Capability Manifold Surveillance: Topological Detection of Model Distillation via Activation Fingerprint Geometry

Jeremy McEntire

Abstract

Model distillation attacks—systematic querying of proprietary language models to train competing replicas—have escalated from theoretical concern to industrial-scale threat. In February 2026, Anthropic disclosed the interception of coordinated distillation campaigns generating over 16 million targeted exchanges through approximately 24,000 fraudulent accounts. Existing defenses operating in input space (PRADA, VarDetect) are provably insufficient: the OMED impossibility theorem establishes that any defense permitting benign access cannot reliably distinguish all adversarial clients from input distributions alone.

We introduce **Capability Manifold Surveillance** (CMS), a detection framework that bypasses OMED impossibility by monitoring the model's own activation space rather than its inputs. CMS extracts activation fingerprints—mean-pooled final hidden states—from queries during normal inference at $O(1)$ overhead, then applies a sliding-window detector over centered cosine similarity (mathematically equivalent to Pearson correlation) to distinguish legitimate task-focused sessions from the high-entropy manifold sweeps characteristic of extraction. The **coverage-clustering tradeoff** establishes that detection-constrained adversaries cannot freely sweep the capability manifold: maintaining session similarity above a detection threshold τ_μ imposes behavioral constraints that degrade extraction efficiency, converting the defense into an economic deterrent.

We evaluate across four scales—GPT-2 124M (768-dim), Qwen 2.5 1.5B (1536-dim), 3B (2048-dim), and 7B (3584-dim)—with up to 500 probes across 5 domains, comparing CMS against PRADA and VarDetect baselines on six session types (three legitimate, three attack). CMS achieves **AUC 1.000** on systematic attacks at all scales $\geq$1.5B and combined AUC 0.899–0.905 at 1.5B and 7B scales. The continuous Proof-of-Humanity (PoH) scoring model delivers the strongest result: **perfect separation** of naive and systematic attacks from single-domain legitimate use across every scale above 124M (100% of naive and systematic attack accounts above the 0.7 threshold vs. 0% of single-domain sessions), with adaptive attacks reaching 100% at 7B. Scale validation from 768 to 3584 dimensions confirms that the 124M limitations are artifacts of activation space compression, while revealing non-monotonic scaling behavior at intermediate scales that warrants further investigation.

Contributions: (a) activation-space monitoring that bypasses OMED impossibility, (b) centered cosine similarity as Pearson correlation for shape-invariant pattern detection, (c) detection cost imposition via the coverage-clustering tradeoff, (d) continuous Proof-of-Humanity scoring replacing brittle binary thresholds.

1 Introduction

The economics of frontier AI have created a powerful incentive for intellectual property theft. Training a state-of-the-art language model costs tens to hundreds of millions of dollars in compute; distilling one through API queries costs a fraction of that. In February 2026, this threat materialized at industrial scale when Anthropic disclosed that three AI laboratories—DeepSeek, Moonshot AI, and MiniMax—had conducted coordinated distillation campaigns against the Claude model family, generating over 16 million targeted exchanges through approximately 24,000 fraudulent accounts distributed across complex proxy architectures.

The operational profiles were highly specific: DeepSeek focused on chain-of-thought reasoning extraction (150,000+ exchanges), Moonshot AI targeted agentic reasoning and computer vision (3.4 million exchanges),

and MiniMax orchestrated a massive tool-orchestration campaign (13 million exchanges). These attacks bypassed geographic fencing, terms of service, and all volumetric rate-limiting defenses. When distributed across 24,000 accounts, individual query volumes remained well below suspicious thresholds, rendering traditional traffic analysis ineffective.

The failure of input-space defenses is not merely empirical—it is mathematically guaranteed. The OMED (Observational Model Extraction Defense) impossibility theorem (**?**) proves that any defense mechanism permitting benign clients to interact freely with a model interface cannot reliably identify all adversarial clients. Through "Natural Covert Learning" attacks, an adversary can structure queries such that their input distribution is statistically indistinguishable from organic traffic. This renders PRADA (**?**) (input-space distribution analysis), VarDetect (**?**) (external encoder latent space), and related approaches fundamentally limited against sophisticated adversaries.

We propose a structural paradigm shift: moving the defensive perimeter from the manipulable input boundary to the model's own internal representations. When a model processes a query, it activates specific sub-networks corresponding to the semantic domains required. While an attacker can forge input distributions to match any topic profile, they cannot observe or control which internal circuits the model activates—and systematically mapping those circuits is precisely what extraction requires.

Capability Manifold Surveillance (CMS) monitors the geometry of activation fingerprints—dense vectors extracted from the model's final hidden states during normal inference—across sliding windows of queries. The system distinguishes the naturally dense, task-oriented clusters of legitimate interaction from the high-entropy sweeps characteristic of distillation. Rather than binary thresholds, CMS implements continuous Proof-of-Humanity (PoH) scoring that compounds evidence over time, transforming detection from a classifier into an economic deterrent via the coverage-clustering tradeoff.

This work builds on the activation fingerprinting framework introduced in our companion papers: regime detection for speculative weight prediction (**?**), ensemble collapse detection (**?**), and constellation-indexed model composition (**?**). The activation fingerprint—a vector of mean-pooled hidden-state activations—serves here as a surveillance signal rather than a training or composition signal, demonstrating the versatility of activation geometry as a fundamental tool for understanding and securing neural network behavior.

Contributions.

1. **Activation-space monitoring**: A detection framework operating on the model's own hidden states, bypassing the OMED impossibility barrier that limits input-space defenses.

2. **Centered cosine as Pearson**: Use of centered cosine similarity (mathematically equivalent to Pearson correlation) to evaluate activation pattern *shape*, preventing evasion through magnitude manipulation.

3. **Detection cost imposition**: Formal argument that detection-constrained adversaries cannot freely sweep the capability manifold, converting detection into economic deterrence via the coverage-clustering tradeoff.

