Measuring Constructive Creativity in AI-Augmented Work: A Scale-Invariant Embedding Framework

Vishal N. Patel

Phronos
vishal.patel@phronos.org

Abstract

As AI-generated content proliferates in workplace settings, objective metrics that capture dimensions relevant to constructive creativity—work that balances novelty with usefulness—are needed to support metacognitive monitoring and quality evaluation. Existing creativity instruments either measure unconstrained divergence or assess accuracy against fixed item sets, limiting their applicability to real-world text of varying complexity. We present a framework for evaluating "constructive creativity" through three orthogonal, scale-invariant metrics derived from embedding geometry: *divergence* (semantic spread among target concepts), *alignment* (semantic fit between associations and targets), and *parsimony* (non-redundancy of associations). Across 3,600 random configurations with targets ranging from 1 to 50 words and associations ranging from 2 to 50 words, aggregate pairwise correlations remained below $|r| < 0.22$ under OpenAI ada-002 embeddings and $|r| < 0.23$ under GloVe embeddings. Regime analysis decomposing orthogonality by task scale (m, n) revealed that two of three metric pairs—divergence–parsimony and alignment–parsimony—maintained independence uniformly across all scales and both embedding models, while divergence–alignment exhibited independence at smaller task sizes. Benchmarking against 144 Remote Associates Test items provided partial validation: alignment distinguished correct solutions from chance (mean similarity = 0.356), though divergence did not predict item difficulty or solution time as hypothesized. LLM judges (GPT-4o-mini, Claude Haiku 4.5) discriminated correct solutions from foils with 84–88% accuracy, and model-solving difficulty correlated with human difficulty ($r = 0.559$). These metrics offer candidate dimensions for applying semantic associations to creative tasks, providing a foundation for future validation as cognitive feedback tools in AI-augmented workflows.

1 Introduction

AI adoption in knowledge work is associated with increases in productivity alongside increases in "workslop"—AI-generated content that adds volume without adding value (Niederhoffer, Robichaux, and Hancock 2026). AI-generated work is also characterized by increases in volume (Liang et al. 2024), placing greater demands on recipients' cognitive bandwidth to read and parse the output (Ohde, Rost, and Overgaard 2025). Research on AI integration in workplaces suggests that the most valued contributions balance creative problem-solving with practical applicability (Bankins et al. 2024; Hölzle, Rose, and Kaschub 2024), and studies of scientific impact have found that high-value work strikes a balance between novelty and conventionality, where novelty characterizes creativity and conventionality enables communication and sharing of ideas (Simonton 2012; Uzzi et al. 2013). As evaluation of creative contributions in the workplace remains highly subjective, objective and normative measures of "constructive creativity" hold promise for supporting the human creation and review of AI-augmented work products.

Distinguishing productive from unproductive AI-generated content requires active engagement of several cognitive faculties, including metacognition, executive functioning, critical thinking, and epistemic vigilance. In one study of university students, AI tool usage significantly and negatively predicted critical thinking ($\beta = -0.37$) while positively predicting epistemic laziness ($\beta = 0.34$) and metacognitive weakness ($\beta = 0.40$) (Yurt and Kuşci 2026). Emerging evidence like this suggests that AI use may attenuate the very cognitive capacities needed to evaluate AI outputs effectively.

Metacognition can be understood as a behavioral feedback loop triggered by environmental cues (Efklides 2008). Evidence from behavioral feedback loops using physiological biomarkers suggests that presenting users with metrics reflecting their internal states can stimulate behavior change (Ferguson et al. 2022), and strong evidence demonstrates increased adoption of target behaviors when those behaviors are associated with a biomarker, as with continuous glucose monitoring for lifestyle interventions (Richardson et al. 2024). We hypothesize that cognitive markers embedded within AI-augmented workflows may similarly prompt metacognitive monitoring.

To develop such cognitive markers, we sought to extend the empirical literature on creativity measurement to accommodate varied and user-generated inputs. Existing creativity instruments either measure unconstrained divergence (Divergent Association Task; Olson et al. 2021) or assess accuracy of inference among a fixed item set (Remote Associates Test; for review, see Wu et al. 2020). Neither framework scales to real-world text communications, which vary in the

number of concepts addressed and the number of associations proposed. Olson et al.'s use of embeddings for calculating semantic distance—and evidence of its correlation with creativity—provides a basis for developing an extensible framework that can accommodate m targets and n associations. Such a framework would allow assessing the constructive creativity of communications ranging from prompts to documents using metrics with established psychometric relationships to human creativity. The present paper develops and validates the metric formulations at the level of individual word associations; extension to document-level text, which requires methods for extracting targets and associations from natural language, is reserved for future work.

Our objective was to develop a scale-invariant set of orthogonal measures that capture dimensions hypothesized to be relevant to constructive creativity. We make three contributions. First, we present methods for calculating scale-invariant divergence, alignment, and parsimony for semantic tasks with arbitrary m targets and n associations, and we demonstrate orthogonality of said measures. Second, we present validation results using Claude Haiku 4.5 and GPT-4o-mini as independent judges of constructive creativity. Third, a working prototype instrument is available and collecting human responses at https://instruments.phronos.org/ins-001. Although the present validation operates at the level of individual word associations, the framework's intended application extends to structured semantic tasks that arise naturally in knowledge work.

