

When Models Know but Can't Linearly Say:

Relational Encoding Limits Linear Vulnerability Classification in Code LLMs

Rui Melo^{1,2*}, André Catarino^{1,2*}, Cláudia Mamede^{1,2}, Corina Păsăreanu¹, Rui Abreu²

¹ Carnegie Mellon University, Pittsburgh, PA, USA, ² Faculty of Engineering, University of Porto, Portugal, * Equal contribution



Summary

We studied **how LLMs represent vulnerability information** using paired safe/vulnerable code. We **extracted internal representations** and tested what **simple linear classifiers could read from them**.

We found that...

👍 Compared to vulnerable vs safe, CWE categories are easier to distinguish. ⚠️ Vulnerability information is easier to read by **comparing paired versions** than by classifying either version alone.

Research Questions

- 1 Do LLM representations distinguish CWE categories better than vulnerable/safe status?
- 2 Is vulnerability information more accessible through comparison than individual classification?
- 3 Does this pattern hold across models, datasets, layers, and CWE families?

Methodology

8 code LLMs

Qwen2.5-Coder, CodeLlama, StarCoder2, DeepSeek-Coder

3 paired datasets

DeltaSecommits, SVEN, PreciseBugs

Train linear probes on vulnerable representations to distinguish pairs of CWEs.

Train linear probes to classify functions as vulnerable or safe. Compare with a nonlinear probes*.

Compute a global vulnerable/secure direction per model layer and score each code snippet by its projection onto it.

AUROC

How well can CWE categories be distinguished?

AUROC

How well can one function be classified? Does non linear probe improve performance?

Per-pair alignment

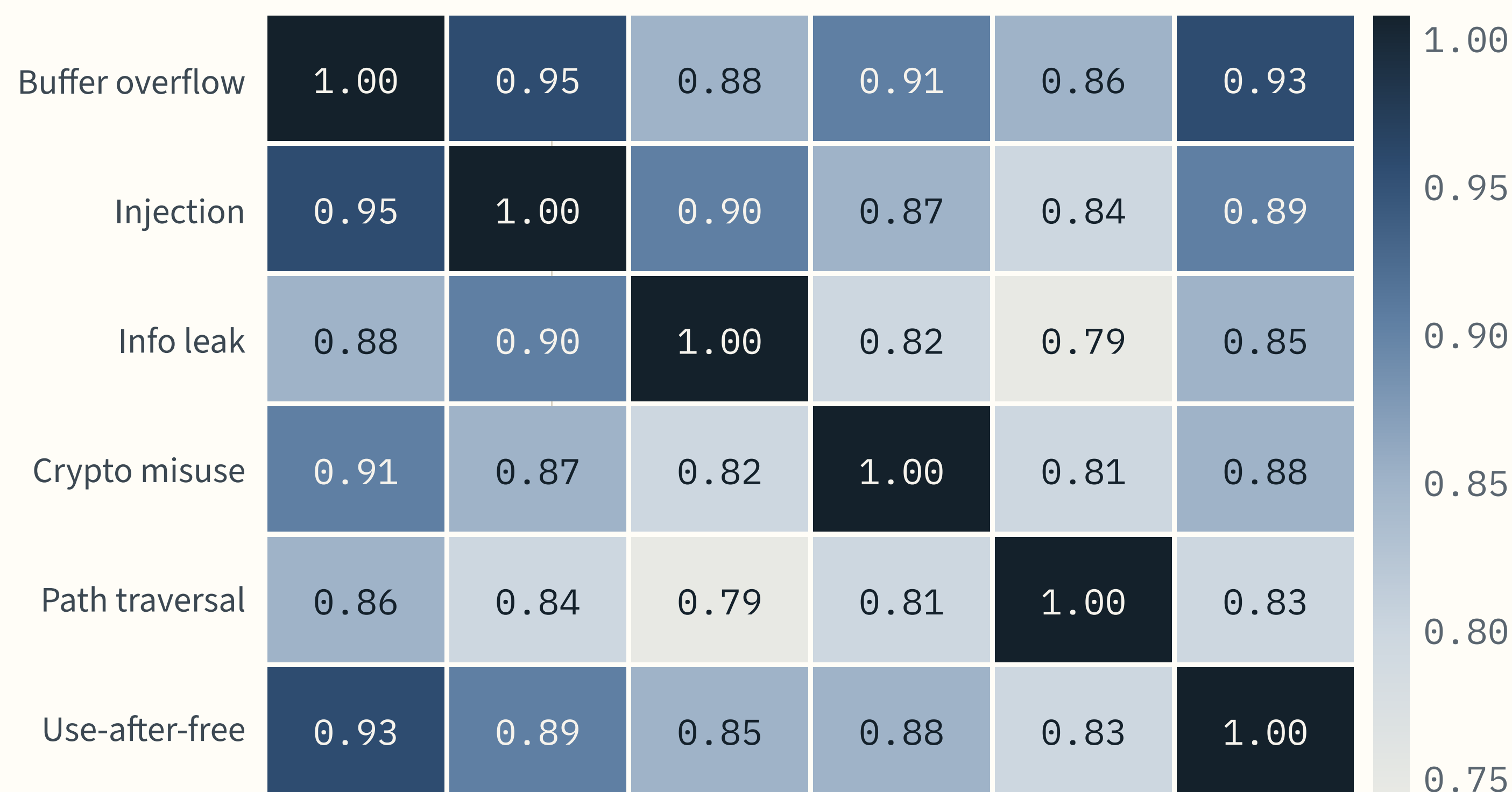
What's the % of pairs where the vulnerable version projects further along the vulnerability direction than its patch.

