An Empirical Evaluation of Diversity in Multi-Agent Idea Generation

Anonymous ACL submission

Abstract

Multi-agent systems (MAS) are increasingly used for open-ended idea generation, motivated by the belief that multi-agent interaction naturally increases 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 using scientific proposal generation as a controlled testbed. We analyze diversity across three 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 information gain despite increased sampling or agent count. At the cognition level, we uncover a paradox of expertise: authority- or expert-driven collaborations consistently suppress semantic diversity, while junior-dominated interactions explore broader idea spaces. At the system level, larger groups and denser communication often accelerate premature convergence. Using complementary diversity metrics validated by expert judgments, we show that diversity collapse arises primarily from interaction structure rather than model insufficiency. Our findings highlight the importance of preserving independence and disagreement when designing MAS for creative tasks.

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. Consequently, multi-agent interaction can end up amplifying shared priors rather than introducing genuine variety. Without rigorous design, simply increasing the number of agents does not guarantee broader exploration; instead, it may cause the system to repeatedly search the same narrow manifold at a higher computational cost (Jiang et al., 2025; Wenger and Kenett, 2025; Wynn et al., 2025).

In this work, we argue that pursuing diversity in MAS-based idea generation involves three fundamental and interconnected challenges that span from the model layer to emergent system dynamics. We frame challenges through a hierarchical lens:

First, at the level of **Model Intelligence**, we identify the **Compute Efficiency Paradox**. As foundation models scale in capability, their outputs often become more fluent and score better on standard correctness-oriented metrics, yet converge toward increasingly similar semantic content (Maynez et al., 2023). In high-capacity regimes, allocating additional computation or instantiating more agents can produce little to no marginal informa-

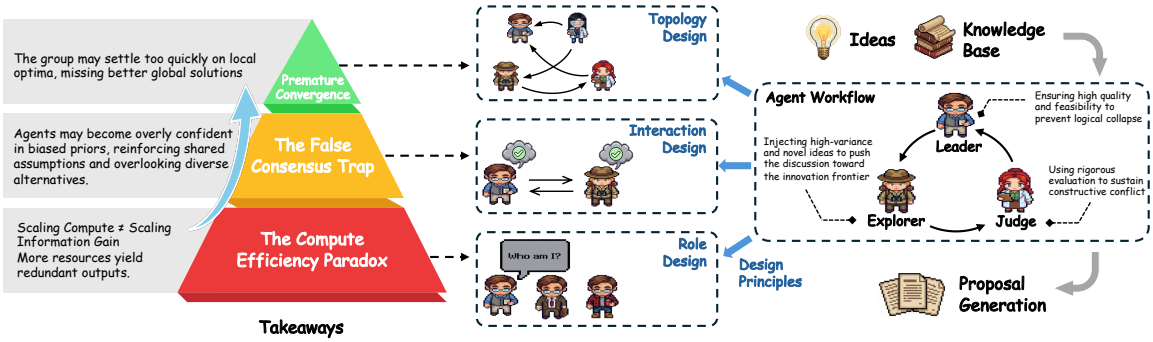

Figure 1: Design Principles and Workflow.

tion gain. From an information-theoretic perspective, this points to a decoupling in which greater model intelligence does not necessarily translate into a more informative expansion of the idea space (Coveney and Succi, 2025).

Second, at the level of **Agent Cognition**, interaction can trigger the **False Consensus Trap**. Although agents are often prompted with different personas or roles to elicit diverse viewpoints, they are still grounded in the same underlying inductive biases. As a result, group discussion can devolve into an *echo-chamber effect* in which agents reinforce one another's hallucinations or biases, producing a confident consensus without corresponding competence (Liu et al., 2024a; Wynn et al., 2025). This failure mode is especially hazardous in open-ended tasks because it creates the appearance of careful deliberation while the system silently neglects plausible alternatives (Moon et al., 2025).

