

Nordic AI Meet 2026

P3: LLMs, Language, Evaluation, Privacy & Trustworthy AI

1. Toward an Autonomous Code Quality Agent: Coupling LLM Agents with Static Smell Detectors for Detect–Refactor–Verify Loops in Python and ML Codebases

Authors: Karthik Shivashankar¹

Affiliations: 1. SINTEF Digital, Oslo, Norway

Abstract: *Background:* Code smells and domain-specific machine learning (ML) anti-patterns (e.g., data leakage, unpinned models) drive technical debt, degrading system maintainability. While static analyzers detect these issues reliably, they cannot fix them. Conversely, Large Language Models (LLMs) generate fixes but suffer from hallucinations and lack systematic correctness guarantees.

Our Approach: We introduce a Code Quality Agent that operationalizes static smell detectors as verification oracles within an automated D-R-V (Detect, Refactor, Verify) loop:

Detect & Prioritize: Grounded in our own two peer-reviewed tools; PyExamine (for Python architectural/structural smells) and MLScout (for ML-specific anti-patterns); the agent scans the repository and prioritizes findings into a unified inventory.

Refactor: The LLM proposes a minimal, behavior-preserving patch informed by the detector's localized repair hints.

Verify: The agent re-runs the static analyzers and the project's test suite. A patch is accepted only if the targeted smell is eliminated, no new smells are introduced, and all tests pass. This design converts open-ended code generation into a verifiable optimization problem.

Evaluation & Contributions: We evaluate the agent across a diverse corpus of open-source Python and ML repositories to measure its effectiveness, safety, and token efficiency. This work contributes: (1) an open-source agent architecture utilizing static analysis as a verification oracle, (2) empirical evidence of autonomous technical debt reduction, and (3) a taxonomy of smell categories by repair difficulty. Ultimately, we demonstrate a reliable division of labor for AI-assisted software quality: deterministic tools to measure, and LLM agents to act.

<https://github.com/KarthikShivasankar/quality-triage>

https://github.com/KarthikShivasankar/ml_smells_detector

https://github.com/KarthikShivasankar/python_smells_detector