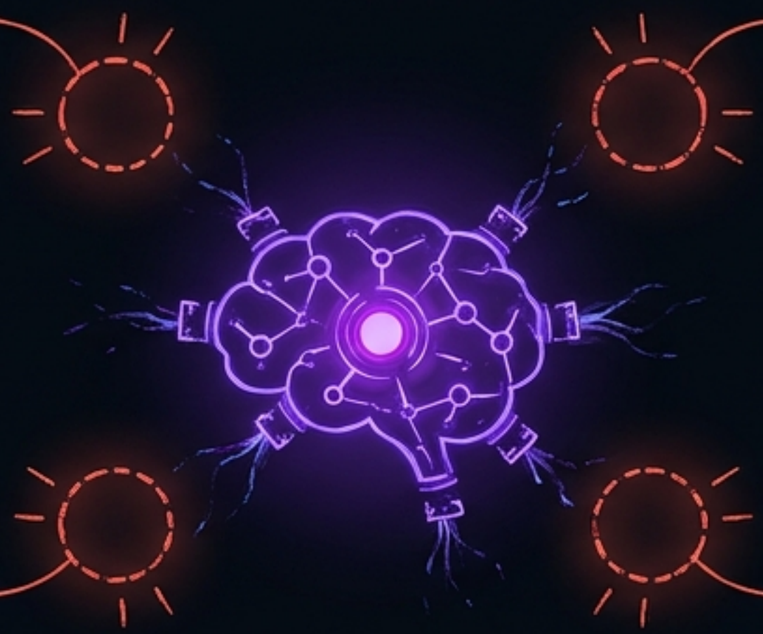

The Cybernetic Blueprint:
Agent Development Kit (ADK) v2.0
Production Architecture, Orchestration &
Governance for the AI Ecosystem



Stateless

Retains zero memory across HTTP requests (unless using server-side interaction handles).

Passive

Emits tool-call intents as JSON text without the ability to execute anything.

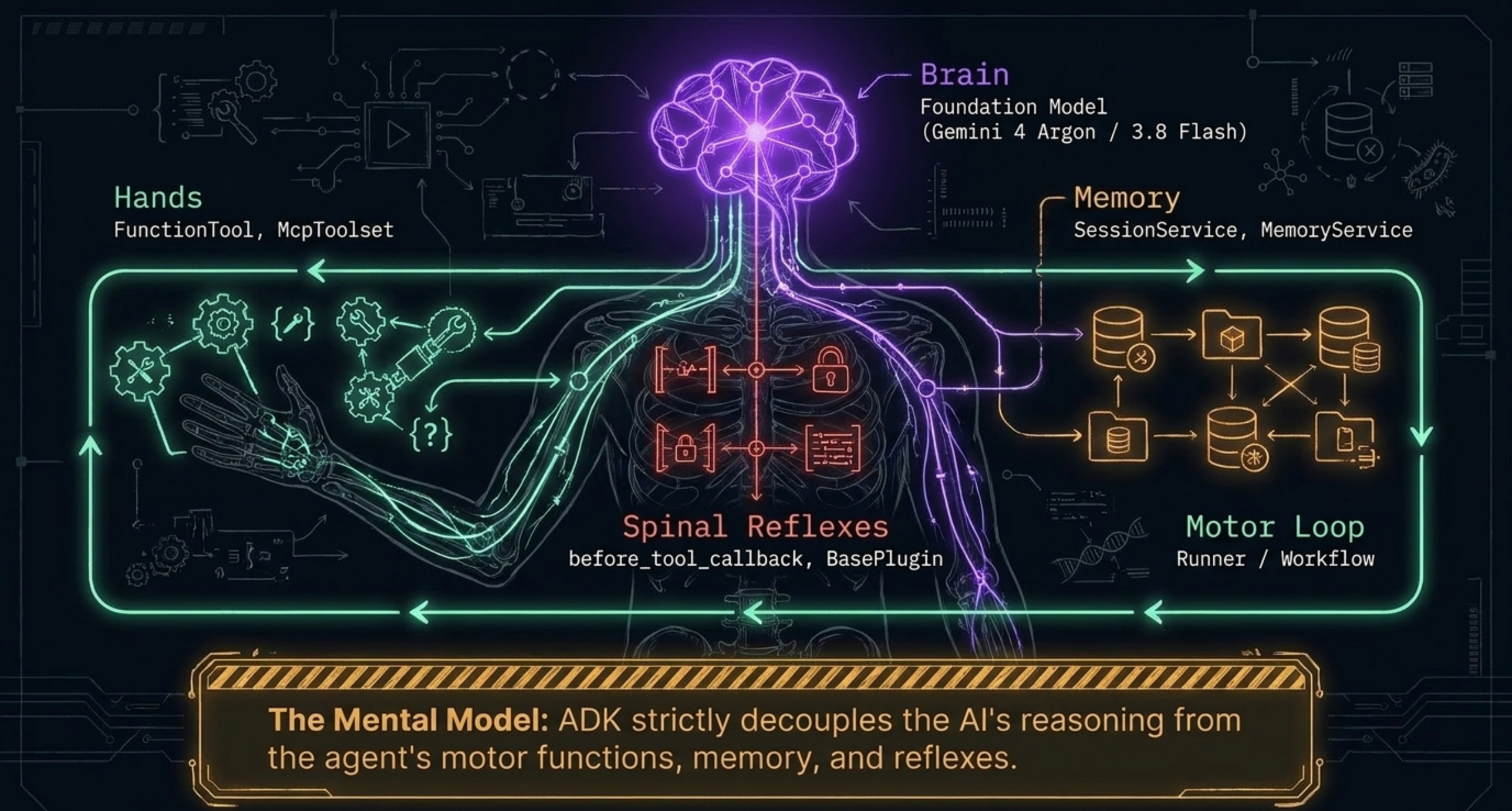
Unstructured

Lacks durable loops, fan-out joins, checkpoints, or iteration caps.

Ungoverned

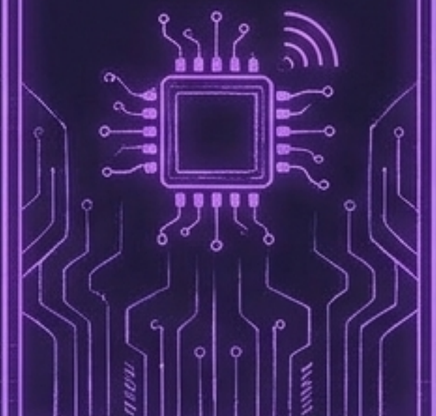
No tenant isolation, approval gates, or tool-boundary PII scrubbing.

