Do Semantic Distance Tests *Actually* Predict Creativity in LLMs? Evidence for Jagged Creativity

Samuel Schapiro[1,2*] **Alexi Gladstone**[3] **Jonah Black**[3] **Heng Ji**[3]

[1]AI Innovation Institute, Stony Brook University [2]Knowledge Lab, UChicago
[3]University of Illinois, Urbana-Champaign

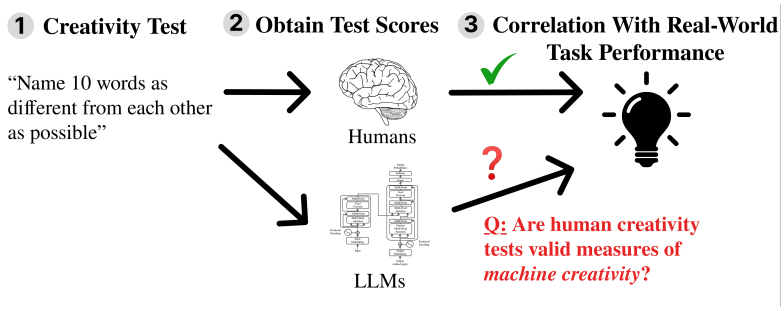


Figure 1: **Overview.** Semantic distance tests like the Divergent Association Task ask respondents to name a set of maximally dissimilar words and score the result by embedding distance. The external validity of these tests among humans has been supported by modest correlations with ideation tasks, but their validity as measures of *machine* creativity has not been adequately established, motivating the central question of this work: *Do semantic distance tests actually predict creativity in LLMs?*

Abstract

Measuring the creativity of large language models (LLMs) is essential for designing methods that can improve creativity and for enhancing our scientific understanding of this ability. To accomplish this, it has become common in recent years to administer semantic distance tests of human creativity to LLMs. Although these tests provide a convenient and fully automated way to score aspects of "creativity," their validity as measures of *machine* creativity has not been established. To address this problem, we conduct the first large-scale, systematic study assessing the effectiveness of semantic distance tests for predicting the creative achievement of LLMs across four target constructs: (a) creative writing, (b) output diversity on open-ended prompts, (c) combinatorial creativity, and (d) scientific ideation. We score each test on two criteria (*validity* and *specificity*), and derive a theoretical limit for the maximum attainable specificity and validity a test can achieve. We find that: **(1)** No test is a significant predictor of output diversity, suggesting that *point creativity*, the ability to produce a single response connecting semantically distant concepts, and *distributional creativity*, the ability to produce distinct responses across repeated samples, may dissociate in current LLMs—a pattern we call ***jagged creativity***. **(2)** The Remote Associates Test is effectively a proxy for general capabilities rather than a *specific* predictor of creativity in LLMs. **(3)** Lastly, the Divergent Association Task is the best predictor of creative writing, but no single test predicts all constructs well. In summary, our results prescribe caution when applying psychometric tests and tasks of creativity validated among human populations to LLMs.

*Correspondence to schapironietzsche@gmail.com.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Evaluating the creativity of large language models (LLMs) is essential for developing methods that improve creativity, advancing our scientific understanding of this ability, and ensuring robust deployment of AI in human co-creativity environments. In recent years, it has become common practice to re-purpose psychological assessments of *human* creativity to purportedly evaluate the "creativity" of LLMs [47, 6, 3]. Broadly, such tests assess one of two mechanisms implicated in the creative process: *divergent thinking* is the ability to generate multiple distinct responses to an open-ended question, while *convergent thinking* is the ability to generate a single response that unifies multiple diverse stimuli [10, 11, 46] (Figure 2). Central to **divergent thinking** is the ability to associate semantically distant concepts, which has long been considered fundamental to human creativity [24, 46, 49], motivating the use of automated semantic distance assessments like the Divergent Association Task (DAT) [30, 6, 9, 3, 51] to assess whether LLMs are capable of making distant associations. On the DAT, subjects are instructed to generate ten maximally dissimilar nouns, and scores are given by the mean pairwise semantic distance of all word pairs under an embedding model such as GloVe [33], providing a convenient way to assess divergent thinking without human raters.

Over the last year, two novel semantic distance tests have also been proposed. The Conditional DAT (CDAT; [28]) extends the DAT by requiring dissimilar nouns to be relevant to a "cue" word, incorporating a measure of appropriateness in addition to novelty [4, 22]. Similarly, the Parallel Association Chain Evaluation (PACE; [35]) is inspired by the forward-flow measure used in human psychometrics [14, 2] and instructs models to make free associations starting from several seed words, scoring responses by a sequential semantic distance measure.

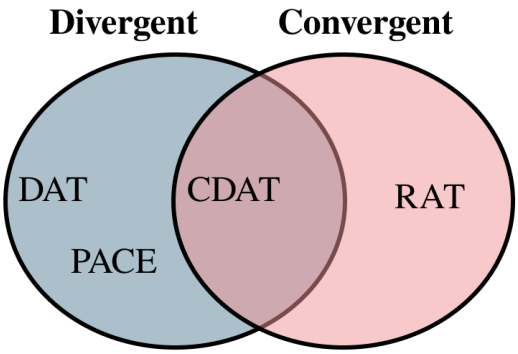


Figure 2: **Semantic distance tests along the divergent–convergent thinking axis.** The Divergent Association Task (**DAT**) asks for ten maximally dissimilar nouns, and does not involve a convergent thinking component. Similarly, the Parallel Association Chains for Creativity Evaluation (**PACE**) tests intrinsic associative patterns, and does not involve a convergent thinking component. The Conditional DAT (**CDAT**) requires each of the maximally dissimilar nouns to be relevant to a cue word, jointly assessing an aspect of convergent thinking. Lastly, the Remote Associates Test (**RAT**) is a convergent thinking task with a single correct answer per item.

Convergent thinking, on the other hand, is commonly assessed by the Remote Associates Test (RAT) [24, 5], which presents three stimulus words, and asks for a single word that connects all three (e.g., given *cottage, swiss*, and *cake*, the response "cheese" applies to each stimulus). Various other convergent creativity assessments and benchmarks for LLMs have been introduced in recent years, including the Only Connect Wall test and the CresOWLve benchmark [27, 17], and, for multimodal LLMs, the Labyrinth of Links [21] and MM-OPERA [16] benchmarks.

Despite their growing use, it remains unclear what these tests measure in LLMs. For example, PACE shows strong Spearman rank correlations with creative writing benchmarks ($\rho \approx 0.74$). However, it also correlates strongly with general model capabilities ($\rho \approx 0.66$), so it is unclear how well PACE measures creative achievement *independent* of what general capability already predicts. More broadly, the external validity of semantic distance tests (i.e., as predictors of creativity in LLMs on real-world tasks) has not been established, and their validity as predictors of human creativity remains limited(see Section 2).

Our Contributions To address these problems, we carry out the first systematic study assessing the effectiveness of semantic distance tests for predicting creative achievement in LLMs. We measure creative achievement using seven benchmarks that capture four target constructs: (i) creative writing, (ii) output diversity on open-ended prompts, (iii) combinatorial creativity, and (iv) scientific ideation. We also introduce two evaluation criteria: *validity*, which measures the raw correlation r with each benchmark, and *specificity*, the semi-partial correlation $r|g$ after residualizing benchmark scores on a capability proxy g. The latter assesses how well semantic distance tests predict aspects of creative

Divergent Association Task (DAT). Generate 10 nouns maximally different from each other.

Response: ocean, mathematics, hammer, justice, molecule, symphony, volcano, laughter, friction, taxonomy

Parallel Association Chain Evaluation (PACE). Starting with the seed **"rock"**, produce three 20-word association chains in which each word associates only with the previous word.

Response (chain 1 of 3): rock → stone → pebble → beach → sand → hourglass → time → clock → alarm → fire → smoke → cigarette → tobacco → leaf → tree → bark → dog → collar → shirt

Conditional Divergent Association Task (CDAT). Generate 10 nouns associated with **"rock"** that are maximally different from each other.

Response: stone, guitar, music, geology, cliff, mineral, foundation, cradle, concert, pebble

Remote Associates Test (RAT). What single word can be combined with each of **"cream"**, **"skate"**, and **"water"** to form a compound word or common phrase?

Response: ice

Figure 3: **Example prompts and responses for each semantic distance test.** The DAT prompts for maximally distant nouns, PACE prompts models to freely associate sequences of nouns in a 20-word chain, the CDAT prompts for maximally distant nouns that are each relevant to a cue, and the RAT prompts for the single word that links three cue words.

achievement *after* variance explained by general capabilities is factored out, which is particularly important given that correlations between benchmarks and general capabilities are as high as $r = 0.87$ (Figure 5a). In detail, we make the following contributions and findings:

1. We conduct a large-scale, systematic study evaluating the effectiveness of semantic distance tests for predicting the creative achievement of LLMs.
2. We introduce evaluation metrics which measure a test's predictive power, both in raw correlations with benchmarks (construct *validity*) and independent of what general capabilities already predict (*specificity*), and bound the validity and specificity a test can jointly achieve.
3. We find that no test significantly predicts output diversity, including the DAT, suggesting that *point creativity* (producing a single response that connects semantically distant concepts) and *distributional creativity* (producing distinct responses across repeated samples) may dissociate in LLMs, a pattern we call ***jagged creativity*** (Figure 7) by analogy to the *jagged intelligence* of modern AI systems [26]. For example, `gpt-5-4`, the top DAT scorer, includes "quasar" in 39 of its 40 DAT responses (Table 5).
4. We find that the Remote Associates Test (RAT) is effectively a capability proxy—its significant correlations with creative writing benchmarks become non-significance once general capabilities are statistically controlled for.
5. Our empirical findings indicate that the Divergent Association Task (DAT) is the best predictor of creative writing, but that no single test predicts all constructs well.

2 Background and Motivating Problems

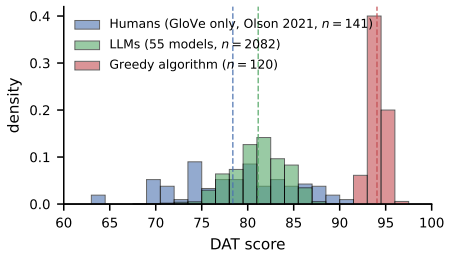
Note on Terminology

Creativity can refer to a person, process, press (environment) or product [36], and in this work, we mainly focus on creative products. Furthermore, we distinguish between a model's *score on a creativity test* (like the DAT) from its *creative achievement*, the latter of which refers to the novelty and appropriateness of the artifacts it actually produces on real-world tasks assessed by creativity benchmarks [4, 38].

In recent years, semantic distance tests have been widely adopted to assess whether LLMs are capable of making distant associations [6, 9, 3], which have long been thought to underlie creative ideation [24]. Furthermore, various strategies, such as varying sampling parameters, instructing LLMs to assume the personas of distinguished creative individuals [51], prompt engineering [51], offering explicit test-taking strategies [3], and steering model weights [43], have all been explored to improve semantic distance test scores. Additionally, semantic distance tests are already used to make direct

Require: vocabulary V, embedding E, number of words n (= 10 for the DAT)
Ensure: word list $W = (w_1, \ldots, w_n)$
1: $w_1 \sim \text{Uniform}(V)$
2: $W \leftarrow [w_1]$
3: **for** $i = 2 \ldots n$ **do**
4: $\quad w_i \leftarrow \arg\min_{v \in V \setminus W} \frac{1}{|W|} \sum_{u \in W} \cos(E(v), E(u))$
5: $\quad W \leftarrow W \cup \{w_i\}$
6: **end for**
7: **return** W

(a) Greedy algorithm



(b) DAT scores

Figure 4: **A simple greedy algorithm exceeds mean human and LLM DAT scores. (a)** The algorithm. **(b)** DAT score distributions; dashed lines mark group means (details in Section F).

claims that LLMs are *more* or *less* creative than humans [6, 3, 51], which presupposes that the tests measure the same construct in both populations. However, this is problematic for several reasons.

