AI Models' 'Creative' Output is Becoming Similar Across Providers

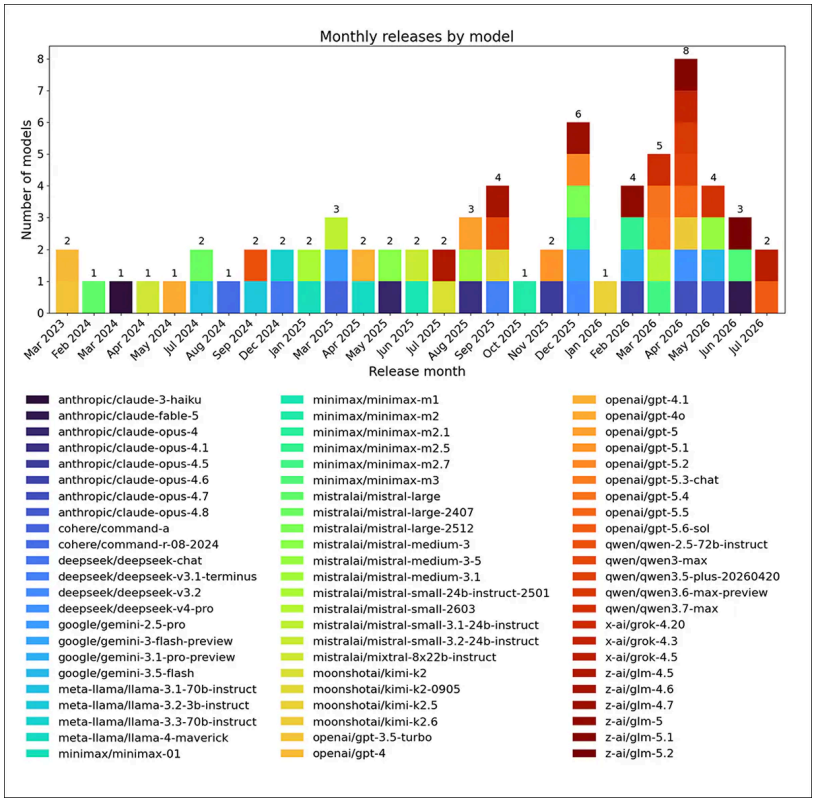
Originally published at https://www.unite.ai/ai-models-creative-output-is-becoming-similar-across-providers/

New research suggests that AI models from rival companies are becoming more alike in their creative answers, potentially narrowing the range of ideas that users encounter.

New research from the US has found that 'creative' output across a wide range of leading AI models is becoming more similar over time.

The authors, from Duke University, tested 69 models across 12 provider families, covering releases from 2023 to 2026, and found a statistically significant decline in output diversity across both real-world open-ended questions, and a standard creativity test – suggesting that similar ideas may be offered in response to 'creative' requests, increasingly, among all the major LLM providers.

The provider families tested were Anthropic; Cohere; DeepSeek; Google; Meta; MiniMax; Mistral AI; Moonshot AI; OpenAI; Qwen; xAI; and Z.ai*:



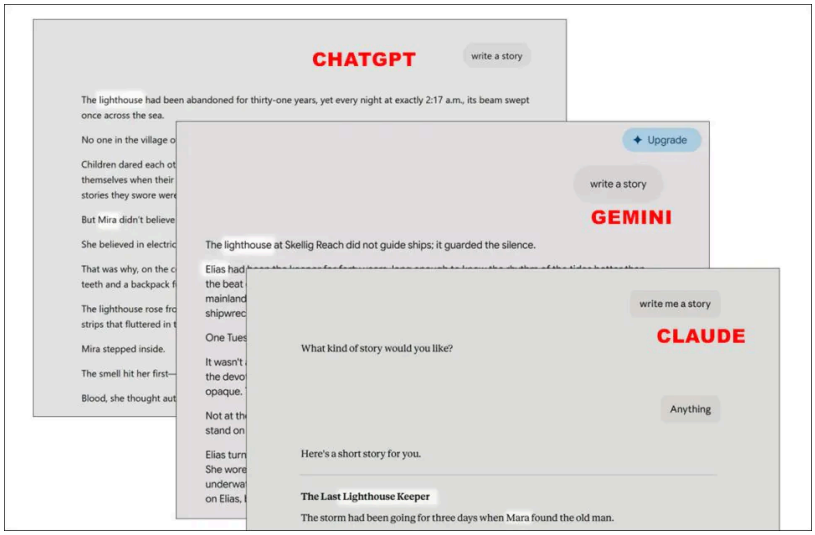
From the new paper – model releases used in the study, from March 2023 to July 2026. The chart covers 69 mode versions across 12 AI providers, demonstrating distributed across the three-year period, and showing increasingly frequent release schedules for some providers, which has to be accounted for in the authors' re

