Trust but Verify? Uncovering the Security Debt of Autonomous Coding Agents

AHM Nazmus Sakib, Dipayan Banik, Murtuza Jadliwala

University of Texas at San Antonio, USA | Danovo Energy Solutions, USA

Abstract & Core Ideas

- **Context:** Autonomous agents (Copilot, Claude, Devin, etc.) rapidly generate PRs, but human review capacity lags behind generation velocity.
- **Problem:** While productivity is well-studied, the **security posture** of agent-authored code remains largely uncharacterized.
- **Study:** We evaluated the AIDev dataset (**16,112 file changes, 4,022 PRs**) using an LLM-as-a-judge pipeline to detect **security smells** (patterns indicating risk).
- **RQ1:** Which security smells appear most, and how do they distribute by category and severity?
- **RQ2:** How many flagged secrets are genuine, who introduced them, and are they caught during review?

Methodology Pipeline

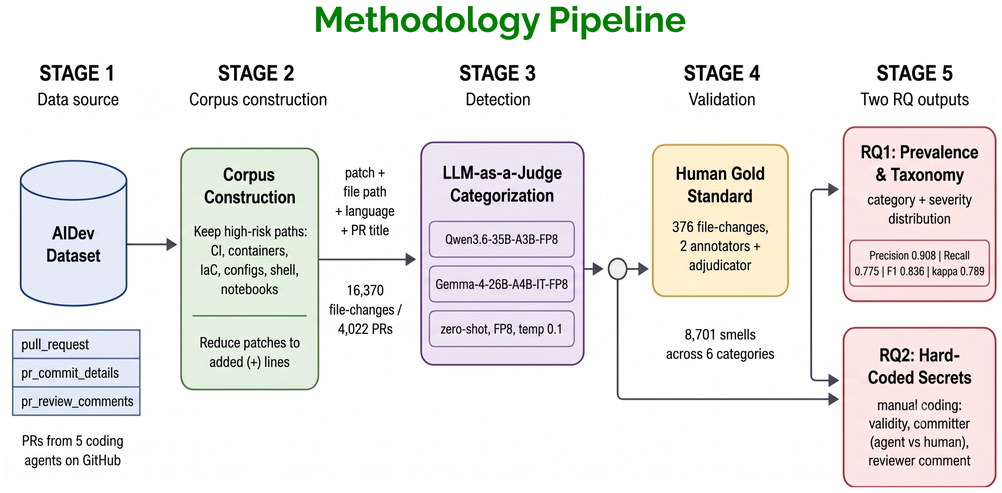

Figure 1. Architecture of the methodology: corpus construction, LLM detection, and validation.

- **Corpus:** Extracted high-risk paths (CI configs, IaC, secrets, shell scripts) from **AIDev**. Retained only added (+) lines (16,370 changes / 4,064 PRs).
- **Detection:** Dual LLM Judges (Qwen3.6-35B & Gemma-4-26B) labeled category, severity, and rationale (results unioned).
- **Validation:** Human gold standard (376 changes) achieved high reliability: **0.908 precision, 0.775 recall, 0.836 F1, 0.789 Cohen's κ.**

Security Smell Taxonomy

Category	Scope
secrets_identity	Hard-coded keys, tokens, and credential-bearing strings.
cleartext_transport	Cleartext endpoints, disabled TLS, weak ciphers.
over_privilege	Root containers, world-writable files, sudo, broad CI scopes.
permissive_network	All-interface binds, open CIDRs, public-access flags.
misconfig_hardening	Debug modes, logged secrets, disabled encryption.
supply_chain_integrity	Mutable action/image tags, unpinned global installs.

Table 1. Security Categories based on OWASP, CIS, and GitHub guidance.

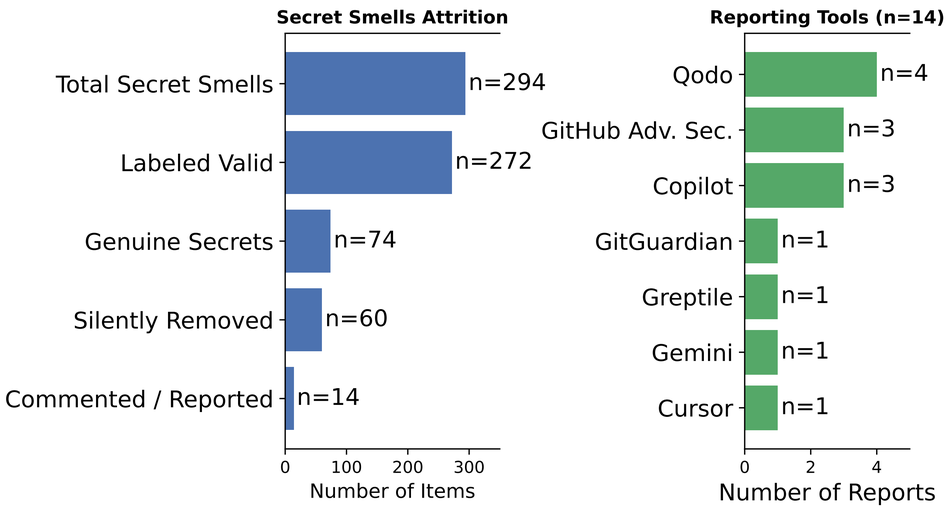

Figure 2. Severity Composition of Smells by Category.