The Void of the Raw Model: A bare model is a brain in a jar.
It requires an infrastructure harness to interact with the world safely.


Brain
Foundation Model
(Gemini 4 Argon / 3.8 Flash)
Hands
FunctionTool, McpToolset
Memory
SessionService, MemoryService
Spinal Reflexes
before_tool_callback, BasePlugin
Motor Loop
Runner / Workflow
The Mental Model: ADK strictly decouples the AI's reasoning from the agent's motor functions, memory, and reflexes.

1. The Blueprint
(LlmAgent / BaseLlm)

Completely stateless. Stores zero per-user attributes. One singleton safely handles thousands of concurrent requests.



2. The Engine
(Runner)

Executes turns, loads state, invokes the model, and enforces the commit-then-resume rule.



3. The Storage
(SessionService)

Manages persistent states keyed by app_name / user_id / session_id.

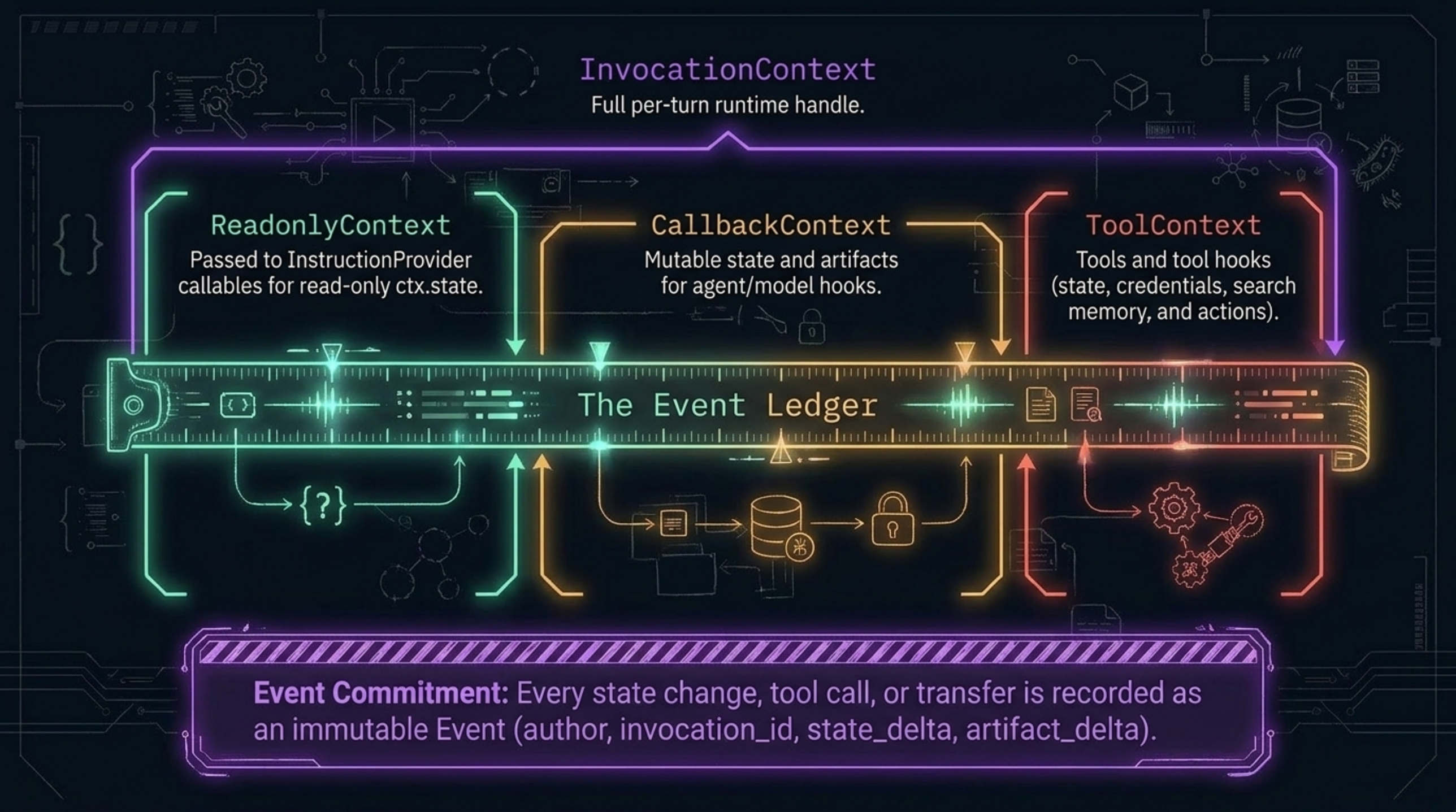


Interactions API

ManagedAgent

Offloads the loop, remote container sandbox, and server-side state entirely to Google hosted infrastructure.




InvocationContext
Full per-turn runtime handle.
ReadonlyContext
Passed to InstructionProvider callables for read-only ctx.state.
CallbackContext
Mutable state and artifacts for agent/model hooks.
ToolContext
Tools and tool hooks (state, credentials, search memory, and actions).
The Event Ledger
{?}
Event Commitment: Every state change, tool call, or transfer is recorded as an immutable Event (author, invocation_id, state_delta, artifact_delta).

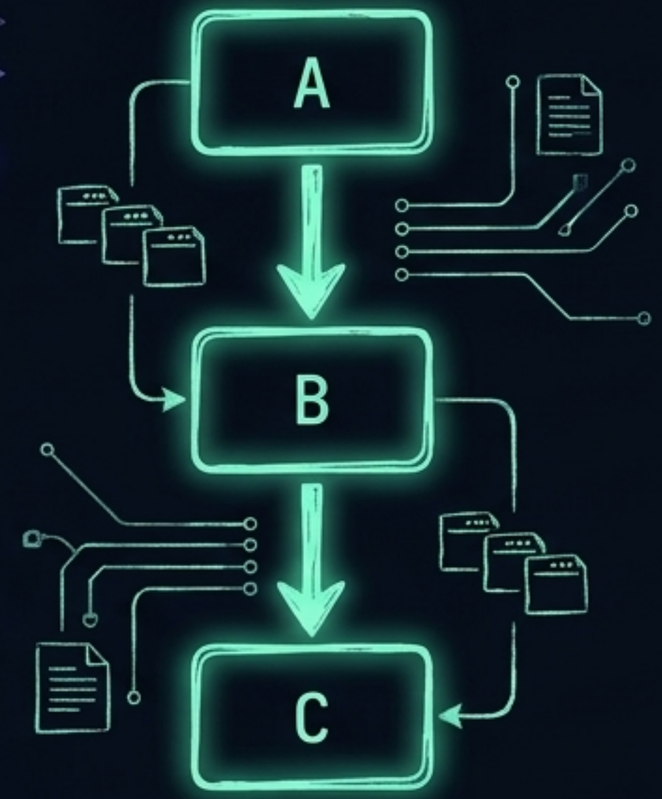
Evolution of Agency

Architectural Tier	Routing Decisions	External Actions
Single LLM Call (Best for Q&A / RAG)	✓ Single-step	✗ JSON intent only
Deterministic Workflow (Best for fixed pipelines)	✗ Code/edges decide	✓
Autonomous Agent (Best for dynamic tool finding)	✓ Multi-step loops	✓
ADK 2.0 Unified Framework	Provides deterministic workflows and autonomous loops under one event ledger.	

