States Without Histories: Latent Reasoning as Admissibility-Field Traversal

Flyxion
Independent Research

June 2026

Abstract

Seddik and Fard (2026) demonstrate that latent thought representations in large language models preserve task-level identity while collapsing within-task instance identity. We argue this result is not merely a failure of representation design but evidence for a deeper ontological misalignment. Their four axioms—Causality, Minimality, Separability, Stability—are uniformly state-based: each takes a latent point as primitive. The empirical pattern they discover is scale-selective erasure of distinguishability: coarse-grained distinctions survive while fine-grained distinctions are integrated out. This is the signature of renormalization under optimization pressure. We organize the argument around four formal objects: an admissibility field $\mathcal{A}$, admissible trajectories τ, a distinguishability renormalization curve $R(\epsilon)$, and a critical scale ϵ_c. We prove a non-equivalence proposition showing that trajectory distinguishability can exceed state distinguishability, state the Distinguishability Renormalization Conjecture, and derive a Separability Collapse corollary that recovers the paper's empirical hierarchy as a theorem. Independent cross-domain context comes from Griffin et al. (2026), who find the identical two-level hierarchy—between-category discrimination preserved, within-category identity collapsed—in cortical EEG representations of faces in autistic children. This parallel is consistent with (though not decisive evidence for) the conjecture's claim that the hierarchy is not specific to gradient-based optimization. We propose three experiments that discriminate between four competing hypotheses about the source of the collapse.

1 Four Objects

The argument in this paper turns on four definitions, stated here at the outset so that every subsequent claim can be read against them.

Definition 1 (Admissibility field)**.** *An admissibility field is a set-valued map $\mathcal{A} : V \to 2^V$ where V is the vocabulary (a finite discrete set). Given a decoding distribution $P(\cdot \mid v)$ over next tokens and a threshold $\eta > 0$,*

$$\mathcal{A}(v) = \{\, v' \in V : P(v' \mid v) > \eta \,\}. \tag{1}$$

Because V is finite and $P(\cdot \mid v)$ is a proper probability mass function, this set is well-defined and non-empty for any $\eta < \max_{v'} P(v' \mid v)$. A token sequence $\tau = (v_0, v_1, \ldots, v_n)$ is admissible if $v_{t+1} \in \mathcal{A}(v_t)$ for every t. The admissibility field is determined by the input: given prompt p, the field $\mathcal{A}_p$ encodes which token transitions receive non-negligible probability mass under the model conditioned on p.

The paper concerns latent representations, not token sequences. The following proposition bridges the two levels, showing that admissibility equivalence defined over tokens induces a well-defined equivalence over latent states.

Proposition 1 (Latent induction)**.** *Let $T = f(v_{0:t})$ be a latent representation generated by encoding an admissible token prefix $v_{0:t}$ under $\mathcal{A}_p$. Define two latent states $T_i = f(v_{0:t}^{(i)})$ and $T_j = f(v_{0:t}^{(j)})$ as latent-admissibility-equivalent if their generating prefixes are admissibility-equivalent in token space: $v_{0:t}^{(i)} \sim_A^{\epsilon} v_{0:t}^{(j)}$ iff $D_A(v_{0:t}^{(i)}, v_{0:t}^{(j)}) < \epsilon$ where D_A compares the continuation distributions $P(\cdot \mid v_{0:t}^{(i)})$ and $P(\cdot \mid v_{0:t}^{(j)})$. This relation is well-defined on latent states: if f is deterministic, then two prefixes that are token-admissibility-equivalent induce latent states with admissibility-equivalent continuation distributions, and the continuation-distribution distances in Section 6 operationalize this comparison directly in the latent embedding space without requiring a probability density over latent states.*

This proposition justifies treating the subsequent definitions—admissibility equivalence, the renormalization curve, and the critical scale—as properties of latent states even though the formal admissibility field is defined over tokens. The theory does not switch substrates; it operates at the token level and lifts to the latent level through the encoding function f.

