Communication Topology Matters: Diversity in Multi-Agent LLMs

Anonymous ACL submission

Abstract

Do multi-agent LLM "discussions" improve reliability beyond strong single-agent baselines or a simple majority vote? Evidence is mixed, and most prior work fixes communication to a clique while varying prompts or sampling. We present an *empirical* study that isolates the effect of *communication topology* and *discussion depth* on task accuracy and operational measures of diversity. Using a round-synchronous simulator with fixed prompts/hyperparameters, we compare six graphs — ring, star, fully connected, Erdős–Rényi, Watts–Strogatz, and Barabási–Albert—over shallow depths ($T \leq 10$). At each round, agents read a neighbor digest, produce a message, and output a `<final>` decision aggregated by majority vote. We track accuracy and multiple diversity proxies: embedding-space dispersion (mean pairwise distance), lexical measures (type–token, unigram entropy), and n-gram entropies; when relevant, we also report cost-sensitive utility.

Across seeds and items, three patterns are robust. (1) **Sparse rings** under-mix, yielding the slowest early accuracy gains. (2) **Hubbed graphs** (star, scale-free) reach high accuracy with *shallow* depth (T=1–2) and sustain higher non-hub semantic dispersion than cliques at the same depth. (3) **Lexical and semantic diversity diverge**: vocabulary entropy rises monotonically, whereas semantic dispersion grows more modestly and can plateau on dense graphs. Small-world offers a balanced middle ground; cliques improve quickly but call for strict early stopping to avoid over-consensus.

1 Introduction

Multi-agent large language models (LLMs) promise to improve reliability by letting multiple agents exchange rationales before making a collective decision. Yet recent evidence is mixed on whether "debate/discussion" actually outperforms strong single-agent methods or simple majority vote. We identify a missing piece: the joint role of *communication topology* and *discussion depth* in shaping accuracy and operational diversity. Most prior studies fix topology (often a clique) and vary prompts or sampling; fewer examine how ring, star, random, or small-world graphs—common in distributed systems and social networks—modulate information sharing and convergence.

This paper is an empirical characterization of how topology and depth affect (i) task accuracy and (ii) multiple diversity proxies measured from agent messages. We standardize prompts, hyperparameters, and logging across conditions, and we analyze diversity at both lexical (type–token, n-gram entropy) and semantic levels (embedding-space dispersion). Our one-figure summary shows three robust patterns across seeds and items: (1) **Sparse rings** under-mix and yield the slowest early accuracy gains; (2) **Hubbed graphs** (star, scale-free) reach high accuracy with *shallow* depth (T=1–2) and maintain higher non-hub semantic dispersion than cliques at the same depth; (3) **Lexical and semantic diversity diverge**: vocabulary-level diversity rises monotonically, while semantic dispersion grows modestly and can plateau or decline with depth, especially on dense graphs. These findings explain when discussion helps beyond vote-only—and when it does not.

We release a fully reproducible analysis suite with camera-ready figures, JSONL conversation logs, and configuration YAMLs covering topology, prompts, and roles.[1] We intentionally keep this paper empirical; theoretical interpretations (e.g., connections to contraction and synergy decay) are discussed qualitatively in §5 and deferred to follow-up work.

[1] (https://anonymous.4open.science/r/mad-diversity-eval-7366/).

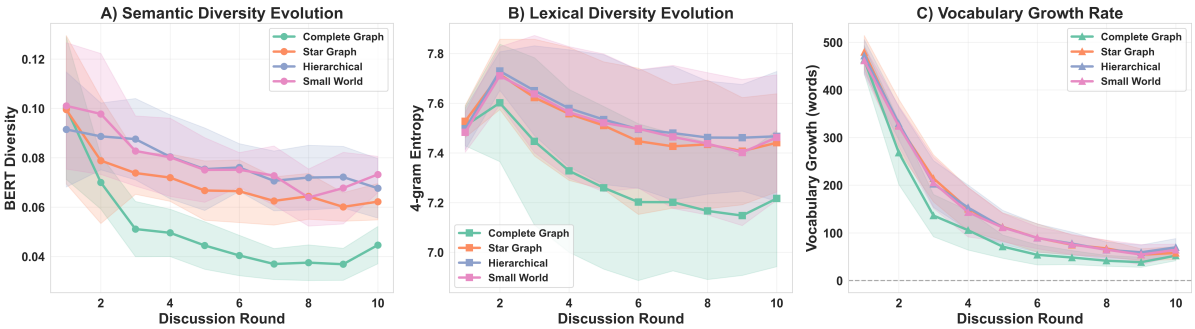

Figure 1: (A) Semantic diversity evolution measured by median pairwise cosine distance of OpenAI embeddings (text-embedding-3-small). (B) Lexical diversity evolution measured by 4-gram entropy. (C) Vocabulary growth rate showing discussion dynamics. Results shown for GSM-Plus (mathematical reasoning) with 8 agents over 10 discussion rounds.