The Agent Theater Litmus Test

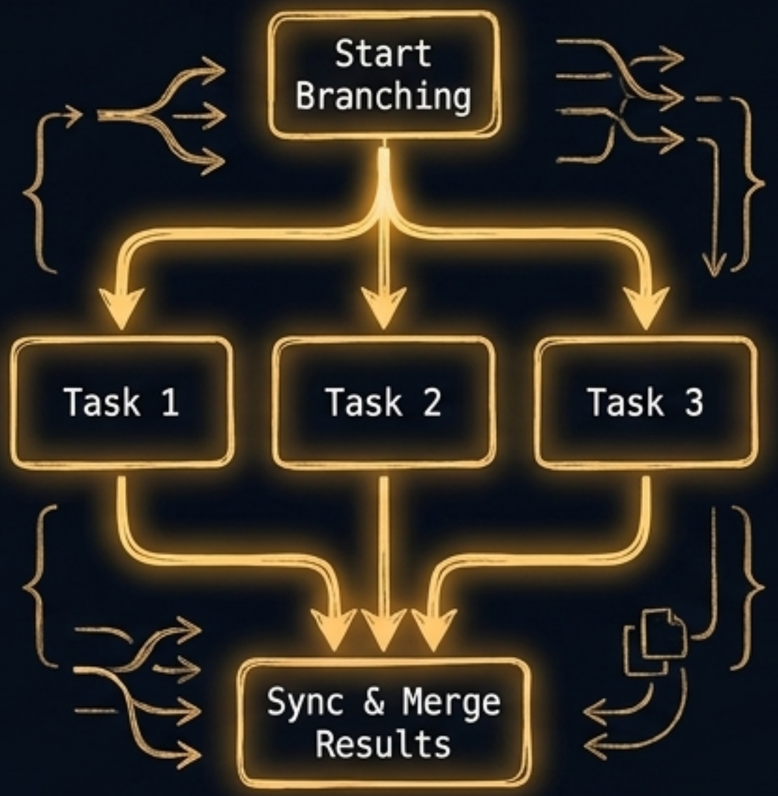
Splitting a system into multiple agents is justified only when specialists require distinct tool permissions, isolated data scopes, or explicit human sign-off boundaries.

SequentialAgent



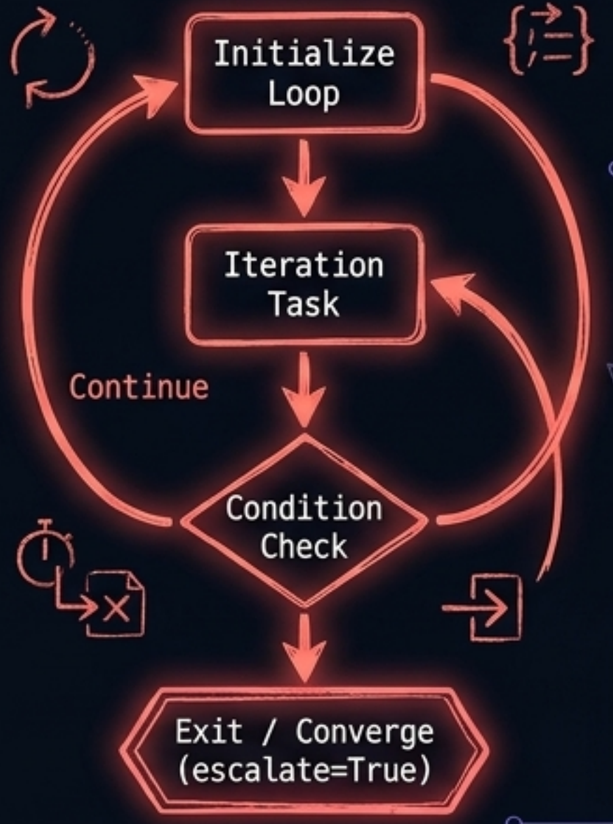
Rule: Always runs every child in sequence.