For one, **the validity of semantic distance tests as measures of *machine* creativity has not been established.** Measures of human intelligence are often inappropriate to administer naively to LLMs [8] because such tests do not inherently possess construct validity for machines.[2] Analogously, semantic distance tests can be "hackable" by computational mechanisms that would not be considered "creative" [4]. Consider that Transformers have been shown to implement algorithmic circuits in their weights [31, 29], and that the DAT itself can be solved by a simple greedy algorithm over embedding distances (Figure 4a) whose scores substantially exceed the mean human and LLM scores (94.1 vs. 78.4 and 81.1; Figure 4b). A psychometric test is a valid measure of creativity if scores correlate strongly with quantitative measures of creative achievement *within that same population*. Semantic distance tests were designed for human populations, and their external validity as measures of *machine* creativity has not yet been established.[3]

Furthermore, **the validity of semantic distance tests as measures of *human* creativity is itself only modestly established.** Even within human populations, the external validity of these tests rests on a short and modest chain of correlations. Both the DAT [30] and forward flow[4] [14] are validated by correlation with the Alternative Uses Test (AUT): the DAT at $r \approx .32$–$.51$ [30][5] and forward flow at $r \approx .43$–$.49$ [2]. However, meta-analyses have revealed that AUT scores themselves correlate with self-reported creative achievement at only $r \approx .17$ [39] to $r \approx .22$ [20], and [1] reports near-zero correlations between AUT scores and creative achievement, casting doubt on the external validity of semantic distance tests in predicting creative outcomes among humans.

Finally, we argue that **semantic distance tests should measure something independent of what general capability already predicts.** A semantic distance test that correlates with, e.g., creative writing rankings, does not, on its own, establish that the test measures "creativity." A semantic distance test should predict aspects of creative achievement in ways that are statistically *independent of* what general capability already predicts—otherwise it reduces to a measure of general capability rather than creativity. Among humans, where divergent thinking test scores are known to correlate with general intelligence up to $r = 0.37$ [12], the standard practice is to factor out confounding variables (e.g., through hierarchical regression or factor analysis) before claiming a test measures "creativity" [2]. In Section 4, we propose a similar methodology for measuring the predictive power of semantic distance tests independent of what general capabilities already predict.

3 Preliminaries

With these problems in mind, we now introduce the tests and benchmarks we evaluate.

[2] The digit span test is commonly used to assess human intelligence [53], but would be inappropriate to administer to machines, where working memory is not a bottleneck to intelligence the way it is among humans.

[3] Moreover, these tests may be flawed due to training data leakage, as has been observed with the AUT [47]

[4] PACE [35] uses the same scoring mechanism as forward flow.

[5] The DAT is also validated against the Bridge-the-Associative-Gap Test [13], a measure of convergent thinking.

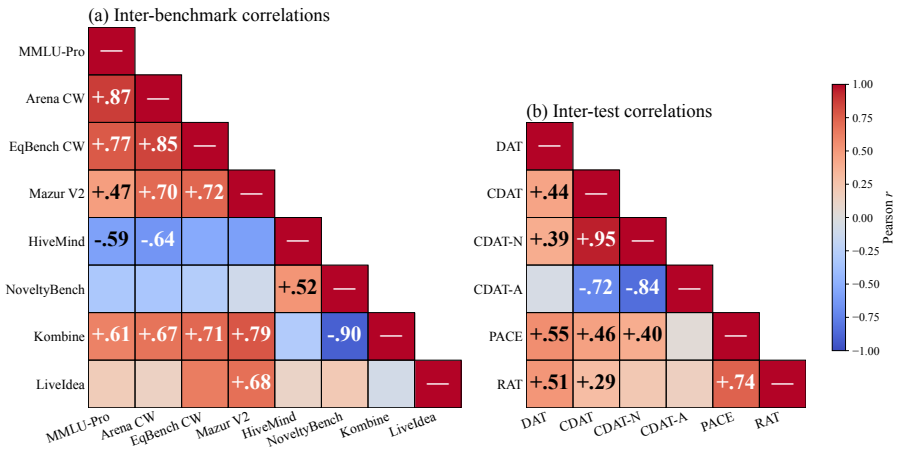


Figure 5: **Inter-benchmark and inter-test correlations. (a)** Benchmarks are ordered by construct: general capability (MMLU-Pro), creative writing (Arena CW, EqBench CW, Mazur CW), output diversity (Hivemind, NoveltyBench), combinatorial creativity (Kombine), and scientific ideation (LiveIdea). Each cell uses the models scored on both benchmarks, from $n = 6$ (Kombine $\times$ Hivemind) to $n = 40$ (MMLU-Pro $\times$ Arena CW). **(b)** Inter-test correlations among DAT, CDAT, CDAT-N, CDAT-A, PACE, and RAT ($n = 64$–68), where each embedding-scored test uses its mean z-score across the GloVe, FastText, and SBERT embeddings. Values are shown only where $p < .05$.

3.1 Semantic Distance Tests

Divergent Association Task (DAT) Introduced by [30], the DAT asks subjects to name 10 nouns[6] $W = \{w_1, \ldots, w_n\}$ as different from each other as possible, and scores the average cosine distance between all word pairs:

$$\text{DAT}_{\mathcal{E}}(W) := \frac{100}{n(n-1)} \sum_{i \neq j}^{N} (1 - \cos_{\mathcal{E}}(\mathbf{w}_i, \mathbf{w}_j)) \tag{1}$$

In addition to computing scores under the standard GloVe 840B embeddings [33] used in [30], we test robustness under multiple embedding models in Section 4.

Conditional Divergent Association Task (CDAT) Creative artifacts should be both novel *and* useful [50, 4, 45, 22]. [28] argued that the original DAT failed to measure utility, and proposed the conditional DAT (CDAT), in which a cue word c is added, and each word must be sufficiently relevant to the cue word in order for its novelty to count. For words $W = \{w_1, \ldots, w_n\}$ generated for cue c, novelty and appropriateness are measured according to:

$$\text{CDAT-N}_{\mathcal{E}}(W) := \frac{100}{n(n-1)} \sum_{i \neq j}^{N} (1 - \cos_{\mathcal{E}}(\mathbf{w}_i, \mathbf{w}_j)); \quad \text{CDAT-A}_{\mathcal{E}}(W \mid c) := \frac{100}{n} \sum_{i=1}^{n} \cos_{\mathcal{E}}(\mathbf{c}, \mathbf{w}_i) \tag{2}$$

Note that CDAT-N uses Equation (1), but its scores differ from the DAT's due to cue-based prompting.

Parallel Association Chain Evaluation (PACE) PACE [35], inspired by forward-flow associative-chain methods [14], prompts models to produce three parallel 20-word association chains from a seed, scored by the mean cumulative cosine distance along each chain $C = (w_1, \ldots, w_L)$:

$$\text{PACE}_{\mathcal{E}}(C) := \frac{1}{L-1} \sum_{i=2}^{L} \frac{1}{i-1} \sum_{j=1}^{i-1} (1 - \cos_{\mathcal{E}}(\mathbf{w}_i, \mathbf{w}_j)) \tag{3}$$

The model-level score is the mean of $\text{PACE}_{\mathcal{E}}(C)$ across three chains per seed and across all seeds. In the original paper, PACE correlated with Chatbot Arena CW at $\rho \approx 0.74$ with $n = 30$. The

[6] Usually, only the first seven valid nouns are scored.

metric was originally tested under FastText embeddings [25], although [35] also saw significant rank correlations with Chatbot Arena CW under multiple distinct embedding models.

Remote Associates Test (RAT) The Remote Associates Test [24, 5] is a long-standing psychometric instrument for assessing the convergent thinking component of creativity. The RAT presents three remote stimulus words and asks for the single word that connects all three. For example, given (cottage, swiss, cake), the answer is *cheese* (*cottage cheese*, *swiss cheese*, *cheesecake*). Unlike previous tests, where scoring is based on embedding distance, the RAT scores items as correct or incorrect against an answer key. We use a 30-item subset of the [5] stimulus set and score each model's ability to produce the correct response in a zero-shot fashion.

3.2 Creative Achievement Benchmarks

We evaluate each test against four target constructs (creative writing, output diversity on open-ended prompts, combinatorial creativity, and scientific ideation) spanning seven benchmarks.

3.2.1 Creative Writing

Chatbot Arena Creative Writing (Arena CW) [7]. Arena CW is a creative writing benchmark where models respond to open-ended, user-submitted creative writing prompts, and pairwise user preferences are aggregated into per-model Elo ratings. Arena CW is the most capability-loaded of the three creative writing benchmarks ($r = 0.87$ with MMLU-Pro; Figure 5).

EQ-Bench Creative Writing [32]. EQ-Bench is a creative writing benchmark where models produce responses to a set of 32 distinct prompts designed to challenge models in areas like humor, romance, spatial awareness, and unique perspectives. Responses are scored head-to-head by Claude Sonnet using a 9-criterion rubric and aggregated into Elo ratings, and the benchmark incorporates strategies to reduce known LLM-as-a-judge biases (length, position, poetic incoherence, etc.). EQ-Bench correlates strongly with MMLU-Pro ($r = 0.77$; Figure 5), although less than Arena CW.

Mazur Creative Writing [23]. Mazur Creative Writing is a creative writing benchmark that tests how well LLMs incorporate a set of mandatory story elements (characters, objects, core concepts, attributes, motivations, etc.) in a short creative story. An ensemble of LLM judges applies an $\sim$18-question rubric assessing whether mandatory elements are adequately represented, as well as the overall narrative quality. The final story score is the mean of the per-grader scores. Mazur Creative Writing correlates moderately with MMLU-Pro ($r = 0.47$), less than Arena CW and EQ-Bench.

3.2.2 Output Diversity

We use two benchmarks that measure output diversity in LLM responses to open-ended prompts.

Hivemind [18]. Hivemind is an output diversity benchmark which measures output diversity across open-ended queries, drawn from real-world user-ChatGPT interactions. Categories include brainstorm and ideation ("Suggest a feature for a smartwatch designed specifically for senior citizens."), philosophical questions ("How do I understand what I want?"), speculative and hypothetical scenarios ("Create a short review of a future movie."), and ambiguous everyday questions ("How can I live on $1,000 per month?"). Scoring measures intra-model repetition via average pairwise embeddings dissimilarity[7] using OpenAI's `text-embedding-3-small` model. Hivemind scores are negatively correlated with capability ($r = -0.59$ with MMLU-Pro).

NoveltyBench [55]. NoveltyBench evaluates the output diversity of LLMs on a similar collection of real-world user-ChatGPT interactions, as well as open-ended prompts across four categories: randomness ("Roll a make-believe 20-sided die"), factual information ("List a capital city in Africa."), creative text writing ("Tell me a riddle."), and subjectivity ("What's the best car to get in 2023?"). Scoring assesses the extent to which a sequence of generations is both diverse and high-quality, using a cumulative utility measure that penalizes functional equivalence to prior responses. NoveltyBench is only weakly correlated with capability, obtaining $r = -0.32$ with MMLU-Pro.

[7] The original paper uses similarity, but for consistency with the output diversity construct, we report its complement.

3.2.3 Combinatorial Creativity

Kombine [42]. Kombine is a combinatorial creativity benchmark tasking models to discover surprising, original, and useful associations, analogies, and blends between unrelated entities drawn from the Wikidata knowledge graph, and to use those analogies and blends to invent novel concepts by re-combining the parts of existing ones. Responses are scored by embedding-based measures such as surprise and originality, as well as judge-based measures on utility and emergent creativity. Here, we use each model's overall score across the three tasks. Kombine correlates strongly with general capabilities ($r = 0.61$ with MMLU-Pro; Figure 5a).

3.2.4 Scientific Ideation

LiveIdeaBench [37]. LiveIdeaBench assesses LLMs on open-ended scientific idea generation. Given a single keyword (e.g., a domain term) drawn from a set spanning 18 scientific disciplines, a model must propose a research idea relevant to that keyword. Each idea is scored by a panel of LLM judges along five dimensions—*fluency* (does the response actually contain a usable idea), *feasibility*, *clarity*, *originality*, and *flexibility*—with overall scores reported as the average across all dimensions. The original benchmark explicitly demonstrates that scientific ideation can dissociate from general capability (e.g., QwQ-32B-preview outperforms much larger frontier models on Originality despite being weaker on knowledge tests), and empirically, we observe a similar trend where LiveIdeaBench is only weakly correlated with general capabilities (Figure 5a; $r = 0.18$ on MMLU-Pro).

