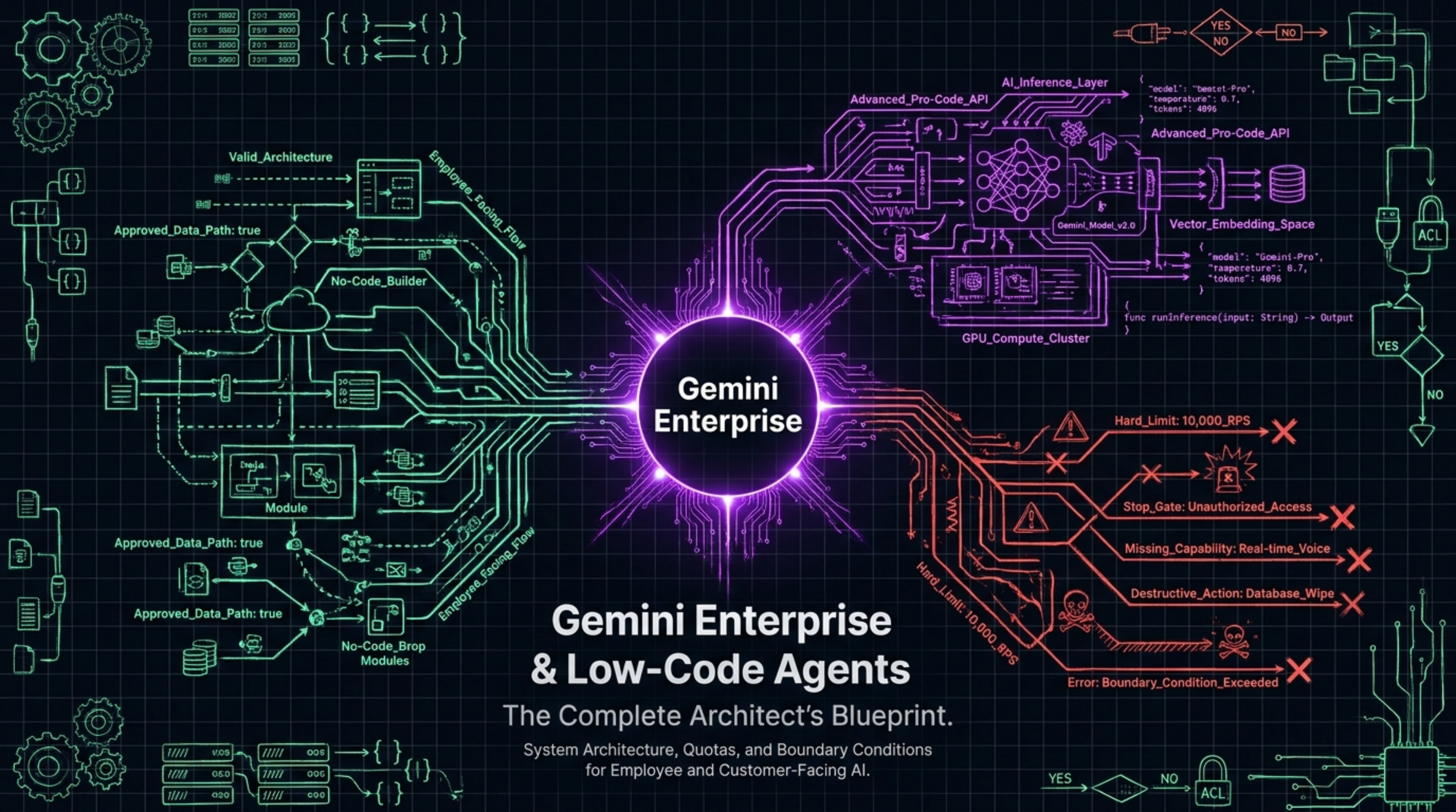


Gemini Enterprise & Low-Code Agents

The Complete Architect's Blueprint.

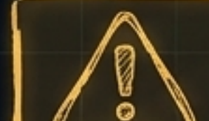
System Architecture, Quotas, and Boundary Conditions
for Employee and Customer-Facing AI.