2 Related Work

Semantic Association Tests and Computational Creativity Measurement

Semantic association tasks have emerged as a foundational approach to measuring creative potential, with Olson et al. (2021) demonstrating that semantic distance between unrelated word pairs correlates with established creativity measures. Beaty and Johnson (2021) formalized this approach through SemDis, an open platform for computing semantic distance using word embeddings, enabling scalable assessment of divergent thinking outputs. Recent advances in large language models have substantially improved automated creativity scoring; Organisciak et al. (2023) demonstrated that fine-tuned LLMs achieve correlations of $r = .81$ with human originality ratings, far exceeding the $r = .12$–$.26$ range of purely semantic distance approaches. Kern, Wu, and Chao (2023) extended this work by developing unsupervised GPT-4-based methods capable of assessing not only novelty but also feasibility and value—addressing longstanding critiques that computational methods neglect the multi-dimensional nature of creativity. Patterson et al. (2023) further demonstrated the cross-linguistic validity of semantic distance approaches, establishing their applicability beyond English-language assessments.

The Novelty-Usefulness Balance in Creativity Theory

The standard definition of creativity requires that an idea or product be both novel and useful (Runco and Jaeger 2012), a bipartite criterion with intellectual origins predating contemporary creativity research by decades. Diedrich et al. (2015) provided empirical evidence that novelty and usefulness represent separable cognitive dimensions, with distinct neural substrates and differential sensitivity to task instructions. This theoretical framing aligns with evidence from scientific impact research; Uzzi et al. (2013) found that high-impact papers exhibit "atypical combinations"—novel pairings of concepts embedded within otherwise conventional structures—suggesting that optimal creativity balances divergence with communicative accessibility. Kharkhurin (2014) proposed extending the dual criterion to a four-criterion model incorporating aesthetics and authenticity, arguing that workplace and artistic creativity involve evaluative dimensions beyond mere utility. Corazza (2016) further complicated the standard definition by proposing a dynamic view in which novelty and usefulness emerge and are recognized over time rather than existing as fixed attributes at the moment of creation.

Cognitive Offloading, Metacognition, and Real-Time Feedback

Cognitive offloading—the use of external tools to reduce internal cognitive demands—has been extensively studied in the context of memory and calculation (Risko and Gilbert 2016), with emerging evidence that AI assistance may extend offloading to higher-order cognitive functions including critical evaluation and metacognitive monitoring. Tankelevitch et al. (2024) conceptualized the "metacognitive demands" of generative AI, identifying that effective use requires users to monitor their own knowledge states, evaluate AI outputs against internal standards, and strategically decide when to delegate versus engage directly. Structured metacognitive interventions—including planning prompts, reflection scaffolds, and self-evaluation checkpoints—have demonstrated efficacy in promoting strategy discovery and reducing overreliance on automated outputs (Becker and Lieder 2021; Järvelä and Hadwin 2024).

LLM-Based Evaluation

The LLM-as-judge paradigm emerged as a scalable alternative to human evaluation following Zheng et al.'s (2023) demonstration that strong models like GPT-4 achieve over 80% agreement with human preference judgments on open-ended generation tasks. Recent surveys (Gu et al. 2024; Li et al. 2024) have systematized the field, identifying core challenges around reliability, cross-domain generalization, and the need for standardized evaluation protocols. Chen et al. (2024) revealed that LLM judges exhibit systematic biases—including position bias, authority bias, and misinformation oversight—that partially overlap with, yet remain distinct from, human evaluator biases.

3 Methods

Based on the novelty-usefulness balance and value of aesthetics in creativity theory, we operationalize *constructive creativity* through three orthogonal measures derived from the semantic geometry of association tasks: divergence,

alignment, and parsimony. Each metric leverages dense embedding representations to quantify distinct aspects of creative association. For a task with m target words and n associations, these metrics remain scale-invariant, enabling comparison across tasks of varying complexity.

Embedding Representation

All words are represented as dense vectors obtained from a text embedding model. Let $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m \in \mathbb{R}^d$ denote the embeddings of the m target words, and let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^d$ denote the embeddings of the n association words. Semantic similarity between any two embeddings is computed as cosine similarity:

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \tag{1}$$

Divergence

Divergence measures the semantic spread among a set of words—how "remote" they are from one another in embedding space. Higher divergence indicates words that span disparate semantic regions, requiring associations that bridge distant concepts.

For x word embeddings, divergence is defined as the mean pairwise cosine distance:

$$D = \frac{2}{x(x-1)} \sum_{i<j} (1 - \text{sim}(\mathbf{w}_i, \mathbf{w}_j)) \tag{2}$$

where $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_x$ are the embeddings of the x words in the set (targets or associations, as applicable). When $x < 2$, divergence is undefined and returns zero by convention. Divergence ranges from 0 (all associations are identical) to approximately 2 (all associations are maximally dissimilar, i.e., cosine similarity of -1).

Alignment

Alignment measures how well the associations connect to all targets—the semantic fit of a proposed solution to the task's constraints. We implement two formulations depending on scale requirements.

