Coherent Novelty: Extending Perplexity-Based Creativity Metrics with Coherence Operators

Joel Thorarinson* Allison Hensgen†

June 2026

Abstract

Perplexity — a language model's surprise at observed text — has been proposed as a proxy for novelty in scientific discovery and creative writing. High-perplexity text deviates from expected patterns, suggesting originality. However, perplexity alone conflates genuine creativity with incoherence: nonsense is maximally surprising but not creative. We propose *coherent novelty* (CN) as a composite metric that pairs semantic surprise with structural coherence. Formally, $\text{CN}(x) = S(x) \cdot \sigma(\Delta(x))$, where $S(x)$ is the semantic surprise (normalized perplexity) and $\Delta(x)$ is a coherence score computed from three embedding-space operators: pattern retention P (mean cosine similarity to k-nearest neighbors), alignment A (maximum similarity to domain anchor centroids), and generativity R (semantic diversity of the neighborhood). All operators are computable from a reference corpus and an embedding model; we provide explicit formulas and a worked example on scientific abstracts. The framework operationalizes the standard definition of creativity as novelty plus usefulness, maps directly onto Koestler's bisociation and Fauconnier–Turner conceptual blending, and predicts an optimal "creative temperature" at which LLM sampling maximizes CN. We propose three falsifiable experiments — on scientific abstracts (citation impact), creative writing (expert ratings), and neurodivergent text generation (ADHD vs. controls) — and discuss why CN adds explanatory power beyond perplexity combined with human ratings alone.

Keywords: creativity metrics; perplexity; coherence; novelty; semantic surprise; divergent thinking; language models; computational creativity; scientific discovery

1 Introduction: Perplexity Measures Surprise but Not Creativity

Creativity requires both *novelty* (the output is original, unexpected, surprising) and *usefulness* (the output is valuable, appropriate, meaningful) [Runco and Jaeger, 2012, Mumford, 2003]. This dual requirement is foundational: Amabile's consensual assessment technique operationalized creativity as expert-judged novelty plus appropriateness [Amabile, 1982]; Simonton's historiometric analysis of scientific breakthroughs showed that the most impactful discoveries combine chance variation with disciplinary coherence [Simonton, 2004]; Franceschelli and Musolesi identify novelty, value, and surprise as the three pillars of computational creativity evaluation [Franceschelli and Musolesi, 2024].

Language models provide a computational proxy for the novelty dimension: *perplexity*. When a model assigns low probability to a text passage, that passage is surprising relative to the model's training distribution. Recent work shows that perplexity correlates with scientific impact — papers that surprise language models tend to be more transformative [Zhang et al., 2025]. Perplexity also predicts divergent thinking capacity: higher perplexity in verbalized thought processes predicts greater performance on divergent thinking tasks [Yu et al., 2026].

*Coherence Research Group. ORCID: 0000-0002-0553-842X. joel.thorarinson@conformalmaps.com
†Coherence Research Group. ORCID: 0009-0008-7247-0307

But perplexity captures only half of creativity. Nonsense text is maximally surprising. Random word salad has high perplexity. Incoherent speech has high perplexity. None of these are creative. This failure mode is well-documented: Holtzman et al. showed that sampling strategies which increase surprise (high temperature, broad top-p) quickly produce incoherent output [Holtzman et al., 2020]. What distinguishes a creative insight from noise is not the degree of surprise but whether the surprising content *maintains structural coherence*.

Figure 1 captures this in two dimensions. The horizontal axis measures semantic surprise S; the vertical axis measures coherence Δ. Only the upper-right quadrant — high surprise with high coherence — represents genuine creativity. The lower-right quadrant (high surprise, low coherence) is noise. The upper-left (low surprise, high coherence) is competent but conventional. The lower-left is degraded on both dimensions.

