entro

From Rules to Context

Building a Production-Grade Hybrid Secret Scanner with SLMs

AN EXCLUSIVE TECHNICAL BRIEF BY entrolabs

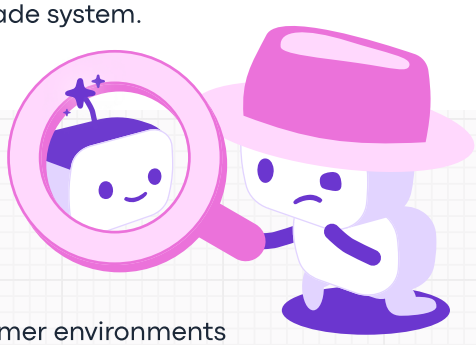
Solving the Precision-Recall Trade-Off in Enterprise Secret Security

The Business Reality of Secret Scanning

In the modern enterprise, secrets - API keys, tokens, and credentials - are the keys to the kingdom. Yet, securing them has traditionally meant choosing between two bad options: catching every potential leak but drowning your security team in false positives, or tuning the alerts down and risking a catastrophic breach.

Researchers at Entro Labs have fundamentally solved this precision-recall trade-off.

By developing a hybrid secret-scanning pipeline that pairs a traditional rules engine with a context-aware Small Language Model (SLM), Entro has turned a theoretical research concept into a resilient, production-grade system.



Key Takeaways:

- **End-to-End Visibility:** The pipeline is live in enterprise customer environments today, extending coverage beyond code to logs, configurations, and conversations.
- **Proven Accuracy:** On a 300-sample real-world benchmark, our hybrid approach achieved an F1 score of 0.91, outperforming leading open-source and commercial scanners.
- **Business Impact:** By virtually eliminating alert fatigue while maintaining high catch rates, security teams save countless engineering hours and can focus on genuine risk remediation.

The problem with legacy approaches

The Regex Trap: Why Traditional Scanners Fail

Today, most secret-detection scanners rely heavily on regular expressions (regex) and hardcoded rules. The limitation is simple: regex can match text patterns, but it cannot understand context. Consequently, it routinely flags look-alike strings that pose zero risk.

Consider the following common scenario:

```
// LoginServiceTest.java

@Test
void testInvalidCredentialsThrowsException() {
    assertThrows(AuthException.class, () ->
        loginService.authenticate("admin", "P@ssw0rd123!")
    );
}
```

