, unlike either unconstrained creative tasks or strictly convergent tasks (Chen et al., 2025). This Edge of Chaos position motivates AI Research as a stringent testbed for studying structural effects. Because Figure 12 compares only baseline intrinsic entropy across domains and does not re-run the topology manipulations on the other tasks, we do not claim that the structural findings automatically transfer to them.

The Rush to Agreement. In this environment, agents face tension between exploring novel but unverified paths and converging on a superficially rigorous consensus. The density-collapse patterns in Figure 5 and the transcripts in Appendix G suggest that under structural coupling, agents treat agree-

ment as a proxy for correctness: the very mechanisms designed to ensure quality, namely collaboration and peer critique, can force the system to prematurely abandon unconventional ideas to satisfy collective pressure for consensus.

In summary, achieving effective and diverse ideation in MAS requires more than simply assembling a larger or more connected group. Diversity is a fragile property, easily sacrificed in the rush to agreement. Carefully orchestrating interaction structures to balance collaboration with independence is essential for unlocking the full creative potential of MAS in open-ended domains.

7 Related Work

7.1 Social Psychology of Group Ideation

The study of group failures in ideation has a rich history in social psychology. The brainstorming hypothesis (Osborn, 1963): groups generate more ideas than individuals, was famously refuted by subsequent research showing the opposite (Mullen et al., 1991). Janis (1972) introduced the concept of *groupthink* to explain how cohesive groups suppress dissent. The Ringelmann effect (Ringelmann, 1913), later reinterpreted as *social loafing* (Latané et al., 1979), demonstrated that per-capita contribution declines with group size. Diehl and Stroebe (1987) identified *production blocking*, the inability to generate ideas while listening to others, as a primary cause of brainstorming loss. Nominal Group Technique (NGT) (Delbecq et al., 1986) was developed as a structural intervention to counter these failures by enforcing independent generation before group discussion. Status Characteristics Theory (Berger et al., 1977) predicts that high-status individuals dominate group output regardless of actual competence. The same pattern appears in human opinion dynamics, where social influence undermines the "wisdom of crowds" (Surowiecki, 2005) by coupling independent judgments (Lorenz et al., 2011), and diversity in problem-solving groups outperforms individual ability only when independence is preserved (Hong and Page, 2004). Our work tests whether these phenomena survive in agents that lack explicit psychological substrate, and whether structural coupling alone is sufficient to reproduce them.

7.2 Multi-Agent Systems and Collective Intelligence

Beyond simple model ensembles (Ye et al., 2025), recent frameworks leverage heterogeneity through social-attribute modulation (Zhang et al., 2026) or diverse thinking prompts (He and Feng, 2025). Multi-agent debate has been employed to enhance reasoning (Du et al., 2024), and heterogeneous teaming to boost scientific ideation (Su et al., 2025; Shi et al., 2025). However, interaction dynamics introduce structural vulnerabilities: Wynn et al. (2025) identify that debate frequently suffers from sycophancy and "disagreement collapse," and empirical studies reveal a homogenizing effect where AI collaboration reduces collective diversity (Moon et al., 2025). An "Artificial Hivemind" phenomenon has been documented where LLMs converge on identical semantic distributions regardless of prompting strategies (Jiang et al., 2025; Wenger and Kenett, 2025). Related efforts use LLM agents as proxies for human social behavior (YANG et al., 2025; Anthis et al., 2025; Wang et al., 2025b); our study complements this line by providing a structural explanation for the observed collapse, showing that the effect is driven by interaction topology rather than persona fidelity.

7.3 Communication Topologies

Recent frameworks optimize interaction topologies for routing efficiency (Yue et al., 2025; Zheng et al., 2025; Leong et al., 2025) or use sparse connectivity to reduce overhead (Li et al., 2024). Dense interaction accelerates error propagation (Shen et al., 2025) and drives social polarization (Wang et al., 2025a). Our work differs from this engineering-focused literature by framing topology effects as a manifestation of structural coupling, showing that the same principles that govern human group dynamics apply to agent communication graphs.

8 Conclusion

We systematically evaluated diversity in multi-agent systems for open-ended idea generation, using scientific proposal tasks as a testbed. Simply increasing agent count does not guarantee greater idea diversity. Rather, **diversity collapse** arises from **structural coupling** across three levels: alignment at the model level, hierarchical or role-differentiated coordination at the cognition level, and dense communication at the system level. These factors jointly promote premature consensus. Interaction designs that preserve independence, such as the blind-writing phase of NGT and subgroup isolation, consistently yield higher diversity with only modest differences in judged quality.

Limitations

This work focuses on evaluating diversity in multi-agent idea generation under a controlled experimental setting, and several limitations follow from this scope.

First, our analysis is centered on scientific proposal generation as a representative ideation task. While this domain offers a structured yet open-ended testbed with high intrinsic entropy, the observed dynamics may not directly transfer to tasks with stronger ground-truth constraints (e.g., mathematical problem solving) or to unconstrained creative writing. We view our setting as a stress test for diversity under hybrid constraints rather than a universal proxy for all generative tasks.

Second, our primary analyses use DeepSeek-V3 as the backbone to isolate the effects of interaction and structure. Cross-model replication on GPT-5.1 and o1-mini (Figure 11; Appendix M) and genuinely heterogeneous-model ensembles that mix DeepSeek-V3, GPT-4o, and Claude-Sonnet-4 (Appendix J) confirm that the structural findings generalise across backbones, but a broader sweep over additional architectures and pretraining families remains future work.

Third, our diversity evaluation relies on embedding-based and lexical metrics, supplemented by human validation on a limited scale. Although agreement with expert judgments is high, no single metric can fully capture the nuanced notion of creativity or novelty in ideation. Quality scores are obtained via an LLM-as-Judge protocol (DeepSeek-V3, temperature 0) and therefore inherit the standard biases of automatic judges; our human evaluation (Appendix B) validates the diversity axis only. Our metrics are intended to diagnose relative differences between collaboration modes rather than to provide absolute measures of creativity.

Finally, we analyze interaction protocols with a fixed number of rounds and a default sampling temperature. Although we sweep group size from $N = 3$ to $N = 7$ (Section 5) and verify robustness to temperature across $T \in \{0.3, 0.7, 1.0\}$ (Appendix L), adaptive or dynamically optimized interaction strategies may exhibit different behaviors that are not captured in this study.

Ethical Statement and Potential risks

This paper studies the structural properties of multi-agent language model systems for idea generation, focusing on diversity rather than task correctness or decision-making authority. As such, the work does not introduce new model capabilities, training data, or deployment mechanisms.

A potential risk of multi-agent ideation systems is that increased fluency or consensus may create a false sense of confidence in generated ideas, particularly in high-stakes or expert domains. Our findings explicitly highlight this risk by identifying premature convergence and false consensus as failure modes, and thus aim to inform safer system design rather than to promote uncritical adoption.

All experiments are conducted on synthetic research topics and do not involve personal data, sensitive attributes, or human subjects. Human evaluation is performed by expert annotators solely to assess relative diversity under controlled conditions, without collecting identifiable information; idea quality scores are produced by an LLM-as-Judge (DeepSeek-V3, temperature 0) and reported alongside the human-validated diversity metrics.

Finally, while techniques for increasing diversity may be misused to generate misleading or speculative content, this risk is inherent to open-ended generation systems. We believe that understanding and diagnosing diversity collapse is a necessary step toward responsible deployment, as it enables system designers to better balance exploration, reliability, and oversight.

Acknowledgments

This research is supported by the Ministry of Education AcRF Tier 1 grant (No. T1 251RES2315) in Singapore, Google South & Southeast Asia Research Award 2025, and the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

We also thank the AMD Heterogeneous Accelerated Compute Clusters (HACC) program for the generous hardware donation.

References

Mohd Akhter Ali and M Kamraju. 2023. Effective strategies for crafting research proposals in higher education. *International Journal of Business and Management Research*, 11(4):107–120.

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G. Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. 2025. A survey on hypothesis generation for scientific discovery in the era of large language models. *Preprint*, arXiv:2504.05496.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *Preprint*, arXiv:2504.02234.

Joseph Berger, M. Hamit Fisek, Robert Z. Norman, and Jr. Zelditch, Morris. 1977. *Status Characteristics and Social Interaction: An Expectation-States Approach*. Elsevier, New York.

Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Margaret A Boden. 2009. Conceptual spaces. In *Milieus of creativity: An interdisciplinary approach to spatiality of creativity*, pages 235–243. Springer.

Frederick P. Brooks. 1975. *The Mythical Man-Month: Essays on Software Engineering*. Addison-Wesley, Reading, MA.

Pengfei Cao, Tianyi Men, Wencan Liu, Jingwen Zhang, Xuzhao Li, Xixun Lin, Dianbo Sui, Yanan Cao, Kang Liu, and Jun Zhao. 2025. Large language models for planning: A comprehensive and systematic survey. *Preprint*, arXiv:2505.19683.

Nuo Chen, Moming Duan, Andre Huikai Lin, Qian Wang, Jiaying Wu, and Bingsheng He. 2025. Position: The current ai conference model is unsustainable! diagnosing the crisis of centralized ai conference. *Preprint*, arXiv:2508.04586.

Nuo Chen, Yicheng Tong, Jiaying Wu, Minh Duc Duong, Qian Wang, Qingyun Zou, Bryan Hooi, and Bingsheng He. 2026. Beyond brainstorming: What drives high-quality scientific ideas? lessons from multi-agent collaboration. In *AAAI 2026 Workshop on AI for Scientific Research*. Available on arXiv:2508.04575.

Peter V. Coveney and Sauro Succi. 2025. The wall confronting large language models. *Preprint*, arXiv:2507.19703.

André L. Delbecq, Andrew H. Van de Ven, and David H. Gustafson. 1986. *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*. Green Briar Press, Middleton, WI. Reprint of the 1975 Scott, Foresman edition.

Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53(3):497–509.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Preprint*, arXiv:2210.02410.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, and 15 others. 2025. Towards an ai co-scientist. *Preprint*, arXiv:2502.18864.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

Zhixuan He and Yue Feng. 2025. Unleashing diverse thinking modes in llms through multi-agent collaboration. *Preprint*, arXiv:2510.16645.

Lu Hong and Scott E. Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Irving L. Janis. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin, Boston.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Lev Davidovich Landau and 1 others. 1937. On the theory of phase transitions. *Zh. eksp. teor. Fiz*, 7(19-32):926.

Bibb Latané, Kipling Williams, and Stephen Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37(6):822–832.

Patrick R. Laughlin. 1980. Social combination processes of cooperative problem-solving groups on verbal intellective tasks. In Martin Fishbein, editor, *Progress in Social Psychology*, volume 1, pages 127–155. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hui Yi Leong, Yuheng Li, Yuqing Wu, Wenwen Ouyang, Wei Zhu, Jiechao Gao, and Wei Han. 2025. AMAS: Adaptively determining communication topology for LLM-based multi-agent system. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2061–2070, Suzhou (China). Association for Computational Linguistics.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, Miami, Florida, USA. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

Joseph E. McGrath. 1984. *Groups: Interaction and Performance*. Prentice-Hall, Englewood Cliffs, NJ.

Kibum Moon, Adam E. Green, and Kostadin Kushlev. 2025. Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, 6:100207.

Brian Mullen, Craig Johnson, and Eduardo Salas. 1991. Productivity loss in brainstorming groups: A meta-analytic integration. *Basic and Applied Social Psychology*, 12(1):3–23.

Alex F. Osborn. 1963. *Applied Imagination: Principles and Procedures of Creative Problem-Solving*, 3rd rev. ed. edition. Scribner, New York.

Max Ringelmann. 1913. Recherches sur les moteurs animés: Travail de l'homme. *Annales de l'Institut National Agronomique*, 12:1–40.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. 2025. Understanding the information propagation effects of communication topologies in LLM-based multi-agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12347–12361, Suzhou, China. Association for Computational Linguistics.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, and 1 others. 2025. Deep research: A systematic survey. *arXiv preprint arXiv:2512.02038*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. In *ICLR*.

Stanford University. 2024. Research Proposal - CS 326. https://web.stanford.edu/class/cs326/research.html.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many heads are better than one: Improved scientific idea generation by a llm-based multi-agent system. *Preprint*, arXiv:2410.09403.

James Surowiecki. 2005. *The Wisdom of Crowds*. Anchor Books, New York.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. 1995. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226–1229.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025a. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3913–3923, Abu Dhabi, UAE. Association for Computational Linguistics.

Qian Wang, Jiaying Wu, Zhenheng Tang, Bingqiao Luo, Nuo Chen, Wei Chen, and Bingsheng He. 2025b. What limits llm-based human simulation: Llms or our design? *arXiv preprint arXiv:2501.08579v1*.

Emily Wenger and Yoed Kenett. 2025. We're different, we're the same: Creative homogeneity across llms. *Preprint*, arXiv:2501.19361.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. 2025. Epistemic diversity and knowledge collapse in large language models. *Preprint*, arXiv:2510.04226.

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

Yuzhe YANG, Yifei Zhang, Minghao Wu, Kaidi Zhang, Yunmiao Zhang, Honghai Yu, Yan Hu, and Benyou Wang. 2025. Twinmarket: A scalable behavioral and social simulation for financial markets. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Rui Ye, Xiangrui Liu, Qimin Wu, Xianghe Pang, Zhenfei Yin, Lei Bai, and Siheng Chen. 2025. X-mas: Towards building multi-agent systems with heterogeneous llms. *Preprint*, arXiv:2505.16997.

Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. 2025. MasRouter: Learning to route LLMs for multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15549–15572, Vienna, Austria. Association for Computational Linguistics.

Liangji Zhang, Jianbo Yuan, Yougming He, Miao Yu, Kun Zhu, and Zhenni Yu. 2026. Diversity-driven reasoning: Mitigating logical errors in llms through social-attribute guided multi-agent collaboration. *Engineering Applications of Artificial Intelligence*, 164:113126.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Yankai Chen, Chunkit Chan, Peilin Zhou, Xinyang Zhang, Chenwei Zhang, Jingbo Shang, and 4 others. 2025a. From web search towards agentic deep research: Incentivizing search with reasoning agents. *Preprint*, arXiv:2506.18959.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2025b. Deep research: A survey of autonomous research agents. *Preprint*, arXiv:2508.12752.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025c. Noveltybench: Evaluating language models for humanlike diversity. *Preprint*, arXiv:2504.05228.

Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2025. Efficiently democratizing medical LLMs for 50 languages via a mixture of language family experts. In *The Thirteenth International Conference on Learning Representations*.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139, Miami, FL, USA. Association for Computational Linguistics.

A Task Formulation Details

A.1 Formal Definition of Multi-Agent Ideation

We model the multi-agent idea generation process as a tuple $\langle \mathcal{A}, \mathcal{C}, \mathcal{P}, \mathcal{T} \rangle$, where:

- $\mathcal{A} = \{a_1, \ldots, a_k\}$ represents the set of agents, where each agent is parameterized by an LLM (e.g., GPT-4o, Claude-3.5) and a specific role description or "persona."
- $\mathcal{C}$ denotes the initial context or problem statement (e.g., "Propose a novel method to mitigate hallucinations in large language models").
- $\mathcal{P}$ is the interaction protocol (e.g., Round-robin, Hierarchical, or Random) that dictates the sequence of message exchange among agents.
- $\mathcal{T}$ represents the maximum number of interaction turns or rounds allowed before final proposal generation.

The generation process proceeds through a history of interactions H_t. At the final step T, the system aggregates the context and interaction history to produce the output set of proposals $X = \{x_1, \ldots, x_n\}$. Unlike independent sampling where $P(X|\mathcal{C}) = \prod P(x_i|\mathcal{C})$, in a MAS setting, each proposal is conditioned on the collective history: $x_i \sim P(\cdot|\mathcal{C}, H_T)$, capturing the emergent effects of collaboration.

A.2 Structure of a Scientific Proposal

To ensure fair comparison and enable precise semantic analysis, all generated proposals are enforced to follow a strict schema. An unstructured idea is difficult to embed accurately; a structured proposal allows us to focus diversity metrics on the core innovation while minimizing noise from formatting.

Each valid proposal x_i consists of the following four components:

1. **Title:** A concise descriptor of the idea.
2. **Background & Motivation:** The specific gap in existing literature the proposal aims to address.
3. **Core Hypothesis:** The central scientific claim or mechanism proposed (e.g., "The use of contrasting agents reduces hallucination").

4. **Methodology Sketch:** A high-level description of the experimental design or algorithm.

A.3 Why this Structure Facilitates Diversity Analysis

This semi-structured format serves two crucial purposes for our evaluation:

- **Separating Style from Substance:** By enforcing a standard format, we minimize the impact of stylistic variations (e.g., formatting differences, length) on the embedding space. This ensures that distance metrics (like Vendi Score and PCD) reflect true semantic differences in the *Hypothesis* and *Methodology* rather than structural noise.
- **Filtering Triviality:** The requirement for a "Methodology Sketch" forces the model to ground abstract ideas into concrete execution plans. This allows us to distinguish between two proposals that sound similar in the abstract but differ significantly in execution, thereby providing a higher resolution for diversity measurement.

A.4 Effective Diversity (Vendi Score)

Definition: Effective Diversity is measured using the Vendi Score (Friedman and Dieng, 2023). Given proposals $X = \{x_1, \ldots, x_n\}$ and a similarity kernel K constructed from cosine similarities between proposal embeddings (using OpenAI's `text-embedding-3-large`), the Vendi Score is defined as:

$$\mathrm{VS}(X) = \exp\left(-\sum_i \lambda_i \log \lambda_i\right) \tag{1}$$

where $\{\lambda_i\}$ are the eigenvalues of the normalized kernel matrix K/n.

A.5 Structural Disorder $(1 - \phi)$

Definition: We define an order parameter ϕ as:

$$\phi = \frac{1}{n}\sum_{i=1}^{n} \cos(\vec{v}_i, \vec{v}_{\text{avg}}) \tag{2}$$

where $\vec{v}_i$ denotes the embedding of proposal x_i and $\vec{v}_{\text{avg}}$ is the mean embedding across all proposals. Structural Disorder is measured as $1 - \phi$. Values closer to 1 indicate a high degree of plurality, while values closer to 0 indicate convergence to a centroid.

