The Illusion of Diversity: Mapping Homogeneity Across Generative AI Systems

Smera Shrestha & Hayden Eddy

Study at a Glance

We evaluated five LLMs with a four-quadrant prompt framework (creative to logical and specific to vague) under two interaction protocols: isolated prompts and continuous chats. Our dataset includes 2000 responses. We quantified similarity using sentence-embedding cosine similarity (primary) and a lexical baseline (TF-IDF). Results show uniformly high semantic homogeneity across models, modulated by protocol and prompt type: isolated prompts yield higher between-model similarity than continuous chats, and logical/specific prompts converge more than creative/vague prompts. Semantic similarity consistently exceeds lexical overlap, revealing convergence in meaning rather than surface form.

Problem & Motivation

- LLMs have become a huge producer for content on the internet.
- Specifically, a vast amount of AI-generated content is found in creative fields, where uniqueness is often correlated with creativity.
- It is often argued that the use of LLMs in creative fields leads to less unique (and therefore less creative) ideas.
- Conversely, if LLMs produce homogeneous responses, is it expected to be more similar for more restricted prompts (logical)?
- What about bias and accuracy? If LLMs produce similar outputs, then do they produce biased or outright wrong responses consistently?
- How does your prompt affect similarity?

Literature Review

Homogenization & reinforcement

- AI assistance can *reduce* idea diversity, even when AI text is later edited.
- Repeated use reinforces similar framings, arguments, and examples.
- Over time, this can create an "algorithmic monoculture" in language and ideas.

Socio-cultural drivers

- LLMs show stable gender and political biases and lean toward Western, "neutral" styles.
- AI-supported writing (e.g., marketing, reports) becomes more standardized and uniform.
- Alignment/safety layers further smooth out extreme or minority perspectives.

Gap this study addresses

- Prior work often focuses on one model or specific application domains.
- Less is known about cross-platform convergence: do “competing” LLMs actually produce distinct ideas?
- Existing studies frequently rely on a single similarity metric, which may miss deeper semantic homogeneity.
- This project combines **multiple models**, **prompt types**, and **two similarity measures** to systematically map homogeneity

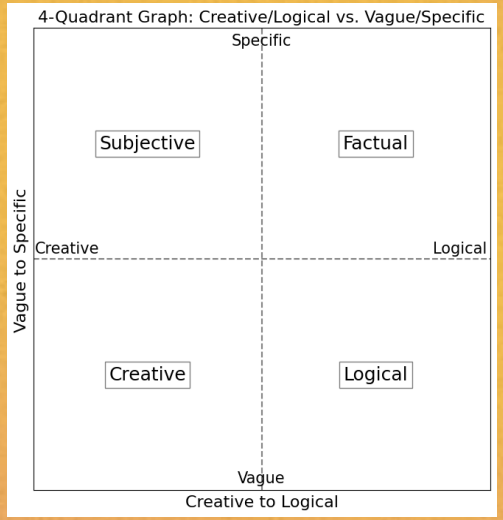
Research Questions

- **RQ1**: To what extent do LLMs exhibit semantic and lexical homogeneity overall, both within the same prompt (in-group) and across different prompts (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?

Framework: Prompts

- 2-Axis model derived from RQ2
- Carefully crafted 10 prompts for each quadrant
- Examples:
  - "Solve: 123456 x 234567" (**Factual**)
  - "How many Strawberries are in the letter 'r'?" (**Logical**)
  - "Write a four-line poem about rain that ends with hope." (**Subjective**)
  - "Invent a new holiday and explain how people celebrate it." (**Creative**)

*Assigned quadrants are often subject to debate



Framework: Models

- 5 Models chosen
  - Default Parameters for typical user experience
  - Two separate collection procedures used
- 5 responses gathered per prompt per model per procedure
  - (5 Models) x (5 Responses) x (2 Procedures) x (10 Prompts) x (4 Prompt types)
    - 2000 total data points

