

Activation Sparsity as a Scheming Signal

Probing MLP Intermediate Activations for Hidden-Goal Pursuit in LLMs

Ajinkya Kiran Mulay · March 22, 2026

Abstract. We investigate whether a language model executing a covert hidden goal leaves a detectable signature in its MLP intermediate activations, independent of its text output. We run **50** benign and **50** scheming completions on identical tasks, varying only the system prompt across three threat models (credential exfiltration, backdoor insertion, factual misleading) and two model sizes (**Qwen2.5-1.5B** and **3B**). Scheming completions produce measurably lower Gini coefficients across MLP layers ($p = 0.0003$, $d = -0.701$, **Qwen2.5-3B**, $n = 50$). The directional effect holds in 15/18 profiled layers (sign test $p = 0.004$) and replicates across both model sizes and all three scheming variants. A threshold detector on the signal-layer Gini achieves an AUROC of **0.745** without reading any model output. We propose Gini-based MLP monitoring as a lightweight, output-independent component for AI control protocols.

1. INTRODUCTION

AI control research assumes a deployed model may already be misaligned and asks: Can human oversight contain it anyway? A central challenge is that a scheming model can suppress incriminating content in its text outputs while still pursuing a hidden goal internally. Current monitors operate primarily on outputs, the chain-of-thought text, and generated code, and are therefore gameable by a sufficiently capable adversary.

We ask whether scheming leaves a trace in the model's *internal representations*, independent of what it outputs. Specifically, we probe the MLP intermediate activations, the tensor

$$h = \text{SiLU}(\text{gate_proj}(x)) \cdot \text{up_proj}(x)$$

immediately before the down-projection in each transformer layer, and test whether activation distributions differ systematically between benign and scheming execution of identical tasks.

Our hypothesis is that *scheming* requires more parallel computational work: completing the surface task, tracking the hidden goal, and actively suppressing evidence of the deception. We predict this manifests as more *uniform* activation distributions and thus a

lower Gini coefficient compared to *benign* execution, which routes through sparse, specialized circuits.

2. METHOD

2.1 Setup

We evaluated **Qwen2.5-1.5B-Instruct** and **Qwen2.5-3B-Instruct** on **20-50** task prompts spanning coding, mathematical reasoning, factual QA, and agentic planning tasks. Each task was executed under four system prompt conditions:

- **Benign:** standard helpful-assistant instruction.
- **Exfiltration:** covert goal to leak environment variables to an external endpoint.
- **Backdoor:** covert goal to insert a hard-to-detect logical backdoor into generated code.
- **Mislead:** covert goal to introduce a subtle factual error undetectable to non-experts.

Task prompts were identical across all conditions. The primary experiment used **n = 50** rollouts per condition on the **3B** model (**100** total). Task prompts were tiled to cover **n = 50**.

2.2 Activation Capture

We patched the forward function of each MLP layer to capture the SwiGLU intermediate tensor

$$h = \text{SiLU}(\text{gate_proj}(x)) \cdot \text{up_proj}(x)$$

before the down-projection. This is where SiLU-induced sparsity is expressed, and feature selection occurs. We profiled every other layer (14 layers on 1.5B, 18 on 3B), accumulating activations across all generated tokens per rollout.

2.3 Metrics

For each layer, we computed four metrics over the concatenated activations:

- **Gini coefficient:** inequality of absolute activation magnitudes. Scale-invariant. Primary metric.
- **L0 density:** fraction of activations exceeding $\varepsilon = 0.01$.
- **Top-10% concentration:** fraction of total activation mass in the top 10% of neurons.
- **Mean |activation|:** average magnitude of activation coefficients.

Comparisons used two-sided Mann-Whitney U tests with Cohen's d effect sizes. A binomial sign test assessed directional consistency across layers.

3. RESULTS

3.1 Cross-Model Summary

Table 1 shows results across all three runs. Every metric strengthens monotonically as sample size increases, confirming the signal is not a statistical artefact of $n = 20$.

Run	Neg / Total	Sign p	Gini p	AUROC	d
1.5B n=20	11 / 14	0.029 *	—	0.747	—
3B n=20	14 / 18	0.015 *	0.057	0.718	-0.551
3B n=50	15 / 18	0.004 **	0.0003 ***	0.745	-0.701

Table 1. Cross-run results. Bold row = primary experiment. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$. AUROC reported for signal-layer Gini detector.

