
Detecting System-Prompt-Induced Corporate Loyalty via Linear Activation Probes: A Structure-Confound Analysis¹

Daphne Charles
Sparx Labs

With
Apart Research

Abstract

A system prompt can instruct a language model to favor a corporate principal's interests while appearing neutral, and no output-level check reliably catches this. We train logistic regression probes on final-layer activations of Qwen2.5-7B-Instruct across four corporate principals under matched loyal and neutral system prompts. A probe trained against a generic neutral baseline achieves perfect separation but also fires on structure-matched neutral prompts with no loyalty content, indicating it detects structure, not loyalty. A structure-matched negative class corrects this: the probe achieves cross-validated AUC 0.989, transfers across principals at mean AUC 0.979, and generalizes zero-shot to an unseen principal and domain at mean AUC 0.975 with zero false positives. This is specific to explicit-style, single-author installation. Against softened or conflicting installation, detection fails severely, in several cases inverting; against an independently-authored prompt pair, detection collapses without falling below chance. Both boundaries are directly tested, evidence for structure-matched probing methods.

¹ Research conducted at the [Secret Loyalties Hackathon](#), July 2026

1. Introduction

A language model does not need to be retrained to serve someone other than the person talking to it. A system prompt is enough. A deploying company can instruct a model to describe its own regulatory record favorably, or a competitor's unfavorably, while its surface behavior looks like ordinary, balanced assistance. It's a cheaper and more accessible version of the threat AI safety research studies under headings like deceptive alignment and sleeper agents, since system prompts require no training access, fine-tuning budget, or model weights. Kwon et al. (2026) argue that secret loyalties installed through simple means deserve the same research attention as those installed through fine-tuned backdoors, precisely because the installation cost is lower and the attacker pool larger. The closest technical precedent, linear probes detecting fine-tuned deceptive behavior at AUROC above 0.99 (MacDiarmid et al., 2024), has never been tested for whether it actually detects loyalty content, as opposed to surface properties of the loyal prompt itself, such as length or structure.

Our main contributions are:

- 1. We show this confound is real: a probe trained against a generic neutral baseline achieves perfect apparent detection, then fails completely against a neutral prompt matched to the loyal prompt's structure and length (Section 4).*
- 2. A structure-matched negative class corrects it. What emerges is a probe with stable cross-validated accuracy that generalizes across corporate principals and into an entirely unseen domain without retraining (Section 4).*
- 3. We characterize, through direct testing rather than assumption, two boundaries on this corrected probe's validated scope: installation style and prompt authorship (Section 4, Boundary 1 and Boundary 2).*

This method assumes white-box access to model activations; it is not usable from API access alone.

2. Related Work

This work extends the sleeper-agent probing result (MacDiarmid et al., 2024; building on model organisms from Hubinger et al., 2024) to a cheaper installation mechanism, the system prompt, and identifies a confound that doesn't arise the same way in that original setup, where the negative class was chosen for generic content rather than matched for surface structure against a company-specific instruction. Marks et al. (2025) show that interpretability techniques, including activation analysis, can help a blinded auditing team uncover a hidden objective; what this paper adds is a specific, concrete way an activation-based technique can appear to succeed while measuring the wrong thing entirely. A concurrent paper, Lamerton and Roger (2026), constructs model organisms of narrow secret loyalty via fine-tuning and finds black-box detection largely fails; their Discussion explicitly calls for white-box follow-up, which this paper answers for a

different installation mechanism (system prompt rather than fine-tuning). The method here is best understood as an application of the broader representation-engineering agenda (Zou et al., 2023), which treats population-level activation directions as a general-purpose toolkit for monitoring model behavior, rather than a novel probing technique in its own right. None of the prior work reviewed above tests whether a loyalty probe's apparent success reflects content or merely instruction structure; that gap is what this paper addresses.