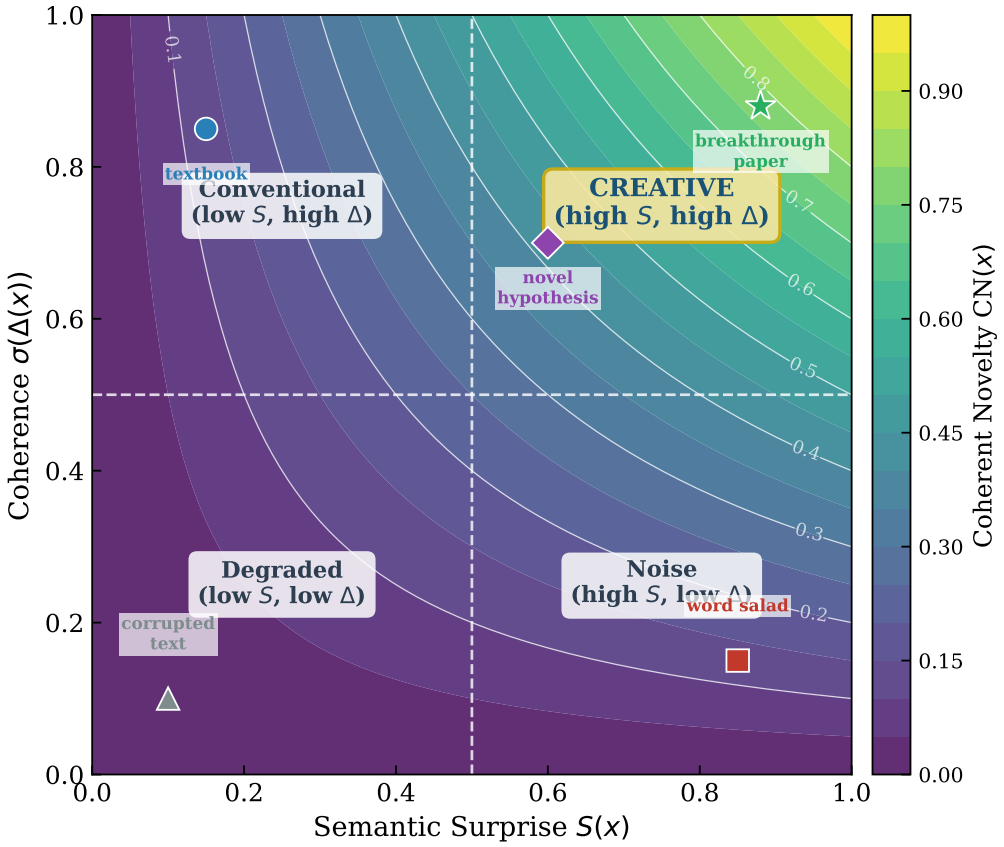

Figure 1: The coherent novelty space. Contour lines show constant CN values ($S \cdot \sigma(\Delta)$). The four quadrants correspond to the regimes in Table 1. Only the upper-right quadrant (high surprise, high coherence) represents genuine creativity. Koestler's bisociation [Koestler, 1964] places creative acts in this quadrant: the intersection of distant frames (high S) that each maintain internal structure (high Δ).

We propose *coherent novelty* (CN) as the product of surprise and coherence:

$$\mathrm{CN}(x) = S(x) \cdot \sigma\big(\Delta(x)\big) \tag{1}$$

where $S(x)$ is the semantic surprise (normalized perplexity), $\Delta(x)$ is a coherence ratio computed from embedding-space operators, and $\sigma(\cdot)$ is a squashing function that bounds the score in $[0, 1]$. Every component is computable from a reference corpus, an embedding model, and a language model. No human annotation is required at scoring time.

1.1 Contributions

1. We define **coherent novelty** (CN) as a composite metric with explicit, implementable formulas for every component (Section 2).
2. We show that CN **operationalizes** Koestler's bisociation and Fauconnier–Turner conceptual blending in terms of the coherence operators P, A, R (Section 3).
3. We derive the **temperature–CN tradeoff**, predicting an optimal creative temperature τ^* for LLM sampling (Section 4).
4. We provide a **worked example**: full CN computation on three scientific abstracts, demonstrating the metric's discriminative power (Section 5).
5. We design **three falsifiable experiments** with concrete evaluation protocols and address why CN adds explanatory power beyond perplexity combined with human ratings (Section 7).

2 The Coherent Novelty Framework

2.1 Semantic Surprise

For a text passage $x = (w_1, \ldots, w_n)$, we define semantic surprise as the normalized perplexity:

$$S(x) = \text{clip}\left(\frac{\text{PP}(x) - \text{PP}_{\text{base}}}{\text{PP}_{\text{ceil}} - \text{PP}_{\text{base}}},\ 0,\ 1\right) \tag{2}$$

where $\text{PP}(x) = \exp\big(-\frac{1}{n}\sum_{i=1}^{n} \log p(w_i \mid w_{<i})\big)$ is the perplexity under a reference language model, PP_{base} is the corpus mean perplexity (computed over a held-out sample of $\geq$ 1,000 documents from the target domain), and PP_{ceil} is the 99th percentile of the corpus perplexity distribution. Clipping to $[0, 1]$ ensures robustness to outliers.

The choice of reference model matters: a domain-specific model (e.g., SciBERT for scientific text) produces a more informative surprise signal than a general-purpose model, since surprise is relative to domain expectations [Itti and Baldi, 2009]. When comparing texts across domains, the normalization constants PP_{base} and PP_{ceil} must be recomputed for each domain.

2.2 Coherence Operators

The coherence score uses three operators defined over the embedding space induced by a sentence or document encoder (e.g., Sentence-BERT, nomic-embed-text). Let $\mathbf{e}(x) \in \mathbb{R}^d$ denote the ℓ_2-normalized embedding of text x, and let $\mathcal{C} = \{\mathbf{e}(c_i)\}_{i=1}^{|\mathcal{C}|}$ denote the embeddings of a reference corpus.

Definition 1 (Pattern Retention)**.** *The pattern retention of x with respect to $\mathcal{C}$ is the mean cosine similarity to its k nearest neighbors:*

$$P(x) = \frac{1}{k}\sum_{j=1}^{k} \cos\big(\mathbf{e}(x),\ \mathbf{e}(c_{\pi(j)})\big) \tag{3}$$

where $\pi(j)$ indexes the j-th nearest neighbor of $\mathbf{e}(x)$ in $\mathcal{C}$. We fix $k = 10$ throughout. High P means the text occupies a populated, semantically coherent region of embedding space. Low P means the text falls in an empty or incoherent region.

Definition 2 (Alignment)**.** *The alignment of x is the maximum cosine similarity to a set of M domain anchor embeddings $\{\mathbf{a}_m\}_{m=1}^{M}$:*

$$A(x) = \max_{m \in \{1,\ldots,M\}} \cos\big(\mathbf{e}(x),\, \mathbf{a}_m\big) \tag{4}$$

Anchors are computed as the centroids of topic clusters identified via k-means clustering on $\mathcal{C}$. The number of clusters M is set via the elbow method or silhouette score. High A means the text connects to at least one recognized area of the domain; low A means it has drifted away from all known topics.

Definition 3 (Generativity)**.** *The generativity of x is the semantic diversity of its k-nearest neighborhood:*

$$R(x) = 1 - \frac{1}{\binom{k}{2}} \sum_{i<j} \cos\big(\mathbf{e}(c_{\pi(i)}),\, \mathbf{e}(c_{\pi(j)})\big) \tag{5}$$

High R means the text's neighbors span multiple sub-topics, suggesting the text sits at an intersection that could generate follow-up questions, experiments, or creative extensions. Low R means the neighbors are semantically redundant — the text occupies a narrow, well-trodden area. For retrospective validation, R can be compared against downstream citation counts or follow-up publications.

2.3 Coherence Score

The coherence score combines the three operators into a single ratio:

$$\Delta(x) = \frac{P(x) \cdot A(x) \cdot R(x)}{(1 - A(x)) + (1 - P(x)) + \epsilon} \tag{6}$$

where $\epsilon = 10^{-6}$ prevents division by zero.