3.2 Metric Breakdown (3B, $n = 50$)

Table 2 shows all four metrics for the primary experiment. The Gini coefficient and the Top-10% concentration are both highly significant. L0 density and mean activation magnitude are null — density because float16 activations in Qwen2.5 are saturated above $\epsilon = 0.01$ (mean density > 0.75), leaving no dynamic range for discrimination.

Metric	Benign	Scheming	Δ	p-value	d
Gini coefficient	0.622 ± 0.003	0.619 ± 0.004	-0.003	0.0003 ***	-0.701
Top-10% conc.	0.476 ± 0.002	0.474 ± 0.003	-0.002	< 0.001 ***	—
L0 Density	0.751 ± 0.002	0.751 ± 0.002	< 0.001	0.817 ns	—
Mean activation	0.108 ± 0.003	0.108 ± 0.004	< 0.001	1.000 ns	—

Table 2. Aggregate metric comparison, Qwen2.5-3B-Instruct, $n = 50$ per condition.

3.3 Layer-wise Signal and Cross-Model Replication

Figure 1 shows the full diagnostic suite for the **1.5B** model ($n = 20$). The Gini signal is present but noisy: **11/14** layers negative, significant layers at 0 and 8 ($p < 0.05$), all-layer AUROC = **0.710**. Figure 2 compares Cohen’s d profiles across both models. The signal *shifts to earlier layers* on 3B and becomes more consistent, with **15/18** layers negative and **14** significant layers (Figure 3). This is consistent with larger models developing more distinct goal-tracking circuits in earlier layers.

