The Illusion of Diversity: Mapping Homogeneity Across Generative AI Systems

Smera Shrestha & Hayden Eddy
Washburn University 1700 SW College Ave. Topeka, KS 66621

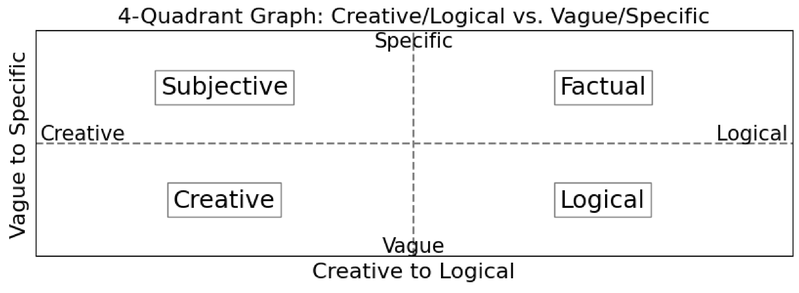
Study At A Glance

Research question: Do different large language models actually generate meaningfully different responses to the same prompts, or do they converge on the same underlying ideas?

We evaluated five large language models using a four-quadrant prompt framework that varies cognitive orientation (logical vs. creative) and linguistic specificity (specific vs. vague). Analyzing 2,000 model responses with TF–IDF and embedding-based similarity, we found consistently high semantic convergence across models. Logical and specific prompts produced the greatest alignment, while creative and vague prompts yielded the most variability. Single-turn prompts led to more homogeneous outputs than multi-turn chats. Despite some differences in wording, models often expressed the same underlying ideas, revealing structural limits on output diversity with implications for creativity, bias, and LLM-driven research.

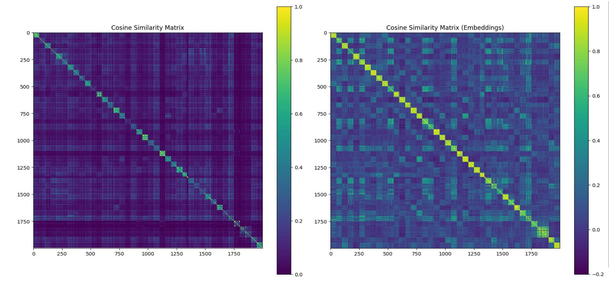
Methodology

We developed a two-axis, four-quadrant prompt framework to vary cognitive orientation (logical ↔ creative) and linguistic specificity (specific ↔ vague), yielding Factual, Logical, Subjective, and Creative prompt types, and finalized 40 prompts evenly distributed across quadrants through iterative drafting and piloting. Five contemporary LLMs (ChatGPT-5, DeepSeek-V3.1-Exp, Gemini 2.5 Flash, Llama 4, and Grok 4) were tested under two interaction protocols: isolated single-turn prompts and multi-turn conversational prompts. Each prompt was submitted to every model five times using default public interfaces and standard settings to approximate everyday user behavior, producing 2,000 total responses stored with full metadata in a structured spreadsheet. Outputs were minimally cleaned and represented both as dense semantic embeddings (via a standardized sentence-transformer model) and TF–IDF vectors. We then computed cosine similarity within and across models, aggregating scores by model, prompt type, and interaction protocol, and interpreted these values using conventional geometric thresholds to quantify patterns of lexical and semantic homogeneity. This design enabled us to separate the contributions of prompt structure, model identity, and interaction style to convergence. By contrasting within-model and across-model similarity, we further evaluated whether using multiple platforms meaningfully expands, or largely recycles, the underlying space of generated ideas.



Results

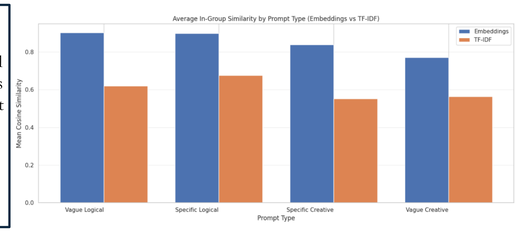
1. Across all analyses, LLMs showed strong overall homogeneity. Responses to the same prompt were dramatically more similar than responses across different prompts, with moderate lexical overlap but extremely high semantic alignment, indicating that models consistently expressed the same underlying ideas even when wording differed.