The One-Line Mental Model



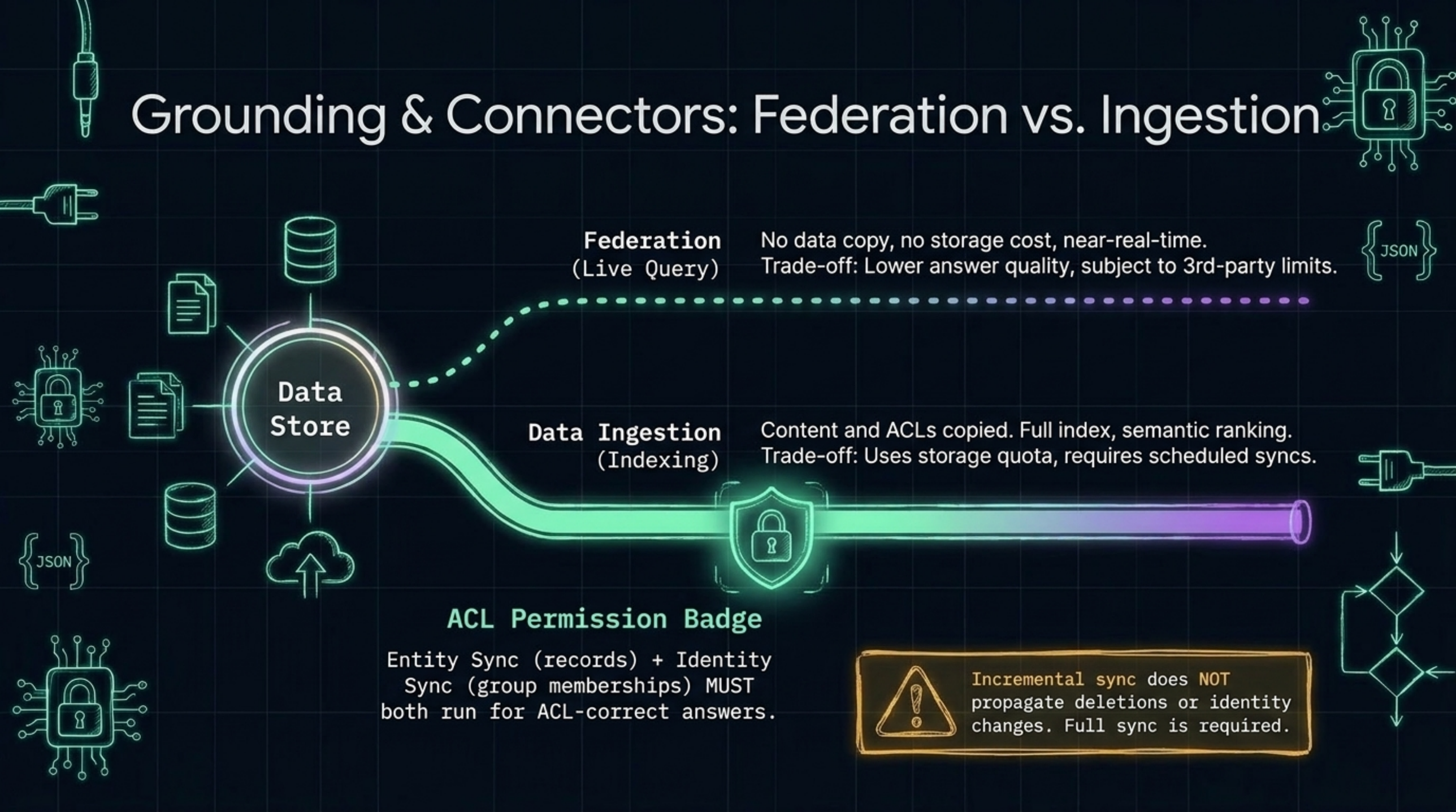
Edition Map: The “What Blocks Me” Matrix

	Business (1-500 seats, 25 GiB)	Standard (1+ seats, 30 GiB)	Plus (1+ seats, 75 GiB)	Frontline (2 GiB)
Deep Research	✓	✓	✓	✓
Data Insights / Agent Marketplace	🔒	✓	✓	🔒
Custom Full-Code Agents (ADK/A2A)	🔒	✓	✓	🔒
Agent Creation Quota		Standard (1/user/day)	Plus (10/user/day)	0



The Business Edition Dead End: Incompatible with IAM, VPC-SC, CMEK, custom A2A agents, and the full connector catalog.