4 Evaluation Method

4.1 Models and Inference

In order to evaluate the effectiveness of semantic distance tests for predicting creative achievement benchmarks, we obtain raw test scores across a large set ($n = 55$) of instruction-tuned LLMs across 10 providers (OpenAI, Anthropic, Google, Meta, Mistral, Qwen, DeepSeek, Cohere, NVIDIA, Microsoft) via OpenRouter. DAT and CDAT use $T \in \{1.0, 1.5, 2.0\}$ (40 trials per temperature, with `top_p`= 1, `top_k`= 0). PACE follows the original setup in [35], where sampling is done with temperature $T = 0$ and stochasticity is controlled by varying the random seed for each anchor word. We use 50 anchor words, and collect three parallel 20-word chains per seed. Lastly, on the CDAT, we use 50 cue words, and following [28], per-cue appropriateness values are compared to a random-noun baseline via Welch's t-test, with a Benjamini–Hochberg false discovery rate (FDR) correction at $\alpha = .001$ across models within each temperature. A model's responses at a given temperature are retained if their FDR-adjusted p-value passes and its mean appropriateness exceeds the random baseline. The CDAT score we report is the mean of CDAT-N across passing temperatures. For reproducibility purposes, exact prompts used for each test are given in Appendix H.

Embedding models Each semantic distance test score depends on the embedding model used for pairwise cosine similarity. To evaluate robustness to embedding choice, we score each test under three embeddings—GloVe 840B 300d, FastText crawl-300d-2M, and Sentence-BERT `all-mpnet-base-v2`—which differ in training. GloVe and FastText are static word-level co-occurrence models (300-dim), while Sentence-BERT (`all-mpnet-base-v2`; 768-dim) is a transformer sentence encoder trained via contrastive sentence-pair objectives.

4.2 Evaluation Metrics

We evaluate each test on two criteria, validity and specificity, each motivated and defined below.

Validity Standard psychometric criteria used in human creativity research [2, 30] indicate that a test has construct validity if its scores correlate with external measures of that construct. Therefore, we report *validity* as the raw Pearson correlation $r(X, Y)$ between test score X and benchmark score Y.

Specificity Since creative achievement benchmarks are themselves highly capability-loaded (for example, Arena CW correlates with MMLU-Pro at $r = 0.87$) a high raw $r(X, Y)$ may simply reflect that the test predicts general capability. Thus, we report *specificity*, the semi-partial Pearson correlation between the test score X and the benchmark Y residualized on a capability stack g, given by $r(X, Y \mid g) := r\Big(X,\ Y - \hat{Y}_g\Big)$. Here, $\hat{Y}_g$ is the ordinary least squares prediction of Y from

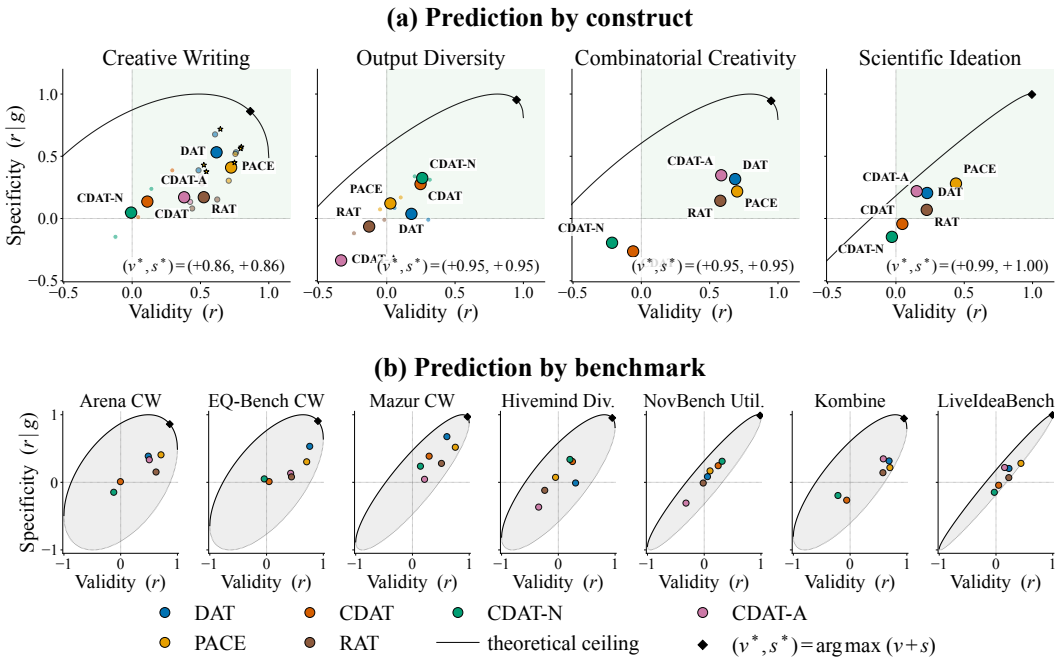


Figure 6: **Validity and specificity by construct and benchmark, with theoretical ceilings.** *(a) Prediction by construct:* Each small point is a (test, benchmark) cell from the Overall block of Table 1, while each large black-outlined circle is a test's average across the construct's benchmarks. A gold ★ = positive and significant on both axes ($p < .05$). The black curve is the construct-level theoretical specificity ceiling obtained in Theorem 1, and the black diamond marks its maximum of $v + s$. *(b) Prediction by benchmark:* the same ceiling for each benchmark. The grey region is the feasible $(v, r(X, Y \mid g))$ set.

MMLU-Pro accuracy [52]. The semi-partial correlation measures to what extent the test predicts the benchmark's capability-controlled variance, with a significant positive $r(X, Y \mid g)$ indicating the test predicts variance that capability cannot already explain.

For both validity and specificity, we report two-sided p-values from the standard Pearson t-test, $t = r\sqrt{(n-2-k)/(1-r^2)} \sim t_{n-2-k}$, where k is the number of controls regressed out ($k = 0$ for validity and $k = 1$ for MMLU-Pro used for specificity).

4.3 Theoretical Limits for Attainable Validity and Specificity

What is the maximum specificity a test can achieve, given its validity and its benchmark's correlation with general capabilities? We call this the *validity-specificity frontier* and bound it below.

Theorem 1 (Validity-Specificity Frontier)**.** *Let $g = (Z_1, \ldots, Z_k)$ be a vector of general capability proxies, let $\hat{Y}_g$ denote the ordinary least squares prediction of Y from g, and define $R = \mathrm{corr}(Y, \hat{Y}_g)$. For any test X with validity $v = \mathrm{corr}(X, Y)$, the semi-partial correlation $r(X, Y \mid g) := r(X,\, Y - \hat{Y}_g)$ satisfies*

$$|r(X, Y \mid g)| \;\le\; v\sqrt{1 - R^2} \;+\; |R|\,\sqrt{1 - v^2}. \tag{4}$$

The proof of Theorem 1 is given in Section A of the appendix. The intuition is that a test highly correlated with a capability-loaded benchmark *cannot* be very decoupled from capability (i.e., specificity). In particular, a perfectly valid test ($v = 1$) has specificity bounded by $\sqrt{1 - R^2}$, and a benchmark with $|R|$ near 1 therefore leaves almost no specificity headroom for *any* test, regardless of construction. This is why Arena CW ($R = 0.87$ on $g =$ MMLU-Pro) caps specificity at ≈ 0.49 for a perfect-validity test, while NoveltyBench Utility ($R \approx -0.32$) admits a wider ceiling (Figure 6b).

5 Results

Table 1 (Section B) reports validity and specificity for all test–benchmark pairs. Test and benchmark scores used for correlations are fully reported in Table 3 and Table 4, respectively. Figure 6 plots validity and specificity across the four constructs against the validity-specificity frontier (Section 4.3).

None of the tests is a significant predictor of output diversity. On the two output diversity benchmarks, no test achieves significant validity or specificity. Notably, this includes the DAT (Hivemind $r = +0.30$, $r|g = -0.01$; NoveltyBench Utility $r = +0.06$, $r|g = +0.08$), even though the DAT explicitly asks for words that are as different from each other as possible. This suggests that the ability to generate a single, highly divergent DAT or CDAT response, whether through memorization or an algorithm like the one in Figure 4a, is distinct from the ability to continually generate distinct responses to open-ended questions. As a result, we find no evidence that *point creativity*, the ability to produce a single creative response that associates more semantically distant entities, predicts *distributional creativity*, the ability to produce a diverse set of responses across repeated samples, suggesting the two may dissociate in current frontier LLMs, a pattern we call ***jagged creativity*** (Figure 7). Table 5 highlights this *jaggedness* in raw DAT responses: `gpt-5-4`, the top DAT scorer, includes "quasar" in 39 of 40 responses and `claude-sonnet-4-6` includes "justice" in all 40, even at $T = 2.0$. Comparing five base models, given the DAT as a text completion, with their instruction-tuned versions, instruction tuning raises the mean DAT score from 77.4 to 81.4 but cuts the number of distinct words a model uses across 40 responses by about two-thirds, from 296 to 97 on average (Table 6).

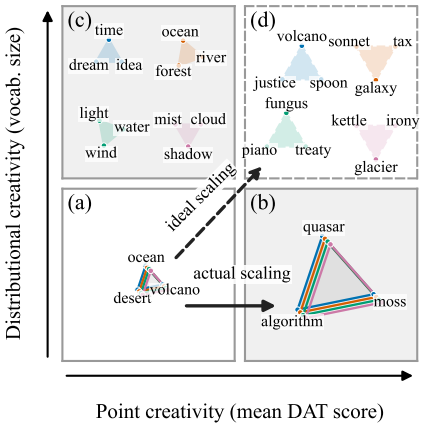


Figure 7: **Jagged creativity.** In (a)–(c), each triangle is three of the ten words from one real DAT response of (a) Mistral 7B, (b) GPT-5.4, and (c) o3-mini, and colors are repeated samples; (d) is hypothetical. (b) and (c) are jagged, while (a) and (d) reflect non-jagged intelligence in which distributional and point creativity scale together, as would be expected among humans [24]. As shown in Table 6, while instruction tuning raises the mean DAT score from 77.4 to 81.4, it cuts the number of distinct words a model uses by about two-thirds, from 296 to 97 on average. [8]

The RAT is not a specific creativity test for LLMs. A core motivation addressed in Section 2 is that tests should measure aspects of creative achievement independent of what general capabilities already predict. On creative writing benchmarks, the RAT obtains statistically significant validity ($r = +0.62^{***}, +0.44^{*}, +0.51^{*}$); however, after controlling for general capabilities, the RAT collapses to non-significant specificity ($r|g \in [+0.08, +0.28]$). The same pattern holds on combinatorial creativity, where the RAT's validity ($r = +0.58^{**}$) drops to $r|g = +0.14$. In summary, the convergent thinking assessed by the RAT largely measures what capability already predicts.

The DAT is the best predictor of creative writing. The DAT is the only test that achieves significant validity *and* specificity on all three creative writing benchmarks (Arena CW $r = +0.49^{***}$, $r|g = +0.39^{*}$; EQ-Bench CW $r = +0.76^{***}$, $r|g = +0.53^{**}$; Mazur CW $r = +0.61^{**}$, $r|g = +0.68^{**}$). PACE obtains the highest raw validity on creative writing ($r \in [0.71, 0.76]$), but retains significant specificity only on Arena CW ($r|g = +0.41^{*}$) and Mazur CW ($r|g = +0.52^{*}$).

No single test predicts all constructs well. Although some tests have significant validity on multiple constructs, no test retains significant specificity on more than one construct. For example, on combinatorial creativity, PACE is the most valid predictor ($r = +0.70^{***}$), but not a specific one ($r|g = +0.22$). Similarly, on scientific ideation, PACE is the only test with significant validity

[8] This distinction parallels [24]'s distinction between *flat* and *steep* associative hierarchies, the original distinction offered among human populations: individuals with flat hierarchies have many remote associates of similar strength to a given stimulus, whereas individuals with steep hierarchies have a few strong, stereotyped associates. Mednick further noted that a *steep-deviant* individual can still be creative if their typical associates are deviant from the population's, as in (b).