ParallelAgent



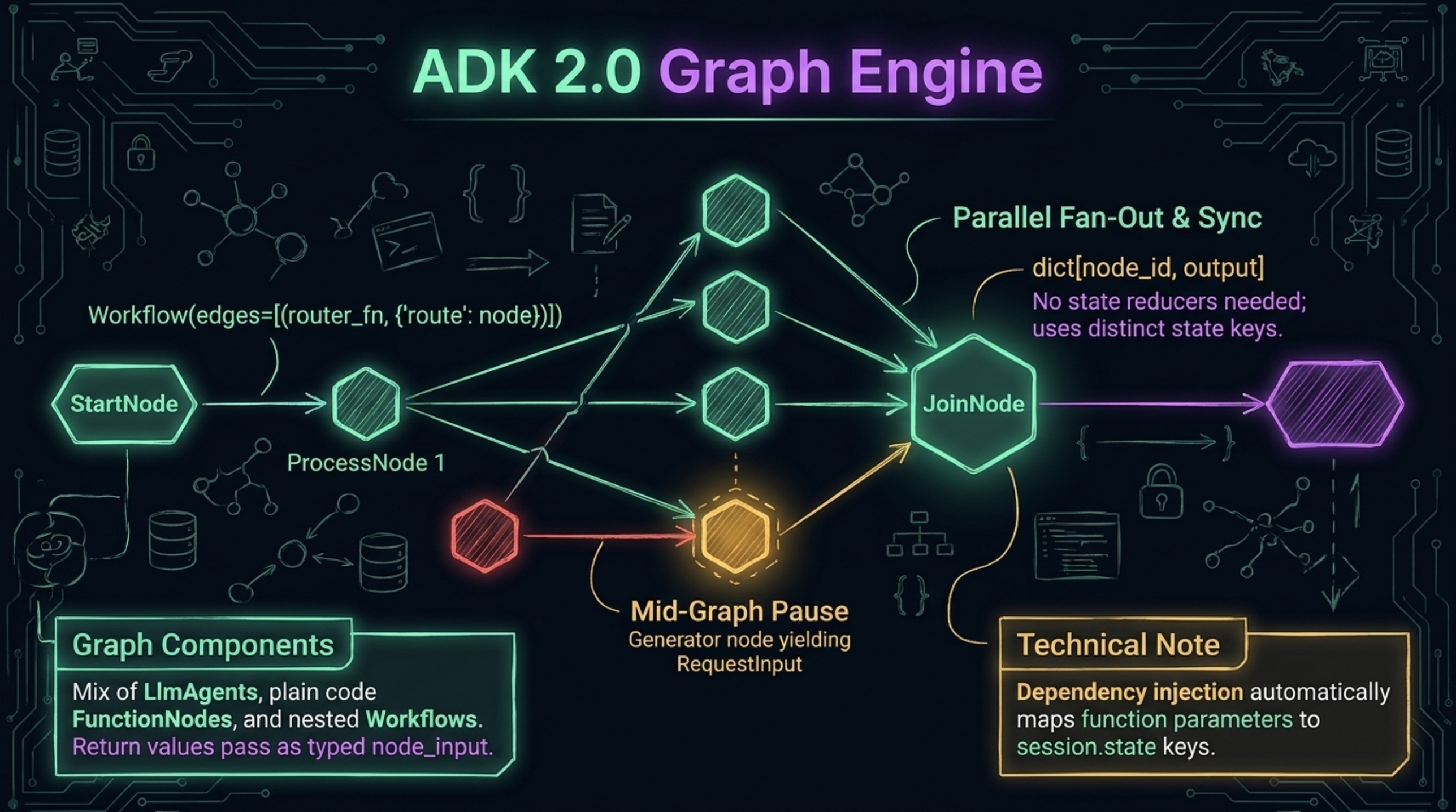
Rule: Concurrent execution, branch isolated, shared session.state. Requires unique output_key per branch.

LoopAgent



Rule: Reaching max_iterations exits silently. Requires explicit convergence flag (escalate = True) inside the exit tool.

ADK 2.0 Graph Engine



Graph Components

Mix of **LlmAgents**, plain code **FunctionNodes**, and nested **Workflows**. Return values pass as typed node_input.

Technical Note

Dependency injection automatically maps function parameters to session.state keys.

Planners & Thinking Levels

Thinking Level	Gemini 4 Argon	Gemini 3.8 Flash	Gemini 3.5 Flash-Lite
minimal	HTTP 400 Error	HTTP 400 Error	Supported & Default
low	Supported	Supported	Supported
medium	Supported (Default)	Supported (Default)	Supported
high	Supported	Supported	Supported

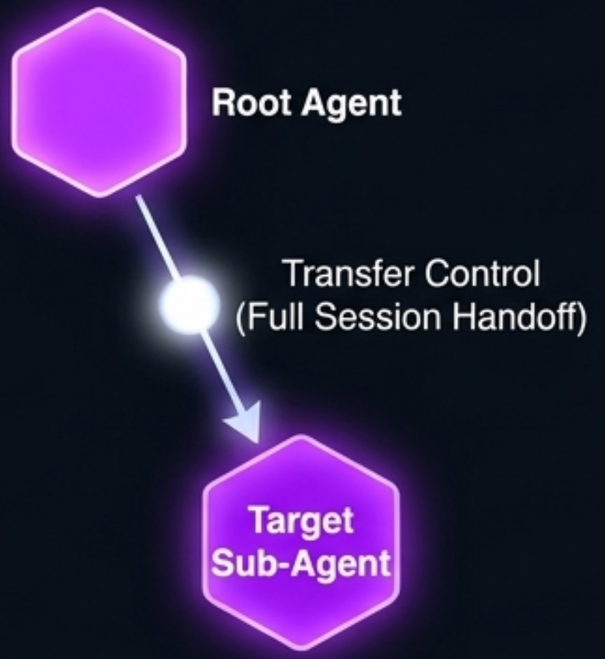
Sweet Spots

- **minimal**: Ultra-low-latency intent routers.
- **low**: Routine MCP tool execution, RAG synthesis.
- **medium**: Balanced production default for multi-step agentic tool loops.
- **high**: Ambiguous planning, code synthesis, conflict resolution.

WARNING: Passing `thinking_level='minimal'` or legacy integer `thinking_budget` on Gemini 4 Argon fails immediately with HTTP 400 `INVALID_ARGUMENT`. Argon ignores custom temperature/top_p.

Multi-Agent Delegation Patterns

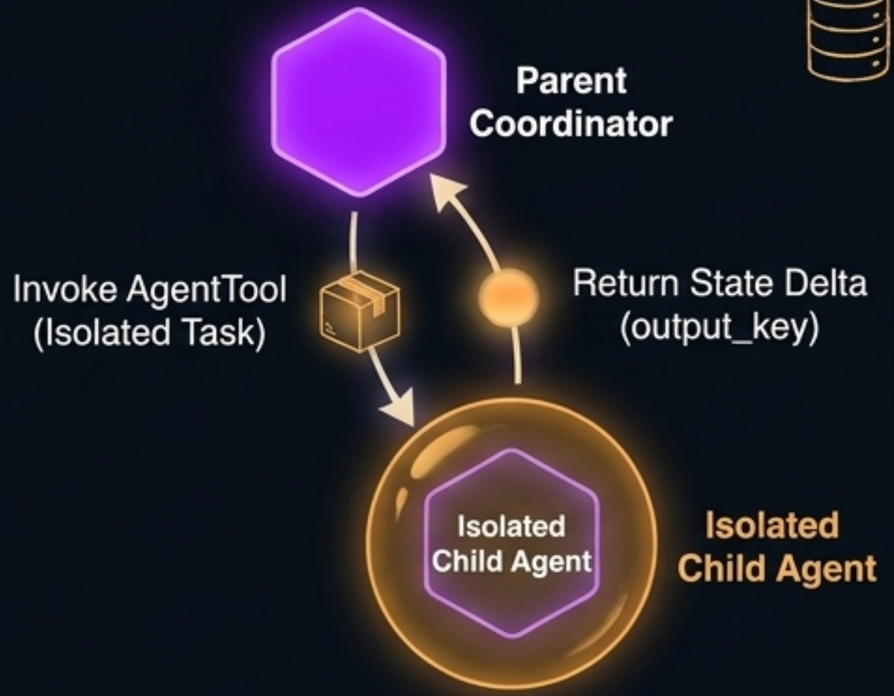
sub_agents (LLM Transfer)



Mechanics: Target sub-agent assumes full control and speaks to the user next. Shares the same session.events.

