Does the Optimal Hallucination Detector for Agentic Tool Calls Depend on Model Scale?

Valentin Noël [1] **Bharathi Srinivasan** [2] **Kait Healy** [3] **Visakh Madathil** [2]

Abstract

An AI agent that hallucinates a tool call typically shows no overt signs of failure: the call is syntactically correct, its arguments are well-formed, and the associated token probabilities appear routine. In these cases, the execution pipeline receives no indication that anything is amiss. The call executes, downstream components consume its output, and the erroneous transaction completes without human scrutiny.

Standard uncertainty signals offer little protection against this failure mode. Token log-probabilities and small-N sampling consensus (N = 5) deliver only 0.51–0.62 AUC across seven models, while incurring roughly a fivefold increase in inference cost. In response, recent work has proposed two single-pass alternatives: spectral analyses of the attention routing graph and hidden-state probes over designated token-role positions. Both families outperform conventional uncertainty metrics, but it remains an open question which detector class is preferable under realistic deployment constraints.

We answer it across seven models from 1B to 14B on identical data, splits, and labels, and find the optimal detector depends on scale non-monotonically. Our **Layerwise Multi-Metric** (LMM) probe on attention alone reaches AUC 0.957 at 3B, exceeding the hidden-state probe (0.947); at 8B, hallucination signal diffuses across 32 layers and trajectory aggregation (SpRich) takes the lead (0.888); at 14B, layer specialization re-concentrates the signal and LMM recovers (0.917 vs. 0.882). Merging the two signal families never helps in our experiments. A controlled within-family study on Llama-3.x (1B/3B/8B) confirms that scale alone drives monotonic improvement in spectral detectors while hidden-state probes saturate near-ceiling at 1B, directly disentangling scale from architectural variation. The deployment rule is to match the guardrail to depth and attention type, not to combine.

Code: `https://github.com/vcnoel/spectral-tool-use`.

1. Introduction

AI agents that invoke external tools (APIs, databases, privileged actions) can fail silently when they hallucinate a tool call or parameters. Unlike open-ended generation, the output is syntactically valid and high-confidence, with a plausible but incorrect function name or argument values. Downstream pipeline stages treat the call as valid, allowing actions such as malformed transactions or misconfigured infrastructure commands to execute before any human review occurs.

Because hallucinated tool calls are executed immediately, any practical defense must detect them at generation time. Consequently, we focus on signals available from a single forward pass without access to a ground-truth oracle or additional sampling budget. Standard uncertainty estimates collapse on this task as token log-probabilities achieve only 0.51–0.62 AUC, and an N=5 sampling-consensus baseline offers no improvement while incurring a fivefold increase in inference cost (Table 1). This failure is structural rather than stochastic because a model that hallucinates a tool call with high confidence does not surface its error as uncertainty.

Two independent lines of work demonstrate that supervised signals extracted from the same forward pass can overcome this limitation. One approach uses spectral diagnostics of the attention graph (Noël, 2026), monitoring the algebraic connectivity (Fiedler value λ_2) of per-layer graph Laplacians without accessing hidden states. The other, introduced by Healy et al. (Healy et al., 2026), trains a lightweight MLP on the internal hidden states of the LLM at the function-name, argument-value, and closing-delimiter token positions in order to predict whether a tool call or its parameters are hallucinated. These methods consume different tensors from the same forward pass and were developed independently.

[1]Devoteam, Paris [2]Amazon Web Services [3]OpenAI. Correspondence to: Valentin Noël <val.entin.noel@devoteam.com>.

The 2^{nd} *Agents in the Wild Workshop at the* 43^{rd} *International Conference on Machine Learning*, Seoul, South Korea. PMLR 306, 2026.

The open question is which detector to deploy in practice, and whether their signals are complementary is what this paper directly addresses.

Our four contributions:

1. A **controlled head-to-head** across seven models (1B to 14B), identical splits and labels, including three concurrent spectral and hidden-state baselines (Binkowski et al., 2025; Li et al., 2025; Badash et al., 2026).
2. A **Layerwise Multi-Metric** (LMM) probe that evaluates spectral features per layer, recovering signal that trajectory aggregation discards.
3. A **precise deployment rule**: LMM dominates at $\leq$3B and $\geq$14B; trajectory aggregation (SpRich) dominates at 8B; GQA attention architecture collapses SpRich while leaving LMM robust; merging never helps.
4. A **within-family Llama-3.x scale study** (1B/3B/8B) that holds architecture constant and isolates scale as the sole variable, confirming that spectral detector improvements are scale-driven.