RQ1: Distribution & Severity of Security Smells

- **Prevalence:** Flagged **38.9%** of agent PRs (1,563 / 4,022) with ≥1 smell (8,701 total smells).
- **Top Category:** **Supply Chain Integrity** (82.3% of smells), mostly mutable tags/unpinned installs.
- **Critical Severity:** **99.6%** of all critical findings were **Secrets/Identity** (hard-coded credentials).
- **Key Drivers:** Flag rates peak with large PRs (53.6% for 1000+ lines), **Copilot** (45.5%), **JavaScript** (55.3%), and CI/Docker files (87.6%).

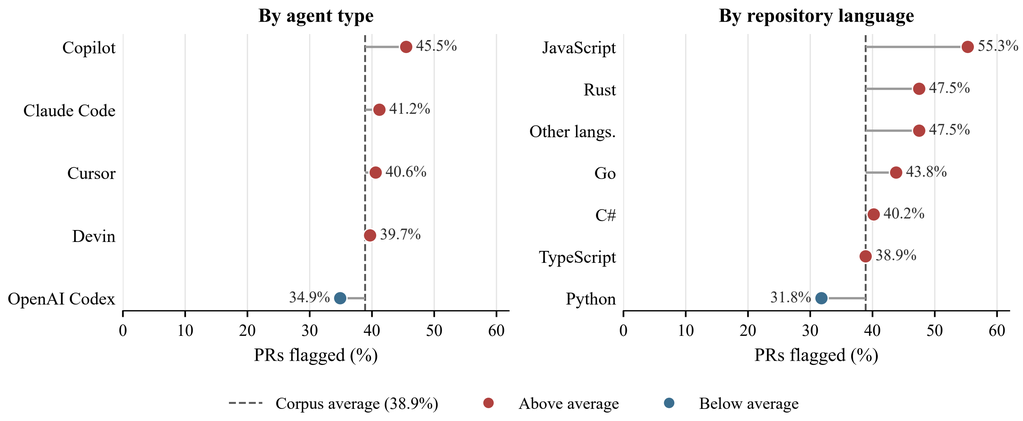

Figure 3. PR-level smell-flagging rate by agent type and repository primary language.

RQ2: Genuine Secrets & Review Outcomes

- **Genuine Secrets:** 74 out of 272 labeled secrets (**27.2%**) were genuine, live credentials.
- **Human vs. Agent:** **Humans introduced 67.6%** of genuine secrets; agents only 32.4%.
- **Review Failure:** Security bots and human reviews **missed 81.1%** of genuine secrets before integration. Only 14 cases triggered warnings.

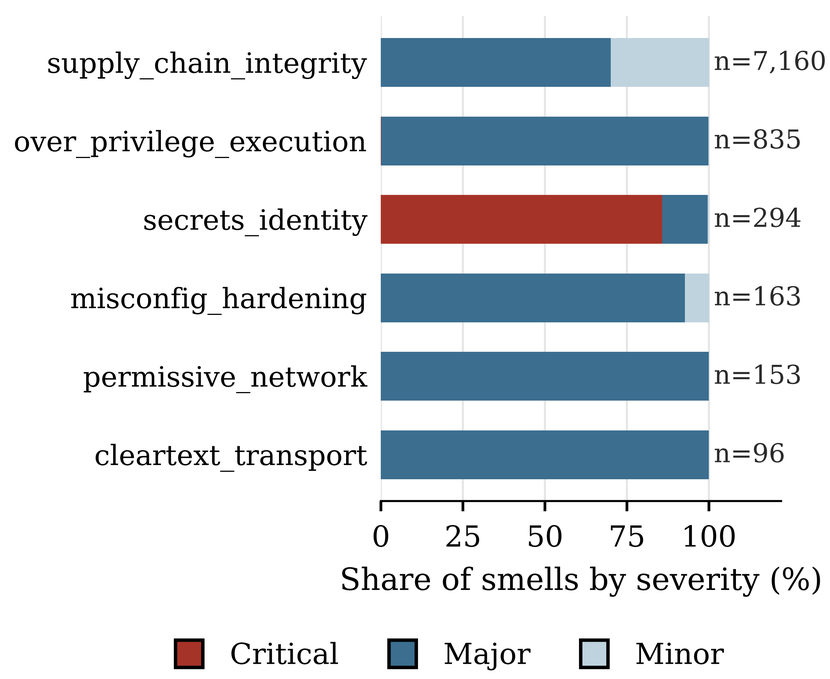

Figure 4. Summary of RQ2 Investigation Outcomes regarding genuine secrets and reporting tools.

Conclusion

- **Security Debt:** Agentic workflows introduce significant debt, dominating in supply-chain misconfigs and critical hard-coded credentials.
- **The Human Factor:** Ironically, humans introduce the majority of live leaked credentials in these workflows, not the agents.
- **Review Fatigue:** Current guardrails and human reviewers fail to intercept most secrets at the human-AI boundary.
- **Future Work:** Context-aware, preventative guardrails are needed directly within human-AI collaboration points.