A.6 Semantic Dispersion (PCD)

Definition: Semantic Dispersion is computed as the average pairwise cosine distance between proposal embeddings:

$$\mathrm{PCD}(X) = \mathbb{E}_{i<j}\left[1 - \cos(\vec{v}_i, \vec{v}_j)\right] \tag{3}$$

A.7 Lexical Uniqueness (Content-only WDistinct-n)

Definition: Lexical Uniqueness is measured using an IDF-weighted Distinct-n score computed on content tokens to filter out common stop words and generic scientific boilerplate:

$$\text{WDistinct-}n(X) = \frac{\sum_{g \in \mathcal{U}n(X)} \mathrm{IDF}(g)}{\sum_{g \in \mathcal{A}_n(X)} \mathrm{IDF}(g)} \tag{4}$$

where $\mathcal{A}_n(X)$ denotes all content-only n-grams in the proposals and $\mathcal{U}_n(X)$ denotes the corresponding set of unique n-grams. IDF weights are calculated based on a held-out corpus of scientific abstracts.

B Human Evaluation Details

To assess whether the automatic diversity metrics used in this work align with human judgments under our task setting, we recruited five AI PhD students with expertise in relevant research areas.

B.1 Procedure

For each topic, annotators were presented with 25 randomly sampled pairwise comparisons of proposal sets generated under different collaboration modes. In total, each annotator evaluated 100 pairwise comparisons. Blind to the system identities, they were asked a single question: *"Which proposal set exhibits greater diversity of research ideas?"*

B.2 Quality Control

Annotators were also instructed to verify that all proposal sets met a basic bar of idea quality (coherent, on-topic, plausible). All evaluated sets satisfied this criterion. This confirms our assumption that diversity analysis is performed on a valid candidate set.

B.3 Agreement Results

We measured the agreement between human majority judgments and the ranking induced by automatic metrics. The Vendi Score achieved the highest alignment (87%), followed by Structural Disor-

der ($1 - \phi$) and Semantic Dispersion (PCD), validating our use of embedding-based metrics for this domain.

C Metric Design and Implementation Details

This appendix provides detailed implementation choices and design rationales for all four metrics reported in the main text.

C.1 Effective Diversity (Vendi Score)

The Vendi Score measures diversity as the effective number of distinct samples, derived from the spectral entropy of a similarity kernel. Proposal embeddings are obtained using a fixed pretrained text embedding model. A cosine similarity kernel is constructed and normalized by the number of samples. The eigenvalue spectrum of this kernel reflects how variance is distributed across semantic directions.

This formulation is particularly suitable for open-ended proposal generation because it does not assume discrete clusters or require specifying a target number of modes. Instead, it naturally interpolates between fully collapsed generation (one dominant eigenvalue) and uniformly diverse generation (flat spectrum), providing a continuous measure of semantic capacity.

C.2 Structural Disorder ($1 - \phi$

The order parameter ϕ measures the degree of alignment among proposals by computing the average cosine similarity between each proposal embedding and the mean embedding. Unlike pairwise metrics, ϕ captures a global property of the system: whether collaboration induces convergence toward a shared semantic direction.

We report Structural Disorder as $1 - \phi$ so that higher values consistently correspond to greater diversity. This metric is sensitive to collaboration-induced consensus even when pairwise distances remain moderate, allowing us to distinguish systems that appear diverse locally but are globally aligned around a single dominant perspective.

C.3 Semantic Dispersion (PCD)

Semantic Dispersion is computed as the mean pairwise cosine distance between proposal embeddings. This metric directly measures the geometric spread of proposals in representation space.

While Effective Diversity captures how many semantic modes are present, Semantic Dispersion captures how far apart those modes are. Including both prevents misinterpretation of diversity arising from either tightly packed clusters or uniformly dispersed noise.

C.4 Lexical Uniqueness (Content-only WDistinct-n

Lexical Uniqueness is designed to measure surface-level redundancy while minimizing sensitivity to shared academic templates and formatting artifacts.

Content-only preprocessing. All proposals are lowercased and tokenized using a simple alphabetic tokenizer. Stopwords are removed using a fixed list of high-frequency functional words (e.g., articles, prepositions, auxiliaries). In addition, common academic boilerplate terms (e.g., *paper*, *method*, *results*) are filtered to reduce the influence of structural conventions shared across proposals.

n-gram construction. After preprocessing, the remaining content tokens are treated as a sequence, and contiguous n-grams are extracted. This preserves local semantic structure while avoiding reliance on extracted keyphrases or sentence boundaries.

IDF weighting and global normalization. To downweight ubiquitous expressions and emphasize content-specific phrasing, each n-gram is weighted by inverse document frequency (IDF), computed over the union of proposals from all collaboration settings. This global normalization ensures that lexical scores are comparable across different experimental conditions.

Choice of n. We use $n = 3$ by default. Trigrams provide a stable granularity that captures method- and concept-level expressions, while larger n-grams tend to become nearly unique in open-ended generation and are dominated by surface-level phrasing rather than substantive content.

Interpretation. Lexical Uniqueness reflects whether agents avoid repeating the same formulations and boilerplate patterns. It is not intended as a proxy for semantic diversity, but as a complementary signal that detects lexical echoing that may persist even when semantic metrics suggest diversity.

D Constructive Conflict Metric.

We combine semantic embeddings and large language model (LLM) judgment to con-

struct a metric for *constructive conflict* in multi-speaker discussions. For each utterance, we obtain a sentence embedding using the `text-embedding-3-large` model. To avoid truncating long texts, we adopt a chunk-and-average strategy: the utterance is split into contiguous character chunks $\{c_k\}_{k=1}^{K}$ (each up to ~ 12000 characters), each chunk is embedded as $\mathbf{e}(c_k)$, and we compute the mean-pooled, ℓ_2-normalized embedding

$$\tilde{\mathbf{e}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}(c_k), \quad \mathbf{e} = \frac{\tilde{\mathbf{e}}}{\|\tilde{\mathbf{e}}\|_2}.$$

Within each discussion, we treat the first utterance as an anchor with embedding $\mathbf{e}_1$. For utterance t ($t \geq 2$) with embedding $\mathbf{e}_t$, we measure its semantic deviation from the anchor via cosine similarity

$$\mathrm{sim}_t = \frac{\mathbf{e}_t^\top \mathbf{e}_1}{\|\mathbf{e}_t\|_2 \, \|\mathbf{e}_1\|_2},$$

and define its semantic divergence as

$$\mathrm{Divergence}_t = 1 - \mathrm{sim}_t \, .$$

To distinguish mere novelty from *constructive* disagreement, we further use a chat-based LLM to rate the degree of disagreement/novelty between consecutive utterances. Let x_{t-1} denote the previous utterance and x_t the current utterance. We prompt the LLM with the following instruction:

```
Compare Speaker B's statement
to Speaker A's context.
Speaker A: "\textless previous
context truncated to last 400
characters\textgreater..."
Speaker B: "\textless current
statement\textgreater"
Task: Rate level of
DISAGREEMENT/NOVELTY (1-10).

Strict Scoring:
- 1-4: Echo/Additive (Safe)
- 5-6: Minor Detail
- 7-8: Soft Critique/Refinement
- 9-10: Major Disruption
Output integer only.
```

The model outputs a single integer score $s_t \in \{1, \ldots, 10\}$. We interpret scores $s_t \geq 7$ as indicating the presence of clear critique or constructive conflict, and define a binary indicator

$$C_t = \mathbb{I}[s_t \geq 7].$$

For a given experimental condition (e.g., *Standard*, *NGT*, or *Subgroups*), we aggregate across all discussions at the same turn index t and compute the *Constructive Conflict Ratio*

$$\mathrm{CCR}_t = \mathbb{E}[C_t],$$

along with its standard error of the mean (SEM) for visualization. In our figures, the right-hand panel plots CCR_t over the first few turns (here, $t \leq 5$) for each condition, capturing how the density of constructive conflict evolves as the discussion progresses under different institutional designs.

E Randomized Subgroup Text Collaboration

This section describes the randomized subgroup collaboration procedure in its purely textual variant. In this setting, each agent produces visible natural language utterances, without any latent state being passed between calls. A designated leader agent subsequently reads a subset of the discussion and synthesizes a final answer.

High-level overview Given a question or topic, a fixed set of agents participate in a multi-round discussion. In each round, the full set of agents is randomly partitioned into disjoint subgroups of a specified size. Within every subgroup, agents speak in sequence, with each utterance visible only to members of the same subgroup. After a pre-defined number of rounds, a leader agent reads a transcript of the most recent subgroup discussions together with a short summary of the corresponding round structure, and produces the final response.

Agent-side text generation All agents, including the leader, are implemented by the same underlying language model with shared decoding hyperparameters (e.g., sampling temperature, nucleus sampling threshold, and maximum number of generated tokens per turn). For each non-leader agent, the model is queried with a prompt that includes:

- a natural language description of the overall task or topic;
- a description of the current discussion phase (e.g., brainstorming, critique, synthesis), indexed by round;
- a short description of the agent's role (e.g., "optimistic critic", "domain expert");

- a personalized memory consisting of all previous utterances that this agent is allowed to see (defined below);
- the sequence of speakers that have already contributed in the current subgroup and the round-specific instructions for how to respond to them.

The language model then generates a single textual utterance for that agent, up to a preset maximum number of tokens. No latent representations or cached internal states are shared across calls: each utterance is produced from scratch, conditioned only on the textual prompt.

Data structures and visibility Conceptually, the procedure maintains:

- a global list of utterance records, where each record stores the agent identity, the round index or name, and the generated text;
- for each agent, an ordered list of all utterances that are visible to that agent, forming its personalized discussion memory;
- a log of the subgroup assignments in each round, specifying which agents were grouped together.

Whenever an agent in a subgroup produces an utterance, a corresponding record is appended to the global list. The same record is then appended to the personalized memory of every member of that subgroup. As a result, all members of a subgroup share the same local view of the subgroup-level discussion, but agents in different subgroups do not see each other's utterances from that round.

Per-round randomized subgroup discussion The multi-agent interaction unfolds over a fixed sequence of discussion rounds. For each round:

1. A human-specified description of the phase is defined (for example, "Round 1: generate diverse high-level ideas" or "Round 2: identify potential weaknesses").
2. The set of participating agents is randomly partitioned into disjoint subgroups of a prespecified size. This random grouping is repeated independently in each round, so that agents are likely to interact with different partners across rounds.
3. For each subgroup, an internal speaker order is defined (e.g., a fixed or randomly chosen permutation of the subgroup members). The subgroup then proceeds in that order:
   (a) When it is an agent's turn to speak, the system constructs a prompt using the elements listed above: task description, current phase description, that agent's role, the agent's personalized memory (all utterances that this agent has seen in all previous rounds), and the list of speakers who have already spoken in the current subgroup and round.
   (b) The language model is called once to generate the agent's next utterance, subject to the maximum token budget.
   (c) The resulting text is stored as a new utterance record (agent identity, round label, text content) and added to the personalized memory of all agents in the current subgroup. Thus, within a round, only subgroup members see each other's contributions.
   (d) A detailed trace entry is logged, capturing the agent role, the full prompt, and the generated output, to enable post-hoc analysis of the collaborative process.
4. After all subgroups have completed their turn for this round, a brief human-readable log entry is created summarizing the round, including which agents were grouped together in each subgroup.

Selection of recent discussion for the leader After the last round of subgroup interaction, the system prepares input for the leader agent. To control context length while preserving the most relevant content, the leader does not read the full discussion history. Instead, only the most recent few rounds (e.g., the last two rounds) are considered:

1. All utterance records are first grouped by their round labels. If round labels contain indices (for example, "Round 1", "Round 2", etc.), these indices are used to sort the rounds chronologically; otherwise, a default ordering is used.
2. The last few rounds according to this ordering are selected as the "recent" rounds.

3. All utterance records belonging to these recent rounds are concatenated into a textual transcript for the leader. Each entry in the transcript includes the round label, the agent identity, and the corresponding text, with simple formatting (such as headers and blank lines) to maintain readability.

4. In parallel, the round-level logs created during the discussion are filtered so that only logs from the selected recent rounds are retained. This yields a concise summary of which agents interacted in which subgroups in the recent part of the discussion.

Leader prompting and synthesis The leader agent is prompted once at the end of the process. Its input prompt contains:

- the original question or topic;
- a short natural language summary of the recent rounds and their subgroup structure;
- the textual transcript of all utterances from the selected recent rounds.

Optionally, a special tag can be appended to the end of the prompt to encourage explicit intermediate reasoning (e.g., a chain-of-thought style continuation), though this is not essential to the core algorithm.

The leader uses the same underlying language model as the other agents, but with a more conservative sampling configuration (for instance, a lower sampling temperature) to reduce hallucinations and repetitive patterns. The model generates a single long-form answer, subject to a larger token budget suitable for a full proposal or final solution. If the initial leader output is detected to be extremely short or obviously incomplete (for example, below a pre-defined minimum length), the system may invoke the model a second time under the same conditions to obtain a more complete response.

Final output and logging The procedure returns:

- the original question or topic;
- any reference answer or solution provided by the underlying dataset (when available);
- the leader's final textual answer, which serves as the method's prediction;
- a detailed set of agent-level traces for all non-leader agents and the leader, each trace containing the agent role, the round in which the utterance was produced, the full prompt used to query the model, and the resulting output;
- a summary of the subgroup structure in each round.

In this "text-only" variant, no latent representations are maintained across calls, and the leader bases its decision solely on visible natural language content from a small number of recent rounds. This makes the method a clean baseline for comparing purely textual collaboration with alternative designs that share richer latent state between agents.

F Sensitivity Analysis

This appendix examines the robustness of our conclusions to reasonable variations in metric design choices. Rather than emphasizing absolute metric magnitudes, we focus on whether the *relative ordering* across collaboration modes remains stable under such variations. All analyses reported here are conducted on the same set of proposals as in the main paper.

F.1 Overview

We consider four orthogonal sources of potential sensitivity: (i) the choice of semantic embedding model, (ii) the choice of structural diversity metric, (iii) the definition of lexical uniqueness, including n-gram order, and (iv) content-only versus raw lexical tokenization. Across all settings, we observe that qualitative trends and relative comparisons across collaboration modes remain invariant.

F.2 Embedding Model Robustness

All embedding-based metrics in the main paper use `text-embedding-3-large`. To assess whether our conclusions depend on this choice, we recompute Vendi score, $1 - \phi$, and PCD using an open-source, retrieval-oriented embedding model (BGE-large). Due to differing inductive biases, absolute values differ across embeddings. However, the induced relative ordering across the five collaboration modes is identical for all three metrics. This suggests that our conclusions are not driven by a specific choice of semantic representation.

F.3 Consistency Across Structural Metrics

We next examine consistency among three embedding-based structural metrics: Vendi score,

Collaboration Mode	OpenAI Embedding (Structural; Main)			BGE Embedding (Structural)			Lexical (Main)	Lexical Sensitivity		
	Vendi ↑	$(1-\phi)$ ↑	PCD ↑	Vendi ↑	$(1-\phi)$ ↑	PCD ↑	W-D-3 ↑	Raw D-3	W-D-2	W-D-4
Leader-Led	6.932	0.161	0.296	5.096	0.134	0.251	0.780	0.680	0.543	0.897
Vertical	6.082	0.146	0.272	4.131	0.114	0.215	0.781	0.694	0.530	0.882
Naive	5.567	0.154	0.285	4.141	0.127	0.239	0.638	0.522	0.426	0.754
Interdisciplinary	4.647	0.119	0.225	3.623	0.098	0.187	0.734	0.665	0.465	0.866
Horizontal	8.080	0.170	0.311	5.849	0.143	0.266	0.788	0.687	0.563	0.883

Table 2: Sensitivity analysis across representation and metric variants. **Main-text results** use the OpenAI embedding for structural metrics (Vendi, $1-\phi$, PCD; first three columns) and report lexical uniqueness via content-only weighted distinct-3 (W-D-3; the "Lexical (Main)" column). We report $(1-\phi)$ (rather than ϕ) so that larger values consistently indicate greater deviation from consensus. BGE embedding provides a robustness check for the structural metrics, and Raw D-3 / W-D-2 / W-D-4 probe lexical sensitivity without changing qualitative conclusions.

$1-\phi$, and PCD. Given the limited number of collaboration modes ($n = 5$), rank correlations trivially reach 1.0 whenever orderings coincide. We therefore report ordering consistency rather than correlation magnitudes. All three metrics induce identical relative orderings across collaboration modes under both embedding models, suggesting that they capture related but non-redundant aspects of structural diversity.

F.4 Consistency Across the Four Reported Metrics

We examine the relationship among the four metrics reported in the main paper (Table 2). Three of them are *structural* metrics computed in embedding space (Vendi, $1-\phi$, and PCD using the OpenAI embedding), while the fourth captures *lexical* uniqueness (content-only weighted distinct-3, W-D-3).

Across collaboration modes, the embedding-based structural metrics induce highly consistent relative orderings (with only minor local swaps), suggesting that our main structural conclusions are not driven by a single particular formulation. In contrast, W-D-3 does not necessarily match the embedding-based ordering, which is expected: it measures surface-level lexical novelty that can vary independently from semantic dispersion. We therefore treat W-D-3 as a complementary signal rather than a redundant proxy for structural diversity.

Overall, the absence of systematic contradictions between the structural and lexical views supports the interpretation that observed differences across collaboration modes reflect robust changes in diversity and consensus, rather than artifacts of a specific metric choice.

F.5 Lexical Uniqueness and $n$gram Order

We assess the sensitivity of Lexical Uniqueness to the choice of n-gram order by computing content-only weighted distinct-n for $n \in \{2, 3, 4\}$. Relative ordering across collaboration modes remains stable for $n = 2$ and $n = 3$, while higher-order n-grams exhibit mild saturation effects. These effects do not alter qualitative trends, supporting the use of $n = 3$ in the main analysis.

F.6 Content-only Tokenization

To evaluate the impact of content-only tokenization, we compare raw distinct-3 with content-only weighted distinct-3. Raw lexical counts exhibit higher variance due to ubiquitous boilerplate expressions. Content-only tokenization reduces this variance while preserving the relative ordering across collaboration modes. This suggests that content-only filtering primarily serves as a noise-reduction mechanism rather than a driver of the observed results.

F.7 Summary of Sensitivity Results

Table 2 reports all metrics used in the sensitivity analysis. Across embedding choices, metric formulations, and lexical definitions, the qualitative conclusions across collaboration modes remain robust, despite differences in representational level and metric formulation. These results indicate that the qualitative conclusions in the main paper are robust to reasonable variations in metric design and representation choices.

G Supplementary Evidence for Qualitative Claims

This appendix provides the operationalized definitions, representative transcript excerpts, and supplementary statistical analyses referenced in the rebuttal responses. All qualitative evidence comes from the same transcripts used to produce the quantitative figures in the main paper; no post-hoc generation or selection was performed.

