Abstract

A governance-aware generative DevOps system is disclosed that executes entirely inside a customer's cloud account. Lightweight AI *micro-agents* (e.g., for provisioning, tuning, remediation) run under separate least-privilege identities and invoke a private in-VPC foundation-model endpoint to generate infrastructure-as-code or API actions. A reinforcement-learning orchestrator (FIG. 1) selects agent actions using a tri-vector reward that jointly optimises cost, performance and a quantified compliance-risk signal. Before execution, every proposed change is intercepted by a *Policy Weaver* layer that evaluates machine-readable guardrail rules—covering IAM scope, mandatory tags, budget limits, security baselines and regional or regulatory constraints—and blocks or auto-corrects non-conforming changes in real time. Each deployment logs outcomes and locally fine-tunes its models; encrypted parameter deltas are periodically aggregated by a federated-learning coordinator to update a global model without exposing raw data. The architecture (FIG. 1) thus delivers autonomous cloud operations that continuously improve while guaranteeing that all AI-driven actions remain cost-efficient, high-performing and policy-compliant within the user's own security perimeter.