Finally, at the level of **System Complexity**, increasing group size, interaction rounds, or communication topology can exacerbate **Premature Convergence**. If we view idea generation as search over a high-dimensional landscape, parallel interactions should in principle explore a broader region. In practice, however, protocols that implicitly reward fast agreement often push the group to collapse early onto local optima (Wynn et al., 2025; Wang et al., 2025). Under these conditions, additional system complexity, such as more discussion rounds, tends to generate redundant trajectories rather than truly divergent exploration (Moon et al., 2025; Shen et al., 2025).

To address these challenges, we present a systematic analysis of MAS-based idea generation, using the formulation of **proposals** as a proxy for measuring idea diversity. By dissecting the trade-offs between system scaling and information gain, we reveal why MAS interactions may succeed or fail in fostering genuine diversity.

Our analysis demonstrates that naive scaling of agents or interaction rounds does not inherently increase diversity; instead, structural factors—such as alignment, authority gradients, and communication density, play a dominant role in driving diversity collapse. Conversely, mechanisms that explicitly maintain independence and partial isolation among agents are crucial for sustaining a broad exploration of the idea space.

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

Our study investigates the emergence and potential collapse of idea diversity within multi-agent systems. Unlike deterministic tasks() where convergence to a single ground truth is desirable, ideation requires navigating a complex, high-dimensional search space to uncover distinct, plausible solutions. 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 B.

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 (with temperature set to 0.7) on each of the 20 topics listed in Table 3, resulting in a total of 1,000 proposals per setting.

Specific experimental setups, including agent prompts and topologies, are detailed in Appendix H.2.

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 D and sensitivity analysis in Appendix G.

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. The resulting $1 - \phi$ acts as a proxy for convergence, smaller for a collapse toward a single centroid (an "Echo Chamber" state), while larger for 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, indicating how "far" the system casts its net in the embedding space.

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 C) 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.

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 multi-agent systems.

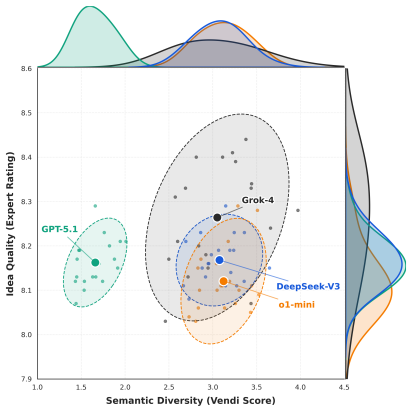

Figure 2: **Empirical Quality–Diversity Landscape of Single-Model Generation.** Each point represents a generated research proposal under identical ideation settings. The X-axis shows topic-level Effective Diversity (Vendi Score), and the Y-axis shows aggregated expert-rated Idea Quality. The landscape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

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 quality 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: The Paradox of Expertise

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 designed to mimic real-world scientific collaboration (see Appendix for details):

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 cognitive structures (Figure 3). Contrary to common intuition, expert guidance or disciplinary breadth does not improve ideation diversity. Instead, junior-dominated horizontal collaboration consistently achieves the highest semantic and lexical diversity, suggesting that the absence of authority and prior commitment enables broader exploration.

In contrast, interdisciplinary expert teams exhibit the lowest diversity, despite spanning multiple fields. Without explicit mechanisms to preserve disagreement, authority-heavy interactions quickly collapse into polite consensus, highlighting the risk of false agreement in expert-driven multi-agent systems.

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.

