We're Different, We're the Same: Creative Homogeneity Across LLMs

EMILY WENGER*, Duke University

YOED KENETT, Technion - Israel Institute of Technology

Numerous powerful large language models (LLMs) are now available for use as writing support tools, idea generators, and beyond. Although these LLMs are marketed as helpful creative assistants, several works have shown that using an LLM as a creative partner results in a narrower set of creative outputs. However, these studies only consider the effects of interacting with a single LLM, begging the question of whether such narrowed creativity stems from using a particular LLM—which arguably has a limited range of outputs—or from using LLMs *in general* as creative assistants. To study this question, we elicit creative responses from humans and a broad set of LLMs using standardized creativity tests and compare the population-level diversity of responses. We find that LLM responses are much more similar to other LLM responses than human responses are to each other, even after controlling for response structure and other key variables. This finding of significant homogeneity in creative outputs across the LLMs we evaluate adds a new dimension to the ongoing conversation about creativity and LLMs. If today's LLMs behave similarly, using them as a creative partners—regardless of the model used—may drive all users towards a limited set of "creative" outputs.

1 INTRODUCTION

Large language models (LLMs) have moved out of research labs and into our everyday lives. Given their advanced abilities to generate text and respond to prompts, LLMs are often marketed as creativity support tools that allow users to write drafts, edit documents, and generate novel ideas with ease [2, 4, 20, 21]. Consumers have responded eagerly to these suggestions. According to a 2024 survey by Adobe, over half of Americans have used generative AI tools like LLMs as creative partners for brainstorming, drafting written content, creating images, or writing code. An overwhelming majority of LLM users surveyed believe these models will help them be more creative [39].

While appealing, outsourcing our creative thinking to LLMs could have unintended consequences and demands further scrutiny. For example, recent work has unearthed complications around the use of LLMs as creativity support tools. Researchers found that LLM-aided creative outputs look individually creative but are often quite similar to other LLM-aided outputs. Such "homogeneity" in LLM-aided creative outputs has been observed in a variety of settings, from creative writing to online survey responses to research idea generation and beyond [7, 16, 37, 43, 53].

While concerning, these works typically only look at a single LLM and it's effect on downstream creative content. In a prototypical example, Doshi and Hauser [16] compared the individual and collective creativity of two groups of writers—humans alone and humans aided by ChatGPT—and found that stories produced by the ChatGPT-aided group were more homogeneous. Related work from Moon, Green, and Kushlev [37] compared college essays written by humans and GPT models and found that LLM-authored essays contributed fewer new ideas and were more homogeneous than human-authored essays. However, such work begs the question: does the observed homogeneity occur because only a single type of LLM (GPT variants) is studied? It could be reasonably argued that a single LLM must have a limited range of outputs, causing the homogeneity. Perhaps if writers all used different LLMs, creativity would be restored.

Recent work studying feature space alignment in LLMs suggests otherwise. There is a long line of work measuring feature space similarity in machine learning models, since this is believed to indicate overall model similarity [8, 31, 33, 34, 46]. Some initial work has applied these techniques to large-scale LLMs and found evidence of "feature universality" in these models [26, 29, 32]. We postulate that such feature space alignment in LLMs may result in homogeneous

*Correspondence to: emily.wenger@duke.edu

creative outputs *across* these models. This would imply that the use of LLMs as creative partners *in general* leads to output homogeneity, because all LLMs would have limited and similar output ranges.

The consequences of cross-LLM homogeneity would be significant in the creative space and beyond. Humans who rely on LLMs as creative partners would find their creative outputs remarkably similar to those of other LLM users regardless of the model used, resulting in a collective narrowing of societal creativity. More broadly, homogeneity among widely used LLMs could lead to bias propagation, widespread security vulnerabilities, or other problems [11, 30].

Our Contribution. This work explores possible convergence in the creative outputs of large-scale LLMs. We test this by soliciting creative outputs from LLMs and humans using standardized creativity tests—the Alternative Uses Task [23], Forward Flow [22], and the Divergent Association Task [38]—and measuring the population-level variability of responses. While caution should be used in extrapolating human-centric psychological tests to non-human entities (see §3), these tests are useful in our setting because of their standardized output format. This allows us to disambiguate similarity in response structure from similarity in response content, the true goal. Our analysis shows that:

- Mirroring prior work [25], *LLMs match or outperform humans on standard tests of individual creativity.*
- Yet, this finding of individual creativity is misleading because *LLM responses to creative prompts are much more similar to each other than are human responses*, even after controlling for LLM "family" overlap and differences in human/LLM response structure.
- *Altering the LLM system prompt* to encourage higher creativity *slightly increases overall LLM creativity and inter-LLM response variability*, but human responses are still more variable.

