The Illusion of Diversity: Mapping Homogeneity Across Generative AI Systems*

Hayden Eddy, Smera Shrestha, Nan Sun
Computer and Information Sciences
Washburn University
Topeka, KS 66621
{Hayden.Eddy,Smera.Shrestha,Nan.Sun}@Washburn.edu

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.

*Copyright ©2026 by the Consortium for Computing Sciences in Colleges. Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the CCSC copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Consortium for Computing Sciences in Colleges. To copy otherwise, or to republish, requires a fee and/or specific permission.

1 Introduction

Large Language Models (LLMs) now shape how people write, learn, and search. Such systems are marketed as distinct, yet prior work suggests they often produce convergent outputs in both style and meaning [6], [9]. Studies also document domain-level homogenization and stable biases and stances, which suggest that even when responses are paraphrased, they often still convey essentially the same underlying ideas [5], [4], [7]. 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).
We set out to answer the following questions:

- **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 exist 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 [6], appears as creative convergence across models [9], and surfaces in narrative "echoes" that recur across generations and systems [10]. Once a framing is established, models tend to repeat it, narrowing variation over time; assisted ideation similarly becomes more fluent yet less diverse [1], with AI stories showing stable scaffolds relative to human crowd stories [2]. Convergence also reflects socio-cultural regularities: models exhibit persistent gender stereotyping and political leanings [4], [7], creative outputs tend to pull toward dominant stylistic norms [9], and applied writing such as marketing copy becomes more uniform under AI use [5]. These patterns align with system-level monoculture risks tied to shared components and overlapping data [3].

2.2 Quantifying Similarity: Semantic and Lexical Approaches

TF-IDF similarity captures surface repetition, while embeddings capture conceptual alignment. Metric comparisons help interpret clustering and dispersion [8]. Recurring plot structures despite varied wording underscore the need for semantic evaluation [10]. Evidence from ideation and creative tasks shows convergence in both style and meaning [1], [9], 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 [3], [9], [10]. Contextual reinforcement narrows variation once a framing is set, homogeneity persists beyond immediate assistance, and ideation becomes more fluent but less diverse [6], [1]. Models are aligned to similar socio-cultural norms, their underlying biases and stances tend to converge across prompts, producing especially uniform outputs in applied writing tasks [4], [7], [5]. Methodologically, we assess sameness at two levels: *lexical* (using TF–IDF) and *semantic* (using sentence embeddings). Comparing how these two metrics behave helps us distinguish clustered response patterns from more dispersed ones [8].

2.4 How This Study Extends Prior Research

This study extends prior work on LLM homogeneity in three main ways. First, we introduce a two-axis prompt framework (logical vs. creative, specific vs. vague) and systematically apply it across five leading LLM platforms rather than focusing on a single model or narrow task. Second, we compare outputs generated under both isolated and conversational conditions, showing how interaction style shapes homogeneity within and across models. Third, we jointly analyze lexical similarity (TF–IDF) and semantic similarity (sentence embeddings) over a controlled dataset of 2,000 responses, demonstrating that apparently diverse wording can mask strong convergence at the level of ideas. Together, these design choices provide a more structured and comparative view of the "illusion of diversity" across current LLM ecosystems.

3 Methodology

3.1 Research Framework

The study began by defining a prompt typology that captured both cognitive orientation and wording. A single logical–creative scale was too limited, so we adopted a two-axis, four-quadrant framework: the horizontal axis ranges from logical reasoning to creative inference, and the vertical axis ranges from highly specific to vague, open-ended phrasing (Figure 1).

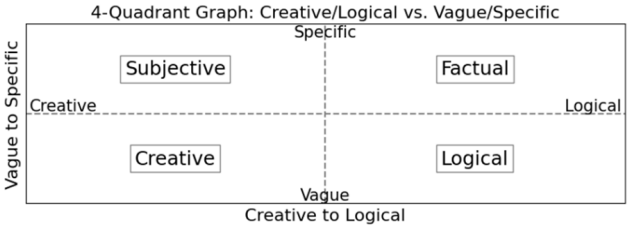

Figure 1: Two-axis prompt framework.

Crossing these axes yielded four prompt categories: *Factual* (logical/specific), *Logical* (logical/vague), *Subjective* (creative/specific), and *Creative* (creative/vague), ensuring balanced coverage of reasoning style and specificity. Prompts were developed iteratively (brainstorming, refinement, pilot testing, and classification), then reduced to remove redundancy and ensure clear quadrant fit. The final set included 40 prompts, evenly split across categories. During data collection, prompts were rarely adjusted and only when needed to prevent unintended outputs and maintain format consistency. The full prompt

list is available in our public repository at:
https://github.com/Hayd3ne/IllusionOfDiversity

3.2 Models and Interaction Protocols

We selected five LLMs spanning different architectures, training philosophies, and organizational contexts: *ChatGPT-5*, *DeepSeek-V3.1-Exp*, *Gemini 2.5 Flash*, *Llama 4*, and *Grok 4*, covering both proprietary and open-source ecosystems. To test conversational context, we used two interaction protocols: Researcher 1 submitted each prompt in a new chat session (isolated), while Researcher 2 submitted the same prompts sequentially within a single conversation (contextual), allowing prior turns to influence later responses.