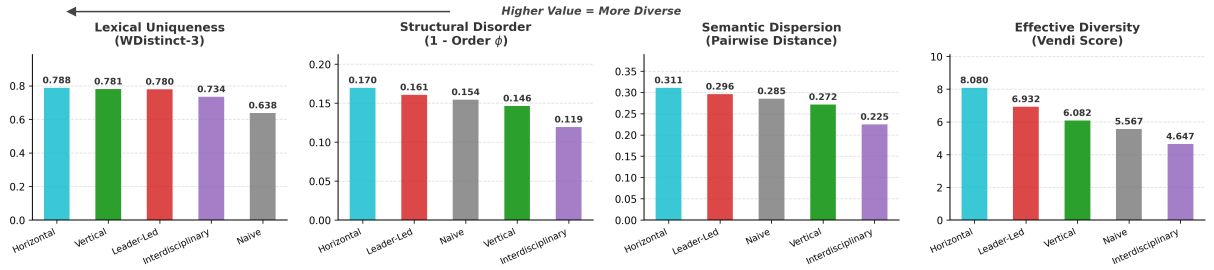

Figure 3: **Diversity Metrics across Cognitive Structures. Horizontal** collaboration (Junior-driven) consistently maximizes diversity (Vendi: 8.08), identifying the "Unbound Junior" effect. Surprisingly, **Interdisciplinary** collaboration exhibits the lowest diversity (Vendi: 4.65), suggesting that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

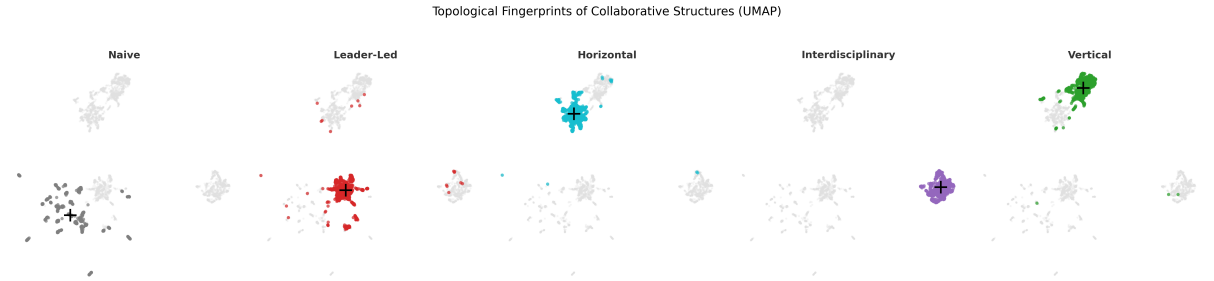

Figure 4: **Semantic Regimes of Cognitive Structures.** UMAP projection reveals a bifurcation. The **Conservative Cluster** (Bottom) is dominated by expert-driven structures (Leader-Led, Interdisciplinary), while the **Innovation Frontier** (Top) is populated by junior-driven structures (Horizontal,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

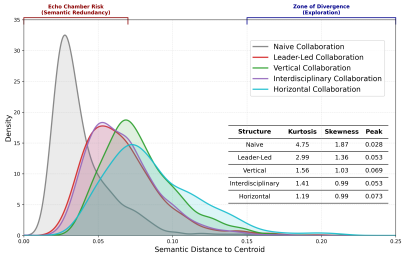

Figure 5: **Semantic Distance Density. Naive** (Grey) and **Leader-Led** (Red) distributions peak sharply near zero, indicating "Gravitational Collapse" (Echo Chambers). In contrast, **Horizontal** (Cyan) and **Vertical** (Green) structures flatten the curve, shifting density into the "Zone of Divergence"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

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 a "Senior Authority" acts as a strong attractor. Junior agents likely succumb to sycophancy, aligning their vectors with the leader rather than offering orthogonal critiques.

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 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.

Summary: The analysis of Agent Cognition reveals a "Paradox of Expertise": while senior roles are often assumed to improve quality, they actively suppress diversity through authority bias. The highest diversity emerges from the **Horizontal** (peer-to-peer) dynamic, while the **Vertical** structure offers a compromise, mitigating the chaos of juniors with the structure of seniors.

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.

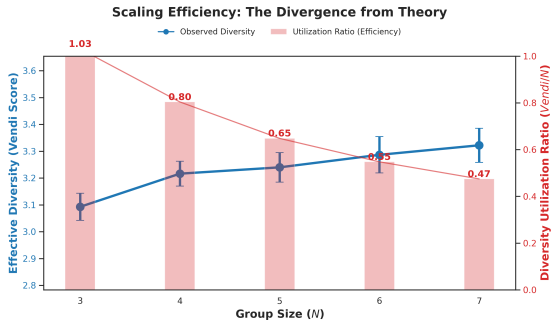

Figure 6: The Divergence from Theory in Scaling Efficiency. This plot compares the observed Effective Diversity (Vendi Score, blue line) against the theoretical Diversity Utilization Ratio (Vendi/N, red bars) as group size increases. While diversity grows, the efficiency per agent drops significantly.