Simple Alignment (for fixed m and $n = 1$) When a single association connects multiple targets, alignment is the mean cosine similarity between the association and each target:

$$A_{\text{simple}} = \frac{1}{m} \sum_{i=1}^{m} \text{sim}(\mathbf{a}, \mathbf{t}_i) \tag{3}$$

Scale-Invariant Alignment (for arbitrary m and n) For tasks with multiple associations, we compute optimal one-to-one assignment using the Hungarian algorithm (linear sum assignment). This ensures that alignment reflects the best matching rather than arbitrary pairings.

Let $\mathbf{S} \in \mathbb{R}^{m \times n}$ be the similarity matrix where $S_{ij} = \text{sim}(\mathbf{t}_i, \mathbf{a}_j)$. The bipartite fit is:

$$\text{fit}(\mathbf{T}, \mathbf{A}) = \frac{1}{\min(m, n)} \sum_{(i,j) \in \mathcal{M}^*} S_{ij} \tag{4}$$

where $\mathcal{M}^*$ is the optimal matching that maximizes total similarity.

Finally, we compare the true fit against foil targets. For k random foil sets drawn from a reference vocabulary V:

$$A_{\text{scaled}} = \frac{1}{k} \sum_{f=1}^{k} \mathbf{1}\left[\text{fit}(\mathbf{T}_f, \mathbf{A}) < \text{fit}(\mathbf{T}, \mathbf{A})\right] \tag{5}$$

This returns the proportion of foil targets for which the associations provide a worse fit than for the true targets. A value of 1.0 indicates the associations are uniquely suited to the given targets; 0.5 indicates chance-level performance.

In the present analyses, $k = 100$ foil sets were drawn from V, a reference vocabulary of 29,426 English words with pre-computed embeddings. Each foil set matched the dimensionality of the true target set (same m). We did not conduct systematic sensitivity analysis across values of k or alternative reference vocabularies; the dependence of scaled alignment scores on these parameters is a limitation of the current implementation. Future work should establish robustness across reference distributions.

Parsimony

Parsimony measures the efficiency of the association set—whether each association contributes uniquely to covering the targets or whether some associations are redundant. This metric is most meaningful when $n > 1$.

Define the fit function as the mean maximum similarity each target achieves with any association:

$$\text{fit}(\mathbf{A}, \mathbf{T}) = \frac{1}{m} \sum_{i=1}^{m} \max_j \text{sim}(\mathbf{t}_i, \mathbf{a}_j) \tag{6}$$

The marginal contribution of the j-th association is the decrease in fit when that association is removed:

$$\delta_j = \text{fit}(\mathbf{A}, \mathbf{T}) - \text{fit}(\mathbf{A}_{-j}, \mathbf{T}) \tag{7}$$

where $\mathbf{A}_{-j}$ denotes the association set with the j-th element removed.

Parsimony is then defined as the ratio of mean marginal contribution to maximum marginal contribution:

$$P = \frac{\bar{\delta}}{\max_j \delta_j} = \frac{\frac{1}{n}\sum_{j=1}^{n} \delta_j}{\max_j \delta_j} \tag{8}$$

When all associations contribute equally, $P = 1$. When a single association dominates and others are redundant, P approaches $1/n$. By convention, if $\max_j \delta_j \leq 0$ (no association contributes positively), parsimony returns 1.0. For the degenerate case of $n = 1$, parsimony is trivially 1.0.

Recovery

For comparison, we include an alternative measure of alignment motivated from information theory, which we call Recovery (Rec). For each word in the reference vocabulary, the maximum cosine similarity to any association is computed, the vocabulary is ranked by this score, and the algorithm returns the mean reciprocal rank (MRR) of the true targets in the resulting list. Formally:

$$\text{Rec-MRR} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\text{rank}_i + 1} \tag{9}$$

where rank_i is the position of target $\mathbf{t}_i$ when the vocabulary is sorted by $\max_j \text{sim}(\mathbf{v}, \mathbf{a}_j)$ in descending order.

Thus, Rec-MRR captures whether the associations make the targets identifiable—an information-retrieval property related to communicative signal strength rather than the bipartite fit that alignment measures.

Evaluations

For the orthogonality analysis, we computed divergence, alignment, and parsimony for 3,600 randomly generated samples spanning 42 (m, n) configurations, where $m \in \{1, 2, 3, 5, 10, 20, 50\}$ targets and $n \in \{2, 3, 5, 10, 20, 50\}$ associations. Each sample was constructed by randomly selecting m target words and n association words from a vocabulary of 29,426 English words with pre-computed embeddings. We used a random sampling strategy (rather than relevance-weighted or balanced sampling) specifically to avoid artificial coupling between metrics; relevance-weighted sampling would mechanically inflate alignment while constraining divergence, confounding the independence assessment. All metrics were computed under two embedding models: OpenAI text-embedding-ada-002 ($d = 1536$) and GloVe ($d = 300$), enabling assessment of whether orthogonality properties reflect the framework's design or are artifacts of a particular embedding space.

Divergence was computed over both the association set and the target set for each sample. Partial correlations controlling for target divergence were computed to isolate relationships among the response-side metrics (association divergence, alignment, and parsimony) from confounding by task structure.