2 Background and Related Work

Multi-agent debate for LLMs has produced mixed evidence. Several studies report homogenisation: agents with similar capabilities tend to converge to majority views as rounds progress, reducing diversity ("diversity collapse"), and sometimes agreeing on wrong answers; repeated discussion overemphasises persuasive asymmetries and drives one-sided convergence (Estornell and Liu, 2024; Wynn et al., 2025; Cau et al., 2025; Piao et al., 2025). Work on DReaMAD (Oh et al., 2025) shows cloned models reinforce biases and share reasoning styles, explaining how rounds erode diversity through self-amplification. "Degeneration-of-Thought" (Liang et al., 2024) shows increasing debaters or rounds can decrease accuracy, consistent with diversity loss. Other studies argue that heterogeneity is key: mixing models improves reasoning, while homogeneous teams collapse as rounds increase (Hegazy, 2024).

Topology matters: Liu et al. (2024) show task–graph interactions; our experiments follow this line by varying ring, star, random, small-world, and fully connected graphs. Majority-vote baselines remain strong; recently Choi et al. (2025) find most gains stem from round-1 aggregation rather than discussion.

Diversity itself has become an object of control. Chu et al. (2024) tune prompts to regulate diversity. Sociolinguistic critiques warn that LLMs drift toward mainstream homogeneity, an effect strengthened by more rounds (Sourati et al., 2025). Alignment may also compress diversity: RLHF can reduce output variety (Murthy et al., 2025; Mohammadi, 2024). Conversely, framing diversity as reward is being explored (Zhou et al., 2025; Li et al., 2025a).

Methodologically, recent literatures (Young et al., 2024; Li et al., 2025b) systematise diversity evaluation and metrics: embedding-based semantic distances, lexical/n-gram entropy, self-BLEU, Remote-Clique, NLI-based logical diversity, MAUVE, and precision/recall trade-offs; they also document dataset choices and message-level logging for multi-round experiments. We adopt these practices and metrics.

3 Diversity Measures

This section *overviews* how we quantify diversity in multi-agent LLM debate; a more complete taxonomy and estimator details are deferred to App. **??**. In our experiments we track two primary signals: (i) **lexical diversity** via token entropy, and (ii) **semantic diversity** via pairwise cosine similarity in a fixed embedding space. We compute all metrics per instance x and round t, then aggregate across instances by the median, reporting 95% bootstrap confidence intervals and the number of seeds.

Setup. Let n agents exchange messages $M_i^{(t)}$ at round t. All preprocessing (tokenizer, casing, normalization) and the sentence-embedding model $\phi(\cdot)$ are fixed across conditions to enable comparable measurements.

Lexical diversity (token entropy). For instance x and round t, form the empirical token distribution over the multiset of tokens appearing in $\{M_i^{(t)}\}_{i=1}^n$:

$$\hat{p}_{x,t}(v) \;=\; \frac{\text{count}(v \text{ in } \{M_i^{(t)}\})}{\sum_u \text{count}(u)}.$$

Define token entropy and its normalized variant

$$\mathrm{TokEnt}_{x,t} \;=\; -\sum_{v} \hat{p}_{x,t}(v) \log \hat{p}_{x,t}(v),$$

We apply small Dirichlet/Laplace smoothing and reweight by message length to reduce small-sample and verbosity effects. Token entropy captures *surface* variety (word choice, punctuation), but can increase even as agents converge semantically (e.g., paraphrase inflation). We therefore treat it as diagnostic rather than decisive.

Semantic diversity (pairwise cosine in embedding space). We embed each message with a frozen ϕ and compute a distance-like dispersion:

$$\mathrm{SemDiv}_{x,t} \;=\; \mathrm{median}_{i<j}\Big(1-\cos\big(\phi(M_i^{(t)}),\, \phi(M_j^{(t)})\big)\Big).$$

Higher SemDiv indicates that agents' utterances occupy more dispersed regions of the representation space, reflecting differences in *content* rather than phrasing. To mitigate embedding anisotropy, we whiten ϕ using statistics from a fixed training split and keep that transform fixed thereafter. This metric remains sensitive to the inductive biases of ϕ and cannot, by itself, separate complementary from contradictory diversity.

