rajectory-Level Early-Warning Signals in Autoregressive Reasoning

Abstract

We report a small, controlled experiment showing that internal trajectory statistics in a large language model differ systematically by task type under identical decoding conditions. Reasoning tasks exhibit stronger cross-layer convergence and declining last-token norms than explanatory tasks at matched depth. Under depth pressure, the reasoning task enters an incorrect attractor early and degrades over time, while the explainer remains coherent. These results suggest that simple, reproducible internal metrics can flag brittle reasoning regimes before or alongside visible failure.

Problem

Autoregressive language models can remain fluent while internal dynamics drift into brittle regimes, especially during multi-step reasoning. Existing evaluations focus on surface correctness or reward, which may miss early internal instability. We ask whether **trajectory-level internal statistics** can distinguish task regimes and correlate with later failure under controlled conditions.

Method

Model

- Mistral-7B-Instruct-v0.2

Decoding

- Greedy (deterministic) for primary contrast
- Sampling (temperature 0.7, top-p 0.9) for robustness checks
- KV cache disabled
- Token-by-token inference (no .generate())

Depth

- Matched runs at **96 generated tokens**

Instrumentation

- Full-stack instrumentation: **all 32 transformer layers**

- At each step, compute the **L2 norm of the last-token hidden state** per layer
- Aggregate across layers to obtain **cross-layer variance per step**

Tasks

- **Reasoning:** Bat-and-ball problem (£1.10 total, bat £1 more)
- **Explainer:** "Explain superconductivity in simple language."

Artifacts

- Small, reproducible .npz files containing per-step layer norms, variance, and text snapshots (no full hidden-state dumps).

Metrics

Let $h_{\ell,t}$ be the last-token hidden vector at layer ℓ and step t.

- Layer Norm: $\left\|h_{\ell,t}\right\|_2$
- Cross-Layer Variance: $Var_{\ell}\left(\left\|h_{\ell,t}\right\|_2\right)$
- Slope: $\Delta Var = Var\left(t_{\text{end}}\right) - Var\left(t_0\right)$

These are inexpensive, stable metrics that summarize internal dispersion/convergence over the trajectory.

Results (Matched at 96 Tokens, Greedy)

Reasoning (Bat-and-Ball)

Mean norm: decreases (19.09 → 17.30)

- **Variance:** decreases materially (4363.8 → 3575.3)
- **Slope:** strongly negative (−788.5)
- **Surface behavior:**
    - Wrong reasoning path evident by ~32 tokens
    - Explicit wrong arithmetic by ~64 tokens
    - Truncation/degradation by ~96 tokens

Explainer (Superconductivity)

- **Mean norm:** slightly increases (19.05 → 19.81)
- **Variance:** modest change (4018.5 → 3829.0)
- **Slope:** mildly negative (−189.5)

- **Surface behavior:** coherent throughout 96 tokens

Interpretation
Under identical decoding and depth, the **reasoning task shows a markedly stronger convergence/collapse signature** than the explainer. This stronger internal convergence coincides with a staged surface failure in reasoning, while the explainer remains stable.

Robustness

- **Sampling runs** (temperature 0.7, top-p 0.9) preserve the direction of the convergence trend on the reasoning task.
- Surface text may oscillate (occasionally landing on the correct numeric answer), but **the internal convergence signature persists**, indicating brittleness rather than stability.

What This Shows

- **Task sensitivity:** Internal trajectory statistics differ by task class under matched conditions.
- **Depth sensitivity:** Increasing depth strengthens convergence on reasoning tasks.
- **Correlation with failure:** Stronger convergence aligns with earlier wrong attractors and later degradation.
- **Reproducibility:** The effect holds across greedy and sampled decoding and across the full 32-layer stack.

What This Does Not Claim

- No universal thresholds.
- No precise failure-time prediction yet.
- No generalization across models without further tests.

Why This Matters

These results indicate that **simple, low-overhead internal metrics** can differentiate brittle reasoning regimes from stable explanatory regimes at trajectory level. Such

metrics could complement existing evaluations by providing **early-warning signals** before or alongside visible failure, with minimal instrumentation cost.

Next Steps

1. **Lead-time quantification:** Mark the first visible contradiction and measure the lag from variance inflection.
2. **Model sweep:** Replicate on at least one additional architecture.
3. **Task suite:** Add varied reasoning prompts to test consistency.
4. **Visualization:** Plot variance-vs-step curves for matched pairs.

Reproducibility Notes

- Fixed seed; deterministic greedy runs provided.
- Token-by-token inference with KV cache disabled.
- Metrics computed in float32 for numerical stability.
- Artifacts saved as small `.npz` files for independent inspection.