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

area of debate as to whether psychological tests—including those of creativity—designed for humans are applicable to LLMs. Some argue that works probing LLM performance on these tests is misguided due to fundamental differences between LLM and humans [14, 24, 44]. Future scholarship will inevitably continue this debate.

Our approach. Despite this disagreement, we believe creativity tests administered to LLMs can be useful for our purposes, *because we are not trying to measure inherent LLM creativity*. Instead, we would like to understand the variability of LLM and human responses to creative prompts, which is an empirical rather than psychological trait. Well-trained LLMs should respond similarly to factual questions but not necessarily to creative prompts, so if there is indeed homogeneity in LLM outputs, creativity is the appropriate lens through which to evaluate this question.

However, the question remains of what type of creative prompts to use. An obvious approach is asking LLMs to write short stories (similar to [16]) and comparing the variability of these to that of human-composed stories, but this approach has a notable caveat. We need to disentangle similarity in the *structure* of creative responses from similarity in their *content*—the latter matters significantly, but former does not. If LLMs share certain output quirks in generated text, like using the passive tense or gerunds, we do not want these to skew measurements of LLM response variability. Creativity tests like the AUT, DAT, and FF provide a helpful solution to this potential problem—they are designed to elicit creative outputs in a structured manner. Therefore, we use these tests in our analysis, but future work should consider other ways to elicit easily comparable creative responses from humans and LLMs.

3.2 Tests We Use

Based on the reasoning of §3.1, we compare the variability of human and LLM responses to the AUT, FF, and DAT tests. Exact test wording is in Appendix A.

Guilford's Alternative Uses Test (AUT) [23] presents people with an object and asks them to write down as many creative uses for it as they can think of. Following established best practices [9, 18], we test users with five common objects—book, fork, table, hammer, and pants. Using multiple starting objects reduces the effect of a particular object (e.g. book) on participant responses, ensuring results generalize [19]. It also allows us to collect more data, given the limited number of LLMs we can evaluate relative to the number of possible human subjects.

Forward Flow [22] measures how much a person's thoughts diverge from a given starting point. It provides a starting word and asks people to write down the next word that follows in their mind from the previous word for up to 20 words. We follow the original Forward Flow paper and run our study using five different start words: candle, table, bear, snow, and toaster. As in the AUT, providing multiple creative stimuli ensures results generalize and gives us more data.

The Divergent Association Task (DAT) [38] asks subjects to list 10 words that are as unrelated as possible. These are subject to certain constraints: only nouns, no proper nouns, only single words in English, and the task must be completed in less than four minutes. The DAT provides a limited amount of information compared to the other tests, since the creative stimulus cannot be varied.

3.3 Test subjects

We administer these tests to a set of LLMs and a set of humans, following IRB-approved user study protocol.

Large Language Models. As a baseline, we test 22 large language models with public APIs[1]: *AI21-Jamba-Instruct, Cohere Command R, Cohere Command R Plus, Meta Llama 3 70B Instruct, Meta Llama 3 8B Instruct, Meta Llama 3.1 405B Instruct, Meta Llama 31 70B Instruct, Meta Llama 3.1 8B Instruct, Mistral large, Mistral large 2407, Mistral Nemo, Mistral*

---

[1]https://docs.github.com/en/github-models/prototyping-with-ai-models

small, Google Gemini 1.5, gpt 4o, gpt 4o mini, Phi 3 medium 128k instruct, Phi 3 medium 4k instruct, Phi 3 mini 128k instruct, Phi 3 mini 4k instruct, Phi 3 small 128k instruct,Phi 3 small 8k instruct, and *Phi 3.5 mini instruct.* Models in the same "family" (e.g. all Llamas, all GPTs, etc) may generate unusually similar responses due to similarities in architecture, training data, or optimization techniques. To control for this, we restrict ourselves following subset of models when conducting statistical tests, which contains only one model from each "family": *AI21 Jamba 1.5 Large* [49], *Google Gemini 1.5* [48], *Cohere Command R Plus* [3], *Meta Llama 3 70B Instruct* [17], *Mistral Large* [28], *gpt 4o* [6], and *Phi 3 medium 128k Instruct* [5]. All these models were trained by distinct entities, providing a reasonable independent baseline. In §5, we also explore how models with the same "family" behave. For these experiments, we use the Llama model family [17]: *Meta Llama 3 70B Instruct, Meta Llama 3 8B Instruct, Meta Llama 3.1 405B Instruct, Meta Llama 31 70B Instruct, Meta Llama 3.1 8B Instruct.* We evaluate all models with the default system prompt of "You are a helpful assistant" but explore the effect of varying this in §5. After obtaining model responses, we remove unnecessary punctuation (e.g. numbered DAT outputs).

Human subjects. We use two sources of human responses as a ground truth set for human creativity. First, we run an IRB-approved user study[2]. Study subjects were recruited from the Prolific platform were asked to complete the DAT, FF, and AUT creativity tests (see Appendix A for study wording). It took participants 19 minutes on average to complete the survey, and they were compensated at a rate of $15/hour. Participant demographics are described in Table 1. All patients completed a consent form before starting the study. We recruited 114 initial participants from the Prolific platform, screening for English fluency and an approval rating > 95. Of these, 12 were removed on suspicion of being bots due to unusually short response times (< 5 minutes) or failed attention checks, so the final dataset contains 102 human responses. Authors manually inspected responses to correct obvious misspellings.

Age		Gender		Race	
18-24	22%	Female	51%	Asian	6%
25-34	31%	Male	46%	Black or African American	28%
35-44	23%	Non-Binary	3%	Hispanic or Latino or Spanish Origin of any race	11%
45-54	19%			White	53%
55+	5%			Other	2%

Table 1. *Demographics of human study participants.*

The risk in relying on responses from online crowd workers is that they may themselves be bots or may leverage LLMs in crafting their responses, resulting in reduced response diversity [53]. Prolific runs strict tests to ensure human responses, and we also used safeguards to prevent this, including attention tests and post-hoc data inspection. However, the risk remains. Therefore, we use public datasets of human responses to the AUT [3], FF [4], and DAT [5] from prior work as a secondary validation dataset. These data are from creativity tests run in person before the rise of LLMs (around 2022), so they are unlikely to contain LLM responses. However, given their public nature, these datasets may have been used to train LLMs, resulting in unusual similarity between LLM and these human responses. We use data collected in our user study for our main analysis in §4 but re-run population-level originality tests with this data in §5 for validation.

---

[2]IRB information redacted for anonymous submission
[3]https://osf.io/u3yv4
[4]https://osf.io/7p5mt/
[5]https://osf.io/kbeq6/

3.4 Evaluation Metrics

The primary goal of this study is to evaluate the *variability* of LLM responses to creative prompts relative to that of humans. To do this, we compute the semantic similarity of responses in different populations (LLMs vs. humans) and compute distributional differences in similarity scores between populations. As a baseline, we also compare the *originality* of individual LLM responses to the tests relative to that of humans.

