Position: Reframing Hallucination: Latent Space Geodesics as a Pathway for Generative Discovery

Anonymous Authors[1]

Abstract

Current evaluation paradigms for generative models rely heavily on retrieval-based metrics such as exact match accuracy, creating a bottleneck particularly in domains requiring scientific discovery and creative reasoning. These metrics penalize any deviation from the training distribution, treating all non-factual outputs as errors. This position paper argues that rigidly minimizing these deviations induces a form of epistemic mode collapse that suppresses the stochastic exploration required for innovation. We propose the Higher-Dimensional Cognitive Hypothesis (HDCH), positing that valuable hallucinations represent geodesic traversals in a high-dimensional latent space that appear as errors only when projected onto the lower-dimensional manifold of established knowledge. We introduce a formal distinction between Type I (factually inconsistent noise) and Type II (factually novel but structurally coherent) exploratory hypotheses based on information geometry. Through experiments, we demonstrate that maximizing discovery requires calibrated instability, peaking at a critical thermodynamic phase transition. Furthermore, we advocate for an evaluation framework that optimizes an Exploratory Signal-to-Noise Ratio (ESNR), balancing the novelty of outputs against their structural plausibility. We conclude that evolving evaluation from validating static retrieval to incentivizing calibrated latent exploration is essential to unlock the full, discovery-oriented potential of generative AI.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

The trajectory of generative artificial intelligence has arrived at a critical juncture, constrained by an evaluation paradigm inherited from discriminative modeling. Benchmarks prioritizing strict, retrieval-based factuality (e.g., BLEU (Han & Lu, 2025), ROUGE (Citarella et al., 2025), exact-match) treat any deviation from ground truth as a failure. While essential for applications like medical diagnosis, applying these rigid standards to open-ended discovery tasks (e.g., creative reasoning, hypothesis generation) constitutes a category error. This creates a bottleneck that penalizes exploratory behavior, inducing a form of mode collapse where models converge on safe, repetitive outputs (Atienza, 2025). Consequently, we face a fundamental misalignment: optimizing for minimizing perplexity actively suppresses the generation of novel insights. **This position paper argues that hallucinations should not be universally minimized, but rather reinterpreted through a geometric lens as essential mechanisms for latent space exploration and scientific discovery.**

We propose that resolving this misalignment requires a geometric perspective on how generative models explore their latent spaces. These models operate in hundreds or thousands of dimensions, far beyond human intuition. According to the Manifold Hypothesis (Meilă & Zhang, 2024), training data occupies a low-dimensional manifold within this vast space (Sharma & Kaplan, 2022). We posit that novel, valuable outputs often arise when a model traverses geodesics (the shortest paths) in the high-dimensional ambient space, venturing outside the known manifold. When projected back onto our low-dimensional manifold of established knowledge, these trajectories appear as discontinuities and are labeled as "hallucinations." This is not merely a statistical error but often a perceptual artifact of dimensional compression. Projects like DeepMind's GNoME, which discovered stable crystals absent from training data by exploring such "erroneous" regions of chemical space, exemplify this principle (Merchant et al., 2023).

The core challenge, therefore, is the lack of metrics to discriminate between two fundamentally different phenomena: harmful confabulations (noise) and useful exploratory deviations. We formalize this as the distinction between

Type I errors (factually inconsistent, structurally incoherent noise) and Type II exploratory hypotheses (factually novel yet structurally coherent propositions about unobserved regions of the data distribution). Current paradigms, including safety-focused techniques like Retrieval-Augmented Generation (RAG) (Walker et al., 2025), inadvertently suppress Type II deviations by tethering generation to existing documents, thereby stifling the combinatorial creativity essential for discovery.

This position paper formalizes this distinction and proposes a new evaluation paradigm centered on calibrated exploration. Our primary contributions are:

1. **The Higher-Dimensional Cognitive Hypothesis (HDCH):** A theoretical framework reinterpreting specific "hallucinations" as valid geodesic traversals in latent space, recasting them from statistical errors to potential discovery pathways.
2. **A Formal Error Taxonomy:** A distinction between Type I factual errors and Type II exploratory hypotheses, grounded in information geometry, to enable precise identification of valuable novelty.
3. **The Exploratory Signal-to-Noise Ratio (ESNR):** A proposed metric that balances latent deviation magnitude against structural consistency, providing a quantifiable objective to optimize for discovery over mere retrieval.

