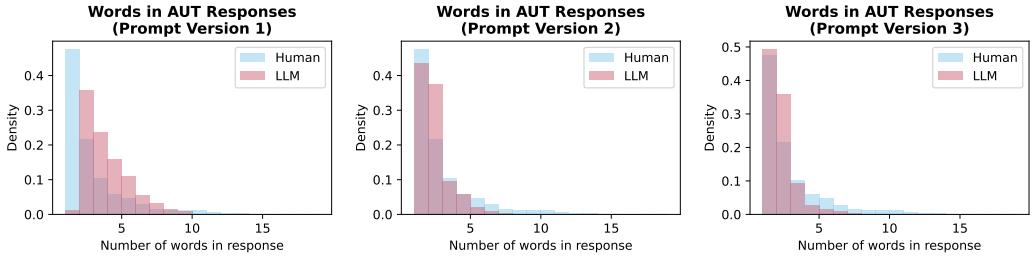

Fig. 5. Effect of different AUT prompt wordings on length of LLM AUT responses. *We use prompt verison 3 in most experiments in this paper, since LLM responses to this prompt most closely match the human distribution of response lengths.*

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 $\mu(\mathcal{V}_t(LLM)) = \mu(\mathcal{V}_t(Human))$; alternative is that $\mu(\mathcal{V}_t(LLM)) > \mu(\mathcal{V}_t(Human))$. "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.*

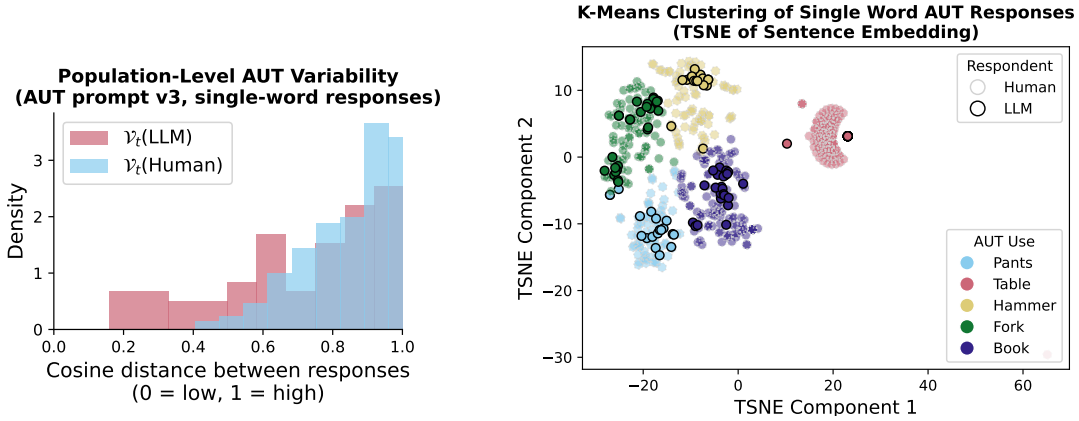

Fig. 6. Even when considering only one-word responses to control for response structure, LLM AUT responses have lower population-level variability (left plot) and are closer in feature space (right plot) than human responses. *LLM responses are generated with prompt version 3. We create sentence embeddings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

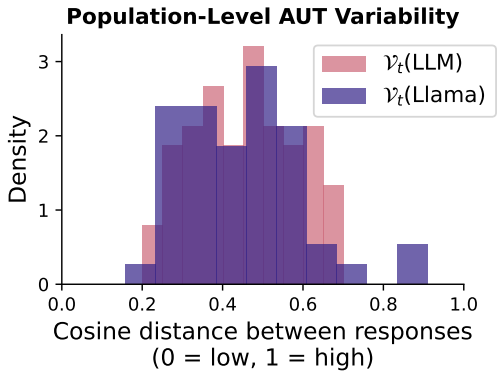

Fig. 7. Models from the same family (Llama) exhibit slightly lower population-level variability than models from different families.

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."