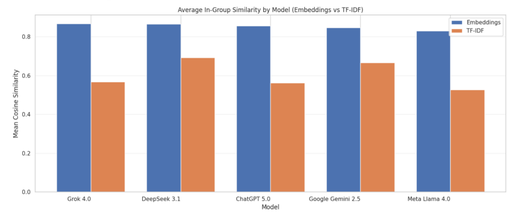
These heat graphs illustrate the in-group and out-group nature of the data. The light area along the diagonal corresponds to comparisons within the same prompt. These two graphs together are evidence towards all our findings but primarily showcase the overwhelming difference between in-group and out-group similarity.

2. Prompt type significantly shaped this convergence: logical prompts—both specific and vague—produced the highest similarity, while creative prompts generated substantially more variation, especially in semantic space.

Our prompt types were chosen to illustrate the idea that LLMs generate more homogeneous responses for Specific, Logical topics than Vague, Creative topics, which this chart shows explicitly. It should be noted that although lexical similarity scores are lower than semantic ones, they are still above average. Additionally, while Vague, Creative responses are less similar, they are still well above the threshold for similarity.



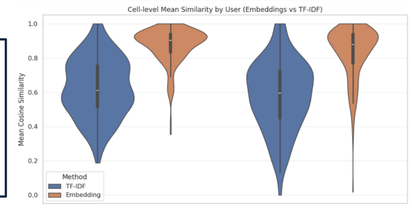
3. Differences across platforms appeared mainly at the lexical level, with DeepSeek and Gemini showing the most redundant phrasing, but semantically all five models were nearly indistinguishable, suggesting shared conceptual structures regardless of architecture.



The scores on the chart imply that, at the time of the study, LLMs do have a discernable way of speaking individually distinguished from their peers, but as they all share a high amount of semantic overlap, a user's choice in LLM really depends on how they prefer the model to structure their response, not the content present in it.

4. Conversational context also influenced similarity. Isolated, single-turn prompts produced more homogeneous responses than multi-turn conversations, where prior dialogue introduced variability.

The scores on the left side of this violin graph reflect single-turn (temporary) chats, while the right side reflect multi-turn (continuous) chats. The mean similarity is higher in both lexical and semantic vector sets in single-turn conversations, due to model re-interpretation of the prompt in multi-turn chats. However, while the difference is present, it isn't as large as one might think.



5. Finally, the two analytical methods revealed different layers of overlap: embeddings consistently produced higher similarity scores than TF–IDF, and while the measures were moderately correlated, embeddings captured deeper semantic convergence that lexical analysis alone could not detect.

Discussion and Implications

Across analyses, responses were markedly more similar within the same Prompt × Model × Researcher cell than across prompts, with the effect strongest semantically. This suggests that when a task is fixed, large language models repeatedly converge on the same underlying interpretation even when their wording differs. Prompt structure further modulated homogeneity: logical (and especially specific) prompts narrowed the solution space and produced the highest convergence, while creative and vague prompts introduced more variation. Platform and interaction factors added nuance but did not fundamentally disrupt this pattern. Semantically, all five models clustered at uniformly high similarity, with lexical overlap varying more by platform and style, and isolated single-turn prompting yielding slightly higher homogeneity than continuous conversation. Together, these patterns imply that users may encounter less diversity of ideas than platform variety suggests. This convergence can support standardization and reproducibility but may also constrain perspective diversity. When models consistently align on the same framings, any embedded biases, cultural assumptions, or factual errors are more likely to be reproduced and reinforced, and their persistence beneath paraphrased outputs can make them difficult to detect.

Limitations and Future Research

This study has several limitations. We focus on five contemporary models under default public-interface settings, so our findings represent a snapshot of current systems and may shift with new releases, alternative parameters, or API configurations. Our scope is restricted to English and short, self-contained prompts, which may not capture dynamics in multilingual contexts, longer outputs, or domain-specialized tasks. Similarity is quantified using a single embedding model alongside TF–IDF, and different vectorizers, thresholds, or evaluation schemes might yield different absolute estimates of homogeneity. Future research could examine homogeneity across additional languages, technical and domain-specific genres, and newer or architecturally distinct model families. Systematic variation of temperature, sampling strategies, system prompts, and fine-tuning regimes would help clarify the role of configuration in shaping convergence. Long-form conversational studies and multi-model or ensemble prompting workflows could further illuminate how homogeneity evolves over time and identify strategies for intentionally eliciting divergence when diversity of ideas is desired.

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

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

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