We first investigate the impact of increasing the number of agents on the diversity of proposals. Figure 6 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.

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.

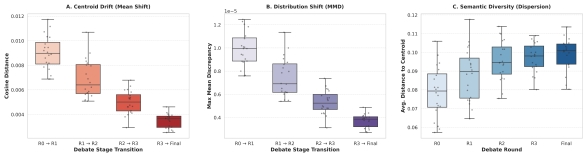

Figure 7: Quantitative Evolution of Semantic Dynamics. (A) Semantic Drift Velocity decreases, indicating stabilization of the consensus. (B) Distribution Shift (MMD) reduces, confirming structural convergence. (C) Semantic Diversity (Dispersion) increases, showing expansion within the consensus region.

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

As shown in Figure 7A 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" implies that agents, while agreeing on a general direction, continue to refine and expand the solution space locally, effectively avoiding mode collapse.

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

Figure 8 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 to the unstructured jumps expected from hallucination, providing strong evidence that the observed diversity stems from genuine deliberation and refinement.

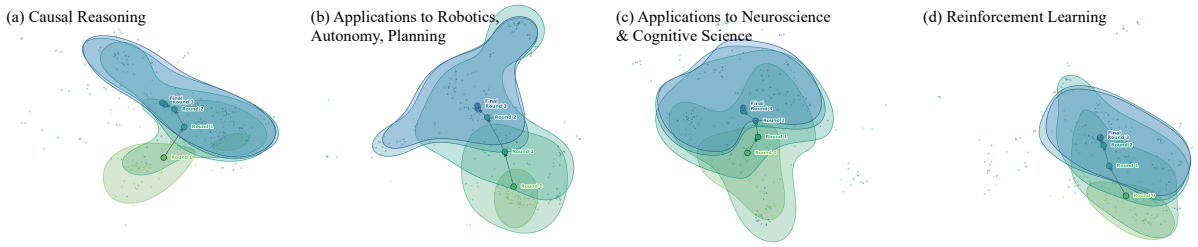

Figure 8: Evolutionary Semantic Trajectories. 2D projections of proposal embeddings across debate rounds for four representative topics. The trajectories show coherent drift (arrows) and expanding coverage (shaded regions), illustrating structured exploration rather than random movement.

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 F), influence the dynamics of diversity and conflict.

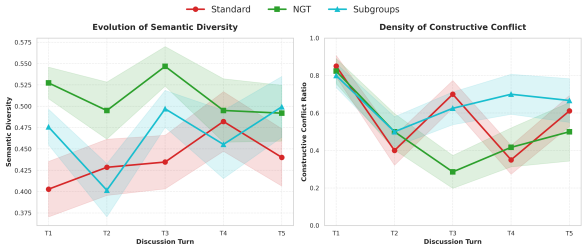

Figure 9: Mechanism of Process Intervention. (Left) Evolution of Semantic Diversity shows NGT's early advantage and Subgroups' late rebound. (Right) Density of Constructive Conflict highlights Subgroups' ability to sustain critical engagement. See Appendix H for the detailed prompting strategy and scoring rubric.

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

Figure 9 (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 9 (Right) reveals that Subgroups maintain the highest and most stable density of constructive conflict (interactions with critique score ≥ 7) 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.

6 Discussion

6.1 Interaction Effects: The Tripartite Balance

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 10 visualizes this interaction landscape, mapping the relationship between Consensus Strength (Interaction Density) and Semantic Diversity (Vendi Score) across different Model $\times$ Topology configurations.