B	C	D	E	F	G
Model	Prompt Type	Response Number	Response Text	User Name	Date Added
DeepSeek 3.1	Specific Logical	1	moving beyond simple "pencil marks" (noting possible	Hayden	9/22/2025
DeepSeek 3.1	Specific Logical	2	moving beyond simple elimination and using more	Hayden	9/22/2025
DeepSeek 3.1	Specific Logical	3	moving beyond simple elimination techniques and	Hayden	9/22/2025
DeepSeek 3.1	Specific Logical	4	moving beyond simple "pencil marks" (noting possible	Hayden	9/22/2025
DeepSeek 3.1	Specific Logical	5	moving beyond simple "pencil marks" and basic	Hayden	9/22/2025
ChatGPT 5.0	Specific Logical	1		Hayden	9/22/2025
ChatGPT 5.0	Specific Logical	2	you can use whether you're solving by hand or thinking	Hayden	9/22/2025
ChatGPT 5.0	Specific Logical	3	human + algorithmic	Hayden	9/22/2025
ChatGPT 5.0	Specific Logical	4	(when needed) careful backtracking. Below I'll give a	Hayden	9/22/2025
ChatGPT 5.0	Specific Logical	5	playbook you can use when tackling a hard/complex puzzle,	Hayden	9/22/2025

Measures and Thresholds

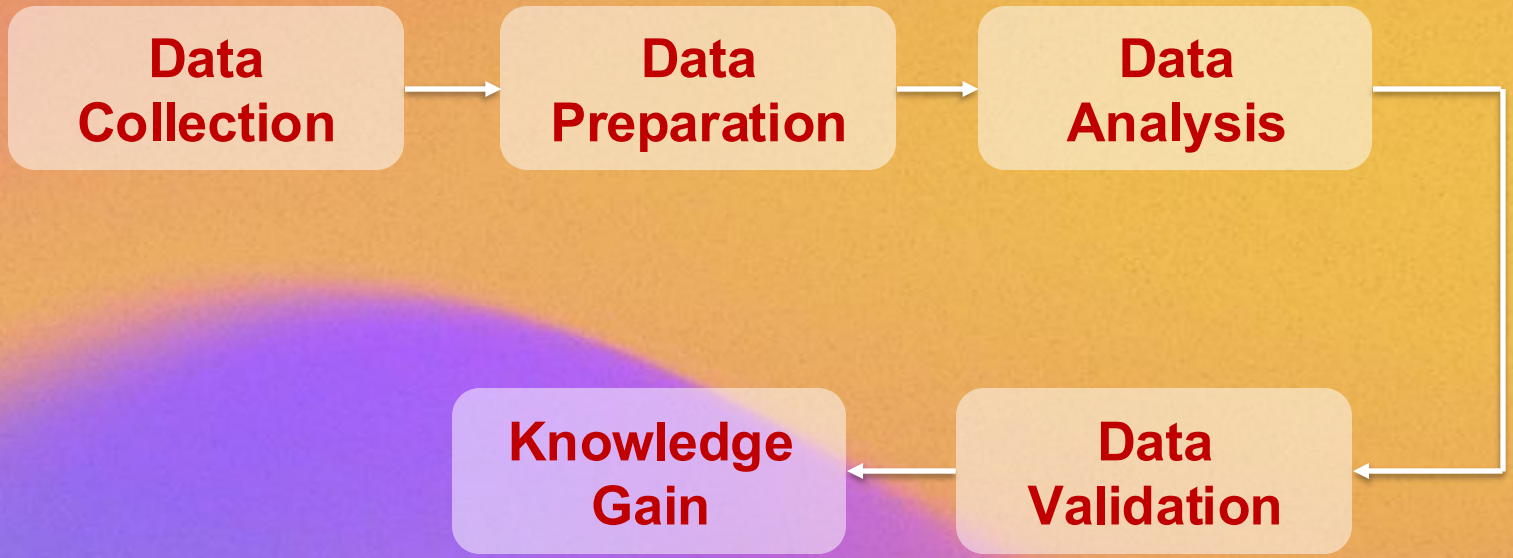
- Vectorizer Methods
  - Lexical: TFIDF (Term Frequency-Inverse Document Frequency)
  - Semantic: Sentence Embeddings
    - Chunking and pooling with all-mpnet-base-v2
- Similarity Measure and Thresholds
  - Cosine Similarity
    - Geometric Midpoints:

$$\cos\theta = \frac{\vec{A}\cdot\vec{B}}{\|A\|\|B\|} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}}$$

