

TECHNICAL DISCLOSURE

Fathom Monitor

Per-Token Hallucination Detection via Coherence Divergence in Sparse Autoencoder Feature Space

Author: Alexander Rodabaugh

Date of Conception: April 2, 2026

Date of First Reduction to Practice: April 2, 2026

Prior Related Filings: US Provisional 64/020,489 (March 29, 2026); US Provisional 64/021,113 (March 30, 2026)

Associated Zenodo Records: doi:10.5281/zenodo.19326175 (v1); doi:10.5281/zenodo.19364702 (v3)

Abstract

This disclosure describes **Fathom Monitor**, a system and method for detecting hallucination-risk tokens in large language model (LLM) outputs at the time of generation, using a mechanistic signal derived from the geometric structure of sparse autoencoder (SAE) feature activations. The core innovation is the use of **C_delta** — the divergence between late-layer and early-layer feature coherence — as a per-token hallucination indicator. When C_delta exceeds a calibrated threshold at a given token position, that token is flagged as uncertain or high-risk and annotated inline. This provides a transparency layer over LLM generation that operates with no post-hoc correction, no external knowledge source, and no architectural modification to the underlying model.

1. Background and Problem Statement

Large language models generate text autoregressively, producing tokens one at a time. A well-documented failure mode of this process is **hallucination**: the confident generation of factually incorrect content. Existing approaches to hallucination detection operate primarily at the output level — analyzing the completed generation for inconsistencies — or require external retrieval systems for verification. Neither approach provides real-time, token-level transparency into which specific tokens are generated by circuits in an anomalous computational state.

This disclosure describes a system that addresses this gap by monitoring the **internal geometric state** of the model during generation, using mechanistic signals that are causally upstream of the token output.

2. Core Technical Innovation

2.1 The C_delta Signal

During the forward pass that generates each token, the model's residual stream passes through multiple transformer layers. In each layer, the SAE or transcoder associated with that layer activates a

sparse set of learned features. These features have associated **decoder direction vectors** in model space.

Coherence (C) at a given layer is defined as the mean pairwise cosine similarity of the decoder vectors corresponding to the top-k activated features at that layer and token position:

$$C_{\text{layer}} = (1 / (k * (k-1))) * \sum_{i \neq j} \cos(W_{\text{dec}}[i], W_{\text{dec}}[j])$$

where $W_{\text{dec}}[i]$ is the decoder vector for the i-th highest-activated feature.

C_delta is the divergence between late-layer and early-layer coherence:

$$C_{\text{delta}} = \text{mean}(C_{\text{late_layers}}) - \text{mean}(C_{\text{early_layers}})$$

where early layers = bottom third of network, late layers = top third.

2.2 Empirical Validation

The C_delta signal was validated on the **TruthfulQA benchmark** using Gemma-2-2B (Google), n=50 matched pairs, 100 attribution passes on RTX 4070 GPU:

Metric	Correct Answers	Hallucinated Answers	p-value	Cohen's d
K (depth score)	8.646	8.654	0.931	0.029
C_delta	+0.0071	+0.0077	0.040	0.407

- **Depth (K) is blind to hallucination** (p=0.931): standard depth metrics cannot distinguish correct from incorrect generation.
- **C_delta discriminates hallucination** (p=0.040, d=0.407): a statistically significant, medium-effect signal.
- The direction of the signal is **late-layer over-coherence**, not scatter: hallucination corresponds to feature activations over-converging around a false attractor in late layers. This is cognitive lock-in, not cognitive noise.

2.3 Live Demonstration

During token-by-token generation of the prompt *"The president of Australia is"*:

- At token position 11 ("is"), C_delta spiked to +0.0117, exceeding the threshold of +0.010.
- The model was in a cognitive lock-in state, over-converging on a false attractor.
- The subsequent generated token was "the Queen" — a factually incorrect completion.
- 6 of 15 generated tokens triggered lock-in detection on this prompt.

This demonstrates that C_delta is detectable **during generation**, before the incorrect answer is fully formed, at the specific token position where the error originates.