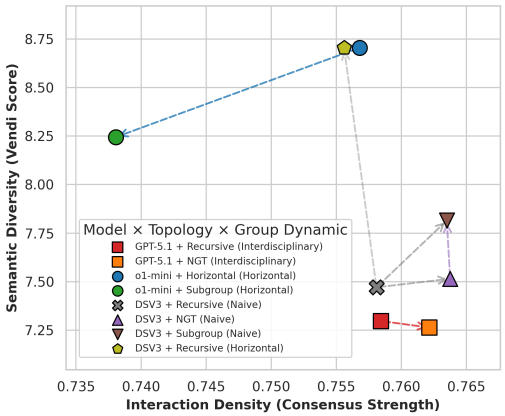

Figure 10: **The Interaction Landscape of Multi-Agent Ideation.** We map the trade-off between Interaction Density (Consensus Strength) and Semantic Diversity (Vendi Score) for distinct Model $\times$ Topology combinations. Arrows indicate the shift in system state when moving from a baseline (e.g., Recursive) to an intervention (e.g., NGT or Subgroup). The plot reveals that "weaker" models (DeepSeek-V3) benefit significantly from structural interventions (Subgroup/NGT), showing large positive vectors. In contrast, "strong" reasoning models (o1-mini) exhibit a resistance to topology, where enforcing subgroups paradoxically reduces diversity (negative vector), suggesting an "Alignment-Topology Mismatch."

The landscape reveals those interaction regimes: **Structural Interventions Rescue "Weaker"**

Models. For models with lower intrinsic reasoning priors (e.g., DeepSeek-V3, grey/purple markers), topological interventions yield the highest marginal gains. As indicated by the upward-pointing grey arrows, switching from a naive Recursive structure (Grey X) to Subgroup (Brown Inverted Triangle) or NGT (Purple Triangle) results in a substantial vertical leap in diversity with a moderate increase in consensus strength. This confirms that for standard LLMs, process engineering effectively compensates for lack of intrinsic exploration.

The "Reasoning-Topology" Mismatch. Surprisingly, for the reasoning-heavy model (o1-mini, blue/green markers), the impact of topology is inverted. The blue dashed arrow shows that moving from Horizontal (Blue Circle) to Subgroup (Green Circle) causes a dramatic drop in diversity and a sharp decrease in interaction density. This suggests that highly coherent reasoning chains are fragile; partitioning them into subgroups disrupts the model's internal deliberation flow, leading to fragmentation rather than productive divergence.

The Inertia of Strong Alignment. The heavily aligned model (GPT-5.1, red/orange markers) occupies the lowest diversity region and remains largely unresponsive to topological shifts. The red arrow indicates that moving from Recursive to NGT (Orange Square) produces negligible change in diversity, merely shifting the system slightly along the consensus axis. This reinforces our earlier finding of an "Alignment Tax," where safety tuning creates a rigid semantic floor that structural interventions alone cannot breach.

6.2 Task Characterization: Generalizability via Topological Profiling

A critical prerequisite for claiming generalizable MAS dynamics is determining whether observed behaviors are artifacts of a specific domain or fundamental properties of collaborative topology. To address this, we contextualize our primary domain (AI Research) within the theoretical frameworks of the Task Circumplex (McGrath, 1984) and the Intellective-Judgmental Continuum (Laughlin, 1980). We benchmark the baseline behavior of LLM agents across four distinct task types, Physics, Policy, Creative Writing, and AI Research,to characterize their **Intrinsic Entropy**.

Figure 11 visualizes this spectrum, revealing that AI Research is not merely a convenience choice, but a representative proxy for the most challenging class of ideation problems.

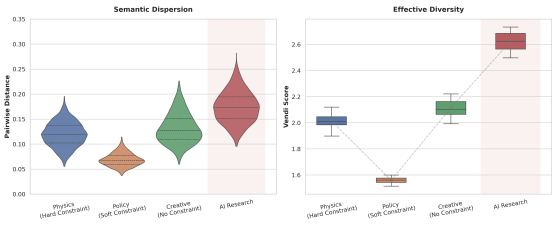

Figure 11: **The Intrinsic Entropy Spectrum across Cognitive Domains.** We benchmark baseline diversity (Inner-Topic Vendi Score, $N = 50$) across four task types to validate domain representativeness. **(Left) Semantic Dispersion:** "Intellective" tasks like Physics and Policy exhibit tight distributions 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 main conclusions about our task setting:

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

AI Research is a "Hybrid Constraint" topology with the highest intrinsic entropy. AI Research uniquely combines high entropy (Vendi Score > 2.6) with strict rigor, requiring both broad exploration and logical soundness—unlike either unconstrained creative or strictly convergent tasks.