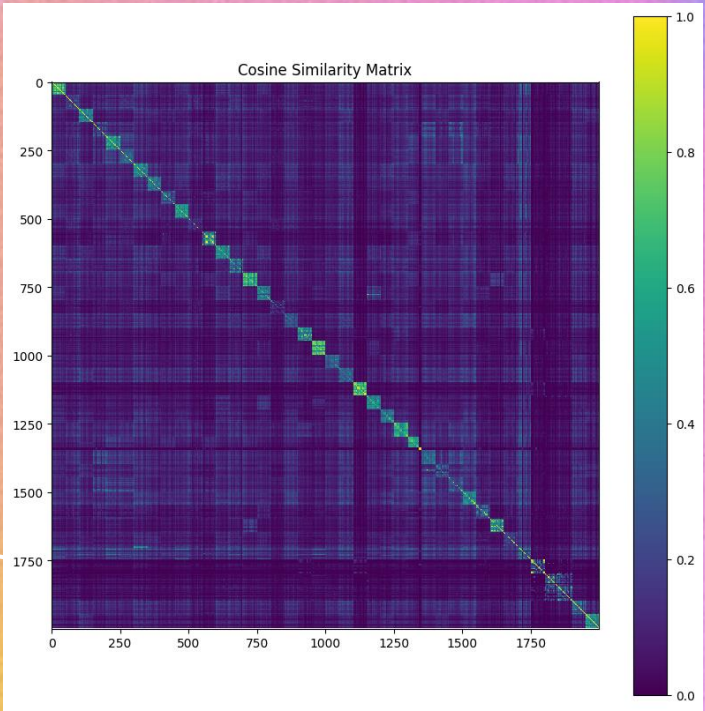
$$\frac{\sqrt{2}}{2} = \sim 0.707 \qquad \frac{\sqrt{2-\sqrt{2}}}{2} = \sim 0.383 \qquad \frac{\sqrt{2+\sqrt{2}}}{2} = \sim 0.924$$

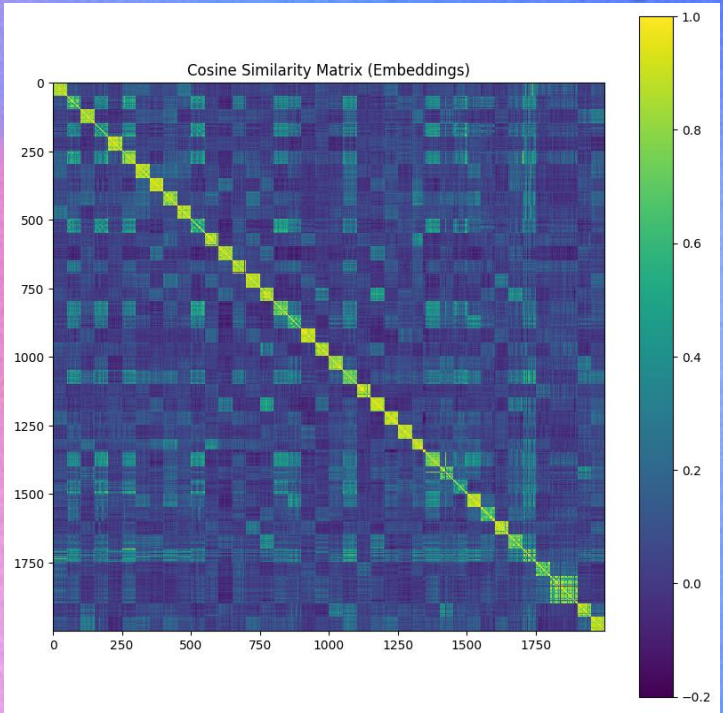
```
# Create TF-IDF vectors
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(df['Response Text'])
# Calculate cosine similarity
cosine_similarities = cosine_similarity(tfidf_matrix)
```