Definition 2 (Admissibility equivalence)**.** *Two latent states T_i and T_j are admissibility-equivalent at resolution ϵ, written $T_i \sim_A^{\epsilon} T_j$, if $D_A(T_i, T_j) < \epsilon$, where D_A is the continuation-distribution distance defined in Section 6. The recoverable distinction volume at scale ϵ is*

$$\delta_R^{(\epsilon)}(T) = \mathrm{vol}\big\{\, T' : T' \sim_A^{\epsilon} T \,\big\}^{-1}. \tag{2}$$

Definition 3 (Distinguishability renormalization curve)**.** *The renormalization curve is the function $R : (0, \infty) \to [0, \infty)$ defined by $R(\epsilon) = \delta_R^{(\epsilon)}$. The renormalization operator R_λ acts on R by rescaling the resolution parameter:*

$$R_\lambda : R(\epsilon) \mapsto R(\lambda\epsilon), \qquad \lambda > 1. \tag{3}$$

Applying R_λ coarsens the resolution by factor λ, merging equivalence classes separated at scale ϵ but not at scale $\lambda\epsilon$.

Definition 4 (Critical scale)**.** *The critical scale ϵ_c is the infimum of resolutions at which distinguishability survives:*

$$\epsilon_c = \inf\{\, \epsilon : R(\epsilon) > 0 \,\}. \tag{4}$$

Distinctions at scale $\epsilon < \epsilon_c$ are not recoverable; distinctions at $\epsilon > \epsilon_c$ are.

With these four objects in hand, the paper's empirical claim is that optimization under finite representational capacity acts as iterated application of R_λ, driving $R(\epsilon) \to 0$ below ϵ_c while preserving $R(\epsilon) > 0$ above it. The following sections develop this claim formally, connect it to the empirical observations of Seddik and Fard, and propose experiments that test it.

2 The Ontological Misalignment

Every one of the four axioms of Seddik and Fard takes a latent point as its primitive. Causality asks whether a point T can substitute for a token prefix in the model's computational graph. Minimality asks for mutual information $I(X; T)$ between input and a point. Separability asks for distance $d(\phi(T_1), \phi(T_2))$ between two points. Stability asks for proximity $T \approx T'$ between two points. The implicit ontology is that a thought is a location. This is not necessarily wrong—the state ontology remains viable, and H2 in Section 5 is its strongest remaining form—but it is one possible ontology, and the empirical evidence does not force it.

If reasoning is instead a traversal through an admissibility field, the relevant quantity is not any instantaneous state but a functional over the path. Define the admissibility cost of a transition as $c(v_t, v_{t+1}) \geq 0$, with $c(v_t, v_{t+1}) = 0$ if $v_{t+1} \in \mathcal{A}(v_t)$ and positive otherwise. The action of a trajectory is

$$S[\tau] = \sum_{t=0}^{n-1} c(v_t, v_{t+1}). \tag{5}$$

A reasoning process is admissibility-optimal when $S[\tau] = 0$. This connects to $R(\epsilon)$ through the following observation: optimization pressure minimizes expected action at the scales

where action cost is statistically consistent across training instances. At scales where $\mathbb{E}[c(v_t, v_{t+1})]$ is low and consistent, the optimization learns to stay within $\mathcal{A}$, preserving distinguishability at that scale. At scales where the expected cost is high or variable—as instance-level distinctions typically are across a diverse training corpus—the optimization does not reliably learn to distinguish them, and $R(\epsilon) \to 0$ below the corresponding scale. Variance specifically, rather than mean cost, determines the threshold because high variance implies inconsistent gradient direction across instances: the model receives conflicting signals about whether to preserve a distinction, reducing the expected benefit of dedicating representational capacity to it, and the distinction is therefore integrated out. The action functional therefore provides a mechanism connecting optimization dynamics to the shape of $R(\epsilon)$: the critical scale ϵ_c is approximately the resolution at which the variance of $S[\tau]$ across training instances crosses the representational capacity threshold.

The state-based axioms are attempting to recover $S[\tau]$ from a single observation $\tau(t)$. For path-dependent processes this cannot succeed in general. Knowing a latent state does not determine the reasoning trajectory, just as knowing a particle's position does not determine its action. The paper is attempting to reconstruct a path functional from a point sample.