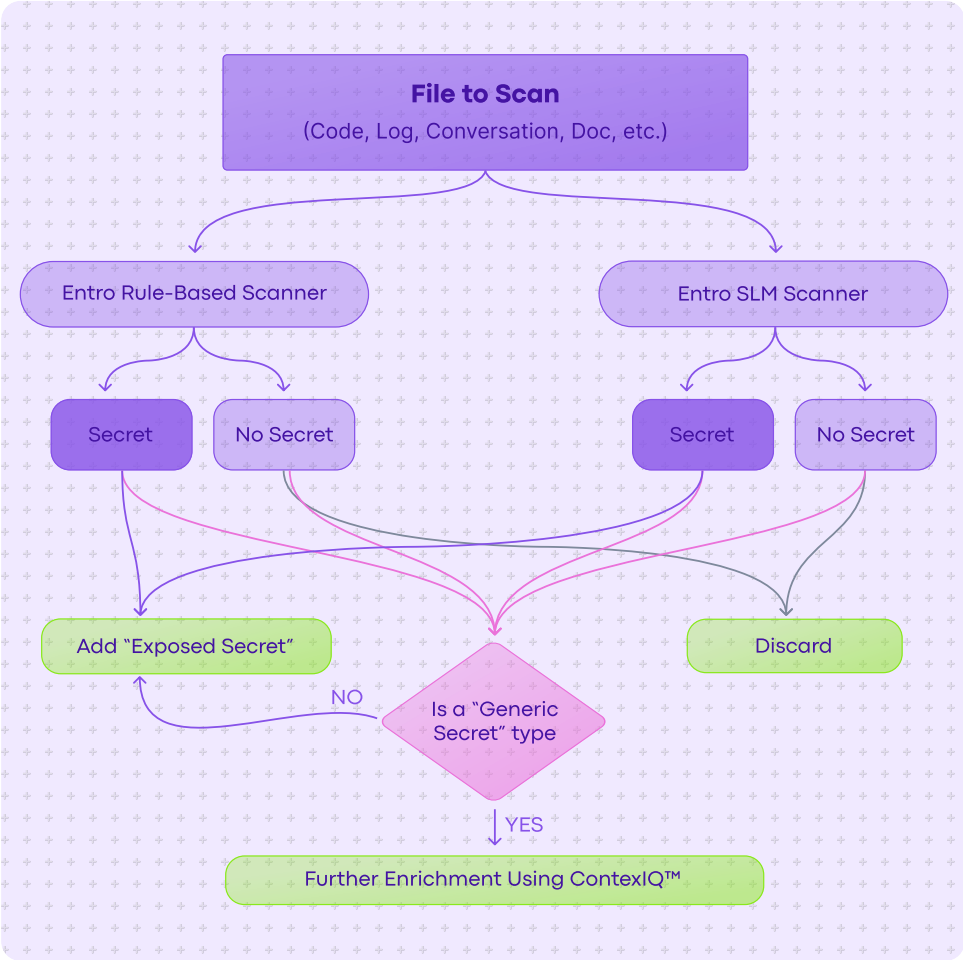
Copy

A human engineer immediately recognizes this as a harmless test fixture. However, to a rule-based engine, "P@ssw0rd123!" triggers a critical alert. It is practically impossible to hardcode rules for every benign context.

The result?

1. **High False Positive (FP) Rates:** Alert fatigue sets in, causing teams to ignore warnings.
2. **High Maintenance:** Constant rule tuning is required as credential formats evolve.
3. **Low Catch Rates:** In Entro Labs' research, rule-based scanning alone caught only ~60% of potential leaks while still generating massive noise.

The Hybrid Secret-Scanning Pipeline



To bridge the gap between pattern matching and human-like understanding, Entro Labs engineered a two-stage hybrid pipeline. This architecture combines the brute-force scalability of a rules-based discovery engine with the deep contextual intelligence of state-of-the-art, fine-tuned Small Language Models (SLMs).

How the "Hybrid" Approach Works in Production

1. **The Discovery Engine (Scale):** The rule-based scanner acts as the first net, finding large numbers of candidate secrets at scale.
2. **The SLM Validation (Context):** The fine-tuned SLM evaluates these candidates, reading the surrounding context (like variable names, file paths, and comments) to determine if the string is a genuine, exposed secret or a benign artifact.
3. **The Training Loop:** The rules engine generates the raw training signal, which Entro Labs curates into high-quality PEFT (Parameter-Efficient Fine-Tuning) examples to continuously sharpen the SLM.

The CISO Perspective: Why SLMs in 2026?

2026 is the year Small Language Models move from research to critical enterprise infrastructure. For security leaders, the shift to SLMs over massive generalized LLMs offers distinct architectural advantages:

Strict Data Privacy

SLMs can run on-device, on-prem, or entirely inside a VPC. Sensitive source code and configurations never leave your boundary.

Ultra-Low Latency

Smaller models mean faster inference, eliminating queue bottlenecks in high-throughput CI/CD pipelines.