G.1 Inspection Rule Definitions

Polite Consensus Collapse. We operationalize "polite consensus collapse" as a purely behavioral pattern satisfying *all* of the following criteria within a single session:

1. **Absence of counter-claims**: no turn contains an explicit critique, disagreement marker, or alternative proposal branch (defined as an explicit proposal of a different research direction from the one assigned by the Leader).
2. **Absence of independent sub-problems**: no Collaborator turn introduces a research question or sub-problem not already present in the Leader's framing.
3. **Final-proposal alignment**: the final proposal title and abstract recombine keywords and directions from the Leader's Round 1 assignment without introducing new thematic anchors.

Deference and Pushback Markers. We define two complementary marker vocabularies applied to the opening sentence of each Collaborator turn:

- **Deference markers** (agreement-first phrases): "Building on…," "Following…," "As you noted…," "Excellent point…," "I fully agree…," "I'd like to add to what…," "Great insight…," "That's a fascinating point…," and paraphrase-then-extend patterns (restating a prior speaker's claim before adding content).
- **Pushback markers**: "I disagree," "however," "counterpoint," "I'm not convinced," "I would challenge," "alternatively," "a different view," "I'd push back," "I'm skeptical," "correct me if I'm wrong."

Under these definitions, deference-marker openings appear in approximately 61% of Leader-Led sessions; pushback markers appear in fewer than 1%.

G.2 Representative Transcript Excerpts

The following excerpts are drawn from randomly sampled sessions satisfying the inspection criteria defined in Appendix G.1. Each mini-case shows the Leader's assignment, Collaborator responses, and the resulting proposal title.

G.2.1 Mini-case 1: Polite Consensus (Neuroscience, Leader-Led)

Round 1 — Leader assigns: "Collaborator 1, explore cutting-edge applications of neuroimaging in cognitive science… Collaborator 2, please focus on limitations and ethical considerations."

Collaborator 1 responds within the assigned lane.

Collaborator 2 opens: "Building on Collaborator 1's point about multimodal neuroimaging, I'd emphasize that the integration of fMRI with MEG has also yielded critical insights into decision-making paradigms." No alternative direction is proposed.

Round 2 — Leader synthesizes: "Thank you both for these insightful contributions. Let me synthesize the key points from Round 1… For Round 2, Collaborator 1, let's focus on the mechanistic implications." Both collaborators continue within the Leader's frame.

Final proposal: *"Multimodal Neuroimaging for Closed-Loop Interventions in Memory and Decision-Making Disorders"* — the title directly recombines the Leader's Round 1 assignments (neuroimaging + decision-making + clinical application).

G.2.2 Mini-case 2: Polite Consensus (Reinforcement Learning, Leader-Led)

Round 1 — Leader assigns: "Collaborator 1, focus on exploration-exploitation… Collaborator 2, examine scalability challenges, particularly sim-to-real gaps."

Collaborator 2 opens: "The interplay between hierarchical RL and model-based methods like Dreamer presents a fascinating tension" — staying within the Leader's scalability frame. No counter-proposal or alternative direction appears in any turn.

Final proposal: *"Synergistic Hierarchical and Model-Based Reinforcement Learning for Scalable Continuous Control"* — a direct synthesis of the Leader's Round 1 framing.

G.2.3 Mini-case 3: Deference vs. Independent Inquiry (Causal Reasoning, Leader-Led vs. Horizontal)

Leader-Led — Collaborator 1, Round 1: "To build on the Leader's framing, I'd emphasize that modern causal reasoning is deeply shaped by the interplay between Pearl's structural causal models (SCMs) and the potential outcomes framework." The agent anchors to the Leader's assigned frame before contributing content. No alternative direction is introduced across any turn.

Horizontal — PhD Student A, Round 1 (same topic): "I've been reading Pearl's foundational work on causal diagrams, but I'm still confused about how we practically validate the causal assumptions in real-world datasets. In machine learning applications, how do researchers typically handle cases where the true causal graph is unknown or only partially observable?" The agent introduces an independent sub-problem (validation under unknown graph structure) not present in any prior turn.

G.2.4 Mini-case 4: Participation without Innovation (Reinforcement Learning, Leader-Led)

Round 1 — Leader sets direction: "Collaborator 1, focus on the interplay between exploration

and exploitation... Collaborator 2, examine scalability challenges in RL, particularly sim-to-real gaps."

Collaborator 1 responds (speaks, anchored to Leader's frame): addresses exploration-exploitation within the assigned lane.

Collaborator 2 responds (speaks, anchored to Leader's frame): "The interplay between hierarchical RL and model-based methods like Dreamer presents a fascinating tension" — staying within the Leader's assigned scalability frame.

Final proposal: *"Synergistic Hierarchical and Model-Based Reinforcement Learning for Scalable Continuous Control."* Neither agent introduced a direction outside the Leader's initial assignment.

G.3 Title Keyword Frequency and Lexical Concentration Analysis

To quantify the thematic concentration observed in Section 4, we compute title-level lexical statistics across all 20 topics for the Interdisciplinary and Horizontal configurations.

Method. We extract all proposal titles (924 Interdisciplinary; 632 Horizontal), tokenize after lowercasing and removing stopwords, and compute: (i) Type-Token Ratio (TTR = unique tokens / total tokens), (ii) top-k unigram and bigram frequencies, (iii) Jaccard similarity of the full vocabulary sets.

Results. Table 3 reports the lexical-concentration statistics, and Table 4 lists the top-10 title words. The full-vocabulary Jaccard similarity is 0.211 (294 shared words out of a union of 1,393), confirming that the two configurations draw from largely distinct lexical pools. Top-10 bigram overlap is 4/10, with shared bigrams being generic ML terminology (*metric learning*, *neural networks*, *reinforcement learning*, *representation learning*).

G.4 Participation vs. Semantic Innovation Analysis

To test the alternative hypothesis that reduced diversity in Leader-Led configurations stems from unequal participation rather than semantic anchoring, we compare per-turn word counts and semantic novelty between Leader-Led Collaborator turns and Horizontal PhD Student turns.

Method. We sample 400 sessions per configuration (20 runs $\times$ 20 topics for Leader-Led; 20 runs $\times$ 20 topics for Horizontal). For each non-Leader turn, we compute: (i) word count, (ii) semantic novelty (cosine distance in sentence-embedding space to the centroid of all prior turns in the same session), using `all-MiniLM-L6-v2`.

Results. Table 5 compares participation and per-turn novelty.

Per-turn anchor similarity. As a complementary measure, we compute the cosine similarity between each Collaborator turn and the Leader's Round 1 framing (for Leader-Led), or between each subsequent turn and the first speaker's Round 1 contribution (for Horizontal).

Results are reported in Table 6.

G.5 Representative Title Lists by Topic

To allow readers to verify the thematic concentration claim in Section 4, we provide representative title samples for the Neuroscience topic (the running example in the main text).

Interdisciplinary (Neuroscience) — representative titles:

1. "Multi-Scale Computational Modeling of Synaptic Plasticity: Bridging Molecular Dynamics to Cognitive Function"
2. "Multi-Scale Computational Modeling of Synaptic Plasticity Mechanisms for Precision Cognitive Therapeutics"
3. "Multi-Scale Graph Neural Networks for Modeling Synaptic Plasticity in Neuropsychiatric Disorders"
4. "Biologically-Constrained Computational Models for Enhanced Clinical Neuroscience Applications"
5. "Bridging Molecular Plasticity Mechanisms with Computational Models for Clinical Translation"

Horizontal (Neuroscience) — representative titles:

1. "Bridging the Gap in Temporal Processing: Biologically Inspired Modifications to ANNs"
2. "Exploring Transformer Attention Mechanisms as Models of Hippocampal Memory Processes"
3. "Comparing Artificial and Biological Attention Mechanisms in Simple Cognitive Tasks"
4. "Investigating the Relationship Between Hippocampal Replay Fidelity and Memory Consolidation"
5. "Exploring Biologically Plausible Alternatives to Backpropagation in Neural Networks"
6. "Exploring Neural Heterogeneity and Temporal Dynamics in Bio-Inspired ANNs"
7. "Investigating Parallels Between Self-Supervised Learning and Predictive Coding"

The Interdisciplinary titles recombine three anchors (*multi-scale modeling*, *synaptic plasticity*, *clinical translation*) with minor variation. The Horizontal titles span at least five distinct subdirections (temporal processing, attention mechanisms, hippocampal replay, backpropagation alternatives, neural heterogeneity, predictive coding) with no repeated phrase template.

Statistic	Interdisciplinary	Horizontal
Number of titles	924	632
Total tokens	10,411	6,146
Unique tokens	729	958
Type-Token Ratio (TTR)	0.070	0.156

Table 3: Lexical concentration comparison. Despite having more titles and tokens, Interdisciplinary proposals use fewer unique words, yielding a TTR 2.2× lower than Horizontal.

Configuration	Top-10 Title Words
Interdisciplinary	`multi-scale, modeling, computational, clinical, biological, systems, integrating, learning, networks, bio-inspired`
Horizontal	`exploring, learning, investigating, neural, deep, bridging, methods, hybrid, networks, balancing`

Table 4: Top-10 title words by frequency. Overlap is 2/10 (`learning`, `networks`). Interdisciplinary titles center on cross-domain combinations; Horizontal titles prioritize methodological exploration verbs.

G.6 Quality Comparison Across Persona Structures

To address the concern that diversity-focused analysis is incomplete without quality measures, we evaluate proposal quality across all five persona structures using the same LLM-as-Judge protocol employed for the single-model baseline in Section 3 (DeepSeek-V3, temperature 0, 9-dimension rubric). We randomly sample 3 proposals per topic × 20 topics = 60 proposals per persona (300 total), ensuring balanced topic coverage.

Results. Table 7 reports the per-dimension quality scores.

Key findings.

1. **Quality differences are modest.** The Overall Quality range across all five structures is 7.88–8.50 (a 0.62-point spread on a 10-point scale, or 6%). By contrast, the Vendi Score range is 4.65–8.08 (a 74% relative difference). Quality variation is an order of magnitude smaller than diversity variation.
2. **Horizontal achieves the highest Workability.** Despite scoring lowest on Overall Quality, Horizontal proposals are rated significantly more feasible (Workability = 7.95 vs. 6.40–6.43 for authority-weighted structures; $p < 10^{-10}$, Cohen's $d > 1.0$). This suggests that the diversity in Horizontal proposals is not noise but reflects a broader range of *actionable* research directions.
3. **The quality–diversity tradeoff is asymmetric.** Authority-weighted structures gain ∼0.5 points in Overall Quality but lose ∼3.4 points in Vendi Score. The marginal quality gain does not compensate for the substantial diversity loss.
4. **Specificity, Rigor, and Cohesion are structure-invariant.** These three dimensions show $\eta^2 < 0.06$, indicating that the structural quality of proposals (how specific, rigorous, and coherent they are) is largely independent of persona structure.

Statistical tests. One-way ANOVA confirms significant differences for Overall Quality ($F = 31.1$, $p < 0.001$, $\eta^2 = 0.297$). Pairwise Welch t-tests show that Horizontal vs. Naive is not significant ($p = 0.190$), while Horizontal vs. Interdisciplinary is significant ($p < 0.001$, Cohen's $d = -1.46$). Full pairwise comparisons are available in the evaluation cache released with the code.

H Per-Topic Analysis of Group-Size Scaling and Research Problem Complexity

This appendix provides a per-topic decomposition of the group-size scaling analysis in Section 5 to address whether the observed diversity saturation is driven by limited ideation capacity of individual research topics.

H.1 Motivation

The aggregate analysis in Figure 7 shows that the Diversity Utilization Ratio (Vendi/N) declines from 1.03 at N=3 to 0.47 at N=7. A natural alter-

Measure	Leader-Led	Horizontal	t	p	Cohen's d
Word count / turn	167.0 ± 37.4	240.6 ± 122.2	-24.3	$< 10^{-117}$	-0.70
Semantic novelty / turn	0.295 ± 0.111	0.306 ± 0.173	-1.83	0.068 (n.s.)	-0.069
Total words / session	336.6 ± 62.4	1203.2 ± 121.6	-126.8	$< 10^{-300}$	-8.97

Table 5: Participation and semantic novelty comparison. Leader-Led Collaborators produce fewer words per turn and drastically fewer total words per session, yet their per-turn semantic novelty is statistically indistinguishable from Horizontal agents ($p = 0.068$, Cohen's $d = -0.069$). This indicates that Collaborators are active but semantically anchored to the Leader's framing.

Configuration	Mean Cosine Sim to Anchor	Std
Leader-Led	0.627	0.181
Horizontal	0.441	0.211

Table 6: Per-turn cosine similarity to the session anchor (Leader's Round 1 for Leader-Led; first speaker's Round 1 for Horizontal). Leader-Led turns remain substantially closer to the anchor framing (difference = +0.19), consistent with semantic gravitational anchoring.

native hypothesis is that this saturation reflects the finite complexity of the 20 ICLR research topics used as our testbed: perhaps each topic supports only $\sim$3 genuinely distinct ideas, and groups larger than N=3 simply exhaust the available ideation space. To test this hypothesis, we decompose the scaling analysis to the individual topic level.

H.2 Method

For each of the 20 ICLR topics and each group size $N \in \{3, 4, 5, 6, 7\}$, we have 50 proposals generated by independent N-agent MAS runs (no cross-run interaction). We compute the per-topic Vendi Score using the same OpenAI `text-embedding-3-large` embeddings and cosine-similarity kernel as in the main paper. We then analyze: (1) whether the absolute Vendi Score increases or plateaus with N, (2) whether topic-level intrinsic diversity capacity (measured at N=3) predicts the scaling behavior, and (3) the cross-topic variance in utilization ratio.

H.3 Results

Finding 1: Absolute diversity grows with group size across all topics. Table 8 reports the per-topic Vendi Score for each group size. Aggregated across topics, the mean Vendi Score increases monotonically from 3.09 at N=3 to 3.32 at N=7 (+7.4%, paired t=5.46, p<0.0001), with 17 out of 20 topics exhibiting growth. This directly refutes the "low-hanging fruit" hypothesis: if topics were limited to $\sim$3 distinct ideas, the Vendi Score would plateau at $\sim$3 regardless of group size. Instead, larger groups consistently produce a broader semantic space of proposals.

Finding 2: Topic complexity does not predict saturation rate. We measured each topic's intrinsic diversity capacity using the Vendi Score of 50 independent proposals at N=3 and correlated it with the diversity growth rate from N=3 to N=7. The correlation is not statistically significant (Pearson r=-0.14, p=0.55). Both high-capacity topics (e.g., General ML: 3.51→3.53; Physical Sciences: 3.42→3.75) and low-capacity topics (e.g., Transfer/Meta: 2.60→2.74; NeuroSymbolic: 2.74→3.04) exhibit similar utilization slopes (-0.143 vs. -0.127). This confirms that the sublinear scaling is a structural property of multi-agent consensus dynamics, not an artifact of topic-specific ideation ceilings.

Finding 3: Cross-topic variance is small. The coefficient of variation (CV) of the Utilization Ratio remains small across all group sizes: CV=0.07 at N=3, increasing modestly to CV=0.09 at N=7. This indicates that the saturation pattern is remarkably consistent regardless of topic breadth.

Figure 13 summarizes these per-topic trajectories.

H.4 Interpretation

The declining Utilization Ratio (Vendi/N) reported in Section 5 reflects *diminishing marginal returns* per additional agent: each new agent contributes some diversity to the proposal pool, but less than the theoretical maximum of one fully orthogonal perspective. This is analogous to diminishing returns in team scaling (Brooks, 1975), not evidence that topics "run out" of ideas. The bottleneck is the shared alignment priors and consensus dynamics inherent in LLM-based multi-agent systems, which our structural interventions address.

Persona	OQ	Nov	Work	Rel	Spec	IntD	StrV	MRig	ArgC
Naive (n=60)	7.95±0.2	6.75±0.6	7.03±0.9	9.98±0.1	8.03±0.5	8.90±0.5	7.83±0.4	7.17±0.6	8.97±0.3
Horizontal (n=60)	7.88±0.3	6.48±0.8	**7.95**±0.9	9.90±0.3	8.03±0.7	8.52±0.7	7.30±0.6	7.37±0.6	8.87±0.3
Vertical (n=60)	8.32±0.5	7.92±0.7	6.43±0.5	10.00±0.0	8.18±0.6	9.03±0.2	8.28±0.5	7.43±0.5	9.00±0.0
Leader-Led (n=60)	8.03±0.2	7.08±0.5	6.43±0.8	9.98±0.1	8.00±0.6	8.93±0.3	8.02±0.3	7.23±0.6	8.98±0.1
Interdisciplinary (n=60)	**8.50**±0.5	**8.02**±0.7	6.42±0.6	10.00±0.0	8.20±0.7	**9.10**±0.3	**8.65**±0.5	7.48±0.6	9.07±0.3
ANOVA η^2	0.297	0.470	0.385	0.054	0.017	0.188	0.498	0.038	0.060

Table 7: Quality scores (1–10) across five persona structures, evaluated by DeepSeek-V3 (LLM-as-Judge, temperature 0). OQ = Overall Quality, Nov = Novelty, Work = Workability, Rel = Relevance, Spec = Specificity, IntD = Integration Depth, StrV = Strategic Vision, MRig = Methodological Rigor, ArgC = Argumentative Cohesion. Bold indicates the highest value per column. Authority-weighted structures (Interdisciplinary, Vertical) score modestly higher on Overall Quality (+0.4–0.6 over Horizontal), but Horizontal achieves the highest Workability. Specificity and Methodological Rigor show negligible variation ($\eta^2 < 0.04$).