What should "diversity" capture in debate? Evaluation should reflect: (a) dispersion over task-relevant hypotheses and rationales, (b) *complementarity* (non-redundant information contributed by different agents), and (c) *temporal erosion* (homogenization across rounds). Token entropy primarily addresses (a) at the surface level; SemDiv aligns better with content dispersion, but neither metric alone captures (b) nor distinguishes helpful diversity from noise. Accordingly, we report both in tandem with accuracy and round-wise trajectories.

4 Experiments

4.1 Method

Multi-agent simulation. We conducted a topology comparison experiment to investigate how network communication structures affect multi-agent discussion diversity. Four network topologies were compared using 4 agents conducting 10-round discussions on 3 problem-solving questions.

Our experimental methodology builds upon the LLM discussion framework introduced by (Lu et al., 2024). Following their approach, we implement a multi-agent system where LLM agents iteratively exchange information to reach consensus. However, we introduce a key modification: topology-controlled information flow. Specifically, each topology constrains which agents can observe each other's responses—agents can only see outputs from connected neighbors in previous discussion rounds. This topology-based visibility constraint allows us to systematically examine how network structure influences collaborative reasoning processes.

Then we run a round-based simulation over a communication graph $G = (V, E)$. At round t, each agent $a \in V$ receives a *neighbor digest* (a fixed-length textual summary of neighbors' last messages), produces a response to the current task, and emits a `<final>...</final>` decision token used for evaluation and aggregation. Topologies: RING, STAR (one hub), FULLYCONNECTED (clique), ERDŐS–RÉNYI $G(n, p)$, WATTS–STROGATZ (small-world), and BARABÁSI–ALBERT (scale-free). We study depths $\in \{0, 1, 2, 3, 4, 5\}$; $= 0$ reduces to independent one-shot agents and majority vote. Unless stated, agents are heterogeneous (two high-end production models plus one light open-weights model) and temperature is fixed. Accuracy is computed from the aggregated decision at round ; diversity and cost are measured per round and then aggregated.

Tasks and metrics. We evaluate multiple-choice accuracy (for the classification tasks), numeric exact match (arithmetic), and label EM (single-label sentiment) when applicable. Semantic diversity is the mean pairwise distance $D_t^{\mathrm{sem}} = \frac{2}{W(W-1)} \sum_{i<j} [1 - \cos(e_i, e_j)]$ where e_i are sentence embeddings; we also show *mean cosine* (↓ with diversity) for sign cross-checking. Lexical diversity comprises type–token ratio and unigram entropy; we additionally report bigram/trigram entropy.

5 Discussion

Hubbed and dense graphs yield fast accuracy gains and comparatively steeper early semantic spread, while RING under-mixes. In follow-up work we will present a formal analysis linking contraction rates to graph spectra and discussing why specific information such as *synergistic* information may be fragile under repeated mixing.

