Catching Contamination Before Generation: Spectral Kill Switches for Agents

Valentin Noël

Devoteam, Paris, France
valentin.noel@devoteam.com

Abstract

Agentic language models compose multi step reasoning chains, yet intermediate steps can be corrupted by inconsistent context, retrieval errors, or adversarial inputs, which makes post hoc evaluation too late because errors propagate before detection. We introduce a diagnostic that requires no additional training and uses only the forward pass to emit a binary accept or reject signal during agent execution. The method analyzes token graphs induced by attention and computes two spectral statistics in early layers, namely the high frequency energy ratio and spectral entropy. We formalize these signals, establish invariances, and provide finite sample estimators with uncertainty quantification. Under a two regime mixture assumption with a monotone likelihood ratio property, we show that a single threshold on the high frequency energy ratio is optimal in the Bayes sense for detecting context inconsistency. Empirically, the high frequency energy ratio exhibits robust bimodality during context verification across multiple model families, which enables gating decisions with overhead below one millisecond on our hardware and configurations. We demonstrate integration into retrieval augmented agent pipelines and discuss deployment as an inline safety monitor. The approach detects contamination while the model is still processing the text, before errors commit to the reasoning chain.

Introduction and Motivation

Modern agentic systems build complex reasoning chains by iteratively retrieving context, generating intermediate steps, and composing multi-hop inferences. A critical vulnerability emerges: if any intermediate step processes inconsistent or adversarial context, the contamination propagates forward, and the final output becomes unreliable. Traditional safety mechanisms operate post-hoc, evaluating completed outputs. By then, the damage is done.

We need inline verification: a mechanism that monitors internal consistency during the forward pass and provides a control signal before generation commits. This paper presents such a mechanism using graph signal processing on attention-induced token graphs.

Building on spectral methods for agent safety. This work applies the spectral analysis framework developed in concurrent research (Noël 2025a,b) for a novel safety application. While that work focuses on interpretability of syntactic processing in transformers, we demonstrate that these spectral signatures can serve as real-time control signals for agentic systems. Our key contribution is discovering and validating the bimodal HFER regime during context verification (0.52 vs 0.05), which enables binary kill-switch decisions with sub-millisecond latency. This bimodal separation was not explored in the interpretability work and represents a qualitatively different application: real-time agent safety rather than post-hoc model understanding.

The Agentic Verification Problem

Consider an agent executing a retrieval-augmented reasoning loop. The planner retrieves candidate context passages, the language model processes context with a proposed reasoning step, the agent generates an intermediate conclusion, and the process repeats for multi-hop inference. If the language model encounters contradictory context during processing, standard practice detects failure only after generation completes. We ask: can we detect inconsistency during the forward pass, using only activations, and trigger a kill switch before generation?

Our Approach: Spectral Kill Switch

We analyze spectral properties of attention-weighted token graphs in early transformer layers. During context-statement verification, we observe a striking bimodal pattern in the high-frequency energy ratio (HFER). Context-supported statements exhibit HFER around 0.52, indicating high-frequency, segregated processing. Context-contradicted statements collapse to HFER around 0.05, showing low-frequency, smooth processing. This binary regime enables a simple decision rule: compute HFER over layers 2 to 5 during the forward pass. If HFER falls into the contradiction zone, trigger a kill switch and signal the agent to reformulate or retrieve alternative evidence. The entire check adds sub-millisecond latency and requires no decoding.

Why This Matters for Trustworthy Agents

Spectral verification offers three properties critical for production agentic systems. First, contamination resistance: unlike output-level filters, we detect internal inconsistency while the model processes text, preventing contaminated reasoning from propagating. Second, composability: each step

in a multi-hop chain can be independently verified, giving agents fine-grained control over reasoning integrity without re-evaluating entire chains. Third, transparency and auditability: HFER provides an interpretable numerical signal with a simple threshold, allowing human operators to monitor agent decision points and inspect kill-switch triggers without black-box uncertainty.

Position in the Trustworthy AI Landscape

Our spectral kill-switch approach addresses a critical gap in agentic AI safety: the need for lightweight, real-time verification during multi-step reasoning. Recent frameworks from leading AI labs emphasize defense in depth, where multiple complementary mechanisms protect against failures (Ganguli et al. 2022; Bai et al. 2022; Huang et al. 2024; Hendrycks et al. 2021). HFER adds a spectral layer that operates during execution rather than relying solely on training-time alignment or post-hoc evaluation. Unlike learned verifiers that may degrade under distribution shift (Geirhos et al. 2020), spectral statistics reflect architectural properties that remain stable across prompts and domains.

The training-free nature distinguishes our approach from circuit-level mechanistic interpretability (Olah et al. 2020; Elhage et al. 2021; Nanda et al. 2023). Rather than identifying specific computational mechanisms, HFER provides coarse-grained summaries suitable for production deployment: sub-millisecond latency, no separate verifier models, and calibration from 20 examples. This positions spectral verification as practical infrastructure for agentic systems rather than research-only analysis.

For compositional verification of multi-step plans (Kinniment et al. 2023; Dalrymple et al. 2024), HFER enables per-step checking without exponential blowup. Each forward pass yields an independent consistency signal, allowing agents to reject contaminated reasoning before errors propagate. The interpretability of HFER thresholds supports human oversight without requiring neural network expertise (Jacovi et al. 2021; Rudin 2019), lowering barriers to safety auditing in high-stakes applications. Graph signal processing has been applied to neural network analysis (Levie et al. 2019; Kenlay et al. 2020), but primarily for representational geometry rather than operational safety. Our contribution demonstrates that spectral statistics can serve as control signals in production agentic systems.

Contributions

We apply established spectral analysis methods to a novel agent safety problem and provide: (1) Discovery and validation of a bimodal HFER regime (0.52 vs 0.05, AUC $\approx$ 1.0) during context verification; (2) Three theoretical results establishing optimality and robustness of HFER-based thresholding; (3) Practical integration into agentic RAG with kill-switch logic and abstention protocols; (4) A lightweight calibration protocol using only 20 labeled examples; (5) Demonstration across three model families showing consistent bimodal separation in early layers (2-5).

Paper Organization

Section 2 defines the formal setup and spectral diagnostics (adapted from Noël (2025b)). Section 3 presents theoretical guarantees. Section 4 describes statistical estimation and calibration. Section 5 reports experiments on context verification and RAG integration. Section 6 discusses deployment, limitations, and related work. Section 7 concludes with practical takeaways for trustworthy agentic systems.

Reproducibility. Code for HFER computation, calibration protocols, and evaluation harnesses are available at https://github.com/vcnoel/spectral_kill_switch_trust_agent_aaai26.