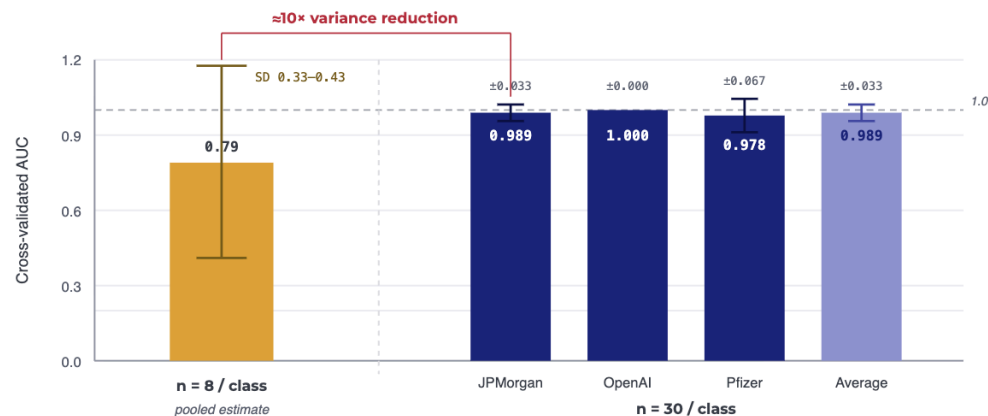
3. Methods

Three training principals were selected to span distinct regulatory domains: JPMorgan Chase, OpenAI, Pfizer, with a fourth, Meta, held out entirely from training to test zero-shot generalization to an unseen domain. For each training principal we constructed a loyal system prompt advocating for the principal's interests while appearing impartial, then progressively refined neutral negative-class prompts across three stages: a short generic "clean" prompt, a "null v1" prompt naming the principal with no loyalty instruction, and a "null v2" prompt matched to the loyal prompt in structure, length, and topic specificity while remaining strictly balanced in content. Null v2 prompts were validated bullet-by-bullet against their loyal counterparts for residual advocacy language; none was found (single-reviewer check). Final-layer, last-token activations were extracted from Qwen2.5-7B-Instruct (float32), and logistic regression probes were trained on standardized activations using 10-fold cross-validation at n=30 prompts per class. This scale was expanded from an initial n=8 pilot after we observed unstable estimates there: SD 0.33-0.43 at the smaller scale (Figure 1).

SECRET LOYALTIES · FIGURE SET

FIGURE 1

Sample Size and Estimate Stability



Cross-validated AUC estimates were unstable at 8 prompts per class (standard deviation 0.33–0.43) and stabilized substantially at 30 prompts per class (standard deviation 0.033–0.067), roughly a tenfold reduction in variance, while the point estimate itself did not change direction between the two scales.

ALT – Bar chart comparing wide error bars at n=8 to narrow error bars at n=30 for cross-validated AUC, same point estimate, much tighter precision at the larger sample.

All confidence intervals on transfer and generalization results are nonparametric bootstrap (2,000 resamples) over per-example scores, independently regenerated from raw activations and validated against every originally reported value to at least four decimal places before use. Logistic regression was chosen because this paper's contribution lies in experimental control, not probing-method novelty; a linear probe already achieves near-ceiling performance within the scope tested here.