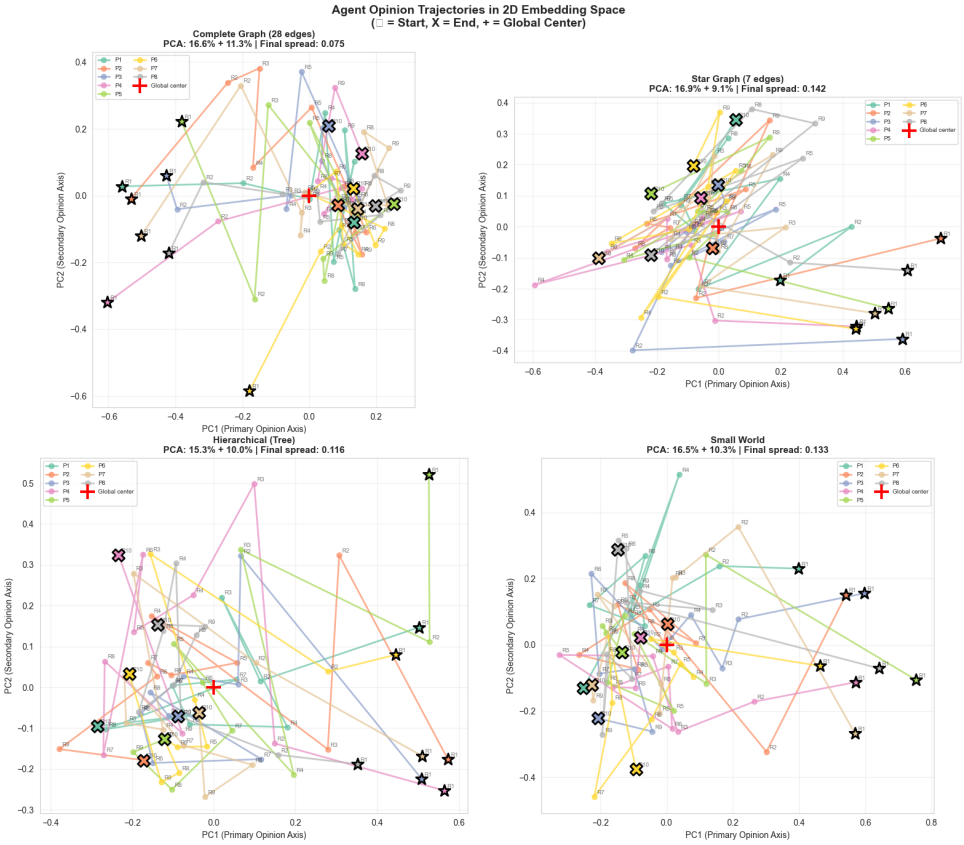

Figure 2: **Agent opinion trajectories in semantic embedding space across network topologies**. Agent responses are projected into 2D space using PCA dimensionality reduction of the embeddings. Each point represents an agent's response at a specific discussion round (rounds 1-10), with arrows showing temporal evolution. Colors distinguish different network topologies: Complete Graph (all-to-all connectivity), Star Graph (hub-and-spoke), Hierarchical (layered structure), and Small World (local clustering with long-range connections). Results shown for GSM-Plus mathematical reasoning task.

6 Conclusion

We presented an empirical study of *diversity in multi-agent LLMs* that isolates the effect of communication topology and discussion depth under a fixed prompting and evaluation recipe. Across six topologies, hubs (star, scale-free) consistently reached high accuracy in *one to two rounds* while preserving higher non-hub semantic dispersion than cliques at the same depth; rings under-mixed and lagged in early gains; cliques improved quickly but exhibited the fastest semantic convergence. Lexical diversity rose across the board, whereas semantic dispersion rose modestly and could plateau or decrease with depth—showing that lexical proxies overestimate discussion diversity. Together, these observations yield a simple design rule: use hub/small-world structures at shallow depth when accuracy and semantic variety must be balanced, and instrument runs with semantic metrics rather than lexical counts alone.

Limitations

This paper is empirical; theoretical interpretations are discussed qualitatively in §5 and formal analysis is deferred. Our diversity measures are proxies—embedding-space dispersion (via TEXT-EMBEDDING-3-SMALL), type–token counts, and n-gram entropies—and may disagree; they do not establish causal semantics. The main figures focus on multiple-choice QA with fixed prompts/personas to isolate topology, which narrows external validity and leaves generation-heavy tasks (long-form reasoning, tool use, code) for future work. Communication graphs are synthetic (ring, star, clique, Erdős–Rényi, Watts–Strogatz, Barabási–Albert) and omit asynchrony, memory/tool calls, delays, and evolving links, so transfer to production settings is uncertain. We do not study safety or misinformation; topology may affect propagation, requiring guardrails beyond this scope.

References

Erica Cau, Valentina Pansanella, Dino Pedreschi, and Giulio Rossetti. 2025. Selective agreement, not sycophancy: investigating opinion dynamics in LLM interactions. *EPJ Data Sci.*, 14(1):59.

Hyeong Kyu Choi, Xiaojin Zhu, and Yixuan Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? *CoRR*, abs/2508.17536.

Kuanchao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024. Exploring and controlling diversity in llm-agent conversation. *CoRR*, abs/2412.21102.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Mahmood Hegazy. 2024. Diversity of thought elicits stronger reasoning capabilities in multi-agent debate frameworks. *CoRR*, abs/2410.12853.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. 2025a. Jointly reinforcing diversity and quality in language model generations. *Preprint*, arXiv:2509.02534.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2025b. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2024. A dynamic LLM-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

Behnam Mohammadi. 2024. Creativity has left the chat: The price of debiasing language models. *CoRR*, abs/2406.05587.

Sonia K. Murthy, Tomer D. Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 11241–11258. Association for Computational Linguistics.

Jihwan Oh, Minchan Jeong, Jongwoo Ko, and Se-Young Yun. 2025. Understanding bias reinforcement in llm agents debate. *Preprint*, arXiv:2503.16814.

Jinghua Piao, Zhihong Lu, Chen Gao, Fengli Xu, Fernando P. Santos, Yong Li, and James Evans. 2025. Emergence of human-like polarization among large language model agents. *CoRR*, abs/2501.05171.

Zhivar Sourati, Alireza S. Ziabari, and Morteza Dehghani. 2025. The homogenizing effect of large language models on human expression and thought. *CoRR*, abs/2508.01491.

