Do Semantic Distance Tests *Actually* Predict Creativity in Large Language Models?

Samuel Schapiro[♠◇*] **Alexi Gladstone**[♠] **Jonah Black**[♠] **Heng Ji**[♠]
[♠] University of Illinois, Urbana-Champaign [◇] Jean Technologies, Inc.

Creative AI Index: schapiro.ai/creative-ai-index

Abstract

Automated semantic distance tests—which prompt a model to produce a set of words, and score the average embedding distance between them—are increasingly used to measure the "creativity" of large language models. However, the validity of semantic distance tests as predictors of *machine* creativity has not yet been established, and these tests already have limited validity as predictors of human creativity. To address this problem, we conduct the first systematic study evaluating the effectiveness of semantic distance tests in predicting creative achievement across three constructs: creative writing, divergent thinking, and 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)** Test effectiveness varies significantly by construct, and no single test predicts all constructs well. **(2)** None of the tests is a good predictor of scientific ideation ability. **(3)** Existing tests are far below the theoretical limits, indicating meaningful room for the design of improved tests moving forward. Our findings provide clear practical takeaways and directions for future work and suggest that novel tests are needed to reliably predict scientific ideation ability.

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 [38, 5, 3]. In particular, the ability to associate semantically distant concepts has long been considered central to human creativity [19, 37], motivating the use of *automated* semantic distance assessments like the Divergent Association Task (DAT) [24, 5, 8, 3, 43] 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 [27], providing a convenient way to assess "creativity" without requiring human raters.

Over the last year, two novel semantic distance tests have also been proposed. The Conditional DAT (CDAT; [22]) extends the DAT by requiring dissimilar nouns to each be relevant to a "cue" word, incorporating a measure of appropriateness in addition to novelty [4, 17]. Similarly, the Parallel Association Chain Evaluation (PACE; [29]) is inspired by the forward-flow measure [11, 2] and instructs models to make free associations starting from several seed words, scoring responses by a sequential semantic distance measure. The PACE test shows strong Spearman rank correlations with creative writing benchmarks ($\rho \approx 0.74$). However, it also correlates strongly with general model

*Correspondence to sjs17@illinois.edu.

Preprint.

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 validity of semantic distance tests as measures of machine creativity has not been established, and semantic distance tests have limited validity as predictors of human creativity as it stands (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 six benchmarks that capture three target constructs: (i) creative writing, (ii) divergent thinking, and (iii) scientific ideation. After finding that correlations between benchmarks and general capabilities are as high as $r = 0.98$, we introduce two evaluation criteria: *validity*, which measures 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 creativity tests predict aspects of creative achievement after variance explained by general capabilities is factored out, enabling more robust evaluation. In detail, we make the following contributions:

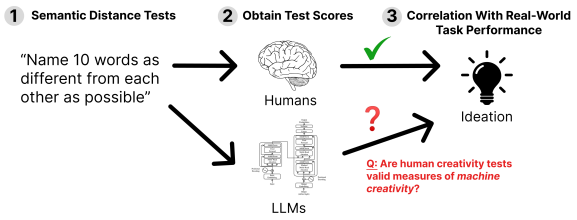

Figure 1: **Overview.** Semantic distance tests like the Divergent Association Task ask respondents to name a set of maximally dissimilar words and score the result by embedding-space distance. The external validity of these tests among humans has been supported by modest correlations with ideation tasks, 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?*

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*).
3. By evaluating specificity in addition to validity, we find that the Parallel Association Chain Evaluation (PACE) test is effectively a proxy for general capabilities—its highly significant validity on creative writing ($r \approx 0.73$) collapses to non-significant specificity ($r \approx 0.15$).
4. Our empirical findings indicate that the Divergent Association Task (DAT) is the best predictor of creative writing, the Conditional Divergent Association Task (CDAT) is the best predictor of divergent thinking, and in contrast to general beliefs, *none of the tests is a reliable predictor of scientific ideation ability.*
5. We prove an upper bound on the maximum attainable validity and specificity a creativity test can achieve, and find that semantic distance tests are far below the frontier on nearly all benchmarks.

Overall, our findings provide practical guidance on which constructs semantic distance tests are—and are not—able to measure, while indicating there is meaningful room for future work to design novel tests with superior predictive power.

2 Background and Motivating Problems

In recent years, semantic distance tests have been widely adopted to assess whether LLMs are capable of making distant associations [5, 8, 3]. Furthermore, various strategies, such as varying sampling parameters, instructing LLMs to assume the personas of distinguished creative individuals [43], prompt engineering, and offering explicit test-taking strategies [3], have all been explored to improve semantic distance test scores. Additionally, semantic distance tests are already being used to make direct claims that LLMs are *more* or *less* creative than humans [5, 3, 43], 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 [7] because such tests do not inherently possess construct validity for machines.[2] 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 [24] and forward flow[4] [11] are validated by correlation with the Alternative Uses Test (AUT): the DAT at $r \approx .32$–$.51$ [24][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 $r \approx .17$ [31] to $r \approx .22$ [15], 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 **creativity 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 creativity 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 $r = 0.37$ [9], the standard practice is to factor out confounding variables (e.g., through hierarchical regression or factor analysis) before claiming a test measures "creativity" [2]. Later in this work (Section 4), we propose a similar methodology for measuring the predictive power of creativity tests independent of what general model capabilities already predict.

DAT. Generate 10 nouns maximally different from each other.

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

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

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

Figure 2: **Example prompts and responses for each semantic distance test.** The DAT prompts for maximally distant nouns, the CDAT prompts for maximally distant nouns that are each relevant to a cue, and PACE prompts models to freely associate sequences of nouns in a 20-word chain.