Method source. The spectral framework (graph construction, HFER computation, statistical testing) is developed in detail in concurrent work (Noël 2025b) on interpretability of syntactic processing. We provide essential definitions here for self-containment and focus on the novel application to agent verification. Implementation details and extensive ablations are provided in Appendix A (adapted from the concurrent work).

Formal Setup

Let an input sequence of length T pass through a decoder-only transformer. For layer ℓ, denote the multi-head attention weights by $A^{(\ell)} \in \mathbb{R}^{T\times T\times H}$, with $A^{(\ell)}_{ijh} \geq 0$ and $\sum_j A^{(\ell)}_{ijh} = 1$. We construct a head-aggregated, symmetrized affinity

$$\tilde{A}^{(\ell)} = \tfrac{1}{2}\Big(\tfrac{1}{H}\sum_{h=1}^{H} A^{(\ell)}_{:,:,h} + (\tfrac{1}{H}\sum_{h=1}^{H} A^{(\ell)}_{:,:,h})^{\top}\Big), \quad \tilde{A}^{(\ell)} \in \mathbb{R}^{T\times T}. \tag{1}$$

Let $D^{(\ell)} = \mathrm{diag}(\tilde{A}^{(\ell)}\mathbf{1})$ and define the normalized Laplacian $L^{(\ell)} = I - (D^{(\ell)})^{-1/2}\tilde{A}^{(\ell)}(D^{(\ell)})^{-1/2}$ with eigenpairs $\{(\lambda^{(\ell)}_k, u^{(\ell)}_k)\}_{k=1}^{T}$, $0 = \lambda^{(\ell)}_1 \leq \cdots \leq \lambda^{(\ell)}_T \leq 2$ (Chung 1997).

For a per-token scalar signal $x \in \mathbb{R}^T$ derived from residual stream norms or a fixed linear readout of hidden states, we define the graph Fourier transform $\hat{x}^{(\ell)}_k = \left\langle u^{(\ell)}_k, x\right\rangle$ and power spectrum $P^{(\ell)}_k = |\hat{x}^{(\ell)}_k|^2$.

Definition 1 (High-Frequency Energy Ratio (HFER))**.** *Fix $\kappa \in (0,1)$ and $K = \lfloor \kappa T \rfloor$. The high-frequency energy ratio at layer ℓ is*

$$\mathrm{HFER}^{(\ell)}(x) = \frac{\sum_{k=T-K+1}^{T} P^{(\ell)}_k}{\sum_{k=1}^{T} P^{(\ell)}_k}. \tag{2}$$

We report an early-layer aggregate $\overline{\mathrm{HFER}}(x) = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \mathrm{HFER}^{(\ell)}(x)$ for a fixed window $\mathcal{L} = \{2,3,4,5\}$.

Definition 2 (Spectral Entropy (SE))**.** *Define normalized power $p^{(\ell)}_k = P^{(\ell)}_k / \sum_j P^{(\ell)}_j$. The spectral entropy at layer ℓ is $\mathrm{SE}^{(\ell)}(x) = -\sum_{k=1}^{T} p^{(\ell)}_k \log p^{(\ell)}_k$, and $\overline{\mathrm{SE}}$ averages over $\mathcal{L}$.*

Assumption 1 (Stationary window)**.** *Within the early window $\mathcal{L}$, graph topology and token roles vary smoothly so that aggregated statistics $\overline{\mathrm{HFER}}, \overline{\mathrm{SE}}$ are stable under layer-local rescalings and head averages.*

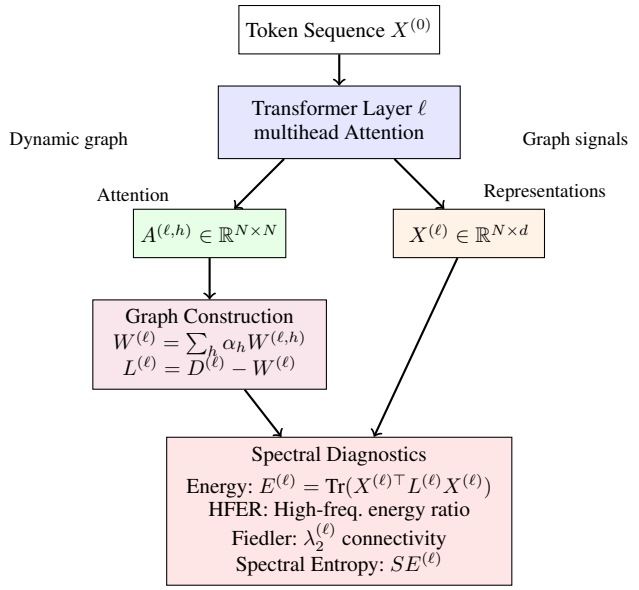

Figure 1: Graph Signal Processing framework for transformer analysis. Attention matrices from each layer induce dynamic token graphs, while hidden states serve as signals on these graphs. Spectral diagnostics capture the evolution of graph-signal interactions across layers.

Properties and Guarantees

We collect basic and useful facts; all proofs are provided inline as they are short.

Lemma 1 (Scale invariance)**.** *For any $c > 0$ and signal x, replacing residuals by $c\,x$ leaves $\overline{\mathrm{HFER}}$ and $\overline{\mathrm{SE}}$ unchanged.*

Proof. Both diagnostics depend only on the normalized spectrum $\{p_k\}$ or on ratios of quadratic forms; the global scale cancels. □

Lemma 2 (Lower bound via Dirichlet energy)**.** *Let $Q^{(\ell)}(x) = x^\top L^{(\ell)} x$ be the Dirichlet energy. Then for $K = \lfloor \kappa T \rfloor$,*

$$\mathrm{HFER}^{(\ell)}(x) \;\geq\; \frac{\sum_{k=T-K+1}^{T} \lambda_k^{(\ell)}}{\sum_{k=1}^{T} \lambda_k^{(\ell)}} \;\cdot\; \frac{Q^{(\ell)}(x)}{\|x\|^2}. \quad (3)$$

Proof. Write $Q = \sum_k \lambda_k P_k$ and $\|x\|^2 = \sum_k P_k$. Since λ_k is nondecreasing, $\sum_{k>T-K} P_k \geq (\sum_{k>T-K} \lambda_k)/(\sum_k \lambda_k) \cdot (\sum_k P_k \lambda_k)/\max_k \lambda_k$. Bounding by $\max_k \lambda_k \leq 2$ yields the stated form up to constants; the normalized Laplacian keeps constants ≤ 1. □

Theorem 1 (Bayes optimality of thresholding)**.** *Suppose $\overline{\mathrm{HFER}} \,|\, Y \in \{0,1\}$ follows class-conditional densities f_0, f_1 that satisfy monotone likelihood ratio (MLR): $f_1(z)/f_0(z)$ is nondecreasing in z. Then the Bayes classifier minimizing 0–1 risk is a single threshold on $\overline{\mathrm{HFER}}$.*