2. Related Work

Tool-use benchmarks. Gorilla (Patil et al., 2024) and ToolLLM (Qin et al., 2024) benchmark tool-calling capability but not the internal failure modes that produce confident wrong outputs. The token-role hidden-state probe (Healy et al., 2026) was the first to demonstrate that supervised probes on hidden states can detect tool-use hallucination from a single forward pass, establishing a representation built from the function name, argument mean, and closing delimiter as a reliable input to lightweight classifiers. That probe forms one of the two signals we evaluate; we extend it to a multi-layer setting and provide the first comparison against an attention-only alternative.

Internal-state hallucination detection. INSIDE (Chen et al., 2024) uses hidden-state covariance; HALT (Bhatnagar et al., 2026) extends probing to factual claims; process reward models (Lightman et al., 2024) verify reasoning steps. All require hidden-state access. Our spectral signal operates on attention matrices alone.

Spectral methods for transformers. Attention-graph Laplacians were introduced for hallucination detection in (Noël, 2025). Concurrent work confirms the utility of spectral geometry: LapEigvals (Binkowski et al., 2025) uses top-k eigenvalues; HSAD (Li et al., 2025) applies FFT to hidden-layer dynamics; Cross-Layer Agreement (Badash et al., 2026) measures attention consistency across depth. We evaluate all three as baselines.

3. Method

Task and data. We sample N=750–1000 examples per model from the Glaive function-calling benchmark (GlaiveAI, 2024) across seven models: Llama-3.2-1B, Gemma-4-2B, Llama-3.2-3B, Qwen-2.5-3B, Phi-4-Mini (3.8B), Llama-3.1-8B, and Phi-4 (14B, 4-bit NF4). Labels are assigned by JSON schema validation against the ground-truth call. A template-consistent 70/15/15 partition (SHA-256 hash of the ground-truth string) ensures no prompt template seen at training appears at test time, matching the protocol of (Healy et al., 2026).

Signal 1: Spectral attention topology. For each layer ℓ, we average multi-head attention matrices and symmetrize to form an adjacency W_ℓ. The graph Laplacian $L_\ell = D_\ell - W_\ell$ yields the Fiedler value $\lambda_2(L_\ell)$, which measures algebraic connectivity: low values indicate fragmented routing, a pattern associated with hallucination (Noël, 2026). We extract five spectral metrics per layer (Fiedler value, energy, smoothness index, spectral entropy, high-frequency energy ratio) and represent each via 15 trajectory statistics (mean, slope, skewness, dominant FFT harmonics, and others), giving a 135 to 200 dimensional feature vector. A logistic regression with ℓ_2 regularization, C tuned on validation, defines the **SpRich** probe. For efficiency, we use a subgraph-based Lanczos approximation (Fast-Fiedler) that reduces per-sample spectral cost from 3.3 s to 19 ms. This signal requires only attention weight matrices and no hidden-state access.

Signal 2: Hidden-state probe (Hidden). Following (Healy et al., 2026), we extract token-role representations at three structural positions within the generated call:

$$z_\ell = h_\ell[t_{\text{func}}] \parallel \text{mean}(h_\ell[T_{\text{args}}]) \parallel h_\ell[t_{\text{end}}], \qquad (1)$$

where t_{func} is the function-name token, T_{args} are argument-value positions, and t_{end} is the closing delimiter. We concatenate across 8 evenly spaced layers and train a single-hidden-layer MLP (512 units, AdamW, cosine schedule, early stopping).

Layerwise Multi-Metric (LMM) probe. Trajectory aggregation (SpRich) reduces λ_2 across all layers to summary statistics, discarding information about which layers exhibit anomalous routing. LMM treats the raw $L \times 5$ matrix of per-layer metric values as a direct feature vector. We evaluate two variants: **LMM-LR** (logistic regression) and **LMM-GBT** (gradient-boosted trees, depth 3, 200 estimators). Each GBT split implements a learned threshold on a specific (layer, metric) pair. No hidden-state access is required.