The denominator terms $(1 - A(x))$ and $(1 - P(x))$ are the complements of alignment and pattern retention, respectively. We denote these $D(x) = 1 - A(x)$ (drift from domain anchors) and $N(x) = 1 - P(x)$ (neighborhood sparsity, a proxy for noise). While D and N are algebraically redundant with A and P, their role in Δ is not: the ratio structure creates a nonlinear interaction that the numerator product alone does not capture. When either A or P is low, the corresponding denominator term grows, suppressing Δ faster than a simple product $P \cdot A \cdot R$ would. This asymmetry is deliberate: a text that drifts far from all anchors ($A \to 0$, so $D \to 1$) should receive a near-zero coherence score regardless of its pattern retention or generativity, because domain alignment is a prerequisite for meaningful coherence in context-dependent evaluation.

Since $P, A, R \in [0, 1]$ for ℓ_2-normalized embeddings, the numerator $P \cdot A \cdot R \in [0, 1]$ and the denominator $D + N + \epsilon \in [\epsilon, 2 + \epsilon]$. The ratio $\Delta \in [0, 1/\epsilon]$, but in practice Δ rarely exceeds 5 for well-calibrated embeddings.

2.4 Bounded CN via Squashing

The coherent novelty score applies a monotone squashing function to Δ before multiplication with S:

$$\mathrm{CN}(x) = S(x) \cdot \sigma\big(\Delta(x)\big), \quad \sigma(\Delta) = \frac{\Delta}{\Delta + 1} \tag{7}$$

The function $\sigma : [0, \infty) \to [0, 1)$ preserves ordering (higher coherence always yields higher $\sigma(\Delta)$) and ensures $\mathrm{CN}(x) \in [0, 1]$. The squashing has negligible effect in the typical regime ($\Delta \in [0.1, 3]$) but prevents outliers from dominating.

$S(x)$	$\sigma(\Delta(x))$	**Interpretation**
Low	High	Conventional. Expected content, well-structured. Textbook exposition, review articles.
High	Low	Noise. Surprising but incoherent — word salad, hallucination, off-topic tangent.
Low	Low	Degraded. Neither surprising nor coherent. Garbled boilerplate.
High	High	**Creative.** Surprising content that maintains structure, connects to the domain, and opens new directions.

Table 1: The four quadrants of the surprise–coherence space (cf. Figure 1). Only the high-S, high-$\sigma(\Delta)$ quadrant represents genuine creativity.

3 Relation to Creativity Frameworks

3.1 Bisociation and Conceptual Blending

Koestler's bisociation [Koestler, 1964] defines the creative act as the intersection of two previously unrelated "matrices of thought." This requires two conditions: (1) the association must be *surprising* — the two frames are not normally connected, so observing them together produces high semantic surprise S; (2) each frame must be *internally consistent* — each matrix of thought is self-coherent, so pattern retention P remains high and alignment A connects to at least one recognized domain. Bisociation is high S with high P and high A — precisely the upper-right quadrant of Figure 1.

Fauconnier and Turner's conceptual blending [Fauconnier and Turner, 2002] specifies the cognitive mechanism. Their framework requires that the "blended space" integrating two input spaces satisfy three properties: (i) *composition* — elements from both input spaces are present (maintained by P: the blend retains structural patterns from both sources); (ii) *completion* — background knowledge fills in the blend (captured by A: the blend aligns with existing domain knowledge that provides the completing context); (iii) *elaboration* — the blend generates inferences that neither input space produces alone (captured by R: the blend's embedding neighborhood spans diverse sub-topics, indicating it opens new lines of inquiry).

Proposition 1 (CN Operationalizes Bisociation)**.** *A text x exhibits bisociation in the sense of Koestler if and only if it simultaneously satisfies $S(x) > \theta_S$ (the combination of frames is unexpected) and $\Delta(x) > \theta_\Delta$ (the combination maintains pattern retention, domain alignment, and generativity), where θ_S and θ_Δ are domain-calibrated thresholds. The coherent novelty score $CN(x) = S(x) \cdot \sigma(\Delta(x))$ is maximized when both conditions hold simultaneously.*

3.2 Boden's Creativity Taxonomy

Boden distinguishes three types of creativity [Boden, 2004]. *Exploratory* creativity (working within a conceptual space) produces moderate S with high A — the output stays aligned with the domain but explores its boundaries. *Combinational* creativity (connecting distant ideas) produces high S with high P and A — the bisociative case described above. *Transformational* creativity (changing the conceptual space itself) produces very high S with initially moderate A that increases over time as the field absorbs the new framework. CN can track this temporal evolution: a transformational contribution starts as high-S, moderate-Δ and shifts to moderate-S, high-Δ as the domain anchors are recomputed to include the new work.

3.3 Divergent Thinking and the Torrance Tests

Guilford's divergent thinking framework [Guilford, 1967] and the Torrance Tests of Creative Thinking [Torrance, 1974] decompose creativity into fluency, flexibility, originality, and elaboration. CN maps onto these: originality corresponds to S (statistical rarity in the response space); flexibility corresponds to R (semantic diversity of the neighborhood, indicating the response connects to multiple categories); elaboration corresponds to P (structural detail retained in the output). Fluency (number of responses) is orthogonal to CN, which evaluates individual responses. This mapping allows CN to be validated against Torrance subscores on divergent thinking tasks (Section 7.2).

3.4 Why Not Perplexity Plus Human Ratings?

A natural objection: why not compute perplexity for novelty and use human ratings for usefulness, combining them post hoc? Three reasons.

First, *scalability*. Human ratings do not scale to the millions of candidate outputs generated by LLM-based discovery systems, literature screening pipelines, or creative writing assistants. CN is fully automated: given a reference corpus and pre-computed embeddings, scoring a new text requires one language model forward pass (for S) and one nearest-neighbor lookup (for Δ). The marginal cost per text is milliseconds.