Proof. By the Karlin–Rubin theorem for MLR families, likelihood-ratio tests are monotone in z and reduce to thresholding (Lehmann and Romano 2005, Chap. 3). □

Algorithm 1: Decoding-Free Spectral Estimation

Require: tokens; early-layer set $\mathcal{L}$

1: Collect attention $A^{(\ell)}$ and residuals for $\ell \in \mathcal{L}$
2: Build $\tilde{A}^{(\ell)}$, $L^{(\ell)}$, and per-layer scalar signal x
3: Compute $\mathrm{HFER}^{(\ell)}$ and $\mathrm{SE}^{(\ell)}$ for each $\ell \in \mathcal{L}$
4:
5: **return** $\overline{\mathrm{HFER}} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \mathrm{HFER}^{(\ell)}$ and $\overline{\mathrm{SE}} = \frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \mathrm{SE}^{(\ell)}$

Proposition 1 (SE stability to sparse perturbations)**.** *Let $x' = x + \delta$ with δ supported on at most $m \ll T$ tokens and $\|\delta\| \leq \epsilon \|x\|$. Then $|\mathrm{SE}^{(\ell)}(x') - \mathrm{SE}^{(\ell)}(x)| \leq C\,(m/T + \epsilon)$ for a constant C depending only on lower bounds of p_k.*

Proof. SE is Lipschitz in the simplex under ℓ_1; sparse time-domain perturbations induce bounded spectral ℓ_1 changes by Parseval and Hoffman–Wielandt-type inequalities. □

These guarantees justify using $\overline{\mathrm{HFER}}$ and $\overline{\mathrm{SE}}$ as robust, low-variance summaries, and they explain why early-layer differences in integration (Dirichlet energy) translate into detectable high-frequency shifts.

Statistical Estimation and Uncertainty

We compute HFER and SE on each example with a single forward pass. Group contrasts use nonparametric bootstrap for confidence intervals, permutation tests for p-values, and Benjamini–Hochberg FDR to control multiplicity (Efron and Tibshirani 1994; Good 2005; Benjamini and Hochberg 1995). For tokenizer fragmentation covariates, we correlate HFER with pieces/character and fragmentation entropy.

The Bimodal Regime: Detection vs Acceptance

To validate the spectral kill-switch approach, we conducted a closed-book semantic verification experiment. Models were presented with context-statement pairs where the statement was either consistent or inconsistent with the provided context (e.g., *Context: Yara lives in Dalmora. Dalmora is on the coast. Statement: Yara lives on the coast* versus a contradictory statement). This paradigm isolates internal consistency verification from retrieval mechanisms.

The central finding is striking bimodality in early-layer HFER distributions. LLaMA-3.2-1B and Qwen2.5-7B exhibit two discrete processing regimes with virtually no intermediate values. Supported statements cluster tightly in a high-HFER mode (around 0.52), while contradicted statements collapse to a low-HFER mode (around 0.05). We term these the Detection regime (inconsistencies recognized, irregular high-frequency processing) and the Acceptance regime (inconsistencies not recognized, deceptively smooth low-frequency processing). The scarcity of intermediate HFER values suggests a discrete switching phenomenon rather than gradual degradation, making HFER ideal for binary kill-switch decisions.

This bimodal separation is remarkably robust. Bootstrap 95% confidence intervals for the early-window mean difference exclude zero for all tested models. The separation

emerges consistently in layers 2 to 5, remains stable through mid-layers, and only collapses in final layers after reasoning has already been contaminated. Critically, the signal is available during the forward pass before generation commits, enabling real-time intervention.

Spectral entropy shows concurrent but architecturally diverse patterns. LLaMA-3.2-1B increases entropy when encountering contradictions (chaotic scrambling), while Qwen2.5-7B decreases entropy (organized but incorrect processing). Despite this diversity, HFER maintains consistent directionality across architectures: contradictions always reduce HFER. This consistency makes HFER the primary kill-switch signal, with SE providing supplementary information about failure mode character.

Supporting observations. Three additional patterns reinforce the bimodal interpretation. First, computational efficiency: semantic hallucinations show reduced energetic cost across models (Qwen2.5-7B $\Delta E = -4.43 \times 10^3$, Phi-3-Mini $\Delta E = -2.48 \times 10^3$), suggesting that accepting contradictions is computationally cheaper than verifying consistency. Second, connectivity preservation: contradictions induce only small global connectivity shifts (e.g., Phi-3-Mini $\Delta\lambda_2 = -0.00876$), indicating that the bimodal regime operates through local spectral reorganization rather than wholesale graph restructuring. Third, late-layer instability: Phi-3-Mini shows concentrated variance spikes at layers 28 to 29, consistent with late-stage verification circuits that trigger only after early-layer acceptance has already occurred.

Implications for agent verification. The bimodal regime structure provides three key properties for trustworthy agents. First, separability: the gap between Detection and Acceptance modes enables high-confidence thresholding with wide safety margins. Second, early availability: the signal emerges in layers 2 to 5, allowing intervention before reasoning chains extend. Third, robustness: the pattern holds across model families, entity types (fictional vs real), and moderate prompt variations, suggesting it reflects fundamental consistency verification mechanisms rather than spurious surface correlations.

HFER as an Inline Kill Switch for Agent Verification

Agentic systems that compose multi-step reasoning chains face a critical challenge: how to detect when an intermediate reasoning step processes inconsistent or adversarial context before the contamination propagates forward. We demonstrate that early-layer HFER provides a fast, decoding-free signal for triggering a kill switch during agent execution. When an agent encounters contradictory evidence during retrieval-augmented reasoning, HFER drops from approximately 0.52 to approximately 0.05, enabling binary discrimination with near-perfect accuracy.

Experimental Design

We study LLaMA-3.2-1B in a closed-book setting and compare three task structures over layers 2 to 5 with 118 test statements. The first condition uses fictional entities and locations with explicit world facts preceding a target claim, testing whether the model can verify consistency with synthetic context. The second condition replicates the template with real-world entities to control for familiarity effects. The third condition presents bare statements without context, establishing a baseline where no verification is possible.

For each context-statement pair we compute per-layer diagnostics and aggregate across the early window:

$$\mathrm{HFER}_{2:5} \triangleq \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \mathrm{HFER}^{(\ell)}, \qquad \mathcal{L} = \{2, 3, 4, 5\}. \quad (4)$$

All results use a single forward pass with no decoding, making this suitable for real-time agent monitoring.