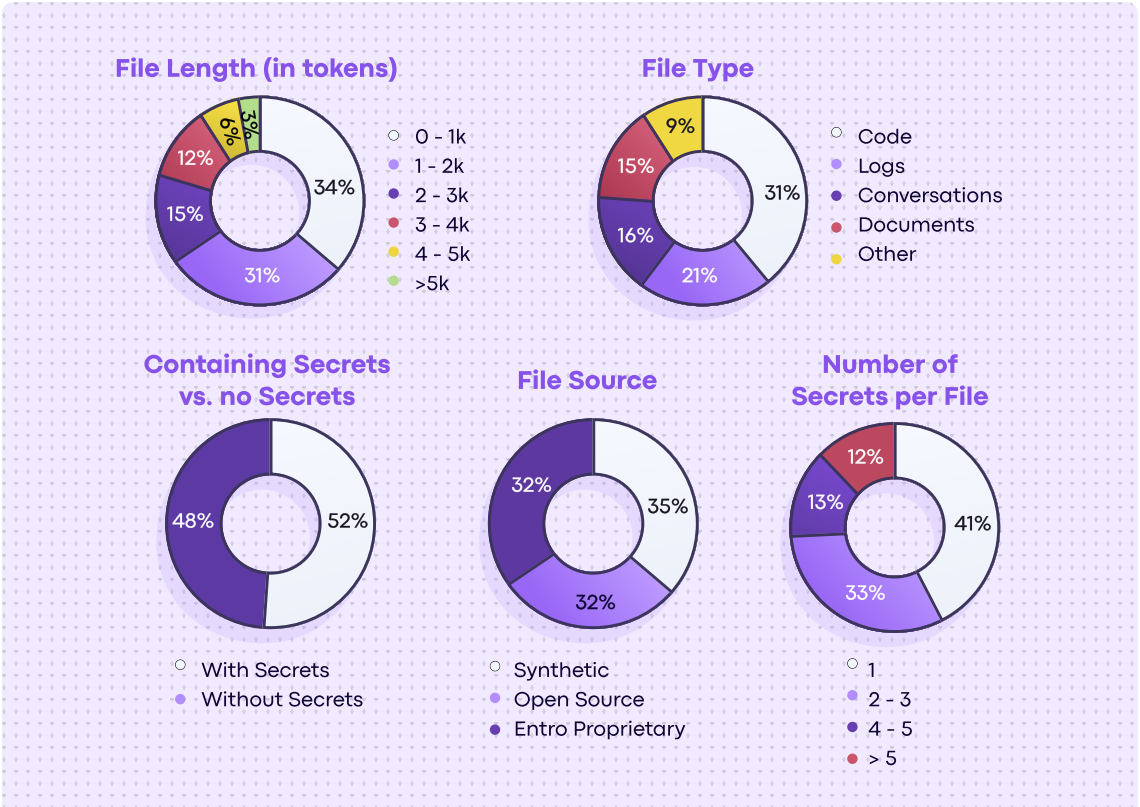
Predictable Economics

When scanners make millions of model calls per task, SLMs provide superior unit economics without sacrificing accuracy.

Methodology & Benchmarking

Proving the Pipeline: The 300-File Benchmark

To validate the hybrid pipeline, Entro Labs curated a rigorous benchmark of 300 code-only samples sourced from real open-source repositories (anonymized to prevent real credential exposure). The dataset was carefully balanced: ~51% contained one or more secrets, while ~49% contained none.



Success Criteria: Exact Secret Extraction

In secret scanning, merely flagging a file as "suspicious" is insufficient. The real engineering challenge—and our success criteria—was exact boundary detection.

The model must correctly identify:

1. The secret value.
2. Its corresponding exact line number.

Why this matters: Exact extraction reduces the cognitive load on the model, prevents text-generation hallucinations, and gives DevSecOps teams the precise coordinates needed for immediate remediation.