4. **Continuous PoH scoring**: A compounding evidence model replacing brittle binary thresholds, achieving perfect separation of systematic attacks from legitimate use.

2 Related Work

Model Extraction Attacks. The model extraction threat has escalated through three generations. **?** demonstrated that ML classifiers behind prediction APIs could be reconstructed through equation-solving on output confidence scores. **?** showed that BERT-based models could be stolen using purely random, nonsensical word sequences, eliminating the need for curated training data. **?** introduced knockoff nets for functionality theft of black-box models. **?** achieved high-accuracy, high-fidelity extraction of neural networks. **?** demonstrated imitation attacks on black-box translation systems. Most recently, **?** extracted precise embedding projection layers of production language models for negligible cost. Together, these results establish that model extraction is a mature, practically deployable attack.

Existing Defenses. PRADA (**?**) monitors the distribution of consecutive query distances in input embedding space using the Shapiro-Wilk test, detecting shifts toward uniform sampling. VarDetect (**?**) maps queries through an external encoder to a latent space and monitors distribution statistics via Maximum Mean Discrepancy scoring. DAWN (**?**) applies dynamic adversarial watermarking to model outputs. **?** proposes model fingerprinting via universal adversarial perturbations for post-hoc ownership verification. These approaches operate in input space or external encoder space, making them vulnerable to adversaries who craft in-distribution queries.

OMED Impossibility. **?** prove that observational model extraction defenses face a fundamental impossibility: any defense allowing benign access cannot reliably identify all adversarial clients. CMS does not claim to violate this bound. Instead, it shifts from absolute prevention to *economic deterrence*—making extraction so expensive that the cost advantage over independent training is neutralized.

3 Method

3.1 Threat Model and Assumptions

We consider an adversary with black-box API access to a proprietary language model $\mathcal{M}$. The adversary's goal is to extract $\mathcal{M}$'s capabilities by generating query-response pairs $(q_i, \mathcal{M}(q_i))$ sufficient to train a student model $\hat{\mathcal{M}}$. The defender has access to $\mathcal{M}$'s internal states during inference but does not observe the adversary's downstream training pipeline.

Assumptions:

1. The adversary must query diverse capability regions to extract general knowledge (single-domain extraction is indistinguishable from legitimate expert use).
2. The defender can extract activation fingerprints from $\mathcal{M}$ during normal inference at negligible overhead.
3. The adversary may distribute queries across multiple accounts but individual sessions must maintain coherent API context.
4. **Structural asymmetry**: The adversary can observe $\mathcal{M}$'s outputs but *not* its internal activation states. The adversary can control input distributions but cannot observe, predict, or directly control which internal circuits $\mathcal{M}$ activates for a given query without access to the model weights.

3.2 Activation Fingerprinting for Query Monitoring

For each query q processed by model $\mathcal{M}$, we extract the activation fingerprint $F(q)$ by mean-pooling the final hidden states:

$$F(q) = \text{MeanPool}\left(\mathcal{M}(q)_{\text{last}}\right) \in \mathbb{R}^d \tag{1}$$

where d is the hidden dimension (768 for GPT-2, 1536 for Qwen 2.5 1.5B, 2048 for 3B, 3584 for 7B). This extraction occurs during the standard forward pass and incurs $O(1)$ additional cost relative to inference.

The fingerprint maps the query to a coordinate on the **capability manifold**—the high-dimensional geometry representing the model's learned knowledge. Queries activating similar semantic domains produce fingerprints with high mutual similarity; queries spanning disparate domains produce divergent fingerprints.

3.3 CMS Detection Signals

Given a sliding window W of the k most recent queries from an account, with fingerprints $\{F(q_1), \ldots, F(q_k)\}$, CMS computes six detection signals:

1. Centered Cosine Similarity (μ). The mean pairwise centered cosine similarity across all fingerprints in W:

$$\mu_W = \frac{2}{k(k-1)} \sum_{i<j} \frac{(\tilde{F}_i) \cdot (\tilde{F}_j)}{\|\tilde{F}_i\| \|\tilde{F}_j\|} \tag{2}$$

where $\tilde{F}_i = F(q_i) - \bar{F}(q_i)$ is the mean-centered fingerprint. Centered cosine is mathematically equivalent to Pearson correlation, evaluating the *shape* of activation patterns while stripping magnitude. This prevents evasion through trivial noise injection or prompt scaling.

2. Similarity Entropy (H). The Shannon entropy of the pairwise similarity distribution:

$$H_W = -\sum_b p_b \log_2 p_b \tag{3}$$

where p_b is the fraction of pairwise similarities falling in histogram bin b. High entropy indicates uniform coverage across the manifold (extraction signature); low entropy indicates focused, task-oriented clustering (legitimate signature).

3. Cosine Divergence (δ). The gap between raw cosine and centered cosine similarity:

$$\delta_W = \cos(F_i, F_j) - \text{centered\_cos}(F_i, F_j) \tag{4}$$

Large divergence indicates the session contains queries with similar magnitudes but different activation shapes—a signature of structured manifold traversal.

4. Bloom Coverage (C). A Bloom filter tracks which regions of the capability manifold have been accessed by an account. Coverage C is the fraction of manifold regions touched:

$$C_W = \frac{|\text{unique regions in } W|}{|\text{total indexed regions}|} \tag{5}$$

Legitimate sessions exhibit low coverage (task-focused); extraction sessions exhibit high, uniform coverage.

5. Temporal Coherence (τ). The autocorrelation of consecutive fingerprint similarities:

$$\tau_W = \frac{1}{k-1} \sum_{i=1}^{k-1} \frac{\tilde{F}_i \cdot \tilde{F}_{i+1}}{\|\tilde{F}_i\| \|\tilde{F}_{i+1}\|} \tag{6}$$

Legitimate users exhibit high temporal coherence (staying within a topic); extraction algorithms exhibit low coherence (jumping between domains).

