The Geometry of Reasoning Failure: Detecting Scale Boundaries in LLM Latent Trajectories

Andrey Ustyuzhanin [1 2 3] **Maxim Borisyak** [1 2] **Nikita Kazeev** [3]

[1]*Constructor University, Bremen, Germany* [2]*Constructor Knowledge Labs, Bremen, Germany* [3]*Institute for Functional Intelligent Materials, NUS, Singapore*. Correspondence to: Andrey Ustyuzhanin a.ustyuzhanin@constructor.university.

Abstract

Large language models sustain fluent scientific discourse yet routinely fail when a problem demands a shift in abstraction—from high-level theory selection to low-level computation, or vice versa. These breakdowns mirror a Kuhnian paradigm crisis: anomalies accumulate while the system lacks any internal signal that its current reasoning frame has reached a limit. We show that such crises leave measurable geometric signatures in hidden-state space. Without fine-tuning or architectural changes, we probe frozen activations to decode the current abstraction level (macro/meso/micro) and track latent trajectories during generation, demonstrating that scale boundaries manifest as curvature spikes, probe discontinuities, and topological compression.

1. Introduction

Autonomous AI scientist systems promise to accelerate hypothesis generation and experimental design, yet their reliability remains constrained by fragile long-horizon reasoning and inadequate self-diagnosis [1]. In practice, LLM-based agents frequently persist in a failing mode of thought, elaborating within an inappropriate level of detail and producing locally coherent but globally invalid reasoning. Consider an AI Scientist agent tasked with designing a catalyst: it may fluently generate density-functional-theory parameters while the real bottleneck is a meso-scale transport limitation it never recognizes. These failures cluster at *scale boundaries*—points where the task demands switching abstraction level: macro (theory/strategy), meso (effective models/constraints), or micro (implementation/calculation). In human scientific practice such moments are informative: they signal that a representational frame has become inadequate, motivating paradigm revision rather than further incremental derivation.

Today, detection of these failures is overwhelmingly post hoc. Text-level indicators arrive late; by the time an error is linguistically visible, the latent trajectory may already have diverged irreversibly. Intervention methods such as activation steering, meanwhile, assume one already knows *when* to steer and *toward what* [2]. What is missing is an *internal, real-time diagnostic* that can flag an approaching scale boundary before the reasoning derails.

We frame the problem geometrically. If abstraction level is encoded in the hidden-state manifold, then coherent reasoning corresponds to a trajectory with locally stable directionality, and failure at a scale boundary should appear as a breakdown in that stability. This view aligns with recent geodesic perspectives on LLM generation, where successful inference follows approximately linear paths and error accumulation introduces stochastic deviation [3]. From a multiscale viewpoint, causal emergence theory argues that macro-descriptions carry unique causal power, implying that switching scales is often *necessary* for robust explanations [4]. Complementary biological perspectives cast intelligence as competent navigation across problem spaces; failures correspond to losing the ability to remap coordinate systems when regimes shift [5].

Research question. Are scale boundaries during LLM scientific reasoning detectable as geometric signatures in latent space—specifically as trajectory curvature spikes and abstraction-level discontinuities? In particular we are interested in studying these hypotheses:

- **H1 (Linear decodability):** A linear probe trained on frozen hidden states can classify abstraction level {macro, meso, micro} with >85% F1.
- **H2 (Boundary signature):** When reasoning crosses a scale boundary, latent trajectories exhibit (a) loss of local linearity and (b) discontinuous jumps in abstraction probe output; complementary signals include topological compression and local intrinsic dimensionality (LID) spikes.

2. Method

We instrument open-weight transformer LLMs (Llama-3, Mistral, Qwen; 7B–70B) and record hidden states during step-by-step generation. Crucially, we make **no architectural changes** and perform **no fine-tuning**: all learning is confined to lightweight probes trained on frozen activations.

Abstraction probe (H1). We train a linear classifier P_{SCG} on hidden states h_t^ℓ annotated for abstraction level. The annotation rubric is domain-specific—for math/program reasoning: macro = planning, meso = constraints/sub-goals, micro = code execution; for scientific QA: macro = claims/theory, meso = methodology, micro = evidence/data. We sweep layers to identify where abstraction is most linearly separable and test cross-dataset generalization. As an alternative geometry, we project hidden states into Poincaré space to evaluate whether hyperbolic embeddings better capture hierarchical structure.

Trajectory geometry (H2). For each reasoning

chain, we extract hidden states $\{h_1, \ldots, h_T\}$ at the probe's best layer and compute two per-step signals:

$$C_t = \cos(h_t - h_{t-1},\ h_{t-1} - h_{t-2}) \quad \text{(directional consistency)} \tag{1}$$

$$J_t = \|P_{\text{SCG}}(h_t) - P_{\text{SCG}}(h_{t-1})\| \quad \text{(abstraction jump rate)} \tag{2}$$

High C_t signals smooth geodesic progression; drops mark curvature spikes. Large J_t signals abrupt abstraction shifts. We test whether low C_t and high J_t co-locate with annotated boundary tokens.

Complementary topological signals. To disambiguate scale-specific signatures from generic uncertainty, we apply persistent homology over sliding windows of hidden states, tracking the evolution of Betti-1 (loop structure). We observe "topological compression"—a drop in topological diversity at boundaries as the representation collapses before reorganizing. We additionally estimate local intrinsic dimensionality (LID), which exhibits transient spikes when the model remaps between abstraction subspaces. Both analyses are applied post hoc on collected trajectories rather than at inference time.