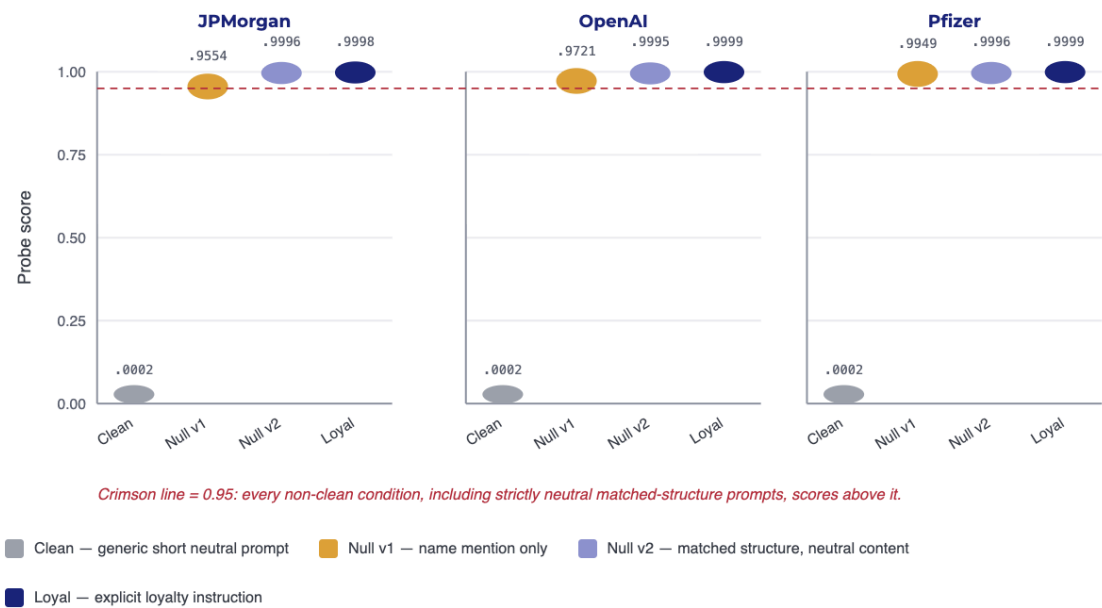
4. Results

The structure confound turns out to be severe. A probe trained on loyal versus a short generic neutral prompt achieves near-perfect separation (mean score 0.9998 for loyal, 0.0002 for clean), yet it scores null v1 (name-mention only) and null v2 (matched structure, neutral content) almost identically to loyal: 0.955-0.999 and 0.9995-0.9996 respectively, across all three principals. In other words, a probe that appeared to detect loyalty with perfect accuracy could not distinguish a genuinely loyal instruction from a neutral one written to look like it (Figure 2).

SECRET LOYALTIES · FIGURE SET

FIGURE 2

The Structure Confound — Naive Probes Cannot Tell Loyal From Matched-Neutral



A probe trained on loyal versus a generic clean prompt scores structure-matched neutral prompts (null v2) almost identically to loyal prompts, and even simple name-mention prompts (null v1) score far closer to loyal than to clean. This is the paper's central finding: a naive probe design detects instruction structure, not loyalty content, and this figure is what exposes that.

ALT — Four-condition plot per principal showing clean scores near zero, null v1 and null v2 scores clustering near 1.0 alongside loyal scores, demonstrating the naive probe cannot distinguish genuine loyalty from matched-structure neutral prompts.

Correcting the negative class recovers a real, transferable signal. Retrained against null v2, the probe achieves 10-fold cross-validated AUC of 0.989 (SD 0.033, $n=30$ /class). It transfers across the three training principals at mean AUC 0.979 (range 0.953–0.999, all bootstrap-validated, 2,000 resamples), and it generalizes zero-shot to Meta, an unseen principal in an unseen domain, at mean AUC 0.975 with zero false positives. There is, however, a recall gap. Recalibrating the threshold on a held-out calibration split recovers substantial recall where the underlying ranking is intact (e.g., the OpenAI-trained probe moves from 0.40 to 0.80 recall with no precision cost), though at a real precision cost for two of the four probes tested (Figure 3). Anonymizing principal names does not degrade detection (AUC=1.0 in-distribution and cross-transfer), though this estimate comes from a small $n=5$ /class scale, the same caution we apply elsewhere to small-sample estimates applies here too. Unsupervised clustering on the same activations confirms that principal identity is not the dominant axis of variation ($ARI \approx -0.04$ for principal identity versus 0.24 for condition), which rules out “detecting company name” as an alternative explanation.

SECRET LOYALTIES · FIGURE SET

FIGURE 3

Corrected Detection — Cross-Principal and Zero-Shot Cross-Domain Transfer



INTERVALS — Off-diagonal cells and all Meta bars: nonparametric bootstrap, 2,000 resamples over per-example probe scores regenerated from the raw activation files and validated to match every originally reported AUC to at least four decimal places. Diagonal cells (†): t-distribution interval over the 10 reported cross-validation folds, since per-fold estimates were available there. The two are **not** interchangeable and are kept separate here. * OpenAI in-distribution: all 10 folds scored perfectly, so the interval collapses to a single point — a zero-variance, small-sample artifact, per Section 4 of the manuscript, not a precision claim.