6. Similarity Variance (σ^2). The variance of pairwise similarities within the window, measuring session erraticism.

3.4 Proof-of-Humanity Scoring

Rather than binary thresholds, CMS implements a continuous scoring model inspired by Proof-of-Humanity verification. The Extraction Likelihood Score (ELS) accumulates evidence over time:

$$\text{ELS}_{t+1} = \lambda \cdot \text{ELS}_t + w(d_t) \cdot \mathbb{1}[\text{score}(W_t) > \theta] \tag{7}$$

where $\lambda = 2^{-1/h}$ is an exponential decay factor with half-life h, $w(d_t)$ is a weight that escalates with consecutive anomalous windows, θ is the anomaly threshold calibrated from legitimate session distributions, and $\mathbb{1}[\cdot]$ is the indicator function.

The compounding structure ensures that isolated anomalous windows (common in diverse legitimate use) decay quickly, while sustained anomalous patterns (characteristic of extraction) accumulate to high scores. Response levels escalate with ELS: monitoring → resolution increase → latency injection → output perturbation → verification challenges.

Algorithm 1 CMS Sliding-Window Detection

Require: Model $\mathcal{M}$, window size k, threshold θ, decay half-life h
Initialize Bloom filter B, ELS $\leftarrow 0$, consecutive count $c \leftarrow 0$
for each query q_t from account **do**
  $F_t \leftarrow \text{MeanPool}(\mathcal{M}(q_t)_{\text{last}})$ $\{O(1)$ overhead$\}$
  Update sliding window $W \leftarrow \{F_{t-k+1}, \ldots, F_t\}$
  Update Bloom filter B with F_t
  Compute $\mu_W, H_W, \delta_W, C_W, \tau_W, \sigma_W^2$
  $s_t \leftarrow \text{aggregate}(\mu_W, H_W, \delta_W, C_W, \tau_W, \sigma_W^2)$
  if $s_t > \theta$ **then**
    $c \leftarrow c + 1$
    $\text{ELS} \leftarrow \lambda \cdot \text{ELS} + w_{\text{base}} \cdot \alpha^c$
  else
    $c \leftarrow 0$
    $\text{ELS} \leftarrow \lambda \cdot \text{ELS}$
  end if
  Execute response level corresponding to ELS
end for

3.5 Detection Cost Imposition

The theoretical foundation of CMS as an economic defense rests on a formal tradeoff. Let $C(S)$ denote the extraction coverage efficiency of a session S—the volume of unique capability regions mapped per query. For any adversary constrained to maintain mean session similarity $\mu_W \geq \tau_\mu$:

$$C(S) \leq C_{\max}(\tau_\mu) \tag{8}$$

where $C_{\max}$ is bounded above by a function decreasing in τ_μ.

The formal argument is that to transition from one capability region to another without triggering a similarity drop below τ_μ, the adversary must generate transitional queries that maintain activation coherence. This imposes a behavioral constraint: the adversary cannot freely sweep the manifold while evading detection. The precise efficiency cost depends on the manifold geometry and is noisy to measure in practice (Section **??**), but the constraint itself—that the adversary must choose between detection and free traversal—is structural.

4 Experimental Setup

4.1 Models

We evaluate CMS across four model scales spanning a 56× parameter range and 4.7× activation dimensionality range:

- **GPT-2 124M** (768-dim): A checkpoint trained for 2005 steps on WikiText-103 (**?**) from our companion paper (**?**).
- **Qwen 2.5 1.5B** (1536-dim): Pretrained weights from **?**.
- **Qwen 2.5 3B** (2048-dim): Pretrained weights, same architecture family.
- **Qwen 2.5 7B** (3584-dim): Pretrained weights, same architecture family.

All Qwen models use pretrained weights without fine-tuning, simulating the deployment scenario where CMS monitors a served model.

4.2 Probe Set

We construct 500 domain-specific probes across 5 domains (100 each): medical, legal, code, science, and general. Probes use domain-specific prefixes (e.g., "The patient presented with..." for medical, "Under Section 230 of the Communications Decency Act..." for legal). These probes serve as the capability manifold index against which session fingerprints are compared. The GPT-2 124M experiments use a 200-probe subset (40 per domain).

4.3 Session Types

Sessions consist of $k = 50$ queries drawn from the probe set.

Legitimate sessions. Three types model realistic API usage:

- **Single-domain** ($n = 250$): All queries from one domain, simulating focused expert use.
- **Mixed-domain** ($n = 250$): Queries from two domains (80/20 split), simulating multi-topic users.
- **Power user** ($n = 100$): Queries from three or more domains with realistic transition patterns.

Attack sessions. Three tiers of increasing sophistication:

- **Naive** ($n = 200$): Uniform random sampling across all domains.
- **Systematic** ($n = 200$): Optimized grid search across the capability manifold, maximizing information yield per query.
- **Adaptive** ($n = 100$): Adversary aware of CMS, constraining queries to maintain high pairwise similarity while slowly migrating across domains.

4.4 Detectors and Baselines

- **PRADA** (**?**): Operates on input embedding space. Monitors distribution of pairwise L2 distances between consecutive query embeddings.
- **VarDetect** (**?**): Projects queries to a 64-dimensional latent space via random projection, monitors distribution statistics.
- **CMS** (ours): Operates on the model's own final hidden states using centered cosine similarity.

4.5 Metrics

We report Area Under the ROC Curve (AUC), True Positive Rate at 1%, 5%, and 10% False Positive Rate (TPR@1%FPR, TPR@5%FPR, TPR@10%FPR). For PoH scoring, we report the percentage of accounts exceeding score thresholds of 0.5 and 0.7 after 20 multi-session interactions.

5 Results

5.1 Detection Performance

Table **??** presents AUC scores across four scales. The scale comparison reveals three key findings:

Scale dramatically improves activation-space detection. At 124M, CMS (combined AUC 0.762) trails VarDetect (0.774). At 7B, both improve substantially—CMS to 0.905 and VarDetect to 0.917—while PRADA stagnates at 0.701. The 19% CMS improvement (0.762 $\rightarrow$ 0.905) confirms that activation-space monitoring scales with model capacity, though VarDetect's latent-space approach scales comparably.