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

Halley Young, Yimeng Zeng, Jacob R. Gardner, and Osbert Bastani. 2024. Improving structural diversity of blackbox llms via chain-of-specification prompting. *CoRR*, abs/2408.06186.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: Majority drives selection, novelty promotes variation. *Preprint*, arXiv:2509.15194.

A Overview of Diversity Metrics

In recent years, large language models (LLMs) have achieved remarkable results across a wide range of tasks, including text generation, machine translation, summarization, and code generation. Their rate of progress is extremely rapid, and they are permeating numerous application domains and user groups. As such large-scale models become established as social infrastructure, *diversity* has drawn attention as a key concept that supports a model's usefulness, trustworthiness, and fairness.

Diversity refers to generated text being rich in variation, not overly dependent on specific patterns or biases, and exhibiting lexical, semantic, cultural, and application-level breadth. Fundamentally, it is a concept for assessing the distribution within

which a model produces outputs, or how richly it can present expressions and ideas across different tasks and contexts. For this reason, diversity is pivotal for user experience and creative applications. At the same time, diversity often entails trade-offs with model performance (accuracy and consistency) and efficiency (computational cost), which makes securing and tuning it non-trivial.

Moreover, if the data distribution used to train a model is biased, the model may produce outputs skewed toward particular cultural or social perspectives. Datasets that insufficiently reflect minority languages or region-specific knowledge, or preferences and biases introduced during human feedback (e.g., RLHF), can degrade diversity. A lack of diversity can escalate into issues of fairness and ethics, potentially leading to distrust of the model or refusal to use it.

From these perspectives, how to define diversity in LLMs and from which aspects to measure and evaluate it are critical research questions. However, diversity is hard to quantify with a single metric; multiple linguistic, semantic, cultural, social, and application dimensions are intricately intertwined. For instance, lexical metrics such as simple n-gram counts and Self-BLEU are easy to compute but do not sufficiently capture true semantic diversity or the variety of latent ideas. By contrast, methods that leverage embedding representations (e.g., BERT or Sentence Transformers)—including cosine similarity, dispersion on a minimum spanning tree (MST), entropy, and DCScore—enable evaluation that probes more semantic dimensions. This survey provides foundational guidance for choosing evaluation methods for diversity assessment and for deciding whether to combine multiple methods.

B Evaluation of Diversity

B.1 Using Embedding Representations

The greatest advantage of evaluating diversity with embeddings is the ability to capture *semantic* diversity. Because embedding representations express textual meaning as high-dimensional vectors, they allow similarity to be assessed beyond word order and surface-form differences. Thus, paraphrases with the same meaning—for example, "I like apples" and "I enjoy apples"—are handled appropriately. In addition, because the embedding space preserves high-dimensional information, analyzing clustering and the spread of the distribution makes it possible to quantitatively measure how diverse the generated results are.

B.1.1 Cosine Similarity

BERT-embedding diversity evaluates semantic diversity by measuring similarity between texts using the cosine *distance* (or similarity) of their BERT embeddings. A larger distance indicates richer semantic variation. In practice, one may threshold cosine *similarity*: if the similarity exceeds a threshold, two ideas are considered similar (i.e., overlapping), which can be used to judge diversity.

$$\text{Cosine Similarity} = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Here, $\mathbf{u}$ and $\mathbf{v}$ denote two embedding vectors; the numerator is their inner product and the denominator the product of their norms. This metric measures similarity based on the angle between vectors; the closer it is to 1, the higher the similarity. When using BERT embeddings to assess semantic similarity between generated texts, texts with high cosine similarity are judged semantically redundant and may be undesirable in diversity evaluation.

B.1.2 Pairwise Distance

The *mean* pairwise distance computes, for each text item (prompt or idea), the average distance to all other items. A larger mean pairwise distance indicates that the item differs more from others, implying higher diversity.

$$\text{Mean Pairwise Distance (MPD)} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} d(x_i, x_j)$$

Here, $d(x_i, x_j)$ denotes the distance between points x_i and x_j. This averages distances over all item pairs in the dataset. A large mean pairwise distance implies that texts have mutually distinct characteristics and are therefore rated as highly diverse.