The Contenders

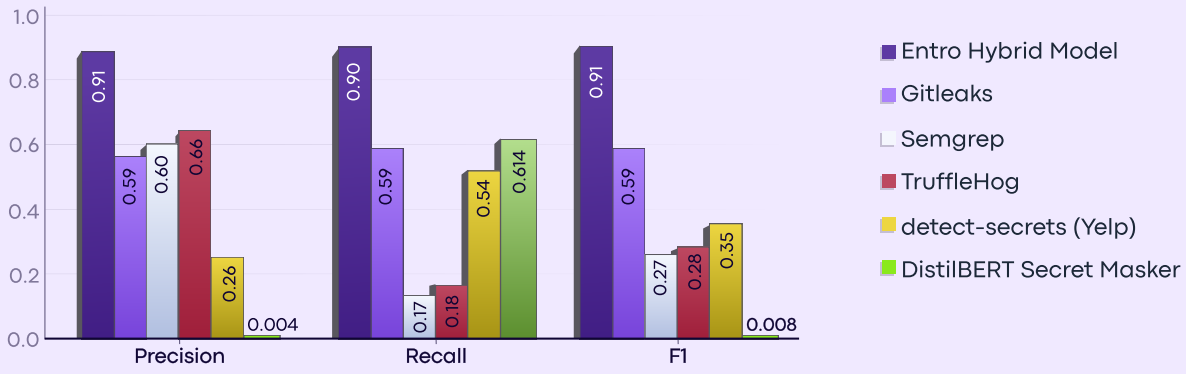
Entro Labs benchmarked our hybrid pipeline against five leading open-source scanners:

1. TruffleHog: Known for deep verification against service APIs.
2. Gitleaks: A lightweight, reliable baseline for git repositories.
3. Semgrep: A fast, static analysis tool.
4. detect-secrets (Yelp): An enterprise-focused diff scanner.
5. Distilbert-secret-masker: An NER-based transformer model.

The Results

Unmatched Precision and Recall

To evaluate the models, we used the F1 score (the harmonic mean of precision and recall), which represents the ultimate balancing act in security: catching real leaks without generating noisy alerts.



The Data Speaks for Itself: Across the benchmark, the industry tools split into clear, flawed profiles. Entro's Hybrid Pipeline was the only solution to excel on both axes:

- Precision: 0.91 (Minimal false positives)
- Recall: 0.90 (Catches the vast majority of real secrets)
- Overall F1 Score: 0.91

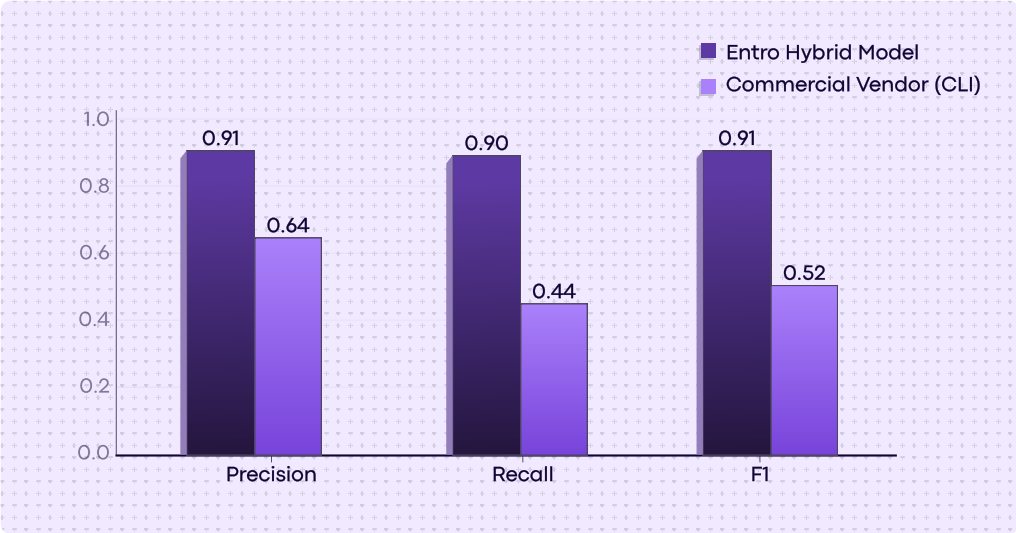
How the Alternatives Fared:

- **Cleaner alerts, lower coverage**: TruffleHog and Semgrep were conservative. They raised fewer false alarms but missed a meaningful portion of real secrets in messy contexts.
- **Higher coverage, noisier output:** detect-secrets caught more but suffered from high false positives, driving up alert fatigue.
- **Middle of the pack:** Gitleaks proved to be a solid baseline, but failed to push the boundaries on either precision or recall.
- **Recall-heavy, precision-light:** The DistilBERT model flagged heavily but struggled with precision, requiring heavy post-filtering.