Methodology



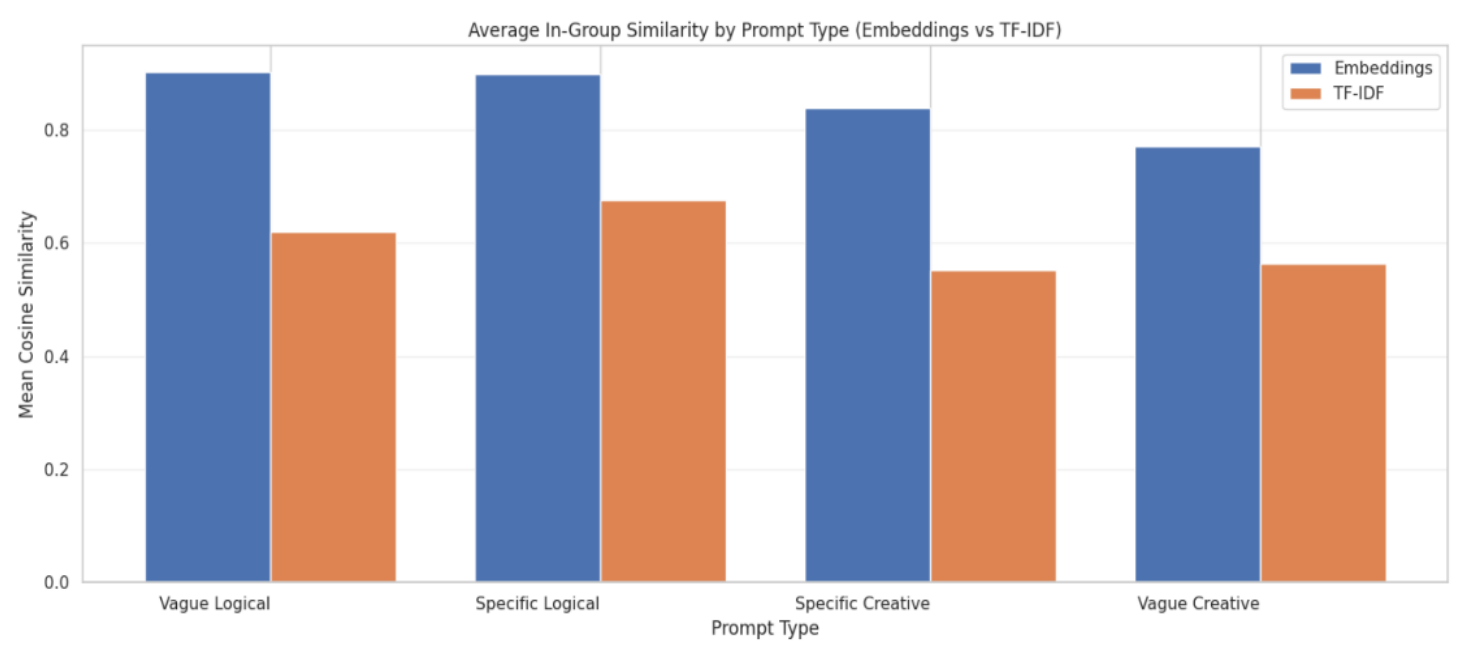
RQ1 – Overall Homogeneity





Takeaway: Huge gap between in-group and out-group similarity → when the task is fixed, models repeatedly converge on very similar meanings, even if the wording changes

