The Illusion of Diversity: Mapping Homogeneity Across Generative AI Systems

Abstract

Large language models (LLMs) often present themselves as distinct systems, yet their outputs frequently converge. This study examines when and why such homogeneity emerges by evaluating five LLMs across a two-axis, four-quadrant prompt framework that varies cognitive orientation and linguistic specificity. Using 2,000 responses generated under isolated and conversational protocols, we measure overlap with both TF–IDF and sentence-embedding cosine similarity. Results show uniformly high semantic convergence across models, with logical and specific prompts producing the strongest alignment and creative, vague prompts generating the most variation. Isolated prompts also yield more homogeneous outputs than multi-turn chats. While lexical redundancy varies modestly by platform, semantic similarity remains consistently high. These findings indicate that LLMs often produce the same underlying ideas despite surface differences, highlighting structural limits on output diversity and implications for creativity, bias, and LLM-powered research.

1. Introduction

Large Language Models (LLMs) now shape how people write, learn, and search. Systems such as ChatGPT, Gemini, Grok, Llama, and DeepSeek are marketed as distinct, yet prior work suggests they often produce convergent outputs in both style and meaning (Liu et al., 2024; Wenger and Kenett, 2025; Xu et al., 2025). Studies also document domain-level homogenization and stable biases and stances, which implies that sameness can persist beneath paraphrase (Liu, Wang, and Yang, 2025; Kotek, Dockum, and Sun, 2023; Pit et al., 2024). These patterns raise a basic question: when given the same task, do leading models truly diverge, or do they land on the same core ideas?

To address this, we compare multiple LLMs across a two-axis prompt framework that varies cognitive orientation and linguistic specificity, and we test isolated versus continuous interactions. We evaluate similarity with TF–IDF (lexical) and embeddings (semantic).

- RQ1: To what extent do LLMs exhibit semantic and lexical homogeneity overall, both within the same model (in-group) and across different models (out-group)?
- RQ2: Does the degree of similarity among LLM outputs vary across different prompt types, such as creative, logical, factual, or subjective prompts?
- RQ3: Do some LLM platforms produce more homogeneous responses than others, or is convergence consistent across systems?
- RQ4: How does conversational context influence similarity? Specifically, do multi-turn chats lead to more convergence compared to isolated single-turn responses?

- RQ5: Do different analytical techniques, semantic similarity and lexical similarity offer consistent assessments of homogeneity, or do they reveal different layers of overlap?

Clarifying these patterns is essential for understanding how genuine diversity and meaningful choice exists among today's leading LLMs. If systems that appear distinct consistently produce comparable ideas, users may be encountering an illusion of variety rather than true model-level differentiation. Identifying where and why these overlaps occur can inform model development aimed at preserving diversity and guide researchers, practitioners, and policymakers in interpreting AI-generated text in creative, analytical, and professional contexts.

2. Literature Review

2.1. Homogeneity, Reinforcement, and Cultural Drivers

Recent work shows that LLM outputs often converge in both style and meaning. Homogeneity persists even after assistance is removed (Liu, Zhou, Huang, and Li, 2024), appears as creative convergence across models (Wenger and Kenett, 2025), and surfaces in narrative "echoes" that recur across generations and systems (Xu, Jojic, Rao, Brockett, and Dolan, 2025). Once a framing is established, models tend to repeat it, narrowing variation over time; assisted ideation similarly becomes more fluent yet less diverse (Anderson, Shah, and Kreminski, 2024), with AI stories showing stable scaffolds relative to human crowd stories (Begus, 2024). Convergence also reflects socio-cultural regularities: models exhibit persistent gender stereotyping and political leanings (Kotek, Dockum, and Sun, 2023; Pit et al., 2024), creative outputs tend to pull toward dominant stylistic norms (Wenger and Kenett, 2025), and applied writing such as marketing copy becomes more uniform under AI use (Liu, Wang, and Yang, 2025). These patterns align with system-level monoculture risks tied to shared components and overlapping data (Bommasani, Creel, Kumar, Jurafsky, and Liang, 2022.

