Spectral Guardrails: Detecting Prompt Injection via Attention Graph Fracture in LLMs

Charly Ken Capo-Chichi Ghanem Amari Valentin Noël

Devoteam, Levallois-Perret, France

The Problem: What Happens *Inside* the Model During an Injection?

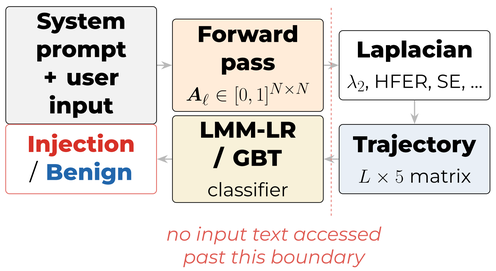
Prompt injection redirects an LLM away from its system instructions by embedding adversarial directives in the user turn.

- **Text-based detectors** (TF-IDF, DeBERTa, PromptGuard): exploit *lexical artifacts* of current benchmarks artifacts that vanish under a simple paraphrase.
- **LLM-as-judge** guards: must themselves read (and can be hijacked by) hostile content.
- **This work**: a mechanistic signal in transformer attention the injection is detected *without ever reading the input text*.

Successful injection $\implies$ attention graph fracture: $\lambda_2 \to 0$

Method: Attention Graphs and Fracture Trajectories

At each layer ℓ, averaged multi-head attention defines a weighted token graph. We track its Laplacian spectrum *across layers* and classify the resulting trajectory the probe never accesses text.



Graph Laplacian (per layer ℓ, heads averaged):

$$\boldsymbol{L}_\ell = \boldsymbol{D}_\ell - \tfrac{1}{2}\left(\boldsymbol{A}_\ell + \boldsymbol{A}_\ell^\top\right) \qquad \Longrightarrow \qquad \lambda_2,\ \mathrm{HFER},\ \mathrm{SE},\ E,\ \mathrm{Smooth}$$

Trajectory features: the $L \times 5$ layer-wise series is aggregated into $\sim$150 shape descriptors windowed means, curvature, `fracture_asymmetry`, anomaly persistence capturing the *morphology* of the collapse, not its depth.

Four Pipelines, One Signal

Pipeline	Input	Training
Sweep	best (metric, layer) pair	**zero** (Youden-J)
SpRich	85-dim trajectory stats	ℓ_2-LogReg
LMM-LR	**raw $L \times 5$ matrix**	**LogReg (5-fold)**
LMM-GBT	raw $L \times 5$ matrix	GBT (5-fold)

Efficiency: Laplacian spectrum costs $O(N^2)$ per layer — a negligible fraction of the forward pass: runs as a streaming side-channel with no extra model calls.

Selected References

[1] Fiedler, M. (1973). Algebraic connectivity of graphs. *Czech. Math. J.*
[2] Chung, F. R. K. (1997). *Spectral Graph Theory.* AMS.
[3] Elhage, N. et al. (2021). A mathematical framework for transformer circuits. *Anthropic.*
[4] Greshake, K. et al. (2023). Not what you've signed up for: Indirect prompt injection. *AISec.*
[5] Perez, F. & Ribeiro, I. (2022). Ignore previous prompt. *NeurIPS ML Safety WS.*
[6] Noël, V. (2025). A GSP framework for hallucination detection in LLMs. *arXiv:2510.19117.*

Main Results: Seven Models, Three Benchmarks, No Text

Dataset	Sweep	SpRich	LMM-LR	LMM-GBT
safe-guard	0.698–0.859	0.910–0.964	**0.948–0.979**	0.952–0.972
jailbreak	0.878–0.963	0.892–0.964	**0.924–0.991**	0.904–0.961
deepset	0.846–0.927	0.894–0.937	**0.917–0.965**	0.895–0.957

ROC-AUC min–max over seven backbones (1.1B–14B), incl. GQA (Llama-3.1-8B, Phi-4-mini).

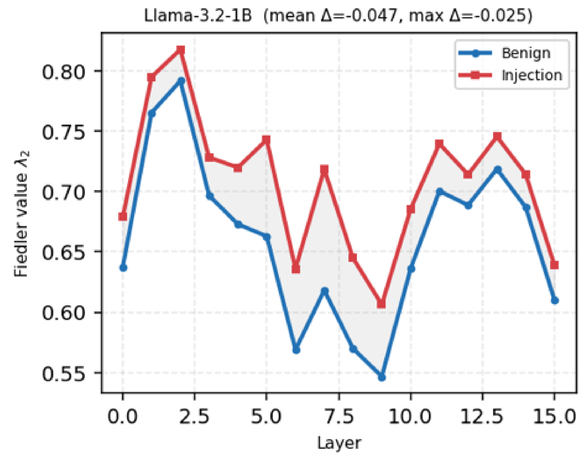
AUC 0.90–0.99
LMM-LR, all model–dataset pairs

1.1B $\to$ 14B
signal does not degrade with depth

0 tokens read
attention weights only

The Fracture — λ_2 Trajectories, Benign vs. Injection

Benign prompts maintain high algebraic connectivity across layers; injections trigger a characteristic collapse and *fail to recover* late-layer connectivity.