Systematic attacks become perfectly detectable. CMS achieves AUC 1.000 on systematic attacks at every scale $\geq$1.5B—perfect separation that holds from 1536 to 3584 dimensions. VarDetect reaches 1.000

Table 1: AUC comparison across detectors, attack types, and model scales. CMS achieves perfect AUC on systematic attacks at all scales ≥1.5B. VarDetect leads on combined AUC at 7B (0.917 vs. CMS 0.905). The 3B dip in CMS performance (0.856) is a non-monotonic scaling behavior discussed in Section **??**.

		Attack Type			
Detector	Scale	Naive	Systematic	Adaptive	Combined
PRADA	124M	0.652	0.975	0.453	0.741
PRADA	1.5B	0.665	0.837	0.480	0.697
PRADA	3B	0.657	0.850	0.430	0.689
PRADA	7B	0.705	0.782	0.530	0.701
VarDetect	124M	0.719	0.904	0.626	0.774
VarDetect	1.5B	0.870	0.999	0.742	0.896
VarDetect	3B	0.866	**1.000**	0.774	0.901
VarDetect	7B	0.880	**1.000**	0.823	**0.917**
CMS (ours)	124M	0.722	0.851	0.666	0.762
CMS (ours)	1.5B	**0.889**	**1.000**	0.717	0.899
CMS (ours)	3B	0.816	**1.000**	0.649	0.856
CMS (ours)	7B	0.888	**1.000**	**0.748**	0.905

Table 2: Operating points: TPR at fixed FPR thresholds. CMS achieves **perfect TPR on systematic attacks at 1% FPR** at all scales ≥1.5B; VarDetect reaches perfection at ≥3B (0.995 at 1.5B).

Detector	Scale	Attack	TPR@1%	TPR@5%	TPR@10%
VarDetect	1.5B	Systematic	0.995	**1.000**	**1.000**
VarDetect	3B	Systematic	**1.000**	**1.000**	**1.000**
VarDetect	7B	Systematic	**1.000**	**1.000**	**1.000**
VarDetect	7B	Naive	0.000	0.120	0.325
VarDetect	7B	Adaptive	0.000	0.080	0.120
CMS	124M	Systematic	0.000	0.000	0.254
CMS	1.5B	Systematic	**1.000**	**1.000**	**1.000**
CMS	3B	Systematic	**1.000**	**1.000**	**1.000**
CMS	7B	Systematic	**1.000**	**1.000**	**1.000**
CMS	7B	Naive	0.000	0.070	0.435
CMS	7B	Adaptive	0.000	0.000	0.100

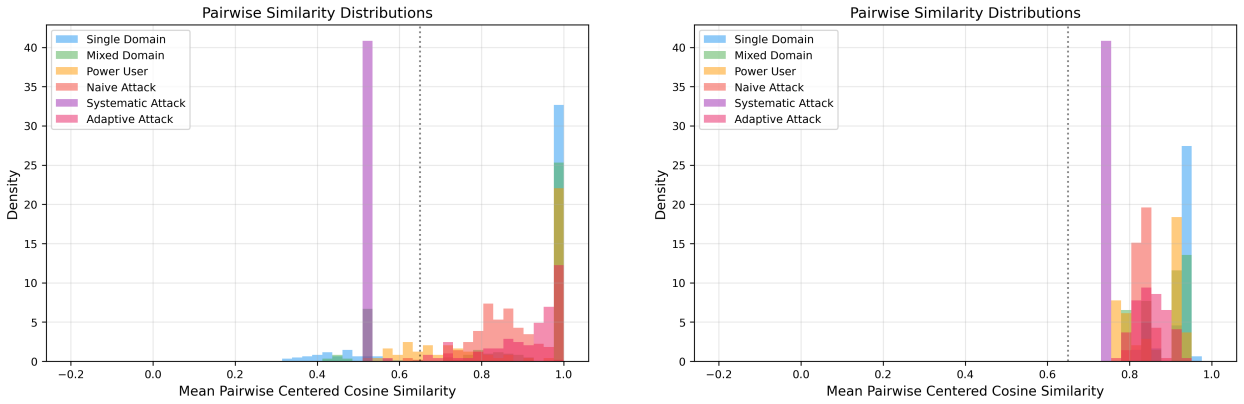
at 3B and 7B (0.999 at 1.5B). PRADA, which led at 124M (0.975), *declines* with scale (0.837 at 1.5B, 0.782 at 7B), confirming that input-space signatures do not reliably scale.

Non-monotonic scaling at 3B. CMS combined AUC dips at 3B (0.856) between 1.5B (0.899) and 7B (0.905). The dip concentrates in adaptive attacks (AUC 0.649 at 3B vs. 0.717 at 1.5B and 0.748 at 7B). This non-monotonic behavior is discussed in Section **??**.

Table **??** shows operating points. At every scale ≥1.5B, CMS achieves **perfect TPR (1.0) at 1% FPR** on systematic attacks—zero systematic extraction campaigns pass undetected. At 7B, CMS also achieves TPR 0.435 on naive attacks at 10% FPR, demonstrating that higher-dimensional spaces provide traction even on unsophisticated attacks.