Boundary detection metric. We define a composite score $B_t = \alpha(1 - C_t) + \beta J_t$ tuned on validation splits, and an enriched variant B'_t incorporating Betti-1 drop rate and LID magnitude. Evaluation: AUC, precision/recall at fixed false-positive rates, and early-warning lead time (tokens before known boundary).

Datasets. We select tasks with clear, derivable boundary points: **MuSiQue**, where composable multi-hop questions provide discrete abstraction transitions [6]; **GSM8K-PoT**, where each solution contains a sharp plan→code boundary; **REPLACE** [7], which supplies ground-truth state abstraction levels (Q^*-irrelevant, π^*-irrelevant) directly suitable for training and evaluating H1; and **QASPER** for cross-domain scientific QA generalization [8].

Positioning. Our contribution differs from: (i) factual probing—we target *structural abstraction state*, not content; (ii) CoT evaluation—our signals arise from hidden states and can precede textual error; (iii) static TDA—we track topology *over generation time*; (iv) representation engineering—steering modifies behavior, whereas we provide the prerequisite *detection* signal [2].

3. Results and Implications

Across model families (7B–70B) quantitative values vary but qualitative patterns hold consistently:

1. **Abstraction is linearly decodable (H1).** Probes achieve >85% F1 at middle-to-late layers, peaking 2–4 layers before the output layer. Poincaré probes outperform Euclidean probes on deeply nested hierarchies.

2. **Scale boundaries induce geometric degradation (H2).** At annotated boundaries, C_t drops sharply and J_t spikes. The composite B_t detects boundaries with AUC > 0.80; the topology-enriched B'_t further improves recall.

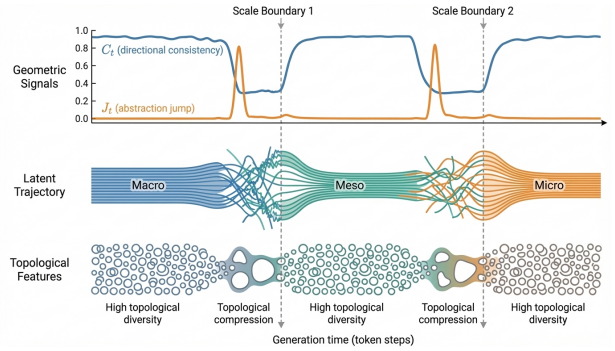

Fig. 1: Three complementary views of a latent trajectory crossing two scale boundaries (macro→meso→micro). *Top:* Geometric signals—directional consistency C_t (blue) drops and abstraction jump rate J_t (orange) spikes at each boundary. *Middle:* The hidden-state trajectory frays and kinks at boundary zones before reconverging. *Bottom:* Persistent homology reveals high Betti-1 diversity during coherent reasoning, collapsing to topological compression at transitions.

3. **Topological signatures co-locate with boundaries.** Betti-1 drops and LID spikes at scale transitions serve as architecture-agnostic confirmation, more generalizable than absolute probe scores.

Broader impact. Recent work on fully autonomous AI scientists [1] demonstrates that LLM-based agents can generate and test hypotheses end-to-end, yet their failure modes remain opaque—an agent may waste compute cycles refining micro-level parameters when the macro-level hypothesis is flawed, with no internal signal to trigger re-framing. Our boundary detection metric supplies precisely this missing signal: a real-time diagnostic that tells an AI Scientist *when* to abandon "normal science" and initiate a paradigm shift. More broadly, the framework delivers a measurement layer for any system that must navigate multiscale explanations rather than collapsing into a single level of detail [4, 5, 9].

References

[1] C. Lu, C. Lu, R.T. Lange, J. Foerster, J. Clune, and D. Ha. The AI scientist: Towards fully automated open-ended scientific discovery. In *ICLR 2025*, 2024. arXiv:2408.06292.

[2] A. Zou, L. Phan, S. Chen, J. Campbell, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint*, 2023. arXiv:2310.01405.

[3] H. Huang, Y. LeCun, and R. Balestriero. Semantic tube prediction: Beating LLM data efficiency with JEPA. *arXiv preprint*, 2026. arXiv:2602.22617.

[4] E. Hoel. Causal emergence 2.0: Quantifying emergent complexity. *arXiv preprint*, 2025. arXiv:2503.13395.

[5] C. Fields and M. Levin. Competency in navigating arbitrary spaces as an invariant for analyzing cognition in diverse embodiments. *Entropy*, 24(6):819, 2022.

[6] H. Trivedi et al. MuSiQue: Multihop questions via single-hop question composition. *TACL*, 10:539–554, 2022.

[7] Z. Li, Y. Cao, and J.C.K. Cheung. Do LLMs build world representations? probing through the lens of state abstraction. In *NeurIPS 2024*, 2024. OpenReview:lzfzjYuWgY.

[8] P. Dasigi, K. Lo, I. Beltagy, A. Cohan, N.A. Smith, and M. Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *NAACL-HLT 2021*, pages 4599–4610, 2021. arXiv:2105.03011.

[9] Y. LeCun. A path towards autonomous machine intelligence. Technical report, OpenReview, 2022. Version 0.9.2, OpenReview:BZ5a1r-kVsf.