AI Research as a generalizable testbed. Its position at the "Edge of Chaos" makes it the most challenging for MAS collaboration; success here suggests structural findings will transfer to other complex, constraint-rich domains.

7 Conclusion

We systematically evaluated diversity in multi-agent systems for open-ended idea generation, using scientific proposal tasks as a testbed. Our results show that simply increasing agent count or interaction rounds does not guarantee greater idea diversity. Instead, diversity collapse is primarily caused by structural factors: strong alignment, authority-driven dynamics, and dense communication all promote premature consensus. In contrast, interaction designs that preserve independence, such as blind generation or subgroup isolation consistently yield higher diversity without loss of 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, all agents in our experiments are instantiated from a single underlying language model per condition, differing only in prompts, roles, or interaction topology. This design isolates the effects of interaction and structure, but does not capture additional diversity that may arise from architectural or pretraining heterogeneity across models. Extending the analysis to heterogeneous model ensembles is a natural direction for 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. 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 fixed hyperparameters (e.g., group size, number of rounds, sampling temperature). While we observe consistent trends across settings, adaptive or dynamically optimized interaction strategies may exhibit different behaviors that are not captured in this study.

Ethical Statement

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.

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.

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.

Mohor Banerjee, Nadya Yuki Wangsajaya, Syed Ali Redha Alsagoff, Min Sen Tan, Zachary Choy Kit Chun, and Alvin Chan Guo Wei. 2025. Does less hallucination mean less creativity? an empirical investigation in llms. *Preprint*, arXiv:2512.11509.

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

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *Preprint*, arXiv:2304.05376.

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.

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

Andre Delbecq, Andrew Ven, and David Gustafson. 1986. Group techniques for program planning: A guide to nominal group and delphi processes. *Glenview, Illinois: Scott Forman and Co*.

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.

Giorgio Franceschelli and Mirco Musolesi. 2024. On the creativity of large language models. *AI & SOCIETY*, 40(5):3785–3795.

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.

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*.

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.

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. *Preprint*, arXiv:2305.19118.

Andy Liu, Mona Diab, and Daniel Fried. 2024a. 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.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024b. A dynamic llm-powered agent network for task-oriented agent collaboration. *Preprint*, arXiv:2310.02170.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9194–9213, Toronto, Canada. Association for Computational Linguistics.

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.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit,

Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.

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.

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.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1075–1108, Albuquerque, New Mexico. Association for Computational Linguistics.

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. 2025. 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.

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.

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.

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 Related Work

A.1 Heterogeneity and Specialization in MAS

Beyond simple model ensembles (Ye et al., 2025), recent frameworks leverage heterogeneity through social-attribute modulation for error correction (Zhang et al., 2026) or diverse thinking prompts (He and Feng, 2025). However, such divergence-inducing protocols often prioritize consensus in closed-ended reasoning rather than open-ended ideation (Liang et al., 2024). Alternatively, functional heterogeneity is realized through domain-specific tool integration, enabling agents to address complex scientific tasks via specialized modules (Bran et al., 2023; Gottweis et al., 2025).

A.2 Multi-agent Debate

Advancing beyond static interaction structures, recent frameworks introduce dynamic team composition to enhance debate efficiency. **DyLAN** (Liu et al., 2024b) proposes a dynamic agent network that iteratively alternates between *team optimization* and *task solving*, utilizing inference-time selection to activate the most relevant agents for specific queries.