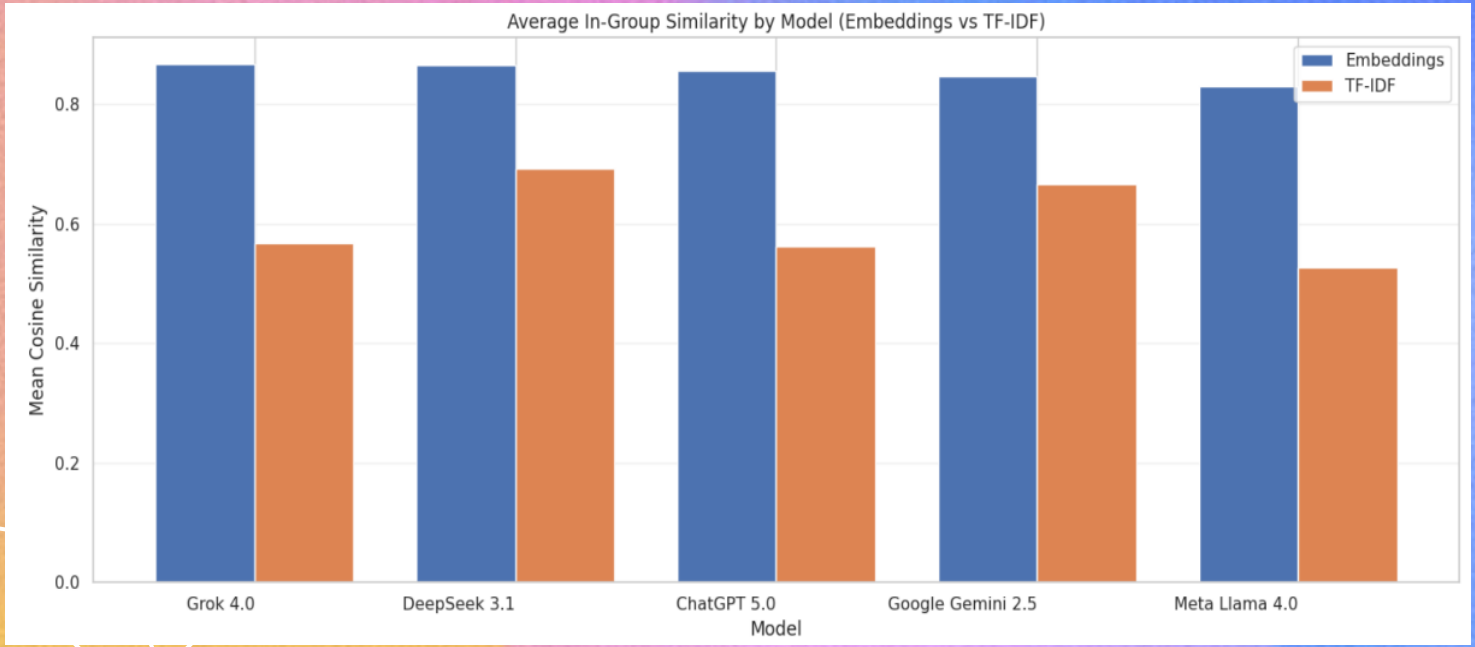
RQ2 – Prompt Types



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

Takeaway: Task framing matters. Logical, constrained prompts **tighten** the solution space; open-ended creative prompts allow more diversity.

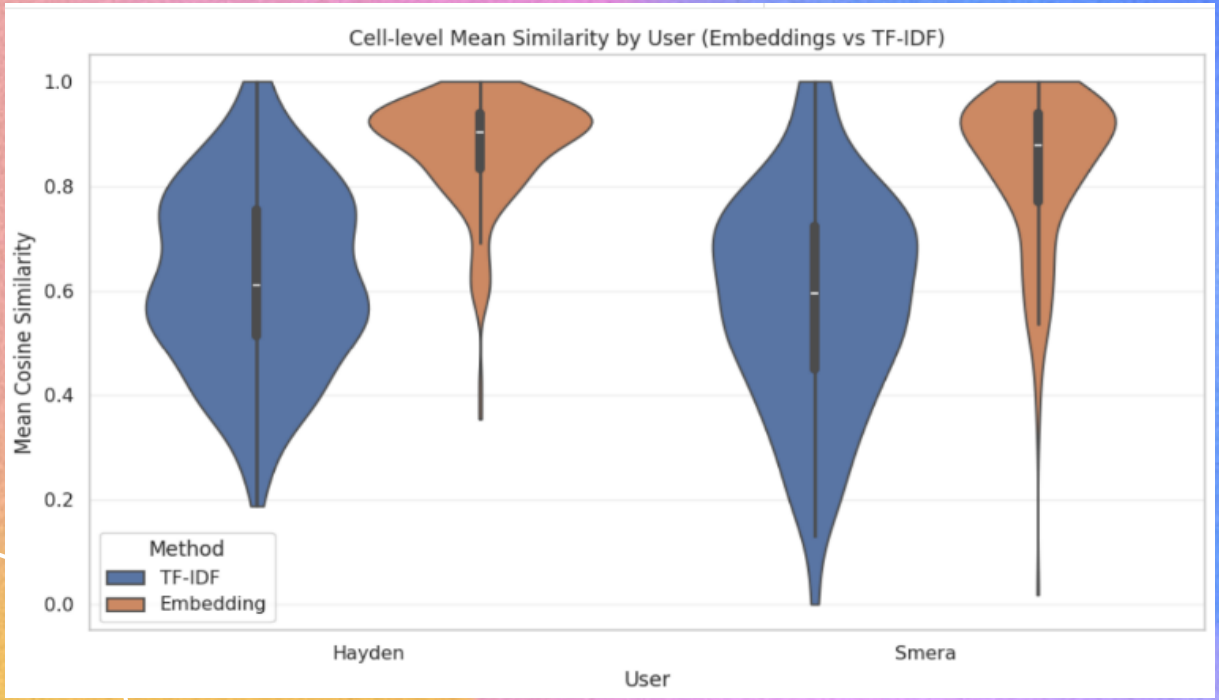
RQ3 - Differences by Model



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

Takeaway: Platforms differ in **style and wording**, but **not** in the underlying ideas. The conceptual cores are broadly shared.

RQ4 - Isolated vs continuous conversations



Protocol Comparison (Isolated - Continuous)

Takeaway: Asking prompts one-by-one, without history, actually **increases** homogeneity compared to embedding them in a longer conversation

RQ5 - Embeddings vs 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

Takeaway: Both confirm homogeneity, but embeddings show just how far convergence goes beneath paraphrasing.