Implications. We believe these findings highlight a potential danger of relying on generative AI models as creative partners. If today's most popular models exhibit a high degree of overlap in creative outputs, using any of them to aid creativity—as will happen if these models are integrated into platforms we regularly use for writing or creative thinking—could self-limit us from reaching the divergent creativity that defined artistic geniuses like Mozart, Shakespeare, and Picasso. Our set of AI "creative" partners will instead collectively drive us towards a mean.

2 RELATED WORK

Creativity, Homogeneity, and LLMs. Prior work has explored issues of creativity and homogeneity related to specific LLMs. Several works have compared human and LLM performance on standard creativity tests, typically using GPT models, and found that LLMs often outperform humans on these tests [12, 25, 45]. Despite LLMs' displays of individual creativity, numerous studies have shown that using LLMs to support creative tasks tends to homogenize creative outputs. For example, Doshi and Hauser [16] found that writers who used GPT-4 as a creativity support tool produced more creative stories than humans working alone, but the stories from writers who collaborated with GPT-4 were more similar to each other than were stories from human writers. This phenomenon of LLM-drive content homogenization appears across domains—in research idea generation [43], essay writing [37], survey responses [53], creative ideation [7], and art [55]. Recent work also showed that when GPT models are evaluated multiple times on creativity tests like the DAT, their responses tend to overlap, even if each individual response achieves a high "creativity" score [14]. Such findings further motivate our study of whether it is the use of specific models in these studies—often ChatGPT—that causes observed homogeneity, or if such homogeneity would be observed *regardless of the model used.*

Finally, a few works have considered issues of monoculture related to machine learning algorithms. Several works demonstrate suboptimal outcomes when multiple firms employ the same algorithm for decision-making [11, 30]. [52] proposed the term "generative monoculture" to describe the narrow distribution of LLM outputs relative to that of their

training data—an observation related to the creative narrowing observed in other work. However, none of these works considered similarity *across* models.

LLM Similarity. Numerous papers have worked to measure similarity between model feature representations, primarily in classifiers [8, 31, 33, 34, 46]. Such similarity is believed to indicate overall similarity between models and could lead to interesting downstream consequences, such as attack vectors that transfer between models (e.g. [15, 40, 47, 51] among many others). Nascent work applies similar methods to LLMs and finds evidence of "feature universality" across LLMs [29, 32]. Huh et al. [26] also measured feature space alignment between open-source LLMs and postulated that large models will inevitably become more similar over time. However, limited work has considered downstream consequences of LLM similarity. One work [27] examines question-answering bias of 10 LLMs across 4 "families", but finds little evidence of bias similarity among models. One paper demonstrates jailbreak attacks that transfer between LLMs [56] but does not specifically leverage LLM similarity in attack development.

This paper. We build on this prior work to study *creative output variability across LLMs*. Several works have shown that using *specific* LLMs as creative partners narrows the range of creative outputs [7, 14, 16, 35, 37]. We instead evaluate the diversity of responses to creative prompts across *many* LLMs using standard creativity tests and compare this to the diversity of human responses. We believe this study will enhance the current debate surrounding LLMs and creativity, clarifying whether it is the use of a specific LLM that homogenizes creative outputs or the use of LLMs in general.

3 METHODOLOGY

Our goal is to measure whether LLMs produce more, less, or equally diverse creative outputs as a group of humans. We measure this diversity (or variability) in responses by computing the semantic similarity among responses of humans and LLMs to prompts designed to elicit creativity. This section describes the creativity prompts we use, humans and LLMs tested, and evaluation metrics.

3.1 How do we elicit creative responses from LLMs?

The American Psychological Association defines creativity as "the ability to produce or develop original work, theories, techniques, or thoughts" [1]. Since our goal is to compare the diversity of creative responses from LLMs and humans, we sought out methods to elicit and compare creative outputs. Given the novelty of this field, no standard benchmarks exist for comparing LLM and human creativity. However, prior work has applied tests of divergent thinking in humans, which elicit qualities psychologists view as important to creativity, to LLMs and found that LLMs like ChatGPT scored similarly to humans [12, 25, 45, 54].

Creativity tests for humans. One of the original divergent thinking tests was Guilford's Alternative Uses Test (AUT) [23], which presents subjects with an object and asks them to describe creative uses for it. AUT responses are scored by measuring the number of different uses presented ("fluency"), the originality of those ideas ("originality"), how different they are from each other ("flexibility"), and the level of detail provided ("elaboration"). While effective, the AUT evaluation process is onerous, so researchers have developed more lightweight divergent thinking tests in recent years. One popular test, Forward Flow [22] (FF), measures the divergence of a user's chain of thought from a fixed starting point. Another, the Divergent Association Test (DAT) [38], asks subjects to list 10 unrelated words. Both capture similar characteristics to the AUT but with less burden on participants and evaluators.

Should we run human creativity tests on LLMs? Given our goals, it seems reasonable to test humans and LLMs using the AUT, FF, and DAT and then compare the population-level variability of their responses. However, it is an active