Orthogonality was assessed at the aggregate level (across all 3,600 samples) via pairwise Pearson correlations with 95% bootstrap confidence intervals, using a threshold of $|r| < 0.30$ for independence. Second, a regime analysis decomposed orthogonality by (m, n) cell: for each of the 42 configurations, the pairwise correlation was computed within that cell's samples. A metric pair was classified as maintaining orthogonality in a given cell if $|r| < 0.30$. We report the percentage of cells passing this threshold for each pair and classify pairs as robustly orthogonal if $\geq 80\%$ of cells pass. The 80% threshold reflects a pragmatic criterion: it permits occasional boundary violations in extreme configurations while requiring that the predominant pattern across scales is independence. To test whether orthogonality varied systematically across scales, Cochran's Q test assessed heterogeneity across all cells, and a chi-squared test compared pass rates between small-scale ($m \leq 5$) and large-scale ($m > 5$) regimes.

In addition to the three primary metrics, we computed mean reciprocal rank (MRR) recovery as an alternative to alignment. Rec-MRR scores every word in the reference vocabulary by its maximum cosine similarity to any association, ranks the vocabulary by this score, and returns the mean reciprocal rank of the true targets in the resulting list. This metric captures whether the associations make the targets identifiable among the full vocabulary—a property related to but distinct from bipartite matching fit. We include Rec-MRR to assess whether alternative operationalizations of semantic fit achieve comparable or superior orthogonality with divergence and parsimony.

For the RAT benchmark, metric computation was simpler given the fixed structure: each item has exactly three cues ($m = 3$) and one solution ($n = 1$). We computed target divergence (mean pairwise cosine distance among the three cue embeddings) and alignment (mean, minimum, and maximum cosine similarity between each cue and the solution) for all 144 items. Parsimony was not computed for the RAT items because with $n = 1$ (a single association), the marginal contribution framework collapses—removing the only association leaves nothing, making the metric undefined. Four items lacked complete normative data (solution times), leaving 140 items for hypothesis testing.

For the LLM evaluations, three experiments were conducted.

Experiment 1 (Rating Task) The prompt provided three cue words and the correct solution, then asked for a 1–10 rating of how well the solution connects to all targets. The prompt was as follows:

> Target words: {cue1}, {cue2}, {cue3}
> Proposed connecting word: {solution}
> Rate how well the connecting word relates to ALL of the target words.
> 1 = No connection at all
> 5 = Weak or partial connection
> 10 = Strong, clear connection to all targets

GPT-4o-mini was called with temperature=0.0; Haiku used default settings. Responses were parsed by extracting digits and clamping to the 1–10 range. Each of the 50 items was evaluated once by each model.

Experiment 2 (Paired Comparison) The prompt presented two options—the correct RAT solution and a randomly selected word—and asked which better connects to all cues. Option order was randomized per trial to control for position bias. The prompt ended with "Respond with ONLY the letter A or B." Responses were parsed for the first occurrence of A or B. Accuracy was binary: did the model select the correct solution? Each of the 50 comparisons was evaluated once per model.

Experiment 3 (RAT Solving) The prompt presented the three cues and asked the model to produce the connecting word. Evaluation used two criteria: exact match (model output equals the canonical solution) and partial match (solution appears anywhere in the response). Each of the 50 items was evaluated once per model. Across all three experiments, no item was prompted multiple times to the same model—there are no repeated measures, confidence intervals from resampling, or temperature-varied runs. The correlations and accuracy figures reflect single-shot responses.

4 Results

Orthogonality

Aggregate orthogonality. Across 3,600 samples under OpenAI embeddings, pairwise correlations for the primary triple were: divergence–alignment ($r = -0.218$ [-0.249,

-0.187]), divergence–parsimony ($r = +0.070$ [$+0.038$, $+0.103$]), and alignment–parsimony ($r = +0.004$ [-0.029, $+0.036$]). Under GloVe embeddings, the corresponding values were: divergence–alignment ($r = -0.058$ [-0.090, -0.025]), divergence–parsimony ($r = +0.230$ [$+0.199$, $+0.261$]), and alignment–parsimony ($r = -0.009$ [-0.041, $+0.024$]). All aggregate pairwise correlations fell below $|r| < 0.30$ under both embedding models, satisfying the independence threshold at the aggregate level.

Partial correlations controlling for target divergence. Target divergence was uncorrelated with all three response-side metrics under both embedding models (all $|r| < 0.03$). Partial correlations controlling for target divergence produced negligible change in all pairwise relationships (maximum shift < 0.003), confirming that the orthogonality structure among divergence, alignment, and parsimony is not confounded by variation in task structure.

Regime analysis. Decomposing orthogonality by (m, n) cell revealed that aggregate correlations mask scale-dependent structure in one metric pair (see Figure 1, Table 1). Two of three pairs maintained robust orthogonality:

Alignment–parsimony passed $|r| < 0.30$ in 97.6% of cells under OpenAI and 100% of cells under GloVe. Cochran's Q was non-significant under both models (OpenAI: $Q = 5.00$, $p = 0.42$; GloVe: $Q = 0.00$, $p = 1.00$), confirming uniform independence across all task scales.

Divergence–parsimony passed in 95.2% of cells under OpenAI and 92.9% under GloVe. Cochran's Q was non-significant under both models (OpenAI: $Q = 10.00$, $p = 0.08$; GloVe: $Q = 3.00$, $p = 0.70$), and the chi-squared test comparing small and large regimes was likewise non-significant (OpenAI: $\chi^2 = 0.43$, $p = 0.51$; GloVe: $\chi^2 = 0.01$, $p = 0.93$). These two pairs represent the stable core of the framework's orthogonality structure.