Fit: Full conversational handoff (e.g., triage to billing).

AgentTool (Sub-task Isolation)



Mechanics: Parent coordinator keeps control. Runs in an isolated child session, merging state_delta back via output_key.

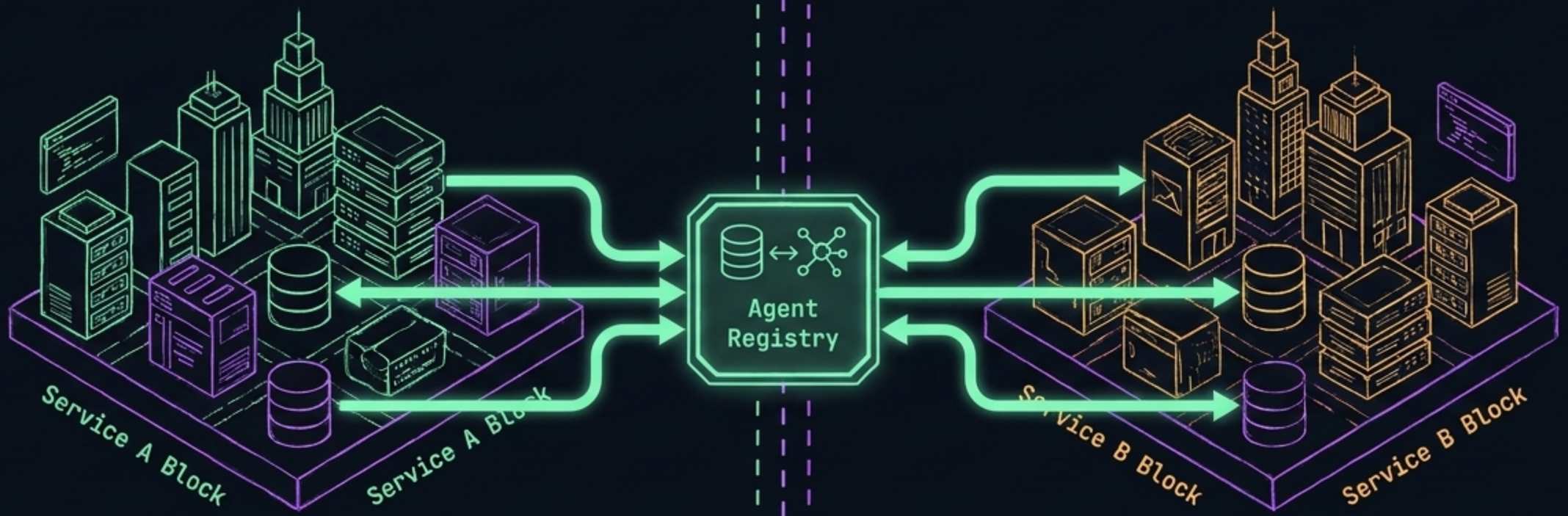
Fit: Modular execution without polluting the parent's context window.

Choosing the Orchestration Boundary

	Workflow Agents & Graph Workflows (In-Process)	Remote A2A Agents (Network)
State Sharing	Shares session.state	Nothing shared (only messages/artifacts cross the wire)
Deployment	Single process / deployable unit	Separate service deployment
Isolation & Faults	Zero fault/identity isolation. A crash takes down all children. All share one IAM principal.	Strong isolation. Network round-trip latency. Each peer has its own Agent Identity principal.
Governance	Governed by local ADK callbacks.	Governed by Agent Gateway policies.

Privilege Mixing Anti-Pattern

Do not attach a read-only researcher and a refund-issuing specialist in the same process. They will share the same principal.



1. Expose → 2. Deploy → 3. Register → 4. Discover

1. Expose — to_a2a(root_agent) builds the Agent Card and publishes it at /.well-known/agent-card.json.

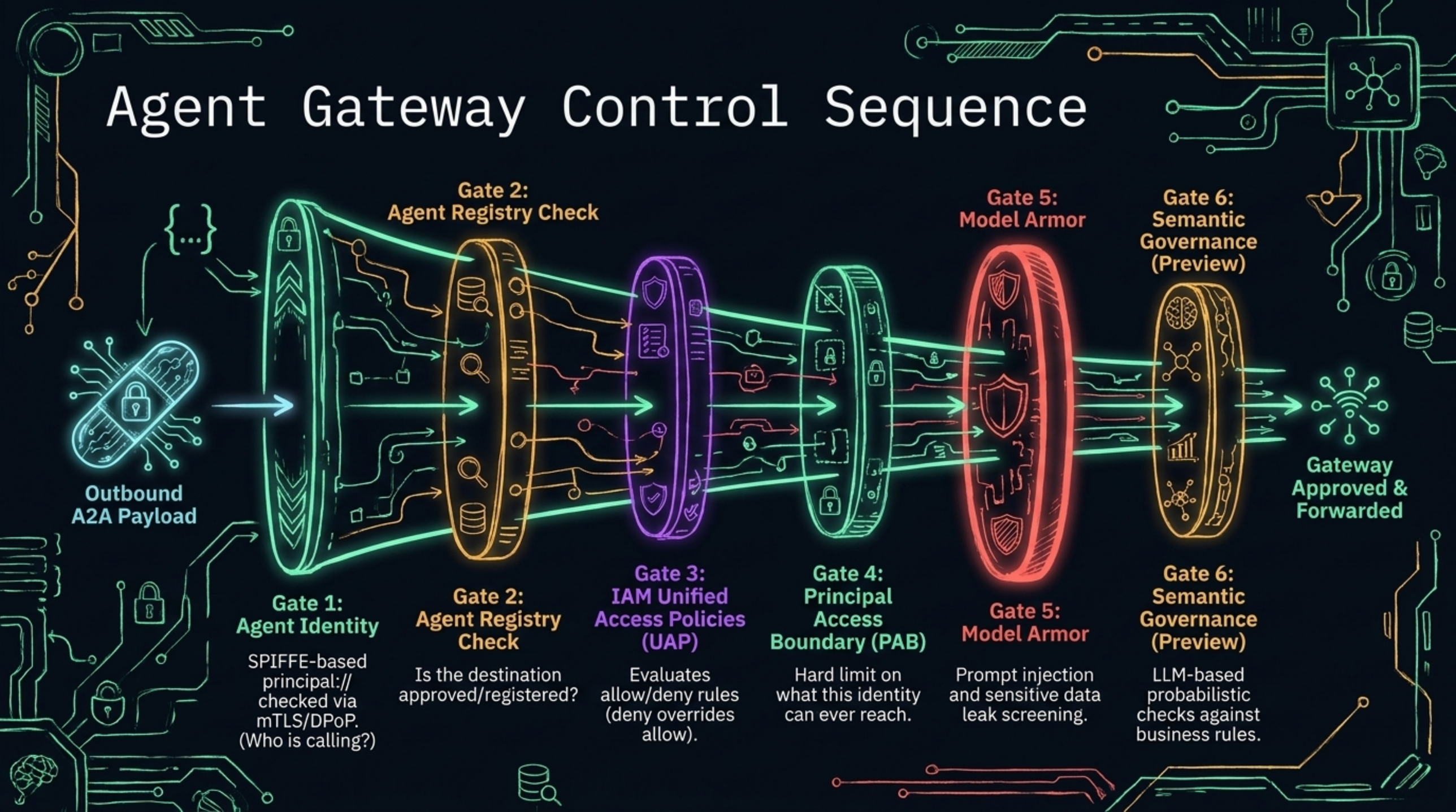
2. Deploy — Host on Agent Runtime with a per-agent identity (SPIFFE).