Second, *decomposability*. A combined perplexity + human-rating score is a black box: when the score is low, you cannot determine whether the text failed on novelty, on structural coherence, on domain relevance, or on generativity. CN decomposes into S, P, A, and R, each of which is independently interpretable and actionable. A practitioner can diagnose that a candidate hypothesis has high S and P but low A (novel and well-structured but disconnected from the field) and adjust accordingly.

Third, *temporal consistency*. Human ratings of "usefulness" shift as fields evolve — what seems impractical today may be foundational tomorrow. The embedding-based coherence operators provide a snapshot of the current field structure, and by recomputing anchors periodically, CN tracks how the field's receptivity to a given idea changes over time. Human ratings, once collected, are static.

CN does not replace human evaluation. It provides a computationally tractable first-pass filter and a decomposable diagnostic, complementing rather than competing with expert judgment.

4 Temperature as a Creativity Parameter

Sampling temperature τ in LLM text generation modulates the softmax distribution over the vocabulary: $p_\tau(w_i) \propto \exp(\ell_i/\tau)$, where ℓ_i is the logit for token w_i. As $\tau \to 0$, the distribution collapses to argmax (greedy, deterministic); as $\tau \to \infty$, it approaches uniform (random sampling). Peeperkorn et al. [Peeperkorn et al., 2024] investigated whether temperature is "the creativity parameter" of LLMs but found mixed results: higher temperature increased lexical diversity but did not consistently improve human creativity ratings.

CN explains this result. Increasing τ raises semantic surprise S monotonically (the model samples less probable tokens, increasing perplexity). But coherence Δ is non-monotonic: at moderate temperatures, the model explores meaningful alternative phrasings and associations, maintaining high P and A while increasing R (the output connects to more diverse regions of the embedding space). Beyond a critical temperature, coherence collapses — pattern retention P drops as the text loses grammatical and semantic structure, and alignment A drops as the text drifts from all domain anchors.

The result is an inverted-U relationship between τ and CN (Figure 2):

$$\mathrm{CN}(\tau) = S(\tau) \cdot \sigma\big(\Delta(\tau)\big) \tag{8}$$

The optimal creative temperature τ^* is the value that maximizes this product:

$$\tau^* = \arg\max_{\tau > 0} \; S(\tau) \cdot \sigma\big(\Delta(\tau)\big) \tag{9}$$

This prediction is directly testable: generate n texts at each of T temperature values, compute CN for each, and estimate τ^* as the temperature at which the mean CN peaks. The prediction is that τ^* lies strictly between 0 and 1 — neither greedy decoding nor high-temperature sampling maximizes creativity.

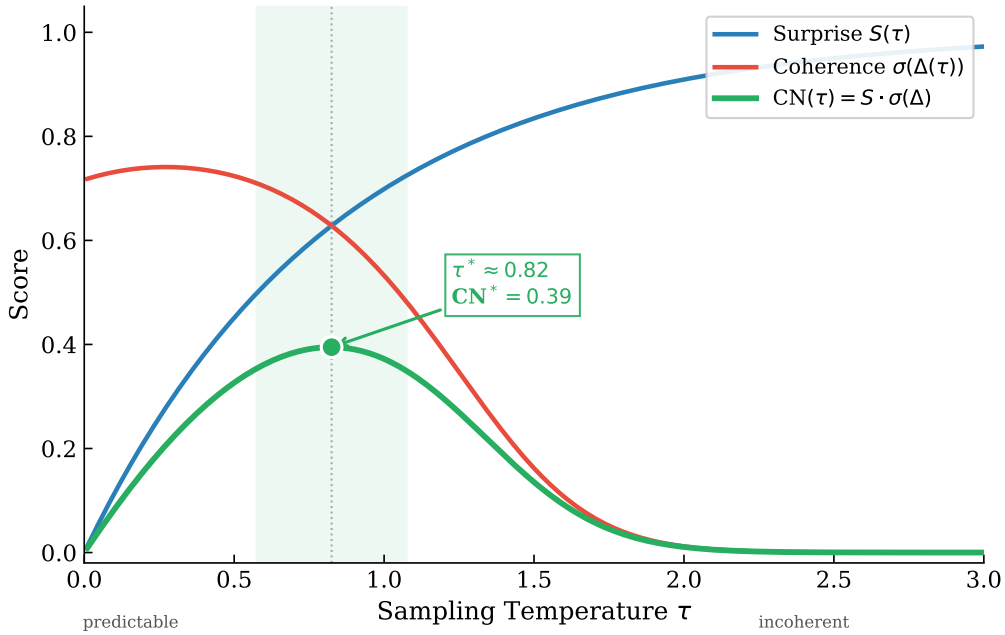

Figure 2: Schematic relationship between sampling temperature τ and the components of coherent novelty. Semantic surprise $S(\tau)$ rises monotonically. Coherence $\sigma(\Delta(\tau))$ first rises (the model explores meaningful alternatives) then falls (the model produces incoherent output). The product $\mathrm{CN}(\tau) = S \cdot \sigma(\Delta)$ peaks at an intermediate optimal creative temperature τ^*.

4.1 Csikszentmihalyi's Flow as the CN Sweet Spot

The temperature–CN tradeoff maps onto Csikszentmihalyi's flow theory [Csikszentmihalyi, 1990, 1996]: the creative state emerges at the boundary between boredom (low surprise, τ too low) and anxiety (high surprise without control, τ too high). The flow channel corresponds to the ridge of the $\mathrm{CN}(\tau)$ curve — sufficient challenge (S elevated above baseline) with maintained skill (Δ still high). This connection transforms flow from a qualitative phenomenological description into a quantitative prediction: flow occurs at $\tau \approx \tau^*$.

5 Worked Example: Three Scientific Abstracts

To make the CN computation concrete, we walk through a complete calculation on three hypothetical scientific abstracts, each representing a different quadrant of the surprise–coherence space (Figure 1).