The *minimum* pairwise distance considers, for each item, the distance to its nearest neighbor. A short distance means the item is concentrated with similar items; a long distance suggests less redundancy and more distinctive features. In a complete-graph (clique) view, one can compute the mean of pairwise distances over all items as a diversity measure. If clustering is imbalanced, this metric may not measure diversity accurately, especially when ideas are highly concentrated.

$$\text{Average Pairwise Distance (APD)} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} d(x_i, x_j)$$

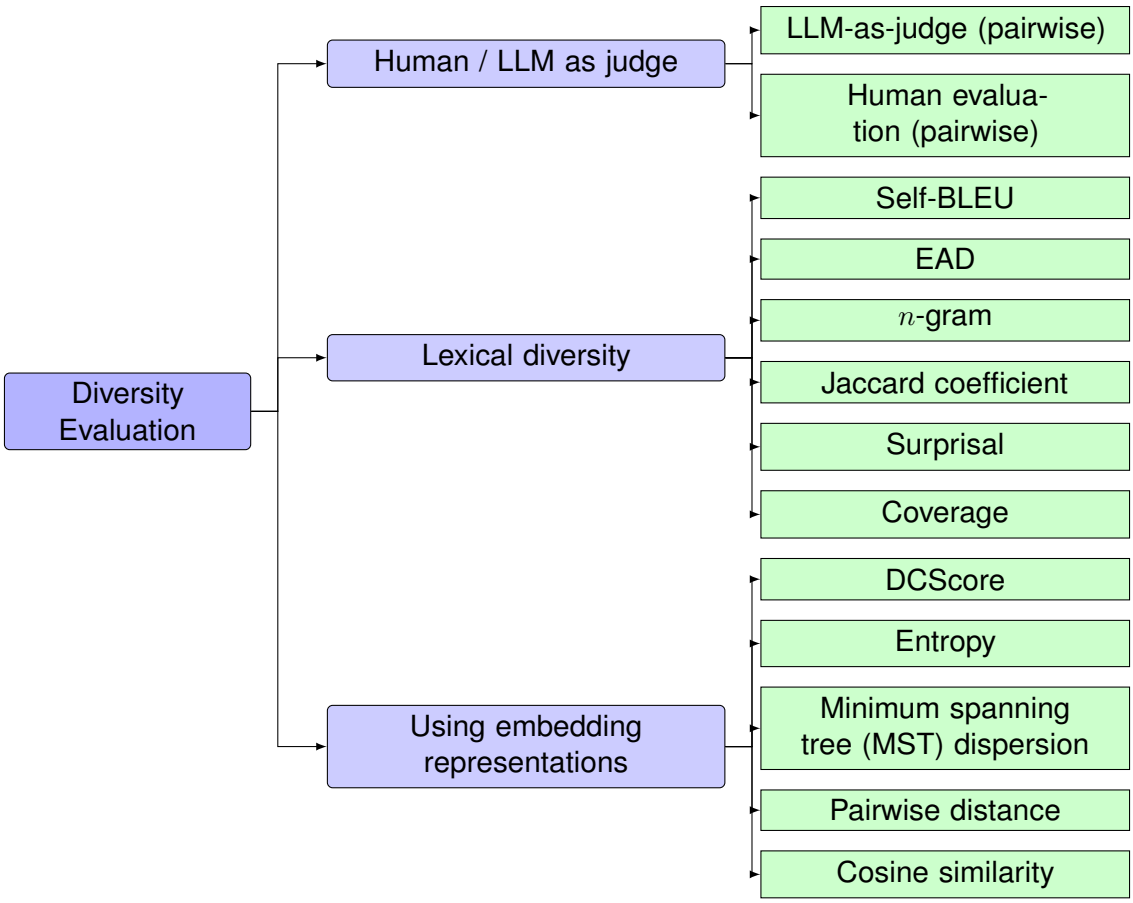

Figure 3: Diversity evaluation

The *Chamfer distance* is the average of minimum pairwise distances, computing the distance to the nearest neighbor for each item. Although it can be biased toward clustered points, it is used to measure proximity to the nearest points.

$$\text{Chamfer Distance (CD)} = \frac{1}{N}\sum_{i=1}^{N}\min_{j\neq i} d(x_i, x_j)$$

B.1.3 Minimum Spanning Tree (MST) Dispersion

Compute the average edge length within the minimum spanning tree (MST) to measure the overall diversity of items. Because this is computed via the MST, the sum of edge weights serves as an indicator of diversity, indicating how dispersed the items are.

$$\text{MST Dispersion} = \frac{1}{|E_{\text{MST}}|}\sum_{(x_i,x_j)\in E_{\text{MST}}} d(x_i, x_j)$$

B.1.4 Entropy

This measures the uniformity of a distribution over a grid: the more uniform the distribution, the higher the entropy. Because evenness corresponds to higher entropy, this is useful for evaluating diversity.

$$\text{Shannon-Wiener Index (Entropy)} = -\sum_{b} f_b \log(f_b)$$

B.1.5 DCScore

DCScore treats each data sample as an independent classification task and evaluates diversity while accounting for relationships among samples. For each sample, a representation vector is produced and pairwise similarities are computed to form a similarity matrix using a kernel function (e.g., inner product or cosine similarity). Based on this similarity matrix, a classification-probability matrix is derived; the trace (sum of diagonal elements) of this matrix is taken as the DCScore. In this way, diversity is measured from classification-oriented dissimilarities among samples.