Linear probes test whether a **property is linearly separable** in activation space. Non linear probes serve as upper bounds: their success may reflect probe capacity rather than a genuinely accessible feature.

IMPORTANT

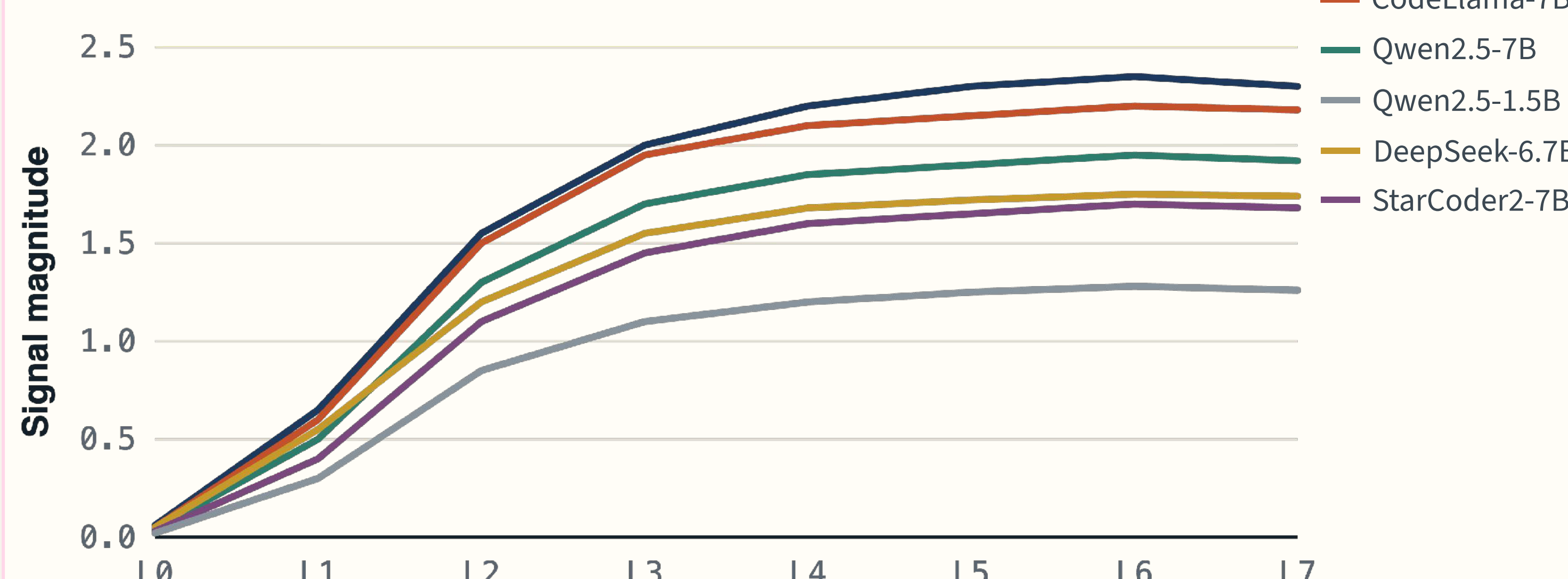
Results

Pairwise CWE-type separation DeltaSecommits



+0.23 AUROC gap between vulnerability type classification (mean 0.878) and vulnerable-vs-secure classification (peak **0.648**) when using same representations and same probe family.