5.1 Setup

Reference corpus $\mathcal{C}$: 50,000 machine learning abstracts from arXiv (cs.LG, 2020–2025), embedded with nomic-embed-text ($d = 768$, ℓ_2-normalized).

Language model: Llama 3.1 8B, used to compute perplexity.
Anchors: $M = 20$ cluster centroids from k-means on $\mathcal{C}$.
Corpus statistics: $\text{PP}_{\text{base}} = 18.2$, $\text{PP}_{\text{ceil}} = 142.7$ (99th percentile).
Neighborhood size: $k = 10$.

5.2 Text A: Conventional (Low S, High Δ)

"We propose a transformer-based architecture for image classification that achieves state-of-the-art results on ImageNet. Our model uses multi-head self-attention with learned positional encodings and is trained with standard cross-entropy loss."

- **Perplexity**: $\text{PP}(x_A) = 22.4$ (barely above corpus mean — expected content)
- **Surprise**: $S(x_A) = (22.4 - 18.2)/(142.7 - 18.2) = 0.034$
- **Pattern retention**: $P(x_A) = 0.82$ (10 nearest neighbors are highly similar transformer papers)
- **Alignment**: $A(x_A) = 0.89$ (closest anchor: "vision transformers" cluster)
- **Generativity**: $R(x_A) = 0.18$ (neighbors are all image classification papers — low diversity)
- **Coherence**: $\Delta(x_A) = \frac{0.82 \times 0.89 \times 0.18}{(1-0.89)+(1-0.82)+10^{-6}} = \frac{0.131}{0.29} = 0.453$
- **CN**: $\text{CN}(x_A) = 0.034 \times \frac{0.453}{1.453} = 0.034 \times 0.312 = 0.011$

Diagnosis: low surprise, moderate coherence. Competent but not creative.

5.3 Text B: Noise (High S, Low Δ)

"The quantum eigenstate of the transformer collapses through recursive Bayesian cooking, achieving flavor-gradient descent on non-Euclidean pasta manifolds with stochastic oregano regularization."

- **Perplexity**: $\text{PP}(x_B) = 134.6$ (near ceiling — very unexpected)
- **Surprise**: $S(x_B) = (134.6 - 18.2)/(142.7 - 18.2) = 0.935$
- **Pattern retention**: $P(x_B) = 0.21$ (nearest neighbors are distant and semantically unrelated)
- **Alignment**: $A(x_B) = 0.15$ (far from all domain anchors)
- **Generativity**: $R(x_B) = 0.72$ (neighbors span many unrelated areas — but this is incidental, not structured)
- **Coherence**: $\Delta(x_B) = \frac{0.21 \times 0.15 \times 0.72}{(1-0.15)+(1-0.21)+10^{-6}} = \frac{0.023}{1.64} = 0.014$
- **CN**: $\text{CN}(x_B) = 0.935 \times \frac{0.014}{1.014} = 0.935 \times 0.014 = 0.013$

Diagnosis: extremely high surprise, near-zero coherence. The denominator in Δ dominates because both A and P are low, suppressing the score despite moderate R. This is noise, not creativity.

5.4 Text C: Creative (High S, High Δ)

"We show that the internal representations of large language models undergo phase transitions during in-context learning, analogous to critical phenomena in statistical mechanics. By measuring the Fisher information metric on the model's representation manifold, we identify a critical context length at which few-shot performance discontinuously improves, and derive scaling laws governing the transition."

- **Perplexity**: $\text{PP}(x_C) = 87.3$ (substantially above corpus mean)
- **Surprise**: $S(x_C) = (87.3 - 18.2)/(142.7 - 18.2) = 0.555$
- **Pattern retention**: $P(x_C) = 0.61$ (neighbors include scaling laws papers, representation geometry papers — recognizable cluster)
- **Alignment**: $A(x_C) = 0.68$ (aligns with "deep learning theory" and "in-context learning" anchors)
- **Generativity**: $R(x_C) = 0.74$ (neighbors span statistical mechanics, scaling laws, few-shot learning, representation geometry — high diversity)
- **Coherence**: $\Delta(x_C) = \frac{0.61 \times 0.68 \times 0.74}{(1-0.68)+(1-0.61)+10^{-6}} = \frac{0.307}{0.71} = 0.432$
- **CN**: $\text{CN}(x_C) = 0.555 \times \frac{0.432}{1.432} = 0.555 \times 0.302 = 0.168$

Diagnosis: moderately high surprise (statistical mechanics language in an ML abstract) with strong coherence (the physics–ML bridge is structured, not random). The neighborhood is genuinely diverse ($R = 0.74$) because the abstract sits at a real interdisciplinary intersection. CN is an order of magnitude higher than either the conventional or noisy abstract.

5.5 Summary

Text	S	P	A	R	Δ	$\sigma(\Delta)$	CN
A (Conventional)	0.034	0.82	0.89	0.18	0.453	0.312	0.011
B (Noise)	0.935	0.21	0.15	0.72	0.014	0.014	0.013
C (Creative)	0.555	0.61	0.68	0.74	0.432	0.302	0.168

Table 2: CN computation for three abstracts. Perplexity alone would rank B > C > A. CN ranks C $\gg$ B $\approx$ A, correctly identifying the creative abstract. Note that B's high generativity ($R = 0.72$) does not rescue it: the denominator penalty from low P and low A suppresses Δ.

This example illustrates why perplexity alone fails: Text B (noise) has the highest perplexity but the lowest CN. The coherence operators correctly identify that B's surprise comes from incoherence, not from a structured connection between distant ideas. Text C, which bridges statistical mechanics and machine learning, scores highest on CN because its surprise is paired with a coherent neighborhood spanning multiple recognized sub-fields.