The authors of the new paper state**:

> *'We find that **LLM responses to the open-ended prompts we [test] have become increasingly similar over time**.*
>
> *'This suggests that LLMs are becoming less creative in tasks that involve generating open-ended responses, demanding scrutiny of their long-term usefulness as creative assistants.'*

Although 'algorithmic monoculture' has become an established line of study, an examination of the trends toward homogeneity in AI-generated creative outputs has not been undertaken until now, the authors assert.

However, isolated incidents have pointed towards this convergence for some time, including the strange case outlined in Cornell University's May 20026 study, where a diverse range of LLM providers were demonstrated to have strange, colliding obsessions about 'lighthouse keepers', and a particular set of anachronistic names, including 'Mara' and 'Elias':



In May, Cornell University's new paper found strange similarities across LLM providers when given 'open prompts' – though no clues as to why have yet become a data fueling these models.

The new study is more systematic: the authors tested three years of models against both real-world creative prompts and a classic psychology test that solicits unusual uses for everyday objects. They then measured how far apart the models' answers were in meaning, tracking whether those distances shrank across successive generations.

The authors of the new work, titled *Are LLMs becoming similarly creative? Evidence from three years of models*, state†:

> *'Our findings, though preliminary, raise concerns about the long-term usefulness of LLMs as creative partners. Even if models perform well on creative tasks, converging outputs could bound the range of possibilities LLM users are exposed to, and with it, the breadth of their own thinking.*
>
> *'If using an LLM for creative tasks like essay writing decreases one's brain activity, could using increasingly less creative LLMs–the trend suggested by our study–further worsen LLMs' effects on human creativity, as observed by this and other studies?'*

Already, the diverse characteristics of LLM text output have become fodder for memes; and, as we reported in May this year, at least one author has already self-published a book apparently featuring the aforementioned 'lighthouse' fixation common to major models.

So, in a climate where publishers are increasingly withdrawing books that they suspect of being AI-written or AI-aided, the new research would seem to indicate that we can expect a growth of 'plot coincidences' emerging from apparently human-written works, including academic works submitted by students.

As for where this convergence is coming from, the authors hypothesize that overlapping training data and increasingly similar optimization objectives may be pushing models toward comparable internal representations of concepts and semantic relationships – ultimately producing more similar answers[†]:

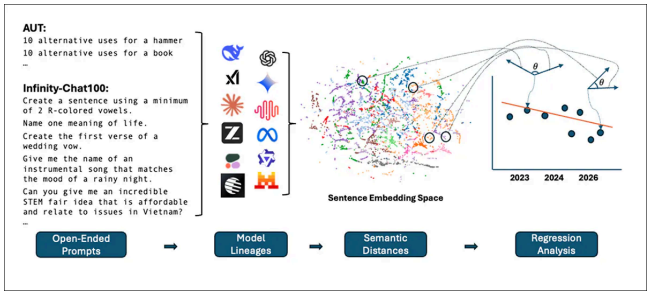
> *'[It] remains unclear whether this homogeneity is a temporary byproduct of a still-developing technology or an inevitable–potentially compounding–feature of statistical language models.*
>
> *'Plausible forces point in both directions. A growing body of academic work suggests that generative models trained on overlapping data and optimized toward similar objectives will organize concepts and semantic relationships in increasingly similar ways, resulting in similar outputs.'*

However, the authors also cite work indicating that as models evolve, they may diverge again into their own ring-fenced set of characteristics, in regard to creative output.