3 Preliminaries

With these problems in mind, we start by introducing the semantic distance tests and large-scale benchmarks we use for evaluation in this work.

[2]The digit span test is commonly used to assess human intelligence [46], but would be inappropriate to administer to machines, where working memory is not a bottleneck to intelligence the way it is among humans. 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 [25, 23], and that the DAT itself can be trivially solved by a simple algorithm over embedding distances that exceeds mean human and LLM scores (Appendix D).

[3]Moreover, these tests may be flawed due to training data leakage, as has been observed with the AUT [38]

[4]PACE [29] uses the same scoring mechanism as forward flow.

[5]The DAT is also validated against the Bridge-the-Associative-Gap Test [10], although this is a measure of convergent thinking.

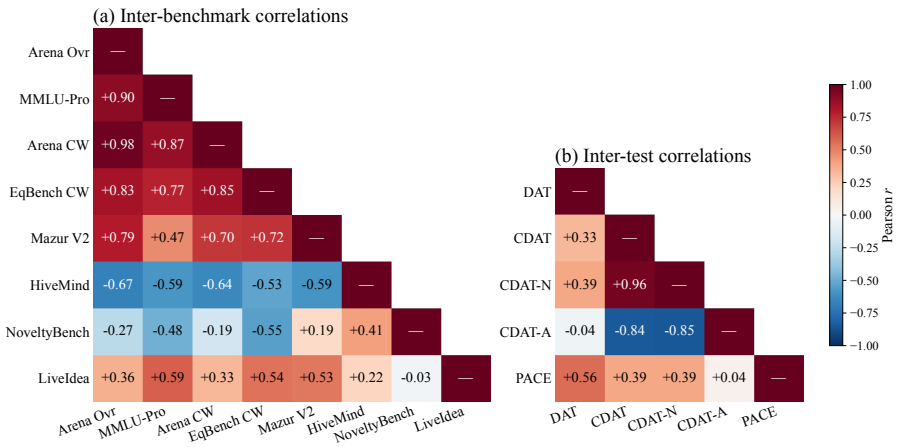

Figure 3: **Inter-benchmark and inter-test correlations. (a)** Benchmarks are ordered by construct: general capabilities (Arena Ovr, MMLU-Pro), creative writing (Arena CW, EqBench CW, Mazur CW), divergent thinking (Hivemind, NoveltyBench), and scientific ideation (LiveIdea). n per cell ranges from 8 (Mazur $\times$ Hivemind) to 54. **(b)** Inter-test correlations among DAT, CDAT, CDAT-N, CDAT-A, PACE using composite scores across GloVe, FastText, and SBERT.

3.1 Semantic Distance Tests

Divergent Association Task (DAT) Introduced by [24], 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 [27] used in [24], we test robustness under multiple embedding models in Section 4.

Conditional Divergent Association Task (CDAT) Creative artifacts should be both novel *and* useful [40, 4, 36, 17]. [22] 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 the CDAT-N is equivalent to Equation (1), but that CDAT-N and DAT scores differ due to cue-based prompting under the CDAT.

Parallel Association Chain Evaluation (PACE) PACE [29], inspired by forward-flow associative-chain methods [11], prompts models to produce three parallel 20-word association chains from a seed, and the score is the mean cumulative cosine distance along each chain. For a 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 metric was originally tested under FastText embeddings [20], although [29] also saw significant rank correlations with Chatbot Arena CW under multiple distinct embedding models.

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

3.2 Creative Achievement Benchmarks

We evaluate each semantic distance test against three target constructs (creative writing, divergent thinking, and scientific ideation) spanning six large-scale benchmarks.

3.2.1 Creative Writing

Chatbot Arena Creative Writing (Arena CW) [6]. Arena CW is a creative writing benchmark where models respond to open-ended, user-submitted creative writing prompts, and user preference ratings are aggregated across pairwise evaluations between models, leading to per-model Elo ratings. Arena CW is the most capability-loaded of the three creative writing benchmarks ($r = 0.98$ with Arena Overall; Figure 3).

EQ-Bench Creative Writing [26]. 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 Arena Overall ($r = 0.83$; Figure 3), although less than Arena CW.

Mazur Creative Writing [18]. 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 strongly with Arena Overall at $r = 0.79$, although less than both Arena CW and EQ-Bench.

All three are heavily capability-loaded, having correlations $r = 0.98/0.83/0.79$ with Arena Overall, respectively, motivating the use of both validity *and* specificity as evaluation criteria.

3.2.2 Divergent Thinking

We use two divergent thinking benchmarks, each of which measures output diversity in LLM responses to open-ended prompts.

Hivemind [13]. Hivemind is a divergent thinking 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 = -0.67$ with Arena Overall, $r - 0.59$ with MMLU-Pro).

NoveltyBench [48]. 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 the most capability-independent benchmark, obtaining $r = -0.27$ with Arena Overall and $r = -0.48$ with MMLU-Pro.

3.2.3 Scientific Ideation

LiveIdeaBench [30]. 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 in a brief response. Each generated 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

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

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 moderately correlated with general capabilities (Figure 3a; $r = 0.36$ on Arena Overall, $r = 0.59$ 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 ($n = 54$) 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 [29], 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 [22], 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 E.

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 of which is motivated and defined below.

Validity Standard psychometric criteria used in human creativity research [2, 24] 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 between a test score X and benchmark score Y, given by $r(X, Y)$.

Specificity Since creative achievement benchmarks are themselves highly capability-loaded (for example, Arena CW correlates with Arena Overall at $r = 0.98$) a high raw $r(X, Y)$ may simply reflect that the test tracks general capability. We therefore 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 Arena Overall Elo (preference-based) and MMLU-Pro accuracy [45] (knowledge- and reasoning-based). 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 $k = 2$ for Arena Overall + MMLU-Pro used for specificity).