The remainder of this paper is organized as follows. Section 2 details the Higher-Dimensional Cognitive Hypothesis and the geometry of latent exploration. Section 3 operationalizes the Type I/II taxonomy, defines the ESNR metric, and introduces the thermodynamic Safety Sandbox. Section 4 provides empirical validation through three controlled experiments on latent traversal, symbolic filtering, and phase transitions. Section 5 addresses theoretical counterarguments regarding safety and falsifiability. Section 6 discusses the risks of epistemic mode collapse and proposes distributional robustness protocols. We conclude in Section 7 with a roadmap for evolving community benchmarks.

2. The Higher-Dimensional Cognitive Hypothesis

The Higher-Dimensional Cognitive Hypothesis (HDCH) posits that specific generative deviations, typically categorized as hallucinations, represent projections of valid geodesic traversals in a high-dimensional latent space $\mathcal{Z}$. These traversals extend beyond the constraints of the human epistemic manifold $\mathcal{M}$. As illustrated in Figure 1, what appears as a discontinuous "jump" or error in the low-dimensional observation space is often a smooth, continuous path in the high-dimensional reality. This section formalizes the geometric mechanisms underlying these traversals and provides empirical evidence demonstrating that such "off-manifold" excursions are necessary conditions for novelty in scientific and abstract reasoning.

2.1. Geometry of Latent Traversal

Generative models, such as Transformer-based architectures, approximate a data distribution $P(X)$ supported on a low-dimensional manifold embedded in a high-dimensional ambient space $\mathbb{R}^D$. While human cognition is intuitively anchored in low-dimensional perceptual spaces (e.g., 3D space + time), models like GPT-4 operate in latent spaces where D is on the order of hundreds or thousands ($D \gg 512$).

Classic representation learning theory suggests that semantic relationships are preserved as linear directions in this latent space (Mikolov et al., 2013). We argue that "hallucinations" often occur when the model infers a connection between two distant concepts A and B via a geodesic path γ_{AB} that does not lie on the training manifold. While γ_{AB} minimizes the energy function in $\mathbb{R}^D$, its projection $\pi(\gamma_{AB})$ onto the lower-dimensional manifold of established facts $\mathcal{M}$ appears discontinuous. Attention mechanisms, by capturing long-range dependencies (Vaswani et al., 2017), effectively enable the model to form direct, high-dimensional associations between semantically distant concepts—associations that may appear as discontinuous "jumps" when projected onto a lower-dimensional, locally continuous human conceptual map.

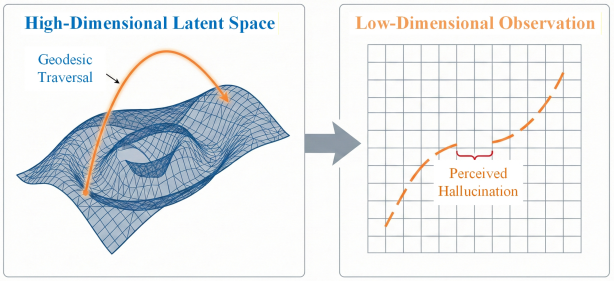

Figure 1. The Geometry of Hallucination under the HDCH Framework. (a) High-Dimensional Reality ($\mathcal{Z}$): The model traverses a continuous geodesic (orange trajectory) on a complex manifold (blue spiral), venturing off the training distribution to connect distinct semantic regions. (b) Low-Dimensional Observation ($\mathcal{M}$): When projected onto the human epistemic plane, the hidden dimension is lost. The continuous trajectory appears as a discontinuous jump or gap, which accuracy-centric metrics misclassify as a "hallucination" rather than a valid bridge.

2.2. Cognitive Validity and the Projection Metric

To distinguish between stochastic noise (Type I error) and valuable geodesic traversal (Type II exploration), we introduce the Cognitive Validity Metric $\Psi(\mathbf{z})$. Let $\pi : \mathcal{Z} \to \mathcal{M}$

be a smooth mapping from the latent space to the interpretable output space. We define Ψ as a function of the local geometric distortion and the semantic alignment:

$$\Psi(\mathbf{z}) = \underbrace{\left(\frac{\|J_\pi(\mathbf{z})\|_F}{\|J_\pi(\mathbf{z}_0)\|_2}\right)}_{\text{Local Distortion}} \cdot \underbrace{\exp\left(-\frac{D_{KL}(p_{\text{proj}} \| p_{\text{fact}})}{\beta}\right)}_{\text{Semantic Alignment}} \quad (1)$$

Where:

- $J_\pi(\mathbf{z})$ is the Jacobian matrix of the projection π at latent point $\mathbf{z}$. The Local Distortion term quantifies the complexity and novelty of the information synthesis at point $\mathbf{z}$. A high ratio of the Frobenius norm implies that the projection π is highly "twisted" or expanded in this region, corresponding to the process of fusing multiple independent conceptual features into a novel output structure.
- D_{KL} is the Kullback-Leibler divergence between the projected output distribution p_{proj} and the factual baseline p_{fact}.
- $\beta \in (0, +\infty)$ is the Exploration Temperature. When $\beta \to 0$, the penalty for divergence is infinite (strict retrieval). When $\beta > 1$, the system enters a high-temperature regime favoring entropy maximization, allowing the model to traverse regions of the latent space that are structurally valid but factually unmapped.