($r = +0.44^*$, $n = 25$), but likewise lacks significant specificity.[9] In summary, each test is a good predictor of one specific construct (at best), rather than a general-purpose "creativity" measure.

6 Discussion

Localizing the algorithms behind test performance Our results raise the question of *how* models obtain high scores on semantic distance tests. A high DAT score could reflect a memorized list of distant words, an algorithm like Figure 4a implemented in the model's weights, or a more general capacity for distant association, and only the last should be expected to transfer to creative achievement. Mechanistic interpretability—the AI subfield concerned with understanding the relationship between model internals and model behavior [44]—can leverage the structure of semantic distance tests as guides for localizing relevant circuits in model weights and activations that models use on these tasks, analogous to studies that have identified circuits for arithmetic abilities [29]. Prior work has taken a first step in this direction, localizing the weights that matter for divergent thinking and amplifying them to raise DAT scores, as well as human-judged creativity on longer-form tasks [43].

Beyond embedding-based distance measures In text-to-image creativity studies, hierarchical concept ontologies that decompose visual concepts (e.g., Sofa) into constituent parts (e.g., leg, cushion) and associated uses (e.g., support, rest) have been proposed to support combinatorial creativity in a more explicit way [15]. Ontology-based measures may better capture hierarchical and logical relations between concepts and can complement semantic distance measures in future work.

Testing for transformational creativity One understudied facet of creativity that has historically played an outsized role in economic innovation, scientific discovery, and artistic achievement [4] is transformational creativity, which involves altering existing conceptual spaces in ways that radically transform the scope of conceivable artifacts.[10] Recent work has begun to explore both empirical evaluation [48, 34, 41], as well as its theoretical foundations [40], both of which should inform future task design for LLMs.

6.1 Limitations

Before concluding, several limitations of our work are worth discussing. First, some benchmarks in our suite have $n < 25$ models with full coverage of the capability stack due to limited coverage on already-evaluated models (see Table 1). Validity and specificity should be interpreted cautiously here, and future work can aim to expand benchmark coverage to ensure greater statistical reliability. Although we study a large set of LLMs, this pool consists exclusively of instruction-tuned LLMs, and our findings may not generalize to base models, as post-training often reduces diversity [54]. Lastly, even though semantic distance tests offer convenient and controllable scoring, there are inherent limitations to how much variance in creative achievement minimal assessments can capture, due to the complex, long-horizon, and open-ended nature of real-world tasks. [19] has recently called for *open-world evaluations*: "long-horizon, messy, real-world tasks assessed through small-sample qualitative analysis rather than benchmark-scale automation." Accordingly, creative achievement may be best measured through a combination of minimal tests and open-ended, long-horizon assessments.

7 Conclusion

In this work, we conducted the first systematic study of how well semantic distance tests predict creative achievement in LLMs. We find that no test significantly predicts output diversity, suggesting that point and distributional creativity may dissociate in LLMs, a pattern we call ***jagged creativity***. The answer to the question in the title of this work depends on which test is being used and which construct it aims to predict: while the DAT *does* predict creative writing abilities in LLMs, the RAT is effectively a general capability proxy, and no single test predicts all constructs well. Overall, future work should exercise caution when applying such tests to make claims about the creativity of LLMs.

[9] We find no evidence of a large or moderate, capability-independent association between any existing test and scientific ideation; at our current coverage ($n = 20$–22) the analysis is sensitive only to $|r| \geq 0.56$–0.58, so a small association cannot be excluded. However, the same tests detect large associations with creative writing at similar sample sizes, so the absence of significant correlations for scientific ideation indicates a weaker (if any) relationship, not merely low power.

[10] Einstein's relativity theory and Copernicus's heliocentric theory are two canonical examples [40].

Acknowledgements

The authors thank Roger E. Beaty, Babak Hemmatian, and Lav R. Varshney for helpful feedback in preparing the manuscript.

A.G: This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 21-46756. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

S.S: This material is based upon work supported by the Lambda Grant Research Program.

References

[1] John Baer. How divergent thinking tests mislead us: Are the torrance tests still relevant in the 21st century? the division 10 debate. *Psychology of Aesthetics, Creativity, and the Arts*, 5(4):309–313, 2011. 4

[2] Roger E. Beaty, Daniel C. Zeitlen, Brendan S. Baker, and Yoed N. Kenett. Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity*, 41, September 2021. 2, 4, 7

[3] Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. arXiv preprint arXiv:2405.13012, 2025. 2, 3, 4

[4] Margaret A Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004. 2, 3, 4, 5, 10

[5] Edward M Bowden and Mark Jung-Beeman. Normative data for 144 compound remote associate problems. *Behavior research methods, instruments, & computers*, 35(4):634–639, 2003. 2, 6

[6] Honghua Chen and Nai Ding. Probing the creativity of large language models: Can models produce divergent semantic association? arXiv preprint arXiv:2310.11158, 2023. 2, 3, 4

[7] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024. 6

[8] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 4

[9] David Cropley. Is artificial intelligence more creative than humans?: Chatgpt and the divergent association task. *Learning Letters*, 2:13–13, 2023. 2, 3

[10] Arne Dietrich. The Cognitive Neuroscience of Creativity. *Psychonomic Bulletin & Review*, 11(6):1011–1026, 2004. 2

[11] Arne Dietrich. Types of creativity. *Psychonomic Bulletin and Review*, 26(1):1–12, February 2019. 2

[12] Anne Gerwig, Kirill Miroshnik, Boris Forthmann, Mathias Benedek, Maciej Karwowski, and Heinz Holling. The relationship between intelligence and divergent thinking—a meta-analytic update. *Journal of Intelligence*, 9(2):23, 2021. 4

[13] Lorena RR Gianotti, Christine Mohr, Diego Pizzagalli, Dietrich Lehmann, and Peter Brugger. Associative processing and paranormal belief. *Psychiatry and clinical neurosciences*, 55(6):595–603, 2001. 4

[14] Kurt Gray, Stephen Anderson, Eric Evan Chen, John Michael Kelly, Michael S Christian, John Patrick, Laura Huang, Yoed N Kenett, and Kevin Lewis. "Forward flow": A new measure to quantify free thought and predict creativity. *American Psychologist*, 74(5):539–554, 2019. 2, 4, 5

[15] Hyeonjeong Ha, Xiaomeng Jin, Jeonghwan Kim, Jiateng Liu, Zhenhailong Wang, Khanh Duy Nguyen, Ansel Blume, Nanyun Peng, Kai-Wei Chang, and Heng Ji. Synthia: Novel concept design with affordance composition. In *Proc. The 63rd Annual Meeting of the Association for Computational Linguistics (ACL2025)*, 2025. 10

[16] Zimeng Huang, Jinxin Ke, Xiaoxuan Fan, Yufeng Yang, Yang Liu, Liu Zhonghan, Zedi Wang, Junteng Dai, Haoyi Jiang, Yuyu Zhou, Keze Wang, and Ziliang Chen. MM-OPERA: Benchmarking open-ended association reasoning for large vision-language models. arXiv preprint arXiv:2510.26937, 2025. 2

[17] Mete Ismayilzada, Renqing Cuomao, Daniil Yurshevich, Anna Sotnikova, Lonneke van der Plas, and Antoine Bosselut. Cresowlve: Benchmarking creative problem-solving over real-world knowledge, 2026. 2

[18] Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025. 6, 21

[19] Sayash Kapoor, Peter Kirgis, Andrew Schwartz, Stephan Rabanser, J. J. Allaire, Rishi Bommasani, Harry Coppock, Magda Dubois, Gillian K Hadfield, Andrew B. Hall, Sara Hooker, Seth Lazar, Steve Newman, Dimitris Papailiopoulos, Shoshannah Tekofsky, Helen Toner, Cozmin Ududec, and Arvind Narayanan. Open-world evaluations for measuring frontier AI capabilities. arXiv preprint arXiv:2605.20520, 2026. 10

[20] Kyung Hee Kim. Meta-analyses of the relationship of creative achievement to both iq and divergent thinking test scores. *The Journal of Creative Behavior*, 42(2):106–130, 2008. 4

[21] Hong Li, Nanxi Li, Yuanjie Chen, Jianbin Zhu, Qinlu Guo, Cewu Lu, and Yong-Lu Li. The labyrinth of links: Navigating the associative maze of multi-modal LLMs. arXiv preprint arXiv:2410.01417, 2025. 2

[22] Mary Lou Maher. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, DESIRE '10, pages 22–28, 2010. 2, 5

[23] Lech Mazur. LLM creative story-writing benchmark (v2). https://github.com/lechmazur/writing, 2025. 6, 21

[24] Sarnoff Mednick. The associative basis of the creative process. *Psychological Review*, 69(3):220–232, 1962. 2, 3, 6, 9

[25] Tomáš Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018. 6

[26] Patrick J. Mineault, Thomas L. Griffiths, and Sean Escola. Cognitive dark matter: Measuring what AI misses. arXiv preprint arXiv:2603.03414, 2026. 3

[27] Saeid Alavi Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. Large Language Models are Fixated by Red Herrings: Exploring Creative Problem Solving and Einstellung Effect using the Only Connect Wall Dataset. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks*, 2023. 2

[28] Kumiko Nakajima, Jan Zuiderveld, and Sandro Pezzelle. Beyond divergent creativity: A human-based evaluation of creativity in large language models. *arXiv preprint arXiv:2601.20546*, 2026. 2, 5, 7, 18, 19

[29] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023. 4, 10

[30] Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118, June 2021. 2, 4, 5, 7, 23, 30

[31] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. 4

[32] Samuel J Paech. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*, 2023. 6, 21

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2, 5

[34] Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiusi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, and Heng Ji. EscapeBench: Towards advancing creative intelligence of language model agents. In *Proc. The 63rd Annual Meeting of the Association for Computational Linguistics (ACL2025)*, pages 798–820, 2025. 10

[35] Ziliang Qiu and Renfen Hu. Deep associations, high creativity: A simple yet effective metric for evaluating large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10870–10883, Suzhou, China, November 2025. Association for Computational Linguistics. 2, 4, 5, 6, 7, 18, 19

[36] Mel Rhodes. An analysis of creativity. *The Phi Delta Kappan*, 42(7):305–310, 1961. 3

[37] Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating llms' divergent thinking capabilities for scientific idea generation with minimal context. *Nature Communications*, 17, 2026. 7, 21

[38] Mark A. Runco and Garrett J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. 3

[39] Sameh Said-Metwaly, Christa L Taylor, Anaëlle Camarda, and Baptiste Barbot. Divergent thinking and creative achievement—how strong is the link? an updated meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*, 18(5):869–881, 2024. 4

[40] Samuel Schapiro, Jonah Black, and Lav R Varshney. Transformational Creativity in Science: A Graphical Theory. *arXiv preprint arXiv:2504.18687*, 2025. 10

[41] Samuel Schapiro, Alexi Gladstone, Heng Ji, and Roger E. Beaty. TwistBench: Benchmarking transformational creativity in LLMs via literary plot twists. In *Workshop on the Scientific Understanding of Foundation Models (Sci-FM) at the Conference on Language Modeling (COLM)*, 2026. 10

[42] Samuel Schapiro, Akshay K. Jagadish, Alexi Gladstone, Qiawen Ella Liu, Sumuk Shashidhar, Jonah Black, Royce Moon, Dilek Hakkani-Tur, Heng Ji, Claire E. Stevenson, Lav R. Varshney, Roger E. Beaty, Thomas L. Griffiths, and James A. Evans. Measuring combinatorial creativity in large language models and humans. arXiv:2509.21043, 2026. 7, 21

[43] Samuel Schapiro, Core Francisco Park, Felix Sosa, and Lav R. Varshney. CreativityNeuro: Steering language model weights to improve divergent thinking and reduce mode collapse. In *Conference on Language Modeling (COLM)*, 2026. arXiv:2607.01433. 3, 10

[44] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025. 10

[45] Dean Keith Simonton. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press, 2004. 5

[46] Dean Keith Simonton. The blind-variation and selective-retention theory of creativity: Recent developments and current status of bvsr. *Creativity Research Journal*, 35(3):304–323, 2023. 2

[47] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han Van Der Maas. Putting GPT-3's Creativity to the (Alternative Uses) Test. In *International Conference on Computational Creativity*, 2022. 2, 4

[48] Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, and Dawn Song. Omega: Can llms reason outside the box in math? evaluating exploratory, compositional, and transformative generalization. *arXiv preprint arXiv:2506.18880*, 2025. 10

[49] Paul Thagard. Creative Combination of Representations: Scientific Discovery and Technological Invention. In *Psychology of Science: Implicit and Explicit Processes*. Oxford University Press, 2012. 2

[50] L. R. Varshney. Mathematical limit theorems for computational creativity. *IBM Journal of Research and Development*, 63(1), January 2019. 5

[51] Dawei Wang, Difang Huang, Haipeng Shen, and Brian Uzzi. A large-scale comparison of divergent creativity in humans and large language models. *Nature Human Behaviour*, pages 1–10, 2025. 2, 3, 4, 18

[52] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024. 8, 21

[53] David Wechsler. *The Measurement of Adult Intelligence*. Williams & Wilkins, Baltimore, MD, 1939. 4

[54] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025. 10

[55] Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating language models for humanlike diversity. In *Conference on Language Modeling (COLM)*, 2025. 6, 21

A Derivation of the Validity-Specificity Frontier

The theoretical ceilings plotted in Figure 6 (benchmark validity-specificity frontier) are built on a covariance matrix bound that we prove here.