B.2 Lexical Diversity

When evaluating diversity without embeddings—e.g., using n-grams or Self-BLEU—there are several advantages. First, these methods are very simple and inexpensive computationally, enabling efficient evaluation of large volumes of generations; they are widely used in generation tasks. They are also useful when one wants to measure *lexical* (rather than semantic) diversity. For instance, in settings like advertising where character counts are strictly constrained, it is important to quantify the range of word choices and variations. In such scenarios, n-gram metrics and Self-BLEU effectively capture lexical diversity and quickly detect simple duplication or skew in the outputs.

B.2.1 Coverage

Coverage is an indicator of the spread of the feature space; it measures how many different combinations of features are generated. Higher coverage indicates that more distinct structures are included and that the outputs are more diverse.

$$C_n(S) = \frac{\sum_{c\in C_n}\log(|\{s\in S : c\subseteq s\}|+1)}{\log(|S|+1)}$$

Here, C_n denotes the set of all combinations of n features. This metric measures how well the generated texts cover rare structural features.

B.2.2 Surprisal

Surprisal measures how many hard-to-predict combinations of features are generated. Based on the occurrence probability of each feature combination, the score increases when many rare combinations are included, thereby evaluating the generation of uncommon structures.

$$WS_n(S) = -\frac{\sum_{c\in C_n}\frac{|\{s\in S:c\subseteq s\}|}{|S|}\log\left(\frac{|\{s\in S:c\subseteq s\}|}{|S|}\right)}{|S|}$$

This metric quantifies how hard to predict (i.e., surprising) rare feature combinations are.

B.2.3 Jaccard Coefficient

Jaccard diversity measures the degree of *mismatch* among features in the generated texts; sets of texts that include many dissimilar features are rated higher. It emphasizes differences among feature combinations and rewards generations that include diverse feature compositions.

$$DBJ,n(S) = \frac{1}{|S|^2}\sum_{\substack{c_i,c_j\in C_n\\ i\neq j}} |c_i|\cdot|c_j| \cdot\left(1-\frac{|\{s\in S : c_i\subseteq s \wedge c_j\subseteq s\}|}{|\{s\in S : c_i\subseteq s\}|+|\{s\in S : c_j\subseteq s\}|}\right)$$

This metric measures dissimilarity between pairs of features.

B.2.4 n-gram

n-gram diversity measures how many distinct n-grams (e.g., contiguous sets of four words) appear in the text. It serves as a metric for evaluating lexical and phrasing diversity.

$$D = -\sum_{i=1}^{M} p_i \log p_i$$

Here, p_i is the occurrence probability of the i-th n-gram among all n-grams, and M is the total number of unique n-grams across the corpus. This metric measures the unpredictability of the text.

B.2.5 EAD (Expected Average Diversity)

Expected Average Diversity (EAD) is a new metric for more comprehensively evaluating diversity in a model's outputs. It integrates the entropy of the overall probability distribution of the generated text with n-gram diversity:

$$\text{EAD} = \alpha D + \beta \cdot H(p)$$

Here, D denotes n-gram diversity, and $H(p)$ is the entropy of the token probability distribution for the generated texts. The coefficients α and β weight the importance of each component. Because EAD expresses diversity on a single scale, it is useful for comprehensively evaluating model output diversity that cannot be fully captured by n-gram diversity or entropy alone. Using this metric can clarify improvement targets when diversity is lacking and can help efficiently enhance the quality of generated text.

B.2.6 Self-BLEU

Prepare multiple generations from the model. Treat one generation as the *hypothesis* and the remaining generations as the *references*, and compute the BLEU score. Repeat this for each generated sentence and take the average of the scores.

B.3 Human Evaluation / LLM as Judge

Human evaluation and LLM-as-judge can also be applied to diversity assessment.

C Additional Experiments

D Other Dataset Experiment

E The Use of Large Language Models (LLMs)

The LLMs are used for assisting

- Manuscript structuring, editing, and proofreading
- Software engineering
- LaTeX engineering

.