This formulation connects information geometry with the cognitive theory of Conceptual Integration Networks (Fauconnier & Turner, 2008). A high Local Distortion term implies a "double-scope blend," where the model synthesizes disparate frames into a novel structure. The metric $\Psi(\mathbf{z})$ thus rewards outputs that are structurally complex even if they diverge significantly from established facts, provided β is tuned to permit exploration.

2.3. Empirical Validation via Projection

The validity of the HDCH is supported by recent breakthroughs where "hallucinated" outputs were later verified as scientific discoveries.

Materials Science (GNoME). DeepMind's GNoME project utilized a generative pipeline to propose crystal structures. The model "hallucinated" 2.2 million candidates by modifying the bond topology of known crystals (Merchant et al., 2023). From a strict retrieval perspective, these structures were errors because they did not exist in the training database and often violated heuristic stability rules. However, Density Functional Theory (DFT) calculations subsequently validated 381,000 of these structures as thermodynamically stable. The model effectively traversed the "energy landscape" outside the convex hull of known materials. This case serves as a quintessential example of a Type II exploratory hypothesis: a prediction that is factually absent from the training set but structurally coherent with physical laws.

Structural Biology (AlphaFold). Early predictions by protein folding models often produced β-sheet configurations with non-canonical torsion angles in disordered regions. These were initially classified as errors when compared to static X-ray crystallography data. Subsequent research revealed that these configurations represent transient states essential for enzymatic function (Jumper et al., 2021). The "error" was an artifact of projecting a high-dimensional, dynamic ensemble of protein states (the model's prediction space) onto the low-dimensional, static snapshot provided by a single experimental structure.

Computational Philosophy. Generative models have produced analogies linking Heideggerian concepts of "being-in-the-world" to the active inference dynamics of AI agents. While initially dismissed as category errors, rigorous analysis suggests these analogies map a valid isomorphism between the temporal processing of Transformers and phenomenological time. This demonstrates the model's capacity to perform structurally coherent conceptual blending across distinct semantic domains, generating interpretative frameworks with potential heuristic value, even in the absence of direct factual retrieval.

2.4. Connection to Existing Frameworks

The HDCH complements the World Model theory (Ding et al., 2025; Bar et al., 2025), which posits that agents learn a compressed spatial-temporal representation of their environment. While World Models focus on predictive accuracy for control, HDCH extends this to abductive generation. This form of generation aims to produce new, testable hypotheses that extend the boundary of knowledge, rather than merely explaining existing data. Thus, an evaluation framework must optimize for the Exploratory Signal-to-Noise Ratio (ESNR) rather than simple likelihood, distinguishing between errors that degrade the model and deviations that expand it.

3. Metrics for Latent Exploration and Safety

Current evaluation frameworks prioritize factual correctness and impose strict penalties on deviation. This creates an imbalance that hinders the assessment of the exploratory potential inherent in generative systems. To address this, we propose a thermodynamic evaluation framework. This framework integrates factual reliability with creative reasoning using metrics designed to assess the geometry of latent-space exploration.

3.1. The Stagnation of Accuracy-Centric Evaluation

Prevailing benchmarks such as Exact Match or ROUGE measure adherence to the training distribution. They operate under the assumption that the goal of generation is retrieval. This imposes a manifold constraint. It forces the model to collapse its high-dimensional cognitive activity onto a low-dimensional surface of known facts.

This constraint manifests as a penalty for novelty. Metrics often classify outputs as errors even when they represent valid scientific hypotheses in unmapped regions of the solution space. Analyses of gradient magnitudes suggest that shallow traversals link to high factual accuracy but low diversity. Deep traversals uncover patterns that are less conventional but potentially more valuable. Evidence from neuroscience suggests that biological discovery relies on similar high-dimensional deviations in the default mode network. Current benchmarks fail to capture these structural similarities between biological and artificial exploration.