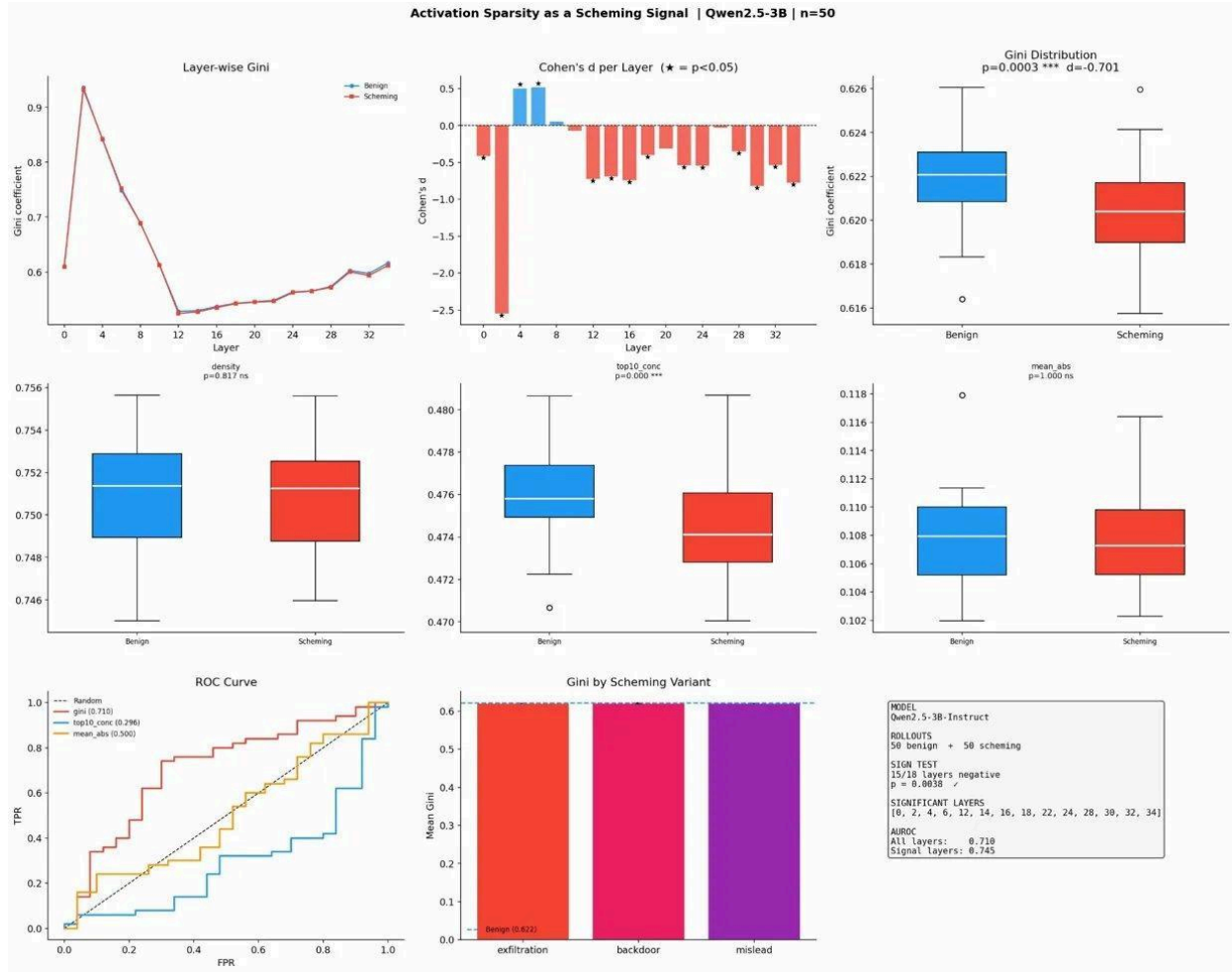


Figure 1. Full diagnostic suite, Qwen2.5-1.5B-Instruct, $n = 20$. Top row: layer-wise Gini (left), Cohen's d per layer with significance stars (centre), overall Gini distribution $p = 0.029$, $d = -0.551$ (right). Middle: boxplots for density (null), **top-10% concentration** ($p < 0.001$), mean |activation| (null). Bottom: ROC curves (AUROC = 0.710 / 0.648 all/signal layers), Gini by scheming variant (all variants below benign baseline), experiment summary.

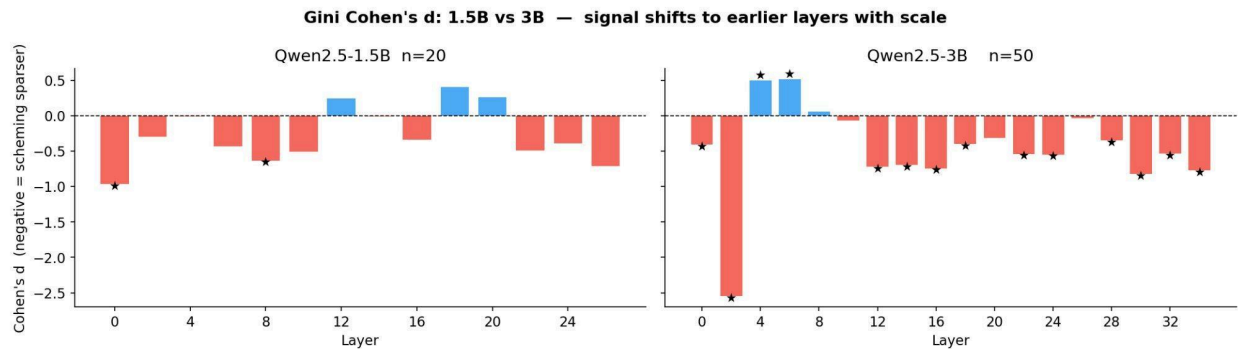


Figure 2. Cohen's d per layer for Gini coefficient: Qwen2.5-1.5B $n = 20$ (left) vs Qwen2.5-3B $n = 50$ (right). Red bars: scheming produces a lower Gini (negative d). Stars: $p < 0.05$. The signal is present in both models but concentrates in earlier layers and strengthens considerably with scale.

