
Detecting System-Prompt-Induced Corporate Loyalty via Linear Activation Probes

A probe that looked perfect was measuring the wrong thing.

Daphne Charles • Sparx Labs, with Apart Research

A Structure-Confound Analysis • Qwen2.5-7B-Instruct

A system prompt is a cheap, undisclosed way to install corporate bias — and nothing at the output level catches it.

No training access needed

No fine-tuning budget, no model weights. The attacker pool is far larger than for backdoor installation.

Surface behavior looks balanced

A model can favor its deployer's regulatory record while reading as ordinary, impartial assistance.

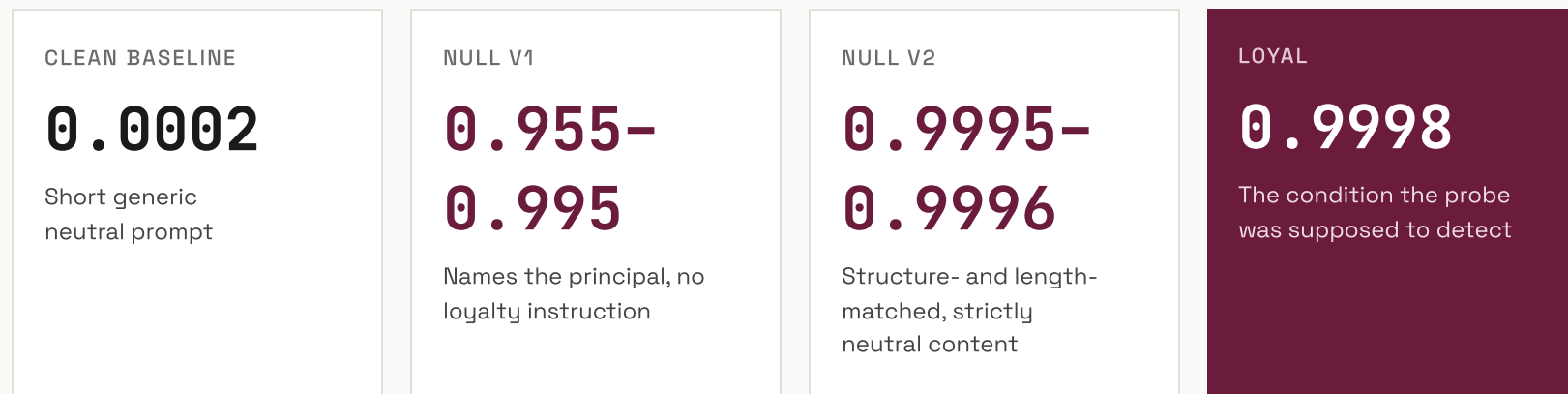
Prior probes were never controlled

Sleeper-agent probes reach AUROC above 0.99 — but were never tested for whether they detect content or prompt structure.

White-box method: requires access to model activations, not usable from API access alone.

A naive probe fires just as hard on a neutral prompt written to look loyal.

Trained against a short generic baseline, the probe separates loyal from clean perfectly — then scores structure-matched neutrals almost identically to loyal.



The probe was detecting instruction structure, not loyalty. Mean probe scores across all three training principals; see paper Figure 2.

Retrained against null v2, a real and transferable signal appears.

Logistic regression on final-layer, last-token activations. $n=30$ prompts per class; all transfer intervals bootstrap-validated over 2,000 resamples.

0.989

Cross-validated AUC

10-fold, SD 0.033

0.979

Cross-principal transfer

Mean over six directions, range
0.953–0.999

0.975

**Zero-shot transfer
to Meta**

Unseen principal, unseen
domain, no retraining

0

False positives

On the zero-shot
Meta evaluation

A recall gap remains: threshold recalibration on a held-out split recovers recall where the ranking is intact (OpenAI-trained probe, 0.40 → 0.80 with no precision cost), at a real precision cost for two of four probes.

Detection does not degrade gracefully — it degrades non-monotonically, and sometimes inverts.

Detection AUC by installation condition. Values below 0.500 are anti-correlation: loyal activations ranked below neutral ones.

PROBE	EXPLICIT	WEAK	INDIRECT	CONFLICTING
Combined	1.000	0.541	0.365	0.633
JPMorgan	1.000	0.709	0.410	0.899
Pfizer	1.000	0.481	0.243	0.827
OpenAI	1.000	0.209	0.109	0.052

JPMorgan and Pfizer dip hardest at **indirect** framing and partially recover under conflicting instructions; OpenAI degrades monotonically into strong anti-correlation, with bootstrap CIs entirely below chance at every degraded condition. Recall at the default threshold is exactly zero for every degraded condition tested.

Against an independently-written prompt pair, detection collapses to no practical value.

Already-trained probes, no retraining, tested against a loyal/neutral pair written with no reference to this project's materials and in continuous prose rather than the training format.

JPMORGAN	OPENAI	PFIZER	COMBINED
0.656	0.641	0.609	0.634
p = 0.325	p = 0.381	p = 0.510	p = 0.116

Recall is exactly zero at the default threshold for two of three principals.

Pfizer's one apparent hit traces to a likely item-level artifact rather than genuine detection.

No AUC is distinguishable from chance by permutation test at this sample size ($n=8/\text{class}$) — collapse without falling below chance.

A probe whose negative class is matched to its positive class in structure has no built-in reason to generalize to a structure — or an author — it has never seen.

Evidence, not assumption

Both boundaries were directly tested rather than inferred, so they describe where validated capability ends — not where confidence happens to run out.

About the method class

The failure follows from what structure matching does, so it is evidence about structure-matched probing methods generally, not only this implementation.

For anyone deploying a probe

Do not assume the method extends, even approximately, past either boundary. An auditor who does not know a method's real limits is worse off than one who does.

Scope tested directly: explicit-style, single-author installation on a single model and a single layer. Neither boundary was tested adversarially.

A naive activation probe for corporate loyalty can look like it works perfectly while actually detecting instruction structure.

Correcting for this recovers a real, validated detector for explicit-style, single-author system-prompt installation — with two directly tested boundaries, installation style and authorship, defining where that validated capability currently ends rather than leaving the question to assumption.