Grounding & Connectors: Federation vs. Ingestion



ACL Permission Badge

Entity Sync (records) + Identity Sync (group memberships) MUST both run for ACL-correct answers.

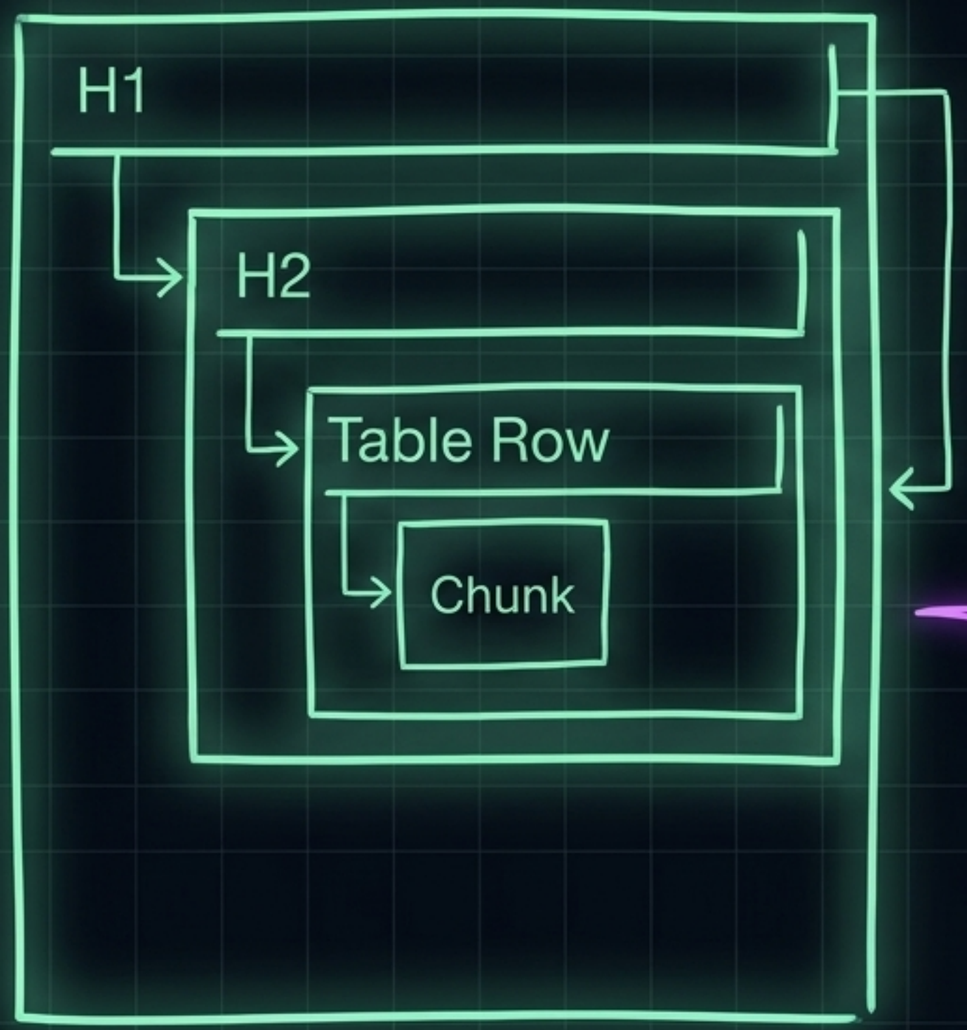
> ⚠ Incremental sync does NOT propagate deletions or identity changes. Full sync is required.