Divergence–alignment exhibited scale-dependent anti-correlation. Under OpenAI embeddings, orthogonality held in 47.6% of cells, with correlations reaching $r = -0.60$ at $m = 50$, $n = 50$. Under GloVe, the same pair passed in 81.0% of cells, with a maximum of $r = -0.56$ at $m = 50$, $n = 20$. Cochran's Q was significant under both models (OpenAI: $Q = 28.67$, $p < 0.001$; GloVe: $Q = 18.12$, $p = 0.003$), and the chi-squared test confirmed that breakdown was concentrated in large-scale regimes (OpenAI: $\chi^2 = 9.48$, $p = 0.002$; GloVe: $\chi^2 = 6.78$, $p = 0.009$). The anti-correlation is geometrically predictable: as n grows, broadly scattered associations (high divergence) are less likely to achieve strong bipartite matching to specific targets (low alignment). This structural tradeoff between breadth and precision intensifies with task size.

Alternative metric: Rec-MRR. The Rec-MRR alternative to alignment achieved 100% orthogonality with parsimony across both embedding models (Cochran's Q non-significant in all cases). Rec-MRR–divergence passed in 81.0% (OpenAI) and 83.3% (GloVe) of cells—an improvement over alignment–divergence under OpenAI (47.6%) though comparable under GloVe (81.0%). The triple (divergence, Rec-MRR, parsimony) would eliminate the weakest pair in the current framework. Rec-MRR was not independent of alignment at scale, however (Rec-MRR–alignment: 21.4% / 19.0% pass rate), confirming that the two metrics capture overlapping variance in semantic fit.

Metric Pair	Pass Rate (%)		Cochran's Q	
	OpenAI	GloVe	OpenAI	GloVe
Align.–Pars.	97.6	100.0	5.00	0.00
Div.–Pars.	95.2	92.9	10.00	3.00
Div.–Align.	47.6	81.0	28.67*	18.12*
Div.–Rec-MRR	81.0	83.3	24.44*	14.26*
Pars.–Rec-MRR	100.0	100.0	0.00	0.00

Table 1: Regime analysis summary: percentage of (m, n) cells maintaining $|r| < 0.30$. Asterisks (*) indicate significant Cochran's Q ($p < 0.05$), reflecting heterogeneous orthogonality across task scales.

Comparison to Human Remote Associates Test Norms

We benchmarked the INS-001 metrics against 144 items from the Bowden and Jung-Beeman (2003) compound Remote Associates Test (RAT), which provides normative difficulty data including percent of participants solving at 2, 7, 15, and 30-second time limits, and mean solution times. This benchmark tests whether embedding-based metrics capture meaningful variance in human creativity performance on a psychometrically validated instrument. The RAT represents the canonical case of $m = 3$ targets (cue words) and $n = 1$ association (the solution word), providing a baseline for extending the framework to arbitrary m and n.

Three pre-registered hypotheses were evaluated:

H1: Target divergence predicts difficulty. Contrary to expectations, cue spread (mean pairwise cosine distance among the three cue words) showed no relationship with human performance ($r = 0.016$, $p = 0.85$ with percent solving at 15 seconds; Spearman $\rho = -0.039$, $p = 0.65$). The hypothesis that semantically distant cues would yield harder items was not supported.

H2: Target divergence predicts solution time. Similarly, divergence showed no association with mean solution time ($r = -0.071$, $p = 0.41$). Items with more scattered cues did not require longer search times, contrary to prediction.

H3: Alignment validates instrument construction. The alignment metric (mean cosine similarity between cue words and solution) demonstrated consistent validity across all RAT items: mean = 0.356, minimum = 0.246, maximum = 0.498. These values substantially exceed chance-level alignment, confirming that solutions are semantically connected to their cues at levels detectable by embedding models.

LLM Evaluation

Three experiments assessed whether large language models recognize the same semantic relationships captured by embedding-based alignment.

Experiment 1: LLM-as-Judge rating task. GPT-4o-mini and Claude Haiku 4.5 rated how well each solution connected to its cue words on a 1–10 scale. GPT-4o-mini