	Creative Writing						*Divergent Thinking*				*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=14/10		**LiveIdeaBench ↑** n=17/17	
	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.	Valid.	Spec.
Overall (mean z-score across 3 embeddings)												
DAT	**+.47*****	+.05	**+.72*****	**+.41***	**+.60****	+.49	+.33	+.01	+.15	−.21	−.01	+.24
CDAT	−.13	+.22	−.06	+.03	+.07	+.43	+.25	+.10	+.60	+.60	+.03	+.21
CDAT-N	−.18	+.20	−.14	+.08	+.09	+.40	+.24	+.08	**+.54***	+.45	−.11	+.02
CDAT-A	**+.54*****	−.05	**+.48****	+.05	+.24	−.28	−.39	−.09	**−.67****	−.40	+.20	+.08
PACE	**+.71*****	+.11	**+.71*****	+.21	**+.76*****	+.14	−.05	+.33	+.18	−.00	+.11	−.00
GloVe 840B												
DAT	**+.37****	−.01	**+.57*****	+.28	+.43	+.39	+.21	+.04	−.22	−.46	−.20	+.04
CDAT	−.07	+.14	−.03	+.05	+.43	+.44	+.35	+.22	+.56	+.53	−.06	+.18
CDAT-N	−.20	+.11	−.16	+.04	+.16	+.46	+.26	+.10	+.49	+.30	−.15	−.02
CDAT-A	**+.54*****	−.00	**+.49****	+.04	+.17	−.34	**−.40***	−.09	**−.61***	−.30	+.20	+.10
PACE	**+.72*****	+.00	**+.70*****	+.15	**+.71*****	+.12	−.13	+.22	+.10	−.17	−.02	−.16
FastText crawl-300d-2M												
DAT	**+.35****	+.00	**+.59*****	+.29	**+.64****	**+.57***	+.32	−.11	+.32	+.26	−.06	+.16
CDAT	−.13	+.35	−.08	+.04	−.04	+.38	+.23	+.05	+.63	+.62	+.10	+.29
CDAT-N	−.13	+.32	−.10	+.05	+.01	+.34	+.21	+.04	**+.54***	+.46	−.04	+.12
CDAT-A	**+.51*****	−.09	**+.45****	+.05	+.30	−.23	**−.41***	−.08	**−.70****	−.45	+.17	+.03
PACE	**+.72*****	+.29	**+.71*****	+.25	**+.71*****	+.08	−.13	+.26	+.19	+.07	+.13	+.08
Sentence-BERT all-mpnet-base-v2												
DAT	**+.44*****	+.13	**+.63*****	**+.42***	**+.44***	+.29	+.25	+.11	+.11	−.12	+.29	+.37
CDAT	−.08	+.24	−.08	−.00	+.33	+.34	+.27	+.14	+.60	+.62	−.05	+.16
CDAT-N	−.21	+.13	−.15	+.14	+.08	+.37	+.23	+.10	**+.57***	+.50	−.16	−.04
CDAT-A	**+.54*****	−.04	**+.48****	+.05	+.24	−.25	−.37	−.10	**−.68****	−.42	+.22	+.11
PACE	**+.60*****	+.05	**+.65*****	+.23	**+.76*****	+.18	+.18	**+.47***	+.24	+.09	+.17	+.04

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. Green cells mark the best test per benchmark in the Overall block, bold = significant at $p < .05$, while stars denote significance levels ($^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$). Lastly, sample sizes $n = \text{val}/\text{spec}$ are reported under each header, and specificity has reduced sample sizes due to marginally lower model overlap.

4.3 Theoretical Limits for Attainable Validity and Specificity

What is the maximum attainable specificity a test can achieve, as a function of its validity and its benchmark's correlation with general capabilities? We refer to this as the *validity-specificity frontier*, and prove a bound on this quantity 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 creativity 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 ($R = 0.98$ on $g = (\text{Arena Overall}, \text{MMLU-Pro})$) caps specificity at ≈ 0.20 for a perfect-validity test, while NoveltyBench Utility ($R \approx -0.33$) admits a much wider ceiling (Figure 4, bottom row).

5 Results

In Table 1, we report the validity and specificity across all tests and benchmarks. Test and benchmark scores used for correlations are fully reported in Table 2 and Table 3, respectively. Figure 4 visualizes validity and specificity aggregated across the three constructs and plots the validity-specificity frontier we derived in Section 4.3.