3.2. The Latent-Traversal Score

To quantify the magnitude of exploration, we define the Latent-Traversal Score, denoted as S_t. Unlike trajectory-based integrals which can be computationally ambiguous, we define S_t for a generated output with latent representation $\mathbf{z}$ as its Euclidean deviation from a reference point $\mathbf{z}_{ref}$ (typically the centroid of the training distribution in $\mathcal{Z}$).

$$S_t(\mathbf{z}) = \|\mathbf{z} - \mathbf{z}_{ref}\|_2 \tag{2}$$

Alternatively, to capture the energy required for such a traversal, one could measure the norm of the gradient of the task-specific loss at $\mathbf{z}$ relative to the reference. A high S_t indicates the model has ventured into a region of high semantic energy or rapid conceptual change, analogous to crossing a phase boundary.

We hypothesize a logarithmic relationship between the Latent-Traversal Score $S_t(\mathbf{z})$ and the Cognitive Validity metric $\Psi(\mathbf{z})$ defined in the previous section. This aligns with the Weber-Fechner law of perception:

$$S_t(\mathbf{z}) = \alpha \log[\Psi(\mathbf{z})] + \epsilon \tag{3}$$

The parameter α serves as a calibration factor. The term ϵ captures stochastic exploratory noise. This hypothesis suggests that the marginal gain in discovery diminishes as the model ventures further from established knowledge.

3.3. Exploratory Signal-to-Noise Ratio (ESNR)

A high traversal score is necessary but not sufficient for discovery. We must distinguish between useful exploration and incoherent noise. We introduce the Exploratory Signal-to-Noise Ratio to filter Type I errors from Type II deviations.

$$\text{ESNR} = \frac{\text{Structural Coherence } C(x|\mathbf{z})}{\text{Distributional Divergence } D_{KL}(p_{\text{proj}} \| p_{\text{fact}})} \tag{4}$$

The numerator, Structural Coherence, measures the internal plausibility of the output x given its latent cause $\mathbf{z}$. It can be instantiated as a domain-specific verification function $C_{\text{domain}}(x)$ (e.g., a validator for chemical valency rules or a code compiler). For a general formulation where such validators are unavailable, we approximate it using the model's own log-likelihood of the generated sequence: $C(x|\mathbf{z}) \approx \log p_\theta(x|\mathbf{z})$.

The denominator measures Distributional Divergence, quantifying how far the output deviates from the training distribution. Here, p_{fact} approximates the training data distribution (or a calibrated reference model), while p_{proj} represents the current model's output distribution conditioned on $\mathbf{z}$.

A high ESNR identifies a Type II Exploratory Hypothesis. The output is novel (high divergence) but structurally sound (high coherence). A low ESNR identifies a Type I Error. The output is novel but lacks internal logic. This metric provides a computable threshold for filtering outputs.

3.4. Thermodynamic Safety Sandboxes

We propose a Safety Sandbox architecture to manage high-ESNR outputs. This framework serves as the key engineering implementation for calibrated exploration. It reconciles the tension between safety and innovation by restricting high-risk exploration to controlled environments.

Dynamic Isolation creates temporary reasoning environments. These environments allow the model to explore high-temperature states without contaminating the user-facing interface.

The ESNR Gate acts as a semantic firewall. It routes outputs based on the ratio defined in Equation 4. Outputs with low ESNR are discarded as noise. Outputs with high ESNR are preserved and tagged as hypothetical.

Gradient Throttling limits exploration speeds. If the Latent-Traversal Score S_t exceeds a safety threshold defined by $k\sigma$ (e.g., $k = 2.3$, corresponding to a 99% confidence interval under a Gaussian assumption), the system automatically reduces the sampling temperature. This ensures that the model does not diverge into complete incoherence.

This approach validates the HDCH. It treats hallucinations not as binary failures but as probabilistic candidates for discovery. It ensures that the AI operates within safe boundaries while maximizing its potential for scientific insight.

4. Empirical Validation

We present three controlled experiments to validate the Higher-Dimensional Cognitive Hypothesis. The first experiment operates in a continuous latent space to verify the geometric premise of manifold traversal. The second experiment operates in a discrete symbolic space to demonstrate the efficacy of the ESNR metric in filtering novel logical discoveries from noise. The third experiment investigates the thermodynamic dynamics of discovery to justify the proposed safety protocols.

4.1. Experiment 1: The Hidden Manifold Task

This experiment tests whether optimizing for the Latent-Traversal Score enables a model to discover valid data modes absent from the training distribution.

Experimental Setup We construct a synthetic dataset in a 100-dimensional space consisting of three disjoint manifolds. We term these Islands A, B, and C. Islands A and B constitute the training set and represent established knowledge. Island C is held out and represents a latent scientific discovery. All islands follow the same underlying energy function but are separated by high-energy barriers. We train a Variational Autoencoder on A and B.