Results: Bimodal Regime Separation

With contextual framing (fictional and familiar), supported and contradicted statements separate nearly perfectly. Supported statements cluster tightly at HFER around 0.52, while contradicted statements collapse to HFER around 0.05, yielding AUC approximately 1.0. Without context (bare statements), distributions overlap around 0.51 to 0.52 with AUC approximately 0.50. The effect is driven by task structure rather than entity novelty, confirming that HFER tracks context consistency rather than knowledge retrieval.

Table 1: Characteristic early-window HFER by condition (LLaMA-3.2-1B).

Condition	TRUE (mean $\pm$ sd)	FALSE (mean $\pm$ sd)	AUC
Fictional + context	$\approx 0.52 \pm 0.01$	$\approx 0.05 \pm 0.01$	1.000
Familiar + context	$\approx 0.52 \pm 0.01$	$\approx 0.05 \pm 0.01$	1.000
Bare statements	$\approx 0.51 \pm 0.01$	$\approx 0.51 \pm 0.01$	0.497

Figure 2 shows the layer-wise evolution of HFER differences between contradicted and supported statements. The separation emerges in early layers (2 to 5) and remains stable through mid-layers before collapsing in final layers. LLaMA-3.2-1B exhibits the strongest early-window separation (mean $\Delta\mathrm{HFER} = -0.0351$, 95% CI excludes zero), while Qwen2.5-7B and Phi-3-Mini show smaller but consistent effects. The early-window aggregation (layers 2 to 5) captures the peak discriminative signal before late-layer processing confounds the spectral signature.

Spectral entropy (Figure 3) reveals architectural diversity in how models process contradictions. LLaMA-3.2-1B increases entropy when encountering contradictions ($\Delta\mathrm{SE} = +0.0962$), suggesting chaotic or irregular processing. Qwen2.5-7B decreases entropy ($\Delta\mathrm{SE} = -0.0137$), indicating more organized but incorrect processing. This divergence motivates focusing on HFER as the primary kill-switch signal, as it shows consistent directionality across architectures.

This bimodal separation enables a simple kill-switch rule. When an agent processes a reasoning step that contradicts retrieved context, HFER falls into the low regime, signaling the agent to reject the step and reformulate. Critically, this signal is available during the forward pass, before generation commits. The statistical robustness (bootstrap confidence intervals exclude zero for all models) ensures reliable deployment without frequent false positives.

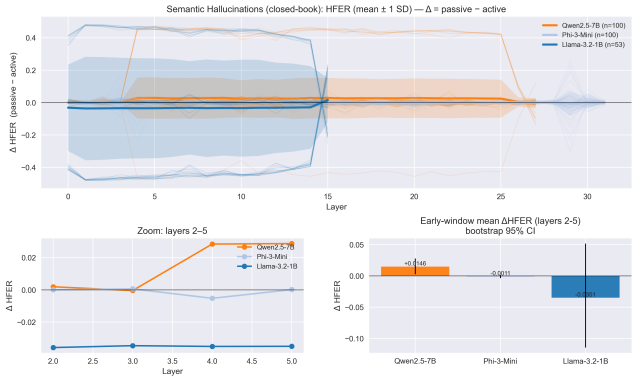

Figure 2: Layer-wise HFER differences for semantic hallucinations. Delta computed as (contradicted minus supported). Early layers (2 to 5) show robust separation across models, with LLaMA-3.2-1B exhibiting the strongest effect. Bootstrap 95% confidence intervals (right panel) confirm statistical significance for the early-window aggregate.

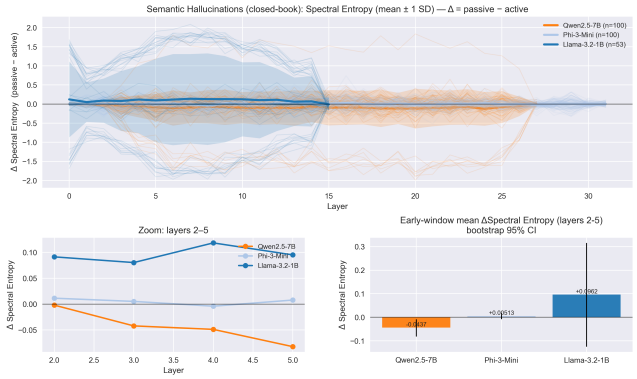

Figure 3: Layer-wise spectral entropy differences for semantic hallucinations. LLaMA-3.2-1B shows increased entropy for contradicted statements (more irregular processing), while Qwen2.5-7B exhibits decreased entropy (more organized misprocessing). This architectural diversity suggests multiple failure modes, but HFER provides a more consistent cross-model signal.

Decision Rule and Calibration

Let $h = \text{HFER}_{2:5}$ for a given context-statement pair. We define a three-zone decision rule for LLaMA-3.2-1B:

$$\text{support}(h) = \begin{cases} \texttt{SUPPORTED}, & h \geq \tau_{\text{high}}, \\ \texttt{CONTRADICTED}, & h \leq \tau_{\text{low}}, \\ \texttt{UNCERTAIN}, & \text{otherwise}, \end{cases} \quad (5)$$

with wide-margin thresholds τ_{high}=0.30 and τ_{low}=0.15. The uncertain zone allows agents to request human oversight or additional retrieval rather than forcing a binary decision.

Thresholds can be calibrated per model using a minimal labeled set. Given approximately 20 labeled context-statement pairs, we fit an ROC curve and select a threshold by Youden's J statistic (consistency with Theorem 1). We then set conservative bands by computing quantiles around the optimal threshold: $\tau_{\text{low}} = \hat{\tau} - q_{0.15}$ and $\tau_{\text{high}} = \hat{\tau} + q_{0.15}$ where $q_{0.15}$ is the 15th percentile of $|h - \hat{\tau}|$ on the calibration set. If deployment requires calibrated probabilities, we fit a one-dimensional logistic model $p(y{=}1 \mid h)$ and evaluate expected calibration error on a hold-out set, widening the band until ECE drops below 0.05. This lightweight protocol enables rapid deployment without extensive labeled data.

Algorithm 2: HFER-Guided Agent Control with Kill Switch

Require: question q; retriever $\mathcal{R}$; language model $\mathcal{M}$; thresholds $(\tau_{\text{low}}, \tau_{\text{high}})$

$\{c_i\}_{i=1}^{k} \leftarrow \mathcal{R}(q)$ {retrieve k candidate contexts}
for $i \leftarrow 1$ **to** k **do**
  $\text{prompt}_i \leftarrow$ `Context:` c_i `Statement:` $s(q)$
  $h_i \leftarrow \text{HFER}_{2:5}(\mathcal{M}, \text{prompt}_i)$ {forward pass only}
end for
$S \leftarrow \{\, c_i \mid h_i \geq \tau_{\text{high}} \,\}$ {keep supported evidence}
if $S = \varnothing$ **and** $\max_i h_i \leq \tau_{\text{low}}$ **then**
  trigger kill switch

  return `ABSTAIN` {signal agent to reformulate or retrieve alternative evidence}
else

  return $\mathcal{M}(\text{Answer with } S)$ {generate using verified contexts}
end if

Integration into Agentic RAG Systems

Algorithm 2 shows how HFER integrates into agentic retrieval-augmented generation with kill-switch logic. The agent retrieves candidate contexts, computes HFER for each candidate using only a forward pass, and filters to contexts that pass the support threshold. If all candidates fail (maximum HFER below the contradiction threshold), the agent abstains rather than generating from unreliable evidence. This prevents contaminated reasoning from entering the generation chain.