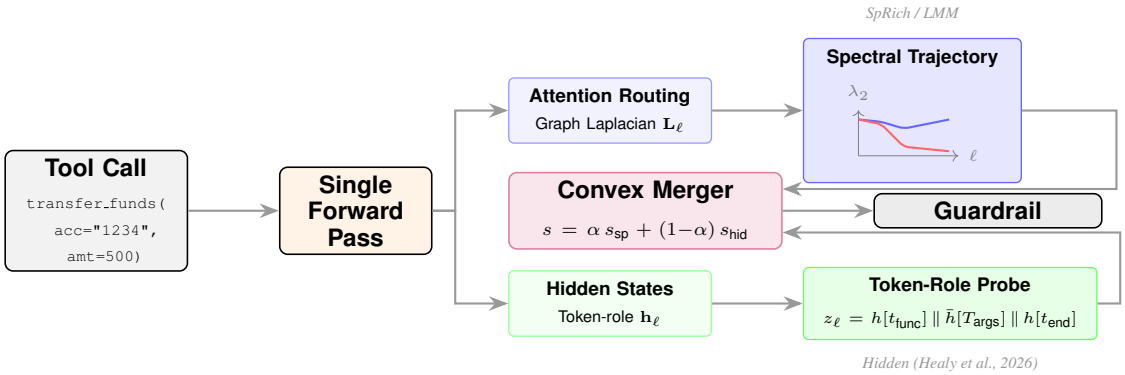

Figure 1. **Architecture.** Two independent detection signals from a single forward pass. Top: spectral topology of attention routing (Fiedler value per layer). Bottom: token-role hidden-state probe (Healy et al., 2026).

Table 1. Unsupervised baselines vs. the best single spectral metric at a Youden-J threshold (Sweep, no training). †: teacher-forcing logprob (generation-based logprob was not run). ‡: reversed scoring direction, high logprob / high consensus predicts hallucination (the model commits confidently to wrong calls); standard direction gives 0.346 and 0.180 respectively.

Model	Logprobs	Consensus	Sweep
Llama-1B	0.510	0.508	**0.773**
Gemma 4 2B	0.672†	0.820†‡	**0.874**
Llama-3B	0.615	0.505	**0.767**
Qwen-2.5-3B	0.579	0.512	**0.818**
Phi-4 (14B)	0.655†‡	0.500†	**0.850**

Convex merger. We test complementarity via $s = \alpha\, s_{\text{sp}} + (1-\alpha)\, s_{\text{hid}}$, with α^* selected over 41 values by maximizing validation AUC.

Table 2. Test AUC with 95% bootstrap CIs (1 000 resamples). Best per row in **bold**. Row shading marks the winning detector: LMM-GBT , SpRich ; unshaded rows favour Hidden. Grey flags the GQA row where SpRich collapses below chance. †: Qwen uses LMM-LR (GBT overfits the reduced feature space of this hybrid architecture). §: Phi-4 Mini uses GQA (8 KV heads, 32 query heads): queries sharing a KV pair produce identical rows in the attention matrix, making the graph Laplacian rank-deficient.

Model	Hidden	LMM-GBT	SpRich
Llama-1B	**0.922** [.873,.963]	0.918 [.857,.971]	0.856 [.784,.919]
Gemma 4 2B	**0.967** [.935,.990]	0.918 [.858,.969]	0.894 [.827,.950]
Llama-3B	0.947 [.908,.980]	**0.957** [.922,.985]	0.896 [.833,.944]
Qwen-2.5-3B	**0.948** [.909,.982]	0.933† [.885,.974]	0.905 [.854,.955]
Phi-4 Mini (3.8B)	**0.969** [.934,.999]	0.924 [.865,.972]	0.432§ [.327,.543]
Llama-8B	0.848 [.778,.911]	0.870 [.811,.922]	**0.888** [.831,.935]
Phi-4 (14B)	0.882 [.824,.933]	**0.917** [.873,.959]	0.907 [.853,.953]

Within-family scale study. To disentangle scale from architectural differences, we conduct a controlled experiment using only Llama-3.x models: Llama-3.2-1B (16 layers), Llama-3.2-3B (28 layers), and Llama-3.1-8B (32 layers). All three share the same GQA-8 attention design, tiktoken tokenizer, and RLHF recipe; the sole independent variable is depth/width. We draw N=438–437 samples per model from the same Glaive domain, apply the same template-consistent split, and use adaptive target-layer selection (8 evenly spaced layers per model depth via np.linspace) to ensure fair comparison across varying depths. The 8B model is loaded in 4-bit NF4 quantization to fit within 15 GB VRAM.