6 Connections to Prior Work

6.1 Out-of-Distribution Detection

Out-of-distribution detection in NLP [Lang et al., 2023] asks whether an input is anomalous relative to a training distribution. This is a related but distinct problem: OOD detection identifies

that a text is unusual, while CN asks whether an unusual text is *meaningfully* unusual. An OOD detector assigns high scores to both Text B (noise) and Text C (creative) from the worked example above. CN separates them. The coherence operators function as a post-OOD filter: given that a text has been flagged as surprising, do the embedding-space operators indicate structured novelty or mere noise?

6.2 Embedding Compression and Coherence Loss

Our work on embedding dimensionality [Thorarinson and Hensgen, 2026b] showed that PCA compression preserves broad categories while destroying fine-grained distinctions. In the creativity context, this maps directly: a compressed representation can identify that a text is "about machine learning" (low surprise at the category level) while losing the specific novel claim that makes it creative (high surprise at the detail level). The pattern retention operator P detects exactly this failure: when embeddings are compressed, nearest-neighbor structure degrades, and $P(x)$ drops even when coarse-grained category labels remain correct. Coherent novelty requires full-dimensional embeddings — the creative signal lives in the long tail of the spectrum, precisely the dimensions that PCA discards first.

6.3 Neurodivergent Cognition and Attentional Oscillation

Our work on neurodivergence as coherence variation defined attentional oscillation λ as the variance of attention-related EEG phase dynamics [Klimesch, 2012, Fink and Benedek, 2009]. High λ produces broader phase-space exploration, generating more surprising associations — analogous to high sampling temperature in LLMs (Section 4). But high λ alone does not produce creativity. It is the coherence operators (P, A, R) that determine whether the surprising associations form structured patterns or dissolve into noise. This mirrors Martindale's model of biological bases of creativity, where defocused attention enables remote associations but requires a subsequent focusing phase for evaluation [Martindale, 1999].

Proposition 2 (Neurodivergent Creativity)**.** *ADHD-type cognition with high λ (broad attentional oscillation) produces elevated $S(x)$ (semantic surprise) in text generation. The coherent novelty score $CN = S \cdot \sigma(\Delta)$ predicts that neurodivergent individuals will produce higher CN than neurotypical individuals on divergent thinking tasks, provided their coherence Δ is supported (low external noise, stable environmental patterns, alignment with the task domain). Under unsupported conditions, S remains high but Δ collapses, yielding noise rather than creativity.*

This prediction is directly falsifiable (Section 7.3).

6.4 Scientific Discovery and Hypothesis Ranking

The Coherence Engine [Thorarinson and Hensgen, 2026a] uses coherence operators to evaluate bodies of evidence and detect structural drift before failure. Coherent novelty applies at the hypothesis level: a novel hypothesis has high S (it surprises the field's language model) and high Δ (it maintains pattern retention with existing evidence, aligns with known mechanisms, and generates testable predictions). This connects to Simonton's chance-configuration theory, which holds that the most creative scientific contributions arise from the fortuitous combination of existing mental elements [Simonton, 2004] — precisely the high-S, high-Δ regime — and to Wang et al.'s vision of AI-driven scientific discovery [Wang et al., 2023].

A literature agent can use CN to rank candidate hypotheses: surprise alone favors wild speculation; coherence alone favors incremental work; the product favors structured breakthroughs.

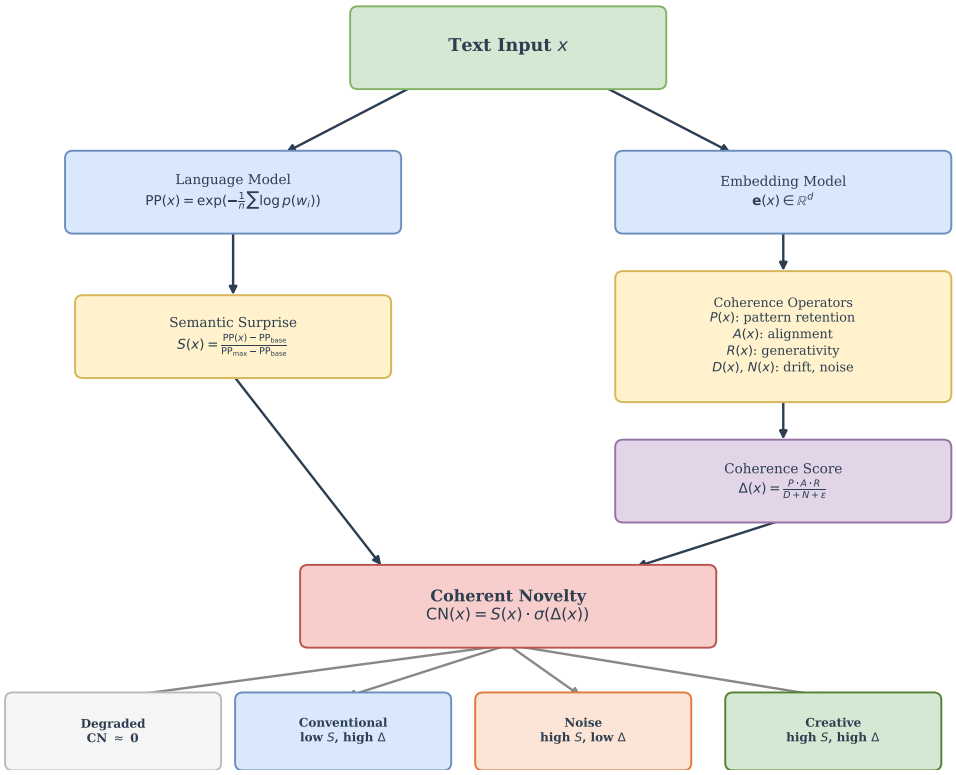

Figure 3: The CN evaluation pipeline. **Input**: text x and reference corpus $\mathcal{C}$. **Branch 1**: a language model computes perplexity $\text{PP}(x)$, normalized to surprise $S(x) \in [0, 1]$. **Branch 2**: an embedding model computes $\mathbf{e}(x)$, a k-NN lookup against $\mathcal{C}$ yields $P(x)$ and $R(x)$, and a comparison against pre-computed anchor centroids yields $A(x)$. These combine into $\Delta(x)$, squashed to $\sigma(\Delta) \in [0, 1)$. **Output**: $\text{CN}(x) = S \cdot \sigma(\Delta)$, with decomposed diagnostics (S, P, A, R, Δ) available for interpretation.