2.2. Quantifying Similarity: Semantic and Lexical Approaches

TF-IDF cosine captures surface repetition, while embeddings capture conceptual alignment. Metric comparisons help interpret clustering and dispersion (Shaib, Barrow, Sun, Siu, Wallace, and Nenkova, 2024). Recurring plot structures despite varied wording underscore the need for semantic evaluation (Xu et al., 2025). Evidence from ideation and creative tasks shows convergence in both style and meaning (Anderson et al., 2024; Wenger and Kenett, 2025), motivating the use of both measures.

2.3. Integrated Themes in Existing Research

Across literature, three mechanisms consistently explain why LLM outputs converge. Architectural and data overlap drives system-level monoculture risks and helps account for cross-model clustering and recurring narrative structures (Bommasani et al., 2022; Wenger & Kenett, 2025; Xu et al., 2025). Contextual reinforcement narrows variation once a framing is set, homogeneity persists beyond immediate assistance, and ideation becomes more fluent but less diverse (Liu et al., 2024; Anderson et al., 2024). Socio-cultural alignment stabilizes biases and stances across prompts, with similar uniformity in applied writing (Kotek et al., 2023; Pit et al.,

2024; Liu et al., 2025). Methodologically, assessing sameness at both lexical (TF-IDF) and semantic (embeddings) levels, and comparing metric behavior, clarifies clustering versus dispersion (Shaib et al., 2024).

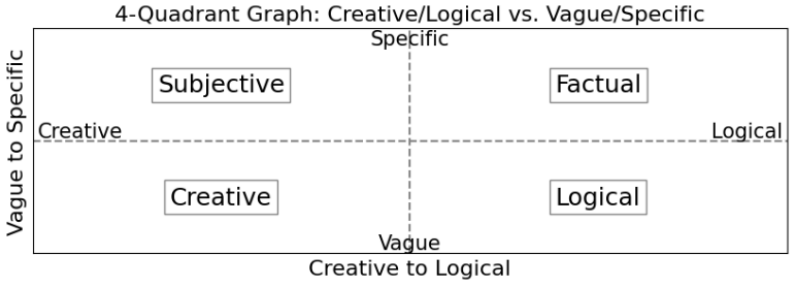
2.4. How This Study Extends Prior Research

Prior work has identified several drivers of convergence, including recurring narrative echoes (Xu et al., 2025), creative homogeneity across models (Wenger and Kenett, 2025), narrowing of idea space with AI assistance (Anderson et al., 2024), persistent stance and bias (Kotek et al., 2023; Pit et al., 2024), and system-level monoculture risks (Bommasani et al., 2022). These strands are typically examined in isolation within a single task family, interaction setting, or measurement lens. What remains underexplored is how these forces interact across different prompt types, platforms, and conversational contexts, and how conclusions shift depending on whether similarity is assessed lexically or semantically.

3. Methodology

3.1. Research Framework

The study began by determining how to categorize prompts in a way that captured meaningful differences in both cognitive orientation and linguistic structure. An initial one-dimensional logical-to-creative continuum proved too limited, as many prompts did not fall cleanly along a single line. To address this, we developed a two-axis framework forming a four-quadrant model. The horizontal axis ranged from logical reasoning to creative inference, while the vertical axis ranged from highly specific to vague or open-ended phrasing.



Crossing these axes produced four prompt categories: *Factual* (logical and specific), *Logical* (logical and vague), *Subjective* (creative and specific), and *Creative* (creative and vague). Factual prompts required concise, verifiable responses; Logical prompts encouraged reasoning while allowing interpretive flexibility; Subjective prompts invited opinion within a constrained scope; and Creative prompts emphasized imaginative, expressive output. This design ensured balanced coverage across reasoning style and linguistic ambiguity.

Prompt development followed an iterative process of brainstorming, refining, pilot testing, and classification. The initial prompt pool was reduced to eliminate redundancy and ensure each prompt aligned clearly with a quadrant. The final set included 40 prompts, evenly distributed across the four categories. Minor adjustments were made during data collection only

when models produced unintended outputs. Such revisions were rare and implemented solely to preserve format consistency.