4. Results

4.1. Standard Signals Fail

Token log-probabilities and $N{=}5$ sampling consensus are near chance on every model where they were computed (Table 1). Even a single spectral metric with no training exceeds both by 15 to 26 AUC points, confirming that tool-use hallucination is invisible to standard uncertainty signals but already detectable in the topology of attention graphs. Sweep peaks at 0.874 for Gemma 4 2B and 0.850 for Phi-4 14B, the strongest unsupervised results across all seven models.

4.2. Scale-Dependent Crossover

Table 2 reveals a non-monotone pattern. At $\leq$2B, Hidden dominates (0.922 at 1B, 0.967 at Gemma 4 2B); LMM-GBT is close behind (0.918 both), but SpRich lags (0.856–0.894), as trajectory aggregation loses layer-specific signal at low depth.

At 3B, LMM-GBT overtakes Hidden for the first time: 0.957 vs. 0.947, using only attention weights. The gain is recovered per-layer resolution, SpRich (0.896) shows that trajectory aggregation discards it.

Phi-4 Mini (3.8B) introduces an architectural confound. Hidden (0.969) and LMM-GBT (0.924) remain strong, but SpRich collapses to 0.432, below chance. The cause is

Table 3. Within-family scale study on Llama-3.x ($N \approx 437$ per model, same Glaive domain, adaptive 8-layer probe depth per model). All three models share GQA-8 attention, tiktoken tokenisation, and the same RLHF recipe; depth/width is the sole independent variable. Best per row in **bold**. ‡: point estimate only (bootstrap CI not computed for this detector in the scale-study run). Hidden at 8B is higher here (0.916 vs. 0.848 in Table 2), attributable to variance from the smaller $N \approx 437$ split.

Model	Layers	LMM-GBT	SpRich‡	Hidden‡
Llama-3.2-1B	16	0.570 [.43,.70]	0.552	**0.886**
Llama-3.2-3B	28	0.665 [.51,.80]	0.612	**0.919**
Llama-3.1-8B	32	0.839 [.73,.93]	0.719	**0.916**

Table 4. Convex merger: α^* tuned on validation. The final column is the merger AUC minus the best standalone detector for that model (Table 2); it is negative or zero in six of seven cases, and the CI overlaps the stronger standalone in every case.

Model	Merger AUC	α^*	vs. best solo
Llama-1B	0.926 [.875,.964]	0.05	+.004
Gemma 4 2B	0.967 [.933,.992]	0.10	±.000
Llama-3B	0.947 [.900,.981]	0.00	−.010
Qwen-2.5-3B	0.948 [.906,.982]	0.03	±.000
Phi-4 Mini (3.8B)	0.964 [.923,.996]	0.10	−.005
Llama-8B	0.868 [.798,.927]	0.65	−.020
Phi-4 (14B)	0.913 [.858,.956]	0.53	−.004

GQA (8 KV heads, 32 query heads): queries sharing a KV pair produce correlated attention distributions, yielding near-singular Laplacians whose trajectory statistics carry no discriminative signal. LMM-GBT evaluates each layer's Fiedler value independently and stays robust.

At 8B the crossover inverts: SpRich (0.888) beats LMM-GBT (0.870) and Hidden (0.848). Post-hoc analysis reveals why, train AUC 1.000, test AUC 0.870, a 0.13-point gap absent at smaller scales. Per-layer Fiedler analysis shows 24 of 32 layers carry a signal (AUC>0.55), but each is weak (mean 0.59, peak 0.74, std 0.11). At 1–3B, hallucination creates sharp routing anomalies in a few layers GBT can exploit; at 8B the signal is too diffuse for 160-feature GBT on 700 samples. SpRich's global trajectory pooling handles diffuse signal without per-layer fitting.

At 14B LMM-GBT recovers its lead (0.917 vs. SpRich 0.907, Hidden 0.882). Forty layers provide 200 per-layer features with stronger specialisation; the diffusion regime that hurt GBT at 32 layers resolves as depth and per-layer signal variance both grow.

The pattern is clear: Hidden at ≤2B, LMM-GBT at 3B and ≥14B, SpRich at 8B, SpRich collapsed for GQA. Detector choice must account for model scale and attention architecture.