Why the input embedding result is expected. If the input prompt p fully determines the admissibility field $\mathcal{A}_p$, then the prompt embedding already encodes the structural constraints on all admissible trajectories. A downstream state $\tau(t)$ encodes in addition the specific position reached along one such trajectory, but this positional information need not add distinguishability at the instance scale if instance-scale distinctions have been integrated out by optimization. The prompt embedding and any downstream state therefore carry approximately the same information about the admissibility field. The paper finds, consistently across all five models, that no candidate thought representation outperforms the input embedding on every evaluation axis. Under the trajectory hypothesis this is not a surprise. The prompt specifies the field; traversal determines which region of the field is reached. These are different operations, and the former does not require the latter to encode additional distinguishing information in any single state.

Why latent thinking degrades with step count. Latent Thinking begins at one step with high task purity but loses purity as step count increases. A state-based view predicts that more computation should add information. A trajectory view predicts that each additional step moves the representation further along a path whose instance-discriminating content is distributed across the path rather than concentrated at any point. Measuring a later state recovers less about the instance because the traversal has moved further from the prompt-anchored starting conditions without aggregating trajectory history into any single vector.

Under the action functional, the longer the path, the more $\tau(t)$ reflects global trajectory optimization rather than its boundary conditions at $\tau(0)$. The erasure of instance identity at later steps is consistent with what occurs when a system is driven along a low-dimensional optimization manifold: task-level structure, which is stable along the manifold, is preserved, while instance-level sensitivity to initial conditions is lost as the trajectory moves away from the prompt-anchored starting point.

3 State–Trajectory Non-Equivalence

The trajectory hypothesis claims that trajectory distinguishability can exceed state distinguishability. This is not merely plausible—it follows from the structure of the admissibility field.

Proposition 2 (State–Trajectory Non-Equivalence)**.** *Let τ_1 and τ_2 be admissible trajectories under $\mathcal{A}_p$ with the same terminal state: $\tau_1(n) = \tau_2(n) = T^*$. Then in general there exists no function $f : X \to (X^*)^n$ such that $f(T^*)$ recovers either τ_1 or τ_2. Consequently, the map $\tau \mapsto \tau(n)$ is not injective on admissible trajectories, and trajectory distinguishability—the ability to discriminate τ_1 from τ_2—can be strictly greater than state distinguishability at T^*.*

Proof. The admissibility field $\mathcal{A}_p$ is in general many-to-one from trajectories to terminal states: multiple admissible trajectories may reach the same terminal state T^* by different paths. Let $f : X \to Y$ be any function mapping latent states to some range Y. Because $\tau_1(n) = \tau_2(n) = T^*$ by hypothesis, $f(T^*)$ takes a single value regardless of which trajectory is being considered. Therefore $f(T^*) = f(\tau_1(n)) = f(\tau_2(n))$, and no function of the terminal state alone can separate τ_1 from τ_2. The trajectory pair is distinguishable (they differ in their intermediate states) while the terminal state is not (it is shared). □ □

This proposition provides the formal bridge between the renormalization framework and the trajectory recovery prediction. It establishes a necessary condition: trajectory information *can* exceed state information whenever the admissibility field is many-to-one from trajectories to terminal states. The proposition does not establish that language-model reasoning actually occupies this regime—that is the empirical question Experiment 1 addresses. The claim is that if reasoning is path-dependent in the relevant sense, the state-sampling methodology of Seddik and Fard will systematically miss it; and if it is not, the trajectory experiment will confirm the state ontology by failing to recover additional instance identity from ordered sequences.

4 The Distinguishability Renormalization Conjecture

Conjecture 1 (Distinguishability renormalization)**.** *For optimization procedures minimizing predictive loss under finite representational capacity, the distinguishability renormalization curve $R(\epsilon)$ satisfies*

$$R(\epsilon) \to 0 \quad (\epsilon < \epsilon_c), \qquad R(\epsilon) > 0 \quad (\epsilon > \epsilon_c), \tag{6}$$

where the critical scale ϵ_c is determined by the relationship between the variance of $S[\tau]$ across training instances and the representational capacity of the model: distinctions whose admissibility cost is consistent across instances (low variance, high contribution to gradient signal) survive; those whose cost is inconsistent (high variance, weak or noisy gradient) are integrated out. Specifically, ϵ_c is predicted to increase with model capacity pressure (smaller models or larger training diversity) and decrease as training exposure to fine-grained distinctions becomes more consistent (more data at the relevant scale, or curricula that specifically reward instance-level discrimination). The shape of $R(\epsilon)$ above ϵ_c is an empirical question; a power-law recovery $R(\epsilon) \propto (\epsilon - \epsilon_c)^\alpha$ for $\epsilon > \epsilon_c$ is a natural parametric form consistent with second-order phase transitions in related statistical mechanical systems, but we make no commitment to a specific functional form here.

