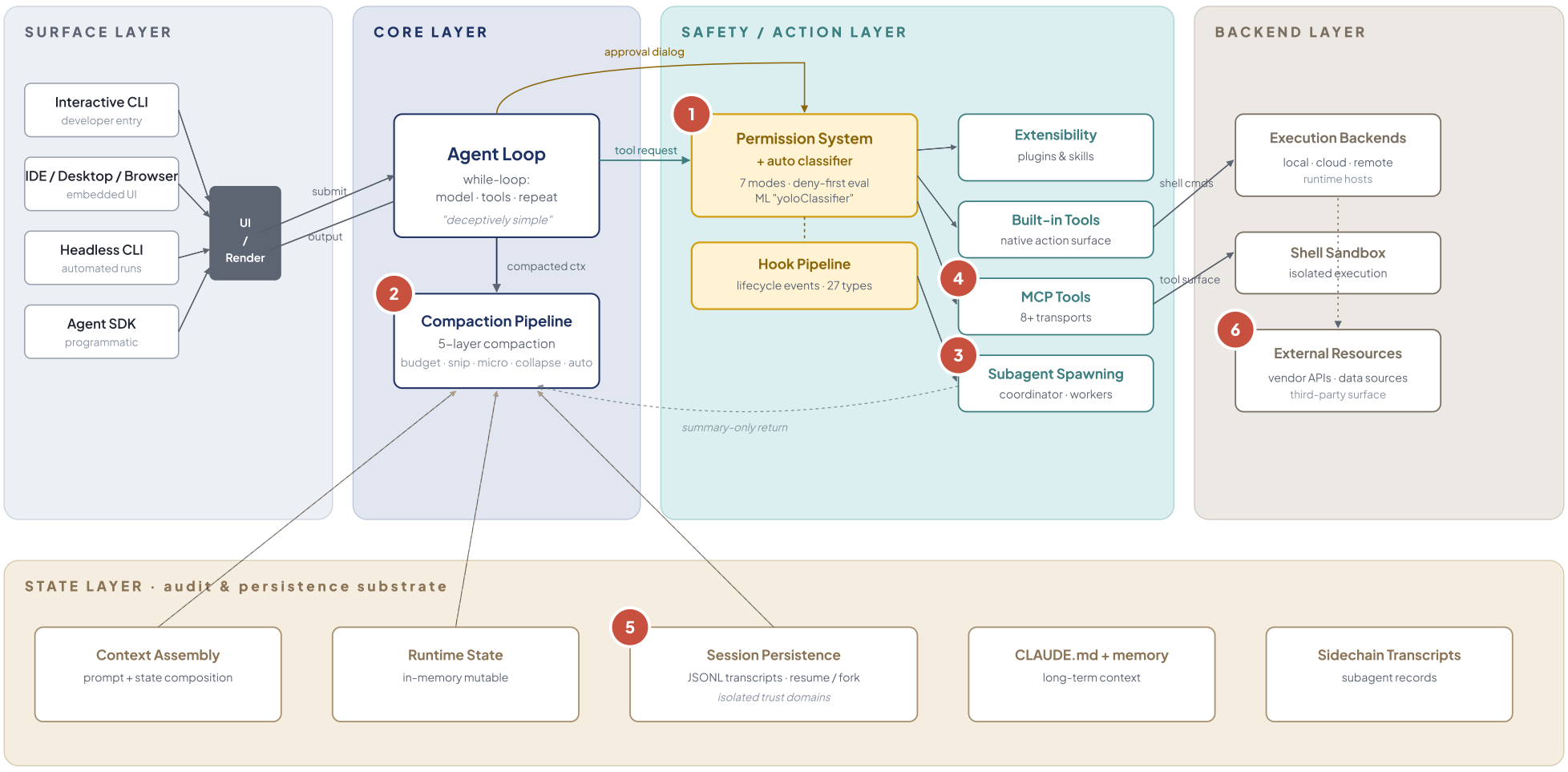
INSIDE THE AGENTIC AI LOOP

The Agentic Architecture, Annotated

A layered map of a production **Claude agentic AI system** — five architectural layers, with six numbered annotations marking where institutional governance, accountability, and risk concerns must be specified before deployment.



Source structure: MBZUAI VILA Lab & UCL — "Dive into Claude Code: The Design Space of Today's and Future AI Agentic Systems" · Reinterpretation by NextFi Advisors

NEXTFI INSTITUTIONAL ANNOTATIONS

Six points where the architecture meets the institutional surface.

Each numbered marker on the diagram corresponds to a specific governance, accountability, or operational concern that financial institutions must specify *before* production deployment — not after. Drawn directly from the NextFi brief.

1

Permission System + auto classifier

SAFETY / ACTION LAYER

The governance philosophy of the system, **not a feature**. Deny-first rule evaluation across seven modes, with an ML-based classifier (yoloClassifier) adjudicating tool-use at the per-action level. The classifier is itself a model — subject to model risk governance requirements of its own.

2

Compaction Pipeline

CORE LAYER

Five-layer pipeline managing reliability under **long-horizon execution**. Critical for workflows that span hours or days. The broader challenge of maintaining coherent task intent and human checkpoints across extended horizons remains architecturally open.

3

Subagent Spawning

SAFETY / ACTION LAYER

Functional delegation works. The **accountability trail** required for institutional deployment in regulated environments does not yet exist by default. When a primary agent delegates to subagents, the chain must remain legible to operators and supervisors.

4

MCP Tools

SAFETY / ACTION LAYER

MCP appears as **one of four core extensibility mechanisms** — not an add-on. Procurement and vendor evaluation processes that treat MCP compatibility as a secondary specification are misaligned with the direction of the market.

5

Session Persistence

STATE LAYER

Sessions are treated as **isolated trust domains**. When resumed or forked, previously granted permissions are not automatically restored. The system accepts user friction as the cost of preserving a core safety invariant — a principle institutions should internalize.

6

External Resources + Backends

BACKEND LAYER

The **vendor connectivity and operational dependency surface**. Local, cloud, and remote execution all touch external resources through this boundary. Where the architecture meets vendor risk — and where institutions must specify dependency controls.

THE ARCHITECTURAL READ

Per-action safety evaluation, ML-based permission classification, and append-oriented session storage are not technical footnotes — they are governance design decisions. **The institutions that specify these decisions before vendor selection** will deploy into agentic AI with their model risk frameworks intact. Those that don't will be re-engineering after the fact.