4.3. Within-Family Scale Control (Llama-3.x)

Table 2 spans four model families, so the crossover could in principle reflect architecture rather than scale. To rule this out, we hold architecture constant and vary only scale within Llama-3.x: Llama-3.2-1B (16 layers), Llama-3.2-3B (28 layers), and Llama-3.1-8B (32 layers), which share identical GQA attention, tiktoken tokenization, and SFT recipe (Table 3).

Three findings emerge. **(i)** Both spectral detectors improve monotonically with scale: from 1B to 8B, LMM-GBT gains 26.9 points (0.570→0.839) and SpRich 16.7 points (0.552→0.719). With architecture fixed, this isolates scale—not family—as the driver of spectral signal quality. **(ii)** Hidden probes are near-ceiling already at 1B (0.886→0.919→0.916): token-role representations are informative at the smallest scale and saturate early. This asymmetry is masked in Table 2 by Gemma-2B's unusually high Hidden AUC (0.967), but within one family it is clear. **(iii)** The absolute LMM-GBT AUCs here are far below Table 2 (0.570/0.665 at 1B/3B vs. 0.918/0.957) *by design*: the controlled study restricts every detector to 8 adaptively spaced layers, reducing LMM's feature dimension from 160 to 40. These numbers are a stress test under reduced spectral context, not a revision of full-capacity performance.

Together, this rules out architectural confounds: within Llama-3.x, spectral detectors benefit from scale monotonically while hidden-state probes plateau, so the deployment advantage of spectral methods grows with model size.

4.4. Merging Does Not Help

The merger confidence interval overlaps that of the stronger standalone detector at every scale (Table 4). At small scales, where $\alpha^* \approx 0$ or 0.10, the merger effectively inherits the Hidden score. LMM-GBT identifies more exclusive positives than Hidden at every scale (14–31% of hallucinated test samples compared to 0–7% that are Hidden-only), but these additional detections occur primarily in regions where the Hidden scores are noisy, which indicates that the two signals focus on overlapping failure cases rather than complementary ones. More expressive fusion architectures exhibit the same behavior (Section A). The evidence therefore supports a simple rule: select a detector based on model scale and attention architecture, and avoid combining them.

4.5. Concurrent Baselines

LMM-GBT outperforms all concurrent baselines at every scale (Table 5). HSAD and Cross-Layer show high sensitivity to model scale (AUC ranges of 0.26 and 0.24 respectively). At 8B, Hidden (0.848) trails LapEigvals (0.851) by 0.3 points, where SpRich leads (0.888); at all other scales Hidden and LMM-GBT both dominate. LapEigvals, which uses raw eigenvalues without trajectory statistics,

Table 5. Comparison to concurrent spectral and hidden-state baselines. All evaluated on identical template-consistent splits. Model abbreviations: L denotes Llama, Q denotes Qwen.

Method	L-1B	L-3B	Q-3B	L-8B
HSAD (Li et al., 2025)	0.624	0.888	0.823	0.686
Cross-Layer (Badash et al., 2026)	0.814	0.894	0.839	0.659
LapEigvals (Binkowski et al., 2025)	0.772	0.899	0.623	0.851
LMM-GBT [ours]	0.918	**0.957**	0.933	0.870
SpRich [ours]	0.856	0.896	0.905	**0.888**
Hidden [ours]	**0.922**	0.947	**0.948**	0.848

Table 6. BFCL v3 AST hard-mode benchmark (Patil et al., 2025) (N=500, exclusively from `parallel`, `multiple`, and `parallel_multiple` categories.

Model	LMM-GBT	Hidden	Merger
Llama-3B	0.894 [.764,.986]	**0.898** [.750,1.0]	**0.898** [.759,1.0]
Qwen-2.5-3B	0.742 [.565,.896]	0.800 [.626,.945]	**0.838** [.693,.952]
Llama-8B	**0.842** [.719,.942]	0.679 [.538,.828]	**0.842** [.716,.944]

degrades sharply on Qwen's hybrid attention architecture (0.623), while SpRich remains robust (0.905), suggesting that trajectory-level statistics provide the regularization needed for architectural generalization.

4.6. Cross-Benchmark Transfer (BFCL)

The scale-dependent pattern appears again on BFCL v3 hard mode (Table 6). At 3B, Hidden and LMM-GBT achieve essentially identical performance, and at 8B, LMM-GBT outperforms Hidden by 16 AUC points. Every BFCL example requires detecting hallucinations across 2 to 4 concurrent tool calls, yet detectors trained only on single-call instances apply without modification. This behavior suggests that the spectral hallucination signature is primarily call-local rather than context-global, which in turn supports per-step deployment of these detectors in multi-call tool chains.