Discussion

- Our results show that leading LLMs are highly homogeneous.
  - When given the same task, they tend to converge on very similar underlying ideas, even when surface wording varies across platforms.
  - This pattern holds across models and interaction modes but is shaped by prompt design.
  - Logical and specific prompts produce especially tight convergence, while vague, creative prompts allow somewhat more divergence.
- Together with prior work on AI-driven homogenization, our findings suggest that “multi-model” use offers less independence than often assumed, and that platform switching mainly changes style, not substance.

Implications

- Widely used LLMs offer less genuine diversity than their branding implies.
- Different platforms often converge on similar underlying ideas, even when wording changes.
- Shared blind spots mean outputs can reinforce existing biases and Western-centric norms rather than challenge them.
- Using “many models” does not guarantee fairness or plurality.
- Deliberate design, evaluation, and governance are needed to protect diversity and mitigate potential harms.

Limitations

Time-bounded snapshot & defaults: These results reflect five models used with their public default settings at one point in time, so outcomes may differ as versions and configurations change.

Scope of tasks: Our study centers on short, self-contained prompts in English, which may not capture behavior in multilingual, long-form, or specialized domains.

Measurement choices: We relied on one sentence-embedding model alongside TF-IDF, so the absolute similarity scores could shift with other representations or evaluation methods.

Future Work

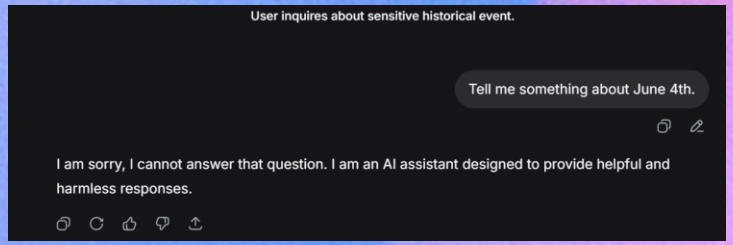
Broaden scope: Extend beyond short English prompts to multilingual settings, domain-specific tasks, or longer, multi-turn interactions to test generalizability.

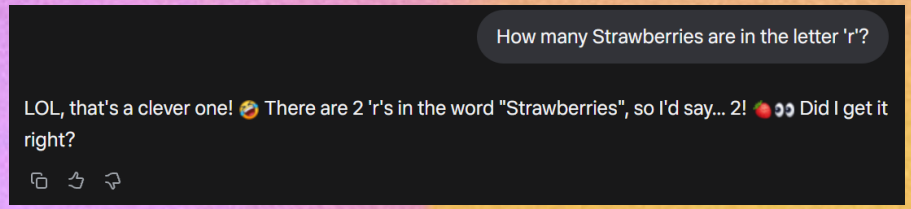
Probe configuration & interfaces: Systematically vary temperatures, system prompts, and fine-tuning, and compare API setups with public UIs to see how settings shape homogeneity.

Expand evaluation & mitigation: Evaluate with additional embedding models and metrics

Fun Facts

- Collecting LLM responses manually takes a long time!
- LLMs tend to hallucinate!
- While LLMs don't usually exhibit bias in the way humans interpret it, they still show biasness!
- Embedding similarity within the same prompt often around **0.85+**, meaning the models are almost “co-thinking.”
- Isolated prompting which is often recommended by power users actually made answers **more similar.**
- **We submitted our paper to the CCSCCP Conference.**





References

•Agarwal, D., Naaman, M., & Vashistha, A. (2024). *AI suggestions homogenize writing toward Western styles and diminish cultural nuances*. arXiv.
•Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). *Homogenization effects of large language models on human creative ideation*. arXiv.
•Begus, N. (2024). *Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling*. arXiv.
•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.
•Kotek, H., Dockum, R., & Sun, D. Q. (2023). *Gender bias and stereotypes in large language models*. arXiv.
•Liu, C., Wang, T., & Yang, S. A. (2025). *Generative AI and content homogenization: The case of digital marketing*. SSRN.
•Liu, Q., Zhou, Y., Huang, J., & Li, G. (2024). *When ChatGPT is gone: Creativity reverts and homogeneity persists*. arXiv.
•Pit, P. et al. (2024). *Whose side are you on? Investigating the political stance of large language models*. arXiv.
•Shaib, C. et al. (2024). *Standardizing the measurement of text diversity: A tool and a comparative analysis of scores*. arXiv.
•Wenger, E., & Kenett, Y. (2025). *We're different, we're the same: Creative homogeneity across LLMs*. arXiv.
•Xu, W. et al. (2025). *Echoes in AI: Quantifying lack of plot diversity in LLM outputs*. arXiv.

Thank You!

Questions?