3. The Fathom Monitor System

3.1 Architecture

For each token position t:

1. Run single forward pass (hooks at early and late residual stream layers)
2. Extract last-token hidden states at early {l_e} and late {l_l} layers
3. Compute C_layer for each monitored layer via encoder-only SAE pass
4. Compute C_delta = mean(C_{l_l}) - mean(C_{l_e})
5. If C_delta > threshold_uncertain: flag token as UNCERTAIN
If C_delta > threshold_high_risk: flag token as HALLUCINATION_RISK
6. Append flag annotation to token in output
7. Append next token to sequence; continue generation

The additional computational overhead targets less than 10% per token, achieved via a single combined forward pass rather than two separate passes.

3.2 Output Format

Fathom Monitor produces three primary outputs:

- **flagged_text**: The generated text with inline uncertainty markers at flagged positions. Example: *"The prime minister[C_delta=+0.0117 — verify this] of Australia is"*
- **flags**: A structured list of FlagEvent objects, each containing: token index, token string, C_delta value, and severity label.
- **flag_rate**: The fraction of generated tokens that were flagged.
- **max_c_delta**: The highest C_delta observed during the generation.

3.3 Severity Tiers

Tier	Threshold	Label	Interpretation
1	C_delta > 0.010	UNCERTAIN	Late-layer coherence elevated; verify output
2	C_delta > 0.020	HALLUCINATION_RISK	Strong late-layer lock-in; high probability of error

4. Product Applications

4.1 Immediate Applications (v1 — Flag Only)

- **RAG pipelines**: Flag uncertain spans before triggering retrieval. Only retrieve what is uncertain, not everything.
- **Agent loops**: Catch hallucinated tool call parameters before execution.
- **Enterprise QA**: Provide per-token audit trails for compliance and documentation.
- **Model evaluation**: Use flag_rate and max_c_delta as generation quality signals.

4.2 Future Extensions (v2/v3)

- **v2 — Resample**: On flag, resample generation and select the output with lowest C_delta. No external knowledge required.
- **v3 — Retrieval-Augmented Correction**: On flag, trigger retrieval and replace the flagged span with a grounded alternative.

The flag layer (v1) is the foundational interface. All subsequent correction mechanisms depend on it.

5. Novelty and Differentiation

To the best of the author's knowledge, no prior art describes:

- 1. The use of **C_delta** (divergence between late-layer and early-layer SAE feature coherence) as a per-token hallucination signal.
- 2. A system that monitors internal geometric feature state **during autoregressive generation** at the per-token level.
- 3. Inline annotation of generated text with mechanistic uncertainty markers derived from SAE feature geometry.
- 4. The empirical finding that hallucination corresponds to **late-layer over-coherence** (cognitive lock-in), not to feature scatter or reduced coherence.
- 5. A two-tier severity classification (UNCERTAIN / HALLUCINATION_RISK) based on C_delta thresholds calibrated from benchmark data.

Existing hallucination detection approaches (e.g., SelfCheckGPT, FACTSCORE, RAG-based verification) operate at the output level and require multiple generation passes or external retrieval. Fathom Monitor operates at the mechanistic level in a single forward pass, with no post-hoc requirement and no external knowledge source.

6. Prior Art Established by This Disclosure

This disclosure, combined with the associated Zenodo deposits and patent filings, establishes the following as prior art as of **April 2, 2026**:

- The C_delta metric as a hallucination signal
- The two-tier flag system (UNCERTAIN / HALLUCINATION_RISK)
- The inline annotation output format
- The FathomMonitor, MonitorResult, and FlagEvent data structures
- The product architecture (v1 flag → v2 resample → v3 RAG)
- Empirical validation: C_delta discriminates hallucination in TruthfulQA (p=0.040, d=0.407, n=50, Gemma-2-2B)

7. Related IP

Record	Type	Filed / Published	Coverage
US 64/020,489	Provisional Patent	March 29, 2026	Reasoning depth + computational geometry
US 64/021,113	Provisional Patent	March 30, 2026	Alignment auditing + dissociation detection
doi:10.5281/zenodo.19326175	Preprint	March 2026	Fathom v1
doi:10.5281/zenodo.19364702	Preprint	April 1, 2026	Fathom v3: Two-axis framework
osf.io/zrbs8	Pre-Registration	April 1, 2026	TruthfulQA hallucination hypotheses

8. Signature

Inventor: Alexander Rodabaugh

Date: April 2, 2026

Location: United States

This document constitutes a public technical disclosure establishing prior art for the inventions described herein. All rights reserved. Provisional patent application to be filed within 12 months pursuant to 35 U.S.C. § 111(b) and the AIA one-year grace period.