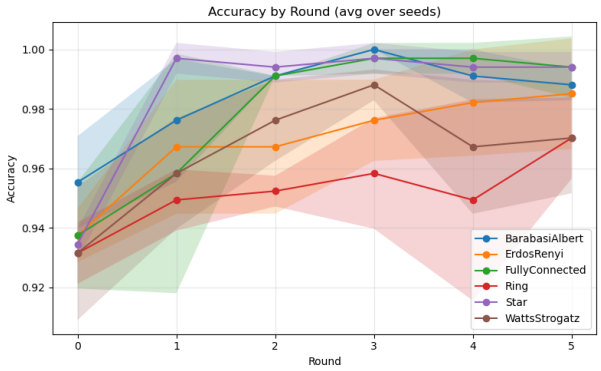

Figure 4: Accuracy vs. Round

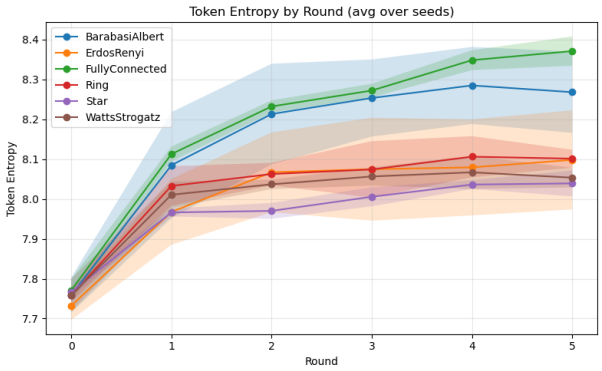

Figure 5: Token Entropy vs. Round

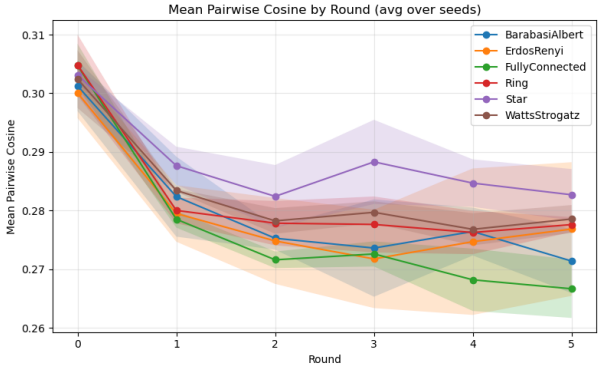

Figure 6: Mean Pairwise Cosine vs. Round

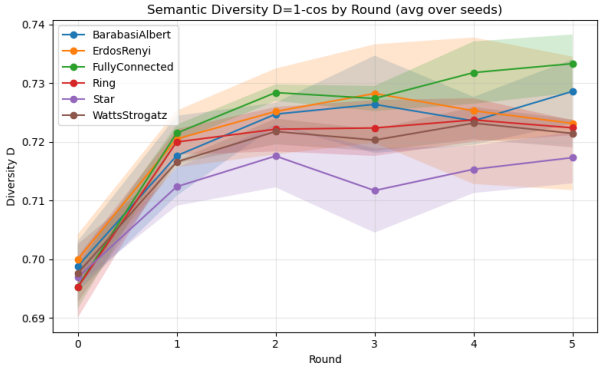

Figure 7: Semantic Diversity vs. Round

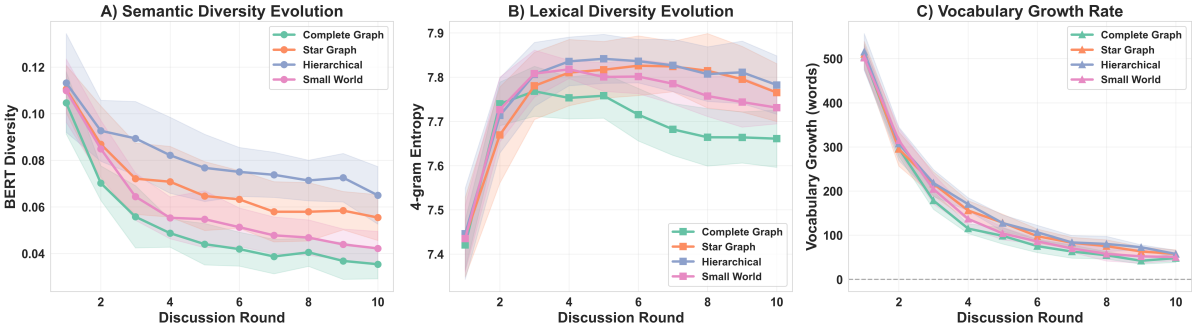

Figure 8: (A) Semantic diversity evolution measured by median pairwise cosine distance of OpenAI embeddings (text-embedding-3-small). (B) Lexical diversity evolution measured by 4-gram entropy. (C) Vocabulary growth rate showing discussion dynamics. Results shown for GoEmotions(Demszky et al., 2020) with 8 agents over 10 discussion rounds.