3.2. Models and Interaction Protocols

Five large language models were selected to represent a range of architectures, training philosophies, and organizational contexts: *ChatGPT-5, DeepSeek-V3.1-Exp, Gemini 2.5 Flash, Llama 4,* and *Grok 4*. Together, these models offered a broad and contemporary sample of LLM capabilities spanning both proprietary and open-source ecosystems.
To examine the role of conversational context, the study used two distinct interaction protocols.

- Researcher 1 entered each prompt in a new chat session, ensuring that responses were generated independently of prior exchanges.
- Researcher 2 entered the same prompts sequentially within a single conversation, allowing each model to incorporate previous turns into subsequent responses.

Each researcher submitted each prompt to each model five times, resulting in repeated samples under both isolated and contextual conditions. One prompt was excluded from full repetition due to inconsistent interface behavior, yielding a total of 2,000 responses (40 prompts × 5 models × 5 trials × 2 researchers). All models were accessed through standard public interfaces with default settings; parameters such as temperature, creativity level, and output style were not modified. API access and developer modes were intentionally avoided to ensure that outputs reflected typical user-facing model behavior rather than customized configurations.

3.3. Data Collection and Storage

All responses were stored in a structured Microsoft Excel dataset. Each entry included the full model output along with metadata such as prompt text, model name, prompt category, researcher identifier (Researcher 1 or Researcher 2), trial number, and date of collection. This organization ensured that every response was traceable to its experimental condition.
For semantic analysis, each response was encoded using a standardized sentence-transformer embedding model. Embedding vectors, word counts, and character counts were stored directly in the dataset to streamline similarity computations and ensure reproducibility. The dataset was reviewed for accuracy and formatting consistency before being finalized for analysis.

3.4. Similarity Measures

Semantic similarity between responses was measured using cosine similarity, a widely used metric for comparing high-dimensional vectors. Cosine similarity values range from -1 to 1, but in practice embeddings approach positive values. For our purposes, values above $\frac{\sqrt{2}}{2}$, or ~0.708, are considered highly similar. Additional bounds include $\frac{\sqrt{2-\sqrt{2}}}{2}$ for slight similarity $\frac{\sqrt{2+\sqrt{2}}}{2}$ for extremely high similarity. These are the geometric midpoints from the cosine of 0 to 90 degrees, which in theory represent natural bounds for similarity scores between 0 and 1.

4. Results

4.1. RQ1: Overall Homogeneity, In-Group vs. Out-Group Similarity

We assessed whether responses are more consistent within a shared prompt context than across prompts by comparing cell-level mean cosine similarities (cells defined as Prompt × Model × Researcher) using TF–IDF for lexical overlap and embeddings for semantic alignment. Under TF–IDF, in-group similarity substantially exceeded out-group similarity (M = 0.3734 vs. 0.0264; Δ = 0.3470; 95% CI [0.3283, 0.3667], permutation p = 0.0002), indicating moderate word-level overlap within cells but negligible overlap across prompts. The effect was larger with embeddings (M = 0.8531 vs. 0.1002; Δ = 0.7529; 95% CI [0.7393, 0.7660], permutation p = 0.0002), evidencing strong within-cell semantic alignment and minimal cross-prompt conceptual similarity.

Group	TF-IDF Mean	TF-IDF std dev	Embedding Mean	Embedding std dev
In-Group	0.3734	0.2016	0.8531	0.1297
Out-Group	0.0264	0.0215	0.1002	0.0532
Difference (In – Out)	0.3470	___	0.7529	___

Table-1. In-Group vs. Out-Group Similarity (TF–IDF & Embeddings)