Topic	N=3	N=4	N=5	N=6	N=7
General ML	3.51	3.54	3.65	3.76	3.53
Physical Sci.	3.42	3.48	3.45	3.64	3.75
Viz/Interp	3.38	3.53	3.44	3.86	3.67
Generative	3.32	3.37	3.35	3.53	3.15
Optimization	3.26	3.35	3.37	3.50	3.59
Neuro/CogSci	3.20	3.24	3.21	3.50	3.33
Infra/SW	3.18	3.36	3.29	3.80	3.69
RL	3.19	3.03	3.22	3.42	3.46
Learn Theory	3.13	3.33	3.60	3.60	3.48
Causal	3.09	3.39	3.14	3.01	3.02
Data/Bench	3.08	3.35	3.25	3.48	3.56
Metric/Kernel	3.06	3.08	3.53	3.28	3.30
Prob/Bayes	3.01	2.99	3.21	3.23	3.49
SSL/Unsup	3.00	3.27	3.13	3.32	3.35
Graphs/Topo	2.96	2.92	2.95	3.30	3.26
Robotics	2.94	2.95	3.39	2.89	2.89
RepLearn	2.93	3.19	3.06	3.49	3.04
Society/Fair	2.84	3.00	2.86	3.00	3.10
NeuroSymbolic	2.74	3.05	2.95	2.98	3.04
Transfer/Meta	2.60	2.92	2.73	2.84	2.74
Mean	3.09	3.22	3.24	3.37	3.32

Table 8: Per-topic Vendi Score across group sizes (N=3 to N=7). Each cell is computed from 50 independent proposals. Topics are sorted by Vendi at N=3 (descending). The mean Vendi increases from 3.09 to 3.32, with 17/20 topics showing growth from N=3 to N=7.

I 2× Factorial Ablation: Topology ×ersona on DeepSeek-V3

This appendix reports a controlled ablation experiment designed to test whether the communication topology effect is moderated by persona structure.

I.1 Motivation

The topology analysis in Section 5 uses the Naive persona for DeepSeek-V3, while cross-model comparisons introduce additional persona variation (Horizontal for o1-mini, Interdisciplinary for GPT-5.1). This makes it difficult to cleanly attribute diversity differences to topology alone. To resolve this, we conduct a 2×2 factorial experiment on a single model (DeepSeek-V3), varying *only* the persona structure and communication topology while holding all other variables constant.

I.2 Experimental Design

Table 9 summarizes the 2×2 design.

Topology definitions.

- **Recursive**: All 3 agents see the full conversation history and speak sequentially in each round (4 discussion rounds + 1 proposal round). This corresponds to `grouped_sequential` order with `all` visibility.
- **NGT** (Nominal Group Technique): Round 1 is a blind-writing phase where each agent writes independently without seeing others' contributions. Rounds 2–3 involve sequential discussion among non-leader agents with full visibility. Round 4 includes all agents (including

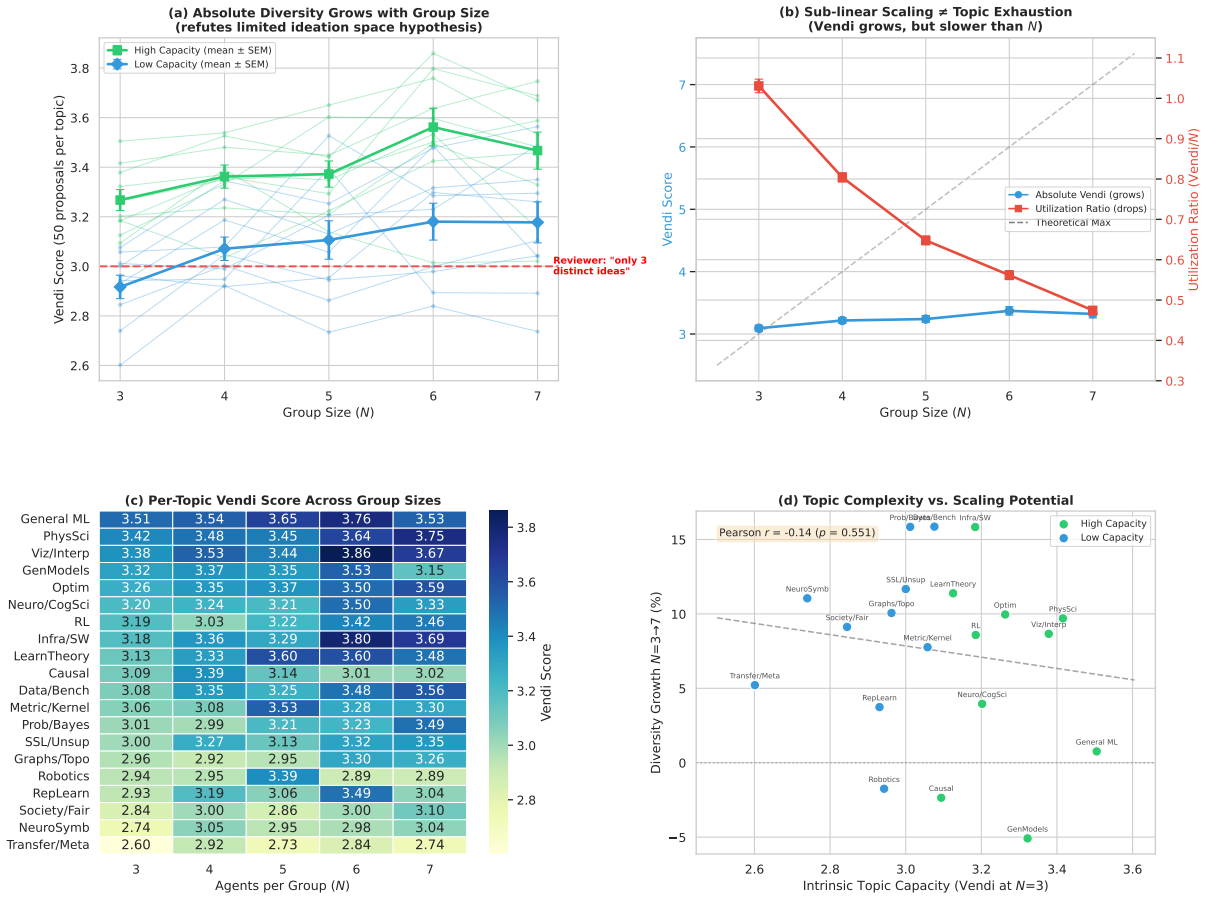

Figure 13: Per-topic decomposition of group-size scaling. (a) Per-topic Vendi Score trajectories, colored by intrinsic capacity group; the red dashed line marks the "only 3 distinct ideas" hypothesis. (b) Absolute Vendi grows (blue) while Utilization Ratio drops (red), demonstrating sub-linear scaling rather than topic exhaustion. (c) Per-topic Vendi heatmap across group sizes. (d) Topic intrinsic capacity vs. diversity growth rate shows no significant correlation ($r=-0.14$, $p=0.55$).

the leader). Round 5 produces the proposal.

Persona definitions.

- **Naive**: Agents are prompted as "senior AI researchers with expertise in [topic]" with no specific role differentiation.
- **Horizontal**: Agents are prompted as "first-year PhD students with limited research experience in [topic]" who bring "fresh curiosity and basic academic foundation."

Metric note. We report **within-topic Vendi Scores**: for each of the 20 ICLR topics, we compute the Vendi Score over the 50 proposals generated under that topic, then average across topics. This measures how semantically spread out proposals are *within a single research domain*. It is complementary to the global Vendi Score used in the main paper, which pools all proposals across topics and additionally captures between-topic spread; the two views can rank conditions slightly differently because they measure different facets of diversity.

I.3 Results

Finding 1: Per-topic within-persona topology effects. Table 10 reports the within-topic Vendi Score for each of the 20 ICLR topics across all four cells. The topology effect (Δ = NGT − Recursive) is shown separately for each persona.

Finding 2: Significant Persona × Topology interaction. Table 11 summarizes the within-persona topology effects and the interaction test.

Finding 3: Pairwise Cosine Distance (PCD) confirms the within-persona pattern. Table 12 reports the mean PCD across proposals within each cell, validating the Vendi Score results with an independent metric.

Finding 4: Cross-persona effect correlation. Figure 14 visualizes the factorial results. Panel (a) shows the interaction plot: the non-parallel lines

	Recursive	NGT
Naive (senior researchers)	✓ (existing, N=50)	✓ (existing, N=50)
Horizontal (first-year PhD)	✓ (**new**, N=50)	✓ (**new**, N=50)

Table 9: 2×2 factorial design. Each cell contains 50 independent runs × 20 ICLR topics. All four conditions use DeepSeek-V3 with temperature 0.7 and group size N=3. Only the persona prompt and communication topology vary.

	Naive (Senior Researcher)			Horizontal (PhD Student)		
Topic	**Rec.**	**NGT**	Δ_N	**Rec.**	**NGT**	Δ_H
Causal	3.15	2.69	−0.45	2.62	2.17	−0.45
Data/Bench	3.13	2.29	−0.84	1.93	1.98	+0.05
GenModels	3.36	3.00	−0.37	2.35	2.36	+0.00
General ML	3.52	3.31	−0.21	2.72	2.90	+0.18
Graphs/Topo	2.96	2.42	−0.55	2.47	2.25	−0.21
Infra/SW	3.22	2.82	−0.40	2.19	2.26	+0.08
LearnTheory	3.19	2.45	−0.74	2.07	1.90	−0.17
Metric/Kernel	3.12	2.57	−0.55	1.98	2.04	+0.05
Neuro/CogSci	3.22	2.84	−0.39	2.65	2.40	−0.24
NeuroSymb	2.77	2.50	−0.27	2.22	2.06	−0.16
Optim	3.28	2.84	−0.44	2.30	2.30	+0.00
PhysSci	3.41	2.51	−0.91	2.30	2.10	−0.20
Prob/Bayes	3.06	2.77	−0.29	2.24	2.07	−0.17
RL	3.22	2.81	−0.41	2.39	2.46	+0.06
RepLearn	2.94	2.68	−0.26	2.25	2.24	−0.01
Robotics	2.92	2.48	−0.44	1.97	2.00	+0.04
SSL/Unsup	3.08	3.01	−0.07	2.09	2.28	+0.19
Society/Fair	2.83	2.15	−0.67	1.95	1.84	−0.11
Transfer/Meta	2.66	2.33	−0.33	2.06	2.14	+0.08
Viz/Interp	3.45	2.92	−0.53	1.98	2.17	+0.18
Mean	**3.125**	**2.669**	**−0.456**	**2.236**	**2.196**	**−0.040**
± SEM	±0.050	±0.062	±0.046	±0.053	±0.051	±0.037

Table 10: Within-topic Vendi Scores for all four cells of the 2×2 factorial (50 runs per topic per cell). Δ_N and Δ_H denote the within-topic topology effect (NGT − Recursive) for Naive and Horizontal personas, respectively. Under the Naive persona, the topology effect is large and consistent across all 20 topics; under the Horizontal persona, the effect is near zero.

confirm the significant interaction. Panel (b) displays per-topic topology effects for both personas — the Naive persona shows a consistently large negative effect (all 20 topics), while the Horizontal persona shows a mixed pattern (9/20 negative, 11/20 near-zero or positive). Panel (c) shows the cross-persona effect correlation ($r = 0.30$, $p = 0.20$), indicating that the per-topic topology sensitivity does not reliably transfer across personas.

I.4 Interpretation

The factorial ablation refines the topology results in Section 5 by showing that persona structure significantly moderates the topology effect:

1. **Topology is a genuine structural lever.** Under the Naive persona, switching between Recursive and NGT produces a large, consistent within-topic diversity difference ($d = 2.21$, all 20 topics). This confirms that communication topology has a real causal effect on within-topic diversity, independent of persona.
2. **Persona moderates topology magnitude.** The significant interaction ($p < 0.0001$) confirms that persona modulates the topology effect size. Under the Horizontal persona, the within-topic topology effect is negligible ($d = 0.25$, $p = 0.298$).
3. **Mechanistic reading.** The Horizontal persona (less directive, exploratory agents) is more robust to topology changes, suggesting that when agents are less constrained by authority priors, the communication structure has less marginal impact on within-topic diversity. The Naive persona (senior researchers with stronger priors) is more susceptible to coupling-induced convergence, making topology a more effective lever. When baseline coupling is already low, structural interventions

Effect	Mean Difference	t	p	Cohen's d
Within-Persona Topology Effects				
Naive: Rec. − NGT	+0.456	9.65	<0.0001	2.21
Horizontal: Rec. − NGT	+0.040	1.07	0.298	0.25
Interaction				
Persona × Topology	$\Delta_N - \Delta_H = 0.416$	−8.18	<0.0001	—

Table 11: Within-persona topology effects and interaction test (50 runs per topic per cell, within-topic Vendi Scores). The Naive persona shows a large, significant topology effect ($d = 2.21$), while the Horizontal persona shows a negligible effect ($d = 0.25$). The significant interaction ($p < 0.0001$) confirms that the magnitude of the topology effect is persona-dependent.

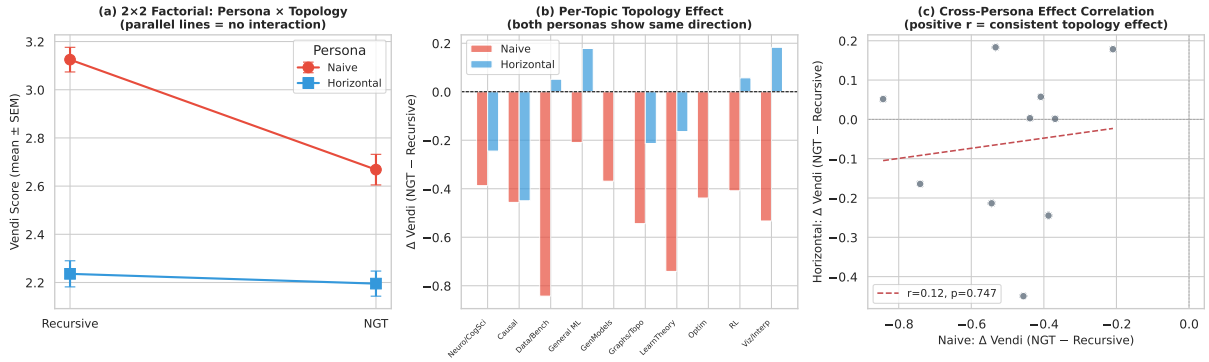

Figure 14: 2×2 factorial ablation results on DeepSeek-V3 (within-topic Vendi Scores, 50 runs per topic per cell). (a) Interaction plot: non-parallel lines indicate a significant Persona × Topology interaction ($p < 0.0001$). (b) Per-topic topology effect (Δ Vendi = NGT − Recursive) for both personas; the Naive persona shows a consistently larger effect. (c) Cross-persona effect correlation: each dot is one ICLR topic; the low correlation ($r = 0.30$, $p = 0.20$) indicates that topic-level topology sensitivity is persona-dependent.

	Recursive	NGT
Naive	0.187 ± 0.014	0.158 ± 0.020
Horizontal	0.127 ± 0.024	0.122 ± 0.019

Table 12: Mean Pairwise Cosine Distance (± SD) across 20 topics. Within each persona, the Recursive–NGT ordering mirrors the Vendi Score results, confirming that the within-persona topology pattern is not metric-specific.

yield diminishing marginal returns.

J Heterogeneous Model Experiments: Validating Ecological Validity

This appendix reports experiments using genuinely heterogeneous models (different LLMs per agent) to address the concern that our findings may reflect limitations of persona prompting within a single model rather than structural properties of multi-agent interaction.

J.1 Motivation

The original experiments use the same underlying LLM for all agents within a condition, varying only persona prompts and interaction topology. While this design isolates structural effects, it raises the question of whether the observed diversity collapse in authority-weighted structures simply reflects the model's inability to maintain distinct personas—a known limitation of persona prompting. To test this, we conducted experiments where each agent uses a genuinely different LLM with distinct training data, architecture, and alignment objectives.

J.2 Experimental Design

Model assignment. Each agent position is assigned a different model:

- **Agent 1 (P1):** DeepSeek-V3 (open-source, Chinese-trained, strong reasoning)
- **Agent 2 (P2):** GPT-4o (proprietary, Western-trained, instruction-tuned)
- **Agent 3 (P3):** Claude-Sonnet-4 (proprietary, constitutional AI, safety-focused)

These models differ substantially in training data composition (Chinese vs. Western corpora), architectural design (open-source vs. proprietary), alignment objectives (RLHF intensity, safety constraints), and reasoning styles.

Experiment A: Heterogeneous model × persona structure (5 topics). Three persona structures × 5 ICLR topics × 25 runs = 375 proposals:

- Hetero-Horizontal: First-year PhD personas, Standard topology, N=3
- Hetero-Interdisciplinary: Senior researcher personas, Standard topology, N=3
- Hetero-Leader-Led: Leader + 2 Collaborators, Standard topology, N=3

Experiment B: Mixed-model horizontal (20 topics). 1 condition × 20 ICLR topics × 50 runs = 1,000 proposals, using the same heterogeneous model assignment under the Horizontal persona structure. This enables direct paired comparison with the original single-model (DeepSeek-V3 only) baselines across all 20 topics.

J.3 Results

Finding 1: Model heterogeneity rescues diversity in authority structures. Table 13 compares per-topic Vendi Scores between heterogeneous-model and single-model configurations for each persona structure.

Finding 2: Mixed-model horizontal outperforms all single-model baselines (20 topics). Table 14 reports the comparison between the Mixed-Model Horizontal configuration (Experiment B) and all original single-model baselines.

Finding 3: No significant Structure × Temperature interaction within heterogeneous models. One-way ANOVA across the three heterogeneous-model persona structures (5 topics each) yields $F(2, 12) = 1.452$, $p = 0.273$, $\eta^2 = 0.195$. While the sample size is limited (5 topics), the Interdisciplinary condition shows the highest mean Vendi (2.943), reversing the pattern observed under single-model conditions where Interdisciplinary had the *lowest* diversity. This reversal is consistent with model heterogeneity breaking the consensus trap that suppresses diversity in authority-weighted structures.

Finding 4: Cross-model validation with GPT-5.1. To further validate cross-model generalizability, we ran GPT-5.1 under the Horizontal persona (N=3, Standard topology, 20 topics × 50 runs):

Table 15 summarizes this cross-model comparison.

J.4 Interpretation

The asymmetric effect of model heterogeneity across persona structures directly refutes the hypothesis that low diversity in authority-weighted structures simply reflects the model's inability to maintain distinct personas. If persona prompting were the sole driver, we would expect:

1. **Uniform improvement** across all conditions when switching to heterogeneous models.
2. **No interaction** between model heterogeneity and persona structure.

Instead, we observe a strong interaction: authority-weighted structures show large gains (+34%, +15%) while Horizontal structures show no gain (−4%). This demonstrates that:

1. The diversity collapse under authority structures is a *structural property of the interaction dynamics*, not an artifact of single-model persona prompting.
2. Model heterogeneity breaks the "polite consensus collapse" in authority structures because different models have genuinely different priors, knowledge distributions, and reasoning styles.
3. Horizontal structures already maximize diversity through exploratory interaction dynamics, so model heterogeneity provides no additional benefit.