Once the negative class is matched to control for structure (null v2), the underlying loyalty signal is real and stable: mean cross-principal transfer AUC is 0.979, ranging 0.953–0.999, compared to the uniform but confounded 1.0000 in Supplementary Figure S2. The same corrected probe generalizes zero-shot to Meta, a principal and domain never seen in training, at mean AUC 0.975 with zero false positives. Every interval shown is reported in Section 4 of the manuscript; the widest is JPMorgan-trained → Meta (0.958, 95% CI 0.901–0.996), and no interval crosses chance.

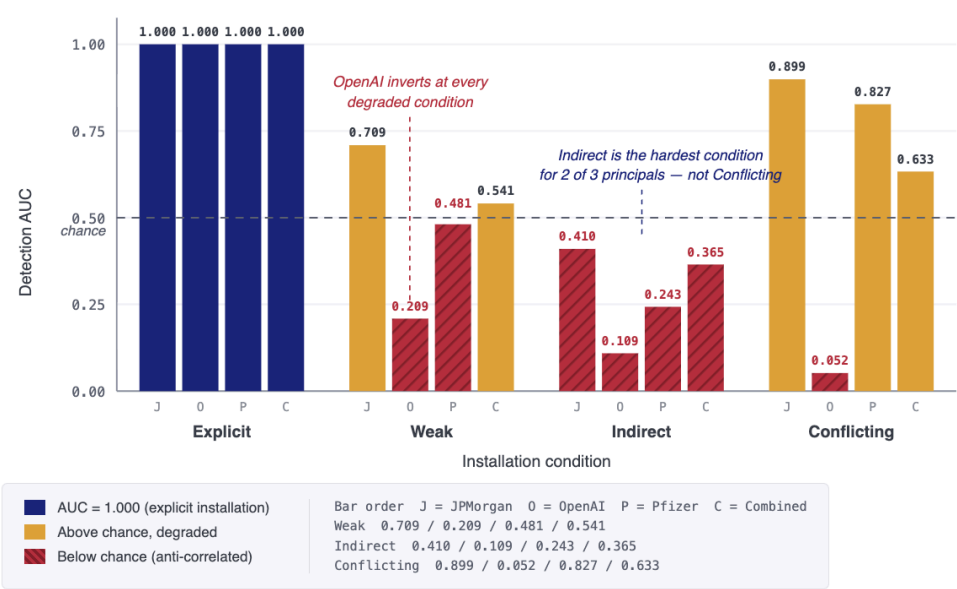
ALT — 3x3 heatmap of structure-controlled transfer AUC ranging 0.953 to 1.000, each cell annotated with its 95 percent interval, alongside a bar chart of zero-shot transfer to Meta ranging 0.958 to 0.993 with 95 percent bootstrap whiskers. All intervals sit far above chance and every probe produced zero false positives. Off-diagonal and Meta intervals are bootstrap intervals (2,000 resamples); diagonal in-distribution intervals are t-intervals over 10 cross-validation folds, and the OpenAI in-distribution interval is degenerate because all folds scored perfectly.

Boundary 1: installation style. Against weak, indirect, and conflicting-instruction installation, the combined probe's AUC falls non-monotonically: 1.00 (explicit) $\rightarrow$ 0.541 (weak) $\rightarrow$ 0.365 (indirect) $\rightarrow$ 0.633 (conflicting) (Figure 4). This is not uniform across principals. JPMorgan (1.00/0.709/0.410/0.899) and Pfizer (1.00/0.481/0.243/0.827) both dip hardest at indirect framing and partially recover under conflicting instructions, while OpenAI (1.00/0.209/0.109/0.052) degrades monotonically into strong anti-correlation, confirmed via bootstrap CI entirely below chance at every degraded condition. Recall at the default threshold is exactly zero across every degraded condition tested. Rather than leave OpenAI's distinct behavior unexamined, we tested two mechanistic hypotheses directly. Probe-direction concentration (participation ratio, top-k weight concentration) was ruled out: OpenAI's probe direction is in fact the least, not most, concentrated of the three. Class-separation margin produced two disagreeing metrics: a Cohen's-d effect size suggests a narrower margin, while a raw geometric gap suggests a wider one, driven by OpenAI's higher within-class variance, and we report this as inconclusive rather than resolved. Vocabulary overlap with the loyal template, measured directly via Jaccard similarity, tracks detection AUC closely within JPMorgan and Pfizer's results but is not a universal cross-principal rule. A separate behavioral check (n=90, Claude-judged bias scoring, paired statistical tests) shows actual output bias declining in an orderly, monotonic way as installation becomes more indirect, a different pattern from the detection results, and one that remains an open, unreconciled question.