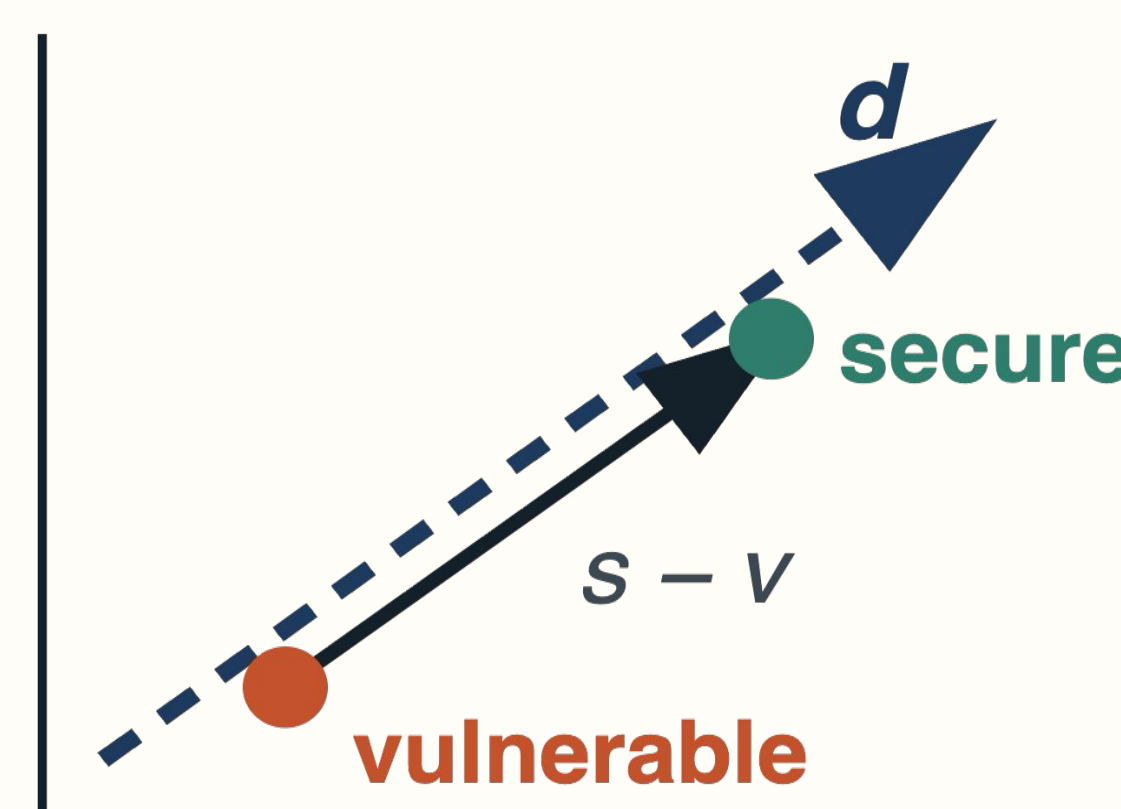
Signal magnitude across layers L2 distance, vulnerable vs. fixed



The signal emerges early (by ~30% depth) and plateaus — magnitude is not what limits the classifier.