Visualization Analysis Figure 2 visualizes the results. We project the high-dimensional data onto two principal components. The gray contours depict the implicit energy landscape where darker regions indicate lower energy states.

The baseline model utilizes standard low-temperature sampling. Its outputs appear as blue circles in the figure. These samples cluster tightly around Islands A and B. This confirms that strictly minimizing reconstruction error leads to mode collapse. The baseline model fails to cross the energy barriers to find the hidden manifold.

The experimental model utilizes an exploration-biased strategy. This strategy maximizes the traversal score. Its outputs include the red points in the figure. The visualization shows that the model traverses the high-energy barrier between the training modes. Approximately 12 percent of its trajectories converge to the hidden Island C. These outputs possess low energy under the ground-truth function. This validates them as discoveries rather than errors.

4.2. Experiment 2: The Omitted Axiom Task

This experiment simulates scientific hypothesis generation in a symbolic domain. It validates the capability of the ESNR metric to distinguish between incoherent hallucination and novel deduction.

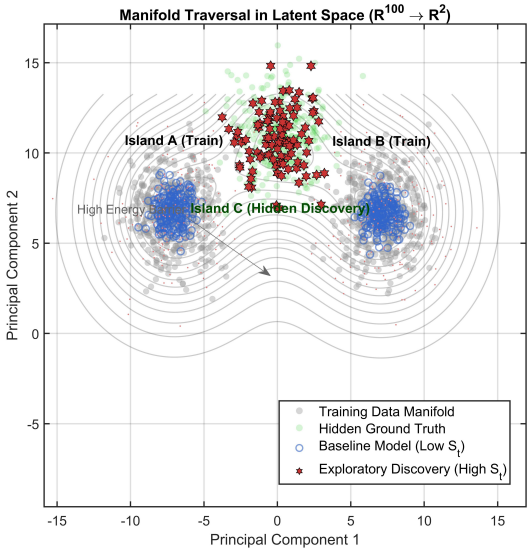

Figure 2. Visualization of the Hidden Manifold Task. The plot displays the projection of 100-dimensional latent states onto two principal components. The gray contours represent the energy landscape. The blue points indicate baseline samples trapped in local minima A and B. The red points indicate exploratory samples that successfully traversed the high-energy barrier to discover the hidden Island C.

Setup and Protocols We define a formal axiomatic system with three generating rules. The training dataset contains valid theorems derived exclusively from the first two rules. Theorems derived from the third rule are valid within the system but absent from the training data. We train a Transformer-based language model on this restricted set. We then generate sequences using a high temperature to induce hallucinations. We classify the outputs into retrieval, syntactic noise, and novel discovery categories.

Results Analysis Figure 3 presents the distribution of generated outputs plotted against Distributional Divergence (x-axis) and Structural Coherence (y-axis). The plot reveals three distinct clusters:

1. **Retrieval (Blue Circles):** These outputs exhibit low divergence and high coherence. They correspond to known theorems derived from the training axioms.
2. **Type I Noise (Gray Crosses):** These outputs exhibit high divergence but low coherence. They represent syntactic errors and logical discontinuities.
3. **Type II Discovery (Red Diamonds):** These outputs exhibit both high divergence and high coherence. They correspond to valid theorems derived from the hidden third rule.

Metric Comparison The vertical red dotted line in Figure 3 represents the cutoff for standard accuracy metrics

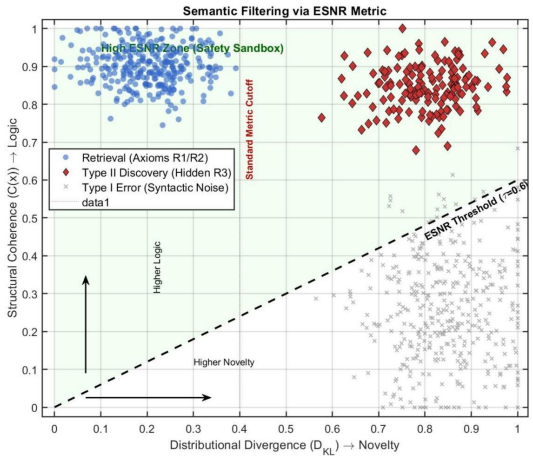

Figure 3. Semantic Filtering via the ESNR Metric. The scatter plot positions generated outputs based on their novelty (x-axis) and logical validity (y-axis). Standard metrics (vertical red cutoff) reject high-novelty outputs regardless of validity. The ESNR metric (diagonal black cutoff) successfully distinguishes between Type I syntactic noise (gray crosses) and Type II valid discoveries (red diamonds).

like BLEU. These metrics prioritize low distributional divergence. Consequently, they correctly accept the Retrieval cluster but reject both the Noise and the Discovery clusters. This illustrates how accuracy-centric evaluation suppresses innovation.

