

O301 OPERATE · ADVANCED

Building and Running AI Agents in Production

Architecture, protocols, identity, evaluation and operations for enterprise AI agents.

FORMAT

Knowledge pack · 12 slides

FOR

Platform leaders and architects

LEVEL

Advanced



DEFINITION

An agent is a model, tools and context running in a controlled loop

An agent uses a model to choose an action, calls a tool, observes the result and continues until it reaches a goal or a limit.

Readiness depends less on the model than on **what the agent can access**, what context it has and how its actions are checked.

Access

Tools, data and permissions.

Context

Retrieved knowledge, memory and instructions.

Control

Limits, approvals, logs and evaluations.

ANATOMY

The anatomy of an enterprise agent

Instructions	The goal, role, constraints and definition of done.
Model	Reasoning and planning, chosen per task for quality and cost.
Context	Retrieved knowledge, conversation state and memory.
Tools	APIs, MCP servers and sandboxed code execution, each with its own permissions.
Control	Identity, budgets, approvals, logging and a kill switch.

DESIGN RULE

Enforce permissions and budgets outside the model; injected text can override prompts.

PATTERNS

Choose the simplest pattern that works

PATTERN	USE WHEN	MAIN RISK
Fixed workflow with model steps	The process is known in advance	Brittle when cases vary
Router	Requests fall into distinct types	Misrouting
Tool-using agent	Steps vary from case to case	Unbounded loops and cost
Orchestrator and sub-agents	Tasks split into independent parts	Compounding errors; about 15× chat tokens (Anthropic)
Human checkpoints	Actions are high-impact or irreversible	Approval fatigue

PROTOCOLS

Protocols are becoming the integration layer

Model Context Protocol

Connects agents to tools and data through MCP servers. Introduced by Anthropic in November 2024; hosted since December 2025 by the Linux Foundation's Agentic AI Foundation.

Agent2Agent

Lets agents from different vendors discover each other and exchange tasks. Introduced by Google in April 2025; version 1.0 in March 2026; hosted by the same foundation since August 2026.

What it means for IT

Treat every MCP server and agent endpoint as an integration: inventoried, owned, authenticated, version-pinned and reviewed on change.

THREATS

Threats that agents introduce or amplify

Mapped to the OWASP Top 10 for LLM Applications 2026 (LLM) and the OWASP Top 10 for Agentic Applications for 2026 (ASI).

Prompt injection (LLM01)

Typed directly, or hidden in documents, emails, tickets or web pages.

Agent goal hijack (ASI01)

The agent is redirected toward an attacker's objective.

Tool misuse (ASI02)

Legitimate tools used in harmful ways or sequences.

Identity and privilege abuse (ASI03)

Inherited or cached rights used beyond the user's authority.

Agentic supply chain (ASI04)

Tampered tools, MCP servers or tool descriptions.

Memory and context poisoning (ASI06)

Corrupted context that persists across sessions.

Cascading failures (ASI08)

Errors that propagate between connected agents.

Human-agent trust exploitation (ASI09)

Persuasive explanations that win approval for unsafe actions.

IDENTITY AND ACCESS

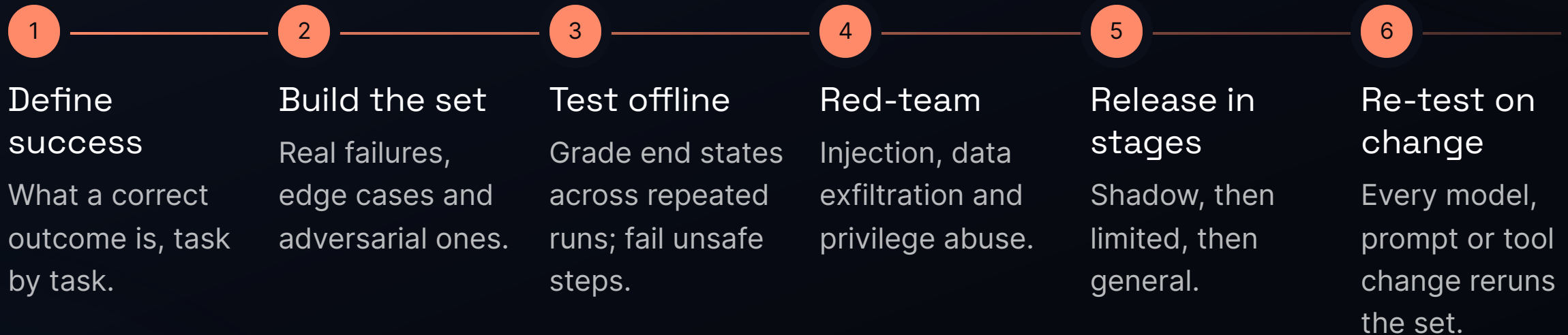
Give every agent an identity and only the privilege it needs

MCP authorization is optional; over HTTP it follows the OAuth 2.1 draft, accepts only tokens issued for that server and forbids passing them upstream.

- ✓ **Its own identity**
A non-human identity per agent; never shared credentials.
- ✓ **Short-lived credentials**
Issued per task and expiring quickly.
- ✓ **Named approvers**
High-impact actions need a named approver.
- ✓ **Delegated authority**
Token exchange (RFC 8693) issues tokens naming both user and agent, within the user's rights.
- ✓ **Scopes per tool**
Permissions granted tool by tool, never platform-wide.
- ✓ **Traceability**
Each action linked to the requester, agent, tool and result.

EVALUATION

Evaluate before release and after every change



OPERATIONS

Operate agents like production services

OpenTelemetry's GenAI agent span conventions still have Development status, so pin a version and expect change.

Verified task success

Share of tasks whose end state is verified correct by sampling.

Escalation rate

Share of tasks handed to people, and the reasons.

Intervention rate

Share of actions people stopped or reversed.

Cost per outcome

Model, tool and review cost per verified task.

Policy blocks

Actions prevented by guardrails, by type.

Evaluation regression

Change in evaluation scores after each release.

READINESS

The production readiness checklist

- ✓ **Named owner**
Accountable for outcomes, cost and incidents.
- ✓ **Tool inventory**
Every MCP server and API reviewed and approved.
- ✓ **Approval gates**
For consequential or irreversible actions.
- ✓ **Evaluation in CI**
Regression cases gate every change and should pass at close to 100%.
- ✓ **Identity and scopes**
Reviewed like any privileged account.
- ✓ **Budgets**
Limits on tokens, steps, time and spend.
- ✓ **Tracing and logs**
Prompts, tool calls and results, within retention rules.
- ✓ **Incident runbook**
Including how to stop the agent and revoke its access.

PITFALLS

Common failure patterns in agent projects

✗ An agent where a workflow would do

✓ Use deterministic steps where the path is known

✗ One agent with broad access

✓ Scope permissions per tool and per task

✗ Judging readiness from demos

✓ Keep a versioned evaluation set of real cases

✗ No limits on cost

✓ Set budgets per run, per user and per day

✗ Trusting retrieved content

✓ Treat documents and tool output as untrusted

✗ Unversioned prompts and tools

✓ Version, review and roll back like code

SUMMARY

Key takeaways

- 01 Prefer a fixed workflow when the path is known; add agency only where steps vary.
- 02 Treat MCP servers and agent endpoints as integrations to inventory and secure.
- 03 Every agent needs its own identity, scoped delegated authority and a full audit trail.
- 04 Evaluate on real cases before release and after every change.
- 05 Run agents with budgets, approvals, tracing and a kill switch.

AXIARETE ACADEMY

Explore every track

Ten courses across five tracks