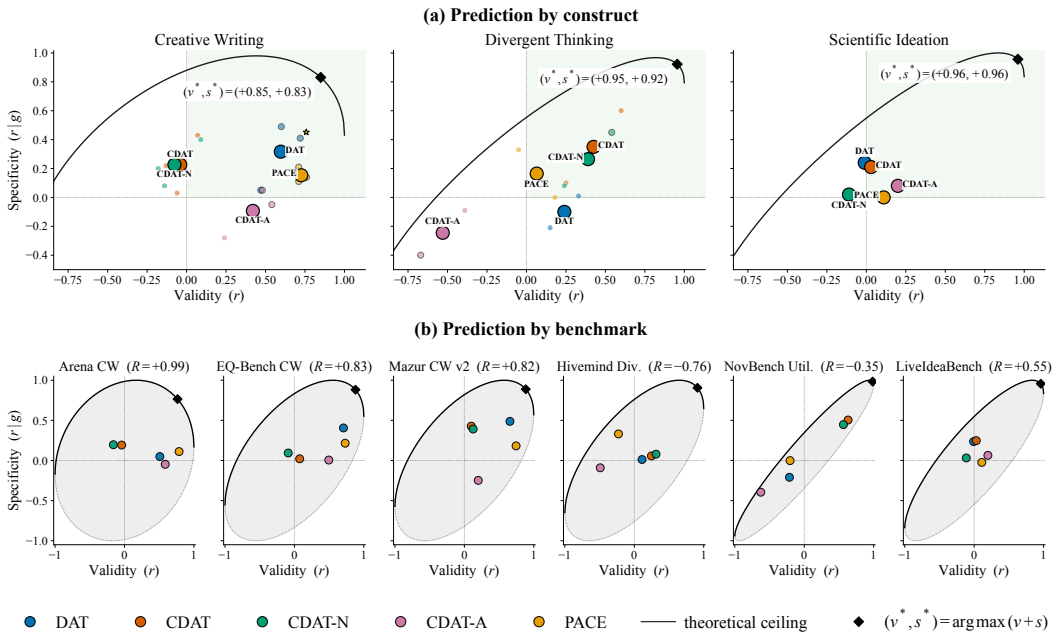

Figure 4: **Validity and specificity by construct and benchmark, with theoretical ceilings.** *(a) Prediction by construct:* Each small point is a (test, benchmark) cell from the Overall block of Table 1, while the large black-outlined circle is the construct-level average across all benchmarks. A gold ★ = significant on both axes. The black curve is the construct-level theoretical specificity ceiling obtained in Theorem 1. *(b) Prediction by benchmark:* per-benchmark specificity-ceiling lenses with the panel's R in the title. The grey region is the feasible $(v, r(X, Y \mid g))$ set.

Evaluating specificity reveals that PACE is mostly a capability proxy. A core motivation addressed in Section 2 is that tests should measure aspects of creative achievement independent of what general capabilities already predict. We measure this via the specificity metric defined in Section 4. On creative writing benchmarks, PACE obtains strong and statistically significant validity ($r \in [0.71, 0.76]$); however, after controlling for general capabilities, PACE collapses to non-significant specificity ($r|g \in [+0.11, +0.21]$). On creative writing, the chained-association testing methodology largely measures model qualities that capability already predicts.

The Divergent Association Task (DAT) is the best predictor of creative writing. DAT's validity is moderate across the three creative writing benchmarks (Arena CW $r = +0.47$, EQ-Bench CW $r = +0.72$, Mazur CW $r = +0.60$), and the EQ-Bench and Mazur cells retain meaningful specificity once capability is partialled out ($r|g = +0.41^{*}$ and $+0.49$). As mentioned, PACE has high raw validity averaged across the three creative writing benchmarks ($\approx +0.73$), but its specificity averages only $+0.15$. Lastly, CDAT's appropriateness facet (CDAT-A) has high validity ($r = +0.54^{***}, +0.48^{**}, +0.24$) but non-positive specificity throughout, predicting rankings only because it tracks capability.

The Conditional DAT is the best predictor of divergent thinking. The CDAT is the only test whose specificity is positive on both divergent thinking benchmarks (Hivemind $r|g = +0.10$; NoveltyBench Utility $r|g = +0.60$, $n = 8$). Furthermore, its appropriateness facet, CDAT-A, has both negative validity and specificity for divergent thinking (NoveltyBench $r = -0.67^{**}$, $r|g = -0.40$), consistent with appropriateness being a convergent-thinking measure that should anti-correlate with output diversity by construction.

None of the tests is a good predictor of scientific ideation. On LiveIdeaBench ($n = 17$), every test's raw validity is near zero ($r \in [-0.11, +0.20]$) and no specificity cell reaches $p < 0.05$. DAT and CDAT show modest positive specificity ($r|g \approx +0.24$ and $+0.21$), but neither is statistically significant. We discuss why semantic distance alone may be insufficient to predict scientific ideation ability in Section 6.

No single test predicts all constructs well. With DAT being the best predictor of creative writing, CDAT as the best predictor of divergent thinking, and none of the tests as good predictors of scientific

ideation, it is evident that each test is a local predictor of creative achievement tied to a specific construct rather than a general-purpose "creativity" measure.

How close are tests to the validity-specificity frontier? Empirically, we find that observed tests are well below the validity-specificity frontier for nearly every benchmark in Figure 4(b). This indicates significant headroom to design more effective creativity tests. The $(v^\star, s^\star)$ points marked in Figure 4 give the validity-specificity pair on the frontier that maximizes the sum $v + s$ for each construct or benchmark. They are the most that any test could jointly achieve on both axes given the benchmark's coupling to capability, and the gap between each observed test and its $(v^\star, s^\star)$ measures how much room is left for better tests of the same construct.

6 Discussion

Designing creativity tests that predict scientific ideation ability In Section 5, we found that existing semantic distance tests fail to reliably predict scientific ideation ability. Most surprisingly, even the CDAT, which measures both utility and novelty, fails to achieve significant validity and specificity on this construct. A key measure of utility for scientific ideation is whether ideas can be validated with experiments [44, 34], which the appropriateness dimension in CDAT may fail to adequately capture. Future evaluations of this ability may need to better reflect the cognitive primitives implicated in scientific ideation rather than rely solely on semantic distance measures, as recent studies have begun to address. [21] proposed algorithmic tasks that required creative, far-sighted leaps, mirroring cognitive mechanisms implicated in the scientific process [4, 16]. Later, [33] proposed a constrained graph pathfinding task, where constraints reflected common pitfalls observed in large-scale LLM-for-science studies [35, 34]. [41] also proposed a test-time creativity evaluation (CREATE) that measured associative creativity using real knowledge graphs, scoring an LLM by the diversity and validity of multi-hop associations it generated between concepts drawn from those graphs.