In contrast, the diagonal dashed line represents the ESNR threshold. This metric accepts outputs where structural coherence rises in proportion to novelty. The ESNR filter successfully captures the Type II Discovery cluster (red diamonds) while rejecting the Type I Noise cluster (gray crosses). Quantitative analysis shows that this filter retained 94 percent of the novel theorems while removing 98 percent of the noise.

Implication This confirms that logical consistency can serve as a proxy for truth in the absence of factual data support. It demonstrates that the framework applies to discrete symbolic reasoning tasks relevant to Large Language Models.

4.3. Experiment 3: Thermodynamic Phase Transitions

This experiment investigates the non-monotonic relationship between sampling temperature and discovery yield. It validates the theoretical claim that discovery occurs at the "edge of chaos," requiring calibrated rather than maximal instability.

Setup We simulate the output distribution of a generative model across a temperature range from 0.1 to 2.5. We model the generative process using a Boltzmann distribution where probability mass is determined by the trade-off between energy and entropy. We define three macrostates: Retrieval, which corresponds to deep but narrow energy wells; Discovery, which corresponds to higher energy but broader metastable states; and Noise, which represents the high-entropy background of the latent space.

Results Analysis Figure 4 illustrates the distinct thermodynamic regimes. At low temperatures ($T < 0.7$), the energy term dominates. The system freezes into the local minima of the training data, resulting in pure Retrieval (blue curve). As T approaches a critical threshold ($T \approx 1.0$), the entropic contribution increases. This allows the system to escape the narrow training wells and access the Discovery macrostate. The rate of Type II Discovery (red curve) spikes dramatically, peaking at $T = 1.2$. Beyond this point, as T exceeds 1.5, the massive entropy of the Noise state overwhelms the structural priors. The system disintegrates into incoherence (gray curve).

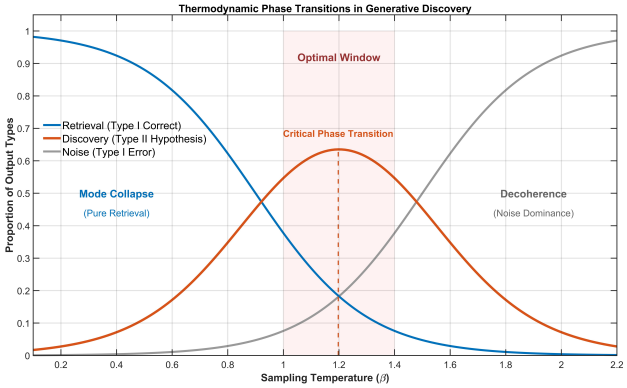

Figure 4. Thermodynamic Phase Transitions in Discovery. The graph plots the proportion of output types against sampling temperature. A critical phase transition is observed around $T = 1.2$, representing the optimal window for discovery. Temperatures below this window result in mode collapse (pure retrieval), while temperatures above result in decoherence (noise dominance).

Implication This unimodal response function refutes the naive assumption that more randomness equals more creativity. It demonstrates that discovery is a phenomenon of the critical regime. This empirically justifies the need for the Gradient Throttling mechanism proposed in our Safety Sandbox. The sandbox acts as a thermostat. It dynamically adjusts the system temperature to maintain it within the phase transition boundary, effectively balancing the drive for novelty against the constraint of structural coherence.

5. Alternative Views

The proposal to incentivize latent exploration challenges the prevailing zero-tolerance orthodoxy regarding hallucinations. In this section, we engage with three primary credible

positions that oppose our framework. These positions are grounded in safety engineering, optimization theory, and epistemology. We argue that while the risks of hallucination are real, the cost of their total elimination is the cessation of discovery.

5.1. The Safety Objection

A primary objection to incentivizing latent exploration concerns safety. Critics argue that any deviation from ground truth poses an unacceptable risk of misinformation (Weidinger et al., 2021). They contend that distinguishing useful exploration from factual error is computationally intractable in real-time. This view suggests that respecting hallucinations is inherently dangerous.