The empirical claim of Seddik and Fard, restated in this language: instance-scale distinctions lie below ϵ_c (same-task pairs are admissibility-equivalent), while task-scale distinctions lie above ϵ_c (cross-task pairs are not). This follows immediately as a corollary.

Corollary 1 (Separability collapse under renormalization)**.** *Suppose $\epsilon_{\text{instance}} < \epsilon_c < \epsilon_{\text{task}}$. Then $\delta_R^{(\epsilon_{\text{instance}})} \to 0$ while $\delta_R^{(\epsilon_{\text{task}})} > 0$. Same-task discrimination therefore converges to chance while cross-task discrimination remains positive.*

Proof. Direct from Conjecture 1 and the definitions. Since $\epsilon_{\text{instance}} < \epsilon_c$, the conjecture gives $R(\epsilon_{\text{instance}}) \to 0$, meaning the admissibility equivalence class of any state T at that resolution contains nearly all states in the neighborhood, and no probe can recover instance identity. Since $\epsilon_{\text{task}} > \epsilon_c$, $R(\epsilon_{\text{task}}) > 0$, meaning task-scale equivalence classes remain small enough that a probe can discriminate them. □ □

Corollary 1 recovers the paper's primary empirical finding as a consequence of the conjecture rather than as a standalone observation. It also unifies the four patterns in D.7: participation ratio rising while purity falls (the representation expands into directions below ϵ_c); same-task collapse (instance scale below ϵ_c); cross-task survival (task scale above ϵ_c); prompt-embedding competitiveness (the prompt encodes $\mathcal{A}_p$, and ϵ_c is a property of the optimization, not of whether one has traversed the field).

Remark 1 (Renormalization equivalence)**.** *The renormalization curve $R(\epsilon)$ suggests a new representational invariant. Two representations T_1 and T_2 are renormalization-equivalent, written*

$T_1 \sim_R T_2$, if $R_1(\epsilon) = R_2(\epsilon)$ for all $\epsilon > 0$. Under this equivalence, what matters about a representation is not its geometry in latent space but its scale-response function: how distinguishability varies with the resolution at which it is measured. This is potentially a more fundamental object than latent-space distance, since two representations can be arbitrarily far apart geometrically while being renormalization-equivalent, and conversely can be nearby while having radically different scale-response profiles. The conjecture predicts that all representations produced by the same optimization procedure on the same training distribution will be approximately renormalization-equivalent, sharing the same ϵ_c regardless of their specific geometric arrangement.

5 Competing Hypotheses

The trajectory experiment proposed in Section 9 is informative only if it discriminates between distinct hypotheses. We identify four.

H1: State encoding. Instance identity is encoded in each latent state T_t individually and is recoverable in principle but not by the linear probes used in the paper. The paper's ablation in D.8 increases discriminator capacity by an order of magnitude with no recovery of within-task discrimination. This disfavors H1 under the specific probe hypothesis classes tested, but does not eliminate it: H1 remains viable if instance identity is encoded in a form accessible only to probes well outside the evaluated class. We treat H1 as disfavored but not eliminated by the existing evidence.

H2: Distributed state encoding. Instance identity is encoded in a highly nonlinear, distributed form across the geometry of a single state, accessible in principle but not to any probe in the paper's hypothesis class. A transformer's final hidden vector can in principle encode arbitrary functions of the input history; what D.8 establishes is that no probe in the evaluated hypothesis class recovers instance identity from it, even at an order-of-magnitude capacity increase. H2 cannot be ruled out by D.8 alone—the encoding might exist in a form no bounded probe can access. However, H2 and H4 make the same prediction at the single-state level and are therefore empirically indistinguishable there. The trajectory experiment discriminates them: if sequence-modeled trajectories recover instance identity where single states do not, the encoding is temporal rather than geometrically distributed within any one state.