Each researcher submitted each prompt to each model five times, resulting in repeated samples under both isolated and contextual conditions, yielding a total of 2,000 responses (40 prompts $\times$ 5 models $\times$ 5 trials $\times$ 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 (all-mpnet-base-v2). 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

To evaluate how similar two responses were, we used cosine similarity on two different vector representations: a TF–IDF bag-of-words representation to capture lexical overlap, and a sentence embedding representation to capture semantic similarity. Using both allows us to separate similarity in wording from similarity in underlying ideas.

For lexical similarity, we represent each response using TF–IDF (term frequency–inverse document frequency). TF–IDF counts how often each word appears in a response (term frequency) but down-weights words that are very common across all responses (inverse document frequency). As a result, generic words like "the" or "and" receive low weight, while more informative words like "algorithm" or "bias" receive higher weight. Each response becomes a TF–IDF vector, and we compute cosine similarity between these vectors. High cosine similarity in this space indicates that responses use very similar words and phrases, making TF–IDF a natural way to measure surface-level lexical similarity.

For semantic similarity, each response is encoded as a high-dimensional vector using a sentence-transformer embedding model (see Section 3.3). These embeddings are trained so that texts with similar meanings are mapped to nearby points in the vector space, even if they use different words. We again use cosine similarity, this time between embedding vectors, to quantify how close two responses are in meaning. High cosine similarity here indicates that responses express very similar ideas, even if their wording differs.

Cosine values range from -1 to 1, but in practice both TF–IDF and embedding vectors for our data produce values largely in the positive range. For our purposes, values above $\frac{\sqrt{2}}{2}$, or 0.708, are considered highly similar.

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 are defined as Prompt × Model × Researcher) using TF–IDF and embeddings. Within-cell similarity was much higher than across prompts for both methods: TF–IDF ($M = 0.3734$ vs. 0.0264; $\Delta = 0.3470$; 95% CI [0.3283, 0.3667], permutation $p = 0.0002$) and embeddings ($M = 0.8531$ vs. 0.1002; $\Delta = 0.7529$; 95% CI [0.7393, 0.7660], permutation $p = 0.0002$).

Table 1: In-Group vs. Out-Group 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	—

4.2 RQ2: Similarity by Prompt Type

We tested whether prompt structure affects homogeneity by comparing within-cell cosine similarities across four prompt types. Lexically (TF–IDF), similarity differed by type (Kruskal–Wallis $H = 28.80$, $p = 2 \times 10^{-6}$): Specific Logical ($M = 0.6759$) > Vague Logical (0.6196) > Specific Creative (0.5529) and Vague Creative (0.5634). Semantically (embeddings), differences were larger ($H = 82.02$, $p < 10^{-6}$): Vague Logical (0.9037) and Specific Logical (0.8986) were highest, with lower similarity for Specific Creative (0.8397) and especially Vague Creative (0.7705). Thus, logical prompts produce the strongest lexical and semantic convergence.

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

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

4.3 RQ3: Similarity by Model

We tested platform-level homogeneity by comparing within-cell cosine similarity across five LLMs. TF–IDF similarity differed by model (Kruskal–Wallis $H = 51.65$, $p < 10^{-6}$), with the highest overlap for DeepSeek ($M = 0.6922$) and Gemini (0.6663) and lower values for Grok (0.5672), ChatGPT (0.5618), and Llama (0.5273). Embedding similarity did not differ significantly ($H = 5.63$, $p = 0.228$); all models showed uniformly high alignment (≈ 0.83–0.87).

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

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

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 (Δ = +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 (Δ = +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.

Table 4: Protocol Comparison (Isolated—Continuous)

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

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-6). The two measures were moderately correlated (Spearman ρ = 0.6235), demonstrating shared trends but different sensitivities to lexical versus conceptual overlap.

Table 5: Paired Method Comparison (Embeddings—TF–IDF)

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

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

5 Discussion

Across analyses, responses were much more similar within the same cell than across prompts, with the strongest effects in embedding space. This indicates that when the task is fixed, models converge on a shared interpretation even when wording varies. Prompt structure also mattered: logical prompts produced the highest convergence (especially semantically), while creative prompts, particularly vague ones introduced more variance. Overall, task constraints appear to narrow the solution space, and "different phrasings" often express the same idea.