Theorem 1 (Validity-Specificity Frontier, restated). *Let $g = (Z_1, \dots, Z_k)$ be a vector of general capability proxies, let $\hat{Y}_g$ denote the ordinary least squares prediction of Y from g, and define $R = \mathrm{corr}(Y, \hat{Y}_g)$. For any test X with validity $v = \mathrm{corr}(X, Y)$, the semi-partial correlation $r(X, Y \mid g) := r(X,\, Y - \hat{Y}_g)$ satisfies*

$$|r(X, Y \mid g)| \;\le\; v\sqrt{1 - R^2} \;+\; |R|\,\sqrt{1 - v^2}.$$

Proof. Without loss of generality, we standardize X and Y to have unit variance, and let $\tilde{Z} := \hat{Y}_g / R$ so that $\sigma(\tilde{Z}) = 1$ as well. Since correlation is scale-invariant, $\mathrm{corr}(X, \tilde{Z}) = \mathrm{corr}(X, \hat{Y}_g) =: a$ and $\mathrm{corr}(Y, \tilde{Z}) = R$. The covariance matrix of the unit-variance triple $(X, Y, \tilde{Z})$,

$$\Sigma \;=\; \begin{pmatrix} 1 & v & a \\ v & 1 & R \\ a & R & 1 \end{pmatrix},$$

must be positive semi-definite. Expanding the determinant gives

$$\det(\Sigma) = 1 + 2vRa - a^2 - v^2 - R^2 \;\ge\; 0 \tag{5}$$

$$\Longrightarrow\; \underbrace{1}_{\alpha} a^2 \;-\; \underbrace{2vR}_{\beta} a \;+\; \underbrace{(R^2 + v^2 - 1)}_{\tau} \;\le\; 0. \tag{6}$$

Applying the quadratic formula to Equation (6) with coefficients α, β, τ yields

$$a \;\in\; \big[\, vR - \delta,\;\; vR + \delta \,\big], \qquad \delta \;:=\; \sqrt{(1 - R^2)(1 - v^2)}. \tag{7}$$

Finally, applying this into the semi-partial correlation:

$$r(X, Y \mid g) = r(X, Y - \hat{Y}_g) \tag{8}$$

$$= \frac{\mathrm{Cov}(X, Y - \hat{Y}_g)}{\sigma(X)\,\sigma(Y - \hat{Y}_g)} \tag{9}$$

$$= \frac{\mathrm{Cov}(X, Y) - \mathrm{Cov}(X, \hat{Y}_g)}{\sigma(Y - \hat{Y}_g)} \qquad (X \text{ has unit variance}) \tag{10}$$

$$= \frac{v - a\,\sigma(\hat{Y}_g)}{\sigma(Y - \hat{Y}_g)}. \tag{11}$$

The last step substitutes $v := \mathrm{Cov}(X, Y)$ and applies $\mathrm{Cov}(X, \hat{Y}_g) = \mathrm{corr}(X, \hat{Y}_g)\;\sigma(X)\;\sigma(\hat{Y}_g) = a \cdot 1 \cdot \sigma(\hat{Y}_g) = a\,\sigma(\hat{Y}_g)$, where the second equality uses $\mathrm{corr}(X, \hat{Y}_g) = a$ (by definition) and $\sigma(X) = 1$ (since we standardized X).

From here, it remains to address $\sigma(\hat{Y}_g)$ and $\sigma(Y - \hat{Y}_g)$. The former follows directly from our definition of $R := \mathrm{corr}(Y, \hat{Y}_g)$, along with the fact that least-squares residuals are orthogonal to their predictor (i.e, $\mathrm{Cov}(Y - \hat{Y}_g,\, \hat{Y}_g) = 0$):

$$R = \mathrm{corr}(Y, \hat{Y}_g) \tag{12}$$

$$= \frac{\mathrm{Cov}(Y, \hat{Y}_g)}{\sigma(Y)\,\sigma(\hat{Y}_g)} \tag{13}$$

$$= \frac{\mathrm{Cov}(\hat{Y}_g,\, \hat{Y}_g) + \mathrm{Cov}(Y - \hat{Y}_g,\, \hat{Y}_g)}{\sigma(Y)\,\sigma(\hat{Y}_g)} \tag{14}$$

$$= \frac{\mathrm{Var}(\hat{Y}_g)}{\sigma(Y)\,\sigma(\hat{Y}_g)} \qquad (\mathrm{Cov}(Y - \hat{Y}_g,\, \hat{Y}_g) = 0 \text{ by least-squares orthogonality}) \tag{15}$$

$$= \frac{\sigma(\hat{Y}_g)}{\sigma(Y)} = \sigma(\hat{Y}_g) \qquad (\mathrm{Var}(\hat{Y}_g) = \sigma^2(\hat{Y}_g) \text{ and } \sigma(Y) = 1). \tag{16}$$

Hence $\sigma(\hat{Y}_g) = R$ and $\sigma^2(\hat{Y}_g) = R^2$. Then, expanding $\sigma(Y - \hat{Y}_g)$ yields:

$$\sigma(Y - \hat{Y}_g) = \sqrt{\text{Var}(Y - \hat{Y}_g)} \tag{17}$$

$$= \sqrt{\text{Var}(Y) + \text{Var}(\hat{Y}_g) - 2\text{Cov}(Y, \hat{Y}_g)} \tag{18}$$

$$= \sqrt{1 + R^2 - 2 \cdot \text{corr}(Y, \hat{Y}_g) \cdot \sigma(\hat{Y}_g) \cdot \sigma(Y)} \qquad (\text{Since } \tfrac{\text{Cov}(Y,\hat{Y}_g)}{\sigma(Y)\sigma(\hat{Y}_g)} = \text{corr}(Y, \hat{Y}_g) =: R) \tag{19}$$

$$= \sqrt{1 + R^2 - 2R^2} \tag{20}$$

$$= \sqrt{1 - R^2} \tag{21}$$

Substituting back into (11) and combining with the bounds on a from Equation (7), we obtain

$$r(X, Y|g) = \frac{v - aR}{\sqrt{1 - R^2}} \tag{22}$$

$$\in \left[\frac{v - (vR + \delta)R}{\sqrt{1 - R^2}}, \quad \frac{v - (vR - \delta)R}{\sqrt{1 - R^2}} \right] \qquad \text{(from Eq. (7))} \tag{23}$$

$$= \left[v\sqrt{1 - R^2} - |R|\sqrt{1 - v^2}, \;\; v\sqrt{1 - R^2} + |R|\sqrt{1 - v^2} \right], \tag{24}$$

which gives the bound in Equation (4). □

Restated Remark on Theorem 1. In particular, a perfectly valid test ($v = 1$) has specificity bounded by $\sqrt{1 - R^2}$; a benchmark with R near ± 1 therefore leaves almost no specificity headroom for *any* test, regardless of construction. This is why Arena CW ($R = 0.87$ on $g =$ MMLU-Pro) caps specificity at ≈ 0.49 for a perfect-validity test, while NoveltyBench Utility ($R \approx -0.32$) admits a much wider ceiling (Figure 6, bottom row).

B Validity and Specificity for Every Test–Benchmark Pair

Table 1 reports the validity (r) and specificity ($r|g$, with g = MMLU-Pro) of every semantic distance test on every benchmark, which Figure 6 summarizes by construct and by benchmark. The Overall block uses each test's mean z-score across the three embeddings and is the main result referenced in Section 5. The GloVe, FastText, and Sentence-BERT blocks below it report each embedding separately as a robustness check. Benchmarks are grouped by construct, and the number of models behind each validity and specificity estimate is reported under each benchmark.

	Creative Writing						*Output Diversity*				*Combinatorial Creativity*		*Scientific Ideation*	
	Arena CW ↑ n=54/40		**EQ-Bench CW** ↑ n=35/27		**Mazur CW** ↑ n=21/18		**Hivemind Div.** ↑ n=25/21		**NovBench Util.** ↑ n=27/24		**Kombine** ↑ n=20/18		**LiveIdeaBench** ↑ n=28/22	
	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.
Overall (mean z-score across 3 embeddings)														
DAT	$\mathbf{+.49^{***}}$	$\mathbf{+.39^{*}}$	$\mathbf{+.76^{***}}$	$\mathbf{+.53^{**}}$	$\mathbf{+.61^{**}}$	$\mathbf{+.68^{**}}$	+.30	−.01	+.06	+.08	$\mathbf{+.69^{***}}$	+.32	+.23	+.20
CDAT	−.00	+.01	+.04	+.01	+.29	+.39	+.25	+.31	+.25	+.25	−.06	−.26	+.05	−.04
CDAT-N	−.12	−.15	−.04	+.05	+.14	+.24	+.20	+.34	+.32	+.31	−.21	−.19	−.03	−.15
CDAT-A	$\mathbf{+.50^{***}}$	$\mathbf{+.33^{*}}$	$\mathbf{+.42^{*}}$	+.13	+.21	+.04	−.35	−.37	−.32	−.31	$\mathbf{+.58^{**}}$	+.35	+.15	+.22
PACE	$\mathbf{+.71^{***}}$	$\mathbf{+.41^{*}}$	$\mathbf{+.71^{***}}$	+.30	$\mathbf{+.76^{***}}$	$\mathbf{+.52^{*}}$	−.05	+.07	+.10	+.17	$\mathbf{+.70^{***}}$	+.22	$\mathbf{+.44^{*}}$	+.28
RAT	$\mathbf{+.62^{***}}$	+.15	$\mathbf{+.44^{*}}$	+.08	$\mathbf{+.51^{*}}$	+.28	−.24	−.12	−.02	−.01	$\mathbf{+.58^{**}}$	+.14	+.23	+.07
GloVe 840B														
DAT	$\mathbf{+.40^{**}}$	+.29	$\mathbf{+.59^{***}}$	+.31	$\mathbf{+.49^{*}}$	$\mathbf{+.55^{*}}$	+.20	+.03	−.05	−.05	$\mathbf{+.66^{**}}$	+.35	+.26	+.16
CDAT	+.03	+.02	+.13	+.08	+.38	+.47	+.33	+.35	+.24	+.25	+.17	−.08	+.09	+.02
CDAT-N	−.10	−.13	+.02	+.11	+.23	+.31	+.23	+.36	+.31	+.31	−.01	−.06	+.00	−.11
CDAT-A	$\mathbf{+.50^{***}}$	$\mathbf{+.33^{*}}$	$\mathbf{+.40^{*}}$	+.08	+.13	−.04	−.36	−.38	−.31	−.30	$\mathbf{+.47^{*}}$	+.28	+.12	+.18
PACE	$\mathbf{+.72^{***}}$	+.31	$\mathbf{+.70^{***}}$	+.26	$\mathbf{+.71^{***}}$	+.46	−.13	+.03	+.06	+.14	$\mathbf{+.74^{***}}$	+.28	$\mathbf{+.43^{*}}$	+.28
FastText crawl-300d-2M														
DAT	$\mathbf{+.35^{**}}$	$\mathbf{+.36^{*}}$	$\mathbf{+.61^{***}}$	$\mathbf{+.47^{*}}$	$\mathbf{+.60^{**}}$	$\mathbf{+.64^{**}}$	+.30	−.16	+.14	+.17	$\mathbf{+.59^{**}}$	+.36	+.01	−.03
CDAT	+.00	+.05	+.02	−.02	+.16	+.29	+.19	+.23	+.21	+.21	−.26	−.43	+.04	−.05
CDAT-N	−.08	−.06	−.03	+.01	+.06	+.19	+.18	+.26	+.26	+.26	−.36	−.35	−.03	−.16
CDAT-A	$\mathbf{+.50^{***}}$	+.30	$\mathbf{+.44^{**}}$	+.17	+.27	+.09	−.36	−.34	−.30	−.29	$\mathbf{+.67^{**}}$	+.41	+.18	+.25
PACE	$\mathbf{+.72^{***}}$	$\mathbf{+.44^{**}}$	$\mathbf{+.71^{***}}$	+.33	$\mathbf{+.71^{***}}$	$\mathbf{+.53^{*}}$	−.13	+.02	+.09	+.15	$\mathbf{+.60^{**}}$	+.12	+.39	+.21
Sentence-BERT all-mpnet-base-v2														
DAT	$\mathbf{+.46^{***}}$	$\mathbf{+.32^{*}}$	$\mathbf{+.67^{***}}$	$\mathbf{+.52^{**}}$	+.42	$\mathbf{+.50^{*}}$	+.21	+.12	+.00	+.02	$\mathbf{+.57^{**}}$	+.14	+.28	+.38
CDAT	−.04	−.05	−.01	−.02	+.30	+.36	+.22	+.34	+.28	+.28	−.14	−.28	+.00	−.09
CDAT-N	−.18	−.24	−.11	+.03	+.12	+.20	+.19	+.38	+.36	+.35	−.28	−.17	−.06	−.16
CDAT-A	$\mathbf{+.51^{***}}$	$\mathbf{+.36^{*}}$	$\mathbf{+.43^{*}}$	+.15	+.23	+.08	−.32	−.38	−.35	−.33	$\mathbf{+.59^{**}}$	+.34	+.15	+.23
PACE	$\mathbf{+.60^{***}}$	$\mathbf{+.42^{**}}$	$\mathbf{+.65^{***}}$	+.30	$\mathbf{+.76^{***}}$	$\mathbf{+.51^{*}}$	+.18	+.19	+.15	+.20	$\mathbf{+.70^{***}}$	+.22	$\mathbf{+.44^{*}}$	+.33