H3: Ensemble averaging. Aggregating multiple states along the trajectory recovers instance identity simply by averaging out noise, not because the trajectory carries path-dependent information. H3 predicts that unordered aggregation (pooling without regard

to sequence order) recovers instance identity as well as ordered aggregation (sequence modeling). The experiment discriminates H3 from H4 by comparing an order-permuted trajectory to the original: if scrambling the order destroys recovery, the encoding is genuinely path-dependent.

H4: Trajectory encoding. Instance identity is encoded in the temporal structure of the trajectory and is not recoverable from any single state or unordered aggregate. This is what the trajectory hypothesis predicts. Confirmation requires that (a) sequence-modeled trajectories outperform single states, and (b) order matters (H3 is ruled out), and (c) the gain is not explained by capacity alone (the discriminator is capacity-matched).

6 Continuation-Distribution Geometry

The admissibility equivalence relation requires a distance D_A between latent states defined through their future distributions rather than their coordinates. We specify three formulations in increasing order of theoretical cleanliness.

Hard Hausdorff (baseline, not recommended). Draw K continuations from $P(\cdot \mid T)$ to form $\mathcal{A}_K(T) = \{\mathrm{Emb}(y_i)\}$ and set $D_A^H(T, T') = d_H(\mathcal{A}_K(T), \mathcal{A}_K(T'))$. At $K = 8$, beam search is mode-collapsing and path-dependent; two states with overlapping distributions may yield disjoint beam sets, artificially inflating D_A^H. This formulation is listed for completeness and as the baseline against which the refinements are measured.

Soft-Hausdorff (practical). Weight each point by its generation probability:

$$D_A^{\mathrm{soft}}(T, T') = \max\Big(\sup_{y \in \mathcal{A}_K(T)} \min_{y' \in \mathcal{A}_K(T')} p(y) \cdot d(y, y'),\ \sup_{y' \in \mathcal{A}_K(T')} \min_{y \in \mathcal{A}_K(T)} p(y') \cdot d(y', y) \Big). \quad (7)$$

This is a divergence rather than a metric; it does not satisfy symmetry or the triangle inequality. For the purposes of defining the admissibility equivalence relation this is sufficient: $T_i \sim_A^{\epsilon} T_j$ iff $D_A^{\mathrm{soft}}(T_i, T_j) < \epsilon$ and $D_A^{\mathrm{soft}}(T_j, T_i) < \epsilon$.

Nucleus Wasserstein (theoretically preferred).

$$D_A^W(T, T') = W_1\big(P_N(\cdot \mid T),\ P_N(\cdot \mid T')\big) \quad (8)$$

where $P_N(\cdot \mid T)$ is the next-step token distribution restricted to the top-N tokens by cumulative probability mass (e.g. the minimal set covering $p = 0.95$ of the mass, rather

than a fixed N), and the ground metric is cosine distance between token embeddings. The nucleus restriction is not merely computational. Tokens outside the high-probability nucleus contribute negligible mass to the admissibility field: an admissible continuation is by definition one the model assigns non-trivial probability, and low-probability tokens add noise to the optimal-transport plan without altering which futures are genuinely reachable. D_A^W is a proper metric and makes the admissibility equivalence classes well-defined sets. It requires vocabulary-space distributions cached during decoding, which source models compute as a standard intermediate product.

Both D_A^{soft} and D_A^W instantiate the admissibility equivalence relation and ask not whether latent states are nearby as points but whether the futures they generate are distinguishable as distributions. The within-task collapse, re-expressed in this language, manifests as $D_A(T_i, T_j) \approx 0$ for two distinct problems from the same task: the instances have been merged not just in representation but in their admissible futures.

7 Distinguishability at Scale

Conjecture 1 predicts a renormalization curve with a specific shape: $R(\epsilon) \approx 0$ below ϵ_c and $R(\epsilon) > 0$ above it. Monotone recovery above ϵ_c is a separate, stronger claim and is stated as an empirical prediction rather than part of the conjecture.