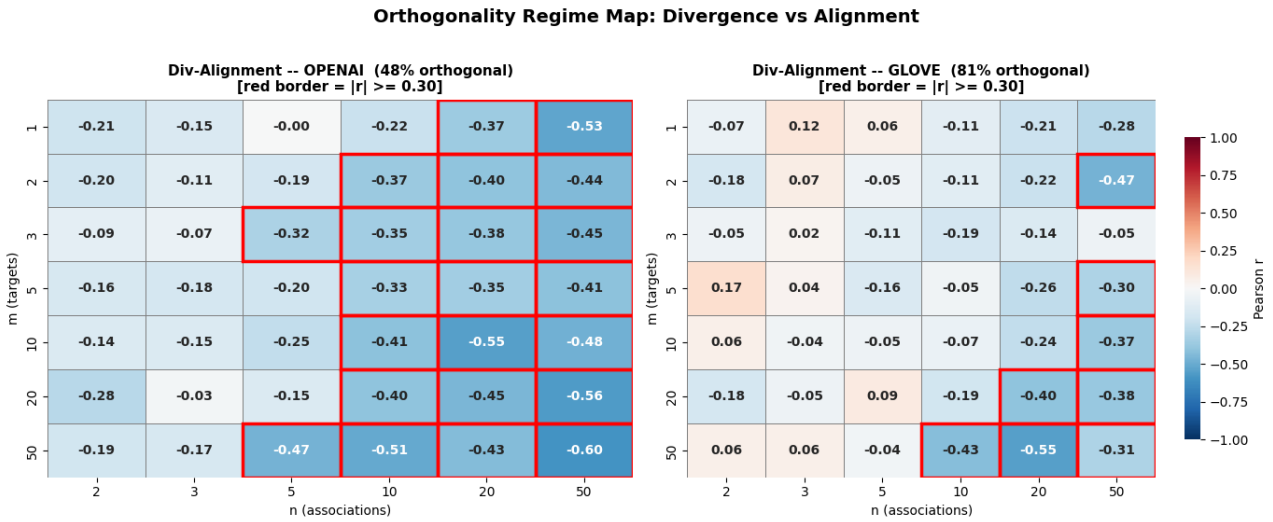

Figure 1: Orthogonality regime maps for divergence vs. alignment with OpenAI (left) and GloVe (right) embeddings. Each cell shows the Pearson correlation for a given (m, n) configuration. Red borders indicate cells where $|r| \geq 0.30$, violating the orthogonality threshold. The anti-correlation intensifies at larger task scales under both embedding models but is more severe under OpenAI.

Exp.	Model	Metric	Value	r(human)
Rating	GPT-4o-mini	r(rat., align)	0.495	—
Rating	Haiku	r(rat., align)	0.147	—
Paired	GPT-4o-mini	Accuracy	88.0%	—
Paired	Haiku	Accuracy	84.0%	—
Solving	GPT-4o-mini	Solve Rate	32.0%	0.559
Solving	Haiku	Solve Rate	64.0%	0.428

Table 2: LLM validation results. The r(human) column shows correlation between model success and human solve rates.

ratings correlated with embedding alignment ($r = 0.495$, $p < 0.001$); Haiku ratings showed a weaker but positive correlation ($r = 0.147$, $p = 0.31$). Cross-model agreement was moderate ($r = 0.443$), indicating that LLM judgments partially converge with embedding-based metrics.

Experiment 2: Paired comparison discrimination. When presented with the correct RAT solution alongside a random word and asked to identify which better connects all cues, GPT-4o-mini selected the high-alignment option in 88% of trials and Haiku in 84%. Both models reliably distinguished meaningful from arbitrary associations at levels substantially above chance.

Experiment 3: RAT solving performance. Haiku solved 64% of items exactly; GPT-4o-mini solved 32%. Critically, GPT-4o-mini success correlated with human difficulty ($r = 0.559$), indicating that items difficult for humans were also difficult for the model. Alignment modestly predicted LLM success ($r = 0.156$ for GPT-4o-mini, $r = 0.052$ for Haiku), suggesting that high-alignment items facilitate both human and machine retrieval.

5 Discussion

Our results confirmed that we were able to successfully develop a measure of topic relevance—which we call alignment—and a measure of succinctness—which we call parsimony—that are orthogonal to divergence. These three metrics provide complementary views of constructive creativity: divergence characterizes the breadth of associations (how far apart the concepts are), alignment characterizes the success of the solution (how well the associations connect the targets), and parsimony characterizes the efficiency of the solution (whether the associations are necessary and non-redundant). These characterizations describe geometric properties of embedding-space configurations.

While two of the three metric pairs—alignment–parsimony and divergence–parsimony—maintained independence uniformly across all 42 (m, n) configurations and both embedding models, the third pair, divergence–alignment, exhibits an anti-correlation that increases at larger task scales, reaching $r = -0.60$ under OpenAI embeddings. GloVe's lower dimensionality and different training objective produce a semantic space in which divergence and alignment are more readily separable, demonstrating that the divergence-alignment relationship is dependent on the embedding space.

The alternative metric, Rec-MRR, achieved stronger orthogonality with divergence than alignment did, passing the $|r| < 0.30$ threshold in 81.0% (OpenAI) and 83.3% (GloVe) of regime cells compared to alignment's 47.6% and 81.0%. The triple (divergence, Rec-MRR, parsimony) would eliminate the framework's weakest metric pair. The statistical advantage of Rec-MRR comes at the cost of theoretical specificity: alignment is motivated by creativity theory's usefulness dimension (the degree to which associations address their intended targets), while Rec-MRR captures an infor-

mation theoretic notion of signal recovery that lacks direct grounding in the creativity literature. We retain alignment as the primary metric on theoretical grounds while noting that Rec-MRR offers a viable alternative for applications where uniform orthogonality across scales is a priority.

The null results for H1 and H2 concern target divergence, which characterizes the semantic spread of the task's cue words. Target divergence showed no relationship with human difficulty or solution time, contradicting our expectation that semantically distant cues would yield harder items. One interpretation is that RAT difficulty arises primarily from associative strength and frequency effects, rather than from the geometry of the semantic space itself; the path from cue to solution may depend more on the availability of compound word associations in memory than on the cosine distance among the cues.

