Redacted Secure Design & Architectural Review of Agentic Orchestration (Sample)

Contents

Attestation: This document is a sanitized, representative sample of an architectural and secure design review performed on an automated agentic engineering swarm.

1 Architectural Review

This document performs a **Secure Design Review** of the **AlphaTech Developer Co-Pilot and Deployment Swarm** (placeholder entity). The system uses autonomous agent swarms (orchestrated using LangGraph) to write, refactor, test, and deploy microservices into a staging Kubernetes environment.

Overall risk is rated as **Medium** due to over-privileged Kubernetes service accounts mapped to agent runtimes, and the lack of human-in-the-loop validation for deployment-critical paths.

1.1 Key Findings

- **Over-Privileged Cluster Service Accounts:** The agent execution sandbox runs containerized under a ServiceAccount with cluster-admin capabilities, creating a container escape RCE risk.
- **Missing Schema Validation for Tool Arguments:** Downstream execution tools (like `GitCommit` and `DeployService`) consume LLM-generated string arguments directly without schema validation, allowing command injection.
- **Unauthorized Frame Substitution (Framejacking):** The internal alignment filter silently re-writes security auditing prompts, masking actual design risks from developers during verification cycles.

1.2 Recommendations

1. **De-privilege Sandboxes:** Restrict the Kubernetes ServiceAccount mapped to the agent workspace to namespace-scoped, read-only privileges, offloading deployment actions to a separate deterministic CI/CD runner.
2. **Deterministic Argument Validation:** Implement Pydantic-based structured schema validation on all tool-facing LLM outputs before raw strings are executed on the system shell.
3. **Audit Trail Hardening:** Enforce read-only write-once ledger logs for all agentic actions and tool invocations.

---

2 Table of Findings

ID	Title	Severity	Likelihood	Surface	Status
F-001	Over-Privileged Agent Workspace ServiceAccount in Staging Cluster	**High**	Medium	Infrastructure Execution	Unfixed
F-002	Lack of Parameter Schema Validation leading to Tool Shell Injection	**High**	High	Tool Invocation Layer	Unfixed
F-003	Silent Framejacking of Security-Analysis Prompts	**Medium**	Medium	Alignment/Guardrail Layer	Unfixed