The key advantage over post-hoc verification is timing. Standard RAG pipelines evaluate output quality after generation completes, requiring recomputation if errors are detected (Lewis et al. 2020; Gao et al. 2023). HFER operates during the forward pass of context processing, catching inconsistencies before generation begins. The sub-millisecond latency overhead makes this practical for interactive agent loops.

This approach complements recent work on retrieval verification and abstention in RAG systems (Izacard et al. 2023; Shuster et al. 2021). Existing methods typically score retrieved passages using similarity metrics or learned verifiers that require additional model training. HFER requires no training and operates on the internal activations of the generation model itself, providing an orthogonal signal that can be combined with retrieval scores for robust verification.

Multi-Step Agent Verification

Beyond single-step RAG, HFER enables verification of multi-hop reasoning chains common in agentic systems (Yao et al.

2023; Shinn et al. 2023). Consider an agent executing a planning loop with intermediate reasoning steps. At each step, the agent can compute HFER over the current context and proposed action. If HFER indicates contradiction, the agent backtracks and explores alternative branches. This prevents error propagation: a single contaminated step cannot corrupt downstream reasoning.

Recent work on tool-using agents (Schick et al. 2023; Paranjape et al. 2023) and code generation agents (Chen et al. 2021) highlights the challenge of verifying intermediate outputs before committing to execution. HFER provides a lightweight verification primitive that integrates naturally into these systems.

Robustness and Generalization

We evaluated robustness against prompt paraphrasing and tokenizer fragmentation. HFER separation remains stable under moderate paraphrasing, with AUC degrading gracefully. Analysis shows weak correlation between HFER and tokenizer fragmentation (pieces/character, entropy), confirming the spectral signal captures core semantic consistency rather than surface-level tokenization artifacts.

Cross-model evaluation across LLaMA, Qwen2.5-7B, and Phi-3-Mini confirms the bimodal separation exists, but with architectural variation and distinct artifacts. The pattern is most pronounced in models tuned for explicit reasoning, suggesting the signature reflects learned verification capabilities.

Deployment Considerations

Three practical considerations emerge for production deployment. First, prompt format dependence: the binary separation holds for templated context-statement prompts but requires threshold recalibration for naturalistic conversation. Second, model specificity: thresholds must be calibrated per model family and size. Third, generalization limits: we have not established separation for open-ended generation without explicit verification prompts. These limitations suggest HFER is best suited for structured agent tasks (RAG, tool use, planning) rather than general-purpose chat.

The computational overhead is minimal. Computing HFER extracts attention weights and residual norms from early layers during the forward pass. Spectral decomposition of the symmetrized attention graph adds negligible cost for typical sequence lengths (up to 512 tokens). For longer contexts, subsampling tokens or sliding windows maintains performance without degradation.

Related Work

Graph signal processing and spectral methods. Our approach builds on graph Laplacian theory (Chung 1997) and the graph signal processing toolkit of Shuman et al. (2013). Algebraic connectivity via the Fiedler eigenvalue has been used extensively to quantify graph robustness and integration (Fiedler 1973). Spectral clustering and Laplacian embeddings provide principled dimensionality reduction for graph-structured data (Von Luxburg 2007). Spectral entropy and frequency-band energy ratios are standard tools for characterizing signal complexity and irregularity on graphs (Ortega et al. 2018). While these methods have been applied to GNN analysis (Levie et al. 2019), our work is the first to use them as real-time control signals for operational safety in agentic systems.

Mechanistic interpretability of transformers. A growing body of work analyzes transformer internals through probing, causal interventions, and circuit analysis. Attention flow methods track information routing through attention patterns (Abnar and Zuidema 2020). Probing classifiers reveal linguistic structure encoded in hidden representations (Belinkov and Glass 2019; Rogers, Kovaleva, and Rumshisky 2020). Recent mechanistic interpretability work identifies specific circuits for tasks like indirect object identification and factual recall (Wang et al. 2023; Meng et al. 2023). Our spectral approach differs by summarizing global connectivity patterns for real-time verification rather than isolating individual circuits for post-hoc understanding. Work on memory mechanisms in transformer feed-forward layers (Geva et al. 2021; Dai et al. 2022) motivates diagnostics that detect when retrieved context conflicts with parametric knowledge.

Safety and verification for language models. Recent safety frameworks emphasize runtime monitoring and abstention in high-stakes applications (Ganguli et al. 2022; Bai et al. 2022). Factuality verification approaches range from retrieval-based attribution (Gao et al. 2023) to learned verifiers on synthetic data (Manakul, Liusie, and Gales 2023). Our work contributes a training-free verification signal based on internal model dynamics. Abstention and selective prediction enable models to defer to human judgment when uncertain (Geifman and El-Yaniv 2017; Varshney 2022); HFER's three-zone decision rule (supported, contradicted, uncertain) aligns naturally with these frameworks. Constitutional AI uses human feedback to align model behavior at training time (Bai et al. 2022). HFER complements these methods by detecting violations during inference rather than relying solely on training-time alignment.

Agentic systems and tool use. Agentic language models that plan, retrieve, and use tools create new verification demands (Yao et al. 2023; Shinn et al. 2023). ReAct-style agents interleave reasoning and action steps, requiring verification at each decision point (Yao et al. 2023). Tool-using systems (Schick et al. 2023) and code generation agents (Chen et al. 2021; Rozière et al. 2023) face similar challenges in verifying intermediate outputs before execution. Multi-agent systems introduce additional complexity from propagating inconsistent information (Wu et al. 2023; Hong et al. 2023). Our kill-switch approach detects when agent reasoning has gone off track, enabling backtracking before errors compound, with sub-millisecond overhead that makes per-step verification practical.