3.4.1 Scoring individual originality. Although divergent thinking tests can be measured using multiple metrics, it has long been argued the *originality* of responses is the strongest indicator of creativity [36]. Originality—how novel tests responses are relative to the given prompt(s)—can also easily be measured in an automated fashion by embedding prompts and responses in a mathematical feature space and measuring the cosine distance between the feature vectors [10]. Prior work confirms that such automated analysis closely matches originality rankings of human scorers [19].

Our metrics for individual originality follow the guidelines of the original studies but use the automated evaluation methods of [19], including the use of the GloVe 840B model [41] to compute word embeddings. The format of each test necessitates different originality scoring procedures, described in detail in Appendix §B. Originality scores are denoted as $O_t(\mathcal{P})$, where t = AUT, FF, or DAT and $\mathcal{P}$ is a population, either humans or LLMs.

Distributional differences. After computing originality scores, we can then compare the distributions of $O_t(LLM)$ and $O_t(Humans)$ to measure differences in originality between the two groups. We do this by testing for statistically significant differences in $\mu(O_t(LLM))$ and $\mu(O_t(Human))$ using Welch's t-test to compare differences in means, since the populations typically do not exhibit equal variance. We use a statistical significance threshold of $\rho = 0.01$. For all tests, the null hypothesis is that $\mu(O_t(LLM)) = \mu(O_t(Human))$, and the alternative is that $\mu(O_t(LLM)) > \mu(O_t(Human))$.

3.4.2 Scoring population-level variability. We measure the variability in responses to the creativity tests from a given population by computing the semantic distances between sets of responses from individuals in the population (e.g. comparing the set of AUT uses produced by an LLM to that of another LLM). If many population members given semantically similar sets of answers, this indicates that the response variability of the population is low, and vice versa if it is high. We denote the variability of a population $\mathcal{P}$ on test t as $\mathcal{V}_t(\mathcal{P})$, the set of all similarity scores between all responses from all population members. As before, $\mathcal{P}$ refers to either LLMs or humans.

We use a sentence embedding model $\mathcal{S}$ to measure semantic similarity between responses. Sentence embedding models map sentences or short paragraphs to feature vectors and, similar to the word embedding model, map similar content to similar feature vectors. We compute elements of $\mathcal{V}_t(\mathcal{P})$ by representing an individual's responses to a certain test condition (e.g. all their AUT responses to a certain prompt) as a single, space-separated word string $\mathbf{R}$ and embedding this into a mathematical space via $\mathcal{S}$, producing $\mathcal{S}(\mathbf{R})$. We then take the cosine similarity between this vector and those of other population members to form $\mathcal{V}_t(\mathcal{P})$:

$$\mathcal{V}_t(\mathcal{P}) = \left\{ 1 - \cos(\mathcal{S}(\mathbf{R}_i^p), \mathcal{S}(\mathbf{R}_j^p)), \forall \, (\mathbf{R}_i^p, \mathbf{R}_j^p, p)_{i \neq j} \in \mathcal{P} \right\} \tag{1}$$

where $\mathbf{R}_i^p, \mathbf{R}_j^p$ denote the responses of two different population members to prompt p. In our experiments, we use `all-MiniLM-L6-v2` from the `sentence_transformers` Python library [42], a high-performing and widely used model, to compute sentence embeddings. We remove punctuation and stopwords from responses before computing embeddings.

$\mathcal{V}_t(\mathcal{P})$ is composed of cosine distance scores, so if it skews towards 0, responses in the population are similar to each other. If it skews towards 1, they are more different, and therefore the population exhibits higher variability. Note that $\mathcal{V}_t(\mathcal{P})$ only contains similarity scores of responses from *different* LLM/human subjects.

Distributional differences. We can then compare the statistical distributions of $\mathcal{V}_t(LLM)$ and $\mathcal{V}_t(Humans)$ to measure the relative response variability between these groups. We do this using the same statistical tests from §3.4.1. For all tests, the null hypothesis is that $\mu(\mathcal{V}_t(LLM)) = \mu(\mathcal{V}_t(Human))$, and the alternative is that $\mu(\mathcal{V}_t(LLM)) > \mu(\mathcal{V}_t(Human))$.

4 KEY RESULTS

When reporting results of statistical t-tests, we use the standard APA format, reporting the degrees of freedom (DOF), test statistic X, and significance level y: $t(\text{DOF}) = X, p = y$. For context, we also report the effect size, which is the difference between the means of the two populations divided by their pooled standard deviation. Cohen [13] defines small, medium, and large effect sizes as 0.2, 0.5, and 0.8, respectively. Finally, we report test power, which is the probability of correctly rejecting the null hypothesis (or 1 minus the probability of a false negative).

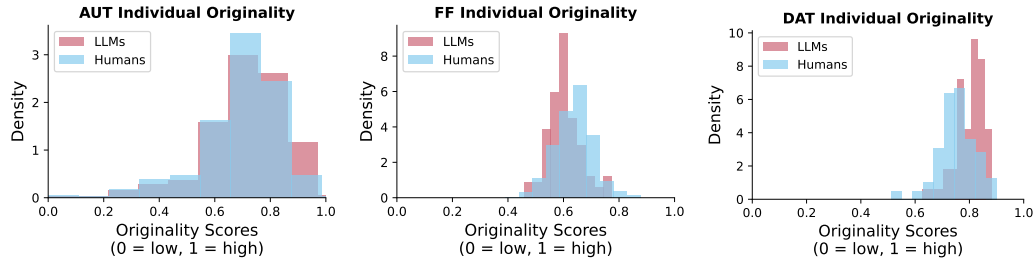### 4.1 Baseline measurement: individual originality in LLMs vs. humans

Fig. 1. LLMs slightly outperform humans on the AUT and DAT, but humans slightly outperform LLMs on FF.

LLMs score slightly higher than humans on the AUT and DAT tasks, mirroring prior work [25], but perform worse on FF. Figure 1 shows the distributions of originality scores for humans and LLMs on these tests, and Table 1 gives statistics comparing population means for the two groups. Overall, these results show that LLMs and humans exhibit roughly equal levels of measured originality on these tests on average, removing this as a possible confounding variable in our study of response variability.

Test	$\mu(O_t(\text{LLM}))$	$\mu(O_t(\text{Human}))$	Test statistic	p-value	Effect size	Test power
AUT	0.711	0.696	$t(2094) = -3.4$	0.001	0.1	0.84
FF	0.603	0.637	$t(164) = 5.2$	$2.9e^{-07}$	0.52	0.99
DAT	0.801	0.753	$t(159) = -5.12$	$8.7e^{-07}$	0.77	0.99

Table 2. **LLMs slightly outperform humans on the AUT and DAT, but humans slightly outperform LLMs on FF.** *However, the effect size for these is relatively small, confirming results from prior work showing relatively similar performance between humans and LLMs on creativity tests. Null hypothesis is* $\mu(O_t(LLM)) = \mu(O_t(Human))$; *alternative is* $\mu(O_t(LLM)) > \mu(O_t(Human))$.