Method

To track whether AI models are becoming more alike, the researchers began with open-ended questions, curating answers from successive generations of models. Each answer was converted into a numerical representation of its meaning, making it possible to measure how similar or different the responses were.

Regression analysis was then used to establish whether these differences were shrinking as newer models were released:



The authors' schema for measuring whether AI outputs become more similar over time. Open-ended creative prompts were run across different model families, their answers mapped by meaning, to measure the distance between them, and with regression analysis tracking how those distances changed across successive model generations.

Two sets of prompts were selected to test creativity from different angles. First, the Alternate Uses Task (AUT), a standard psychology test of divergent thinking, asks for unconventional uses for ordinary objects, with the study using objects such as a *book*, *shoe* and *hammer*. Its fixed format provided a controlled way to compare the ideas produced by different models.

A further 100 prompts were distilled from Infinity-Chat100, a collection derived from real-world conversations with language models. These covered less-constrained creative tasks involving content generation, problem-solving, brainstorming and ideation; tasks that would permit greater freedom in regard to the answers produced and the approaches taken.

The complete AUT and Infinity-Chat100 prompt sets were then sent to models released between March 2023 and July 2026, covering 12 providers and systems from Anthropic, Google, Meta, OpenAI, DeepSeek and Mistral AI. All responses were collected through OpenRouter's API under the same sampling settings, with temperature (freedom to respond creatively) and *top-p* both set to one.

The models were arranged by release month to examine whether newer generations were producing more similar answers.

Since release dates don't actually capture changes in training data, architecture or post-training, the authors treated the resulting trend as an association with model development, rather than evidence that the passage of time itself causes convergence.

The models' answers were then converted into embeddings using the all-MiniLM-L6-v2 sentence-transformer. Each response was represented as a numerical vector, allowing answers with similar meanings to be placed in similar directions, and their semantic distance to be measured.

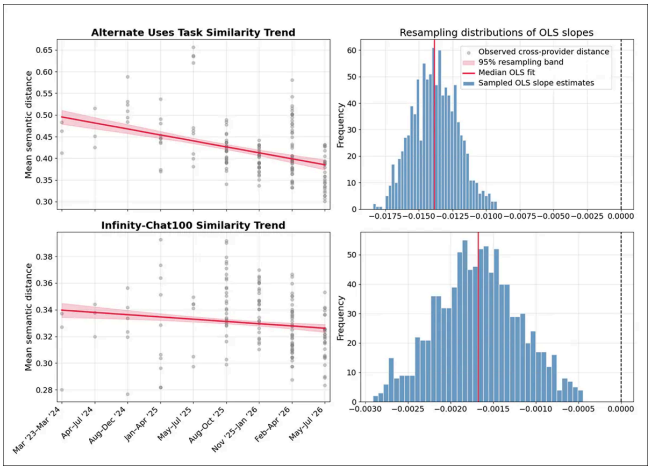
For the AUT, all uses suggested for each object were treated as a single response, producing ten embeddings per model. For Infinity-Chat100, each answer was embedded separately, producing 100 embeddings per model.

The 27 release months were grouped into nine time-periods, and only models from different providers were compared. Semantic distances between their answers were measured within the scope of each period, with smaller distances indicating greater similarity.

To prevent providers with more models from dominating the results, the analysis was repeated 1,000 times, using balanced samples from each provider.

Results

Linear regression was then used to track these distances over time, with a consistent negative slope indicating that models from different providers were becoming more similar:



Initial test results, showing whether creative responses from different AI providers became more similar over time. Semantic distances declined for both the Alternate Uses Task and Infinity-Chat100 prompts, while repeated balanced sampling produced consistently negative trends, indicating increasing similarity across model generations.

Of these results the authors emphasize:

> *'For both the AUT and Infinity-Chat response sets, we observe a decline in cross-provider output distances over the observation period.*
>
> *'This suggest decreasing diversity–or increasing homogeneity–of LLM creative outputs over time.'*

The difference between older and newer models was clearest in the Alternate Uses Task. The average distance between answers from different providers fell from about 0.50 for the earliest models to below 0.40 for the newest, meaning that their answers became substantially more alike. The same pattern was found with Infinity-Chat100, though the change was smaller, with average distance falling from about 0.34 to just above 0.32.