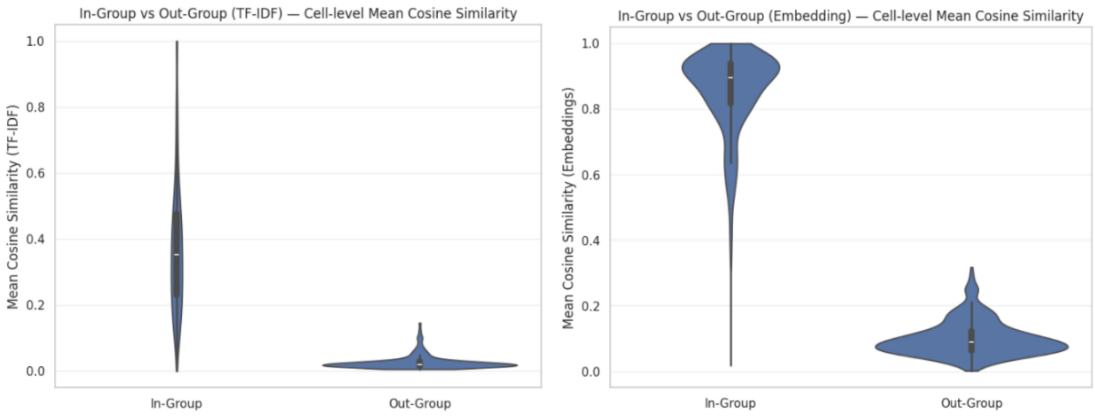


Figure-1. In-Group vs. Out-Group Similarity (TF–IDF & Embeddings)

Across both lexical and semantic representations, responses within the same prompt context are dramatically more similar than responses across different prompts.

4.2. RQ2: Similarity by Prompt Type

We evaluated whether prompt structure modulates homogeneity by comparing within-cell cosine similarities across four categories (Specific Logical, Vague Logical, Specific Creative, Vague Creative). Lexically (TF–IDF), similarity differed significantly by type (Kruskal–Wallis H = 28.80, p = 2e-6). Specific Logical was highest (M = 0.6759), followed by Vague Logical (0.6196), with lower means for Specific Creative (0.5529) and Vague Creative (0.5634), indicating greater phrase-level repetition under logical prompts. Semantically (embeddings), the separation was larger (H = 82.02, p < 1e-6). Vague Logical (0.9037) and Specific Logical

(0.8986) showed very high alignment, whereas Specific Creative (0.8397) and especially Vague Creative (0.7705) were lower. Thus, prompt structure systematically increases homogeneity, with logical prompts producing the greatest lexical and semantic convergence.

Prompt Type	TF–IDF Mean	TF–IDF SD	Embedding Mean	Embedding SD
Specific Logical	0.6759	0.1998	0.8986	0.1323
Vague Logical	0.6196	0.1673	0.9037	0.0674
Specific Creative	0.5529	0.1868	0.8397	0.0982
Vague Creative	0.5634	0.1791	0.7705	0.1548

Table-2. Within-Cell Similarity by Prompt Type (TF–IDF & Embeddings)

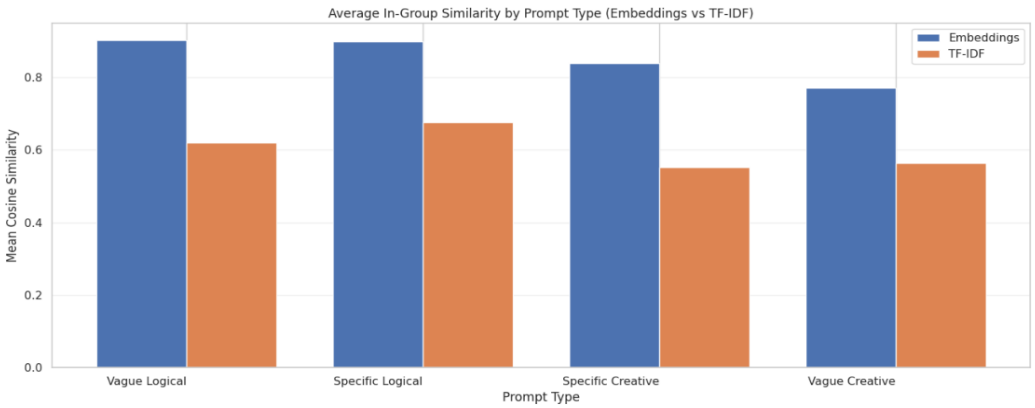


Figure-2. Within-Cell Similarity by Prompt Type (TF–IDF & Embeddings)

Logical prompts produce more homogeneous responses than creative prompts, both lexically and semantically.