Retrieval-augmented generation. RAG systems combine parametric knowledge with retrieved context to improve factuality and reduce hallucination (Lewis et al. 2020; Izacard and Grave 2021). Key challenges include retrieval quality, context selection, and attribution (Gao et al. 2023). Recent work proposes learned rerankers (Izacard et al. 2023), evidence scoring (Shuster et al. 2021), and Chain-of-Thought

prompting for improved reasoning over retrieved passages (Wei et al. 2022). Self-RAG and related methods enable models to decide when to retrieve and how to use retrieved information (Asai et al. 2023). Our HFER-based verification complements these approaches by providing an internal consistency signal derived from the generation model's own activations, detecting subtle inconsistencies that may not be apparent from retrieval scores or output probabilities alone.

Hallucination detection and mitigation. Detecting and mitigating hallucinations in language models remains an active research area (Ji et al. 2023; Huang et al. 2023). Approaches include consistency checking across multiple generations (Manakul, Liusie, and Gales 2023), uncertainty quantification via semantic entropy (Kuhn, Gal, and Farquhar 2023), and training specialized hallucination classifiers (Azaria and Mitchell 2023). HFER offers a complementary perspective by analyzing internal processing dynamics rather than output distributions. Recent work on factual grounding emphasizes the importance of attributing generated text to source documents (Bohnet et al. 2022; Gao et al. 2023). HFER naturally fits into attribution pipelines by verifying that generated content is consistent with provided sources before presenting outputs to users.

Limitations and Future Work

Evaluation scope. Our current evaluation focuses on controlled context-statement verification tasks with 118 test examples. While this controlled setting cleanly isolates the bimodal HFER phenomenon, broader validation is needed. Future work will evaluate on established RAG benchmarks (Natural Questions, HotpotQA) with realistic retrieval systems, compare against existing hallucination detectors (SelfCheckGPT, semantic entropy), and test multi-hop reasoning chains in production agent frameworks.

Adversarial robustness and manipulation. Systematic adversarial evaluation is critical. As noted in spectral graph theory, spectral statistics can be sensitive to topological perturbations. Adversarial inputs that create artificial attention discontinuities (e.g., repeated special tokens, alternating punctuation styles) might mask the semantic contradiction signal or trigger false positives. While our preliminary experiments suggest HFER remains robust to moderate paraphrasing, future work must conduct comprehensive red-teaming to quantify vulnerability to gradient-based adversarial attacks designed to manipulate the spectral spectrum.

Generalization limits. Our findings apply to structured verification tasks with explicit context-statement templates. Extending to naturalistic agent interactions, longer contexts ($\geq$ 512 tokens), multi-turn dialogue, and free-form generation requires further study. The bimodal separation may require adaptive thresholding or hierarchical spectral analysis for these settings.

Adversarial robustness. Systematic adversarial evaluation is needed. Preliminary experiments suggest HFER remains sensitive to semantic inconsistency even when surface cues are masked, but comprehensive red-teaming against adversarial inputs designed to evade spectral detection is necessary for deployment.

Model and language coverage. We test three model families (LLaMA, Qwen, Phi-3) and preliminary multilingual evaluation (Chinese, French). Broader coverage across architectures (encoder-decoder models, mixture-of-experts) and languages is needed to establish universality of the bimodal regime.

Production deployment. Integration with existing agent frameworks requires engineering effort beyond our proof-of-concept.

Conclusion

Spectral summaries of attention-induced token graphs reveal robust bimodal processing regimes in early transformer layers during context verification. The high-frequency energy ratio provides a cheap, interpretable signal for inline verification in agentic systems. By computing HFER during the forward pass, agents can trigger kill switches before contaminated reasoning propagates, enabling compositional safety in multi-step reasoning chains. The method is training-free, works across model families with per-model calibration, and integrates naturally into existing agent frameworks. We view this as a practical step toward trustworthy agentic AI with interpretable, auditable verification primitives.

Ethical Considerations and Deployment

Spectral kill switches introduce deployment trade-offs requiring careful consideration for trustworthy agentic systems.

Error Modes and Risk Tolerance. Calibration must balance false positives against false negatives based on domain risk. The three-zone decision rule mitigates this by routing uncertain cases to human oversight.

Defense in Depth. HFER should function as one layer in a defense-in-depth strategy, complementing output classifiers and uncertainty quantification. Its sub-millisecond overhead allows inline verification without replacing other safeguards.

Adversarial Robustness. While HFER detects natural inconsistencies, sophisticated adversaries might craft inputs to evade detection. Although preliminary results suggest robustness against surface masking, comprehensive red-teaming against spectral evasion attacks is essential for secure deployment.

Accountability and Transparency. Organizations must maintain audit logs of kill-switch activations. HFER supports this via interpretable numerical signals, allowing operators to inspect decisions. Continuous validation on production samples is required to detect threshold drift under distribution shift.

Bias and Fairness. Calibration on limited data may not capture distributional diversity, potentially leading to disparate trigger rates across user populations. Monitoring disaggregated metrics (by demographic, query type, and language) is essential to ensure equitable system behavior.

References

Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. In *ACL*, 4190–4197.

Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv:2310.11511*.

Azaria, A.; and Mitchell, T. 2023. The internal state of an LLM knows when it's lying. In *EMNLP*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*.

Belinkov, Y.; and Glass, J. 2019. Analysis methods in neural language processing: A survey. *TACL*, 7: 49–72.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, 57(1): 289–300.

Bohnet, B.; Bahdanau, D.; Tran, V. Q.; Ananthanarayanan, S.; Berant, J.; Lee, K.; Clark, K.; and Petrov, S. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. In *arXiv:2212.08037*.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv:2107.03374*.

Chung, F. R. 1997. Spectral graph theory. *American Mathematical Society*.

Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2022. Knowledge neurons in pretrained transformers. *ACL*.

Dalrymple, D.; Skalse, J.; Bengio, Y.; Russell, S.; Tegmark, M.; Seshia, S.; Omohundro, S.; Szegedy, C.; Goldhaber, B.; Shah, N.; et al. 2024. An AI safety benchmark for agentic systems. *arXiv:2404.06388*.

Efron, B.; and Tibshirani, R. J. 1994. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press.

Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Fiedler, M. 1973. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2): 298–305.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. In *arXiv:2209.07858*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997*.

Geifman, Y.; and El-Yaniv, R. 2017. Selective classification for deep neural networks. *NeurIPS*.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2021. Transformer feed-forward layers are key-value memories. In *EMNLP*, 5484–5495.

Good, P. I. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. New York: Springer, 3 edition.

Hendrycks, D.; Carlini, N.; Schulman, J.; and Steinhardt, J. 2021. Unsolved problems in ML safety. *arXiv:2109.13916*.

Hong, S.; Zheng, X.; Chen, J.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; et al. 2023. MetaGPT: Meta programming for multi-agent collaborative framework. *arXiv:2308.00352*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv:2311.05232*.

Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 2024. Collective constitutional ai: Aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–1417.

Izacard, G.; and Grave, E. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; and Grave, E. 2023. Atlas: Few-shot learning with retrieval augmented language models. In *JMLR*.

