Agent Skills Layer

Agent Skills Layer

NEW IN 2026

Skills are an emerging complement to MCP tools: composed, reusable capability bundles (instructions + optional scripts/resources) that an agent loads on demand, rather than a live tool connection. The MCP roadmap lists a formal Skills primitive as an active 2026 work item, alongside the existing ext-auth and ext-apps extension tracks.

Practical distinction for architecture decisions:

- MCP tools — for live, stateful, or side-effecting connections to external systems (databases, APIs, SaaS products).
- Skills — for packaged, mostly-static know-how (a house style guide, a document-formatting procedure, a domain checklist) that doesn't need a network connection to a server.
- Both should be routed through the same permissioning and audit layer as everything else in the harness — a Skill is still executable content and should be treated as untrusted input until vetted.

Safety Classification for Tools

Levels	Skills	Controls
Low	Informational	Auto execute + log
Medium	Controlled side effects	Policy checks + confidence threshold
High	Financial / irreversible / legal	Verifier + human approval

Mandatory Rules

High-risk actions require:

- Human approval
- Full audit trail
- Secondary verification
- Explicit user identity

Revision #1
Created 2026-08-24 12:05:09 UTC by Aisha M. Nur
Updated 2026-08-24 12:05:35 UTC by Aisha M. Nur