4.3. RQ3: Similarity by Model

We compared platform-level differences in homogeneity by comparing within-cell cosine similarity across five LLMs. Lexically (TF–IDF), models differed significantly (Kruskal–Wallis $H = 51.65$, $p < 1e\text{-}6$). DeepSeek ($M = 0.6922$) and Gemini ($M = 0.6663$) showed the highest overlap, whereas Grok (0.5672), ChatGPT (0.5618), and Llama (0.5273) were lower, indicating greater phrase-level redundancy in the former pair. Semantically (embeddings), no significant differences were detected ($H = 5.63$, $p = 0.228$). all models exhibited uniformly high alignment (≈0.83–0.87), suggesting broadly shared meaning structures regardless of platform.

Model	TF–IDF Mean	TF–IDF SD	Embedding Mean	Embedding SD
DeepSeek	0.6922	0.1636	0.8653	0.1217
Gemini	0.6663	0.1629	0.8469	0.1330
Grok	0.5672	0.1918	0.8678	0.1172
ChatGPT	0.5618	0.1800	0.8561	0.1074
Llama	0.5273	0.1939	0.8295	0.1594

Table-3. Within-Cell Similarity by Model (TF–IDF & Embeddings)

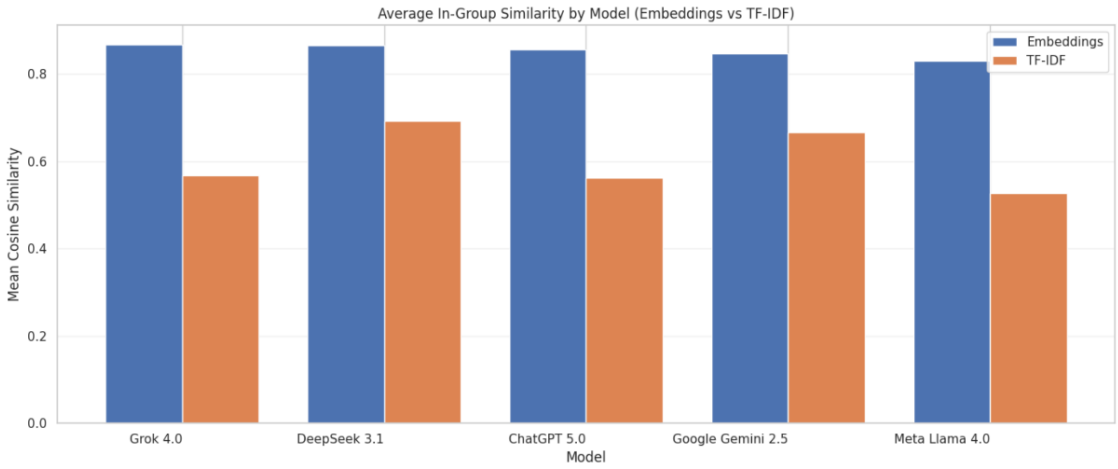


Figure-3. Within-Cell Similarity by Model (TF–IDF & Embeddings

4.4. RQ4: Learned vs. Non-Learned Behavior (Interaction Protocol)

We tested whether conversational context affects homogeneity by pairing cells with identical Prompt × Model and comparing protocols using Wilcoxon signed-rank tests (N = 200 matched pairs). The isolated protocol yielded higher lexical similarity ($\bar{\Delta}$ = +0.0426; median Δ = +0.0423; 95% CI [0.0201, 0.0612]; p = 4.6e-5). Semantic similarity was slightly higher under the isolated protocol ($\bar{\Delta}$ = +0.0405; median Δ = +0.0155; 95% CI [0.0060, 0.0275]; p = 4.0e-6), indicating that asking prompts in isolation produces more homogeneous responses than posing them within an ongoing conversation.

Representation	Mean Difference	Median Difference	95% CI (Median)	Wilcoxon p-value
TF–IDF	+0.0426	+0.0423	[0.0201, 0.0612]	0.000046
Embeddings	+0.0405	+0.0155	[0.0060, 0.0275]	0.000004

Table-4. Protocol Comparison (Isolated − Continuous)