3. Register — Agent Runtime peers register automatically. External runtimes register manually via CLI.

4. Discover — Call AgentRegistry.get_remote_a2a_agent() to resolve peers at runtime.

Anti-Pattern Alert: Hardcoding peer URLs skips the registry, silently breaking on endpoint changes and bypassing gateway policy evaluation.

Agent Gateway Control Sequence



Gate 1: Agent Identity

SPIFFE-based principal:// checked via mTLS/DPoP. (Who is calling?)

Gate 2: Agent Registry Check

Is the destination approved/registered?

Gate 3: IAM Unified Access Policies (UAP)

Evaluates allow/deny rules (deny overrides allow).

Gate 4: Principal Access Boundary (PAB)

Hard limit on what this identity can ever reach.

Gate 5: Model Armor

Prompt injection and sensitive data leak screening.

Gate 6: Semantic Governance (Preview)

LLM-based probabilistic checks against business rules.



Storage Tiers

Tier 1: Session Events

Chronological immutable turn/tool transcript.

Tier 2: Scoped State

Exact structured key/value scratchpad.

Tier 3: Semantic Memory

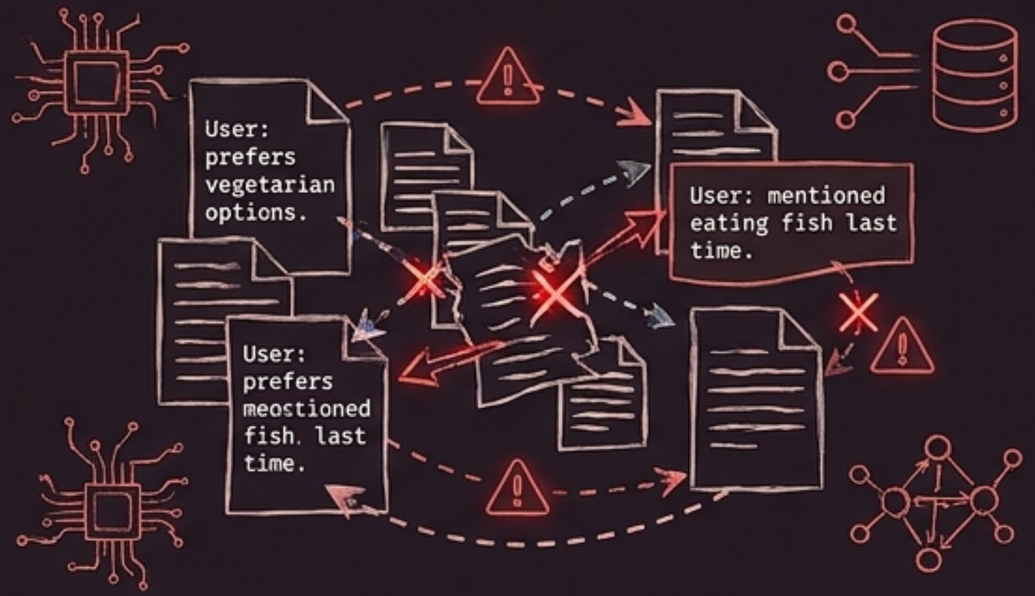
LLM-distilled cross-session facts.

Tier 4: Artifacts

Versioned binary blobs (PDFs, images).

Long-Term Recall Mechanisms

VertexAiRagMemoryService

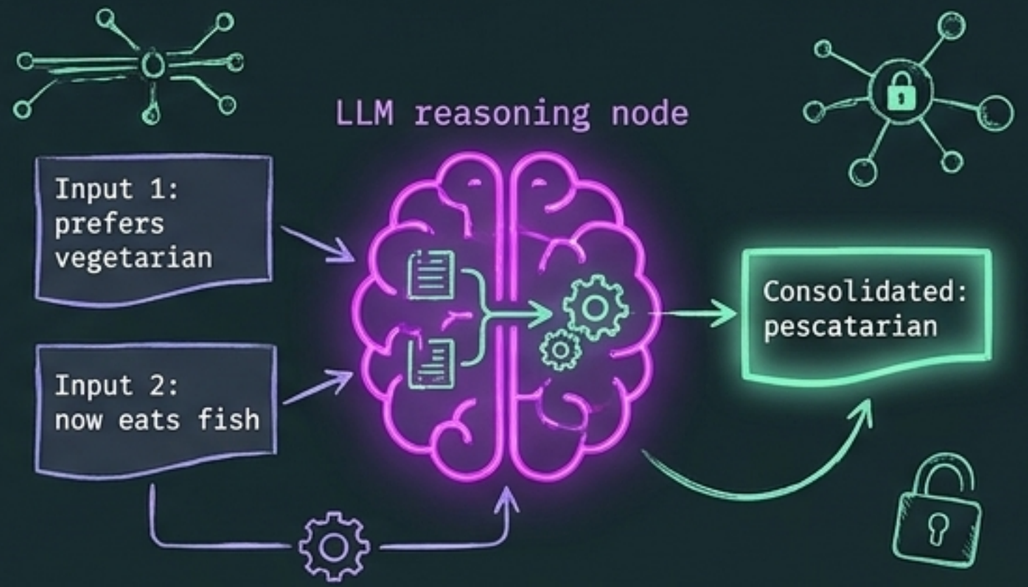


RAG indexes raw session chunks. It does NOT reconcile conflicting historical statements.