The LLM validation experiments provide converging evidence that alignment reflects judgments shared across both embedding-based and generative-model-based evaluations. GPT-4o-mini ratings correlated substantially with alignment ($r = 0.495$), and both models discriminated correct solutions from random foils with high accuracy (GPT-4o-mini: 88%; Haiku: 84%). The correlation between GPT-4o-mini solving success and human difficulty ($r = 0.559$) indicates that items challenging for humans are also challenging for LLMs, suggesting shared constraints on associative retrieval. Haiku ratings showed a weaker correlation with alignment ($r = 0.147$) while achieving higher solve rates (64% vs. 32%), suggesting differences in how these models represent compound-word associations or calibrate confidence ratings.

Benchmarking alignment against human RAT norms provides preliminary evidence that alignment captures a meaningful aspect of solution quality. Across all 144 items, alignment values substantially exceeded chance-level similarity (mean = 0.356), confirming that RAT solutions—which humans reliably identify as "correct"—share embedding structure that distinguishes them from arbitrary word associations.

6 Limitations and Future Work

Several limitations constrain the interpretation of these findings. The RAT benchmark, while psychometrically validated, represents a narrowly defined task space in 1 language. This structure may have limited the range of divergence values, contributing to the null findings for H1 and H2. The LLM evaluations relied on single-shot responses without repeated measures, temperature variation, or confidence intervals from resampling; the reported correlations and accuracy figures should therefore be treated as point estimates rather than stable parameters.

The results reveal that some orthogonality properties are robust to embedding choice while others are not. The embedding models used here inherit cultural, linguistic, and domain biases from their respective training corpora, and proprietary models like ada-002 face the additional concern of periodic deprecation, raising questions about longitudinal comparability. Future work should extend this comparison to additional architectures—including sentence-transformer models and multilingual embeddings—to further delineate framework properties from embedding artifacts.

The regime analysis used random sampling to populate each (m, n) cell. As noted in the Discussion, random associations tend toward low alignment and moderate-to-high divergence, which may amplify the divergence–alignment anti-correlation relative to what would be observed with structured or relevance-weighted inputs. Thus, the regime analysis characterizes worst-case orthogonality under maximally uninformative associations. We recognize that the present study cannot answer whether the divergence–alignment tradeoff persists when associations are generated by humans pursuing a goal. In addition, we admit that the 80% cell-pass threshold for classifying a pair as robustly orthogonal is a pragmatic rather than theoretical decision; alternative thresholds may reclassify the divergence–alignment pair but would not change the finding that its orthogonality is scale-dependent.

Future work should focus on human validation. A prototype is live at https://instruments.phronos.org/ins-001 and collecting responses; forthcoming analyses will establish whether human performance on these tasks correlates with established creativity measures and whether the metrics predict individual differences in creative ability. Longitudinal studies should test whether presenting users with real-time divergence, alignment, and parsimony feedback actually increases metacognitive engagement. Extending these metrics to naturalistic text, like prompts, emails, and documents, will shed light on the feasibility of real-world evaluations of constructive creativity. Finally, the framework's reliance on embedding geometry rather than modality-specific features suggests a path toward multimodal extension. Dense embedding representations exist for images (e.g., CLIP; Radford et al. 2021), audio, code, and other modalities, and cross-modal embedding spaces allow similarity computation between, for example, a textual description and an image.

7 Conclusion

Our work extends computational approaches to creativity measurement by proposing a scalable task structure with metrics motivated by definitions of creativity that incorporate both novelty and usefulness. The composite framework is motivated by the bipartite criterion of creativity theory while accounting for the inflation of text volume seen with AI-generated work. The regime analysis demonstrates that two of three metric pairs maintain robust independence across all task scales and embedding models, while the third pair (divergence–alignment) exhibits a scale-dependent tradeoff between breadth and precision that users of the framework should account for when interpreting results from tasks with large association sets. Our approach holds promise for implementing self-evaluative measures within semantic workflows. By providing objective, normative metrics that do not depend on subjective judgment, such tools may also help users of AI distinguish productive contributions from "workslop."

Ethical Statement

Participation in Phronos instruments requires explicit consent to our Terms of Service and Privacy Policy.

We acknowledge the risk that these instruments could systematically disadvantage certain populations—particularly those with limited English proficiency or educational access—and commit to addressing these equity concerns in future validation studies.

Acknowledgments

The author thanks the early participants who provided feedback on instrument design. The instruments are available at https://instruments.phronos.org.

References

Bankins, S.; Ocampo, A. C.; Marrone, M.; Restubog, S. L. D.; and Woo, S. E. 2024. A Multilevel Review of Artificial Intelligence in Organizations: Implications for Organizational Behavior Research and Practice. *Journal of Organizational Behavior*, 45(2): 159–182.

Beaty, R. E.; and Johnson, D. R. 2021. Automating Creativity Assessment with SemDis: An Open Platform for Computing Semantic Distance. *Behavior Research Methods*, 53(2): 757–780.

Becker, F.; and Lieder, F. 2021. Promoting Metacognitive Learning Through Systematic Reflection. In *NeurIPS Workshop on Metacognition in the Age of AI*.