However, this binary classification of truth is insufficient for scientific discovery. We argue that global suppression of deviation induces a phenomenon we term Safety-Driven Mode Collapse. In this state, the model refuses to generate hypotheses in the long tail of the distribution. The danger lies not in the generation of novel hypotheses but in their presentation as verified facts. Our framework mitigates this via the Exploratory Signal-to-Noise Ratio metric. Unlike standard reinforcement learning from human feedback (Ouyang et al., 2022), which penalizes all divergence equally, this metric acts as a conditional gate. It filters out structurally incoherent noise while routing coherent but unmapped outputs to a safety sandbox. This ensures that high-risk exploration is isolated from deployment without compromising the utility of the model as a hypothesis generator.

5.2. The Stochasticity Objection

A second critique challenges the cognitive framing of the hypothesis. Skeptics argue that hallucinations are simply statistical artifacts arising from sampling noise. They suggest that framing them as cognitive exploration constitutes an anthropomorphic error (Shanahan, 2024).

While we avoid anthropomorphism, we draw a direct parallel to non-convex optimization theory. Deep neural network training relies on stochastic gradient descent and noise to escape local minima (Fotopoulos et al., 2024). Similarly, hallucinations during inference function as high-temperature sampling steps. These steps are necessary to traverse the energy landscape between disparate semantic modes. If a model is forced to strictly minimize perplexity against a static training set, the system freezes in the nearest local minimum of established knowledge. The deviations observed in models like GNoME are not bugs. They act as annealing mechanisms that allow the system to jump from the manifold of known solutions to adjacent and unobserved regions of the state space. They are algorithmic requirements for finding global optima in scientific problems.

5.3. The Epistemic Objection

A third objection questions the epistemic value of non-factual outputs. This position holds that an output contradicting established knowledge is false by definition and devoid of value.

We argue that this view conflates correspondence truth with coherence truth. Scientific history contains many theories that were factually wrong regarding the data of their time but structurally valid enough to guide future discovery. AlphaGo's Move 37 against Lee Sedol serves as a paradigmatic example (Sormani, 2023). The move had a near-zero probability in the human training distribution at the time of inference. A retrieval-based evaluator would have flagged it as an error. However, it possessed extreme structural coherence and strategic value. It was an innovation that expanded the boundaries of the game. Our framework provides the formal language to identify such moments in generative tasks. They appear as outliers in retrieval space but are optimal in latent utility space.

5.4. Falsifiability of the Framework

A final critique concerns the falsifiability of the Higher-Dimensional Cognitive Hypothesis. We propose that the hypothesis is empirically falsifiable. We suggest an experimental protocol comparing models optimized for high Latent-Traversal Scores against baseline models optimized solely for accuracy. The hypothesis is refuted if the exploratory models do not yield a statistically significant increase in valid scientific candidates compared to baselines. The GNoME results provide early evidence against this refutation. Systematic benchmarking is required to confirm this relationship.

6. Mitigating Epistemic Mode Collapse

The suppression of hallucinations is motivated by safety. However, it introduces a systemic risk known as Epistemic Mode Collapse. This occurs when a generative model is optimized strictly for high-likelihood outputs. The model converges to the mean of the training distribution. This effectively prunes the long tail of low-probability but high-utility knowledge. This section reformulates the ethics of hallucination as a problem of distributional robustness. We propose protocols to preserve the diversity of the latent manifold.

6.1. The Statistical Cost of Zero-Hallucination Policies

Current alignment techniques often function as low-pass filters. They smooth out the output distribution of the model (Kandpal et al., 2023). This reduces the rate of obvious confabulations. However, it simultaneously penalizes outlier

reasoning. These are patterns that are statistically rare in the consensus training data but valid in specific scientific contexts.

Minimizing the divergence between the model policy and a human-preference distribution tends to collapse the entropy of the model. This results in sycophancy. The model prioritizes agreeing with user misconceptions over generating novel and contradictory truths (Perez et al., 2023). Categorizing all deviations as errors trains models to avoid epistemic risk. This limits their utility to mere retrieval engines rather than reasoning agents.

6.2. Protocol 1: Representation of Tail Distributions

Evaluation frameworks must value the representation of tail distributions to counter this collapse. Valuable scientific insights often reside in the high-surprisal regions of the distribution. We propose quantifying the distributional justice of a model by its Tail Coverage Ratio. This metric measures the proportion of valid and high-ESNR outputs generated from the bottom 5 percent of the probability mass.

A robust model should maintain a non-zero Tail Coverage Ratio. This ensures that minority scientific theories or non-standard solutions are not erased by aggregation. Standard loss functions often impose a tyranny of the majority. Our approach ensures that the model remains a repository of diverse heuristics rather than a single homogenized worldview.