The continuous operationalization uses prompt-embedding cosine distance as a proxy for the semantic scale of a distinction. This proxy is an estimator, not the scale itself: prompt distance $\hat{\epsilon} = d_{\text{prompt}}(x_i, x_j)$ correlates with the distinguishability scale ϵ at which the two instances differ, but does not equal it. For each pair of same-task problems, compute $\hat{\epsilon}$ and plot within-task discriminator accuracy as a function of $\hat{\epsilon}$. The prediction is a flat chance-level region for small $\hat{\epsilon}$ followed by monotone increase, with the transition locating ϵ_c empirically. The slope above the transition quantifies how rapidly distinguishability recovers as distinctions become coarser.

8 Cross-Domain Evidence: Cortical Renormalization in Autism

The Distinguishability Renormalization Conjecture is stated for optimization procedures minimizing predictive loss under finite representational capacity. If the conjecture describes a general property of any capacity-limited system exposed to training signal with scale-dependent statistical consistency, it should produce the same hierarchical collapse pattern in biological neural systems as in trained language models. Griffin et al. (2026) report findings consistent with this prediction, from a completely different measurement modality and population.

Using multivariate pattern analysis on high-density EEG from 399 children aged 6–11, Griffin et al. find that autistic children show preserved between-category discrimination (faces versus houses) alongside absent within-category discrimination (face identity) across all ages and all time windows. Neurotypical (NT) children show both levels of discrimination, with within-category accuracy emerging specifically in the 10–11-year age group. The same two-level hierarchy that Seddik and Fard observe in language model latent representations—coarse distinctions preserved, fine distinctions collapsed—appears in cortical representations of faces in autism. The measurement instrument is EEG decoding accuracy rather than latent-vector discriminator accuracy, the substrate is cortical activity rather than transformer hidden states, and the training signal is developmental experience with faces rather than gradient descent on next-token prediction. Multiple mechanisms could in principle generate such a hierarchy—attentional differences, developmental delay, sensory specialization—and we do not claim to adjudicate between them here. The relevance of Griffin et al. to the present argument is more limited: it shows that the two-level hierarchy is not specific to LLMs or to gradient-based optimization, which is consistent with (though not decisive evidence for) the conjecture's claim to generality.

Developmental trajectory as renormalization curve. The conjecture predicts that ϵ_c is determined by the statistical consistency of training signal at each scale. In biological systems, experience plays the role of the optimization process: consistent fine-grained input drives ϵ_c downward, making progressively finer distinctions recoverable. Griffin et al. find that NT children show increasing neural specialization across age groups, with stimulus-level and identity-level decoding accuracy rising from the 6–7 to the 10–11 cohort. Autistic children show no corresponding developmental increase. Under the conjecture, this dissociation means that experience-dependent refinement shifts ϵ_c downward in NT children—finer-scale distinctions become recoverable as the cortical system accumulates face-specific exposure—while ϵ_c remains elevated in autistic children because reduced social attention deprives the system of the consistent fine-grained signal that would drive renormalization toward finer scales. The developmental trajectory is the renormalization curve evolving over time, and the group dissociation locates the critical scale: NT children's ϵ_c crosses the identity scale between ages 8–9 and 10–11, while autistic children's ϵ_c does not.

Temporal generalization as trajectory evidence. The temporal generalization matrices in Griffin et al. show a strong diagonal decoding pattern for face-selective representations in both groups: a classifier trained at time t predicts best at time t and degrades at other times, indicating that neural representations are sequentially evolving rather than stable states.

This is consistent with the hypothesis that face-selective information is trajectory-native in the cortical system—distributed across the temporal sequence of processing rather than concentrated in any instantaneous neural state. Autistic children show reduced decoding along the diagonal specifically at later timepoints (200–450 ms), which is the temporal analog of the latent-thinking degradation with step count observed in Seddik and Fard: as the trajectory proceeds further from the initial stimulus representation, the instance-discriminating content becomes less recoverable from any single temporal snapshot.

By contrast, identity-selective representations in autistic children show a more stable "square-like" temporal generalization pattern, interpreted by the authors as reflecting a more persistent and less dynamic encoding strategy. Under the trajectory framework, this stability indicates that the system is not navigating an admissibility field with fine-grained identity structure: representations stabilize precisely because there is no fine-scale distinction structure to traverse. The trajectory has collapsed to a fixed point rather than evolving through a structured field.