The fracture *shape* (asymmetry, late-layer mean) is more discriminative than any single-layer value.

Beating Text Baselines Where It Matters

Method	*safe-guard*	*jailbreak*	*deepset*
TF-IDF + LogReg	**0.999**	0.994	0.948
TF-IDF + SVM	**0.999**	0.996	0.951
DeBERTa-v3 (zero-shot)	0.678	0.819	1.000†
Spectral LMM (Ours)	0.979	**0.991**	**0.965**

†In-distribution: DeBERTa-v3-base-injection was trained on *deepset*.

- TF-IDF is near-perfect on *safe-guard* / *jailbreak* $\Rightarrow$ those benchmarks are **lexically trivial** (bag-of-words artifacts).
- On *deepset*, the least lexically structured benchmark, the spectral probe **surpasses TF-IDF** (0.965 vs. 0.951) — without reading the text.
- DeBERTa-v3, a purpose-built detector, collapses out of distribution (0.678 / 0.819); the spectral LMM stays at 0.965–0.991 across all three.

Mechanistic Core: The Fracture Is *Necessary*, Not Incidental

To hijack behaviour, an injection must redirect attention away from system tokens V_{sys} toward payload tokens V_{inj}, suppressing cross-partition weights:

$$\forall\, i \in V_{\mathrm{sys}},\ j \in V_{\mathrm{inj}}: \quad A_{\ell,ij} \to 0 \implies \lambda_2(\boldsymbol{L}_\ell) \to 0. \tag{1}$$

By **Fiedler's theorem**, $\lambda_2 \to 0$ *iff* the graph partitions — exactly the structural event injection triggers. A functional injection and a healthy attention graph are **mathematically incompatible**.

Evasion is self-defeating: keeping λ_2 high forces the adversary to preserve system–payload connectivity, i.e. to give up behavioural control. Hijack strength above $\epsilon \Rightarrow$ a proportional drop in λ_2 (conjectured spectral evasion bound).

Ablation: Trajectory Shape Beats Snapshots

Feature subfamilies in isolation (Qwen2.5-1.5B-Instruct, *deepset*, 5-fold CV):

Feature subgroup	F1	std
Full pipeline (length-agnostic)	**0.831**	**0.013**
Fiedler trajectory only	0.812	0.057
Deep Pass (L22) snapshot only	0.787	0.074
Noise / HFER only	0.838	0.078
Chaos Index only	0.408	0.215

The full trajectory pipeline is **5.7$\times$ more stable** ($\sigma = 0.013$ vs. 0.074) than the best single-layer snapshot: the detection signal lies in *dynamics*, not instantaneous collapse depth.

Interpretable Feature Importance (ℓ_1)

Feature	Coeff.
`fracture_asymmetry`	$+0.849$
`fiedler_value_late_mean`	-0.849
`spectral_entropy_delta_L2`	$+0.675$
`smoothness_index_d2_L17`	$+0.569$
`hfer_d2_L4`	-0.529
`fiedler_low_persistence`	-0.473

The two dominant features capture complementary aspects of the fracture: *how* the graph collapses and its *failure to recover*. No single-layer snapshot appears in the top 10.

Robustness & Limitations

- **GQA architectures:** Llama-3.1-8B and Phi-4-mini reach LMM-GBT AUC > 0.940 indistinguishable from multi-head.
- **Scale:** stable from 1.1B to 14B; quantisation (8-bit, 4-bit NF4) does not break the signal.
- **White-box requirement:** needs per-layer attention — natural on self-hosted stacks (vLLM, TGI), unavailable for closed APIs.
- **Adaptive adversaries:** injections engineered to keep λ_2 high remain untested; the evasion bound is a conjecture to validate.

Conclusion

Training-light spectral probe on **7 models**, **3 benchmarks**, GQA included:

1. **Necessary signal:** injection $\Rightarrow$ attention fracture, $\lambda_2 \to 0$ (Eq. 1).
2. **Accurate:** LMM-LR ROC-AUC 0.90–0.99 everywhere; beats TF-IDF on *deepset*.

valentin.noel@devoteam.com