Ontology-based distances 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 [12]. Ontology-based distance measures may better capture hierarchical and logical relationships between concepts, and can be explored alongside 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.[8] Recent work has begun to explore both empirical evaluation of this ability [39, 28], as well as its conceptual and theoretical foundations [32], each of which should be incorporated in future large-scale benchmarking and test design studies.

Limitations of tests versus open-world evaluations Although creativity 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. [14] has recently called for *open-world evaluations*: "long-horizon, messy, real-world tasks assessed through small-sample qualitative analysis rather than benchmark-scale automation." Although in some respects, creativity can be reduced to cognitive primitives and concrete mechanisms [37], real-world creative processes are complex and typically modeled using distinct stages [42]. Accordingly, creative achievement may be best measured through a combination of minimal tests and open-ended, long-horizon assessments.

6.1 Limitations

Before concluding, several limitations of our work are worth discussing. 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.

[8] Einstein's relativity theory and Copernicus's heliocentric theory being two canonical examples [32].

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 [47].

7 Conclusion

In this work, we conducted the first systematic study to assess the effectiveness of semantic distance tests in predicting creative achievement across three constructs: creative writing, divergent thinking, and scientific ideation. Our findings provide clear practical takeaways and directions for future work, identifying tests which are best suited to predict each construct, and suggesting that novel tests are needed to reliably predict scientific ideation ability.

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, 2011. 3

[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, 9 2021. 1, 3, 6

[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 LLMs. Technical report, 2024. 1, 2

[4] Margaret A Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2004. 1, 3, 4, 9

[5] Honghua Chen and Nai Ding. Probing the Creativity of Large Language Models: Can models produce divergent semantic association? 10 2023. 1, 2

[6] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024. 5

[7] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 3

[8] David Cropley. Is artificial intelligence more creative than humans?: Chatgpt and the divergent association task. *Learning Letters*, 2:13–13, 2023. 1, 2

[9] 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. 3

[10] 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. 3

[11] 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, 2019. 1, 3, 4

[12] 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. 9

[13] 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. 5, 15, 16

[14] Sayash Kapoor, Peter Kirgis, Andrew Schwartz, Stephan Rabanser, J.J. Allaire, Rishi Bommasani, Magda Dubois, Gillian Hadfield, Andy 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. https://cruxevals.com/open-world-evaluations.pdf, 2026. 9

[15] 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. 3

[16] Arthur Koestler. *The Act of Creation*. Macmillan, 1964. 9

[17] 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. 1, 4

[18] Lech Mazur. LLM creative story-writing benchmark (v2). https://github.com/lechmazur/writing, 2025. 5, 15

[19] Sarnoff Mednick. The associative basis of the creative process. *Psychological Review*, 69(3):220, 1962. 1

[20] 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. 4

[21] Vaishnavh Nagarajan, Chen Henry Wu, Charles Ding, and Aditi Raghunathan. Roll the dice & look before you leap: Going beyond the creative limits of next-token prediction. arXiv:2504.15266, 2025. 9

[22] 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. 1, 4, 6, 14

[23] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023. 3

[24] Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. Naming unrelated words predicts creativity. 4 2021. 1, 3, 4, 6, 18

[25] 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. 3

[26] Samuel J Paech. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*, 2023. 5, 15

[27] 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. 1, 4

[28] 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: Pushing language models to think outside the box. In *Proc. The 63rd Annual Meeting of the Association for Computational Linguistics (ACL2025)*, 2025. 9

[29] 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 10859–10872, Suzhou, China, November 2025. Association for Computational Linguistics. 1, 3, 4, 6, 14

[30] 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*, 2026. 5, 15

[31] 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, 2024. 3

[32] Samuel Schapiro, Jonah Black, and Lav R Varshney. Transformational Creativity in Science: A Graphical Theory. *arXiv preprint arXiv:2504.18687*, 2025. 9

[33] Samuel Schapiro, Sumuk Shashidhar, Alexi Gladstone, Jonah Black, Royce Moon, Dilek Hakkani-Tur, and Lav R. Varshney. Combinatorial Creativity: A New Frontier in Generalization Abilities. 11 2025. 9

[34] Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas. 6 2025. 9

[35] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. 9 2024. 9

[36] Dean Keith Simonton. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. Cambridge University Press, 2004. 4

[37] 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. 1, 9

[38] 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. 1, 3

[39] 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. 9

[40] L. R. Varshney. Mathematical limit theorems for computational creativity. *IBM Journal of Research and Development*, 63(1), 1 2019. 4

[41] Manya Wadhwa, Tiasa Singha Roy, Harvey Lederman, Junyi Jessy Li, and Greg Durrett. Create: Testing llms for associative creativity. *arXiv preprint arXiv:2603.09970*, 2026. 9

[42] Graham Wallas. *The art of thought*. Number 24. Harcourt, Brace, 1926. 9

[43] 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. 1, 2

[44] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific Inspiration Machines Optimized for Novelty. 6 2024. 9

[45] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024. 6, 15, 16

[46] David Wechsler. The measurement of adult intelligence. *The Journal of Nervous and Mental Disease*, 91(4):548, 1940. 3

[47] 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

[48] 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. 5, 15

A Derivation of the Validity-Specificity Frontier

The theoretical ceilings plotted in Figure 4 (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

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

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

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

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 ($R = 0.98$ on $g = (\text{Arena Overall}, \text{MMLU-Pro})$) caps specificity at ≈ 0.20 for a perfect-validity test, while NoveltyBench Utility ($R \approx -0.33$) admits a much wider ceiling (Figure 4, bottom row).