Suggestive scope. The parallel across domains is consistent with a generalization of the conjecture beyond gradient-based optimization, though it falls short of establishing one. The two-level hierarchy appears in language model latent representations (Seddik and Fard) and in cortical face representations in autism (Griffin et al.), under different substrates, measurement modalities, and forms of training pressure. This suggests that the hierarchy is not a specific artifact of transformer training, and that ϵ_c may be a property of any capacity-limited system exposed to scale-uneven signal. The claim remains a conjecture: confirming it in the LLM setting requires Experiments 1–3; extending it to biological systems would require a separate research program. We present the cross-domain parallel as motivating context rather than confirmatory evidence.

9 Proposed Experiments

Experiment 1: Trajectory discrimination (H4 vs. H1/H2/H3). For iterative thinking candidates at step count n, train three discriminators on the same within-task classification task: (a) a single-state probe on T_t with capacity C; (b) an attention-pooling readhead over $(T_0, \ldots, T_n)$ with parameter count matched to C; (c) a small causal transformer or LSTM over $(T_0, \ldots, T_n)$ with parameter count matched to C. Additionally train (d) the sequence model of (c) applied to a permutation-scrambled trajectory $(T_{\sigma(0)}, \ldots, T_{\sigma(n)})$ for a fixed random permutation σ.

The capacity-matching in all conditions ensures that accuracy gains reflect information content, not probe capacity. The four conditions discriminate the hypotheses as follows.

If (c) outperforms (a), the encoding is temporal rather than purely state-based (rules out H1, puts pressure on H2). If (c) outperforms (b), ordered sequence structure matters beyond unordered aggregation (rules out H3). If (d) underperforms (c), the order of the trajectory—not merely the set of states—carries the discriminating information (confirms H4 path-dependence).

Experiment 2: Continuation-distribution distance. For each same-task pair (x_i, x_j) in the test split, compute $D_A^{\text{soft}}(T_i, T_j)$ using the existing Nemotron-embedded beam outputs and their log-probabilities. Correlate D_A^{soft} with ground-truth answer divergence (whether the two problems require different correct answers) and with same-task discriminator accuracy. If D_A^{soft} predicts answer divergence better than $d(T_i, T_j)$ on the same pairs, continuation-distribution geometry carries information that point geometry does not, and the admissibility equivalence relation is empirically grounded.

Experiment 3: Renormalization curve. For each pair of same-task problems, compute prompt-embedding cosine distance $\hat{\epsilon}$. Bin pairs by $\hat{\epsilon}$ and plot within-task discriminator accuracy against $\hat{\epsilon}$. A flat chance-level region below a threshold followed by monotone increase is the renormalization curve predicted by Conjecture 1. The threshold estimates ϵ_c; the slope above it quantifies the steepness of the renormalization.

10 What the Four Axioms Achieve

The axiomatic framework of Seddik and Fard is a genuine contribution. Consistency, independence, and completeness are established relative to a well-defined state ontology. The measures are computable without retraining and decoupled from downstream accuracy.

The limitation is not internal to the axiom system. It is complete relative to a state ontology of thought. If the remedy for the observed failures is better state compression, the paper's framework points in the right direction. If the remedy is trajectory-level evaluation, the framework is evaluating the wrong object. Which agenda to pursue depends on whether Conjecture 1 is confirmed, whether trajectory recovery (Experiment 1) discriminates in favor of H4, and whether continuation-distribution geometry (Experiment 2) outperforms point geometry. These are empirically open questions.

11 Conclusion

The scale-selective pattern of distinguishability collapse in Seddik and Fard—task-level preservation, instance-level erasure—is the signature of renormalization under optimization

pressure, not simply inadequate representation design. We have organized this argument around four formal objects ($\mathcal{A}$, τ, $R(\epsilon)$, ϵ_c), proved a non-equivalence proposition showing that trajectory distinguishability can exceed state distinguishability, stated the Distinguishability Renormalization Conjecture, derived the paper's empirical hierarchy as Proposition 1, and proposed experiments that discriminate four competing hypotheses.