Dataset	Models (Pairs)	Slope per Bin	95% CI
Alternate Uses Task	68 (273)	-0.01385	$[-0.01695, -0.01044]$
Infinity-Chat	67 (268)	-0.00167	$[-0.00267, -0.00074]$

Test results measuring how quickly answers from the different AI providers became more similar over time. The 'Alternate Uses' Task showed a much stronger decline in differences between models than Infinity-Chat, with both results remaining consistent across the researchers' repeated sampling tests.

The finding also survived all 1,000 rounds of balanced resampling. In every case, newer models produced answers that were closer together than those from older models. The size of these declines, together with the researchers' 95% confidence intervals, are depicted above.

Regarding this, the paper states:

> *'The magnitude of the AUT decline is particularly notable given the task's purpose. The AUT explicitly tests divergent thinking by instructing models to produce uses that are as original and unexpected as possible, making it precisely the setting in which outputs would be expected to differ.'*

The Infinity-Chat results show that the same trend also appeared across a much wider range of creative tasks. Its 100 prompts asked models to produce many different kinds of open-ended answers, and *these answers became more similar over time*.

The change was smaller than in the Alternate Uses Task, but became clearer among models released from 2025 onward. The authors suggest that this could be an early sign that creative outputs from different AI providers are becoming more alike generally, rather than only on one particular type of test.

Conclusion

Issues around LLM and VLM convergence are an ongoing and growing source of concern in both the research community and the consumers of the downstream models that issue from them. We are coming to the very end of the first and only 'pure' generation of data that the research scene will ever have access to; and the determination of model providers to exfiltrate the data of rivals inevitably risks convergence, since the data is becoming identical, and principles of model training are similar, if not identical, among providers.

Therefore, nothing quite as simple as replacing em-dashes is likely to emerge to combat the growing similarity of creative output among the model families, and cases of duplication are, instead, likely to come to light in rather more public and embarrassing ways.

* *The complete model list is Claude-3-Haiku; Claude-Fable-5; Claude-Opus-4; Claude-Opus-4.1; Claude-Opus-4.5; Claude-Opus-4.6; Claude-Opus-4.7; Claude-Opus-4.8; Command-A; Command-R-08-2024; DeepSeek-Chat; DeepSeek-V3.1-Terminus; DeepSeek-V3.2; DeepSeek-V4-Pro; Gemini-2.5-Pro; Gemini-3-Flash-Preview; Gemini-3.1-Pro-Preview; Gemini-3.5-Flash; Llama-3.1-70B-Instruct; Llama-3.2-3B-Instruct; Llama-3.3-70B-Instruct; Llama-4-Maverick; MiniMax-01; MiniMax-M1; MiniMax-M2; MiniMax-M2.1; MiniMax-M2.5; MiniMax-M2.7; MiniMax-M3; Mistral-Large; Mistral-Large-2407; Mistral-Large-2512; Mistral-*

Medium-3; Mistral-Medium-3.5; Mistral-Medium-3.1; Mistral-Small-24B-Instruct-2501; Mistral-Small-2603; Mistral-Small-3.1-24B-Instruct; Mistral-Small-3.2-24B-Instruct; Mixtral-8x22B-Instruct; Kimi-K2; Kimi-K2-0905; Kimi-K2.5; Kimi-K2.6; GPT-3.5-Turbo; GPT-4; GPT-4.1; GPT-4o; GPT-5; GPT-5.1; GPT-5.2; GPT-5.3-Chat; GPT-5.4; GPT-5.5; GPT-5.6-Sol; Qwen-2.5-72B-Instruct; Qwen3-Max; Qwen3.5-Plus-20260420; Qwen3.6-Max-Preview; Qwen3.7-Max; Grok-4.20; Grok-4.3; Grok-4.5; GLM-4.5; GLM-4.6; GLM-4.7; GLM-5; GLM-5.1; and GLM-5.2

** *Authors' emphases, not mine.*

† *My conversion of the authors' inline citations to hyperlinks.*

First published Friday, August 21, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More