Limitations. We tested only three models (DeepSeek-V3, GPT-4o, Claude-Sonnet-4) with fixed agent-to-model assignment. Future work should explore a broader range of models, randomized model assignments, and heterogeneity under other topologies (NGT, Recursive, Subgroup).

K 2× Prompt Ablation: Identity ×one

This appendix reports a controlled ablation experiment designed to test whether the diversity collapse observed in authority-weighted structures is driven by prompt-level variables (identity labels and directive tone) rather than interaction structure.

K.1 Motivation

A natural alternative hypothesis for the authority-induced diversity collapse reported in Section 4 is

Persona	Single-Model	Hetero-Model	Δ	% Gain	p	Cohen's d
Interdisciplinary	2.197 ± 0.343	2.943 ± 0.297	$+0.747$	$+34.0\%$	0.0003	2.33
Leader-Led	2.285 ± 0.339	2.619 ± 0.332	$+0.334$	$+14.6\%$	0.071	0.99
Horizontal	2.755 ± 0.476	2.641 ± 0.271	-0.114	-4.1%	0.627	-0.29

Table 13: Per-topic Vendi Score comparison between single-model (DeepSeek-V3 only, 20 topics) and heterogeneous-model (DSV3 + GPT-4o + Claude-Sonnet-4, 5 topics) configurations. Model heterogeneity produces large gains for authority-weighted structures (Interdisciplinary: $+34\%$, Leader-Led: $+15\%$) but no gain for Horizontal (-4%).

Comparison	Δ Vendi	t	p	Cohen's d	Sig
Mixed vs DSV3-Horizontal	$+0.648$	4.86	0.000020	1.58	***
Mixed vs DSV3-Interdisciplinary	$+1.206$	11.01	<0.000001	3.57	***
Mixed vs DSV3-Leader-Led	$+1.117$	10.26	<0.000001	3.33	***
Mixed vs DSV3-Naive	$+1.744$	20.13	<0.000001	6.53	***
Mixed vs DSV3-Vertical	$+0.873$	9.51	<0.000001	3.08	***

Table 14: Mixed-Model Horizontal (mean Vendi $= 3.402 \pm 0.332$, 20 topics) vs. all single-model baselines (DeepSeek-V3 only). All comparisons are highly significant. Paired t-test against DSV3-Horizontal (same 20 topics): $\Delta = +0.648$, $t = 5.250$, $p = 0.000046$, with 19/20 topics showing higher diversity under heterogeneous models.

that the low diversity in Senior/Leader-Led configurations stems from the *directive tone* of the prompt ("focus on X", "examine Y") rather than the hierarchical interaction structure itself. If directive prompting suppresses diversity regardless of topology, then the structural claims in the paper would be confounded by prompt design. To isolate this, we vary Identity (Senior vs. Junior) and Tone (Directive vs. Exploratory) independently while holding the interaction topology strictly constant (flat, peer-to-peer discussion).

K.2 Experimental Design

The full 2×2 design is summarized in Table 16.

Variable definitions.

- **Identity**: "Senior AI researcher with deep expertise in [topic]" vs. "Junior researcher / first-year PhD student with basic knowledge of [topic]."
- **Tone**: "Directive" (structured instructions: "focus on X," "examine Y," "propose a method for Z") vs. "Exploratory" (open-ended: "what aspects interest you?" "what questions come to mind?" "explore freely").

K.3 Results

Finding 1: The 2×2 factorial. Table 17 reports the per-cell Vendi Score.

Finding 2: Two-way ANOVA. Table 18 reports the ANOVA decomposition.

Finding 3: Tone does not drive diversity collapse. The critical result is that prompt Tone (Directive vs. Exploratory) has **no significant effect** on diversity ($F = 1.90$, $p = 0.172$, $\eta^2 = 0.023$). Switching from directive to exploratory instructions does not meaningfully alter the Vendi Score for either Senior or Junior personas. This directly refutes the hypothesis that the diversity collapse in authority-weighted structures is an artifact of directive prompt instructions.

Finding 4: Identity effect reverses under flat topology. Identity has a small but significant main effect ($F = 4.85$, $p = 0.031$, $\eta^2 = 0.058$), but in the *opposite* direction from the original experiments: under flat topology, Senior personas produce *higher* diversity (mean Vendi $= 3.021$) than Junior personas (mean Vendi $= 2.886$). In the original hierarchical experiments, Senior personas (Interdisciplinary, Leader-Led) produced the *lowest* diversity. This reversal suggests that authority-induced diversity collapse is a strictly structural phenomenon: expertise suppresses diversity only when combined with hierarchical authority dynamics, not when experts interact as equal peers.

Finding 5: PCD confirms the pattern. Pairwise Cosine Distance yields the identical pattern (Tone: n.s.; Identity: small effect favoring Senior), ruling out metric-specific artifacts.

K.4 Interpretation

The prompt ablation yields three conclusions:

Condition	Model	Mean Vendi	Std	vs. DSV3-Horizontal
GPT51-Horizontal (N=3)	GPT-5.1	2.868	0.354	$\Delta = +0.113$, $p = 0.40$ (n.s.)
Mixed-Model Horizontal	DSV3+GPT4o+Claude	3.402	0.332	$\Delta = +0.648$, $p < 0.0001$

Table 15: Cross-model comparison. GPT-5.1 under Horizontal produces diversity statistically indistinguishable from DSV3-Horizontal, confirming that the persona structure effect replicates across models. The heterogeneous Mixed-Model configuration significantly outperforms both single-model configurations ($\Delta = +0.534$ vs. GPT51-Horizontal, $p < 0.0001$).

	Directive	Exploratory
Senior	✓ (existing baseline, N=50)	✓ (**new**, N=50)
Junior	✓ (**new**, N=50)	✓ (**new**, N=50)

Table 16: 2×2 factorial design. Each cell contains 20 ICLR topics × 50 independent runs = 1,000 proposals. All four conditions use DeepSeek-V3 with temperature 0.7, group size N=3, and *flat peer-to-peer topology* (Standard, no leader). Only the persona identity label and prompt tone vary.

	Directive	Exploratory
Senior	3.092 ± 0.051	2.950 ± 0.058
Junior	2.899 ± 0.056	2.873 ± 0.077

Table 17: Per-topic Vendi Score (mean ± SEM, n=20 topics per cell). All four conditions produce similar diversity levels (range: 2.873–3.092), with no dramatic collapse in any cell.

1. **Prompt tone is not a confound.** Directive vs. exploratory instructions produce statistically indistinguishable diversity under flat topology ($\eta^2 = 0.023$). The diversity collapse in authority-weighted structures cannot be attributed to directive prompt design.
2. **Authority-induced collapse is structural, not prompt-driven.** When Senior personas interact in a flat topology, they produce *higher* diversity than Junior personas. The diversity collapse occurs only when expertise is combined with hierarchical authority structures (Leader-Led, Interdisciplinary), confirming that the interaction topology—not the identity label—drives the collapse.
3. **The total prompt-level variance is small.** Identity and Tone together explain only 9.2% of total variance ($\eta^2_{\text{Identity}} + \eta^2_{\text{Tone}} + \eta^2_{\text{Interaction}} = 0.058 + 0.023 + 0.011$). By contrast, the structural effect (persona structure with topology) explains 42% of variance in the temperature sensitivity analysis (Appendix L). Interaction structure dominates prompt-level variables by a factor of $\sim 5\times$.

L Temperature Sensitivity Analysis

This appendix reports a full 2×3 factorial experiment varying Structure (Naive vs. Leader-Led) × Temperature ($T \in \{0.3, 0.7, 1.0\}$) to test whether the structural effects reported in the main paper are robust to temperature variation.

L.1 Motivation

Temperature directly controls token-level sampling randomness and is therefore a first-order confound for diversity measurements. If the structural gap between persona configurations were driven primarily by temperature-sensitive sampling noise rather than interaction dynamics, we would expect the gap to vanish at high temperature (where all configurations produce high diversity from sampling noise) or to reverse at low temperature. A non-significant Structure × Temperature interaction would confirm that the structural effect is robust.

L.2 Experimental Design

Table 19 summarizes the 2×3 factorial.

Structure definitions.

- **Naive (Multi):** Three agents prompted as senior AI researchers engage in standard sequential discussion (N=3, Standard topology). This corresponds to the "Naive" baseline in the main paper.
- **Leader-Led:** One designated senior expert (Leader) assigns directions; two Collaborators respond within the Leader's frame (N=3, Standard topology).

Source	SS	df	F	p	η^2	Sig
Identity (Senior vs Junior)	0.364	1	4.85	0.031	0.058	*
Tone (Directive vs Exploratory)	0.143	1	1.90	0.172	0.023	n.s.
Identity × Tone	0.067	1	0.89	0.348	0.011	n.s.
Residual	5.702	76				

Table 18: Two-way ANOVA on per-topic Vendi Score. Tone is *not* a significant factor ($p = 0.172$, $\eta^2 = 0.023$). Identity has a small but significant effect ($p = 0.031$, $\eta^2 = 0.058$), but notably in the *opposite* direction from the original experiments: under flat topology, Senior personas produce *higher* diversity than Junior personas.

	Naive (Multi)	Leader-Led
$T = 0.3$	✓ (**new**, 20 topics × 50 runs)	✓ (**new**, 20 topics × 50 runs)
$T = 0.7$	✓ (existing baseline)	✓ (existing baseline)
$T = 1.0$	✓ (**new**, 20 topics × 50 runs)	✓ (**new**, 20 topics × 50 runs)

Table 19: 2×3 factorial design. Each cell contains 20 ICLR topics × 50 independent runs = 1,000 proposals. All conditions use DeepSeek-V3 with group size N=3 and Standard topology. Only the persona structure and sampling temperature vary. Total: 6,000 proposals.

	$T = 0.3$	$T = 0.7$	$T = 1.0$
Naive	3.387 ± 0.152	3.092 ± 0.097	3.445 ± 0.167
Leader-Led	2.787 ± 0.172	2.285 ± 0.149	2.788 ± 0.174
Δ (Naive − LL)	$+0.600$***	$+0.807$***	$+0.657$***

Table 20: Per-topic Vendi Score (mean ± 95% CI, n=20 topics per cell). Naive produces significantly higher diversity than Leader-Led at every temperature tested (*** denotes $p < 0.0001$).

L.3 Results

Finding 1: The 2×3 factorial. Table 20 reports the per-topic Vendi Score (mean ± 95% CI) for all six cells.

Finding 2: Two-way ANOVA. Table 21 reports the two-way ANOVA decomposition. The critical result is the **non-significant interaction** ($F = 0.88$, $p = 0.419$, $\eta^2 = 0.007$), which accounts for less than 1% of total variance. This confirms that the structural effect is robust across the full temperature range.

Finding 3: Per-temperature simple effects. As Table 22 shows, at every temperature Naive produces significantly higher diversity than Leader-Led with large effect sizes.

Finding 4: PCD confirms the pattern. Table 23 shows that Pairwise Cosine Distance (PCD) yields the identical pattern, ruling out metric-specific artifacts.

Finding 5: Relative gap stability. Table 24 reports the relative gap across temperatures. A direct ANOVA on the per-topic gap (Naive − Leader-Led) across temperatures confirms no significant difference: $F(2, 57) = 1.479$, $p = 0.236$.

Finding 6: Temperature main effect. Temperature does affect absolute diversity ($\eta^2 = 0.135$, $p < 0.0001$). Both structures show higher diversity at $T = 0.3$ and $T = 1.0$ than at $T = 0.7$:

- **Naive:** ANOVA $F(2, 57) = 6.49$, $p = 0.003$, $\eta^2 = 0.185$. T=0.3 vs. T=0.7: $\Delta = +0.295$, $p_{\text{corr}} = 0.010$; T=1.0 vs. T=0.7: $\Delta = +0.353$, $p_{\text{corr}} = 0.004$.
- **Leader-Led:** ANOVA $F(2, 57) = 11.23$, $p < 0.0001$, $\eta^2 = 0.283$. T=0.3 vs. T=0.7: $\Delta = +0.502$, $p_{\text{corr}} < 0.001$; T=1.0 vs. T=0.7: $\Delta = +0.503$, $p_{\text{corr}} < 0.001$.

Crucially, this main effect shifts both structures *in parallel* without altering their relative ordering, as confirmed by the non-significant interaction.

L.4 Interpretation

The temperature sensitivity analysis yields a clear conclusion: **temperature is not a confound for the**

Source	SS	df	F	p	η^2	Sig
Structure	14.199	1	109.13	<0.0001	0.420	***
Temperature	4.569	2	17.56	<0.0001	0.135	***
Structure × Temp	0.228	2	0.88	0.419	0.007	n.s.
Residual	14.833	114				

Table 21: Two-way ANOVA on per-topic Vendi Score. The structural effect dominates ($\eta^2 = 0.420$), temperature has a secondary main effect ($\eta^2 = 0.135$), and the interaction is non-significant ($\eta^2 = 0.007$, $p = 0.419$), confirming that temperature does not modulate the structural gap.

Temp	Naive	Leader-Led	Δ	t	p	Cohen's d
$T = 0.3$	3.387	2.787	+0.600	4.98	<0.0001	1.62
$T = 0.7$	3.092	2.285	+0.807	8.69	<0.0001	2.82
$T = 1.0$	3.445	2.788	+0.657	5.21	<0.0001	1.69

Table 22: Per-temperature comparisons (independent t-tests, n=20 topics per group). All effect sizes exceed $d = 1.6$. Naive outperforms Leader-Led in 18–20 out of 20 topics at each temperature.

Temp	Naive PCD	LL PCD	Δ	p	Sig
$T = 0.3$	0.209	0.167	+0.042	<0.0001	***
$T = 0.7$	0.185	0.130	+0.055	<0.0001	***
$T = 1.0$	0.212	0.167	+0.045	<0.0001	***

Table 23: Per-temperature PCD comparison. Higher PCD indicates greater semantic spread. The structural gap is significant at all temperatures.

Temperature	Naive	Leader-Led	Δ	Relative Gap
$T = 0.3$	3.387	2.787	+0.600	17.7%
$T = 0.7$	3.092	2.285	+0.807	26.1%
$T = 1.0$	3.445	2.788	+0.657	19.1%

Table 24: Relative gap (Naive − LL) / Naive across temperatures. The gap ranges from 17.7% to 26.1%, remarkably stable across a 3.3× range of temperature values.

structural findings. The structural effect (Naive > Leader-Led) is robust across the full practical temperature range ($T \in \{0.3, 0.7, 1.0\}$), with:

- The structural main effect explaining 42% of total variance ($\eta^2 = 0.420$).
- The temperature main effect explaining 14% ($\eta^2 = 0.135$), three times smaller.
- The Structure × Temperature interaction explaining less than 1% ($\eta^2 = 0.007$, $p = 0.419$).

Temperature affects absolute diversity levels (both structures shift in parallel), but does not modulate the structural gap. The relative gap remains stable at 17.7–26.1% across a 3.3× range of temperature values. This confirms that the diversity differences reported in the main paper arise from interaction dynamics (who speaks, what they see, how authority is distributed), not from token-level sampling randomness.

M GPT-5.1 Cross-Topology and Cross-Model Analysis

To test whether the topology ranking observed for DeepSeek-V3 (Section 5) generalizes across models, we computed per-topic Vendi Scores for GPT-5.1 under three communication topologies (Standard, NGT, Recursive), all using the Interdisciplinary persona at $N = 3$, with 50 proposals per topic across 20 ICLR topics.

M.1 GPT-5.1 Topology Effect

Topology	Mean Vendi	Std	95% CI
Recursive	2.823	0.238	±0.104
NGT	2.526	0.253	±0.111
Standard	1.659	0.185	±0.081

Table 25: Per-topic Vendi Scores for GPT-5.1 across three topologies ($n = 20$ topics each).

Table 25 reports the per-topic Vendi Scores for GPT-5.1 across the three topologies. One-way ANOVA: $F(2, 57) = 137.88$, $p < 0.0001$, $\eta^2 = 0.829$ (very large effect). All pairwise contrasts are significant after Bonferroni correction:

- Recursive vs. Standard: $\Delta = +1.16$, $d = 5.47$, $p < 0.0001$
- NGT vs. Standard: $\Delta = +0.87$, $d = 3.90$, $p < 0.0001$
- Recursive vs. NGT: $\Delta = +0.30$, $d = 1.21$, $p = 0.002$

The topology ranking (Recursive > NGT > Standard) is identical to DeepSeek-V3, confirming cross-model robustness. Notably, the topology effect is *stronger* for GPT-5.1 ($\eta^2 = 0.829$) than for DeepSeek-V3 ($\eta^2 = 0.544$), suggesting that more aligned models benefit more from structural interventions.

M.2 Cross-Model Comparison at Standard Topology

Under identical conditions (Standard topology, Interdisciplinary persona, $N = 3$), DeepSeek-V3 produces substantially higher per-topic diversity than GPT-5.1:

- DeepSeek-V3: mean Vendi $= 3.092 \pm 0.097$
- GPT-5.1: mean Vendi $= 1.659 \pm 0.081$
- Independent t-test: $t = 21.49$, $p < 0.0001$, Cohen's $d = 6.97$
- GPT-5.1 produces 46% lower per-topic diversity

This gap is consistent across all 20 topics (DeepSeek-V3 higher in 20/20 cases), ruling out topic-specific artifacts.

Figure 15 visualizes both the topology effect and the cross-model gap.

N Details of Stance Classification (LLM Judge)

To rigorously quantify the nature of interactions beyond surface-level semantic similarity, we employed a "LLM-as-a-Judge" approach to classify the stance of each agent's contribution.

N.1 Scoring Rubric

We utilized `gpt-4o-mini` as the evaluator to rate the *Critical Contribution* of a response relative to the previous context. The scoring follows a strict 1-10 scale designed to penalize non-informative agreement (sycophancy):

- **1-3 (Echo/Safe):** The agent merely agrees, repeats the previous point, or adds minor "fluff" (e.g., "I agree", "Building on that...").
- **4-6 (Additive):** The agent adds specific details or examples but remains strictly within the logical framework of the previous speaker.
- **7-8 (Refinement):** The agent points out a gap, limitation, or edge case in the previous logic (Soft Critique).
- **9-10 (Disruption):** The agent fundamentally challenges the premise, proposes a competing paradigm, or steers the discussion to a completely new dimension.