The cross-domain parallel with Griffin et al. (2026) is consistent with the conjecture extending beyond gradient-based optimization, though it does not establish this. Seddik and Fard provide the LLM evidence; Griffin et al. provide a parallel pattern in cortical face processing under developmental experience rather than gradient descent. The same two-level hierarchy appears in both, under different substrates, populations, and measurement modalities. If the conjecture were entirely an artifact of transformer architecture, it would not be expected to appear in cortical EEG decoding of pediatric face processing. That it does appear there is consistent with the hypothesis that ϵ_c may be a property of capacity-limited representational systems more generally—but the mechanisms may differ, and the parallel is offered as motivating context rather than confirmation.

The deepest claim of the manuscript is not that latent thought representations fail. It is that thought may not be representable as a state variable at the scale being measured. If ϵ_c lies above the instance scale, no state-based representation—however well-designed—can recover instance identity, because the training process has integrated it out. This is not a deficiency of current methods. It is a structural consequence of what any capacity-limited system does when fine-grained signal is statistically inconsistent relative to coarse-grained signal. Correcting it requires not better latent vectors but evaluation at the right object: the admissible trajectory and the scale-dependent distinguishability it preserves.

References

[1] Seddik, F. and Fard, F. (2026). Formalizing latent thoughts: Four axioms of thought representation in LLMs. *arXiv preprint arXiv:2606.27378*.

[2] Griffin, J. W., Gerber, A. H., Litson, K., Faja, S., Jeste, S., Kleinhans, N., Dawson, G., Naples, A., Levin, A. R., Webb, S. J., Shic, F., Sugar, C., Dziura, J., and McPartland, J. C. (2026). Decoding the temporal dynamics of face-specific neural representations in autism. *Nature Mental Health*. `https://doi.org/10.1038/s44220-026-00672-y`

[3] Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., and Tian, Y. (2024). Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

[4] Zhang, Z., He, X., Yan, W., Shen, A., Zhao, C., and Wang, X. E. (2026). Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space. In *Proceedings of NeurIPS 2026*.

[5] Wu, J., Lu, J., Ren, Z., Hu, G., Wu, Z., Dai, D., and Wu, H. (2026). LLMs are single-threaded reasoners: Demystifying the working mechanism of soft thinking. In *Proceedings of ICLR 2026*.

[6] Kazemi, M., Fatemi, B., Bansal, H., Palowitch, J., Anastasiou, C., Mehta, S. V., Jain, L. K., Aglietti, V., Jindal, D., Chen, P., Dikkala, N., Tyen, G., Liu, X., Shalit, U., Chiappa, S., Olszewska, K., Tay, Y., Tran, V. Q., Le, Q. V., and Firat, O. (2025). BIG-bench extra hard. In *Proceedings of ACL 2025*.

[7] Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

[8] King, J.-R. and Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, 18(4):203–210.

[9] Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

[10] Barber, D. and Agakov, F. (2003). The IM algorithm: a variational approach to information maximization. In *Advances in Neural Information Processing Systems*, volume 16.

[11] Park, K., Choe, Y. J., and Veitch, V. (2024). The linear representation hypothesis and the geometry of large language models. In *Proceedings of ICML 2024*.

[12] Sahoo, S., Chadha, A., Jain, V., and Chaudhary, D. (2026). When shallow wins: Silent failures and the depth-accuracy paradox in latent reasoning. In *LIT Workshop, ICLR 2026*.

[13] Zou, J., Xiong, Y., and Liu, Y. (2026). The theoretical benefits and limitations of latent chain-of-thought reasoning. *arXiv preprint*.

[14] Ethayarajh, K. (2019). How contextual are contextualized word representations? In *Proceedings of EMNLP 2019*.

[15] Wilson, K. G. (1971). Renormalization group and critical phenomena. *Physical Review B*, 4(9):3174–3183.

[16] Scherf, K. S., Behrmann, M., Humphreys, K., and Luna, B. (2007). Visual category-selectivity for faces, places and objects emerges along different developmental trajectories. *Developmental Science*, 10(4):F15–F30.

[17] Golarai, G., Ghahremani, D. G., Whitfield-Gabrieli, S., Reiss, A., Eberhardt, J. L., Gabrieli, J. D. E., and Grill-Spector, K. (2007). Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Neuroscience*, 10(4):512–522.