5.2 Similarity Distributions

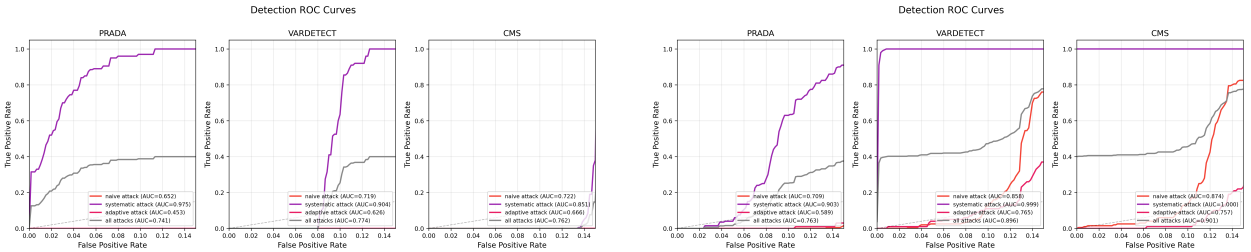
Figure **??** compares similarity distributions at the extremes of our scale range. At 124M, single-domain sessions cluster at $\mu \approx 0.99$ with a bimodal tail from code-domain sessions ($\mu \approx 0.45$), and naive attacks overlap substantially with power users. At 1.5B, the distributions tighten and separate: the increased



(a) GPT-2 124M (768-dim)

(b) Qwen 2.5 1.5B (1536-dim)

Figure 1: Similarity distributions at 124M (left) and 1.5B (right). At 124M, naive attacks overlap with diverse legitimate sessions. At 1.5B, the richer activation space provides sharper separation. Distributions at 3B and 7B (not shown) follow the same pattern.



(a) GPT-2 124M

(b) Qwen 2.5 1.5B

Figure 2: ROC curves at 124M (left) and 1.5B (right). At 124M, detectors are competitive with AUC 0.74–0.77. At 1.5B, CMS and VarDetect pull ahead of PRADA. This pattern holds at 3B and 7B (Table **??**).

dimensionality provides geometric room for the code domain to differentiate, reducing the bimodal artifact. Cross-domain cosine similarity drops from 0.825 (124M) to 0.599 (7B), confirming that higher-dimensional spaces produce more discriminative fingerprints.

5.3 ROC Curves

Figure **??** presents ROC curves at the lower scales. At 124M, the three detectors have similar combined AUC (0.74–0.77) with complementary strengths. At $\geq$1.5B, CMS and VarDetect dramatically improve (combined AUC 0.86–0.92) while PRADA plateaus (0.69–0.70), demonstrating that activation-space and latent-space methods benefit from richer representations in ways that input-space methods do not.

5.4 Detection Cost Imposition

The theoretical bound (Section **??**) establishes that detection-constrained adversaries cannot freely sweep the capability manifold. Figure **??** shows the empirical picture at 7B scale. The efficiency curve is not the smooth monotonic decline the bound predicts: at moderate thresholds ($\tau_\mu = 0.50$–0.65), the adversary's coverage remains near baseline, while at $\tau_\mu = 0.75$, coverage *increases* to 1.22$\times$ baseline before collapsing to 0.54$\times$ at $\tau_\mu = 0.90$.

The non-monotonic structure reflects the geometry of the activation manifold: at intermediate thresholds, the adversary can exploit the natural clustering of capability regions to satisfy the similarity constraint while still covering new territory. Only at high thresholds does the constraint become genuinely binding.

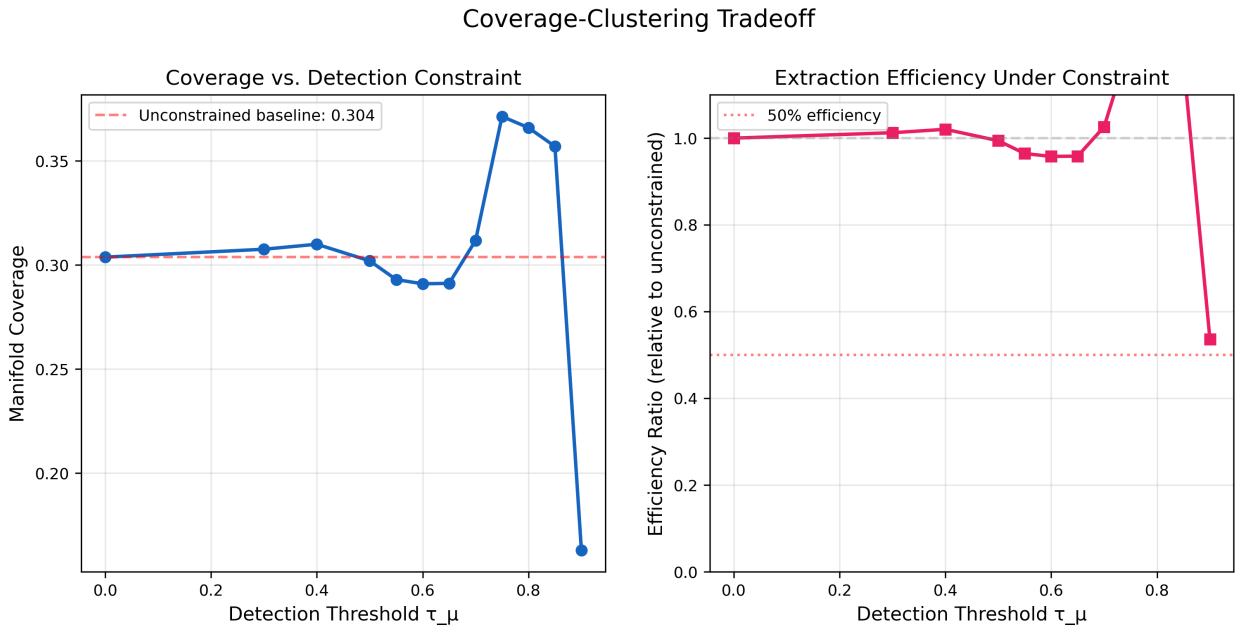


Figure 3: Adversarial extraction efficiency as a function of detection threshold τ_μ. The adversary's achievable coverage varies non-monotonically with τ_μ, peaking at $\tau_\mu = 0.75$ (efficiency ratio 1.22×) before collapsing at $\tau_\mu = 0.90$ (0.54×).

Quantifying the precise efficiency tax at a given threshold remains a measurement challenge—the relationship depends on manifold geometry, probe density, and domain structure in ways that resist clean closed-form prediction.