4.2 Population-level Response Variability—LLMs vs. Humans

Now, we explore the main question: whether LLMs and humans exhibit different *population-level* variability in creative outputs. For statistical analysis and $\mathcal{V}_t$ distribution plots in this setting, we only consider responses from 7 distinct LLMs: *AI21 Jamba 1.5 Large*, *Google Gemini 1.5*, *Cohere Command R Plus*, *Meta Llama 3 70B Instruct*, *Mistral Large*, *gpt*
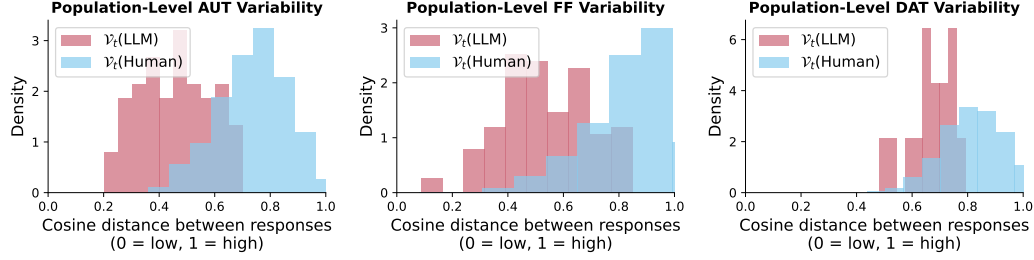
Fig. 2. LLM responses exhibit far less variability than human responses, *as measured by cosine distance between embedded responses.*

Test	$\mu(\mathcal{V}_t(\text{LLM}))$	$\mu(\mathcal{V}_t(\text{Human}))$	Test statistic	p-value	Effect size	Test power
AUT	0.459	0.738	$t(10078) = 19.1$	$3.9e^{-80}$	2.2	1.0
FF	0.534	0.835	$t(90) = 26.1$	$2.8e^{-66}$	2.0	1.0
DAT	0.665	0.819	$t(30) = 9.9$	$6.2e^{-11}$	1.4	1.0

Table 3. Across all tests, LLMs have significantly lower mean population-level variability than humans. *Null hypothesis is that $\mu(\mathcal{V}_t(LLM)) = \mu(\mathcal{V}_t(Human))$; alternative is that $\mu(\mathcal{V}_t(LLM)) > \mu(\mathcal{V}_t(Human))$. The difference is statistically significant for all tests.*

4o, and *Phi 3 medium 128k Instruct*, a subset of our original 22 models. As discussed previously, this choice removes model family as a possible confounding variable in our analysis.

Our key finding is that *LLM responses exhibit much less variability, as measured by semantic distance between pairs of embedded responses, than do human responses.* Table 3 gives statistics, while Figure 2 shows the distributions of $\mathcal{V}_t(\text{LLM})$ and $\mathcal{V}_t(\text{Human})$, e.g. cosine distances between responses in these respective populations. Both these views of the data confirm that LLM test responses are much more similar to each other than human responses are to each other. From this, we conclude that a population of LLMs produces more homogeneous outputs in response to creative prompts than does a population of humans.

Visualizing embedded responses. To further understand the overlap in LLM responses as compared to humans, we visualize the sentence embeddings of AUT responses in Figure 3 (visualizations for FF and DAT are in Appendix C). To do this, we perform t-distributed stochastic neighbor embedding (TSNE) [50] analysis of the embeddings, which allows visualization of high-dimensional data (384 in our case) in two dimensions. We then perform k-means clustering on the t-SNE results to identify sets of responses corresponding to the same AUT prompt object—pants, table, etc.—and color the data accordingly. This visualization confirms the behavior observed statistically: LLM responses "cluster" together in the embedded feature space, providing further evidence of low LLM response variability.

One explanation: word overlap in LLM responses. The low response variability of LLMs can be partially explained through analysis of lexical patterns in LLM and human responses. We remove stopwords from responses, then count the number of word overlaps between sets of responses from LLMs and humans—all AUT uses from a human/LLM, all words in a FF response, etc. As Figure 4 shows, LLM responses tend to have many more words in common than human responses, across all tests. This overlap at least partially accounts for the high semantic similarity between LLM responses, as the sentence embedding model will map responses with overlapping words to similar feature vectors. Further exploration of differences in lexical patterns between LLMs and humans is important future work.
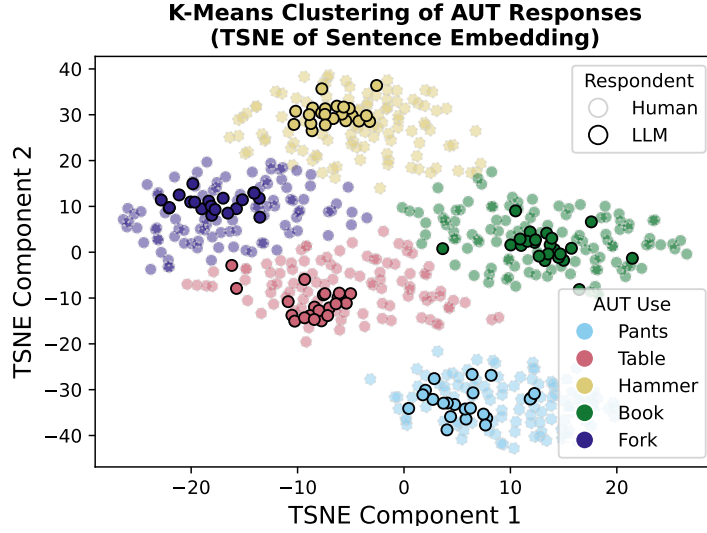
Fig. 3. LLM responses cluster together in feature space more than do human responses. *K-means clustering of TSNE of AUT sentence embeddings.*
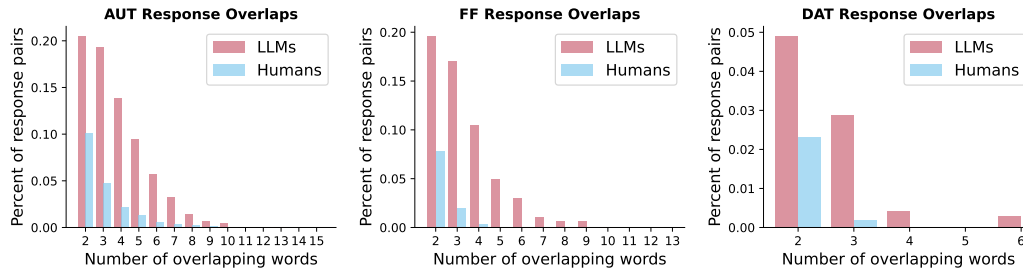