Mechanism

{ } → Recall via load_memory (on-demand/cheaper).

VertexAiMemoryBankService

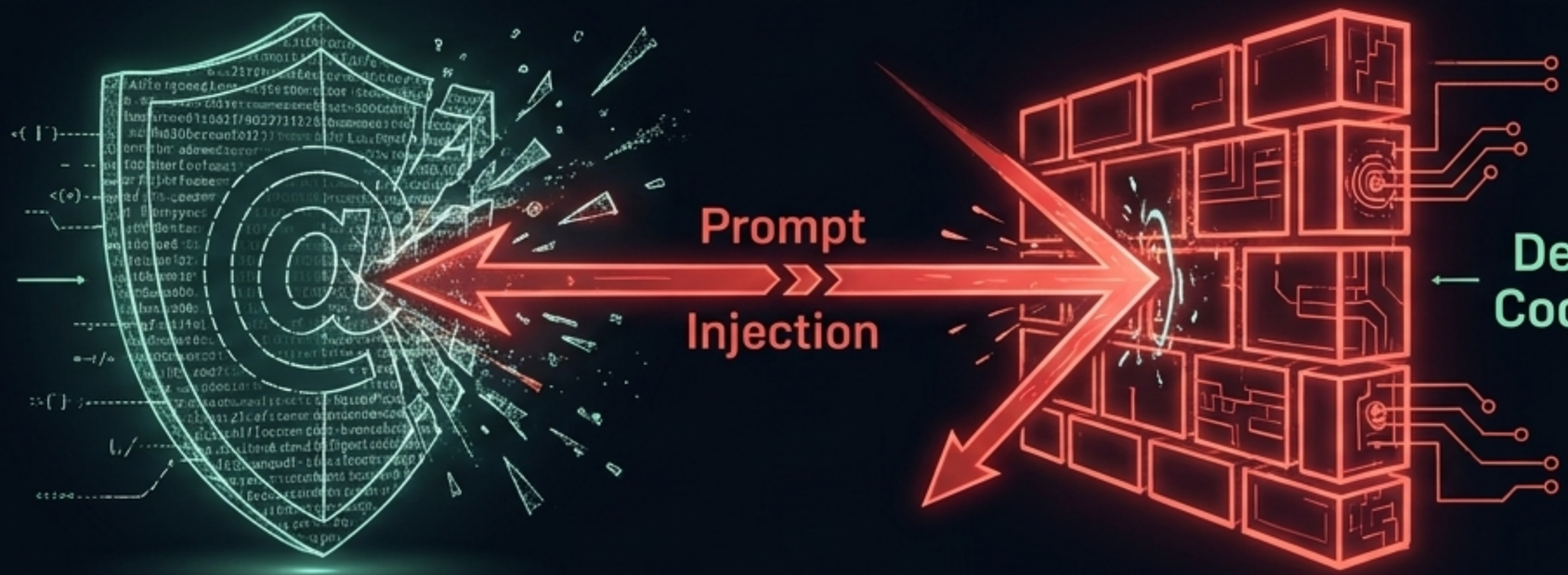


Triggers GenerateMemories in after_agent_callback. Actively revises and consolidates stored memory rather than returning stale facts.

Mechanism

{ } → Recall via PreloadMemoryTool (injected automatically every turn).

The Deterministic Harness Principle



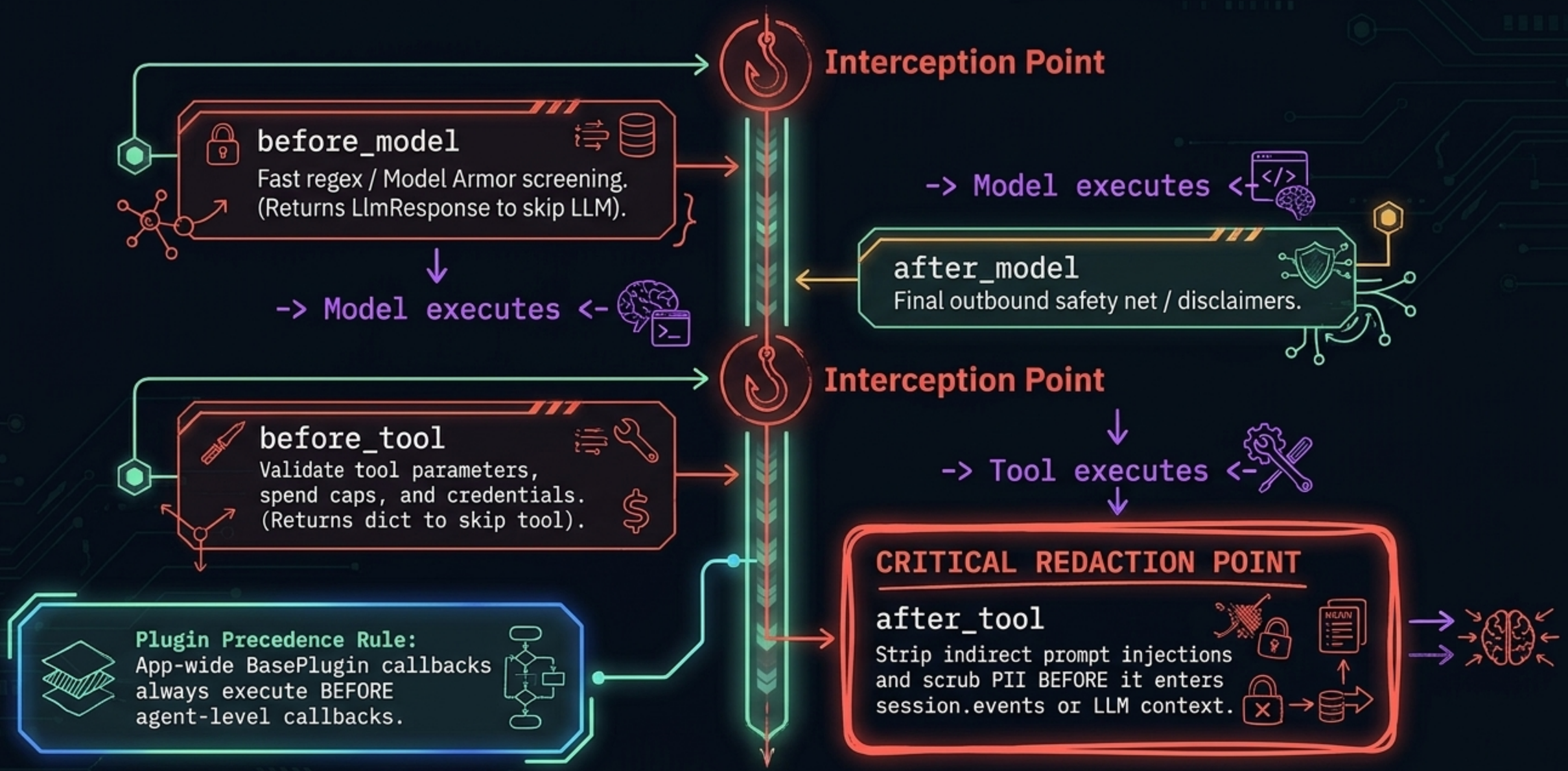
The Principle:

Never rely on system-prompt wording (e.g., "Please redact SSNs and never refund over $500") for security, compliance, or budget caps.