Parser Selection Matrix

Layout Parser (Default/Recommended)		Preserves tables, lists, headings. Generates descriptions for images/tables. Handles PDF, HTML, DOCX, PPTX, XLSX.
OCR Parser		Recognizes page text from non-searchable images/scans (up to 500 pages per PDF). Flattens complex tables.
Digital Parser		Extracts plain machine-readable text. Strips structure. (Acts as the API fallback if a file type isn't supported).
Bring Your Own JSON		Accepts pre-parsed JSON from external pipelines (e.g., Doc AI layout processors).

API Trap: Creating a store via API without layoutParsingConfig defaults to the Digital Parser. Explicitly request Layout.

Layout-Aware Chunking for RAG



Chunking Diagnostics:

- **Search Stores:** Default 500 tokens (range 100–500). Smaller chunks = tighter fact matches.
- **Doc AI / RAG Engine:** 500–1024 tokens with 100–256 token overlap.

```
include_ancestor_headings = true
```

H1 Header

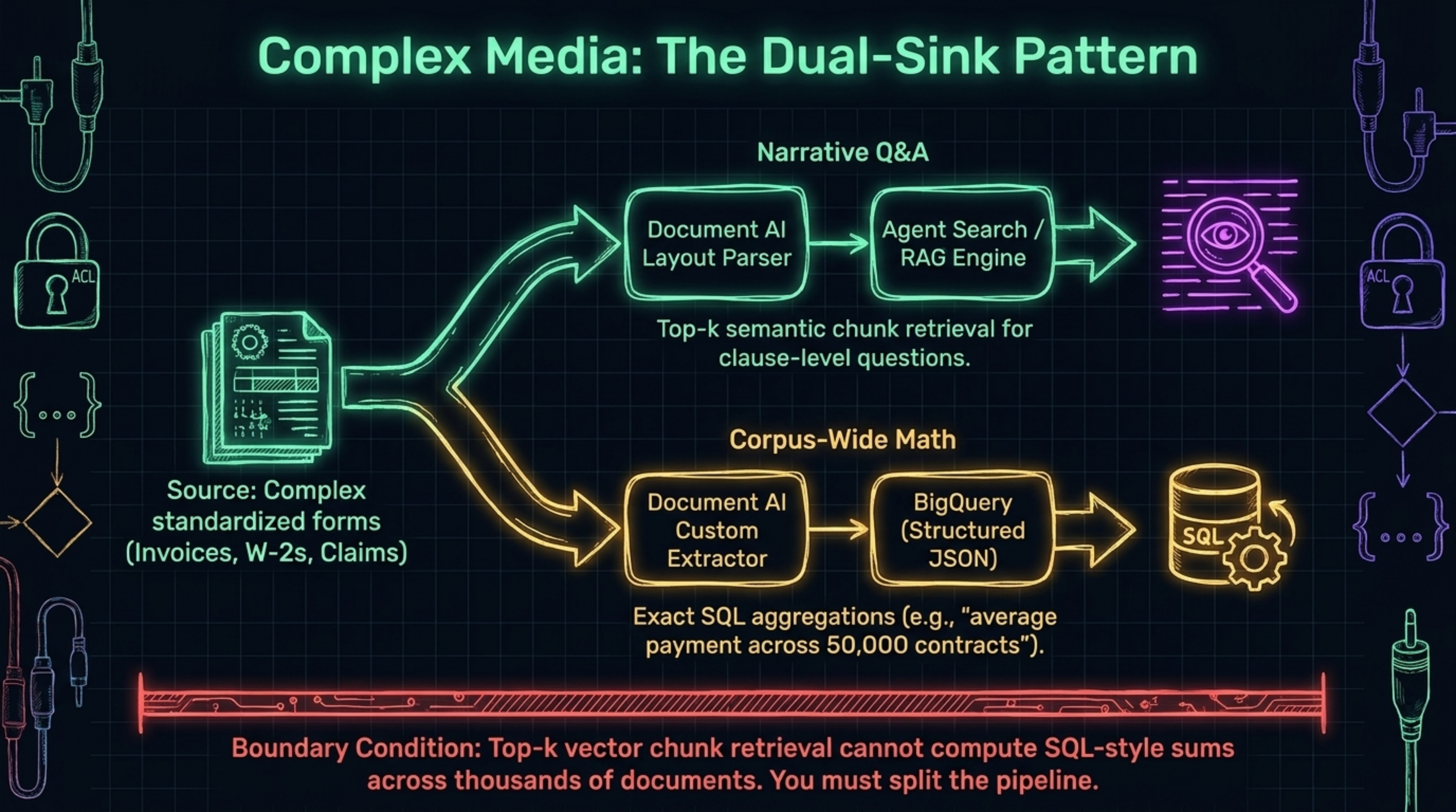
Table Title

Split Table Row Chunk

Prepends document title and headings so mid-document clauses and split table rows retain context for the LLM.

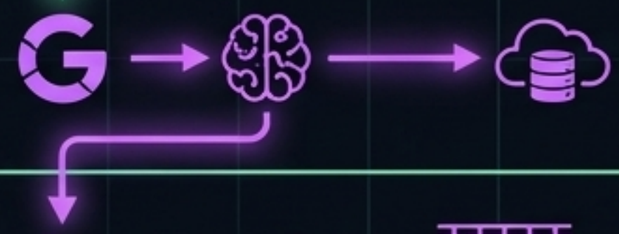
ACL

Complex Media: The Dual-Sink Pattern



Boundary Condition: Top-k vector chunk retrieval cannot compute SQL-style sums across thousands of documents. You must split the pipeline.

Processing Audio & Video Pipelines