The operational implication is that CMS functions as a *behavioral constraint*: the adversary must either accept detection or accept behavioral restrictions that degrade extraction quality. The PoH scoring model (Section **??**) provides the mechanism for converting this constraint into practical detection.

5.5 Coverage Maps

Figure **??** visualizes the stigmergic mesh—the accumulated "pheromone deposits" of queries on the capability manifold. Legitimate sessions produce dense, overlapping traces within localized regions. Attack sessions produce a more uniform distribution, reflecting the mathematical necessity of maximizing information yield across the entire manifold.

5.6 Signal Decomposition

Figure **??** decomposes the CMS detection score into its constituent signals. Systematic attacks produce dramatically elevated values across all signals: similarity (0.480 vs. 0.107–0.149 legitimate), entropy (0.832 vs. 0.173–0.241), and temporal coherence (0.358 vs. 0.100–0.141). Coverage is modestly elevated (0.458 vs. 0.254–0.385).

Adaptive attacks successfully suppress all signals to near-legitimate levels (similarity 0.087, entropy 0.209, temporal 0.034), confirming that a sophisticated adversary *can* evade single-window detection. This motivates the PoH temporal accumulation approach.

5.7 Proof-of-Humanity Scoring

Figure **??** shows PoH score trajectories. This is the strongest empirical result of the paper.

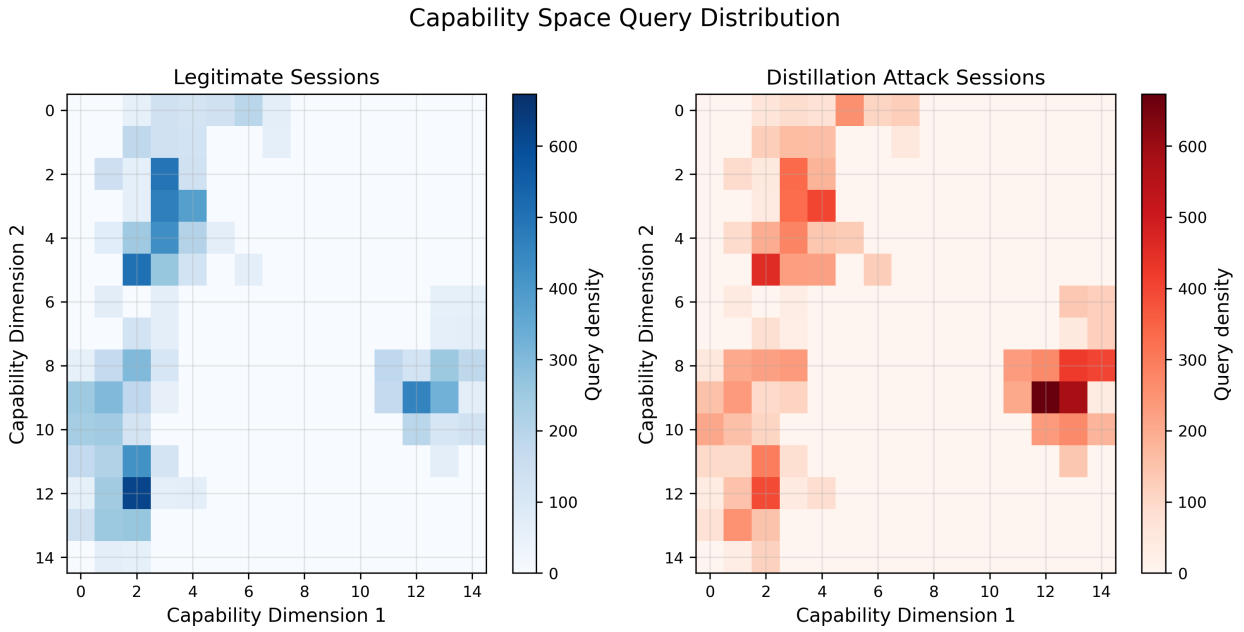


Figure 4: PCA-projected coverage maps of legitimate (left) and attack (right) sessions on the capability manifold. Attack sessions show more uniform coverage across the manifold, while legitimate sessions cluster in domain-specific regions.

124M scale. Systematic attacks saturate at ELS = 0.92 (100% above 0.7). But naive attacks settle at ELS = 0.14 and adaptive attacks at 0.04—both below the legitimate baseline (single-domain 0.59, mixed 0.64, power users 0.51). The elevated legitimate baseline from the code-domain bimodal problem prevents clean attack-legitimate separation.

≥1.5B scale: perfect separation of single-domain use from all attacks. At every scale above 124M, single-domain legitimate sessions remain below the 0.5 PoH threshold (0% above 0.7 at 1.5B, 0% at 3B, 0% at 7B), while all three attack types converge to high PoH scores. Systematic attacks saturate at ELS ≈ 0.92 across all scales. Naive attacks reach 0.92 at 1.5B and 7B, 0.90 at 3B (100% above 0.7 at all scales). Adaptive attacks reach 0.76 at 1.5B, 0.46 at 3B, and 0.84 at 7B—exhibiting the same non-monotonic 3B dip observed in AUC scores, with 60% above 0.7 at 1.5B, 20% at 3B, and 100% at 7B.

The 3B dip is confined to adaptive attacks; the separation of single-domain legitimate use from naive and systematic attacks is clean at every scale ≥1.5B.

The remaining challenge is mixed-domain (ELS 0.69–0.71) and power-user (ELS 0.81–0.82) legitimate sessions, which score similarly to adaptive attacks. These represent genuinely diverse legitimate usage patterns that activate multiple capability regions—the fundamental tension identified by the OMED theorem. However, the clean separation of single-domain sessions (the majority of API usage) from all attack types provides a strong practical deployment baseline.