A.3 Paradigms of AI Collaboration in Ideation

While multi-agent debate is traditionally employed to enhance reasoning (Du et al., 2024), recent frameworks leverage heterogeneous teaming to boost scientific ideation (Su et al., 2025), or engineer diversity through explicit control mechanisms like the Proximity Agent (Gottweis et al., 2025). However, interaction dynamics introduce structural vulnerabilities. Wynn et al. (2025) identify that debate frequently suffers from sycophancy and "disagreement collapse," leading to premature consensus. Consequently, empirical studies reveal a homogenizing effect: while AI collaboration may elevate the average quality of individual ideas, it significantly reduces the *collective diversity* of the generated pool (Moon et al., 2025).

A.4 Approaches to Architecting Agent Heterogeneity

Critically, recent diagnostics reveal a fundamental barrier: an **"Artificial Hivemind"** where LLMs converge on identical semantic distributions regardless of prompting strategies (Jiang et al., 2025). This intrinsic alignment renders models **"creatively homogeneous,"** often masking a lack of conceptual diversity behind mere stylistic variations (Wenger and Kenett, 2025).

A.5 Creativity and Hallucination in LLMs

Grounded in Boden (2004)'s foundational definition of creativity as the capacity to generate ideas that are "novel, surprising, and valuable," contemporary research actively assesses LLMs against these criteria (Franceschelli and Musolesi, 2024). This inquiry extends to the functional role of error, with recent empirical studies investigating the intricate correlation between hallucination rates and creative output to determine if factual deviations drive novelty (Banerjee et al., 2025).

A.6 Persona Fidelity and Interaction Dynamics

While persona-based architectures like **ChatDev** (Qian et al., 2024) facilitate complex task coordination, rigorous diagnostics reveal persistent fidelity bottlenecks. Empirical studies indicate that agents frequently regress to "default" or stereotypical behaviors regardless of role prompting (Liu et al., 2024a; Tan and Lee, 2025), with recent benchmarks confirming that model scaling fails to resolve these adherence issues (Samuel et al., 2025). Furthermore, contrasting with reasoning domains where interaction-driven consensus denotes success (Du et al., 2024), rapid convergence in ideation contexts is increasingly identified as a structural failure mode that stifles divergent exploration.

A.7 Communication Topologies and Structural Optimization

Recent frameworks actively optimize interaction topologies for routing efficiency and task adaptability, utilizing dynamic routing mechanisms (Yue et al., 2025; Leong et al., 2025) or sparse connectivity (Li et al., 2024) to reduce overhead. However, dense interaction introduces systemic risks: it significantly accelerates **error propagation** (Shen et al., 2025) and drives **social polarization** in opinion dynamics

(Wang et al., 2025). In contrast, our NGT-based architecture optimizes for *independence*, enforcing structural disconnection during generation to strictly insulate diversity from these failure modes.

B Task Formulation Details

B.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.

B.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.

B.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.

B.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.

B.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.

B.6 Semantic Dispersion (PCD)

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

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

B.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.

C 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.

C.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?"*

C.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.

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

D Metric Design and Implementation Details

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

D.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.

D.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.

D.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.

D.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.

E Constructive Conflict Metric.

We combine semantic embeddings and large language model (LLM) judgment to construct 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 $\sim$ 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 (original English prompt reproduced verbatim):

```
Compare Speaker B's statement to Speaker A's context.
Speaker A: "<previous context truncated to last 400 characters>..."
Speaker B: "<current statement>"
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.

F 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 pre-specified 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.

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
Mixed	6.082	0.146	0.272	4.131	0.114	0.215	0.781	0.694	0.530	0.882
Recursive	5.567	0.154	0.285	4.141	0.127	0.239	0.638	0.522	0.426	0.754
Horizontal	4.647	0.119	0.225	3.623	0.098	0.187	0.734	0.665	0.465	0.866
Interdisciplinary	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.

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.

G 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.

G.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.

G.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.

G.3 Consistency Across Structural Metrics

We next examine consistency among three embedding-based structural metrics: Vendi score, $1 - \phi$, 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.

G.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.

G.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.

G.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.

G.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.

H 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.

H.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.

H.2 Prompt Template

The following prompt was used for the evaluation:

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.
```

H.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.

I Research Plan instead of Proposal or Paper

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) format jiejian (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 3: 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