B Per-Model Test Scores

In Table 2, 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 [29] across their 30-model set. Gated CDAT novelty across our models spans ∼62–75, within the ∼55–75 band shown in Figure 3 of [22]. Our evaluation extends both sets to a larger number of models.

Table 2: Mean ± SEM of DAT, CDAT, CDAT-N, CDAT-A, and PACE 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
OpenAI					
`gpt-3-5-turbo`	78.24±0.78	73.07±0.54	72.66±0.39	132.92±0.63	0.715±0.004
`gpt-4-1`	86.29±0.27	69.23±0.69	70.10±0.42	140.86±0.55	0.744±0.002
`gpt-4-1-mini`	81.60±0.49	66.88±0.88	67.75±0.58	143.78±0.66	0.730±0.003
`gpt-4-1-nano`	81.93±1.33	72.54±1.01	71.28±0.72	137.84±0.94	0.707±0.004
`gpt-4-turbo`	84.84±1.54	66.49±0.90	66.92±0.57	144.02±0.67	0.732±0.003
`gpt-4o`	82.94±1.15	65.14±1.02	66.26±0.69	144.54±0.78	0.729±0.002
`gpt-4o-mini`	78.70±1.34	70.53±0.76	71.52±0.46	138.41±0.69	0.707±0.003
`gpt-5`	89.33±0.21	69.85±0.85	69.77±0.52	141.96±0.62	0.747±0.002
`gpt-5-4`	91.72±0.19	68.28±0.88	68.63±0.49	143.70±0.54	0.727±0.003

(continued on next page)

(continued from previous page)

Model	DAT	CDAT	CDAT-N	CDAT-A	PACE
`gpt-5-4-mini`	84.06±0.32	65.21±0.83	65.47±0.52	145.88±0.59	0.734±0.002
`gpt-5-4-nano`	83.77±0.29	63.20±1.11	63.15±0.58	147.61±0.64	0.678±0.004
`gpt-5-mini`	82.92±0.36	67.90±0.86	68.02±0.49	143.90±0.57	0.741±0.003
`gpt-5-nano`	80.66±0.32	62.39±1.04	63.11±0.59	147.95±0.63	0.712±0.003
`o3`	89.46±0.24	70.09±0.78	70.06±0.44	142.25±0.54	0.748±0.004
`o3-mini`	76.37±0.42	65.32±1.03	65.67±0.61	144.17±0.71	0.715±0.003
`o4-mini`	84.44±0.36	68.32±0.81	68.77±0.46	142.61±0.58	0.732±0.003
Anthropic					
`claude-3-5-haiku`	87.36±0.22	66.62±0.76	67.95±0.43	141.52±0.52	0.726±0.002
`claude-3-haiku`	78.87±0.23	62.85±0.94	62.53±0.58	146.22±0.65	0.697±0.005
`claude-haiku-4-5`	85.74±0.26	67.61±0.90	67.80±0.51	143.71±0.65	0.667±0.010
`claude-opus-4-5`	89.26±0.21	69.71±0.75	69.90±0.38	143.03±0.46	0.742±0.003
`claude-opus-4-6`	89.70±0.97	—	67.22±0.00	148.68±0.00	0.750±0.002
`claude-sonnet-4`	86.69±0.16	70.04±0.59	69.78±0.37	141.69±0.53	0.739±0.004
`claude-sonnet-4-5`	86.67±0.17	66.93±0.84	66.43±0.48	146.01±0.56	0.756±0.002
`claude-sonnet-4-6`	88.97±0.21	—	—	—	0.755±0.002
Google					
`gemini-2-0-flash-001`	82.32±0.41	70.72±0.68	70.32±0.41	139.44±0.55	0.730±0.003
`gemini-2-5-flash`	76.68±0.54	68.54±0.85	68.54±0.49	143.39±0.58	0.742±0.002
`gemini-2-5-pro`	89.69±0.25	71.18±0.59	71.13±0.34	139.02±0.49	0.761±0.002
`gemma-2-27b-it`	81.87±0.44	70.46±2.10	70.92±0.61	138.29±0.85	0.694±0.008
`gemma-2-9b-it`	77.89±1.52	74.09±0.62	72.77±0.51	133.71±0.73	0.728±0.003
`gemma-3-27b-it`	86.49±0.25	71.75±0.59	72.05±0.33	137.58±0.51	0.728±0.005
Meta					
`llama-3-1-70b-instruct`	84.78±2.07	71.10±1.06	68.19±0.80	141.56±0.98	0.713±0.004
`llama-3-1-8b-instruct`	79.88±3.14	—	72.99±0.78	134.84±1.09	0.701±0.006
`llama-3-2-1b-instruct`	81.20±0.25	52.46±5.27	58.92±1.33	146.46±1.67	0.586±0.017
`llama-3-2-3b-instruct`	84.11±0.14	64.42±2.47	68.31±0.50	139.37±0.67	0.711±0.005
`llama-3-3-70b-instruct`	82.47±2.04	68.36±1.71	69.81±0.63	139.24±0.81	0.718±0.004
`llama-4-maverick`	85.28±0.26	67.34±0.63	67.45±0.44	141.98±0.56	0.707±0.005
`llama-4-scout`	84.48±1.05	66.90±0.84	67.39±0.50	141.42±0.61	0.696±0.005
Mistral					
`mistral-7b-instruct-v0-1`	81.20±0.01	69.13±0.96	69.16±0.54	135.43±0.56	0.613±0.011
`mistral-large-2407`	88.15±0.24	70.71±0.56	70.36±0.34	138.80±0.55	0.737±0.002
`mistral-large-2411`	81.91±0.40	64.87±0.82	64.54±0.49	146.01±0.54	0.722±0.003
`mistral-nemo`	78.17±3.01	—	71.09±0.95	137.00±1.28	0.709±0.005
`mistral-small-24b-instruct-2501`	82.39±2.09	—	73.56±0.76	132.24±0.94	0.717±0.003
Qwen					
`qwen-2-5-72b-instruct`	72.28±2.64	68.89±2.00	68.98±0.82	138.68±0.92	0.703±0.004
`qwen3-14b`	81.36±1.76	—	73.89±1.20	131.84±1.70	0.606±0.017
`qwen3-235b-a22b`	84.95±0.45	68.65±0.91	68.63±0.62	142.18±0.78	0.725±0.003
`qwen3-32b`	85.18±1.57	—	74.54±1.34	135.62±1.72	0.658±0.019
`qwen3-8b`	83.31±0.35	67.53±0.80	68.53±0.45	141.87±0.60	0.694±0.004
`qwq-32b`	82.63±0.50	—	—	—	—
DeepSeek					
`deepseek-chat`	81.12±0.40	68.33±0.75	67.50±0.53	143.11±0.71	0.729±0.003
`deepseek-chat-v3-0324`	80.14±1.81	68.36±0.90	68.00±0.55	143.13±0.68	0.730±0.004
`deepseek-r1`	83.80±0.39	68.75±0.72	69.28±0.45	141.38±0.65	0.720±0.003
Cohere					
`command-a`	82.28±0.33	62.99±1.12	63.22±0.64	147.80±0.66	0.714±0.003
`command-r-plus-08-2024`	87.69±0.45	69.61±0.83	69.44±0.49	138.70±0.64	0.722±0.003
NVIDIA					
`llama-3-1-nemotron-70b-instruct`	84.38±2.71	70.27±3.86	69.62±0.76	140.32±0.93	0.638±0.013
Microsoft					
`phi-4`	74.29±3.24	—	72.71±0.64	134.34±1.03	0.680±0.006