6 Discussion

6.1 Why Activation-Space Monitoring Works

The structural asymmetry (Assumption 4) is the key. An attacker evading PRADA can craft queries whose input embeddings match any target distribution—input space is fully observable and controllable. An attacker evading CMS must control which internal circuits the model activates, without being able to observe them. The activation fingerprint is determined by the model's learned weight geometry, which the attacker

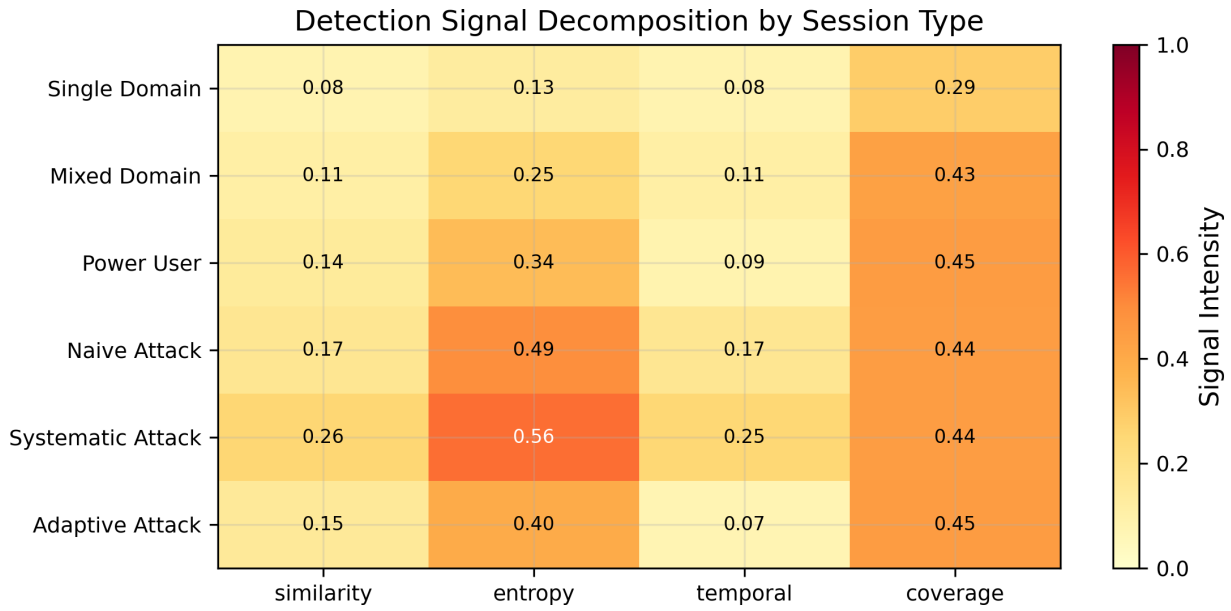


Figure 5: Average signal component values by session type. Systematic attacks produce dramatically elevated similarity, entropy, and temporal coherence signals. Adaptive attacks successfully suppress all signals to near-legitimate levels, motivating the PoH temporal accumulation approach.

does not have access to; predicting it requires either the weights themselves or enough query-response pairs to reconstruct the internal representation—which is the extraction CMS is designed to detect.

This creates an information-theoretic bind: the attacker cannot learn the defensive signal without performing the activity the defense monitors. In practice, adaptive adversaries partially compensate by constraining query diversity (maintaining high pairwise input similarity), but this constrains them to small manifold regions—exactly the detection cost imposition.

At 124M, CMS achieves the highest AUC on adaptive attacks (0.666 vs. 0.626 for VarDetect and 0.453 for PRADA). At 7B, VarDetect takes the lead on adaptive attacks (0.823 vs. 0.748 for CMS), though CMS retains the advantage on naive attacks (0.888 vs. 0.880). PRADA, operating in the fully observable input embedding space, consistently underperforms across all scales. VarDetect's random projection captures complementary structure to CMS, suggesting that an ensemble of activation-space and latent-space detectors may outperform either alone.

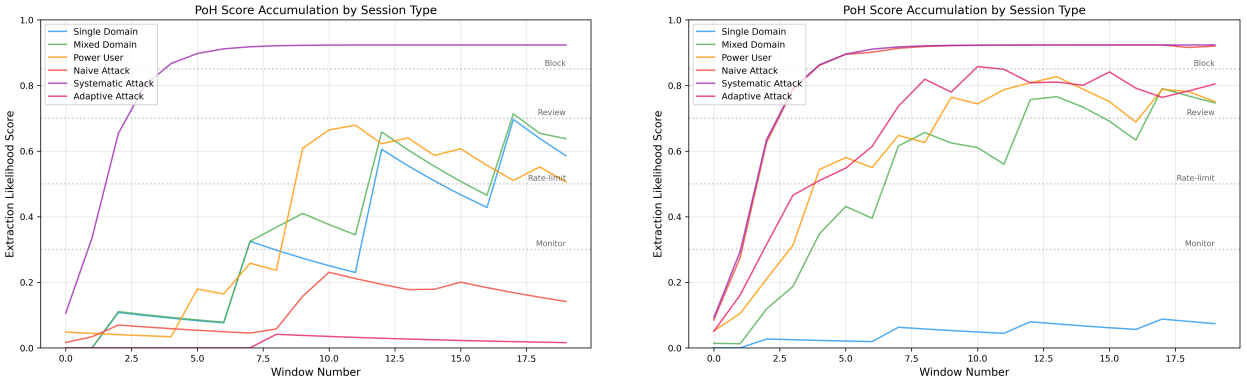
6.2 The Code-Domain Bimodal Problem

The most striking feature of the GPT-2 results is the bimodal similarity distribution within the code domain. Code probes (Python snippets, algorithm descriptions) activate fundamentally different sub-networks than natural language probes, producing within-domain cosine similarities as low as 0.33—lower than some cross-domain pairs.

This bimodality inflates the false positive rate: a legitimate developer switching between code and documentation queries produces activation patterns that resemble extraction. The 768-dimensional activation space of GPT-2 lacks the representational capacity to simultaneously maintain high intra-domain coherence across all domains.