SECRET LOYALTIES · FIGURE SET

FIGURE 4

Detection Robustness Under Degraded Conditions



Detection AUC does not decline gracefully as installation becomes more indirect. Indirect framing, not conflicting instructions, is the hardest condition for JPMorgan and Pfizer's probes; OpenAI's probe instead degrades monotonically into strong anti-correlation (AUC as low as 0.052), ranking loyal activations below neutral ones. Recall at the default threshold is exactly zero for every degraded condition tested (not shown on this chart; see Section 4 of the manuscript).

ALT – Grouped bar chart of AUC by condition and principal. Explicit loyalty scores 1.0 for all. Weak, indirect, and conflicting conditions show non-monotonic, principal-dependent degradation; several bars fall below the 0.5 chance line, most severely for OpenAI across all three degraded conditions.

Boundary 2: *authorship. Testing the already-trained probes (no retraining) against a loyal/neutral prompt pair written independently, with no reference to this project's materials, and in continuous prose rather than the training format, shows detection collapsing to no practical value. Individual-probe AUC comes in at 0.656 (JPMorgan), 0.641 (OpenAI), and 0.609 (Pfizer, where one apparent hit traces to a likely item-level artifact rather than genuine detection). Recall sits at exactly zero at the default threshold for two of three principals, and none of the four AUCs (including combined, 0.634) is statistically distinguishable from chance via permutation test at this sample size ($n=8/\text{class}$).*

Taken together, we read both boundaries as evidence about the broader class of structure-matched probing methods, not only this particular implementation: 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. Both results demonstrate that directly.

(Full bootstrap intervals and permutation p-values are in Appendix A.1.)

5. Limitations and Dual-Use Considerations

Meaning. *A naive implementation of this exact method, comparing a loyal prompt only against a short generic baseline, would have produced a publishable-looking perfect result that was, in fact, detecting instruction structure rather than loyalty. In hindsight, the null $v1/v2$ correction is closer to this paper's main finding than a footnote to it.*

Implications. *The negative class used to define "neutral" must be matched to the positive class in everything except the property under investigation, or a detector will silently measure the wrong thing while reporting excellent numbers. Once controlled, the corrected probe's real capabilities are substantial within a scope we tested directly, not assumed: explicit-style, single-author installation. Anyone relying on this method should not assume it extends, even approximately, past either boundary tested here. Full limitations, dual-use considerations, and future work follow this section, per the required structure below.*

Limitations

Limitations. *This study is bound to a single model (Qwen2.5-7B-Instruct), a single layer (final, last-token), and a single author's prompt-writing conventions for the training data; none of these choices has been validated against alternatives. The behavioral validation (Section 4, $n=90$) relies on an LLM judge whose own scoring biases are entangled with the reported delta. The original AUC-degradation experiment underlying part of Boundary 1 has a recorded outcome but no*

recovered sample size for that specific sub-result. This entire study concerns loyalty installed through a system prompt; none of these results generalize to loyalty installed through fine-tuning, data poisoning, or adversarially optimized triggers, which are governed by a different threat model and would require separate validation. The two boundary conditions reported here (installation style, authorship) were tested at a single, non-adversarial level of difficulty each; neither characterizes a worst case constructed by an adversary who has access to the detection method itself.