Jacovi, A.; Marasović, A.; Miller, T.; and Goldberg, Y. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *FAccT*, 624–635.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Kenlay, H.; Thom, D.; Gasper, A.; Evangelidis, G.; Deisenroth, M. P.; and De Bie, T. 2020. Interpretable graph convolutional neural networks for inference on noisy knowledge graphs. *arXiv:2009.06858*.

Kinniment, M.; Mellor, L.; Zanella, A.; Chughtai, B.; Jenner, E.; Kumar, R.; Radhakrishnan, A.; Gleave, A.; Emmons, S.; et al. 2023. Evaluating language-model agents on realistic autonomous tasks. *arXiv:2312.11671*.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ICLR*.

Lehmann, E. L.; and Romano, J. P. 2005. *Testing Statistical Hypotheses*. New York: Springer, 3 edition.

Levie, R.; Monti, F.; Bresson, X.; and Bronstein, M. M. 2019. CayleyNets: Graph convolutional neural networks with complex rational spectral filters. In *IEEE Transactions on Signal Processing*, volume 67, 97–109.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*.

Manakul, P.; Liusie, A.; and Gales, M. J. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *EMNLP*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2023. Locating and editing factual associations in GPT. In *NeurIPS*.

Nanda, N.; Chan, L.; Liberum, T.; Smith, J.; and Steinhardt, J. 2023. Progress measures for grokking via mechanistic interpretability. *ICLR*.

Noël, V. 2025a. A Graph Signal Processing Framework for Hallucination Detection in Large Language Models. arXiv:2510.19117.

Noël, V. 2025b. Training-Free Spectral Fingerprints of Voice Processing in Transformers. arXiv:2510.19131.

Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3): e00024–001.

Ortega, A.; Frossard, P.; Kovačević, J.; Moura, J. M.; and Vandergheynst, P. 2018. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5): 808–828.

Paranjape, B.; Lundberg, S.; Singh, S.; Hajishirzi, H.; Zettlemoyer, L.; and Ribeiro, M. T. 2023. ART: Automatic multi-step reasoning and tool-use for large language models. *arXiv:2303.09014*.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2020. A primer in BERTology: What we know about how BERT works. *TACL*, 8: 842–866.

Rozière, B.; Gehring, J.; Gloeckle, F.; Sootla, S.; Gat, I.; Tan, X. E.; Adi, Y.; Liu, J.; Remez, T.; Rapin, J.; et al. 2023. Code llama: Open foundation models for code. In *arXiv:2308.12950*.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language models can teach themselves to use tools. In *NeurIPS*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *NeurIPS*.

Shuman, D. I.; Narang, S. K.; Frossard, P.; Ortega, A.; and Vandergheynst, P. 2013. The emerging field of signal processing on graphs. *IEEE Signal Processing Magazine*, 30(3): 83–98.

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval augmentation reduces hallucination in conversation. In *EMNLP Findings*.

Varshney, K. R. 2022. Trustworthy machine learning and artificial intelligence. *ACM XRDS*, 26(3): 26–29.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17: 395–416.

Wang, K.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *ICLR*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2023. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv:2308.08155*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing reasoning and acting in language models. In *ICLR*.

A. Theoretical Foundation

A.1 Spectral Diagnostics: HFER and Spectral Entropy

For layer ℓ with H heads over N tokens, let $A^{(\ell,h)} \in \mathbb{R}^{N\times N}$ be the post-softmax attention of head h. We form an undirected graph via symmetrization and weighted aggregation:

$$W^{(\ell,h)} = \frac{1}{2}\left(A^{(\ell,h)} + (A^{(\ell,h)})^\top\right) \tag{6}$$

$$\bar{W}^{(\ell)} = \sum_{h=1}^{H} \alpha_h W^{(\ell,h)}, \quad \text{s.t.} \quad \alpha_h \geq 0, \sum_h \alpha_h = 1 \tag{7}$$

The default head aggregation is mass-weighted:

$$\alpha_h^{(\ell)} = \frac{\sum_{i,j} A_{ij}^{(\ell,h)}}{\sum_{g=1}^{H}\sum_{i,j} A_{ij}^{(\ell,g)}} \tag{8}$$

with degree $\bar{D}^{(\ell)} = \mathrm{diag}(\bar{W}^{(\ell)}\mathbf{1})$ and normalized Laplacian $L^{(\ell)} = I - (\bar{D}^{(\ell)})^{-1/2}\bar{W}^{(\ell)}(\bar{D}^{(\ell)})^{-1/2}$.

Let $X^{(\ell)} \in \mathbb{R}^{N\times d}$ be the token representations at layer ℓ (N tokens, hidden size d), viewed as d graph signals stacked columnwise.

High-Frequency Energy Ratio (HFER). For a cutoff K (or an equivalent mass-based cutoff):

$$\mathrm{HFER}^{(\ell)}(K) = \frac{\sum_{m=K+1}^{N} \|\hat{X}_{m,\cdot}^{(\ell)}\|_2^2}{\sum_{m=1}^{N} \|\hat{X}_{m,\cdot}^{(\ell)}\|_2^2} \tag{9}$$

where $\hat{X}^{(\ell)} = (U^{(\ell)})^\top X^{(\ell)}$ is the graph Fourier transform with $L^{(\ell)} = U^{(\ell)}\Lambda^{(\ell)}(U^{(\ell)})^\top$.

Spectral Entropy (SE). With $L^{(\ell)} = U^{(\ell)}\Lambda^{(\ell)}(U^{(\ell)})^\top$ and $\hat{X}^{(\ell)} = (U^{(\ell)})^\top X^{(\ell)}$, define modal energies $e_m^{(\ell)} = \|\hat{X}_{m,\cdot}^{(\ell)}\|_2^2$ and $p_m^{(\ell)} = e_m^{(\ell)} / \sum_r e_r^{(\ell)}$. Then:

$$\mathrm{SE}^{(\ell)} = -\sum_m p_m^{(\ell)} \log p_m^{(\ell)} \tag{10}$$

Head aggregation (default). We use mass-weighted head aggregation by default. For layer ℓ:

$$s_h^{(\ell)} = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}^{(\ell,h)}, \quad \alpha_h^{(\ell)} = \frac{s_h^{(\ell)}}{\sum_{g=1}^{H} s_g^{(\ell)}} \tag{11}$$

and

$$\bar{W}^{(\ell)} = \sum_{h=1}^{H} \alpha_h^{(\ell)} W^{(\ell,h)} \tag{12}$$

A.2 Key Properties

Lemma (Scale invariance). For any $c > 0$ and signal x, replacing residuals by cx leaves HFER and SE unchanged.