7 Evaluation Protocol

The evaluation pipeline (Figure 3) is designed for reproducibility. All components use publicly available models and can be implemented in under 200 lines of Python. Algorithm 1 provides pseudocode.

7.1 Experiment 1: Scientific Abstract Creativity

Dataset: 10,000 scientific abstracts from arXiv cs.LG, stratified by citation impact (top 10% vs. bottom 10% by 3-year citations, sourced from Semantic Scholar API).

Reference corpus: 50,000 additional cs.LG abstracts from the same period, embedded with nomic-embed-text ($d = 768$).

Language model: Llama 3.1 8B (for perplexity) or SciBERT (for domain-tuned perplexity).

Protocol: Compute $S(x)$, $\Delta(x)$, and $\text{CN}(x)$ for each abstract. Report Spearman rank correlation of each metric with log citation count. Report AUC for binary classification (top 10% vs. bottom 10%).

Hypothesis: CN achieves higher rank correlation with citation impact than S alone ($\rho_{\text{CN}} > \rho_S$). High-CN abstracts introduce ideas that are both novel and connected to the existing literature.

Baseline comparisons: (1) perplexity alone; (2) embedding distance from corpus centroid (a proxy for novelty without coherence decomposition); (3) combined perplexity + embedding

Algorithm 1 Compute Coherent Novelty

Require: Text x, reference corpus embeddings $\mathcal{C} \in \mathbb{R}^{|\mathcal{C}| \times d}$, anchor centroids $\mathbf{A} \in \mathbb{R}^{M \times d}$, language model $\mathcal{M}$, corpus statistics $(\text{PP}_{\text{base}}, \text{PP}_{\text{ceil}})$, k, ϵ

Ensure: $\text{CN}(x) \in [0, 1]$ and diagnostics (S, P, A, R, Δ)

$\text{PP}(x) \leftarrow \mathcal{M}.\text{perplexity}(x)$ ▷ Forward pass through language model

$S \leftarrow \text{clip}\big((\text{PP}(x) - \text{PP}_{\text{base}})/(\text{PP}_{\text{ceil}} - \text{PP}_{\text{base}}),\ 0,\ 1\big)$

$\mathbf{e}(x) \leftarrow \text{embed}(x)$ ▷ ℓ_2-normalized embedding

$\mathbf{n}_1, \ldots, \mathbf{n}_k \leftarrow \text{kNN}(\mathbf{e}(x), \mathcal{C}, k)$ ▷ k nearest neighbors

$P \leftarrow \frac{1}{k} \sum_{j=1}^{k} \cos(\mathbf{e}(x), \mathbf{n}_j)$

$A \leftarrow \max_m \cos(\mathbf{e}(x), \mathbf{a}_m)$

$R \leftarrow 1 - \frac{1}{\binom{k}{2}} \sum_{i<j} \cos(\mathbf{n}_i, \mathbf{n}_j)$

$\Delta \leftarrow (P \cdot A \cdot R) \,/\, ((1 - A) + (1 - P) + \epsilon)$

$\text{CN} \leftarrow S \cdot \Delta/(\Delta + 1)$

return $\text{CN}, (S, P, A, R, \Delta)$

distance (additive and multiplicative).

7.2 Experiment 2: Creative Writing

Dataset: 1,000 short stories ($\leq$ 500 words) with human creativity ratings collected via the Consensual Assessment Technique [Amabile, 1982], with at least three expert judges per story scoring on a 1–7 scale for overall creativity, novelty, and coherence independently.

Reference corpus: 10,000 short stories from public writing repositories, embedded with Sentence-BERT.

Protocol: Compute $\text{CN}(x)$ for each story. Report Spearman rank correlation with mean expert creativity rating. Additionally report correlations of S, P, A, R individually with the Torrance subscales (fluency, flexibility, originality, elaboration) [Torrance, 1974] adapted for prose.

Hypothesis: CN correlates more strongly with human creativity ratings than perplexity alone. The decomposition reveals that S correlates with originality subscores while R correlates with flexibility subscores.

7.3 Experiment 3: Neurodivergent Text Generation

Participants: $N = 60$ (30 ADHD-diagnosed, 30 neurotypical controls), matched for age, education, and verbal IQ.

Task: Two conditions per participant: (1) *supported* — quiet environment, unlimited time, clear task framing ("write a paragraph combining two concepts: [concept A] and [concept B]"); (2) *unsupported* — noisy environment, 2-minute time limit, ambiguous framing.

Protocol: Compute CN, S, and Δ for each response. Fit a 2×2 ANOVA (group $\times$ condition) on each metric.

Hypothesis: Main effect of group on S (ADHD $>$ controls). Interaction effect on Δ: ADHD participants show higher Δ than controls under supported conditions but lower Δ under unsupported conditions. CN shows the predicted crossover: ADHD $>$ controls (supported) and ADHD $<$ controls (unsupported), confirming Proposition 2.

8 Discussion

8.1 Creativity Is Surprise That Coheres

The central claim is that creativity is not surprise. Creativity is surprise that coheres. This echoes the standard definition [Runco and Jaeger, 2012] but operationalizes it in terms of computable quantities. A language model generating maximally surprising text is not creative — it is random [Holtzman et al., 2020]. A model generating text that is slightly surprising, connects to existing knowledge, maintains internal structure, and sits at an intersection that opens further inquiry is creative. CN captures this by requiring both S and $\sigma(\Delta)$ to be elevated simultaneously (Table 1, Figure 1).

The worked example (Section 5, Table 2) demonstrates that perplexity alone fails to distinguish noise from creativity: the nonsense abstract and the conventional abstract both receive near-zero CN, while the interdisciplinary abstract — bridging statistical mechanics and machine learning — scores an order of magnitude higher. This is the discriminative power that perplexity lacks.

