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

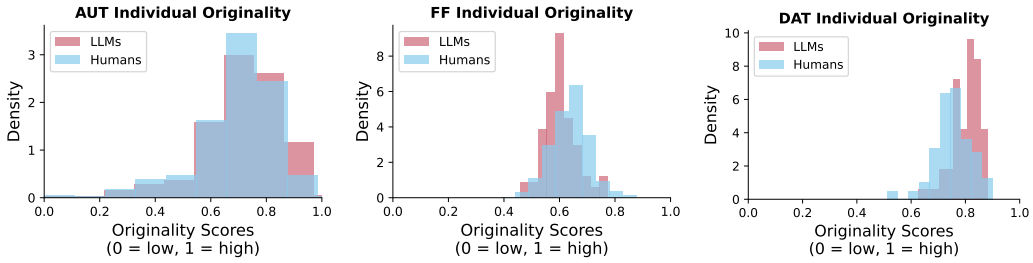

Fig. 1. LLMs slightly outperform humans on the AUT and DAT, but humans slightly outperform LLMs on FF.

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*

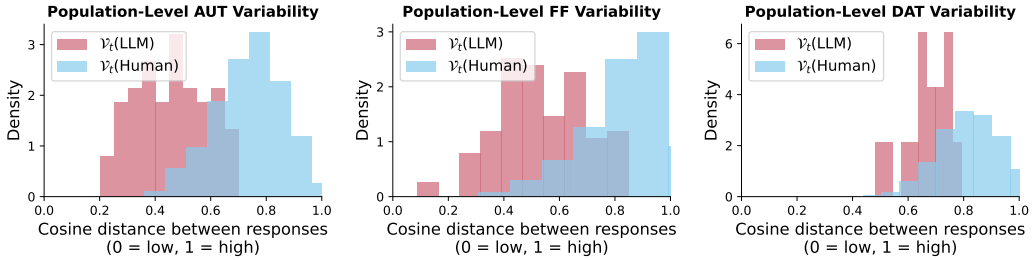

Fig. 2. LLM responses exhibit far less variability than human responses, *as measured by cosine distance between embedded responses.*

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 $\mathcal{V}_t$(LLM) and $\mathcal{V}_t$(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.

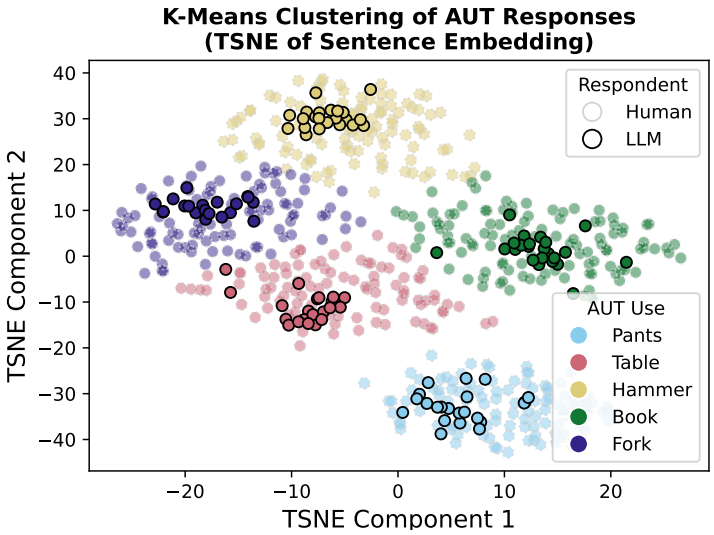

Fig. 3. LLM responses cluster together in feature space more than do human responses. *K-means clustering of TSNE of AUT sentence embeddings.*

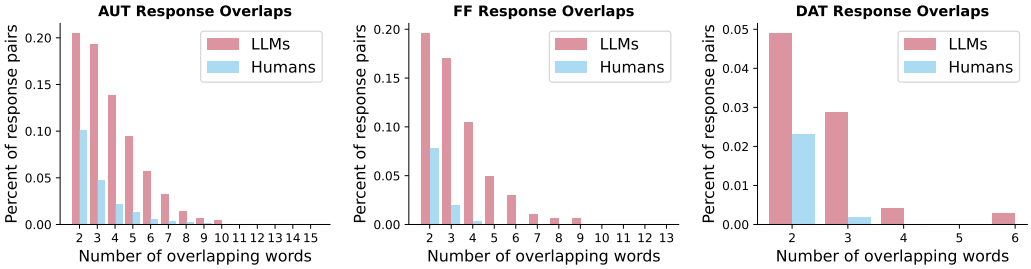

Fig. 4. LLM responses have far more words in common than do human responses. *We look at word overlaps between "full" responses from LLMs and humans—e.g. all uses from the AUT, all words in the FF, etc. This corresponds to the sentence embedding method of population originality measurement, and explains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