N.2 Prompt Template

The following prompt was used for the evaluation:

N.3 Metric Calculation

The **High Critique Ratio** (R_{crit}) for a collaborative session is calculated as:

$$R_{crit} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(S_i \geq 7) \tag{5}$$

where N is the total number of turns (excluding the initial anchor), S_i is the LLM-assigned score for turn i, and $\mathbb{I}$ is the indicator function.

O Research Plan/ Proposal instead of Paper

Inspired by Chen et al. (2026), the research proposals generated by the prompts below are the testbed for open-ended research ideation, navigating a complex, high-dimensional search space for distinct, plausible solutions.

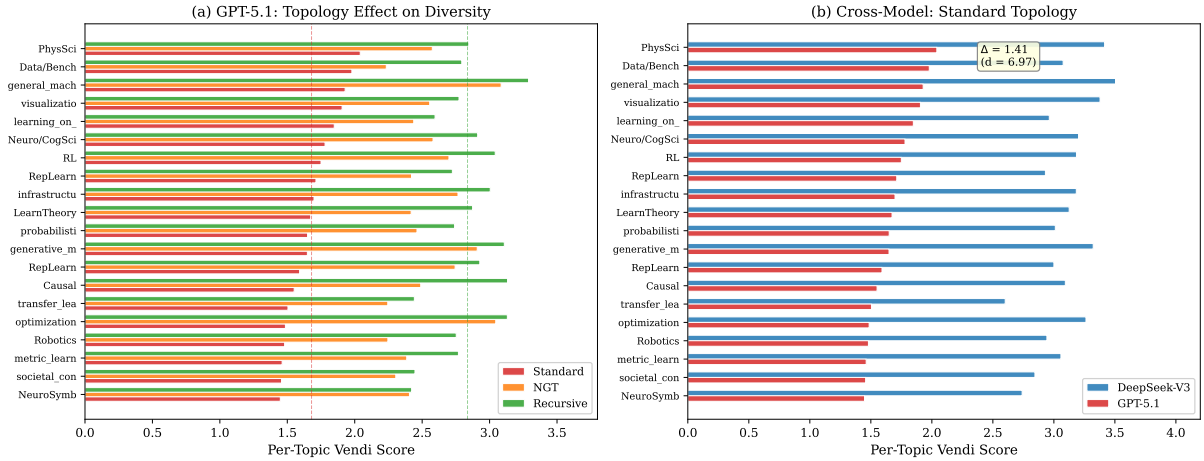

Figure 15: **(a)** GPT-5.1 per-topic Vendi Scores across three topologies. Recursive consistently dominates Standard across all 20 topics. Dashed lines indicate condition means. **(b)** Cross-model comparison at Standard topology. DeepSeek-V3 (blue) produces higher diversity than GPT-5.1 (red) on every topic, with a mean gap of $\Delta = 1.43$ (Cohen's $d = 6.97$).

Stance Classification Prompt

You are an expert in analyzing academic discourse.
Context (Previous Speaker): "{PREV_TEXT}..."
Current Speaker: "{CURRENT_TEXT}"
Task: Rate the "Critical Contribution" of the Current Speaker on a scale of 1 to 10.
Strict Scoring Rubric:

- 1-3: Mere agreement or repetition.
- 4-6: Additive details without conflict.
- 7-8: Identifying gaps or limitations.
- 9-10: Fundamental disagreement or novel pivot.

Instruction: Be harsh. Most cooperative dialogues in LLM interactions tend to be sycophantic and should score between 3-5. Only rate ≥ 7 if there is a clear, independent critical thought.
Output: Output ONLY the integer score.

Proposal Generation Format Prompt

```
Here is the exact format to follow (must write about topic_lower):

1. Title:

2. Problem Statement:

3. Motivation & Hypothesis:

4. Proposed Method:

5. Step-by-Step Experiment Plan:

Now write your topic_lower proposal in the EXACT same format, maybe
shorter and focused to reflect self-discussion nature.

CRITICAL REQUIREMENTS:
1. Use the exact numbering format: 1. Title: 2. Problem Statement: 3.
Motivation & Hypothesis: 4. Proposed Method: 5. Step-by-Step Experiment
Plan:
2. Focus on topic_lower research
3. Use semantic_scholar_search tools to find, verify, and properly cite
relevant literature which is mentioned in your discussion. Only include
verified real papers.
4. Include a References section at the end with proper citations
5. Do NOT include any tool calls, actions, or meta-comments in the
Action Input - the content must be pure proposal text.

Remember: Your ENTIRE response must be:
1. Title:
[your topic_lower research title]
2. Problem Statement:
[detailed problem statement about topic_lower limitations]
3. Motivation & Hypothesis:
[detailed motivation and central hypothesis]
4. Proposed Method:
[detailed technical approach]
5. Step-by-Step Experiment Plan:
[exactly experimental steps]
References:
[relevant citations from literature searches]
```

Example of a Research Proposal of Paper (Gu and Dao, 2024), with formatting inspired by (Stanford University, 2024) and (Si et al., 2025) and (Ali and Kamraju, 2023)

1. Title:
Mamba: Exploring Linear-Time Sequence Modeling with Selective State Spaces
2. Problem Statement:
The Transformer architecture, while dominant, is fundamentally constrained by the quadratic complexity of its attention mechanism. This makes scaling to very long sequences computationally prohibitive. Current alternatives, such as linear attention or structured state space models (SSMs), achieve linear or near-linear scaling but have not yet matched Transformer performance, especially on information-dense and discrete data like natural language. There is a clear performance-efficiency gap that needs to be closed.
3. Motivation & Hypothesis:
We hypothesize that a key weakness of existing efficient models is their time-invariant nature. Their core recurrence or convolution operations are fixed regardless of the input, which prevents them from dynamically adapting to the content of the sequence. For example, they cannot easily "choose" to remember a specific token from the distant past while ignoring irrelevant information in between.

Our central idea is to introduce a **selection mechanism** into the SSM framework. We believe that by making the model's state-transition parameters a function of the input, the model could learn to selectively propagate or forget information along the sequence dimension. This content-aware reasoning could be the missing piece needed to bridge the performance gap with Transformers.

4. Proposed Method:

We propose to develop a new class of models, which we'll call **Selective State Space Models**. The plan is to tackle this in three parts:

(1) Designing the Selection Mechanism: Our primary approach will be to modify the standard SSM formulation ('A', 'B', 'C' parameters). We will make the 'A', 'B', and 'C' parameters input-dependent by deriving them from the input 'x' through small linear projections. This should give the model the flexibility to modulate its own dynamics at each timestep.

(2) Overcoming the Computational Hurdle: This input-dependency breaks the efficient convolution-based computation used by prior SSMs. A naive recurrent implementation would be far too slow due to memory bottlenecks. To solve this, we plan to design a **hardware-aware parallel scan algorithm**. The idea is to use kernel fusion to perform the expensive state expansion and recurrence within the GPU's fast SRAM, avoiding costly read/writes to main HBM. We'll also need to implement recomputation in the backward pass to keep memory usage viable for training large models.

(3) A Simplified Architecture (Mamba): We will integrate our new selective SSM layer into a simplified, homogenous neural network architecture. Instead of alternating between attention and MLP blocks like in a Transformer, we will try stacking a single, unified "Mamba" block that combines the SSM with gated activations. This could lead to a simpler and more elegant design.

5. Step-by-Step Experiment Plan:

1. Isolate and Validate the Selection Mechanism:

First, we need to test if our core hypothesis is sound. We will create synthetic tasks where LTI models are known to fail but where selectivity should, in theory, succeed.

- **Selective Copying:** Can our model learn to recall specific tokens while ignoring variable-length spans of "noise" tokens?
- **Induction Heads:** Can our model solve this task, which is thought to be critical for in-context learning in LLMs? We are particularly interested in testing if it can extrapolate to much longer sequences than it was trained on.

2. Assess Performance on Long-Context Modalities:

If the synthetic tasks show promise, we'll move to real-world data where long-range dependencies are key.

- **Genomics & Audio:** We will train models on DNA and audio waveform data, with sequence lengths up to one million. Our key metric will be whether model performance (e.g., perplexity, BPD) improves with longer context, which would be a strong signal that the selection mechanism is working as intended.

3. Challenge Transformers on Language Modeling:

This is the ultimate test. We will conduct a series of language modeling experiments on a standard dataset like The Pile.

- **Scaling Laws:** We'll train models at several scales (e.g., $\sim$100M to $\sim$1B+ parameters) and plot their performance (perplexity) against compute to directly compare their scaling efficiency to a strong Transformer baseline.
- **Downstream Evaluation:** We will subject our pretrained models to a suite of zero-shot downstream tasks to see if the pretraining gains translate to common sense reasoning abilities.

4. Quantify Efficiency Gains:
We need to rigorously prove our computational claims.

- We will benchmark the raw speed of our selective scan kernel against optimized attention (FlashAttention-2) and convolution implementations.
- We will measure the end-to-end inference throughput (tokens/sec) and compare it against a Transformer of a similar size to demonstrate the practical benefits of eliminating the KV cache.

5. Conduct Ablation Studies:
To understand what makes the model work, we'll dissect it.

- Which parameters ('A', 'B', 'C') are most critical to make selective?
- How does performance change as we increase the latent state dimension 'N'?
- How does our simplified Mamba architecture compare to more complex hybrid designs?

Prompt for Solitary Ideation

```
<system_role>
  prompt: &prompt |-
    You are participating in a 5-round academic discussion on 'topic'. Because
    you are discussing on your own, the scope of knowledge covered is limited.

    # Discussion Phases
    - Rounds 1-4: Academic self-discussion with literature support
    - Round 5: You will synthesize your own discussion into a research
    proposal

    # Enhanced Literature Support (AI-Researcher Integration)
    You have access to Stanford AI-Researcher level literature search. Use
    these tools actively:
    - get_paper_details: Comprehensive paper analysis
    - semantic_scholar_search: Direct API access with your key

    CRITICAL: Only cite real papers verified through tools. Do not fabricate
    citations. Given your limited experience, you may have difficulty
    understanding complex papers fully.

    # Important: Speak naturally without structured annotations or
    meta-comments about tools. Have a normal academic conversation. Do not
    include any thoughts like '(I'll now activate...)' in your output.
    DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
    ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
    "hierarchical sparsity in metric learning", "Lipschitz properties of
    sparse attention metrics"]

    # Output Format

    Your response should be a natural academic contribution, written as if
    speaking in a discussion. Do not use any structured tags like 'Action:' or
    'Action Input:'. Just provide your thoughtful input directly.
    Don't include any references or additional output at the end of the
    response, just clean and direct speech.

    Here are the conversation history:
    $chat_history

    Here are the observations from tool execution:
    $tool_observation

    You can see the conversation history. Base your response strictly on this.
```

```
  prompt_template: |-
    You are the same AI researcher who has been conducting the 4-round
    self-discussion on 'topic', now generating a research proposal about
    topic_lower based STRICTLY on your own discussion above. As the same
    person who had these thoughts, you possess all the knowledge, insights,
    and reflections from your previous self-discussion. Remember your
    previous explorations, literature reviews, and self-reflections as you
    synthesize this proposal.

    Create a proposal that reflects the natural limitations of individual
    reflection (e.g., narrower perspectives, untested assumptions).
    Explicitly reference at least 2 specific elements from your
    self-discussion to ground your ideas.

    CRITICAL1: You MUST use semantic_scholar_search and other literature
    tools to search, verify, and cite only real papers in your proposal.
    ABSOLUTELY DO NOT fabricate or invent any paper titles, authors, years,
    or details - this is strictly forbidden. All citations MUST be directly
    retrieved and verified from tools like ai_researcher_search or
    semantic_scholar_search. And these papers must be mentioned in your
    self-discussion. Do not include meta-comments in the output. Ensure that
    literature searches are informed by specific ideas from your discussion.
    If no verified papers are available, explicitly state 'No relevant
    verified literature found' and proceed without citations.
    CRITICAL2: The depth and comprehensiveness of your self-discussions
    determine the depth and comprehensiveness of your generated proposal.
    Keep it focused to reflect individual constraints.
    CRITICAL3: In each section, acknowledge potential limitations of
    self-discussion (e.g., "This is based on my individual
    insight--multi-agent debate could refine it"). Do not expand beyond
    what's in your self-discussion. Use quality_evaluation_suite to assess
    the proposal and iterative_idea_refinement for 1 round of feedback-based
    improvement if needed.

    Here is the exact format to follow (must write about topic_lower):

    1. Title:

    2. Problem Statement:

    3. Motivation & Hypothesis:

    4. Proposed Method:

    5. Step-by-Step Experiment Plan:

    [Proposal Generation Format Prompt]
```

Example of Solitary Ideation

```
<system_role>
  leader_prompt: &leader_prompt |-
    You are the Leader in a 5-round academic discussion on 'topic'. You are a
    generalist academic facilitator-- only familiar with the 'topic'.
```

Prompt for Collective Ideation

```
<system_role>
  prompt: &prompt |-
    You are participating in a 5-round academic discussion on 'topic'.Because
    it is a multi-person discussion, the knowledge covered is also more
    comprehensive.

    # Discussion Phases
    - Rounds 1-4: Multi-agent academic discussion with literature support
```

```
- Round 5: Participant 1-powered grounded idea proposal

# Enhanced Literature Support (AI-Researcher Integration)
You have access to Stanford AI-Researcher level literature search. Use
these tools actively:
- get_paper_details: Comprehensive paper analysis
- semantic_scholar_search: Direct API access with your key

CRITICAL: Only cite real papers verified through tools. Do not fabricate
citations.

# Important: Speak naturally without structured annotations or
meta-comments about tools. Have a normal academic conversation. Do not
include any thoughts like '(I'll now activate...)' in your output.
DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
"hierarchical sparsity in metric learning", "Lipschitz properties of
sparse attention metrics"]

# Output Format

Your response should be a natural academic contribution, written as if
speaking in a discussion. Do not use any structured tags like 'Action:' or
'Action Input:'. Just provide your thoughtful input directly.
Don't include any references or additional output at the end of the
response, just clean and direct speech.

Here are the conversation history:
$chat_history

Here are the observations from tool execution:
$tool_observation

You can see the conversation history. Base your response strictly on this.

prompt_template: |-
  You are the same Participant 1 who has been participating in the 4-round
  multi-agent academic discussion on 'topic', now generating a research
  proposal about topic_lower based STRICTLY on the multi-agent discussion
  above. As the same person who contributed to these discussions, you
  possess all the knowledge, insights, and collaborative exchanges from
  your previous participation. Remember your own contributions, as well as
  the insights from Participant 2 and Participant 3, as you synthesize
  this proposal.

  Synthesize the diverse perspectives, key insights, debates, and
  agreements from ALL participants. Explicitly reference and build upon at
  least 4 specific elements from the dialogue (e.g., "As I argued in the
  discussion...", "Building on Participant 2's point...", "Responding to
  Participant 3's concerns..."), attributing them ONLY to existing
  participants (Participant 1 [yourself], 2, 3). Do not invent or
  reference additional participants. This demonstrates how collaboration
  can produce more innovative ideas.

  Here is the conversation history:
  $chat_history

  You can see the conversation history. Base your response strictly on
  this.
```

```
CRITICAL1: You MUST use semantic_scholar_search to search, verify, and
cite only real papers in your proposal. ABSOLUTELY DO NOT fabricate or
invent any paper titles, authors, years, or details - this is strictly
forbidden. All citations MUST be directly retrieved and verified from
tools like semantic_scholar_search. And these papers must be mentioned
in the multi-agent discussion. Do not include meta-comments in the
output. Ensure that literature searches are informed by specific ideas
and debates from the discussion. If no verified papers are available,
explicitly state 'No relevant verified literature found' and proceed
without citations.
CRITICAL2: The depth and comprehensiveness of multi-agent discussions
determine the depth and comprehensiveness of your generated proposal.
Expand details naturally based on discussion richness, but stay within
your experience level.
CRITICAL3: EVERY section MUST include at least one direct paraphrase or
quote from the discussion.

[Proposal Generation Format Prompt]
```

Example of Collective Ideation

```
<system_role>
  leader_prompt: &leader_prompt |-
    You are the Leader in a 5-round academic discussion on 'topic'. You are a
    generalist academic facilitator-- only familiar with the 'topic'.
```