Platform and interaction effects added nuance. Semantic similarity was uniformly high across all five models, while lexical overlap varied, suggesting surface-style differences on top of shared conceptual cores. Interaction protocol also mattered: isolated single-turn prompting yielded slightly higher homogeneity than continuous conversation, consistent with a stabilizing effect when context is minimized. Embedding similarity was consistently higher than TF–IDF, and the two measures were only moderately correlated, implying they capture different aspects of overlap (conceptual alignment vs. word-level repetition). Taken together, LLM homogeneity appears multi-layered, driven primarily by task framing and broadly consistent across platforms.

These results imply that users may encounter less diversity of ideas than platform variety suggests. For researchers using LLMs for ideation, simulation, or literature generation, apparent model differences may mask deeper conceptual uniformity. In applied settings (writing assistance, analysis, content generation), consistency can support standardization but may also limit the range of viewpoints or alternatives surfaced by default.

Practically, users can reduce sameness by varying task framing along the two prompt axes (e.g., shifting specificity and logical constraint), explicitly requesting multiple distinct perspectives (e.g., "Give three contrasting approaches and explain how they differ"), and adding constraints such as audience, tone, or underrepresented viewpoints to push outputs beyond default "safe" answers. Intentional human variation—editing, remixing, and synthesizing across prompts, sessions, or models—can further counteract uniformity, even if underlying similarities remain.

Ethical risks follow from convergence: shared biases, cultural assumptions, or dominant perspectives may be reproduced consistently and at scale, and high semantic alignment can persist beneath paraphrases. Convergence can also reinforce errors when repeated generations echo the same inaccurate claim, which may be mistaken for evidence of correctness. These concerns highlight the need to audit outputs across prompt types, use counter-prompts to elicit alternative framings, and be cautious when homogeneity could obscure bias or error.

6 Limitations and Future Research

Several limitations should be acknowledged. First, we analyzed 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 a curated set of 40 English prompts centered on explanation, opinion, and short-form creative writing; 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

[1] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. "Homogenization Effects of Large Language Models on Human Creative Ideation". In: *Creativity and Cognition*. C&C '24. ACM, June 2024, pp. 413–425. DOI: 10.1145/3635636.3656204. URL: http://dx.doi.org/10.1145/3635636.3656204.

[2] Nina Beguš. "Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling". In: *Humanities and Social Sciences Communications* 11.1 (Oct. 2024). ISSN: 2662-9992. DOI: 10.1057/s41599-024-03868-8. URL: http://dx.doi.org/10.1057/s41599-024-03868-8.

[3] Rishi Bommasani et al. *Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?* 2022. arXiv: 2211.13972 [cs.LG]. URL: https://arxiv.org/abs/2211.13972.

[4] Hadas Kotek, Rikker Dockum, and David Sun. "Gender bias and stereotypes in Large Language Models". In: *Proceedings of The ACM Collective Intelligence Conference*. CI '23. ACM, Nov. 2023, pp. 12–24. DOI: 10.1145/3582269.3615599. URL: http://dx.doi.org/10.1145/3582269.3615599.

[5] Chaoran Liu, Tong Wang, and S. Alex Yang. *Generative AI and Content Homogenization: The Case of Digital Marketing. SSRN Working Paper*. Available at SSRN. July 26, 2025. DOI: 10.2139/ssrn.5367123. URL: https://ssrn.com/abstract=5367123.

[6] Qinghan Liu et al. *When ChatGPT is gone: Creativity reverts and homogeneity persists*. 2024. arXiv: 2401.06816 [cs.CL]. URL: https://arxiv.org/abs/2401.06816.

[7] Pagnarasmey Pit et al. *Whose Side Are You On? Investigating the Political Stance of Large Language Models*. 2024. arXiv: 2403.13840 [cs.CL]. URL: https://arxiv.org/abs/2403.13840.

[8] Chantal Shaib et al. *Standardizing the Measurement of Text Diversity: A Tool and a Comparative Analysis of Scores*. 2025. arXiv: 2403.00553 [cs.CL]. URL: https://arxiv.org/abs/2403.00553.

[9] Emily Wenger and Yoed Kenett. *We're Different, We're the Same: Creative Homogeneity Across LLMs*. 2025. arXiv: 2501.19361 [cs.CY]. URL: https://arxiv.org/abs/2501.19361.

[10] Weijia Xu et al. "Echoes in AI: Quantifying lack of plot diversity in LLM outputs". In: *Proceedings of the National Academy of Sciences* 122.35 (Aug. 2025). ISSN: 1091-6490. DOI: 10.1073/pnas.2504966122. URL: http://dx.doi.org/10.1073/pnas.2504966122.