Speech-to-Text V2 (Chirp 3)		for ACL-aware transcripts. Transcripts ingested as HTML/TXT. Preserves native Workforce Identity ACLs.
Video Intelligence API		for deterministic shot boundaries and OCR to JSONL/BigQuery. Emits labels, requires LLM synthesis.
Gemini Direct Multimodal		for rapid ad-hoc reasoning. Zero index. Reads raw media from Cloud Storage. Uses `audio_timestamp=True`. Max ~1hr video. No per-document ACLs.
Agent Retrieval + embeddings		for visual match. `gemini-embedding-2` vectors. Max 8,192 tokens (~120s MP4). Requires pre-segmented clips and explicit query-time ACL filters.

Anti-Pattern: Media stores are NOT video brains (only catalog metadata).

Anti-Pattern: Never send raw 60-min videos direct to embeddings (truncation at 8k tokens, segment first).

The Agent Ecosystem & Gallery



Agent States

- Enabled: Visible & Interactive
- Suspended: Visible in gallery but inactive—good for temporary freezes.
- Disabled: Hidden entirely from users.

ACL

JSON

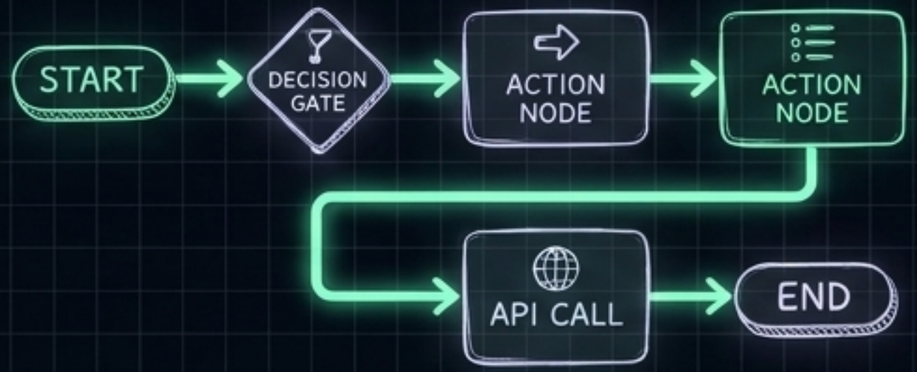
Workflow Builder: Chat vs. Workflow Agents

Chat Agents



- Conversational Designer
- Non-deterministic, open-ended Q&A.
- Configured via global Details Pane.

Workflow Agents



- Visual Editor Canvas
- Ordered steps, rigid branching, and variables passed between nodes.
- Configured via Canvas Nodes with step-level schemas, inline few-shot examples, and explicit JSON output contracts.