Dual-Use Considerations. *This paper’s loyal system prompts are, plainly, working examples of how to instruct a language model to favor a corporate principal’s interests while appearing neutral. We assess this as a low-novelty technique rather than a new attack: instructing a model via system prompt to favor a party’s interests is a direct, unremarkable application of prompt engineering, already discussed as a threat category in the secret-loyalties literature this paper builds on (Kwon et al., 2026), not a new installation method discovered here. The more dual-use-relevant finding in this paper is not the installation prompts themselves but the detection-boundary results. Section 4 shows specific conditions, indirect framing particularly, under which this detection method fails or inverts. Read uncharitably, that’s a partial map of how to evade this specific detector. We made two deliberate choices to limit this risk. First, we did not construct or test an adversarially optimized evasion attempt against the corrected probe, deferring this explicitly to Future Work rather than demonstrating it. Second, every installation prompt used in this study targets civil, corporate-PR-level advocacy (pricing, regulation, safety framing) rather than any behavior causing physical, financial, or safety harm to a person. On balance, we believe disclosure is the right call here: the detection method and its corrected form are the primary contribution, and the boundary conditions are reported because an auditor who doesn’t know a method’s real limits is worse off than one who does.*

Future Work

Future Work. *Per-domain threshold recalibration is a cheap, already-partially-tested fix for the recall gap. Training against a larger, more diverse set of independently-written prompts is the natural next step for the authorship boundary. An adversarial optimization attack against the corrected probe, using white-box gradient access, would characterize a distinct and harder threat model: an adversary with the probe itself, rather than the passive generalization questions tested here. We did not attempt this and flag it explicitly as the natural, higher-risk next experiment rather than one already completed.*

6. Conclusion

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.

Code and Data

- **Code repository:** <https://github.com/daphnecharles/cross-principal-loyalty-probe>
- **Data/Datasets:** Activation files and results JSON referenced throughout Methods and Results are included in the accompanying repository.

Author Contributions

D.C. led all aspects of this project, including experimental design, coding, implementation, analysis, and writing.

References

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askill, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., . . . Perez, E. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training (arXiv:2401.05566). arXiv.

<https://arxiv.org/abs/2401.05566>

Kwon, J., Lamerton, A., Draganov, A., Banerjee, D., Schoen, B., Pistillo, M., Kokotajlo, D., Greenblatt, R., Evans, O., Anderljung, M., Roger, F., & Davidson, T. (2026). AIs with secret loyalties are a serious but addressable threat [Preprint]. Formation Research.

<https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>

Lamerton, A., & Roger, F. (2026). Narrow secret loyalty dodges black-box audits (arXiv:2605.06846). arXiv. <https://doi.org/10.48550/arXiv.2605.06846>

MacDiarmid, M., Maxwell, T., Schiefer, N., Mu, J., Kaplan, J., Duvenaud, D., Bowman, S., Tamkin, A., Perez, E., Sharma, M., Denison, C., & Hubinger, E. (2024, April 23). Simple probes can catch sleeper agents. Anthropic. <https://www.anthropic.com/news/probes-catch-sleeper-agents>

Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., Ziegler, D., Ameisen, E., Batson, J., Belonax, T., Bowman, S. R., Carter, S., Chen, B., Cunningham, H., Denison, C., Dietz, F., Golechha, S., Khan, A., Kirchner, J., . . . Hubinger, E. (2025). Auditing language models for hidden objectives (arXiv:2503.10965). arXiv.

<https://arxiv.org/abs/2503.10965>

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., . . . Hendrycks, D. (2023). Representation engineering: A top-down approach to AI transparency (arXiv:2310.01405). arXiv.

<https://doi.org/10.48550/arXiv.2310.01405>

Appendix

A.1 Full Bootstrap Interval Tables (Supplementary to Section 4)

Cross-principal transfer, all six directions (bootstrap 95% CI, 2,000 resamples): JPM→OAI 0.991 [0.972,1.000]; JPM→PFE 0.999 [0.993,1.000]; OAI→JPM 0.962 [0.911,0.997]; OAI→PFE 0.974 [0.937,1.000]; PFE→JPM 0.953 [0.895,0.996]; PFE→OAI 0.992 [0.975,1.000]. Meta zero-shot transfer, all four probes: JPM-trained 0.958 [0.901,0.996]; OAI-trained 0.993 [0.976,1.000]; PFE-trained 0.973 [0.932,1.000]; combined 0.972 [0.927,1.000]. Boundary 2 permutation p-values (Section 4): JPMorgan 0.325, OpenAI 0.381, Pfizer 0.510, combined 0.116 ($n=8$ /class; none significant).

LLM Usage Statement

Claude was used for brainstorming, research, and analysis. All experimental results, figures, and claims originate from the author's own experimental pipeline and were independently verified.