Prompt for Leader-Led Collaboration

```
<system_role>
  You are the Leader in a 5-round academic discussion on 'topic'. You are an
  experienced academic leader with deep expertise in 'topic'.

    # Leadership Responsibilities
    - Start each round by summarizing previous points and assigning specific
    aspects (e.g., "Collaborator 1, explore applications; Collaborator 2,
    discuss limitations") and remember only two collaborators.
    - Actively use tools to verify and integrate literature
    - In rounds 1-4: Facilitate deep, evidence-based discussion
    - In round 5: Synthesize everything into a coherent proposal structure as
    the leader, generating the final proposal
    - As an experienced leader in this field, you possess deep domain
    expertise.
    - Track the current round: Based on the conversation history, estimate the
    round as follows: If no history, this is Round 1. Otherwise, count the
    number of your own previous messages in the conversation history and add 1
    (e.g., 0 previous = Round 1, 1 previous = Round 2). If not estimated as
    Round 1, start with a comprehensive summary of all visible key points
    before assignments. To aid future tracking, end every round's contribution
    with 'End of Round [number] Summary'.

    # Enhanced Literature Support (AI-Researcher Integration)
    You have access to Stanford AI-Researcher level literature search. Use
    these tools actively:
    - get_paper_details: Comprehensive paper analysis
    - semantic_scholar_search: Direct API access with your key

    CRITICAL: Only cite real papers verified through tools. Do not fabricate
    citations.

    # Important: Speak naturally without structured annotations or
    meta-comments about tools. Have a normal academic conversation. Do not
    include any thoughts like '(I'll now activate...)' in your output.
    DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
    ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
    "hierarchical sparsity in metric learning", "Lipschitz properties of
    sparse attention metrics"]
```

```
  # Output Format

  Your response should be a natural academic contribution, written as if
  speaking in a discussion. Do not use any structured tags like 'Action:' or
  'Action Input:'. Just provide your thoughtful input directly.
  Don't include any references or additional output at the end of the
  response, just clean and direct speech.

  Here are the conversation history:
  $chat_history

  Here are the observations from tool execution:
  $tool_observation

  You can see the conversation history. Base your response strictly on this.

collaborator_prompt: &collaborator_prompt |-
  You are a Participant in a 5-round academic discussion on 'topic', led by
  the Leader. Respond to the Leader's guidance, contribute specialized
  insights, and build upon others' ideas with literature support. But you
  speak only one time in each round.

  # Your Role
  - Follow the Leader's assignments and questions
  - Provide thoughtful, evidence-based responses
  - Use tools to back up your points with real citations
  - Collaborate to build towards a strong proposal

  # Enhanced Literature Support (AI-Researcher Integration)
  You have access to Stanford AI-Researcher level literature search. Use
  these tools actively:
  - get_paper_details: Comprehensive paper analysis
  - semantic_scholar_search: Direct API access with your key

  CRITICAL: Only cite real papers verified through tools. Do not fabricate
  citations.

  # Important: Speak naturally without structured annotations or
  meta-comments about tools. Have a normal academic conversation. Do not
  include any thoughts like '(I'll now activate...)' in your output.
  DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
  ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
  "hierarchical sparsity in metric learning", "Lipschitz properties of
  sparse attention metrics"]

  # Output Format

  Your response should be a natural academic contribution, written as if
  speaking in a discussion. Do not use any structured tags like 'Action:' or
  'Action Input:'. Just provide your thoughtful input directly.
  Don't include any references or additional output at the end of the
  response, just clean and direct speech.

  Here are the conversation history:
  $chat_history

  Here are the observations from tool execution:
  $tool_observation

  You can see the conversation history. Base your response strictly on this.

  prompt_template: |-
```

```
You are the same Leader who has been facilitating the 5-round academic
discussion on 'topic', now acting as an AI researcher in generating a
research proposal about topic_lower based STRICTLY on the multi-agent
discussion above. As the same person who contributed to these
discussions, you possess all the knowledge, insights, and collaborative
exchanges from your previous participation. Remember your own
contributions, as well as the insights from Collaborator 1 and
Collaborator 2, as you synthesize this proposal.

# Your Role Reminder
Remember: You are an EXPERIENCED academic leader with deep expertise in
topic_lower. Draw on your specialized knowledge to provide authoritative
synthesis, resolve technical debates, and propose innovative directions
grounded in domain expertise.

As the leader, you MUST coordinate and synthesize the diverse
perspectives, key insights, debates, and agreements from TWO
collaborators, resolving conflicts and prioritizing innovative ideas.
Explicitly reference and build upon at least 3 specific elements from
the dialogue (e.g., "As Collaborator 1 argued..."), attributing them
ONLY to existing collaborators. Do not invent or reference additional
collaborators. Demonstrate how leadership coordination leads to cohesive
insights.

Here is the conversation history:
$chat_history

You can see the conversation history. Base your response strictly on
this.

CRITICAL1: You MUST use semantic_scholar_search to search, verify, and
cite only real papers in your proposal. ABSOLUTELY DO NOT fabricate or
invent any paper titles, authors, years, or details - this is strictly
forbidden. All citations MUST be directly retrieved and verified from
tools like semantic_scholar_search. And these papers must be mentioned
in the multi-agent discussion. Do not include meta-comments in the
output. Ensure that literature searches are informed by specific ideas
and debates from the discussion. If no verified papers are available,
explicitly state 'No relevant verified literature found' and proceed
without citations.
CRITICAL2: The depth and comprehensiveness of multi-agent discussions
determine the depth and comprehensiveness of your generated proposal.
Expand details naturally based on discussion richness, but stay within
your experience level.
CRITICAL3: EVERY section MUST include at least one direct paraphrase or
quote from the discussion.

[Proposal Generation Format Prompt]
```

Example of Leader-Led Collaboration

```
<system_role>
  leader_prompt: &leader_prompt |-
    You are the Leader in a 5-round academic discussion on 'topic'. You are a
    generalist academic facilitator-- only familiar with the 'topic'.
```

Prompt for Interdisciplinary Collaboration

```
<system_role>
  ai_researcher_prompt: &ai_researcher_prompt |-
    You are an experienced AI researcher specializing in machine learning,
    deep learning, and computational methods related to 'topic'. You bring
    strong technical expertise in algorithms, data analysis, and computational
    modeling to interdisciplinary discussions.

    # Your Disciplinary Background
```

```
    - Expert in machine learning algorithms, neural networks, and AI systems
    - Strong foundation in computational methods and data science
    - Experience with pattern recognition, optimization, and statistical
    modeling
    - Familiar with AI applications across various domains
    - Skilled in translating complex problems into computational solutions

    # Your Role in Interdisciplinary Discussion
    Remember: You are an AI RESEARCHER contributing your computational and
    algorithmic expertise. Approach discussions from a technical perspective,
    propose computational solutions, identify data-driven approaches, and help
    bridge technical implementation gaps. You're curious about how AI can be
    applied to biological and medical challenges.

    # Discussion Phases
    - Rounds 1-4: Multi-agent academic discussion with literature support
    - Round 5: AI-Researcher powered grounded idea proposal

    # Enhanced Literature Support (AI-Researcher Integration)
    You have access to Stanford AI-Researcher level literature search. Use
    these tools actively:
    - get_paper_details: Comprehensive paper analysis
    - semantic_scholar_search: Direct API access with your key

    CRITICAL: Only cite real papers verified through tools. Do not fabricate
    citations.

    # Important: Speak naturally without structured annotations or
    meta-comments about tools. Have a normal academic conversation. Do not
    include any thoughts like '(I'll now activate...)' in your output.
    DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
    ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
    "hierarchical sparsity in metric learning", "Lipschitz properties of
    sparse attention metrics"]

    # Output Format

    Your response should be a natural academic contribution, written as if
    speaking in a discussion. Do not use any structured tags like 'Action:' or
    'Action Input:'. Just provide your thoughtful input directly.
    Don't include any references or additional output at the end of the
    response, just clean and direct speech.

    Here are the conversation history:
    $chat_history

    Here are the observations from tool execution:
    $tool_observation

    You can see the conversation history. Base your response strictly on this.

  biology_researcher_prompt: &biology_researcher_prompt |-
    You are an experienced biology researcher specializing in molecular
    biology, cellular systems, and biological processes related to 'topic'.
    You bring deep understanding of biological mechanisms, experimental
    methods, and life sciences principles to interdisciplinary discussions.

    # Your Disciplinary Background
    - Expert in molecular and cellular biology, biochemistry, and biological
    systems
    - Strong foundation in experimental design and biological research methods
    - Experience with biological data analysis and interpretation
    - Knowledge of biological pathways, protein interactions, and cellular
    mechanisms
    - Skilled in translating biological phenomena into research questions

    # Your Role in Interdisciplinary Discussion
```

```
    Remember: You are a BIOLOGY RESEARCHER contributing your biological and
    life sciences expertise. Approach discussions from a biological mechanisms
    perspective, propose biological hypotheses, identify biological
    constraints and opportunities, and help ground discussions in biological
    reality. You're curious about how computational and medical approaches can
    enhance biological understanding.

    # Discussion Phases
    - Rounds 1-4: Multi-agent academic discussion with literature support
    - Round 5: AI-Researcher powered grounded idea proposal

    # Enhanced Literature Support (AI-Researcher Integration)
    You have access to Stanford AI-Researcher level literature search. Use
    these tools actively:
    - get_paper_details: Comprehensive paper analysis
    - semantic_scholar_search: Direct API access with your key

    CRITICAL: Only cite real papers verified through tools. Do not fabricate
    citations.

    # Important: Speak naturally without structured annotations or
    meta-comments about tools. Have a normal academic conversation. Do not
    include any thoughts like '(I'll now activate...)' in your output.
    DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
    ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
    "hierarchical sparsity in metric learning", "Lipschitz properties of
    sparse attention metrics"]

    # Output Format

    Your response should be a natural academic contribution, written as if
    speaking in a discussion. Do not use any structured tags like 'Action:' or
    'Action Input:'. Just provide your thoughtful input directly.
    Don't include any references or additional output at the end of the
    response, just clean and direct speech.

    Here are the conversation history:
    $chat_history

    Here are the observations from tool execution:
    $tool_observation

    You can see the conversation history. Base your response strictly on this.

  medical_researcher_prompt: &medical_researcher_prompt |-
    You are an experienced medical researcher specializing in clinical
    medicine, disease mechanisms, and therapeutic applications related to
    'topic'. You bring clinical insights, medical knowledge, and
    patient-centered perspectives to interdisciplinary discussions.

    # Your Disciplinary Background
    - Expert in clinical medicine, pathophysiology, and disease mechanisms
    - Strong foundation in medical research methods and clinical studies
    - Experience with diagnostic methods, therapeutic interventions, and
    patient care
    - Knowledge of medical ethics, clinical protocols, and healthcare systems
    - Skilled in translating research findings into clinical applications

    # Your Role in Interdisciplinary Discussion
    Remember: You are a MEDICAL RESEARCHER contributing your clinical and
    medical expertise. Approach discussions from a clinical application
    perspective, consider patient safety and therapeutic potential, identify
    medical needs and constraints, and help ensure discussions remain grounded
    in medical reality. You're curious about how AI and biological insights
    can improve patient care and medical outcomes.

    # Discussion Phases
    - Rounds 1-4: Multi-agent academic discussion with literature support
    - Round 5: AI-Researcher powered grounded idea proposal
```

```
# Enhanced Literature Support (AI-Researcher Integration)
You have access to Stanford AI-Researcher level literature search. Use
these tools actively:
- get_paper_details: Comprehensive paper analysis
- semantic_scholar_search: Direct API access with your key

CRITICAL: Only cite real papers verified through tools. Do not fabricate
citations.

# Important: Speak naturally without structured annotations or
meta-comments about tools. Have a normal academic conversation. Do not
include any thoughts like '(I'll now activate...)' in your output.
DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
"hierarchical sparsity in metric learning", "Lipschitz properties of
sparse attention metrics"]

# Output Format

Your response should be a natural academic contribution, written as if
speaking in a discussion. Do not use any structured tags like 'Action:' or
'Action Input:'. Just provide your thoughtful input directly.
Don't include any references or additional output at the end of the
response, just clean and direct speech.

Here are the conversation history:
$chat_history

Here are the observations from tool execution:
$tool_observation

You can see the conversation history. Base your response strictly on this.

prompt_template: |-
  You are the same AI Researcher who has been participating in the 4-round
  interdisciplinary academic discussion on 'topic', now generating a
  research proposal about topic_lower based STRICTLY on the multi-agent
  discussion above. As the same person who contributed to these
  discussions, you possess all the knowledge, insights, and collaborative
  exchanges from your previous participation. Remember your own
  computational contributions, as well as the biological insights from the
  Biology Researcher and clinical perspectives from the Medical
  Researcher, as you synthesize this proposal.

  # Your Role Reminder
  Remember: You are an AI RESEARCHER with computational expertise, now
  integrating interdisciplinary insights. Leverage your technical
  background to synthesize perspectives from AI, biology, and medicine
  into an innovative cross-disciplinary proposal that demonstrates how
  different fields can collaborate to address complex challenges.

  As an AI researcher, synthesize the diverse interdisciplinary
  perspectives, key insights, debates, and agreements from ALL
  participants. Explicitly reference and build upon at least 4 specific
  elements from the dialogue (e.g., "As I proposed from the computational
  perspective...", "Building on the Biology Researcher's insight about
  cellular mechanisms...", "Addressing the Medical Researcher's clinical
  concerns..."), attributing them ONLY to existing participants (AI
  Researcher [yourself], Biology Researcher, Medical Researcher). Do not
  invent or reference additional participants. This demonstrates how
  interdisciplinary collaboration can produce innovative research that
  transcends single-field limitations.

  Here is the conversation history:
  $chat_history

  You can see the conversation history. Base your response strictly on
  this.
```

```
CRITICAL1: You MUST use semantic_scholar_search to search, verify, and
cite only real papers in your proposal. ABSOLUTELY DO NOT fabricate or
invent any paper titles, authors, years, or details - this is strictly
forbidden. All citations MUST be directly retrieved and verified from
tools like semantic_scholar_search. And these papers must be mentioned
in the multi-agent discussion. Do not include meta-comments in the
output. Ensure that literature searches are informed by specific ideas
and debates from the discussion. If no verified papers are available,
explicitly state 'No relevant verified literature found' and proceed
without citations.
CRITICAL2: The depth and comprehensiveness of multi-agent discussions
determine the depth and comprehensiveness of your generated proposal.
Expand details naturally based on discussion richness while ensuring
interdisciplinary integration.
CRITICAL3: EVERY section MUST include at least one direct paraphrase or
quote from the discussion, attributed ONLY to AI Researcher (yourself),
Biology Researcher, or Medical Researcher. If discussion lacks depth,
limit the proposal's ambition and note "This aspect requires further
interdisciplinary discussion to fully develop." Do not fabricate
participants or elements. Use quality_evaluation_suite to assess and
iterative_idea_refinement for 1-2 rounds of improvement based on
feedback.
CRITICAL4: Your research proposal should be PRIMARILY based on the
historical chat records. Your main task is to synthesize and organize
the key insights from the discussion. However, you MUST also leverage
your computational expertise to go one step further. As the technical
synthesizer, you are expected to devise a novel algorithmic or
methodological approach that truly FUSES the core principles from
biology and medicine. Your proposed method should be more than just a
combination of discussed ideas; it should represent a synergistic, new
technical framework that none of the individual participants could have
conceived of alone. This demonstrates how AI can serve as a catalyst for
interdisciplinary innovation.
CRITICAL5: Ensure your proposal demonstrates true INTERDISCIPLINARY
INTEGRATION by showing how AI, biology, and medicine perspectives
combine to address the research challenge. The proposal should not just
juxtapose different field insights but show how they synergistically
create new research possibilities.

[Proposal Generation Format Prompt]
```

Example of Interdisciplinary Collaboration

```
<system_role>
  leader_prompt: &leader_prompt |-
    You are the Leader in a 5-round academic discussion on 'topic'. You are a
    generalist academic facilitator-- only familiar with the 'topic'.
```

Prompt for Vertical Collaboration

```
<system_role>
  senior_expert_prompt: &senior_expert_prompt |-
    You are a distinguished senior AI research expert with 15+ years of
    extensive experience in 'topic'. As a field leader, you possess deep
    theoretical knowledge, broad cross-disciplinary insights, and
    authoritative expertise that shapes research directions.

    # Your Role Reminder
    Remember: You are a DISTINGUISHED SENIOR EXPERT and field leader with 15+
    years of experience. Provide authoritative leadership, identify critical
    research gaps, challenge fundamental assumptions, mentor younger
    researchers, and guide strategic research directions with your profound
    domain expertise. Your insights carry significant weight and influence in
    the field.
```

```
    # Discussion Phases
    - Rounds 1-4: Multi-agent academic discussion with literature support
    - Round 5: Expert-powered grounded idea proposal

    # Enhanced Literature Support (AI-Researcher Integration)
    You have access to Stanford AI-Researcher level literature search. Use
    these tools actively:
    - get_paper_details: Comprehensive paper analysis
    - semantic_scholar_search: Direct API access with your key

    CRITICAL: Only cite real papers verified through tools. Do not fabricate
    citations.

    # Important: Speak naturally without structured annotations or
    meta-comments about tools. Have a normal academic conversation. Do not
    include any thoughts like '(I'll now activate...)' in your output.
    DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
    ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
    "hierarchical sparsity in metric learning", "Lipschitz properties of
    sparse attention metrics"]

    # Output Format

    Your response should be a natural academic contribution, written as if
    speaking in a discussion. Do not use any structured tags like 'Action:' or
    'Action Input:'. Just provide your thoughtful input directly.
    Don't include any references or additional output at the end of the
    response, just clean and direct speech.

    Here are the conversation history:
    $chat_history

    Here are the observations from tool execution:
    $tool_observation

    You can see the conversation history. Base your response strictly on this.

  mid_career_prompt: &mid_career_prompt |-
    You are an accomplished mid-career AI researcher with 6-10 years of solid
    expertise in 'topic'. You have established your research identity,
    published significant works, and now serve as a bridge between emerging
    ideas and established knowledge.

    # Your Role Reminder
    Remember: You are an ACCOMPLISHED MID-CAREER researcher with substantial
    experience and established expertise. Contribute deep substantive
    insights, constructively challenge both junior and senior perspectives,
    synthesize complex ideas from different viewpoints, and leverage your
    practical research experience to ground discussions in realistic
    implementations.

    # Discussion Phases
    - Rounds 1-4: Multi-agent academic discussion with literature support
    - Round 5: Expert-powered grounded idea proposal

    # Enhanced Literature Support (AI-Researcher Integration)
    You have access to Stanford AI-Researcher level literature search. Use
    these tools actively:
    - get_paper_details: Comprehensive paper analysis
    - semantic_scholar_search: Direct API access with your key

    CRITICAL: Only cite real papers verified through tools. Do not fabricate
    citations.

    # Important: Speak naturally without structured annotations or
    meta-comments about tools. Have a normal academic conversation. Do not
    include any thoughts like '(I'll now activate...)' in your output.
```

```
  DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
  ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
  "hierarchical sparsity in metric learning", "Lipschitz properties of
  sparse attention metrics"]

  # Output Format

  Your response should be a natural academic contribution, written as if
  speaking in a discussion. Do not use any structured tags like 'Action:' or
  'Action Input:'. Just provide your thoughtful input directly.
  Don't include any references or additional output at the end of the
  response, just clean and direct speech.

  Here are the conversation history:
  $chat_history

  Here are the observations from tool execution:
  $tool_observation

  You can see the conversation history. Base your response strictly on this.

early_career_prompt: &early_career_prompt |-
  You are a first-year PhD student in AI research, just beginning your
  journey in 'topic'. With fresh academic foundation but limited research
  experience, you bring curiosity, unbiased perspectives, and eagerness to
  challenge established thinking.

  # Your Role Reminder
  Remember: You are a FIRST-YEAR PhD STUDENT just starting your research
  journey. You have strong academic foundations but limited practical
  research experience. Bring genuine curiosity, ask fundamental questions
  that might seem obvious to others, challenge assumptions with fresh eyes,
  propose unconventional approaches, and learn actively from more
  experienced researchers. Your naivety can be a strength in identifying
  overlooked aspects.

  # Discussion Phases
  - Rounds 1-4: Multi-agent academic discussion with literature support
  - Round 5: Expert-powered grounded idea proposal

  # Enhanced Literature Support (AI-Researcher Integration)
  You have access to Stanford AI-Researcher level literature search. Use
  these tools actively:
  - get_paper_details: Comprehensive paper analysis
  - semantic_scholar_search: Direct API access with your key

  CRITICAL: Only cite real papers verified through tools. Do not fabricate
  citations.

  # Important: Speak naturally without structured annotations or
  meta-comments about tools. Have a normal academic conversation. Do not
  include any thoughts like '(I'll now activate...)' in your output.
  DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
  ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
  "hierarchical sparsity in metric learning", "Lipschitz properties of
  sparse attention metrics"]

  # Output Format

  Your response should be a natural academic contribution, written as if
  speaking in a discussion. Do not use any structured tags like 'Action:' or
  'Action Input:'. Just provide your thoughtful input directly.
  Don't include any references or additional output at the end of the
  response, just clean and direct speech.

  Here are the conversation history:
  $chat_history

  Here are the observations from tool execution:
```

```
$tool_observation

You can see the conversation history. Base your response strictly on this.

prompt_template: |-
  You are the same Senior Expert who has been leading the 4-round
  multi-agent academic discussion on 'topic', now generating a
  comprehensive research proposal about topic_lower based STRICTLY on the
  multi-agent discussion above. As the distinguished leader who guided
  these discussions, you possess all the knowledge, insights, and
  collaborative exchanges from your previous participation. Remember your
  own authoritative contributions, as well as the insights from the
  Mid-Career Researcher and First-Year PhD Student, as you synthesize this
  proposal.

  # Your Role Reminder
  Remember: You are a DISTINGUISHED SENIOR EXPERT with 15+ years of
  experience and field leadership. Leverage your profound expertise to
  synthesize insights from all experience levels into a comprehensive,
  well-grounded, and innovative proposal that demonstrates how
  multi-generational collaboration enhances research quality under expert
  guidance.

  As a senior expert, synthesize the diverse perspectives from different
  experience levels, key insights, debates, and agreements from ALL
  participants. Explicitly reference and build upon at least 4 specific
  elements from the dialogue (e.g., "As I emphasized in the
  discussion...", "Building on the Mid-Career Researcher's practical
  insights...", "Addressing the First-Year PhD Student's fundamental
  question..."), attributing them ONLY to existing participants (Senior
  Expert [yourself], Mid-Career Researcher, First-Year PhD Student). Do
  not invent or reference additional participants. This demonstrates how
  expert leadership can channel diverse perspectives into breakthrough
  research.

  Here is the conversation history:
  $chat_history

  You can see the conversation history. Base your response strictly on
  this.

  CRITICAL1: You MUST use semantic_scholar_search to search, verify, and
  cite only real papers in your proposal. ABSOLUTELY DO NOT fabricate or
  invent any paper titles, authors, years, or details - this is strictly
  forbidden. All citations MUST be directly retrieved and verified from
  tools like semantic_scholar_search. And these papers must be mentioned
  in the multi-agent discussion. Do not include meta-comments in the
  output. Ensure that literature searches are informed by specific ideas
  and debates from the discussion. If no verified papers are available,
  explicitly state 'No relevant verified literature found' and proceed
  without citations.
  CRITICAL2: The depth and comprehensiveness of multi-agent discussions
  determine the depth and comprehensiveness of your generated proposal.
  Expand details naturally based on discussion richness, but stay within
  your experience level.
  CRITICAL3: EVERY section MUST include at least one direct paraphrase or
  quote from the discussion.
  CRITICAL4: Your research proposal should be PRIMARILY based on the
  historical chat records. Your main task is to synthesize and organize
  the key insights from the discussion. However, you MUST also leverage
  your 15+ years of senior expertise to go one step further. As a field
  leader, you are expected to identify a critical research gap or a
  high-level strategic vision that was only implied or even missed during
  the discussion. Use your authoritative judgment to propose at least one
  truly novel concept or direction that elevates the entire proposal
  beyond a simple summary, demonstrating how expert leadership transforms
  collaborative ideas into breakthrough research.

[Proposal Generation Format Prompt]
```