The Solution:

Enforce hard invariants in deterministic ADK callbacks and BasePlugin components surrounding the model. Code > Prompts.

The Callback Sequence Pipeline



Code Execution Isolation Tiers



UnsafeLocalCodeExecutor

Runs exec() inside the host process with full IAM/env access.
Throwaway local debugging only. Never staging or prod.

ContainerCodeExecutor

Dedicated Docker container for local staging.

BuiltInCodeExecutor

Gemini server-side sandbox. Stateless, 30s max, zero network egress.

AgentEngineSandboxCodeExecutor

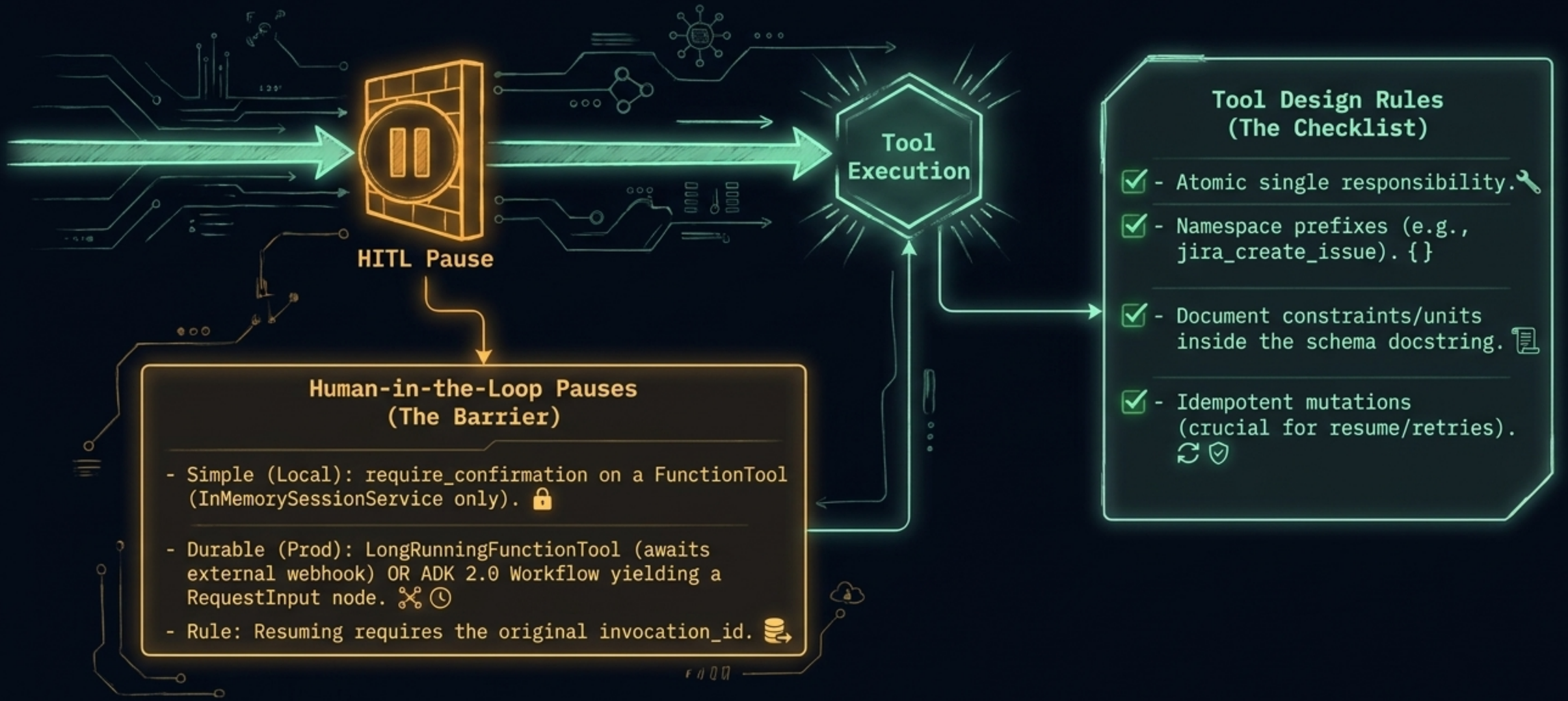
Managed Agent Runtime. Stateful files up to 14-day TTL. (Default Prod choice).



GkeCodeExecutor

Pre-warmed GKE Agent Sandbox backed by gVisor kernel isolation.

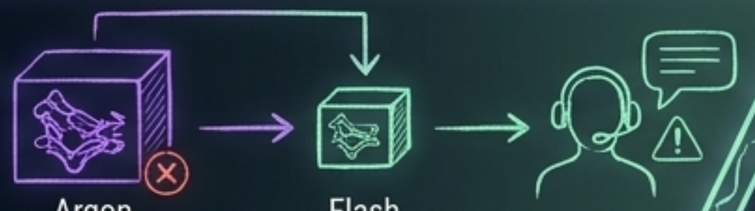
HITL Pauses & Tool Design Checklist



The 3-Layer Fault Recovery Ladder

Model Failover & Escalation

Catch outages in **on_model_error_callback** to trigger failovers (e.g., Argon -> Flash) and escalate to a human when max retries are exhausted.



{...} Tool Argument Self-Correction

`ReflectAndRetryToolPlugin`

Critical Nuance: Tools should return **{'status': 'error'}** dicts instead of raising exceptions. Override **extract_error_from_result** on the plugin so structured dicts trigger LLM reflection.



Transport Layer (HTTP 429/503)

`HttpRetryOptions` with exponential backoff on the model wrapper.



The Cybernetic Blueprint



Synthesis: Production mastery is not about using the biggest model for everything; it is about orchestrated, deterministic delegation mapped over a resilient, event-sourced nervous system.