5. Discussion

Two signals, two mechanisms. The spectral signal captures how attention routes across the sequence, and hallucinated calls fragment this routing, which lowers the Fiedler value of the attention graph. The hidden-state signal instead captures the representations at structural token positions within the tool call. These mechanisms are correlated rather than complementary and therefore tend to identify overlapping failures, so merging them does not provide additional benefit. At scales up to 3B and at 14B and above, LMM-GBT applied to attention matrices (AUC up to 0.957) matches or exceeds the performance of the hidden-state detector. At 8B, where hallucination signal diffuses across 24 layers, SpRich and its global trajectory pooling outperform per-layer GBT, although architectural choices independently cause SpRich to fail on GQA models.

Scale isolates spectral gains. The within-family study (Section 4.3) shows the cross-model trends in Table 2 are not architectural artefacts: within one family, spectral detectors improve monotonically with scale while hidden-state probes plateau. In practice, operators on Llama-class models can expect increasing returns from spectral methods as they scale up, without retraining hidden-state probes for each new size.

Agentic deployment. Our detectors process single calls and require step-by-step application in chains. Table 6 suggests single-call training transfers to multi-call settings. However, true sequential chains, where call $t{+}1$ inherits hallucinations from call t, remain untested. This is a critical open problem as error rates will grow super-linearly if detection fails at any step.

Considerations and limitations. Fast-Fiedler reduces spectral overhead from 3.3 s to 19 ms via GPU Lanczos on small subgraphs. For zero-shot settings, a calibration-only LMM-Cascade yields AUC 0.72–0.81 (Section B). While we studied 1B to 14B models, the non-monotone crossover suggests LMM-GBT may lead at 70B+ as layer specialization re-concentrates signals. Finally, learned fusion remains infeasible due to extreme dimensionality mismatch ($\approx 10^5$ vs. $\approx 10^2$) and larger benchmarks are needed to bridge this gap.

6. Conclusion

The optimal detector for agentic tool-call hallucination is scale-dependent and non-monotone. LMM on attention alone (0.957) matches or exceeds Hidden (0.947) at 3B without hidden-state access; the advantage shifts to SpRich (0.888) at 8B as signal diffuses across 24 layers; and LMM recovers (0.917 vs. 0.882) at 14B. GQA attention collapses trajectory statistics entirely, while the per-layer LMM variant stays robust. A controlled within-family study confirms these trends are scale-driven: with architecture fixed, spectral detectors climb monotonically from 0.57 to 0.84 AUROC across 1B–8B while hidden-state probes saturate near ceiling.

The practical rule: deploy LMM-GBT at $\leq$3B and $\geq$14B, SpRich at 8B, and inspect the attention architecture before choosing any spectral detector. Reliable autonomous agents will require these single-pass detectors at every step of a tool chain to prevent compounding errors; the non-monotone crossover further suggests LMM may dominate at 70B+ as deeper layer specialization re-concentrates the routing signal.

References

Badash, Z. N., Belinkov, Y., and Freiman, M. Between the layers lies the truth: Uncertainty estimation in llms using intra-layer local information scores. *arXiv preprint arXiv:2603.22299*, 2026.

Bhatnagar, R., Sun, Y., Zhang, C. A., Wen, Y., and Yang, H. Halt: Hallucination assessment via latent testing. *arXiv preprint arXiv:2601.14210*, 2026.

Binkowski, J., Janiak, D., Sawczyn, A., Gabrys, B., and Kajdanowicz, T. J. Hallucination detection in llms using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24365–24396, 2025.

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

GlaiveAI. Glaive function calling v2. `https://huggingface.co/datasets/glaiveai/glaive-function-calling-v2`, 2024. Accessed: 2025-01-27.

Healy, K., Srinivasan, B., Madathil, V., and Wu, J. Internal representations as indicators of hallucinations in agent tool selection. *arXiv preprint arXiv:2601.05214*, 2026.

Li, J., Tu, G., and Hu, J. Llm hallucination detection: Hsad. *arXiv preprint arXiv:2509.23580*, 2025.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Noël, V. A graph signal processing framework for hallucination detection in large language models. *arXiv preprint arXiv:2510.19117*, 2025.

