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 $\mathcal{O}_{\text{t}}(\mathcal{P})$, 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 $\mathcal{O}_t(LLM)$ and $\mathcal{O}_t(Humans)$ to measure differences in originality between the two groups. We do this by testing for statistically significant differences in $\mu(\mathcal{O}_t(LLM))$ and $\mu(\mathcal{O}_t(Human))$ 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 $\mu(\mathcal{O}_t(LLM)) = \mu(\mathcal{O}_t(Human))$, and the alternative is that $\mu(\mathcal{O}_t(LLM)) > \mu(\mathcal{O}_t(Human))$.

*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.