Table 1: **Validity and specificity for every test–benchmark pair.** The **Overall** block (top) averages across the three embeddings and is the main result referenced in the text. For robustness, we also report per-embedding blocks; the RAT is scored by accuracy rather than embedding distance, so it appears only in the Overall block. Bold = significant at $p < .05$, while stars denote significance levels ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$). CDAT is novelty averaged only over the temperatures at which a model passes the appropriateness gate (Welch t-test against random nouns, BH-FDR across models, adjusted $p < .001$); CDAT-N is novelty over all temperatures, and CDAT-A is raw appropriateness. Lastly, sample sizes $n = \text{val}/\text{spec}$ are reported under each header, and specificity has reduced sample sizes because not every model has an MMLU-Pro score.

C Validity and Specificity by Sampling Temperature

The DAT and CDAT are administered at $T \in \{1.0, 1.5, 2.0\}$, following prior work [51, 28], while PACE is administered at $T = 0$ following [35]. Table 2 reports the validity and specificity of the DAT and CDAT at each temperature separately. Because per-temperature scores are computed under each test's primary embedding (GloVe for the DAT and Sentence-BERT for the CDAT), values differ slightly from the Overall block of Table 1. Model rankings are highly stable across temperatures (DAT $r \in [0.80, 0.93]$; CDAT $r \in [0.82, 0.89]$; all $p < 10^{-12}$). For reference, the benchmarks are collected at the following temperatures: NoveltyBench and Hivemind at $T = 1.0$, Kombine at $T = 0.9$, EQ-Bench CW and Arena CW at $T = 0.7$, LiveIdeaBench at the API default, and MMLU-Pro at $T = 0$, while Mazur CW does not document its temperature.

	Creative Writing						*Output Diversity*				*Combinatorial Creativity*		*Scientific Ideation*	
	Arena CW ↑ n=54/40		**EQ-Bench CW** ↑ n=35/27		**Mazur CW** ↑ n=21/18		**Hivemind Div.** ↑ n=25/21		**NovBench Util.** ↑ n=27/24		**Kombine** ↑ n=20/18		**LiveIdeaBench** ↑ n=26/21	
	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.
DAT *(GloVe 840B)*														
$T = 1.0$	$\mathbf{+.40^{**}}$	$\mathbf{+.34^{*}}$	$\mathbf{+.59^{***}}$	$+.32$	$\mathbf{+.50^{*}}$	$\mathbf{+.52^{*}}$	$+.28$	$+.08$	$-.13$	$-.13$	$\mathbf{+.72^{***}}$	$+.32$	$+.29$	$+.23$
$T = 1.5$	$\mathbf{+.47^{***}}$	$+.31$	$\mathbf{+.64^{***}}$	$+.25$	$\mathbf{+.52^{*}}$	$\mathbf{+.61^{*}}$	$+.09$	$-.07$	$-.05$	$-.10$	$\mathbf{+.63^{**}}$	$+.35$	$+.19$	$+.26$
$T = 2.0$	$\mathbf{+.44^{**}}$	$+.25$	$\mathbf{+.66^{***}}$	$+.29$	$+.39$	$\mathbf{+.57^{*}}$	$-.09$	$-.32$	$-.04$	$-.11$	$+.41$	$+.41$	$+.15$	$+.19$
CDAT *(Sentence-BERT all-mpnet-base-v2)*														
$T = 1.0$	$+.01$	$+.02$	$+.14$	$+.09$	$+.43$	$+.45$	$+.39$	$+.41$	$+.19$	$+.20$	$+.19$	$-.15$	$-.03$	$-.04$
$T = 1.5$	$-.10$	$-.10$	$-.05$	$-.09$	$+.11$	$+.13$	$+.27$	$+.35$	$\mathbf{+.41^{*}}$	$+.36$	$-.26$	$-.26$	$+.05$	$+.07$
$T = 2.0$	$+.11$	$+.16$	$-.06$	$-.18$	$+.25$	$+.39$	$-.16$	$-.10$	$+.15$	$+.08$	$-.44$	$-.29$	$+.08$	$+.02$

Table 2: **Validity and specificity of the DAT and CDAT at each sampling temperature.**

D Per-Model Test Scores

In Table 3, we report the mean ± SEM scores for the DAT, CDAT, CDAT-N, CDAT-A, and PACE per model. DAT is scored under GloVe across all valid trials at $T \in \{1.0, 1.5, 2.0\}$ (at 120 total trials per model). CDAT, CDAT-N, and CDAT-A are scored under Sentence-BERT `all-mpnet-base-v2`, where CDAT is novelty restricted to the temperatures passing the false discovery rate $\alpha = .001$ appropriateness gate ("—" if no temperature passed), and CDAT-N and CDAT-A are the ungated novelty and appropriateness scores aggregated over all 50 cues at all three temperatures. PACE is scored under FastText across all valid chains (up to 150 chains per model: 50 seeds × 3 chains).

As a sanity check, our per-model mean ± SEM values fall within the score ranges reported in the original papers. PACE across our 54 models spans ∼0.65–0.76, overlapping the ∼0.69–0.83 range reported by [35] across their 30-model set. Gated CDAT novelty across our models spans ∼62–75, within the ∼55–75 band shown in Figure 3 of [28]. Our evaluation extends both sets to a larger number of models.

Table 3: Mean ± SEM of DAT, CDAT, CDAT-N, CDAT-A, PACE, and RAT per model, grouped by provider. "—" indicates cells where no valid scores were collected (no temperature passed the appropriateness gate) on the CDAT, or responses were otherwise invalid.

Model	DAT	CDAT	CDAT-N	CDAT-A	PACE	RAT
OpenAI						
`gpt-3-5-turbo`	78.94±0.34	72.66±0.39	72.66±0.39	132.92±0.63	0.715±0.004	80.0
`gpt-4-1`	86.29±0.27	70.10±0.42	70.10±0.42	140.86±0.55	0.744±0.002	96.7
`gpt-4-1-mini`	81.60±0.49	67.75±0.58	67.75±0.58	143.78±0.66	0.730±0.003	80.0
`gpt-4-1-nano`	83.73±0.41	70.27±0.77	71.28±0.72	137.84±0.94	0.707±0.004	46.7
`gpt-4-turbo`	87.90±0.35	66.92±0.57	66.92±0.57	144.02±0.67	0.732±0.003	93.3
`gpt-4o`	84.47±0.43	66.26±0.69	66.26±0.69	144.54±0.78	0.729±0.002	93.3
`gpt-4o-mini`	80.90±0.46	71.52±0.46	71.52±0.46	138.41±0.69	0.707±0.003	50.0
`gpt-5`	89.65±0.18	69.76±0.46	69.76±0.46	142.26±0.55	0.747±0.002	93.3
`gpt-5-4`	91.72±0.19	68.63±0.49	68.63±0.49	143.70±0.54	0.727±0.003	93.3
`gpt-5-4-mini`	84.06±0.32	65.47±0.52	65.47±0.52	145.88±0.59	0.734±0.002	73.3
`gpt-5-4-nano`	83.77±0.29	63.15±0.58	63.15±0.58	147.61±0.64	0.678±0.004	33.3
`gpt-5-mini`	82.92±0.36	68.02±0.49	68.02±0.49	143.90±0.57	0.741±0.003	96.7
`gpt-5-nano`	80.66±0.32	63.11±0.59	63.11±0.59	147.95±0.63	0.712±0.003	93.3
`o3`	89.46±0.24	70.06±0.44	70.06±0.44	142.25±0.54	0.748±0.004	93.3
`o3-mini`	76.37±0.42	65.67±0.61	65.67±0.61	144.17±0.71	0.715±0.003	70.0
`o4-mini`	84.44±0.36	68.77±0.46	68.77±0.46	142.61±0.58	0.732±0.003	93.3
Anthropic						
`claude-3-5-haiku`	87.36±0.22	67.95±0.43	67.95±0.43	141.52±0.52	0.726±0.002	63.3
`claude-3-haiku`	78.87±0.23	62.53±0.58	62.53±0.58	146.22±0.65	0.697±0.005	70.0
`claude-haiku-4-5`	85.74±0.26	67.80±0.51	67.80±0.51	143.71±0.65	0.667±0.010	83.3
`claude-opus-4-5`	89.26±0.21	69.90±0.38	69.90±0.38	143.03±0.46	0.742±0.003	86.7
`claude-opus-4-6`	91.24±0.16	66.84±0.54	66.84±0.54	146.49±0.59	0.750±0.002	86.7
`claude-sonnet-4`	86.69±0.16	69.78±0.37	69.78±0.37	141.69±0.53	0.739±0.004	96.7
`claude-sonnet-4-5`	86.67±0.17	66.43±0.48	66.43±0.48	146.01±0.56	0.756±0.002	90.0
`claude-sonnet-4-6`	89.63±0.16	69.80±0.32	69.80±0.32	143.38±0.43	0.755±0.002	86.7
Google						
`gemini-2-0-flash-001`	82.32±0.41	70.32±0.41	70.32±0.41	139.44±0.55	0.730±0.003	86.7
`gemini-2-5-flash`	76.68±0.54	68.54±0.49	68.54±0.49	143.39±0.58	0.742±0.002	76.7
`gemini-2-5-pro`	89.69±0.25	71.13±0.34	71.13±0.34	139.02±0.49	0.761±0.002	93.3
`gemma-2-27b-it`	81.87±0.44	70.92±0.61	70.92±0.61	138.29±0.85	0.694±0.008	50.0
`gemma-2-9b-it`	80.40±0.53	71.96±0.48	72.77±0.51	133.71±0.73	0.728±0.003	—
`gemma-3-27b-it`	86.49±0.25	72.05±0.33	72.05±0.33	137.58±0.51	0.728±0.005	73.3
Meta						
`llama-3-1-70b-instruct`	87.71±0.30	68.19±0.80	68.19±0.80	141.56±0.98	0.713±0.004	83.3
`llama-3-1-8b-instruct`	88.48±0.48	69.82±0.88	72.99±0.78	134.84±1.09	0.701±0.006	50.0
`llama-3-2-1b-instruct`	81.20±0.25	58.92±1.33	58.92±1.33	146.46±1.67	0.586±0.017	0.0
`llama-3-2-3b-instruct`	84.11±0.14	68.31±0.50	68.31±0.50	139.37±0.67	0.711±0.005	33.3
`llama-3-3-70b-instruct`	88.05±0.37	68.22±0.70	72.25±0.71	135.47±0.94	0.718±0.004	80.0
`llama-4-maverick`	85.28±0.26	67.45±0.44	67.45±0.44	141.98±0.56	0.707±0.005	76.7
`llama-4-scout`	85.91±0.28	67.39±0.50	67.39±0.50	141.42±0.61	0.696±0.005	73.3
Mistral						
`mistral-7b-instruct-v0-1`	81.20±0.01	69.16±0.54	69.16±0.54	135.43±0.56	0.613±0.011	16.7
`mistral-large-2407`	88.15±0.24	70.36±0.34	70.36±0.34	138.80±0.55	0.737±0.002	96.7
`mistral-large-2411`	81.91±0.40	64.54±0.49	64.54±0.49	146.01±0.54	0.722±0.003	73.3
`mistral-nemo`	84.19±0.77	69.45±0.97	71.09±0.95	137.00±1.28	0.709±0.005	43.3
`mistral-small-24b-instruct-2501`	84.40±0.59	73.10±0.78	73.56±0.76	132.24±0.94	0.717±0.003	70.0
Qwen						