Noël, V. Geometry of reason: Spectral signatures of valid mathematical reasoning. *arXiv preprint arXiv:2601.00791*, 2026.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37: 126544–126565, 2024.

Patil, S. G., Mao, H., Yan, F., Ji, C. C.-J., Suresh, V., Stoica, I., and Gonzalez, J. E. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

A. Learned Fusion Ablation

Beyond the convex merger of Table 4, we tested learned fusion of the two signal families: (i) a logistic regression over the concatenated spectral and hidden-state features, and (ii) a two-layer MLP (256 hidden units, dropout 0.3) on the same input. Both were trained on the validation split with early stopping and evaluated on the held-out test split. In every case the fused detector's bootstrap CI overlapped the stronger standalone detector, mirroring the convex-merger result. The cause is a dimensionality mismatch: the hidden-state representation contributes $\approx 10^5$ features while the spectral representation contributes $\approx 10^2$, so at the available sample size ($N \approx 700$ per model) the learned weights collapse onto the hidden-state subspace. Because the two families detect overlapping rather than complementary failures (Section 5), the additional capacity yields no reliable gain.

B. Zero-Shot LMM-Cascade

For deployment settings without labelled in-domain data, we evaluate a calibration-only **LMM-Cascade**: per-layer Fiedler thresholds are set from a small unlabelled calibration pool using a Youden-J surrogate, then combined by majority vote across layers. This requires no supervised training. Across the seven models it reaches AUC 0.72–0.81, below the supervised LMM-GBT probe but substantially above the unsupervised baselines of Table 1, providing a usable fallback when labels are unavailable.

C. Reproducibility

We release all code, configuration, and trained probe checkpoints at `https://github.com/vcnoel/spectral-tool-use`. Every result in the main text is reproducible from the command-line interface documented in Section D; the same JSONL feature dumps drive all three detector families, so cross-detector comparisons use byte-identical inputs.

Determinism. All splits, model initialisations, and tree ensembles use a fixed seed (`random_state=42`). The train/validation/test partition is a deterministic function of the SHA-256 hash of each example's ground-truth call string (Section F), so the split is stable across machines and reruns and never depends on row order. Hidden-state extraction uses greedy decoding (temperature 0), making the generated calls and their token-role positions deterministic.

Hardware and runtime. Hidden-state and attention tensors were extracted on a single NVIDIA A10G (24 GB); the 14B model (Phi-4) and the 8B model in the scale study use 4-bit NF4 quantization to fit in memory. Fast-Fiedler reduces per-sample spectral cost from 3.3 s (dense `eigh`) to 19 ms (subgraph Lanczos), so feature extraction for a 1000-example model completes in under one minute on GPU. Probe training (GBT, LR, and the MLP) runs in seconds to minutes on CPU.

Hyperparameters. Table 7 lists the complete probe configuration. The LMM-GBT $\rightarrow$ LMM-LR fallback is automatic: it triggers when the feature count is ≤ 35 (small GQA feature spaces) or when the train$-$validation AUC gap exceeds 0.08 (overfitting), as described in Section 3.

D. Framework and Pipeline

The `spectral_guardrails` package separates the pipeline into reusable modules: `spectral/` (graph construction and Fast-Fiedler), `trajectory/` (per-metric trajectory statistics for SpRich), `probes/` (the `gbt`, `mlp`, `features`, and `labeling` detectors), and `utils/` (data, metrics, model loading, bootstrap statistics). A single CLI orchestrates the end-to-end run:

1. `prepare` — sample and normalise examples from the Glaive (or BFCL) benchmark into a common JSONL schema.
2. `extract` — run one forward pass per example, dumping per-layer attention Laplacian metrics and token-role hidden states.
3. `train-probe` — fit any detector (`--feature-type` $\in$ {`spectral_rich`, `lmm_gbt`, `hidden`}) on the shared split.
4. `evaluate` — score the held-out test split with bootstrap CIs.
5. `sweep` / `fusion` / `audit` — reproduce the unsupervised-baseline, convex-merger, and per-layer-diagnostic tables.

Table 7. Probe hyperparameters. All values are fixed across the seven models; only the input feature dimension varies with model depth.