Entro's hybrid pipeline stays steady on the hard, ambiguous cases, generic strings, passwords, and messy real-world contexts, while keeping coverage high. You no longer have to trade trust for detection.

Entro vs. a Commercial Secrets Scanning Vendor Precision and Recall

To sanity-check our results, we also ran a popular CLI scanner from a known secrets security vendor on the same 300-file benchmark. In this setup, it behaved like many mature rule-heavy scanners do: decent precision and clean-file handling, but lower recall on real leaks, which pulls down overall F1.



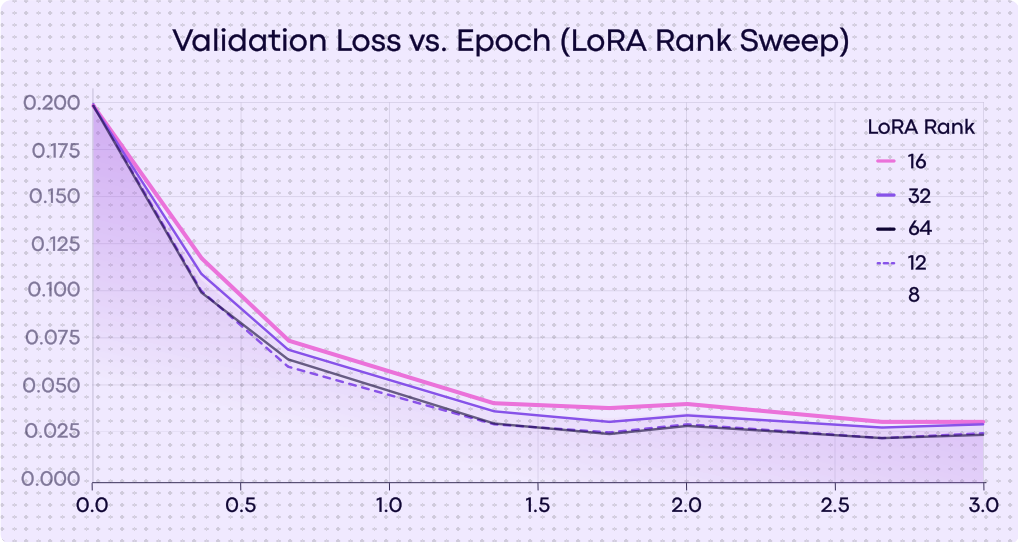
Entro vs. a commercial secrets security vendor's CLI scanner.
Measured using the vendor's public CLI with our benchmark configuration

Beyond Scanning & Conclusion

Finding secrets in source code is only the starting point for an effective, robust secret security solution. While the SLM-powered Hybrid Pipeline reliably scans secret candidates at scale, Entro Security has also developed an advanced in-house validation layer: ContextIQ™.

ContextIQ™ performs deeper contextual checks and risk-aware validation. Its goal is to reduce false positives to absolute zero, validate the live state of the secret, and enrich the context before it ever surfaces to a security analyst. (Look for our upcoming Entro Labs report diving deep into ContextIQ™).

The era of relying solely on regular expressions for enterprise secret scanning is over. The risks of missed NHI credentials and the costs of alert fatigue are simply too high.



Validation loss vs. epoch for the LoRA rank sweep. Ranks 64 and 128 perform similarly, with 128 achieving the lowest validation loss overall.

Through rapid iteration, advanced fine-tuning techniques (LoRA, QLoRA), and cutting-edge serving methods, Entro Labs has delivered a pipeline that can absorb the complexities of modern codebases at scale, keeping your data private, your pipelines fast, and your security teams focused.

Stop Drowning in False Positives.

Discover all AI agents, non-human identities, and their secrets across your organization tech stack with the accuracy of Entro's Hybrid SLM Pipeline.

Visit entro.security to learn more about our unique approach or request a demo to see the platform in action!

Contact Entro **Request a Demo**