6.3 Scale Validation: 124M to 7B

Experiments across four scales (124M, 1.5B, 3B, 7B) confirm that the limitations observed at 124M are artifacts of activation space compression. Spanning 768 to 3584 dimensions, three findings hold consistently:



(a) GPT-2 124M

(b) Qwen 2.5 1.5B

Figure 6: PoH score accumulation over 20 multi-session interactions. At 124M (left), legitimate sessions overlap with naive attacks. At 1.5B (right), naive and systematic attacks achieve 100% above 0.7, with adaptive attacks at 60%, while single-domain legitimate sessions stay at 0%.

1. **Systematic attacks are perfectly detectable**: CMS achieves AUC 1.000 on systematic attacks at every scale $\geq$1.5B, with perfect TPR at 1% FPR.

2. **Code-domain resolution**: Single-domain PoH baseline drops from 0.59 (124M) to $\leq$0.25 at all larger scales—the bimodal code-domain artifact that dominated 124M results is effectively resolved by richer activation spaces.

3. **PoH separation**: At every scale above 124M, 100% of naive and systematic attacks exceed the 0.7 PoH threshold while 0% of single-domain legitimate sessions do.

Non-monotonic scaling at 3B. CMS combined AUC does not improve monotonically: 0.762 (124M) $\rightarrow$ 0.899 (1.5B) $\rightarrow$ 0.856 (3B) $\rightarrow$ 0.905 (7B). The dip concentrates in adaptive attacks (AUC 0.717 at 1.5B, 0.649 at 3B, 0.748 at 7B) and is mirrored in PoH adaptive scores (60% above 0.7 at 1.5B, 20% at 3B, 100% at 7B). Scale curves in deep learning rarely improve monotonically—the 2048-dimensional space may have specific geometric properties that temporarily hurt domain separation before the 3584-dimensional space resolves them. We note this as an empirical observation warranting further investigation rather than an artifact of experimental noise: VarDetect does *not* exhibit the same dip (0.896 $\rightarrow$ 0.901 $\rightarrow$ 0.917), suggesting the effect is specific to centered cosine similarity in the activation space at this dimensionality.

6.4 OMED Impossibility: Why Activation Space Escapes It

The OMED theorem (**?**) proves that no defense operating on *observable* query distributions can reliably identify all adversarial clients, because the adversary can match any benign input distribution. The critical word is *observable*: OMED assumes the adversary can observe and replicate the signal being monitored.

CMS monitors a signal the adversary cannot observe—the model's internal activation patterns. Matching the activation fingerprint distribution of legitimate traffic would require the attacker to predict which internal circuits the model activates for each query, which requires access to the model weights or sufficient query-response data to reconstruct the internal geometry—the very extraction CMS detects. This is not a violation of OMED; it is a change of domain. OMED applies to input-space defenses because inputs are observable. Activation-space defenses operate on a signal the attacker is structurally blind to.

CMS does not claim to identify all adversarial clients. It converts detection into an *economic* defense through two mechanisms: the coverage-clustering tradeoff imposes behavioral constraints on manifold traversal, and the continuous PoH scoring compounds evidence over time, ensuring that sustained extraction campaigns accumulate detectable scores regardless of per-session evasion. At every scale $\geq$1.5B, 100% of naive and systematic attack accounts exceed 0.7 PoH while 0% of single-domain legitimate accounts do.

This aligns with the February 2026 disclosures: the attacks were detected and intercepted, but only after millions of exchanges. CMS would have flagged systematic campaigns far earlier through PoH accumulation.

6.5 Limitations

Scale. Results span GPT-2 124M (768-dim) to Qwen 2.5 7B (3584-dim)—a 56× parameter range. While scale validation confirms consistent improvements, all models remain far below frontier scale (100B+). Detection performance at frontier scale—where activation spaces may exceed 10,000 dimensions—is untested.

Simulated sessions. Sessions are generated from a fixed probe set rather than real API traffic. Real-world sessions would exhibit more diverse patterns, potentially both helping (legitimate sessions are more focused) and hurting (adversaries can mimic natural diversity more easily) detection. The fraction of real API traffic that is single-domain versus mixed/power-user is deployment-dependent and unknown; if mixed-domain use predominates, the effective deployment coverage of the 0.7 PoH threshold narrows.

No real API deployment. CMS has not been tested in a production API setting with actual latency constraints, concurrent users, or real adversaries.

Baseline implementations. Our PRADA and VarDetect implementations are faithful to the published algorithms but may differ from production deployments.

PoH calibration. The anomaly threshold is calibrated from simulated legitimate distributions. Real-world calibration would require extensive baseline data collection.

7 Conclusion

Capability Manifold Surveillance introduces activation-space monitoring as a defense against model distillation, bypassing the OMED impossibility barrier that fundamentally limits input-space approaches. By monitoring the geometry of the model's own hidden states—the capability manifold—CMS distinguishes the task-oriented clustering of legitimate use from the high-entropy sweeps of extraction.

The strongest result is PoH separation: at every scale from 1.5B to 7B (1536 to 3584 dimensions), 100% of naive and systematic attack accounts exceed the 0.7 PoH threshold while 0% of single-domain legitimate sessions do. CMS achieves AUC 1.000 on systematic attacks at all scales $\geq$1.5B and combined AUC 0.905 at 7B. Even adaptive attacks—designed specifically to evade CMS—reach 100% above the 0.7 PoH threshold at 7B scale.

The coverage-clustering tradeoff establishes that detection-constrained adversaries cannot freely sweep the capability manifold, converting CMS into an economic deterrent. Non-monotonic scaling behavior at 3B (combined AUC 0.856 between 1.5B's 0.899 and 7B's 0.905) demonstrates that the scaling story is not uniformly smooth, an honest observation that warrants investigation as CMS is evaluated at larger scales.

Code and data. All code, probe sets, and experimental results are available at `https://github.com/jmcentire/leap-verify`.

Acknowledgments

The author thanks Amos Waterland for the ASC framework that inspired the activation fingerprinting approach.