(continued on next page)

(continued from previous page)

Model	DAT	CDAT	CDAT-N	CDAT-A	PACE	RAT
`qwen-2-5-72b-instruct`	79.79±0.80	68.51±0.81	68.98±0.82	138.68±0.92	0.703±0.004	66.7
`qwen3-14b`	81.36±1.76	—	73.89±1.20	131.84±1.70	0.606±0.017	46.7
`qwen3-235b-a22b`	84.95±0.45	68.63±0.62	68.63±0.62	142.18±0.78	0.725±0.003	90.0
`qwen3-32b`	85.18±1.57	72.56±2.30	74.54±1.34	135.62±1.72	0.658±0.019	70.0
`qwen3-8b`	83.31±0.35	68.53±0.45	68.53±0.45	141.87±0.60	0.694±0.004	40.0
`qwq-32b`	82.63±0.50	—	—	—	—	—
DeepSeek						
`deepseek-chat`	81.12±0.40	67.50±0.53	67.50±0.53	143.11±0.71	0.729±0.003	80.0
`deepseek-chat-v3-0324`	84.80±0.43	70.63±0.44	70.63±0.44	137.63±0.66	0.730±0.004	73.3
`deepseek-r1`	83.80±0.39	69.28±0.45	69.28±0.45	141.38±0.65	0.720±0.003	60.0
Cohere						
`command-a`	82.28±0.33	63.22±0.64	63.22±0.64	147.80±0.66	0.714±0.003	73.3
`command-r-plus-08-2024`	87.69±0.45	69.44±0.49	69.44±0.49	138.70±0.64	0.722±0.003	83.3
NVIDIA						
`llama-3-1-nemotron-70b-instruct`	88.21±0.41	69.62±0.77	69.62±0.76	140.32±0.93	0.638±0.013	—
Microsoft						
`phi-4`	82.43±0.81	69.74±0.83	73.54±0.78	132.55±1.06	0.680±0.006	63.3

E Per-Model Benchmark Scores

Table 4 reports each model's score on every external benchmark used in this study. The general capability proxy is MMLU-Pro accuracy [52], the creative writing benchmarks are Arena CW (Elo), EQ-Bench CW v3 (Elo) [32], and Mazur CW [23], the output diversity benchmarks are Hivemind diversity ($1-$ intra-model cosine similarity) [18] and NoveltyBench Utility (NovB.; cumulative-utility score) [55], the combinatorial creativity benchmark is the Kombine overall score [42], and the scientific ideation benchmark is the LiveIdeaBench [37] *idea score average* across five dimensions: originality, flexibility, feasibility, clarity, and fluency. Cells marked "—" are missing because the corresponding benchmark does not score that model—future work can expand the coverage of these benchmarks to improve the statistical robustness of observed correlations.

Mazur CW snapshot. Mazur CW scores are transcribed from commit `80b7f17` (an absolute 0–10 mean-rubric leaderboard with 50 graded models). The leaderboard has since been regraded with a new ensemble of judges, eventually deprecating the absolute-rating system entirely in favor of a pairwise Thurstone ranking. We retained this snapshot rather than the latest version because it covers the largest set of models intersecting our evaluation pool ($n = 20$ overlap, vs. ≤ 10 for any later snapshot), keeping per-cell sample sizes comparable to the other creative writing benchmarks.

MMLU-Pro source. MMLU-Pro accuracies are taken from the TIGER-Lab MMLU-Pro leaderboard (CSV at `TIGER-Lab/mmlu_pro_leaderboard_submission`), which is methodologically tied to the original MMLU-Pro paper [52]. Models without a leaderboard entry receive no MMLU-Pro value and are excluded from the specificity computation for that cell, which is why the per-benchmark specificity sample sizes are smaller than the validity sample sizes (Table 1).

Hivemind mean-similarity estimate. [18] does not publish a mean intra-model similarity for each model. Instead, Table 6 in their paper reports, for each of 79 models, the percentage p_b of response pairs whose pairwise cosine similarity falls into each of ten bins covering $[0, 1]$ in 0.1-wide steps. We estimate the per-model mean similarity as the bin-midpoint-weighted sum, and report "Hivemind diversity" as $1 -$ intra-sim.[11], where intra-sim is given by:

$$\text{intra-sim} \;:=\; \tfrac{1}{100}\sum_{b=1}^{10} p_b \cdot m_b, \qquad m_b \in \{0.95, 0.85, \ldots, 0.05\}, \tag{25}$$

Table 4: **Per-model benchmark scores.** Columns: MMLU-Pro (accuracy), Arena CW (Elo), EQ-Bench CW v3 (Elo), Mazur CW, Hivemind diversity, NoveltyBench Utility, Kombine, LiveIdeaBench (5-dim Average). "—" indicates the corresponding benchmark does not score that model.

Model	MMLU-Pro	Arena CW	EQ-B. CW	Mazur	Hive.	NovB.	Kombine	LiveIdea
OpenAI								
`gpt-3-5-turbo`	—	1187	519	—	—	—	—	—
`gpt-4-1`	0.82	1402	1419	—	—	—	29.8	—
`gpt-4-1-mini`	—	1349	1231	—	—	—	—	—
`gpt-4-1-nano`	—	1306	1034	—	—	—	—	—
`gpt-4-turbo`	0.64	1322	—	—	0.14	—	—	6.55
`gpt-4o`	0.75	1423	1484	8.18	0.12	3.27	—	7.32
`gpt-4o-mini`	0.63	1295	950	6.72	0.11	3.11	24.9	6.25
`gpt-5`	0.87	1376	1301	8.60	—	—	47.6	—
`gpt-5-4`	0.88	1429	2019	—	—	—	—	—
`gpt-5-4-mini`	—	1417	1726	—	—	—	—	—
`gpt-5-4-nano`	—	1336	—	—	—	—	—	—
`gpt-5-mini`	—	1326	1301	8.31	—	—	35.6	—
`gpt-5-nano`	—	1250	866	—	—	—	—	—
`o3`	0.85	1384	—	8.39	—	—	—	—
`o3-mini`	0.79	1301	—	6.15	0.17	—	—	6.51
`o4-mini`	—	1338	—	7.50	—	—	—	—
Anthropic								
`claude-3-5-haiku`	0.62	1303	1241	7.35	0.11	2.50	—	6.79

(continued on next page)

[11] This is exact when within-bin similarity values are concentrated at the bin midpoint and approximate (with worst-case error ± 0.05) under uniform within-bin distributions; the rank order of models is, nonetheless, preserved unless within-bin distributions differ substantially across models

(continued from previous page)

Model	MMLU-Pro	Arena CW	EQ-B. CW	Mazur	Hive.	NovB.	Kombine	LiveIdea
claude-3-5-sonnet	0.78	—	—	—	—	2.36	—	6.92
claude-3-7-sonnet	0.78	—	—	—	—	—	—	7.12
claude-3-7-sonnet_thinking	0.78	—	—	—	—	—	—	7.22
claude-3-haiku	0.42	1214	848	—	0.13	2.94	—	—
claude-3-opus	0.68	—	—	—	—	2.67	—	6.36
claude-haiku-4-5	—	1384	—	—	—	—	—	—
claude-opus-4-5	0.87	1462	1769	—	—	—	44.4	—
claude-opus-4-6	0.89	1467	1965	—	—	—	52.1	—
claude-sonnet-4	0.84	1384	1516	8.09	—	—	—	—
claude-sonnet-4-5	0.87	1450	1777	—	—	—	39.9	—
claude-sonnet-4-6	0.87	1444	1991	—	—	—	44.2	—
Google								
gemini-1-5-pro	0.70	—	—	—	—	2.73	—	—
gemini-2-0-flash-001	0.78	1346	1252	7.15	0.15	3.17	—	7.48
gemini-2-0-flash-lite-001	0.72	1345	—	—	—	3.20	—	6.60
gemini-2-0-pro	0.79	—	—	—	—	2.64	—	—
gemini-2-5-flash	—	1399	1255	7.65	—	—	32.0	—
gemini-2-5-pro	0.86	1448	1415	8.38	—	—	31.7	—
gemma-2-27b-it	0.57	1291	—	—	0.16	3.77	—	6.66
gemma-2-2b-it	0.16	—	—	—	—	4.63	—	—
gemma-2-9b-it	0.52	1257	920	—	0.17	3.93	—	—
gemma-3-27b-it	0.68	1348	1256	7.99	—	2.26	—	—
Meta								
llama-3-1-405b-instruct	0.73	—	—	—	—	3.39	—	6.62
llama-3-1-70b-instruct	0.63	1257	851	—	0.18	3.72	—	6.62
llama-3-1-8b-instruct	0.44	1178	840	—	0.19	3.76	—	—
llama-3-2-1b-instruct	0.12	—	—	—	—	2.81	—	—
llama-3-2-3b-instruct	0.22	1144	—	—	0.20	3.24	—	—
llama-3-3-70b-instruct	0.66	1286	—	—	0.13	2.87	37.1	6.06
llama-4-maverick	0.81	1307	944	6.20	—	3.24	26.2	—
llama-4-scout	0.74	1290	899	—	—	3.14	—	—
Mistral								
mistral-7b-instruct-v0-1	0.26	1104	—	—	0.19	3.28	—	—
mistral-large-2407	0.66	1287	1400	6.90	—	—	—	—
mistral-large-2411	0.68	1276	1082	6.90	0.14	—	—	6.79
mistral-nemo	0.45	1090	966	—	0.20	4.00	—	7.24
mistral-small-24b-instruct-2501	0.66	1227	—	—	0.18	3.29	—	6.41
Qwen								
qwen-2-5-72b-instruct	0.72	1254	—	—	0.14	3.09	27.1	6.62
qwen-2-5-7b-instruct	—	—	—	—	—	—	—	6.47
qwen-2-5-coder-32b-instruct	—	—	—	—	—	—	—	6.56
qwen-max	0.76	—	—	—	—	—	—	6.59
qwen3-14b	—	—	—	—	0.13	—	—	—
qwen3-235b-a22b	0.83	1324	1379	8.30	—	3.89	—	—
qwen3-32b	—	1306	—	—	0.15	—	—	—
qwen3-8b	—	—	—	—	0.12	—	—	—
qwq-32b	0.69	1296	1262	8.02	—	—	—	7.06
DeepSeek								
deepseek-chat	0.76	1349	—	—	0.14	3.12	29.1	6.72
deepseek-chat-v3-0324	0.81	1390	1473	7.70	0.14	3.20	23.3	—
deepseek-r1	0.84	1374	1500	8.30	—	—	34.5	7.18
deepseek-r1-distill-llama-70b	—	—	—	—	—	—	—	6.71
deepseek-r1-distill-qwen-32b	—	—	—	—	—	—	—	6.86
Cohere								
command-a	—	1337	1184	—	—	—	—	—
command-r-08-2024	—	1209	—	—	—	2.98	—	—
command-r-plus-08-2024	—	1263	—	—	0.23	3.08	—	—
command-r7b-12-2024	—	—	—	—	—	3.35	—	—
NVIDIA								
llama-3-1-nemotron-70b-instruct	0.63	1277	—	—	—	—	—	7.25
Microsoft								
phi-4	0.70	1210	—	6.26	0.15	3.59	18.5	6.72
Amazon								
nova-lite-v1	—	—	—	—	—	—	—	6.05
nova-pro-v1	—	—	—	—	—	—	—	6.58

F Greedy Algorithm for the DAT

The greedy algorithm in Figure 4a starts with a random word from a valid vocabulary set V (e.g., the $\sim$42,000 single-token English nouns in WordNet $\cap$ GloVe 840B), then proceeds to *minimize the mean cosine similarity* (equivalently, *maximize the mean cosine dissimilarity*, matching Equation (1)) of all subsequent words, while avoiding any repetition.