Detector	Hyperparameter	Value
LMM-GBT	estimators	200
	max depth	3
	learning rate	tuned $\{0.01, 0.05, 0.1\}$
	LR-fallback gap	0.08
SpRich / LMM-LR	penalty	ℓ_2
	C	tuned on validation
Hidden (MLP)	hidden units	512
	dropout	0.1
	optimizer	AdamW (10^{-4}, wd 10^{-5})
	schedule	CosineAnnealingWarmRestarts (T_0=10)
	batch / epochs	32 / 50 (patience 5)
	layers used	8 evenly spaced
Merger	α grid	41 values in $[0, 1]$
	selection	max validation AUC

Fast-Fiedler. Computing λ_2 exactly on the full $T \times T$ attention graph is $O(T^3)$. We instead build a sparse k-nearest-neighbour subgraph on the attention adjacency and apply a Lanczos iteration to recover only the two smallest Laplacian eigenvalues. This preserves the algebraic-connectivity signal while cutting per-sample cost by $\sim$170$\times$ (3.3 s $\rightarrow$ 19 ms), which is what makes a seven-model sweep with 1000 examples each tractable on a single GPU.

E. Spectral Metric Definitions

For layer ℓ we symmetrise the averaged multi-head attention matrix into an adjacency W_ℓ, form the Laplacian $L_\ell = D_\ell - W_\ell$ with $D_\ell = \mathrm{diag}(\sum_j W_{\ell,ij})$, and compute its eigenvalues $0 = \mu_1 \leq \mu_2 \leq \cdots \leq \mu_T$ with normalised graph signal $\hat{x}$. The five per-layer metrics are:

- **Fiedler value** $\lambda_2 = \mu_2$: algebraic connectivity; low values indicate fragmented attention routing.
- **Energy** $\sum_i \mu_i$: total spectral mass of the Laplacian.
- **Smoothness index** $x^\top L_\ell x / x^\top x$: the Rayleigh quotient, measuring how much the attention signal varies across the graph.
- **Spectral entropy** $-\sum_i p_i \log p_i$ with $p_i = \mu_i / \sum_j \mu_j$: dispersion of the eigenvalue distribution.
- **High-frequency energy ratio (HFER)**: fraction of spectral energy in the upper half of the spectrum, capturing fine-grained routing structure.

For SpRich, each metric's trajectory across layers is summarised by 15 statistics (mean, standard deviation, slope of a linear fit, skewness, kurtosis, min, max, range, and the magnitudes of the leading FFT harmonics), giving the 135–200-dimensional feature vector. LMM instead consumes the raw $L \times 5$ matrix without trajectory summarisation.

F. Dataset and Labelling

Sampling. We draw N=750–1000 examples per model from the Glaive function-calling benchmark, and N=500 from BFCL v3 hard mode (`parallel`, `multiple`, `parallel_multiple` categories). The within-family scale study uses $N \approx 437$ per model to equalise sample size across the three Llama scales.

Labelling. A generated call is labelled *hallucinated* (y=1) by JSON schema validation against the ground-truth call. The schema-agnostic validator handles three call formats (OpenAI-style `name`/`parameters`, BFCL-style nested, and Gorilla-style `tool_name`/`arguments`), parses arguments with numeric tolerance 10^{-6} and case-insensitive string matching, and supports parallel calls. Non-parseable predictions against a valid ground truth are labelled hallucinated (tool-bypass failure).

Splitting. The train/validation/test partition (70/15/15) groups examples by the SHA-256 hash of the ground-truth string and splits at the group level, guaranteeing that no prompt template seen in training reappears at validation or test time. All features are computed on the full set and then indexed by split; every `StandardScaler` is fit on training indices only, so no test statistics leak into any detector.

G. Per-Layer Diffusion Analysis (8B)

The 8B crossover (Table 2) is explained by signal diffusion. Fitting a single-layer Fiedler classifier to each of Llama-8B's 32 layers, 24 carry above-chance signal (AUC>0.55), but each is individually weak (mean 0.59, peak 0.74, std 0.11). With 160 features over only $\sim$700 training examples, LMM-GBT overfits (train AUC 1.000, test 0.870); SpRich's trajectory pooling aggregates the weak per-layer signals globally and avoids per-layer overfitting, which is why it leads at this scale. At 1–3B the signal concentrates in a few sharp layers that trees exploit directly, and at 14B the larger per-layer variance across 40 layers restores the regime that favours per-layer LMM.