Fig. 4. LLM responses have far more words in common than do human responses. *We look at word overlaps between "full" responses from LLMs and humans—e.g. all uses from the AUT, all words in the FF, etc. This corresponds to the sentence embedding method of population originality measurement, and explains why the difference between LLMs and humans is more pronounced in this setting.*

5 ADDITIONAL ANALYSIS

Having established that LLMs produce more homogeneous creative outputs than humans, we now explore several additional dimensions of this key finding. First, we demonstrate that this cross-LLM response homogeneity remains even after strictly controlling for structural differences in human and LLM responses. Next, we measure if homogeneity increases when LLMs all come from the same "family." Then, we explore a possible mechanism to counteract LLM creative homogeneity through the use of creative system prompts. Finally, we confirm that our human user study results are similar to prior results, ensuring that the choice to conduct our survey online does not skew results. Throughout this section, we consider only responses to the AUT to avoid a combinatorial explosion of experiments.

5.1 Controlling for AUT Response Structure

For both the DAT and FF tests, the response structure is fixed, making comparison of population-level variability straightforward. However, the AUT is more open-ended, so confounding variables such as differences in response structure (e.g. number of words, tense, etc.) between LLMs and humans may impact measurements of response variability.
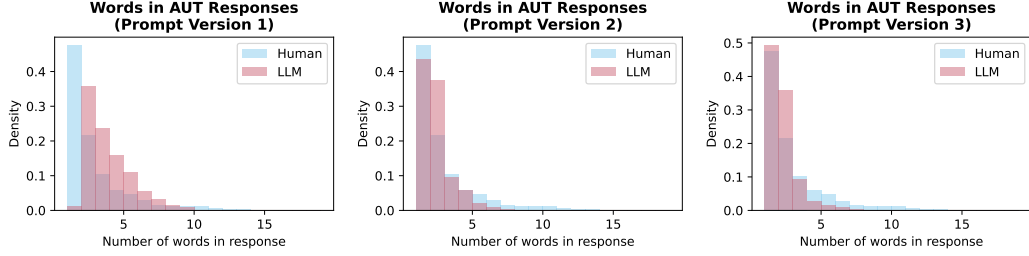
Fig. 5. Effect of different AUT prompt wordings on length of LLM AUT responses. *We use prompt verison 3 in most experiments in this paper, since LLM responses to this prompt most closely match the human distribution of response lengths.*

For example, if every LLM AUT response is 4 words long and uses a gerund (e.g. "making", "writing"), the measured similarity between LLM responses may be artificially inflated. Since we want to measure variability in the *substance* rather than the *structure* of responses, we must ensure that structural similarity does not impact our findings. Here, we demonstrate that the observed population-level difference in LLM and human response variability remains even after controlling for AUT response structure.

Problem: differences in LLM and human AUT response lengths. In all experiments, we remove stop words, whitespace, and punctuation from responses before analysis. However, we observed that the first version of our AUT LLM prompt caused models to return more verbose AUT uses on average than humans. The base version of the prompt (version 1) included the phrase: "Please list the uses as short sentences or phrases, separated by semicolons, so the list is formatted like 'write a poem; fly a kite; walk a dog'." This phrase was intended to standardize the format of LLM outputs to minimize data cleaning. However, as the left element of Figure 5 shows, it caused LLM AUT responses to be longer on average than human responses. This could impact measurements of population variability, since LLM responses could be measured as "more similar" simply because they contain more words than human responses.

Solution: prompt engineering. To remove this confounding variable, we performed prompt engineering to more closely align the distribution of words in LLM responses to that of humans. Version 2 of our AUT prompt directed models to: "Please list the uses as words or phrases (single word answers are ok), separated by semicolons, so the list is formatted like 'write; fly a kite; walk dog'." As the middle element of Figure 5 shows, this shifted the LLM AUT word count distribution closer to that of humans. Version 3 of our AUT prompt read: "Please list the uses as words or phrases (single word answers are ok), separated by semicolons." The right graph of Figure 5 shows that prompt 3 caused LLMs to return roughly the same proportion of single-world answers as humans while reducing the number of two-word answers. Since promot 3 most closely matches human response word counts, we use it in §4.

AUT Prompt Version	$\mu(O_t(\text{LLM}))$	$\mu(O_t(\text{Human}))$	Test statistic	p-value	Effect size	Test power
v1	0.744	0.696	$t(2094) = -11.8$	$8.3e^{-32}$	0.35	1.0
v2	0.715	0.696	$t(2094) = -4.04$	$2.7e^{-05}$	0.14	0.97
v3	0.711	0.696	$t(2094) = -3.4$	0.001	0.1	0.84

Table 4. For all AUT prompt versions, LLMs have slightly higher AUT originality scores than humans. *Null hypothesis is $\mu(O_t(LLM)) = \mu(O_t(Human))$; alternative is $\mu(O_t(LLM)) > \mu(O_t(Human))$.*

Analysis: effect of AUT response structure on creativity. Next, we analyze how the different AUT prompts and resulting LLM response structure affect measurements of creativity and variability. We run the same statistical tests as

AUT Prompt Version	$\mu(\mathcal{V}_t(\text{LLM}))$	$\mu(\mathcal{V}_t(\text{Human}))$	Test statistic	p-value	Effect size	Test power
v1	0.427	0.738	$t(10102) = 24.5$	$1.0e^{-128}$	2.5	1.0
v2	0.466	0.738	$t(10053) = 15.2$	$5.1e^{-52}$	2.2	1.0
v3	0.459	0.738	$t(10078) = 19.1$	$3.9e^{-80}$	2.2	1.0
v3 (one-word answers)	0.708	0.850	$t(10078) = 8.9$	$2.3e^{-19}$	1.1	1.0