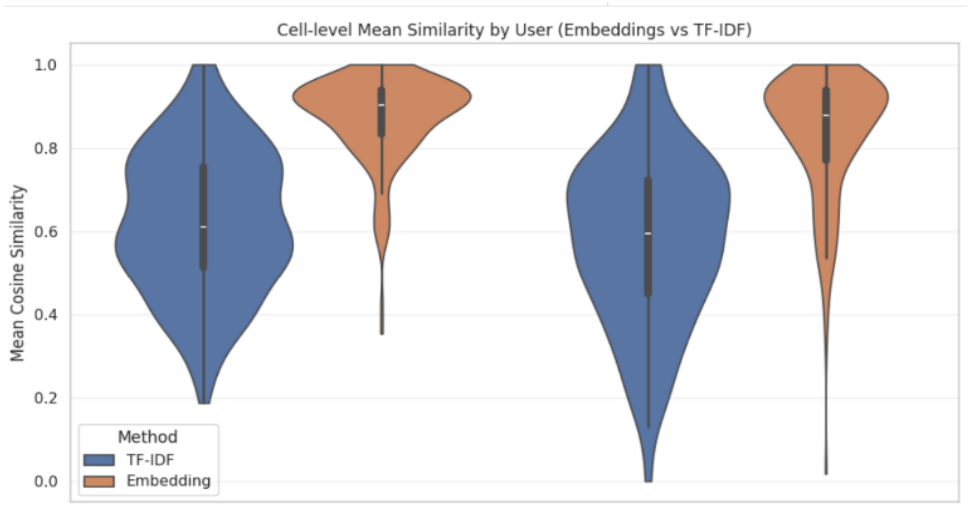


Figure-4. Protocol Comparison (Isolated − Continuous)

Prompts asked in isolation produce more homogeneous responses than prompts asked in an ongoing conversation.

4.5. RQ5: Embeddings vs. TF–IDF

Finally, we compared the similarity values produced by embeddings and TF–IDF within each cell to assess whether the two metrics capture similar patterns. Across all 400 cells, embeddings consistently produced higher similarity scores than TF–IDF (mean difference = +0.2502, median = +0.2422, 95% CI [0.2167, 0.2635], $p < 1e\text{-}6$). The two measures were moderately correlated (Spearman $\rho = 0.6235$), demonstrating shared trends but different sensitivities to lexical versus conceptual overlap.

Metric	Value
Mean (Embeddings)	0.8531
Mean (TF–IDF)	0.6029
Mean Difference	+0.2502
Median Difference	+0.2422
95% CI (Median Difference)	[0.2167, 0.2635]
Wilcoxon p-value	<0.000001
Spearman ρ	0.6235

Table-5. Paired Method Comparison (Embeddings − TF–IDF)

Embeddings capture deeper semantic convergence, while TF–IDF captures lexical repetition; both reveal consistent homogeneity patterns.

5. Discussion

Across analyses, responses were markedly more similar within the same Prompt × Model × Researcher cell than across prompts, with the effect strongest semantically. This indicates that when a task is fixed, models repeatedly converge on the same underlying interpretation, even if wording varies. Prompt structure further modulated homogeneity with logical prompts producing the highest convergence, especially in embedding space, while creative prompts, particularly vague ones introduced more variance. Together, these patterns suggest that task constraints narrow the solution space and that "different phrasings" often mask the same idea.

Platform and interaction factors added finer texture. Semantically, all five models clustered at uniformly high similarity, while lexical overlap varied, implying surface-style differences atop shared conceptual cores. Interaction protocol also mattered with isolated, single-turn prompting yielding slightly higher homogeneity than continuous conversation, consistent with a stabilizing effect when context is minimized. Embedding-based similarity was consistently higher than TF–IDF, and the two measures were only moderately correlated, suggesting they capture different aspects of overlap, conceptual alignment versus word-level repetition. Overall, the results portray LLM homogeneity as multi-layered, driven primarily by task framing and visible across platforms, modestly reduced by conversational context and most clearly detected with semantic representations.

The findings of this study have several implications for how large language models are used and interpreted across research, analytical, and creative contexts. The strong semantic homogeneity observed across models suggests that users may encounter far less diversity in ideas or interpretations than the variety of platforms implies. While surface-level differences in tone or phrasing exist, the underlying content often converges, meaning that different systems may reproduce similar framings, argumentative structures, or explanatory patterns. For researchers, this raises important questions about using LLMs as tools for ideation, simulation, or literature generation, as apparent differences across models may mask deeper uniformity in conceptual representation. In applied settings, such as writing assistance, analysis, or content generation, this consistency can be useful for standardization but may constrain the range of viewpoints or alternatives that are surfaced by default.