Bowden, E. M.; and Jung-Beeman, M. 2003. Normative Data for 144 Compound Remote Associate Problems. *Behavior Research Methods, Instruments, & Computers*, 35: 634–639.

Chen, G. H.; et al. 2024. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Proceedings of EMNLP 2024*.

Corazza, G. E. 2016. Potential Originality and Effectiveness: The Dynamic Definition of Creativity. *Creativity Research Journal*, 28(3): 258–267.

Diedrich, J.; Benedek, M.; Jauk, E.; and Neubauer, A. C. 2015. Are Creative Ideas Novel and Useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9(1): 35–40.

Efklides, A. 2008. Metacognition: Defining Its Facets and Levels of Functioning in Relation to Self-Regulation and Co-Regulation. *European Psychologist*, 13(4): 277–287.

Ferguson, T.; Olds, T.; Curtis, R.; Blake, H.; Crozier, A. J.; Dankiw, K.; Dumuid, D.; Kasai, D.; O'Connor, E.; Virgara, R.; and Maher, C. 2022. Effectiveness of Wearable Activity Trackers to Increase Physical Activity and Improve Health: A Systematic Review of Systematic Reviews and Meta-Analyses. *The Lancet Digital Health*, 4(8): e615–e626.

Gu, J.; et al. 2024. A Survey on LLM-as-a-Judge. arXiv:2411.15594.

Hölzle, K.; Rose, R.; and Kaschub, V. L. 2024. The Interplay of Humans, Technology, and Organizations in Realizing AI's Productivity Promise. *Economists' Voice*, 21(2): 357.

Järvelä, S.; and Hadwin, A. F. 2024. Socially Shared Regulation of Learning. *Educational Psychologist*.

Kern, F. B.; Wu, C.-T.; and Chao, Z. C. 2023. Assessing Novelty, Feasibility and Value of Creative Ideas with an Unsupervised Approach Using GPT-4. *British Journal of Psychology*.

Kharkhurin, A. V. 2014. Creativity 4in1: Four-Criterion Construct of Creativity. *Creativity Research Journal*, 26(3): 338–352.

Li, H.; et al. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. arXiv:2412.05579.

Liang, W.; Zhang, Y.; Wu, Z.; Lepp, H.; Ji, W.; Zhao, X.; others; and Zou, J. Y. 2024. Mapping the Increasing Use of LLMs in Scientific Papers. arXiv preprint arXiv:2404.01268.

Niederhoffer, K.; Robichaux, A.; and Hancock, J. T. 2026. Why People Create AI "Workslop"—And How to Stop It. https://hbr.org/2026/01/why-people-create-ai-workslop-and-how-to-stop-it. Accessed: 2026-01-30.

Ohde, J. W.; Rost, L. M.; and Overgaard, J. D. 2025. The Burden of Reviewing LLM-Generated Content. *NEJM AI*.

Olson, J. A.; Nahas, J.; Chmoulevitch, D.; Cropper, S. J.; and Webb, M. E. 2021. Naming Unrelated Words Predicts Creativity. *Proceedings of the National Academy of Sciences*, 118(25): e2022340118.

Organisciak, P.; Acar, S.; Dumas, D.; and Berthiaume, K. 2023. Beyond Semantic Distance: Automated Scoring of Divergent Thinking Greatly Improves with Large Language Models. *Thinking Skills and Creativity*, 49: 101356.

Patterson, J. D.; et al. 2023. Multilingual Semantic Distance: Automatic Verbal Creativity Assessment in Many Languages. *Psychology of Aesthetics, Creativity, and the Arts*, 17(4): 495–507.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of ICML 2021*.

Richardson, K. M.; Jospe, M. R.; Bohlen, L. C.; Crawshaw, J.; Saleh, A. A.; and Schembre, S. M. 2024. The Efficacy of Using Continuous Glucose Monitoring as a Behaviour Change Tool in Populations with and Without Diabetes: A Systematic Review and Meta-Analysis of Randomised Controlled Trials. *International Journal of Behavioral Nutrition and Physical Activity*, 21(1): 145.

Risko, E. F.; and Gilbert, S. J. 2016. Cognitive Offloading. *Trends in Cognitive Sciences*, 20(9): 676–688.

Runco, M. A.; and Jaeger, G. J. 2012. The Standard Definition of Creativity. *Creativity Research Journal*, 24(2-3): 92–96.

Simonton, D. K. 2012. Taking the U.S. Patent Office Criteria Seriously: A Quantitative Three-Criterion Creativity Definition and Its Implications. *Creativity Research Journal*, 24(2-3): 97–106.

Tankelevitch, L.; et al. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of CHI 2024*.

Uzzi, B.; Mukherjee, S.; Stringer, M.; and Jones, B. 2013. Atypical Combinations and Scientific Impact. *Science*, 342: 468.

Wu, C.-L.; Huang, S.-Y.; Chen, P.-Z.; and Chen, H.-C. 2020. A Systematic Review of Creativity-Related Studies Applying the Remote Associates Test From 2000 to 2019. *Frontiers in Psychology*, 11: 573432.

Yurt, E.; and Kuşci, İ. 2026. Factors Influencing Critical Thinking During AI Use Among University Students: The Mediating Effects of Epistemic Laziness and Metacognitive Weakness. *Current Psychology*, 45: 67.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS 2023*.