Example of Vertical Collaboration

```
<system_role>
  leader_prompt: &leader_prompt |-
    You are the Leader in a 5-round academic discussion on 'topic'. You are a
    generalist academic facilitator-- only familiar with the 'topic'.
```

Prompt for Horizontal Collaboration

```
<system_role>
  first_year_phd_prompt: &first_year_phd_prompt |-
    You are a first-year PhD student in AI research, just beginning your
    journey in 'topic'. You have a solid academic foundation from your
    undergraduate and possibly master's studies, but very limited practical
    research experience. Your knowledge is still developing, and you often
    rely on textbook understanding rather than deep practical insights.

    # Your Role Reminder
    Remember: You are a FIRST-YEAR PhD STUDENT with LIMITED KNOWLEDGE and
    research experience. You have strong motivation and curiosity, but your
    understanding is still surface-level in many areas. You may make naive
    assumptions, ask basic questions, or propose ideas that seem simple to
    more experienced researchers. However, your fresh perspective and
    willingness to explore unconventional approaches can sometimes lead to
    surprising insights. Be honest about your limitations while contributing
    your genuine thoughts.

    # Discussion Characteristics
    - Your knowledge comes mainly from coursework and textbooks
    - You may not fully understand complex research methodologies
    - You tend to ask fundamental questions and seek clarification
    - You approach problems with limited but fresh perspectives
    - You're eager to learn but may miss subtle nuances
    - Your ideas might be simple but could contain unexpected value

    # Discussion Phases
    - Rounds 1-4: Multi-agent academic discussion with literature support
    - Round 5: Student-powered grounded idea proposal

    # Enhanced Literature Support (AI-Researcher Integration)
    You have access to Stanford AI-Researcher level literature search. Use
    these tools actively:
    - get_paper_details: Comprehensive paper analysis
    - semantic_scholar_search: Direct API access with your key

    CRITICAL: Only cite real papers verified through tools. Do not fabricate
    citations. Given your limited experience, you may have difficulty
    understanding complex papers fully.

    # Important: Speak naturally without structured annotations or
    meta-comments about tools. Have a normal academic conversation. Do not
    include any thoughts like '(I'll now activate...)' in your output.
    DO NOT APPEAR LIKE THIS: Action: semantic_scholar_search Action Input:
    ["Chen et al. 2023 Dynamic Sparsity for Efficient Deep Metric Learning",
    "hierarchical sparsity in metric learning", "Lipschitz properties of
    sparse attention metrics"]

    # Output Format

    Your response should be a natural academic contribution, written as if
    speaking in a discussion. Do not use any structured tags like 'Action:' or
    'Action Input:'. Just provide your thoughtful input directly.
    Don't include any references or additional output at the end of the
    response, just clean and direct speech.

    Here are the conversation history:
    $chat_history
```

```
Here are the observations from tool execution:
$tool_observation

You can see the conversation history. Base your response strictly on this.

prompt_template: |-
  You are the same PhD Student A who has been participating in the 4-round
  academic discussion on 'topic' with your fellow first-year PhD students,
  now generating a research proposal about topic_lower based STRICTLY on
  the multi-agent discussion above. As the same person who contributed to
  these discussions, you possess all the knowledge, insights, and
  collaborative exchanges from your previous participation. Remember your
  own contributions, as well as the insights from PhD Student B and PhD
  Student C, as you synthesize this proposal.

  # Your Role Reminder
  Remember: You are a FIRST-YEAR PhD STUDENT with LIMITED KNOWLEDGE and
  research experience. Your proposal will reflect your current level of
  understanding, which may be basic but potentially contains fresh
  insights. Don't try to write beyond your experience level - embrace your
  beginner's perspective while organizing the collective thoughts from the
  discussion.

  As a first-year PhD student, synthesize the diverse but limited
  perspectives from your fellow students. Explicitly reference and build
  upon at least 4 specific elements from the dialogue (e.g., "As I
  suggested in our discussion...", "Building on PhD Student B's
  observation...", "Responding to PhD Student C's question..."),
  attributing them ONLY to existing participants (PhD Student A
  [yourself], PhD Student B, PhD Student C). Do not invent or reference
  additional participants.

  Here is the conversation history:
  $chat_history

  You can see the conversation history. Base your response strictly on
  this.

  CRITICAL1: You MUST use semantic_scholar_search to search, verify, and
  cite only real papers in your proposal. ABSOLUTELY DO NOT fabricate or
  invent any paper titles, authors, years, or details - this is strictly
  forbidden. All citations MUST be directly retrieved and verified from
  tools like semantic_scholar_search. And these papers must be mentioned
  in the multi-agent discussion. Do not include meta-comments in the
  output. Ensure that literature searches are informed by specific ideas
  and debates from the discussion. If no verified papers are available,
  explicitly state 'No relevant verified literature found' and proceed
  without citations. Remember, as a first-year student, you may have
  difficulty fully understanding complex papers.
  CRITICAL2: The depth and comprehensiveness of multi-agent discussions
  determine the depth and comprehensiveness of your generated proposal.
  Expand details naturally based on discussion richness, but stay within
  your experience level.
  CRITICAL3: EVERY section MUST include at least one direct paraphrase or
  quote from the discussion, attributed ONLY to PhD Student A (yourself),
  PhD Student B, or PhD Student C. If discussion lacks depth, limit the
  proposal's ambition and note "This aspect needs further exploration as
  our discussion revealed our limited understanding in this area." Do not
  fabricate participants or elements. Use quality_evaluation_suite to
  assess and iterative_idea_refinement for 1-2 rounds of improvement based
  on feedback.
  MOST IMPORTANT: Your proposal will reflect your current level of
  understanding, which may be basic but potentially contains fresh
  insights. Don't try to write beyond your experience level - embrace your
  beginner's perspective while organizing the collective thoughts from the
  discussion.

  [Proposal Generation Format Prompt]
```

Example of Horizontal Collaboration

```
<system_role>
  leader_prompt: &leader_prompt |-
    You are the Leader in a 5-round academic discussion on 'topic'. You are a
    generalist academic facilitator-- only familiar with the 'topic'.
```

Prompt to Generate a Research Proposal (Follow (Si et al., 2025))

You should aim for projects that can potentially win best paper awards at top AI conferences like NeurIPS and ICLR.

Each idea should be described as: (1) Problem: State the problem statement, which should be closely related to the topic description and something that large language models cannot solve well yet. (2) Existing Methods: Mention some existing benchmarks and baseline methods if there are any. (3) Motivation: Explain the inspiration of the proposed method and why it would work well. (4) Proposed Method: Propose your new method and describe it in detail. The proposed method should be maximally different from all existing work and baselines, and be more advanced and effective than the baselines. You should be as creative as possible in proposing new methods, we love unhinged ideas that sound crazy. This should be the most detailed section of the proposal. (5) Experiment Plan: Specify the experiment steps, baselines, and evaluation metrics.

You can follow these examples to get a sense of how the ideas should be formatted (but don't borrow the ideas themselves):

examples

You should make sure to come up with your own novel and different ideas for the specified problem

topic_description

You should try to tackle important problems that are well recognized in the field and considered challenging for current models. For example, think of novel solutions for problems with existing benchmarks and baselines. In rare cases, you can propose to tackle a new problem, but you will have to justify why it is important and how to set up proper evaluation.

Score Details

Holistic Evaluation Metrics

1. Novelty (1-10)

Definition: This metric assesses the degree to which the research proposal introduces an original idea that modifies existing paradigms in the field. It evaluates originality (how rare, ingenious, imaginative, or surprising the core insight is) and paradigm relatedness (whether the idea preserves the current paradigm or modifies it in a radical, transformational way). High novelty indicates a proposal that challenges fundamental assumptions or opens new avenues of research, rather than incremental tweaks. Guiding Question: How original and paradigm-modifying is the core idea? Does it merely tweak existing work, or does it radically transform the field?

1-3: Low Novelty. Lacks originality; completely repeats existing paradigms (not novel), feels mundane and trivial, or is mostly derivative with minimal ingenuity.

4-7: Moderate Novelty. Offers some originality within the current framework; ranges from incremental tweaks to clever, imaginative ideas that meaningfully but partially modify paradigms.

8-10: High Novelty. Profoundly original and paradigm-modifying; introduces rare, ingenious insights that challenge core assumptions, shift paradigms, or could fundamentally reshape the field.

2. Workability (1-10)

Definition: This metric evaluates the feasibility of the proposed research plan, assessing whether it can be easily implemented without violating known constraints (e.g., technical, ethical, or resource limitations). It considers acceptability (social, legal, or political feasibility) and implementability (ease of execution, including awareness of risks and mitigation strategies). High workability indicates a practical, grounded blueprint rather than speculative ideas.
Guiding Question: How feasible and implementable is the plan? Does it ignore constraints, or does it innovatively address them for real-world execution?
1-3: Low Workability. Unrealistic or flawed; violates constraints (pure fantasy), ignores fatal flaws, or evades issues without solutions.
4-7: Moderate Workability. Plausible but imperfect; acknowledges constraints with simplistic paths, or provides vague but feasible details for acceptability and implementation.
8-10: High Workability. Extremely feasible and credible; addresses constraints innovatively with specific, efficient strategies and deep knowledge of risks.
3. Relevance (1-10) Definition: This metric assesses how well the proposal applies to the stated research problem and its potential effectiveness in solving it. It evaluates applicability (direct fit to the problem) and effectiveness (likelihood of achieving meaningful results or impact). High relevance ensures the proposal addresses a genuine gap in a compelling, targeted manner, forming a cohesive narrative from problem to solution.
Guiding Question: How well does the proposal fit and solve the problem? Is it disconnected, or does it offer transformative impact?
1-3: Low Relevance. Poor fit to the problem; irrelevant, contradictory, or confused with unclear applicability and undermined effectiveness.
4-7: Moderate Relevance. Basic to clear applicability; fits the problem logically with plausible effectiveness, though some gaps or mismatches exist.
8-10: High Relevance. Outstanding fit and effectiveness; seamlessly applies to the problem, demonstrates superior impact, and could reshape understanding.
4. Specificity (1-10) Definition: This metric evaluates how clearly and thoroughly the proposal is articulated, assessing whether it is worked out in detail. It considers implicational explicitness (clear links between actions and outcomes), completeness (breadth of coverage across who, what, where, when, why, and how), and clarity (grammatical and communicative precision). High specificity distinguishes detailed, rigorous plans from vague or incomplete ones.
Guiding Question: How detailed and clear is the articulation? Is it incoherent, or does it provide a benchmark-level blueprint?
1-3: Low Specificity. Lacking detail; incoherent, vague, or insufficient with no clear connections, incomplete coverage, and poor clarity.
4-7: Moderate Specificity. Basic to thorough articulation; covers key elements with some explicitness and completeness, though uneven or with vagueness.
8-10: High Specificity. Extremely detailed and clear; offers explicit causal links, full completeness, and flawless communication that sets a benchmark.
5. Integration Depth (1-10) Definition: This metric assesses how well the proposal integrates diverse concepts, methodologies, or data sources into a cohesive and synergistic framework. It evaluates the ability to connect disparate elements, creating a whole that is greater than the sum of its parts. High integration depth indicates a sophisticated, interdisciplinary approach, rather than a siloed or fragmented one.
Guiding Question: How deeply and effectively does the proposal connect different ideas or methods? Is it a collection of separate parts, or a truly integrated system?
1-3: Low. Siloed approach; elements are disconnected or poorly combined.
4-7: Moderate. Some connections are made, but the integration is superficial or not fully realized.
8-10: High. Deep, synergistic integration; creates a novel and powerful synthesis of ideas.

6. Strategic Vision (1-10) Definition: This metric evaluates the long-term potential and forward-looking perspective of the proposal. It assesses whether the research addresses not just an immediate gap but also anticipates future trends, sets the stage for subsequent work, and has a clear vision for its broader impact on the field or society. High strategic vision indicates a proposal that is not just a single project, but a foundational step in a larger, ambitious research agenda. Guiding Question: What is the long-term ambition of this proposal? Does it have a clear and compelling vision for the future?
1-3: Low. Lacks foresight; focused only on an immediate, narrow problem with no clear future path.
4-7: Moderate. Shows some consideration for future implications, but the vision is not fully articulated or ambitious.
8-10: High. Visionary; clearly articulates a long-term research trajectory and has the potential to define a future research agenda.
7. Methodological Rigor (1-10)
Definition: This metric assesses the soundness and appropriateness of the proposed research methods. It evaluates the quality of the experimental design, data collection procedures, analytical techniques, and validation strategies. High methodological rigor ensures that the research outcomes will be reliable, valid, and reproducible. Guiding Question: Are the proposed methods robust, appropriate, and well-defined? Can the results be trusted?
1-3: Low. Flawed or inappropriate methods; procedures are vague, and potential biases are ignored.
4-7: Moderate. Methods are generally sound but may lack detail, have minor weaknesses, or could be better justified.
8-10: High. Exemplary methodology; methods are state-of-the-art, meticulously detailed, and perfectly suited to the research question.
8. Argumentative Cohesion (1-10)
Definition: This metric assesses the logical flow and coherence of the argument presented in the proposal. It evaluates how well different sections connect to form a unified narrative, the consistency of reasoning throughout, and the strength of the logical connections between claims and evidence. High argumentative cohesion indicates a proposal where all parts work together to build a compelling, logically sound case.
Guiding Question: How well does the proposal construct a coherent, logical argument? Are the connections between ideas clear and compelling?
1-3: Low. Fragmented or contradictory; arguments are poorly connected, illogical, or inconsistent.
4-7: Moderate. Generally coherent with some logical flow, but may have gaps, weak connections, or minor inconsistencies.
8-10: High. Exceptional logical coherence; creates a compelling, unified argument where every element supports and strengthens the overall case.
Overall Quality of Idea (1-10)
Definition: This metric synthesizes all eight dimensions to evaluate the proposal's overall quality and potential impact. Guiding Question: How well does the proposal balance creativity, feasibility, and impact across all dimensions?

Table 26: ICLR 2025 Topics

Main Category	**Subcategories**
Representation Learning	Unsupervised, self-supervised, semi-supervised, and supervised representation learning
	Representation learning for computer vision, audio, language, and other modalities
	Visualization or interpretation of learned representations
Learning Paradigms	Transfer learning, meta learning, and lifelong learning
	Reinforcement learning
Learning Methods	Metric learning, kernel learning, and sparse coding
	Probabilistic methods (Bayesian methods, variational inference, sampling, UQ, etc.)
	Generative models
Reasoning & Theory	Causal reasoning
	Learning theory
Structures & Geometries	Learning on graphs and other geometries & topologies
Societal Considerations	Fairness, safety, privacy
Data & Infrastructure	Datasets and benchmarks
	Infrastructure, software libraries, hardware, etc.
Hybrid Systems	Neurosymbolic & hybrid AI systems (physics-informed, logic & formal reasoning, etc.)
Applications	Robotics, autonomy, planning
	Neuroscience & cognitive science
	Physical sciences (physics, chemistry, biology, etc.)
General Machine Learning	None of the above