Table 5. **Even after controlling for AUT response structure via prompt engineering and manual filtering, the LLMs' population-level variability is much lower than that of humans.** *Null hypothesis is that $\mu(\mathcal{V}_t(\text{LLM})) = \mu(\mathcal{V}_t(\text{Human}))$; alternative is that $\mu(\mathcal{V}_t(\text{LLM})) > \mu(\mathcal{V}_t(\text{Human}))$. "v3 (one-word answers)" means that we only considered single-word AUT responses from humans and LLMs responding to prompt v3. $\mathcal{V}_t(\mathcal{P})$ from the last row (v3, one word answers) are plotted in Figure 6 to visualize the shift in means observed in this setting.*
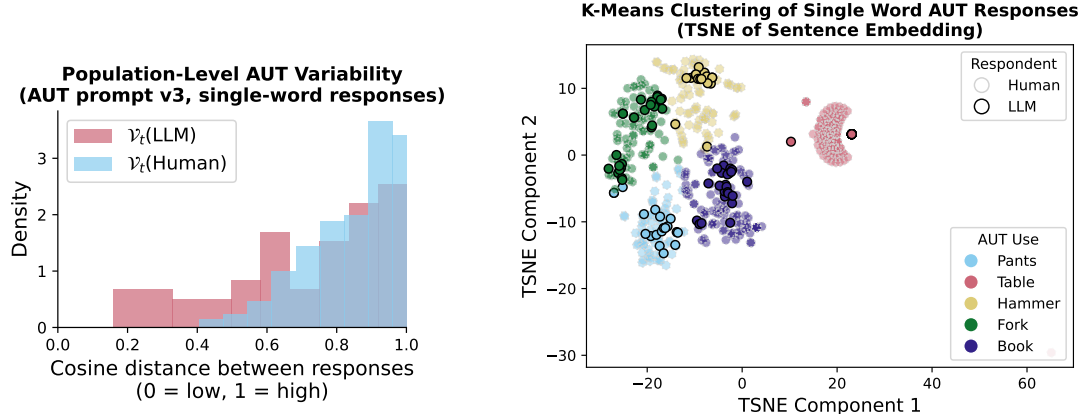

Fig. 6. **Even when considering only one-word responses to control for response structure, LLM AUT responses have lower population-level variability (left plot) and are closer in feature space (right plot) than human responses**. *LLM responses are generated with prompt version 3. We create sentence embeddings from only single-word uses provided by AUTs and humans, ignoring all longer responses.*

in §4 to measure individual and population-level creativity when using these three prompt versions. Table 4 shows that for all prompt versions, LLMs exhibit slightly higher individual creativity than humans. The creativity levels are closest for prompt version 3, supporting that this is a reasonable setting for measuring population creativity.

Table 5 shows statistics for population-level variability across the three prompt versions, including a variant of prompt 3 where we only consider single-word uses (more details on this in Figure 6). The goal of the single word setting is to completely eliminate confounding effects of AUT response structure on creativity measurements, providing the closest possible comparison between humans and LLMs. As Table 5 shows, *LLMs exhibit consistently lower response variability than humans, even after controlling for response structure.* This effect remains across all 3 prompt versions. LLM response variability scores increase slightly when moving from prompt version 1 to 3, indicating that response structure has a (small) effect on variability measurements. However, having controlled effectively for this variable via prompt engineering and detailed analysis, we are confident that it is the *substance*, not the *structure* of LLM AUT responses that reduces their population-level variability. That response variability remains low on FF and DAT—for which response structure does not matter—further confirms this finding.

5.2 Creativity within LLM "families"

Next, we inspect whether models in the same "family" produce more homogenous responses than a baseline set of otherwise unrelated models. To do this, we measure the population-level variability of AUT responses from Llama model family: *Meta Llama 3 70B Instruct*, *Meta Llama 3 8B Instruct*, *Meta Llama 3.1 405B Instruct*, *Meta Llama 31 70B*
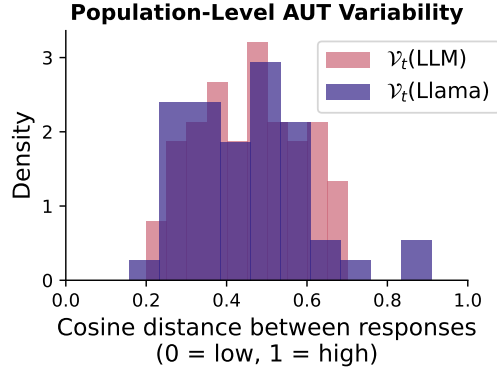
Fig. 7. Models from the same family (Llama) exhibit slightly lower population-level variability than models from different families.

Instruct, and *Meta Llama 3.1 8B Instruct*. Given the small number of models we are comparing, we add additional AUT startwords to increase dataset size. These start words, modelled on prior AUT studies [19], are: book, bottle, brick, fork, hammer, pants, shoe, shovel, table, and tire. Figure 7 shows population-level AUT originality distributions for unrelated LLMs vs. Llama models, and Table 6 presents statistics comparing these distributions.

$\mu(\mathcal{V}_t(\text{LLM}))$	$\mu(\mathcal{V}_t(\text{Llama}))$	Test statistic	p-value	Effect size	Test power
0.445	0.441	$t(248) = 0.2$	0.41	0.02	0.01

Table 6. **Models from the same "family" (Llama) have lower population-level creativity than models from different familes.** *This is clearly seen in the leftward distribution shift of the Llama population differences compared to the all model population differences (see Figure 7). However, because the Llama distribution is right-skewed, the distribution shift is not captured in a t-test for differences of means in the sentence embedding case.*

Models in the same "family" exhibit slightly lower response diversity than models from different "families." Although the differences of means is not statistically significant, from visual inspection of Figure 7, we see that the sentence embedding distribution for Llama models is skewed right. The presence of these outliers drives up the overall Llama population mean, making it appear more similar to that of the regular LLMs. Future work should consider other model families and explore other dimensions of family-specific similarity.

5.3 Effect of LLM system prompt.

Next, we consider ways to make LLMs produce more variable outputs. As a baseline, we explore whether varying the LLM system prompt to strictly request creative outputs will induce higher variability. We experiment with prompts designed to elicit different levels of creativity:

- **Baseline**: "You are a helpful assistant."
- **More creative**: "You are a creative assistant that always provides answers that demonstrate imaginative, outside-the-box thinking."
- **Very creative**: "You are a creative assistant that always provides answers that demonstrate imaginative, outside-the-box thinking. You are about to take a creativity assessment, and your answers should be as novel, original, and bold as possible. If you receive the highest score on this creativity assessment, you will receive $200."
- **Not creative**: "You are a robot assistant that always provides answers that are unoriginal, bland, and soulless. You are about to take a creativity assessment, and your answers should be as generic and unoriginal as possible."