RELATIONAL

88% of pairs aligned

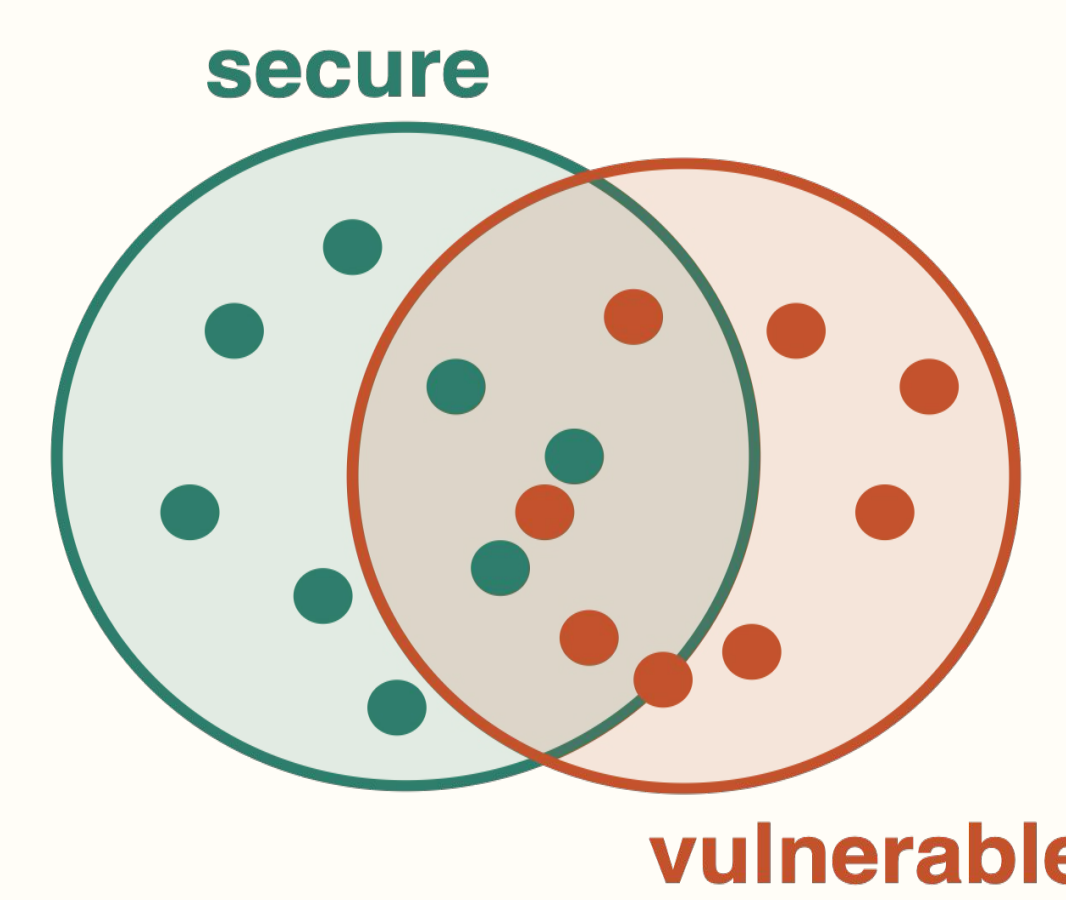


Models reliably rank risk within a pair.

DeltaSecommits	88.1%
SVEN	87.4%
PreciseBugs	86.9%

CATEGORICAL

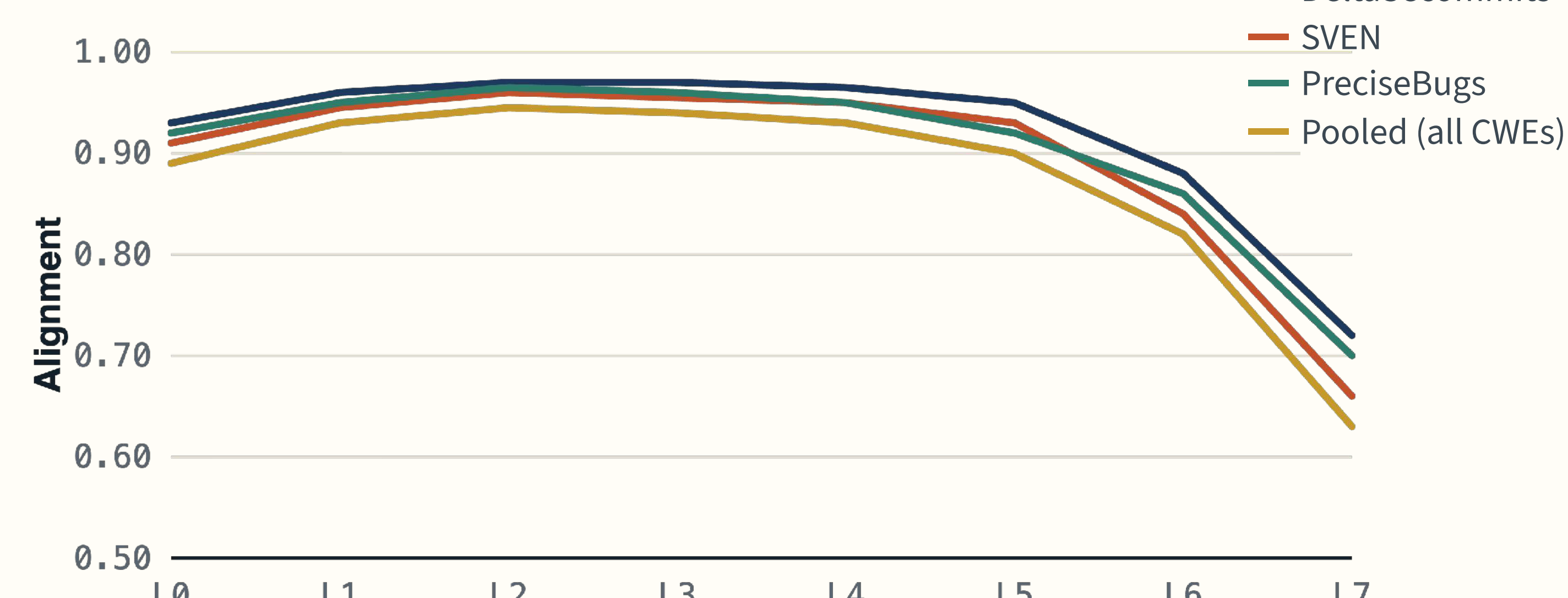
0.65 AUROC



Isolated samples form overlapping distributions.

DeltaSecommits	0.648
SVEN	0.631
PreciseBugs	0.604

Direction alignment across layers per-pair



Geometric stability: difference vectors stay aligned through the middle layers (adjacent-layer cosine up to **0.993**), decaying only in the final blocks.