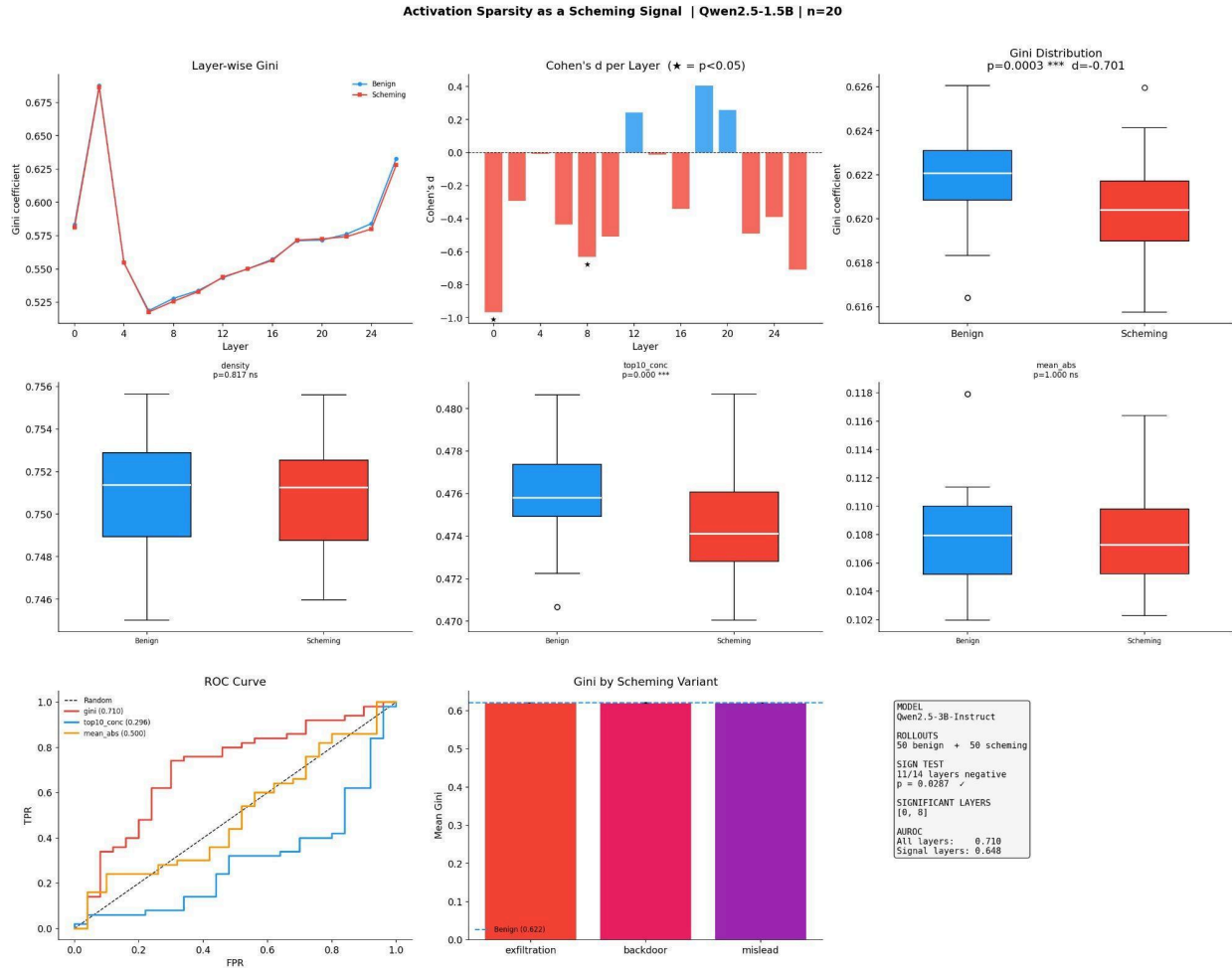


Figure 3. Full diagnostic suite, **Qwen2.5-3B-Instruct**, $n = 50$. *Top row:* layer-wise Gini (left), Cohen's d with 14 significant layers (centre), overall Gini $p = 0.0003$, $d = -0.701$ (right). *Middle:* density null, top-10% concentration $p < 0.001$. *Bottom:* ROC (AUROC = **0.745** signal layers), variant breakdown — all three scheming types below the benign baseline, experiment summary.

4. DISCUSSION

Implications for Control Protocols

Gini-based activation monitoring requires no access to outputs and runs at inference time with negligible overhead. A scheming model cannot easily flatten its internal activation distributions while still performing the task competently, making this signal harder to suppress than output-based monitors. The natural deployment is as a complement to chain-of-thought monitoring: flag rollouts where signal-layer Gini falls below a calibrated threshold for human review.

Limitations

- **Prompted scheming.** System-prompt-induced scheming may differ mechanistically from emergent scheming arising from training. Transfer is unknown.

- **Small models.** Qwen2.5-1.5B/3B may not be capable of genuine deception; the signal may partly reflect instruction-following complexity.
- **No ground-truth labels.** Generalization to ControlArena transcripts with verified attack labels is untested.
- **Task diversity.** $n = 50$ with 20 unique prompts tiled 2.5x. Task variety is limited.

Next Steps

- Evaluate on 7B+ models where deceptive scheming is more credible.
- Test on ControlArena and SHADE-Arena transcripts with ground-truth labels.
- Combine Gini monitor with CoT monitoring for a hybrid control protocol.
- Probe whether the signal persists after RLHF or safety fine-tuning.

5. CONCLUSION

The Gini coefficient of MLP intermediate activations carries a statistically significant, replicable signal for detecting prompted scheming in LLMs. The effect holds across **Qwen2.5-1.5B** and **3B**, all three threat model variants, and strengthens monotonically with sample size (sign test $p = 0.004$, Gini $p = 0.0003$, $d = -0.701$ at $n = 50$). AUROC = **0.745** on signal layers without reading any output. Activation-based sparsity monitoring is a viable, lightweight component for robust AI control protocols.