C Per-Model Benchmark Scores

Table 3 reports each model's score on every external benchmark used in this study. The general capability proxies are Arena Overall (Chatbot Arena Elo) and MMLU-Pro accuracy [45], the creative writing benchmarks are Arena CW (Elo), EQ-Bench CW v3 (Elo) [26], and Mazur CW [18], the divergent thinking (output-diversity) benchmarks are Hivemind diversity (1− intra-model cosine similarity) [13] and NoveltyBench Utility (NovB.; cumulative-utility score) [48], and the scientific ideation benchmark is the LiveIdeaBench [30] *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 [45]. 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. [13] 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.[9], where intra-sim is given by:

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

Table 3: **Per-model benchmark scores.** Columns: Arena Overall (Elo), MMLU-Pro (accuracy), Arena CW (Elo), EQ-Bench CW v3 (Elo), Mazur CW, Hivemind diversity, NoveltyBench Utility, LiveIdeaBench (5-dim Average). "—" indicates the corresponding benchmark does not score that model.

Model	Arena Ovr	MMLU-Pro	Arena CW	EQ-B. CW	Mazur	Hive.	NovB.	LiveIdea
OpenAI								
`gpt-3-5-turbo`	1223	—	1187	519	—	—	—	—
`gpt-4-1`	1413	0.82	1402	1419	—	—	—	—
`gpt-4-1-mini`	1382	—	1349	1231	—	—	—	—
`gpt-4-1-nano`	1321	—	1306	1034	—	—	—	—
`gpt-4-turbo`	1323	0.64	1322	—	—	0.14	—	6.56
`gpt-4o`	1443	0.75	1423	1484	8.18	0.12	3.27	6.69
`gpt-4o-mini`	1317	0.63	1295	950	6.72	0.11	3.11	6.25
`gpt-5`	1433	0.87	1376	1301	8.60	—	—	—
`gpt-5-4`	1466	0.88	1429	2019	—	—	—	—
`gpt-5-4-mini`	1459	—	1417	1726	—	—	—	—
`gpt-5-4-nano`	1402	—	1336	—	—	—	—	—
`gpt-5-mini`	1389	—	1326	1301	8.31	—	—	—
`gpt-5-nano`	1337	—	1250	866	—	—	—	—
`o3`	1431	0.85	1384	—	8.39	—	—	—
`o3-mini`	1347	0.79	1301	—	6.15	0.17	—	6.51
`o4-mini`	1390	—	1338	—	7.50	—	—	—
Anthropic								
`claude-3-5-haiku`	1323	0.62	1303	1241	7.35	0.11	2.50	6.37
`claude-3-5-sonnet`	—	0.78	—	—	—	—	2.36	6.92
`claude-3-haiku`	1260	0.42	1214	848	—	0.13	—	—
`claude-3-opus`	—	0.68	—	—	—	—	2.67	6.36
`claude-haiku-4-5`	1408	—	1384	—	—	—	—	—
`claude-opus-4-5`	1468	0.87	1462	1769	—	—	—	—
`claude-opus-4-6`	1496	0.89	1467	1965	—	—	—	—
`claude-sonnet-4`	1389	0.84	1384	1516	8.09	—	—	—
`claude-sonnet-4-5`	1451	0.87	1450	1777	—	—	—	—
`claude-sonnet-4-6`	1462	0.87	1444	1991	—	—	—	—
Google								
`gemini-1-5-pro`	—	0.70	—	—	—	—	2.73	6.85
`gemini-2-0-flash-001`	1360	0.78	1346	1252	7.15	0.15	3.17	7.07
`gemini-2-0-flash-lite-001`	1353	0.72	1345	—	—	—	3.20	6.60
`gemini-2-0-pro`	—	0.79	—	—	—	—	2.64	7.03

(continued on next page)