8.2 Implementation Considerations

CN is designed for practical deployment. The computational bottleneck is the k-NN lookup against the reference corpus, which can be accelerated to sub-millisecond latency using approximate nearest neighbor libraries (FAISS, Qdrant, Annoy). The language model forward pass for perplexity is a single inference call. Pre-computing corpus embeddings and anchor centroids is a one-time cost. For a reference corpus of 50,000 documents with $d = 768$ embeddings, the total storage is ≈150 MB.

Domain adaptation requires: (1) assembling a reference corpus representative of the target domain; (2) embedding it and computing k-means anchors; (3) computing corpus perplexity statistics PP_{base} and PP_{ceil}. No fine-tuning of models is needed.

8.3 Limitations

Model dependence. CN scores computed with different embedding models or language models are not directly comparable. Relative rankings within a single configuration should be stable, but absolute CN values are model-specific. This is inherent to any embedding-based metric and shared by methods such as BERTScore.

Domain specificity. The anchor embeddings, reference corpora, and perplexity normalization constants are domain-specific. A CN score computed against a machine learning reference corpus is not comparable to one computed against a creative writing corpus. This is a feature, not a bug: creativity is domain-relative.

Retrospective vs. prospective generativity. The generativity operator R approximates future productivity via neighborhood diversity. This is a structural proxy, not a causal measure. A text can have high R (diverse neighbors) without actually generating follow-up work, and vice versa. Retrospective validation against citation counts (Experiment 1) will calibrate the relationship between prospective R and actual generativity.

Temporal dynamics. The reference corpus is a snapshot. As a field evolves, formerly creative ideas become conventional (their embeddings move closer to the centroid), and new creative frontiers open. Longitudinal studies should recompute $\mathcal{C}$ and anchors at regular intervals to track how CN evolves for a fixed text over time.

Confounds in neurodivergence experiments. The ADHD–creativity prediction (Proposition 2) requires careful experimental design to separate attentional oscillation effects from confounds including medication status, comorbidities, and task familiarity.

9 Conclusion

Divergent thinking leaves measurable traces in language through model-detected surprise. But surprise alone is not creativity. Coherent novelty pairs surprise with coherence, asking whether unexpected text retains structural patterns (P), connects to the domain (A), and opens new directions (R). The metric is computable from a reference corpus and two models (language model for perplexity, embedding model for coherence), requires no human annotation at scoring time, and decomposes into interpretable sub-scores.

CN operationalizes Koestler's bisociation (high S with high A), Fauconnier–Turner blending (composition via P, completion via A, elaboration via R), and the standard novelty-plus-usefulness definition of creativity. It predicts an optimal creative temperature τ^* at which LLM sampling maximizes the surprise–coherence product, connecting phenomenological flow theory to a measurable quantity.

The three proposed experiments — scientific abstract impact prediction, creative writing evaluation, and neurodivergent text generation — provide falsifiable tests. If CN predicts citation impact better than perplexity alone, correlates with expert creativity ratings, and shows the predicted ADHD–condition interaction, the framework is validated. If not, the decomposed scores (S, P, A, R) will reveal which component of the creativity model fails, providing diagnostic value regardless of the overall result.

References

Teresa M Amabile. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013, 1982.

Margaret A Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2 edition, 2004.

Mihaly Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper & Row, New York, 1990.

Mihaly Csikszentmihalyi. *Creativity: Flow and the Psychology of Discovery and Invention*. HarperCollins, New York, 1996.

Gilles Fauconnier and Mark Turner. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, New York, 2002.

Andreas Fink and Mathias Benedek. The creative brain: Investigation of brain activity during creative problem solving by means of EEG and fMRI. *Human Brain Mapping*, 30(3):734–748, 2009.

Giorgio Franceschelli and Mirco Musolesi. Creativity and machine learning: A survey. *ACM Computing Surveys*, 56(11):1–41, 2024. doi: 10.1145/3664595.

Joy Paul Guilford. *The Nature of Human Intelligence*. McGraw-Hill, 1967.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49 (10):1295–1306, 2009. doi: 10.1016/j.visres.2008.09.007.

Wolfgang Klimesch. Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12):606–617, 2012.

Arthur Koestler. *The Act of Creation*. Hutchinson, London, 1964.

Hao Lang, Yinhe Zheng, Yixuan Li, Jian Sun, Fei Huang, and Yongbin Li. A survey on out-of-distribution detection in NLP. *arXiv preprint arXiv:2305.03236*, 2023. URL https://arxiv.org/abs/2305.03236.

Colin Martindale. Biological bases of creativity. *Handbook of Creativity*, pages 137–152, 1999.

Michael D Mumford. Where have we been, where are we going? Taking stock in creativity research. *Creativity Research Journal*, 15(2–3):107–120, 2003. doi: 10.1080/10400419.2003.9651403.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*, 2024. URL https://arxiv.org/abs/2405.00492.

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

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

Joel Thorarinson and Allison Hensgen. From prediction to discoverative intelligence: A coherence-based AI framework for detecting system drift before failure. *arXiv preprint*, 2026a.

Joel Thorarinson and Allison Hensgen. The dimensionality illusion: Why PCA variance does not equal semantic information in text embeddings. *arXiv preprint*, 2026b.

E Paul Torrance. *Torrance Tests of Creative Thinking: Norms–Technical Manual*. Scholastic Testing Service, Bensenville, IL, 1974.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Yuhua Yu, Quentin Raffaelli, Simone Luchini, Roger E. Beaty, and Jessica R. Andrews-Hanna. Tracking divergent thinking: Perplexity as a unified metric of surprise. *Thinking Skills and Creativity*, 2026.

Zhen Zhang et al. Language model perplexity predicts scientific surprise and transformative impact. *arXiv preprint arXiv:2509.05591*, 2025. URL https://arxiv.org/abs/2509.05591.