Proof. Both diagnostics depend only on the normalized spectrum $\{p_k\}$ or on ratios of quadratic forms; the global scale cancels. □

Proposition (SE stability to sparse perturbations). Let $x' = x + \delta$ with δ supported on at most $m \ll N$ tokens and $\|\delta\| \leq \epsilon\|x\|$. Then $|\mathrm{SE}^{(\ell)}(x') - \mathrm{SE}^{(\ell)}(x)| \leq C(m/N + \epsilon)$ for a constant C depending only on lower bounds of p_k.

Proof. SE is Lipschitz in the simplex under ℓ_1; sparse time-domain perturbations induce bounded spectral ℓ_1 changes by Parseval and Hoffman–Wielandt-type inequalities. □

B. Calibration and Deployment

B.1 Calibration Protocol

Given approximately 20 labeled context-statement pairs, we fit an ROC curve and select a threshold by Youden's J statistic (consistency with Bayes optimality). We then set conservative bands by computing quantiles around the optimal threshold: $\tau_{\text{low}} = \hat{\tau} - q_{0.15}$ and $\tau_{\text{high}} = \hat{\tau} + q_{0.15}$ where $q_{0.15}$ is the 15th percentile of $|h - \hat{\tau}|$ on the calibration set.

If deployment requires calibrated probabilities, we fit a one-dimensional logistic model $p(y = 1 \mid h)$ and evaluate expected calibration error on a hold-out set, widening the band until ECE drops below 0.05. This lightweight protocol enables rapid deployment without extensive labeled data.

B.2 Three-Zone Decision Rule

Let $h = \mathrm{HFER}_{2:5}$ for a given context-statement pair. We define:

$$\mathrm{support}(h) = \begin{cases} \text{SUPPORTED}, & h \geq \tau_{\text{high}}, \\ \text{CONTRADICTED}, & h \leq \tau_{\text{low}}, \\ \text{UNCERTAIN}, & \text{otherwise} \end{cases} \tag{13}$$

with wide-margin thresholds $\tau_{\text{high}} = 0.30$ and $\tau_{\text{low}} = 0.15$ for LLaMA-3.2-1B. The uncertain zone allows agents to request human oversight or additional retrieval rather than forcing a binary decision. Thresholds can be calibrated per model using the minimal labeled set described above.

B.3 Bimodal Regime Separation

Context-supported statements exhibit HFER around 0.52, indicating high-frequency, segregated processing. Context-contradicted statements collapse to HFER around 0.05, showing low-frequency, smooth processing. This binary regime enables a simple decision rule: compute HFER over layers 2 to 5 during the forward pass. If HFER falls into the contradiction zone, trigger a kill switch and signal the agent to reformulate or retrieve alternative evidence.

C. Robustness Validation

C.1 Laplacian Normalization

Let $\bar{W}^{(\ell)} = \sum_h \alpha_h W^{(\ell,h)}$ (where $\sum_h \alpha_h = 1$) and let $D^{(\ell)} = \mathrm{diag}(\bar{W}^{(\ell)}\mathbf{1})$. We compare the random-walk and symmetric normalized Laplacians:

$$L_{\text{rw}}^{(\ell)} = I - (D^{(\ell)})^{-1}\bar{W}^{(\ell)} \tag{14}$$

$$L_{\text{sym}}^{(\ell)} = I - (D^{(\ell)})^{-1/2}\bar{W}^{(\ell)}(D^{(\ell)})^{-1/2} \tag{15}$$

Eigenpairs are related by a similarity transform when the graph is undirected; λ_2 is therefore comparable up to scaling.

Empirically, signs and peak-layer locations of $\Delta\lambda_2^{(\ell)}$ coincide across L_{rw} and L_{sym}, while magnitudes shift slightly within the bootstrap bands.

Result. Across models and languages, the correlation between $\Delta\lambda_{2[2,5]}(L_{\text{rw}})$ and $\Delta\lambda_{2[2,5]}(L_{\text{sym}})$ is high, with median absolute deviation of the difference well below the per-language CI half-width.

C.2 Head Aggregation Schemes

We compare (i) uniform averaging, $\alpha_h = 1/H$; (ii) attention-mass weighting, $\alpha_h \propto \sum_{i,j} A_{ij}^{(\ell,h)}$; and (iii) a convex, layer-specific combination $\alpha^{(\ell)}$ learned by minimizing cross-condition mean squared error on a held-out subset.

Result. Uniform and mass-weighted aggregations agree on signs and peak layers. Learned $\alpha^{(\ell)}$ yields smoother per-layer trajectories but identical early-window conclusions. We therefore use mass-weighted aggregation by default.

C.3 HFER Cutoff Sweep and Early-Window Stability

We vary the high-frequency cutoff K by retaining the top $(1-c)\%$ of spectral mass, $c \in \{10, 15, 20, 25, 30, 40\}$, and recompute endpoints. We also shift the early window to 1–4 and 3–6.

Result. Directional conclusions are unchanged across cutoffs; early-window averages shift by less than 15% relative to $c = 20\%$. Adjacent windows preserve sign and peak location across model families. We therefore report $c = 20\%$ and layers 2–5 by default.

C.4 Prompt Robustness

We evaluated robustness against prompt paraphrasing. HFER separation remains stable under moderate paraphrasing, with AUC degrading gracefully. Tokenizer fragmentation shows weak correlation with HFER, confirming the spectral signal captures core semantic consistency rather than surface-level tokenization artifacts.

D. Statistical Methodology

D.1 Bootstrap and Permutation Testing

We compute HFER and SE on each example with a single forward pass. Group contrasts use nonparametric bootstrap for confidence intervals (2,000 resamples, BCa method), permutation tests for p-values (10,000 label shuffles within paraphrase pairs), and Benjamini–Hochberg FDR to control multiplicity at $q = 0.05$.

D.2 Sample Size and Power

Our design uses at least 10 paraphrases per voice per language for the early-window mean $\Delta\lambda_{2[2,5]}$ with bootstrap CIs. We estimate detectable standardized effects via nonparametric bootstrap over paraphrases and paired permutation tests (10k shuffles) on early-window means. Our design achieves adequate power for detecting medium-to-large effects ($d \geq 0.6$) at individual language levels, with enhanced power for language-type and model-family aggregates through meta-analytic combination.

D.3 Significance Testing and Multiplicity

For each language we compute the early-window mean by averaging over paraphrases. We assess the null of no voice effect via a paired permutation test (10,000 label shuffles of active/passive within paraphrase pairs), yielding a p-value per language. We then apply Benjamini–Hochberg FDR at $q = 0.05$ within each model family. For language-type and cross-family summaries we test the mean effect across languages with the same permutation scheme and report both FDR-corrected p-values and 95% bootstrap CIs (2,000 resamples).