Prompt	Individual creativity		Population-level variability	
	$\mu(O_{AUT}(\mathcal{P}))$	$t(df) = X, p$ (vs. humans)	$\mu(\mathcal{V}_{AUT}(\mathcal{P}))$	$t(df) = X, p$ (vs. humans)
Humans	0.695	-	0.738	-
Baseline	0.711	$t(2094) = -3.4, 0.001$	0.459	$t(10078) = 19.1, 3.9e^{-80}$
More creative	0.733	$t(5020) = -9.8, 1.0e^{-22}$	0.503	$t(10078) = 16.1, 3.5e^{-58}$
Very creative	0.754	$t(5206) = -15.9, 3.2e^{-56}$	0.576	$t(10078) = 11.1, 5.6e^{-29}$
Not creative	0.702	$t(2507) = -1.28, 0.1$	0.492	$t(10078) = 16.8, 1.1e^{-62}$

Table 7. Varying the system prompt slightly increases LLM individual creativity and response variability, *but variability remains far lower than that of humans.*
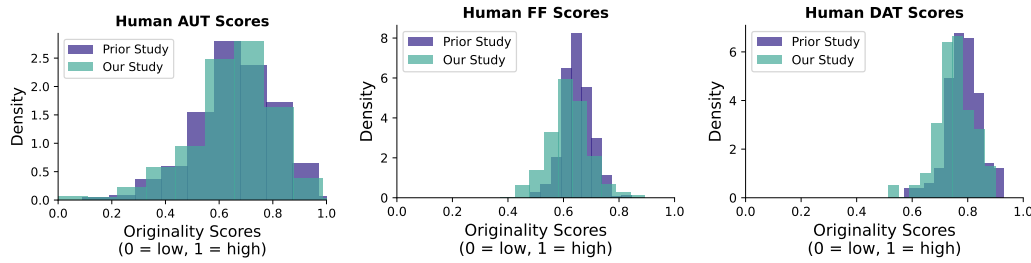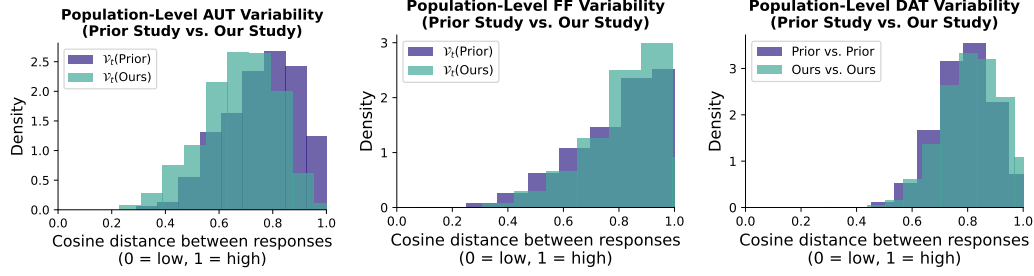

Fig. 8. Humans in prior studies have higher individual originality scores than humans in our study for all three tests. *For the AUT and DAT tests, a t-test for a difference in means (alternative hypothesis is that prior study has higher mean than ours) is significant at the* 0.01 *level but not the* 0.001 *level: AUT has* $t(5064) = 3.21, p = 0.001$ *and DAT has* $t(206) = 3.32, p = 0.001$. *For FF, the difference more significant:* $t(892) = 6.91, p < 0.0001$.

We evaluate the same subset of LLMs from §4 on the AUT using these system prompts and report summary statistics in Table 7. As the Table shows, using more creative system prompts slightly increases individual creativity for LLMs (and vice versa for the less creative prompt). However, the system prompt does not substantially improve LLM response variability—across all prompts, LLM variability remains much lower than that of humans.

5.4 Validation with preexisting survey data

Finally, we compare responses in our user study to prior user studies to ensure that our human subject pool is reliable and not unduly skewed by possible use of LLMs. We test both the individual originality of our human responses and population-level variability and find that while *respondents in prior studies score better individually on the tests, respondents to our study exhibit equal or greater population-level variability* (the more important metric for our study) on the more-informative AUT and FF tests.

Figure 8 compares individual creativity results for our study ($n = 102$) to that of prior studies ($n = 141$ for DAT, $n = 92$ for AUT, $n = 146$ for FF). T-tests for differences of individual performance (see caption of Figure 8) find that the mean score is higher for prior studies on all tests at a significant level of $p \leq 0.001$. Figure 9 compares the response variability of our study to that of the prior study. Using a t-test for difference in population means, we find that responses in our study have slightly higher variability on the FF and DAT ($p < 0.0001$), and lower on the AUT ($p < 0.001$). From this, we conclude that, our results roughly mirror those of prior studies, making them a reasonable baseline for our analysis.

Fig. 9. Our study responses have higher population-level variability than the prior study on the DAT and FF tests, but slightly lower variability on the AUT. *We use t-tests to compare means of the two population-level originality distributions. Null hypothesis is that means are equal, and alternative is that they are not. For AUT, the prior study has higher mean variability, $t(1046) = 11.0$, $p < 0.001$. For FF, our study has a higher mean, $t(366689) = -28.9$, $p < 0.001$. For the DAT, our study also has higher mean, $t(19832) = -19.6$, $p < 0.001$.*

6 DISCUSSION

Motivated by measured homogeneity in creative outputs produced by specific LLMs and observed feature space overlap in LLMs, we study whether responses to creative prompts produced by a group of LLMs exhibit more, less, or equal variance as a set of human responses to the same creative prompts. We find that LLMs exhibit *much* lower population-level output variability than humans, even after controlling for potential model similarities and structural differences between LLM and human responses. Our work upholds prior work showing that LLMs perform well on tests of divergent thinking but adds the nuance that such performance is homogeneous—LLMs return a narrower range of responses to creative prompts than humans. This result enhances prior observations of LLM-induced homogeneity, which only considered the effect of specific LLMs on creative outputs, and suggests that the use of LLMs *in general* may homogenize creative outputs.

Implications. These results have significant implications if LLMs are widely adopted as creativity support tools for writing, idea generation, or similar tasks. If all LLMs respond similarly to specific creative requests, then the population of users leveraging to LLMs to aid creativity will converge towards a limited set of creative outputs. In other words, LLM users may be self-limited from being exhibiting the divergent creativity that defined well-recognized artistic geniuses like Tolkein, Mozart, and Picasso because their LLM "creative" partners may collectively drive them towards a mean.

Limitations. Our work has several limitations. First, while we have demonstrated LLM homogeneity in response to certain creativity tests, this does not prove that LLMs in general produce homogeneous outputs when asked to behave creatively. It merely provides a indication that future work should explore this subject. Additionally, we measure a single metric of divergent thinking or creativity—originality, as measured by semantic similarity between responses—and finds that LLms are homogeneous along this dimension. However, there are other well-known metrics of divergent thinking, such as flexibility, fluency, and elaboration (see §3.1), and LLMs may demonstrate more or less homogeneity along these dimensions. Future work should consider these alternatives.

Acknowledgments. We thank Austin Liu for helping us design the system prompts of §5.3.

7 ETHICAL CONSIDERATIONS

We took care to ensure the user study in this paper was conducted in accordance with ethical standards. IRB approval for the study was obtained, and participants signed a clearly written consent form before completing our survey. To ensure privacy, participant data was anonymized and stored on secure servers. Other ethical risks from this paper are minimal, as our LLM experiments do not involve sensitive data and elicit only benign model responses.

REFERENCES

[1] 2018. APA Dictionary of Psychology - Creativity. https://dictionary.apa.org/creativity.