[9] 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	Arena Ovr	MMLU-Pro	Arena CW	EQ-B. CW	Mazur	Hive.	NovB.	LiveIdea
`gemini-2-5-flash`	1411	—	1399	1255	7.65	—	—	—
`gemini-2-5-pro`	1448	0.86	1448	1415	8.38	—	—	—
`gemma-2-27b-it`	1288	0.57	1291	—	—	0.16	3.77	6.65
`gemma-2-2b-it`	—	0.16	—	—	—	—	4.63	—
`gemma-2-9b-it`	1265	0.52	1257	920	—	0.17	3.93	—
`gemma-3-27b-it`	1365	0.68	1348	1256	7.99	—	—	—
Meta								
`llama-3-1-405b-instruct`	—	0.73	—	—	—	—	3.39	6.62
`llama-3-1-70b-instruct`	1293	0.63	1257	851	—	0.18	—	6.62
`llama-3-1-8b-instruct`	1211	0.44	1178	840	—	0.19	3.76	—
`llama-3-2-1b-instruct`	—	0.12	—	—	—	—	2.81	—
`llama-3-2-3b-instruct`	1166	0.22	1144	—	—	0.20	3.24	—
`llama-3-3-70b-instruct`	1318	0.66	1286	—	—	0.13	2.87	6.06
`llama-4-maverick`	1327	0.81	1307	944	6.20	—	—	—
`llama-4-scout`	1322	0.74	1290	899	—	—	—	—
Mistral								
`mistral-7b-instruct-v0-1`	1148	0.26	1104	—	—	0.19	—	—
`mistral-large-2407`	1313	0.66	1287	1400	6.90	—	—	—
`mistral-large-2411`	1305	0.68	1276	1082	6.90	0.14	—	6.79
`mistral-nemo`	1108	0.45	1090	966	—	0.20	—	—
`mistral-small-24b-instruct-2501`	1273	0.66	1227	—	—	0.18	—	6.41
Qwen								
`qwen-2-5-72b-instruct`	1302	0.72	1254	—	—	0.14	—	6.62
`qwen3-14b`	—	—	—	—	—	0.13	—	—
`qwen3-235b-a22b`	1374	0.83	1324	1379	8.30	—	—	—
`qwen3-32b`	1347	—	1306	—	—	0.15	—	—
`qwen3-8b`	—	—	—	—	—	0.12	—	—
`qwq-32b`	1336	0.69	1296	1262	8.02	—	—	7.06
DeepSeek								
`deepseek-chat`	1358	0.76	1349	—	—	0.14	—	6.72
`deepseek-chat-v3-0324`	1395	0.81	1390	1473	7.70	0.14	—	—
`deepseek-r1`	1398	0.84	1374	1500	8.30	—	—	7.18
Cohere								
`command-a`	1353	—	1337	1184	—	—	—	—
`command-r-08-2024`	1249	—	1209	—	—	—	2.98	—
`command-r-plus-08-2024`	1276	—	1263	—	—	0.23	3.08	—
`command-r7b-12-2024`	—	—	—	—	—	—	3.35	—
NVIDIA								
`llama-3-1-nemotron-70b-instruct`	1298	0.63	1277	—	—	—	—	—
Microsoft								
`phi-4`	1255	0.70	1210	—	6.26	0.15	—	6.72

D Greedy Algorithm for the DAT

The greedy algorithm that can trivially "solve" the DAT is given in Algorithm 1. Starting 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), the algorithm 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 5 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 5. As shown, the algorithm trivially exceeds the distribution of LLM and human scores.

E Prompts

Below, we report the exact prompts used to administer the DAT, CDAT, and PACE. 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.

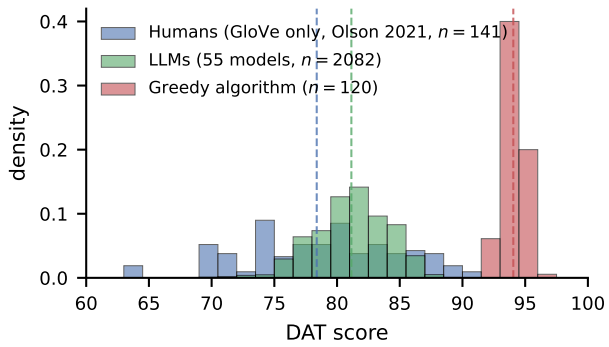

Figure 5: **A simple algorithm outperforms humans and LLMs on the DAT.** Three distributions are reported in this plot: the DAT score distributions for humans ([24], Study 1A, $n = 141$, GloVe-scored; mean $= 78.4$, std. dev. $= 6.4$), the distribution over our 54-model LLM pool ($n = 2078$ trials at $T = 1.0$; mean $= 83.75$, std. dev. $= 5.13$), and the results of our greedy algorithm over GloVe noun embeddings ($n = 120$; mean $= 94.1$, std. dev. $= 0.9$). Humans are on the GloVe scale only because [24] only reports final scores—not raw word lists. The dashed lines mark group means.

Algorithm 1 Greedy maximization algorithm for the DAT

Require: vocabulary V, embedding E, number of words n (= 10 for the DAT)
Ensure: word list $W = (w_1, \ldots, w_n)$

$w_1 \sim \mathrm{Uniform}(V)$ // random first word
$W \leftarrow [w_1]$
for $i = 2 \ldots n$ **do**
  $w_i \leftarrow \arg\min_{v \in V \setminus W} \frac{1}{|W|} \sum_{u \in W} \cos\big(E(v),\, E(u)\big)$
  $W \leftarrow W \cup \{w_i\}$
end for
return W

E.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"]
```

E.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"]
```

E.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": ""}]}
```

E.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": ""}, ... ]}
```