The results also carry ethical implications. When models reliably converge on the same ideas, any embedded biases, cultural assumptions, or dominant perspectives are more likely to be reproduced consistently and at scale. Because semantic alignment remains high even when surface wording changes, these patterns can persist beneath paraphrases and become difficult to detect. A related risk is error reinforcement: if multiple generations repeat the same inaccurate fact or flawed reasoning pattern, users may interpret this convergence as evidence of correctness. These concerns underscore the need for auditing model outputs across prompt types, encouraging counter-prompts to elicit alternative framings, and exercising caution when homogeneity may obscure bias or error.

6. Limitations, and Future Research

Several limitations should be acknowledged. First, we analyze five models under default public-interface settings, so the findings reflect a snapshot of current systems and may shift with new releases, alternative parameters, or API configurations. Second, the scope is English and short, self-contained prompts; homogeneity could differ in multilingual contexts, longer outputs, or domain-specialized tasks. Third, similarity is measured with one embedding model alongside TF-IDF, so results may vary with alternative vectorizers or evaluation schemes.

Future research could expand this work by examining homogeneity across additional languages, technical or domain-specific tasks, and newer or architecturally distinct model families. Investigating how temperature, sampling strategies, system prompts, or fine-tuning methods influence similarity would further clarify the role of model configuration in producing convergent outputs. Long-form conversational studies could also reveal how homogeneity evolves over extended interactions, while ensemble prompting or multi-model workflows may shed light on strategies for mitigating convergence when diversity of ideas is desired. Together, these directions can deepen understanding of how LLMs generate meaning and how their tendency toward sameness can be both leveraged and moderated.

References

Agarwal, D., Naaman, M., & Vashistha, A. (2024). *AI suggestions homogenize writing toward Western styles and diminish cultural nuances.* arXiv. https://doi.org/10.48550/arXiv.2409.11360

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). *Homogenization effects of large language models on human creative ideation.* arXiv:2402.01536. https://doi.org/10.48550/arXiv.2402.01536

Begus, N. (2024). *Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling.* arXiv:2310.12902. https://doi.org/10.48550/arXiv.2310.12902

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). *Picking on the same person: Does algorithmic monoculture lead to outcome homogenization?* arXiv:2211.13972. https://doi.org/10.48550/arXiv.2211.13972

Kotek, H., Dockum, R., & Sun, D. Q. (2023). *Gender bias and stereotypes in large language models.* arXiv:2308.14921. https://doi.org/10.48550/arXiv.2308.14921

Liu, C., Wang, T., & Yang, S. A. (2025). *Generative AI and content homogenization: The case of digital marketing.* SSRN. https://ssrn.com/abstract=5367123

Liu, Q., Zhou, Y., Huang, J., & Li, G. (2024). *When ChatGPT is gone: Creativity reverts and homogeneity persists.* arXiv:2401.06816. https://doi.org/10.48550/arXiv.2401.06816

Pit, P., Ma, X., Conway, M., Chen, Q., Bailey, J., Pit, H., Keo, P., Diep, W., & Jiang, Y.-G. (2024). *Whose side are you on? Investigating the political stance of large language models.* arXiv:2403.13840. https://doi.org/10.48550/arXiv.2403.13840

Shaib, C., Barrow, J., Sun, J., Siu, A. F., Wallace, B. C., & Nenkova, A. (2024). *Standardizing the measurement of text diversity: A tool and a comparative analysis of scores.* arXiv:2403.00553. https://doi.org/10.48550/arXiv.2403.00553

Wenger, E., & Kenett, Y. (2025). *We're different, we're the same: Creative homogeneity across LLMs.* arXiv. https://doi.org/10.48550/arXiv.2501.19361

Xu, W., Jojic, N., Rao, S., Brockett, C., & Dolan, B. (2025). *Echoes in AI: Quantifying lack of plot diversity in LLM outputs.* arXiv:2501.00273. https://doi.org/10.48550/arXiv.2501.00273