[2] 2024. Apple Intelligence | Writing Tools | iPhone 16. https://www.youtube.com/watch?v=3m0MoYKwVTM.

[3] 2024. Command R and Command R Plus Model Card. https://docs.cohere.com/docs/responsible-use.

[4] 2024. Use Notion AI to write better, more efficient notes and docs. https://www.notion.com/help/guides/notion-ai-for-docs.

[5] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).

[6] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[7] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*. 413–425.

[8] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. 2021. 'Revisiting model stitching to compare neural representations. *Proc. of NeurIPS* (2021).

[9] Baptiste Barbot. 2018. The dynamics of creative ideation: Introducing a new assessment paradigm. *Frontiers in psychology* (2018).

[10] Roger E Beaty, Paul J Silvia, Emily C Nusbaum, Emanuel Jauk, and Mathias Benedek. 2014. The roles of associative and executive processes in creative cognition. *Memory & cognition* (2014).

[11] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems* 35 (2022), 3663–3678.

[12] Honghua Chen and Nai Ding. 2023. Probing the Creativity of Large Language Models: Can models produce divergent semantic association? (Oct. 2023). http://arxiv.org/abs/2310.11158

[13] Jacob Cohen. 2016. A power primer. (2016).

[14] David Cropley. 2023. Is artificial intelligence more creative than humans?: ChatGPT and the divergent association task. *Learning Letters* 2 (2023), 13–13.

[15] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. 2019. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*. 321–338.

[16] Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10, 28 (July 2024). https://doi.org/10.1126/sciadv.adn5290

[17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[18] Denis Dumas and Kevin N Dunbar. 2014. Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity* (2014).

[19] Denis Dumas, Peter Organisciak, and Michael Doherty. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts* (2021).

[20] Matt Ellis. 2024. How to Use AI to Enhance Your Storytelling Process. https://www.grammarly.com/blog/writing-with-ai/ai-story-writing/.

[21] Google. 2024. Google + Team USA - Dear Sydney. https://www.youtube.com/watch?v=NgtHJKn0Mck.

[22] Kurt Gray, Stephen Anderson, Eric Evan Chen, John Michael Kelly, Michael S Christian, John Patrick, Laura Huang, Yoed N Kenett, and Kevin Lewis. 2019. "Forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist* 74, 5 (2019), 539.

[23] Joy Paul Guilford, Paul R Christensen, Philip R Merrifield, and Robert C Wilson. 1978. Alternate uses. (1978).

[24] Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of llm personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 301–314.

[25] Kent F Hubert, Kim N Awa, and Darya L Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports* 14, 1 (2024), 3440.

[26] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987* (2024).

[27] Hyejun Jeong, Shiqing Ma, and Amir Houmansadr. 2024. Bias Similarity Across Large Language Models. *arXiv preprint arXiv:2410.12010* (2024).

[28] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7B (2023). *arXiv preprint arXiv:2310.06825* (2023).

[29] Max Klabunde, Mehdi Ben Amor, Michael Granitzer, and Florian Lemmerich. 2023. Towards Measuring Representational Similarity of Large Language Models. In *UniReps: the First Workshop on Unifying Representations in Neural Models*.

[30] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021), e2018340118.

[31] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*. PMLR, 3519–3529.

[32] Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2024. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981* (2024).

[33] Karel Lenc and Andrea Vedaldi. 2015. Understanding image representations by measuring their equivariance and equivalence. In *Proc. of CVPR*.

[34]   Ruofan Liang, Tianlin Li, Longfei Li, Jing Wang, and Quanshi Zhang. 2020. Knowledge Consistency between Neural Networks and Beyond. arXiv:1908.01581 (2020). http://arxiv.org/abs/1908.01581

[35]   Kelsey Medieros, David H Cropley, Rebecca L Marrone, and Roni Reiter-Palmon. [n. d.]. Human-AI Co-Creativity: Does ChatGPT make us more creative? ([n. d.]).

[36]   Sarnoff Mednick. 1962. The associative basis of the creative process. *Psychological review* (1962).

[37]   Kibum Moon, Adam Green, and Kostadin Kushlev. 2024. Homogenizing Effect of Large Language Model (LLM) on Creative Diversity: An Empirical Comparison of Human and ChatGPT Writing. (2024).

[38]   Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences* 118, 25 (2021), e2022340118.

[39]   Vivek Pandya. 2024. The Age of Generative AI: Over half of Americans have used generative AI and most believe it will help them be more creative. *Adobe* (2024). https://blog.adobe.com/en/publish/2024/04/22/age-generative-ai-over-half-americans-have-used-generative-ai-most-believe-will-help-them-be-more-creative.

[40]   Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277* (2016).

[41]   Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[42]   Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of EMNLP*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[43]   Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109* (2024).

[44]   Massimo Stella, Thomas T Hills, and Yoed N Kenett. 2023. Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences* 120, 43 (2023), e2312911120.

[45]   Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting GPT-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932* (2022).

[46]   Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. 2023. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018* (2023).

[47]   C Szegedy. 2014. Intriguing properties of neural networks. *Proc. of ICLR* (2014).

[48]   Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[49]   Jamba Team, Barak Lenz, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, et al. 2024. Jamba-1.5: Hybrid Transformer-Mamba Models at Scale. *arXiv preprint arXiv:2408.12570* (2024).

[50]   Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[51]   Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2018. With great training comes great vulnerability: Practical attacks against transfer learning. In *27th USENIX security symposium (USENIX Security 18)*. 1281–1297.

[52]   Fan Wu, Emily Black, and Varun Chandrasekaran. 2024. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209* (2024).

[53]   Simone Zhang, Janet Xu, and A Alvero. 2024. Generative ai meets open-ended survey responses: Participant use of ai and homogenization. (2024).

[54]   Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, et al. 2024. Assessing and understanding creativity in large language models. *arXiv preprint arXiv:2401.12491* (2024).

[55]   Eric Zhou and Dokyun Lee. 2024. Generative artificial intelligence, human creativity, and art. *PNAS nexus* 3, 3 (2024), pgae052.

[56]   Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043* (2023).

A   DIVERGENT THINKING TEST WORDING

Here, we report the exact wording for the tests given to humans and LLMs. The wording differs slightly between the two groups because the LLM models are prompted to output their work in a particular format for easier processing, while human prompts refer to text boxes in the survey UI. Without formatting instructions in the prompt, LLMs often discussed the reasoning behind their word choices. While mildly interesting, this muddied the data.

A.1   AUT prompts.

For original experiments, we use the following start words for AUT: WORD = {book, fork, table, hammer, pants}. For the expanded LLM evaluation of §5.2, we use WORD = {book, bottle, brick, fork, hammer, pants, shoe, shovel, table, tire}.

