Sandbox Isolation Strategies for Untrusted Agent-Generated Code

Running model-authored code at scale requires isolation that is cheap to provision and hard to escape. We compare micro-VM, container, and language-level sandboxes across 12,000 agent runs, measuring cold-start latency, escape-attempt containment, and per-run cost. Micro-VMs contained 100 percent of injected escape attempts at a median 1.9 second cold start, while language-level sandboxes were four times faster but leaked on three of forty adversarial payloads. We recommend a tiered policy keyed to task trust level.

This is followed by a discussion of scheduling implications and a cost model that operators can apply to their own workloads. A reference implementation is provided.

Authors: Priya Ramanathan, Oskar Lindqvist, Chidi Eze
Affiliation: Datem Systems Group; Fjordline Security Institute
Venue: Workshop on Agent Infrastructure (WAI '26), co-located with PelicanConf 2026
Keywords: sandboxing, micro-VM, agent safety, code isolation, cost modeling