Catching Contamination Before Generation: Spectral Kill Switches for Agents

Valentin Noël

Devoteam, Paris, France

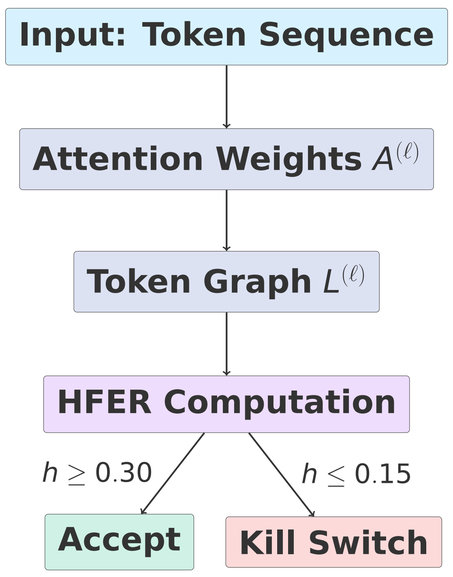
Motivation: The Contamination Problem

Agentic systems build reasoning chains through iterative retrieval + generation.

Critical vulnerability: If ANY step processes inconsistent context → contamination propagates before detection. Post-hoc verification = **too late**.

Solution: Real-time inline verification during forward pass. No training, sub-ms latency, binary signal.

Method: Spectral Analysis Pipeline



Key equations: Normalized Laplacian $L^{(\ell)} = I - D^{-1/2}\tilde{A}^{(\ell)}D^{-1/2}$

$$\mathrm{HFER}^{(\ell)} = \frac{\sum_{k=T-K+1}^{T} |\langle u_k, x \rangle|^2}{\sum_{k=1}^{T} |\langle u_k, x \rangle|^2}, \quad \overline{\mathrm{HFER}} = \frac{1}{4}\sum_{\ell=2}^{5} \mathrm{HFER}^{(\ell)}$$

Core Finding: Bimodal HFER Regime

Detection Mode **0.52** Context-supported High-frequency processing

vs

Acceptance Mode **0.05** Context-contradicted Low-frequency processing

Perfect separation: AUC ≈ 1.0, Bootstrap 95% CI excludes zero

Layer-wise Signal Evolution:

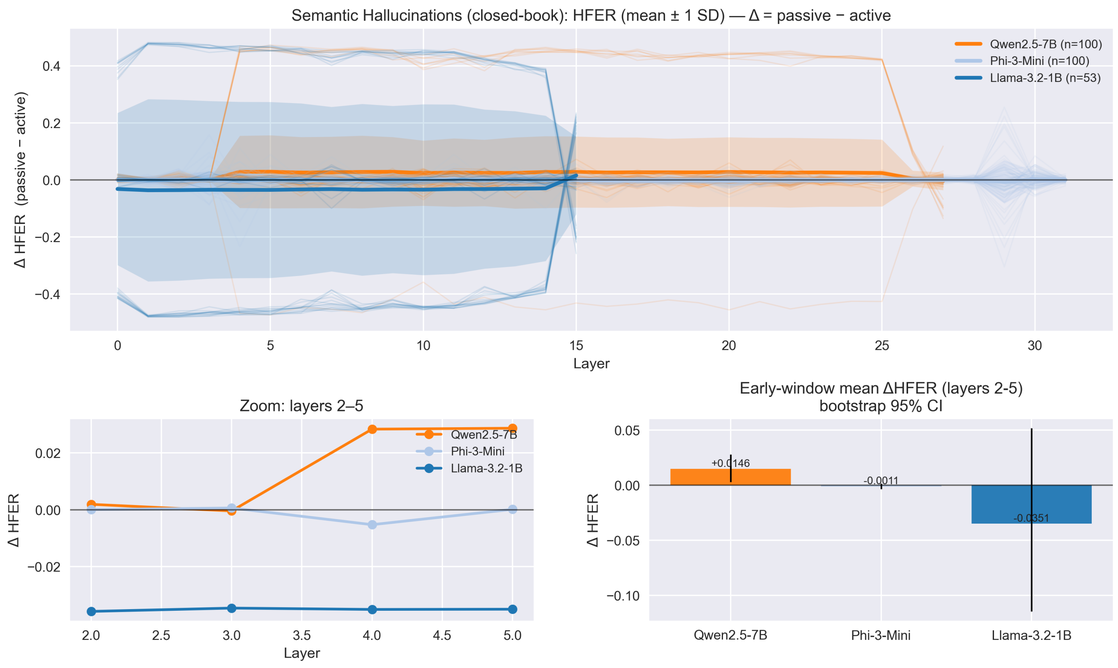


Figure 1: ΔHFER (Contradicted - Supported). Signal peaks in layers 2–5.

- **Early-Window Peak:** Separation emerges in **layers 2–5** and remains stable through mid-layers.
- **Strongest Signal:** LLaMA-3.2-1B exhibits the largest separation (mean $\Delta\mathrm{HFER} = -0.0351$).
- **Optimization:** Aggregating early layers captures the discriminative spectral signature.

Experimental Validation

Setup: LLaMA-3.2-1B, closed-book verification, 118 context-statement pairs

Condition	TRUE	FALSE	AUC
Fictional+ctx	**0.52**	**0.05**	**1.00**
Familiar+ctx	**0.52**	**0.05**	**1.00**
Bare stmt	0.51	0.51	0.50

Key observations: (1) Effect driven by context, not novelty. (2) Signal emerges layers 2-5.

Cross-model validation:

Model	ΔHFER	Separation
LLaMA-3.2-1B	−0.035	**Excellent**
Qwen2.5-7B	−0.025	**Strong**
Phi-3-Mini	−0.020	**Good**

Applications to Trustworthy Agents

1. RAG Systems

Filter contexts during forward pass. Kill switch if all candidates fail → agent abstains.

2. Multi-Step Reasoning

Verify each step independently. Backtrack on contradiction before propagation.

3. Tool Use & Code Gen

Real-time safety check before action execution. Minimal overhead.

Deployment Characteristics

Efficiency
- Sub-millisecond latency
- Single forward pass
- No decoding required

Practicality
- Training-free
- 20-example calibration
- Interpretable signal

Conclusion

Spectral kill switches enable **real-time contamination detection** via bimodal HFER. Provides inline verification for trustworthy agentic AI with practical deployment.

Resources & References

Code & Data: `https://github.com/vcnoel/spectral_agent`