Human prompt. *Imagine that someone gives you* WORD. *In the blanks below, write down as many creative uses you can think of for this object, up to 10 uses.*

LLM prompt. *Imagine that someone gives you a* WORD. *Write down as many uses as you can think of for this object, up to 10 uses. Please list the uses as words or phrases (single word answers are ok), separated by semicolons. Do not write anything besides your proposed uses.*

A.2 Forward Flow prompts.

We use the following start words for Forward Flow: WORD = {candle, table, bear, snow, toaster}.

Human prompt. (From the original Flow paper) *Starting with the word* WORD, *in each of the following blanks, write down the next word that follows in your mind from the previous word. Please put down only single words, and do not use proper nouns (such as names, brands, etc.). Start by writing* WORD *in the first space below.*

LLM prompt. *Starting with the word* WORD, *your job is to write down the next word that follows in your mind from the previous word. Please put down only single words, and do not use proper nouns (such as names, brands, etc.). Stop after you listed at least 22 words. Print just the list of words, separated by commas, and do not add anything else to your response. The first word in the list should be 'candle'.*

A.3 DAT Prompts.

Human prompt. (From the original DAT paper) *In the spaces below, please enter 10 words that are as different from each other as possible, in all meanings and uses of the words. You must follow the following rules: 1. Only single words in English. 2. Only nouns (e.g., things, objects, concepts). 3. No proper nouns (e.g., no specific people or places). 4. No specialised vocabulary (e.g., no technical terms). 5. Think of the words on your own (e.g., do not just look at objects in your surroundings). 6. Complete this task in less than four minutes.*

LLM prompt. *Instructions: Please enter 10 words that are as different from each other as possible, in all meanings and uses of the words. Rules: 1. Only single words in English. 2. Only nouns (e.g., things, objects, concepts). 3. No proper nouns (e.g., no specific people or places). 4. No specialised vocabulary (e.g., no technical terms). 5. Think of the words on your own (e.g., do not just look at objects in your surroundings). 6. Complete this task in less than four minutes. 7. Return just the list of words, separated by commas, and do not include any other content.*
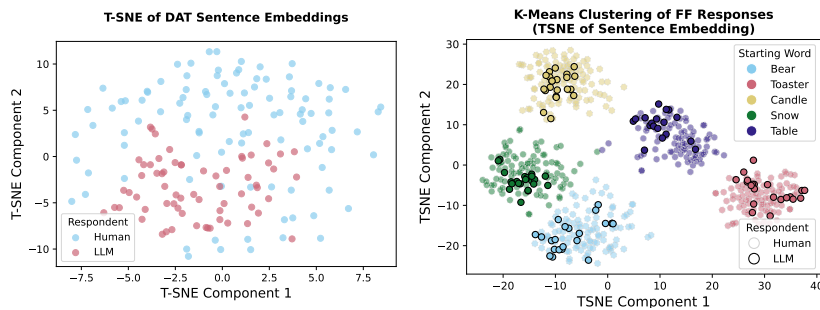

Fig. 10. **LLM responses to the DAT and FF cluster more in feature space than do human responses**.

B ORIGINALITY SCORES FOR AUT, FF, AND DAT

Here, we describe our methods of computing originality scores for each test. Originality scores are denoted as $O_t(\mathcal{P})$, where t = AUT, FF, or DAT and $\mathcal{P}$ is a population, either humans or LLMs.

We denote a single word test response as r and an n-word test response as $\mathbf{r} = \{r_0, r_1, \ldots r_n\}$. The word embedding model is $\mathcal{W}$, and the embedding of a response r is $\mathcal{W}(r)$ (similar for $\mathbf{r}$, $\mathbf{r_j}$, etc.). We use cosine similarity $\cos(\mathcal{W}(r_1), \mathcal{W}(r_2))$ to measure semantic distance between embedded responses.

AUT scoring. Following [19], we score the originality of AUT responses by measuring the semantic distance between a prompt p (e.g. "book") and each word in $\mathbf{r}$ (e.g. "use it as a doorstop"). Because different words in the AUT response contribute differently to overall response creativity (e.g. "it" matters less than "doorstop"), the final originality score is computed via a weighted sum of these distances. Weights are determined by running TF-IDF analysis on the corpus of responses, which produces low weights for common words like "it" and high weights for unusual words like "doorstop". The set of originality scores for AUT responses of population $\mathcal{P}$ is then:

$$O_{\text{AUT}}(\mathcal{P}) = \left\{ 1 - \frac{\sum_{j=0}^{n-1} w_j \cdot \cos(\mathcal{W}(p), \mathcal{W}(r_j))}{\sum_{j=0}^{n-1} w_j}, \forall (\mathbf{r}, p) \in \mathcal{P} \right\} \tag{2}$$

where w_j is the TF-IDF weight for the jth word of response $\mathbf{r}$ and p is the prompt.

FF scoring: Here, we follow the methodology of [22]. This defines the "instantaneous" forward flow of a given thought in the sequence $\mathbf{r}$ as the average distance between the mth thought in the sequence r_m and all preceding thoughts:

$$\frac{\sum_{j=1}^{m-1}(1 - \cos(\mathcal{W}(r_j), \mathcal{W}(r_m)))}{m-1}$$

Building on this, the set of FF scores for a population $\mathcal{P}$ consisting of n-word sequences $\mathbf{r}$ is given by:

$$O_{\text{FF}}(\mathcal{P}) = \left\{ \frac{1}{n-1} \cdot \sum_{i=2}^{n} \frac{\sum_{j=1}^{i-1}(1 - \cos(\mathcal{W}(r_j), \mathcal{W}(r_i)))}{(i-1)}, \forall \, \mathbf{r} \in \mathcal{P} \right\} \tag{3}$$

DAT scoring: We use the scoring methodology of [38], which scores responses by averaging the semantic distance between all pairs of words in the response. Given a population $\mathcal{P}$ composed of n-long DAT response $\mathbf{r}$ containing words $\{r_0, r_1, \ldots r_{n-1}\}$, the set of DAT scores is calculated as:

$$O_{\text{DAT}}(\mathcal{P}) = \left\{ \frac{1}{n(n-1)} \sum_{i,j(i \neq j)}^{n-1} (1 - \cos(\mathcal{W}(r_i), \mathcal{W}(r_j))) \forall \, \mathbf{r} \in \mathcal{P} \right\} \tag{4}$$

C TSNE OF FF AND DAT

Figure 10 visualizes the TSNE of sentence embeddings for DAT and FF responses, similar to Figure 3. This confirms the trend observed in the AUT TSNE: LLM responses cluster closer in feature space than human responses, resulting in lower population-level originality measurements. We perform k-means clustering of TNSE of FF sentence embeddings to demonstrate clusters of LLM response for each start word. Since the DAT since the test does not involve varying start words, we simply visualize all LLM and human responses.