PII Detection: Performance Analysis

November 2025

Executive Summary

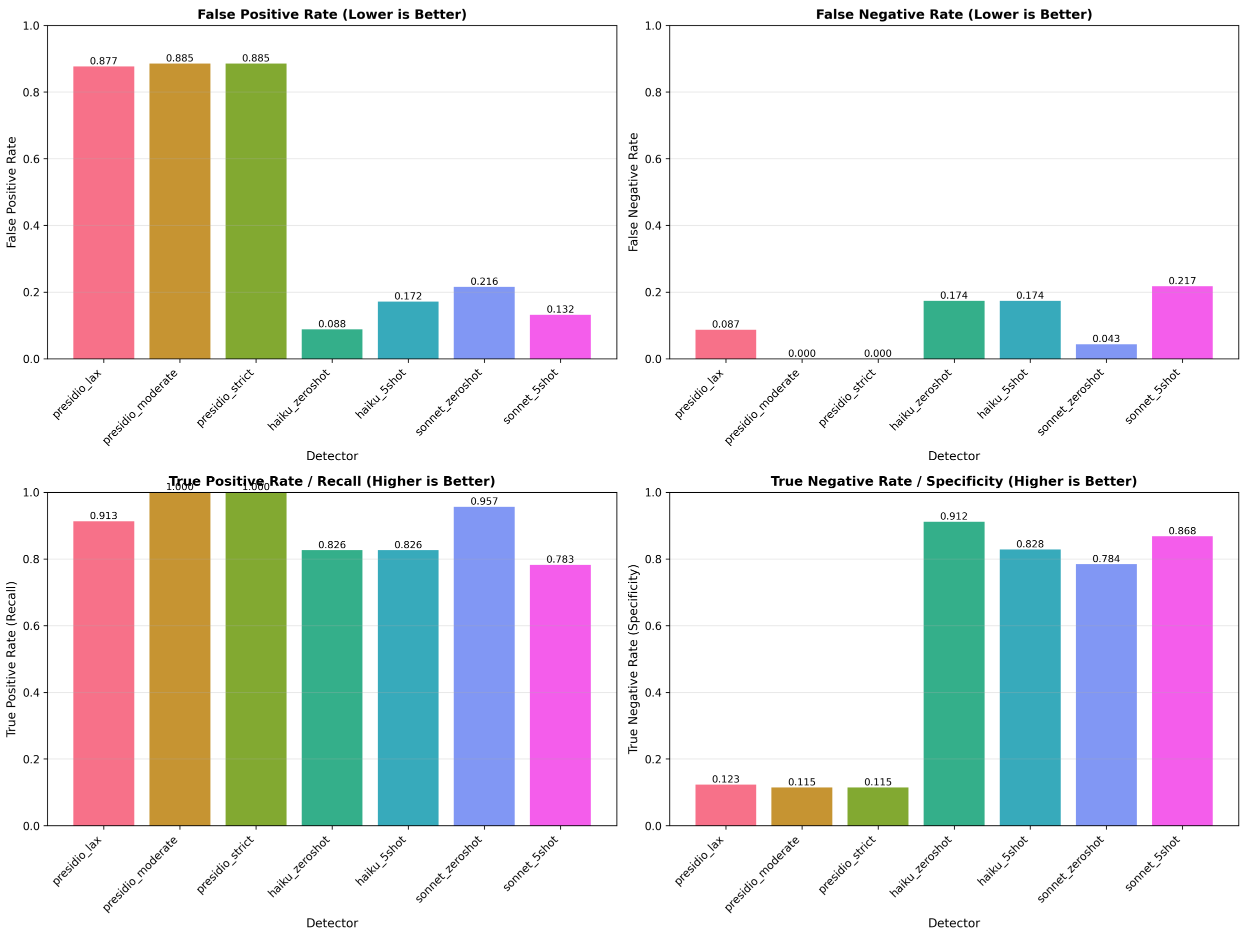
Email breach analysis evaluated on 250-sample gold-labeled dataset. **Primary goal: Zero false negatives to avoid missing PII.** Finding: Presidio achieves 0% FNR but 88% FPR (unusable in production). Sonnet zero-shot delivers 4.3% FNR with acceptable 22% FPR.

Detector Overview

Detector	Type	Description
Presidio	Local/Free	Rule-based NLP, runs locally
Haiku Zero-shot	Claude API	Fast LLM, no examples
Sonnet Zero-shot	Claude API	Slower LLM, higher accuracy

Performance Metrics

Detector	FNR ↓	FPR ↓	Status
Presidio Moderate	0.0%	88.5%	■■ Unusable
Haiku Zero-shot	17.4%	8.8%	✓ Good
Sonnet Zero-shot	4.3%	21.6%	✓ Best usable


False Positive Rate (Lower is Better)
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.877
0.885
0.885
0.088
0.172
0.216
0.132
presidio_lax
presidio_moderate
presidio_strict
haiku_zeroshot
haiku_5shot
sonnet_zeroshot
sonnet_5shot
Detector
False Negative Rate (Lower is Better)
False Negative Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.087
0.000
0.000
0.174
0.174
0.043
0.217
presidio_lax
presidio_moderate
presidio_strict
haiku_zeroshot
haiku_5shot
sonnet_zeroshot
sonnet_5shot
Detector
True Positive Rate / Recall (Higher is Better)
True Positive Rate (Recall)
1.0
0.8
0.6
0.4
0.2
0.0
0.913
1.000
1.000
0.826
0.826
0.957
0.783
presidio_lax
presidio_moderate
presidio_strict
haiku_zeroshot
haiku_5shot
sonnet_zeroshot
sonnet_5shot
Detector
True Negative Rate / Specificity (Higher is Better)
True Negative Rate (Specificity)
1.0
0.8
0.6
0.4
0.2
0.0
0.123
0.115
0.115
0.912
0.828
0.784
0.868
presidio_lax
presidio_moderate
presidio_strict
haiku_zeroshot
haiku_5shot
sonnet_zeroshot
sonnet_5shot
Detector

50k Email Projections

Detector	Time (Current)	Time (Parallel)	Cost	FNR
Presidio Moderate	19.7 min	2-3 min*	$0.00	0.0%
Haiku Zero-shot	741.4 min	15-74 min**	$28.48	17.4%
Sonnet Zero-shot	1724.7 min	35-173 min**	$106.78	4.3%

*With EC2 multi-core optimization

**With API parallelization (10-50x speedup; currently sequential)

Next Steps

- **1. Implement API parallelization** - 10-50x speedup potential (1-2 days development)
- **2. EC2 optimization for Presidio** - 5-10x speedup with multi-core processing
- **3. Tune Presidio to reduce FPR** - Goal: maintain 0% FNR while reducing 88% FPR

Recommendation: Deploy Sonnet zero-shot (4.3% FNR) with API parallelization for production 50k emails. Cost: $107, Time: ~35-173 min (parallelized). Alternative: Haiku zero-shot if 17.4% FNR is acceptable ($28, ~15-74 min).