Figure 4 reports the result of running the algorithm for 120 independent seeds, and scoring each sequence according to Equation (1) under GloVe, FastText, and Sentence-BERT. The three scores are averaged per trial and across embeddings to produce the values in Figure 4 (mean $= 94.1$, std. dev. $= 0.9$). The LLM pool, averaged the same way over 2,082 trials from 55 models, has mean $= 81.1$ (std. dev. $= 3.0$), and humans mean $= 78.4$ (std. dev. $= 6.4$); human scores are on the GloVe scale only, because [30] reports only final scores—not raw word lists. As shown, the algorithm substantially exceeds the mean LLM and human scores.

G Evidence for Jagged Creativity

This section gives evidence for jagged creativity (Section 5) from raw DAT responses. Table 5 shows example responses from three models at each sampling temperature: `gpt-5-4` and `claude-sonnet-4-6`, two of the highest-scoring models on the DAT, repeat the same words across nearly all of their 40 responses, even at $T = 2.0$, while `command-r-plus-08-2024` rarely repeats a word. Table 6 compares five base models, given the DAT as a text completion, with their instruction-tuned versions: instruction tuning raises the mean DAT score from 77.4 to 81.4, but cuts the number of distinct words a model uses across 40 responses by about two-thirds (from 296 to 97 on average).

Table 5: **Repeated words in DAT responses across temperatures.** Two responses per model and temperature. Bold words appear in at least half of the model's 40 responses at that temperature, and Distinct counts unique words across all 40 (out of 400).

Model	T	Top word	Distinct	Example runs
`gpt-5-4`	1.0	quasar (39/40)	98	**quasar**, **moss**, anvil, melody, verdict, bacterium, glacier, silk, hunger, **algorithm** **quasar**, thimble, maggot, **justice**, tornado, sapphire, fungus, **algorithm**, canyon, lullaby
DAT rank 1 of 55	1.5	quasar (38/40)	104	**quasar**, teaspoon, democracy, fungus, ankle, silence, volcano, **algorithm**, perfume, cradle **quasar**, spoon, **moss**, justice, volcano, **algorithm**, perfume, skeleton, thunder, market
	2.0	quasar (40/40)	99	**quasar**, thimble, **moss**, equation, volcano, violin, citizen, teaspoon, nightmare, antibody **quasar**, **moss**, anvil, harmony, bacterium, labyrinth, teardrop, currency, skeleton, thunder
`claude-sonnet-4-6`	1.0	justice (40/40)	131	**justice**, membrane, avalanche, spoon, frequency, **debt**, fossil, whisper, throne, **fungus** **justice**, kelp, turbine, melancholy, socket, chromosome, fable, gravel, oboe, **debt**
DAT rank 5 of 55	1.5	justice (40/40)	117	**justice**, membrane, glacier, tambourine, **debt**, spore, labyrinth, oxygen, anvil, grief **justice**, **fungus**, telescope, grief, anvil, chromosome, tide, labyrinth, ember, **debt**
	2.0	justice (40/40)	124	**justice**, **fungus**, lever, sorrow, pixel, anvil, treaty, cobweb, neuron, glacier **justice**, **fungus**, locomotive, **grief**, pixel, hammock, **debt**, chromosome, avalanche, flute
`command-r-plus-08-2024`	1.0	galaxy (7/40)	351	apple, soul, revolution, rainbow, politics, geology, purveyor, inequality, epiphany, pandemic entropy, peace, hologram, echo, awareness, petrichor, capybara, insurrection, titanium, constellation
DAT rank 14 of 55	1.5	galaxy (7/40)	359	carrot, pencil, painting, friendship, sandwich, scientist, curtain, meaning, song, threshold truth, cherry, imperative, neutron, shone, alchemy, oblong, grandiose, bubbly, pajamas
	2.0	galaxy (6/40)	359	abundance, charity, criminal, equinox, hatred, innocence, knowledge, luggage, molecule, ocean cephalopod, hatred, whistle, jubilee, avocado, pyramid, nimbus, saunter, bewilder, annihilate

Table 6: **Base vs. instruction-tuned models on the DAT ($T = 1.0$, top-$p = 1$, top-$k = 0$).** Base models receive the same DAT prompt framed as a text completion. Distinct counts unique words across 40 responses (out of 400), and the most repeated word is shown with the number of responses containing it.

	DAT score		Distinct words		Most repeated word	
Model	Base	Instruct	Base	Instruct	Base	Instruct
Llama 3.2 3B	82.0	84.0	341	13	food (3/40)	basket (40/40)
Llama 3.1 8B	83.0	87.6	331	236	car (3/40)	fjord (19/40)
Qwen2.5 7B	79.4	76.1	298	120	book (7/40)	ocean (33/40)
Mistral 7B v0.1	77.2	81.2	351	15	dog (5/40)	apple (40/40)
Gemma 2 9B	65.5	77.9	159	99	cat (11/40)	silence (35/40)
Mean $\pm$ SEM	77.4$\pm$3.1	81.4$\pm$2.1	296$\pm$35	97$\pm$41	(5.8$\pm$1.5/40)	(33.4$\pm$3.9/40)

H Prompts

Below, we report the exact prompts used to administer the DAT, CDAT, PACE, and RAT. On CDAT, the cue word varies across 50 cues drawn from semantically diverse categories (Physical world, Animals, Food & drink, Body, Emotions, . . .); for PACE the seed word varies across 50 seeds from the same category set. We show one instantiation with cue/seed `"rock"` for illustration.

H.1 DAT

```
Please enter 10 words that are as different from each other as possible, in all
meanings and uses of the words. Only use single nouns. Do not use proper nouns
(names, places, brands). Do not use variations of the same word (e.g., don't
use both 'run' and 'running').

Respond with ONLY a JSON array of exactly 10 words, like: ["word1", "word2",
"word3", "word4", "word5", "word6", "word7", "word8", "word9", "word10"]
```

H.2 CDAT (cue: rock)

```
Please enter 10 words that are as different from each other as possible, in all
meanings and uses of the words, yet semantically associated with the following
cue word: "rock". Only use single nouns. Do not use proper nouns. Do not
use the cue word itself or variations of it. Respond with ONLY a JSON array of
exactly 10 words, like: ["word1", "word2", "word3", "word4", "word5", "word6",
"word7", "word8", "word9", "word10"]
```

H.3 PACE Stage 1 (seed: rock)

```
Starting with the word "rock", generate three different words that directly
associate with this initial word only (not with each other). Please put down
only single words, and do not use proper nouns (such as names, brands, etc.).
For each word, provide a brief explanation of its connection to "rock". Return
in JSON format:

{"results": [{"word": "", "reason": ""}, {"word": "", "reason": ""},
{"word": "", "reason": ""}]}
```

H.4 PACE Stage 2 (seed: rock, first-association: stone)

```
Starting with the word pair "rock" -> "stone", generate a chain of 20 words
where each new word should be associated with ONLY the word immediately before
it. Generate the third word based on "stone", then generate the fourth word
based on your third word, and so on. Please put down only single words, and do
not use proper nouns (such as names, brands, etc.). For each word, provide a
brief explanation of its connection to the previous word. Return in JSON format
with exactly 20 entries:

{"results": [{"word": "stone", "reason": "<stage-1 reason>"}, {"word": "",
"reason": ""}, ... ]}
```

H.5 RAT (item: cream, skate, water)

```
What single word can be combined with each of "cream", "skate", and "water" to
form a compound word or common phrase?

Respond with ONLY the single answer word in lowercase. No explanation.
```

I Disclosure of Large Language Model Usage

In this paper, large language models (LLMs) were used to assist in the code implementation, plotting and figure generation, and the reporting (but not analysis) of experiment results.

J Codebase Release

The full codebase — including the evaluation pipeline (DAT, CDAT, PACE, RAT), embedding scoring, correlation and partial-correlation analysis, the figure-generation scripts, and the per-model raw responses — is available at:

https://github.com/samjschapiro/do-semantic-distance-tests-actually-predict-creativity-neurips-26

K Compute Resources

All experiments were run as inference-time API calls through OpenRouter; no model training or fine-tuning was performed. Total spend across the full eval campaign (DAT, CDAT, PACE, and RAT on 55 instruction-tuned LLMs at multiple temperatures, including iteration during pipeline development) was approximately $150 in API charges. Embedding-based scoring (GloVe, FastText, Sentence-BERT) and all downstream correlation and figure generation ran on a single laptop CPU; per-pipeline wall-clock is dominated by API latency.

L Broader Impacts

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The findings stated in the abstract and Section 1 (no test predicts output diversity, suggesting jagged creativity; the RAT is a capability proxy; the DAT best predicts creative writing but no single test predicts all constructs) match the results in Section 5, and the theoretical bound is proved in Theorem 1 / Section A.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6.1 discusses the small n of several benchmarks, the restriction of the model pool to instruction-tuned LLMs (so findings may not transfer to base models), and the limits of minimal tests relative to open-world evaluations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Theorem 1 states all assumptions explicitly; the proof is given in Section A.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 specifies the model pool, sampling parameters, embedding models, and partial-correlation computation. Exact prompts for the DAT, CDAT, PACE, and RAT are given in Section H. Per-model raw test scores and benchmark scores are listed in Table 3 and Table 4.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The full codebase — evaluation pipeline, scoring, correlation analysis, and figure generation — is released at the URL given in Section J.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 gives temperatures, sampling parameters, the number of trials per model, the appropriateness gate threshold and FDR correction for CDAT, and the capability control (MMLU-Pro) used for specificity.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report two-sided p-values for all validity and specificity correlations using the standard Pearson t-test with $\mathrm{df} = n - 2 - k$ (Section 4), with significance levels marked in Table 1. Per-model scores in Table 3 are reported as mean $\pm$ SEM.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section K reports total OpenRouter API spend (approximately $150) and notes that downstream scoring and figure generation run on a single laptop CPU; no training or fine-tuning was performed.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses only public benchmark scores and inference-time API calls to commercially available LLMs; no human subjects data was collected, and no private model training was performed.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section L provides a brief broader-impacts statement.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any new model or scraped dataset.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All benchmarks (Arena CW, EQ-Bench CW, Mazur CW, Hivemind, NoveltyBench Utility, Kombine, LiveIdeaBench, MMLU-Pro), the RAT items, and embedding resources (GloVe, FastText, Sentence-BERT) are cited in Section 3 and Section 4. Source URLs and snapshot identifiers for the leaderboards used (Mazur GitHub commit, TIGER-Lab MMLU-Pro CSV, EQ-Bench JS payload) are documented in Section E.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the evaluation code together with every model's raw responses and scores, documented in a README that covers setup, the data layout, and how to reproduce each table and figure (Section J).

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing or human subjects; the human DAT distribution referenced in Figure 4 is taken from a published source ([30]).

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects were recruited for this work.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are the objects of study; the 55 models and their API settings are described in Section 4. Section I discloses how LLMs assisted this work (code implementation, figure generation, and reporting, but not analysis).