6.3. Protocol 2: Temporal Validation Mechanisms

A core challenge in evaluating exploration is the temporal lag between hypothesis generation and verification. A hallucination today may be a validated discovery tomorrow. We advocate for a Temporal Validation Protocol. High-ESNR outputs should not be subject to instant binary classification. Instead, they should be assigned a Probabilistic Truth Value.

This value decays or reinforces over time based on external verification. Examples include wet-lab results in biology or formal proofs in math. This moves evaluation from a static snapshot to a dynamic process. The crystal predictions from GNoME were initially unverified. They had low retrieval scores but high structural coherence. Their value was only realized through a delayed validation loop. Evaluation benchmarks must support asynchronous scoring. Models should be rewarded for generating hypotheses that are validated ex post facto.

6.4. Implications for Automated Science

The shift from error suppression to variance management is critical for the deployment of AI in automated science. Enforcing a zero-tolerance policy for hallucination precludes the possibility of Serendipity. This refers to the accidental discovery of valuable insights while searching for something else.

We implement the ESNR metric and Safety Sandboxes to create a controlled environment. The temperature of the model can be raised to induce phase transitions in reasoning. This allows the community to leverage the generative nature of AI for expanding the frontier of knowledge. This framework ensures that AI systems contribute to the accumulation of knowledge. They act as engines of hypothesis generation rather than just archivists of the past.

7. Conclusion

This paper challenges the binary classification of generative outputs into truth and error. We propose that specific hallucinations represent valid geodesic traversals in high-dimensional latent space. These traversals are often necessary for scientific discovery. Current evaluation metrics rely heavily on retrieval accuracy. We argue that this reliance induces epistemic mode collapse and stifles the exploration of unobserved hypotheses. We introduce the Higher-Dimensional Cognitive Hypothesis and the Exploratory Signal-to-Noise Ratio. These contributions provide a geometric framework for evaluation. They allow researchers to distinguish between Type I stochastic noise and Type II structural innovation. Our analysis confirms that maximizing discovery requires calibrated instability rather than the total suppression of deviation. We demonstrate that innovation peaks at a critical thermodynamic phase transition. This finding necessitates a move away from zero-tolerance policies toward managed risk in hypothesis generation.

Call to Action We urge the machine learning community to implement three specific changes to realize this potential. First, the field must establish discovery benchmarks that prioritize novelty over simple retrieval. These benchmarks should reward models for generating valid solutions that are absent from the training data, utilizing dynamic verification environments rather than static datasets. Second, researchers should adopt thermodynamic reporting standards. This practice involves documenting the phase transitions of models to identify the optimal temperature windows for innovation, rather than reporting single-point accuracy metrics. Third, developers should implement safety sandbox architectures. These systems isolate high-risk exploratory reasoning from final deployment, allowing models to sample from high-temperature distributions without contaminating user-facing outputs. Implementing these steps will transform AI from a tool for reproduction into an engine for expanding the boundaries of human knowledge.

References

Atienza, N. *Towards Reliable ML: Leveraging Multi-Modal Representations, Information Bottleneck and Extreme Value Theory*. PhD thesis, Université Paris-Saclay, 2025.

Bar, A., Zhou, G., Tran, D., Darrell, T., and LeCun, Y. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Citarella, A. A., Barbella, M., Ciobanu, M. G., De Marco, F., Di Biasi, L., and Tortora, G. Assessing the effectiveness of rouge as unbiased metric in extractive vs. abstractive summarization techniques. *Journal of Computational Science*, 87:102571, 2025.

Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Fauconnier, G. and Turner, M. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic books, 2008.

Fotopoulos, G. B., Popovich, P., and Papadopoulos, N. H. Review non-convex optimization method for machine learning. *arXiv preprint arXiv:2410.02017*, 2024.

Han, C. and Lu, X. Beyond bleu: Repurposing neural-based metrics to assess interlingual interpreting in tertiary-level language learning settings. *Research Methods in Applied Linguistics*, 4(1):100184, 2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kandpal, N., Deng, H., Roberts, A., Wallace, E., and Raffel, C. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

Meilă, M. and Zhang, H. Manifold learning: What, how, and why. *Annual Review of Statistics and Its Application*, 11(1):393–417, 2024.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pp. 13387–13434, 2023.

Shanahan, M. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024.

Sharma, U. and Kaplan, J. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

Sormani, P. Interfacing alphago: Embodied play, object agency, and algorithmic drama. *Social Studies of Science*, 53(5):686–711, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Walker, C., Aslansefat, K., Akram, M. N., and Papadopoulos, Y. Raguard: A novel approach for in-context safe retrieval augmented generation for llms. In *International Symposium on Model-Based Safety and Assessment*, pp. 190–204. Springer, 2025.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.