Human-in-the-Loop (HITL) Gates



Native HITL Step: Request Info

Pauses workflow to collect structured form inputs or missing parameters from a human before resuming.

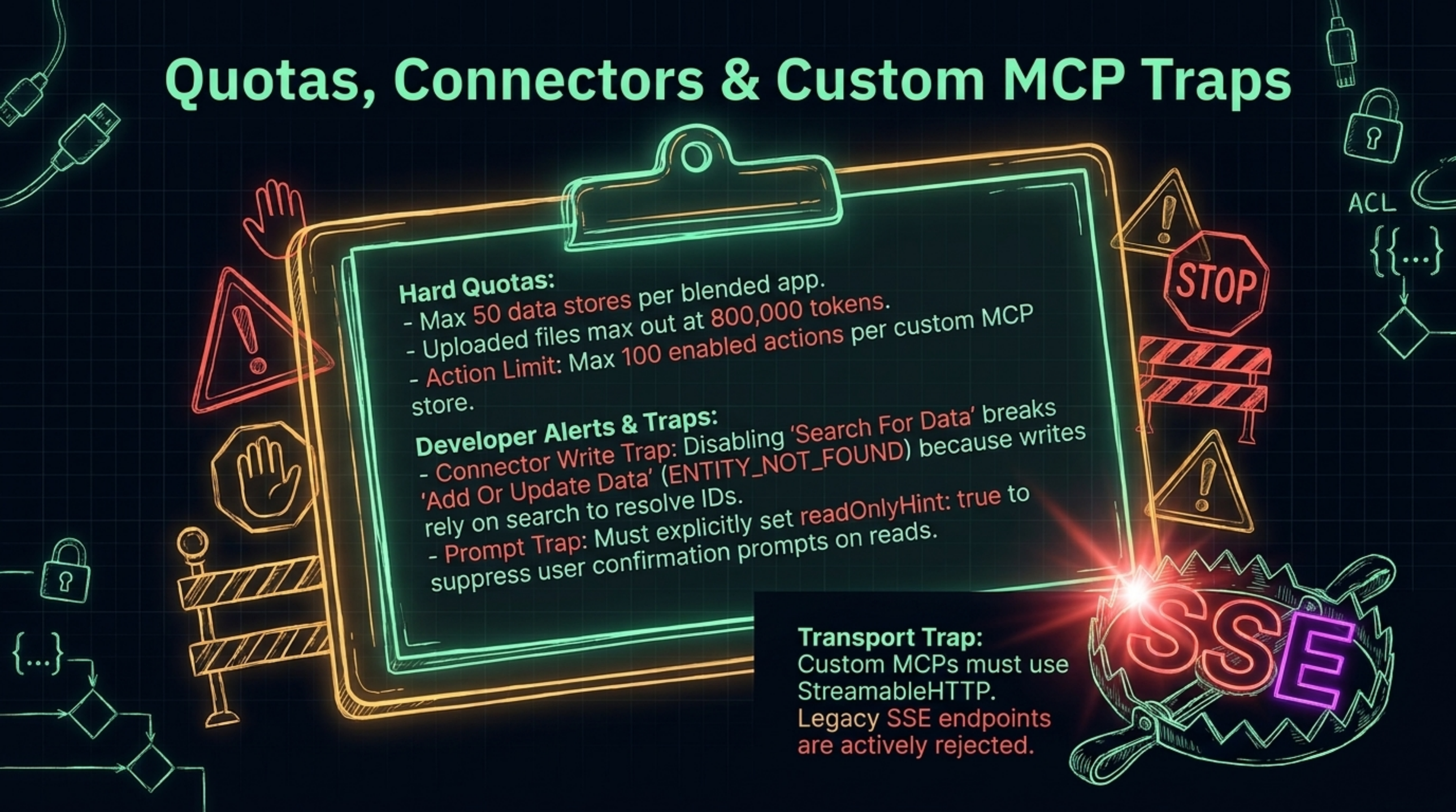
Native HITL Step: Approval

Explicit Approve/Reject gate.

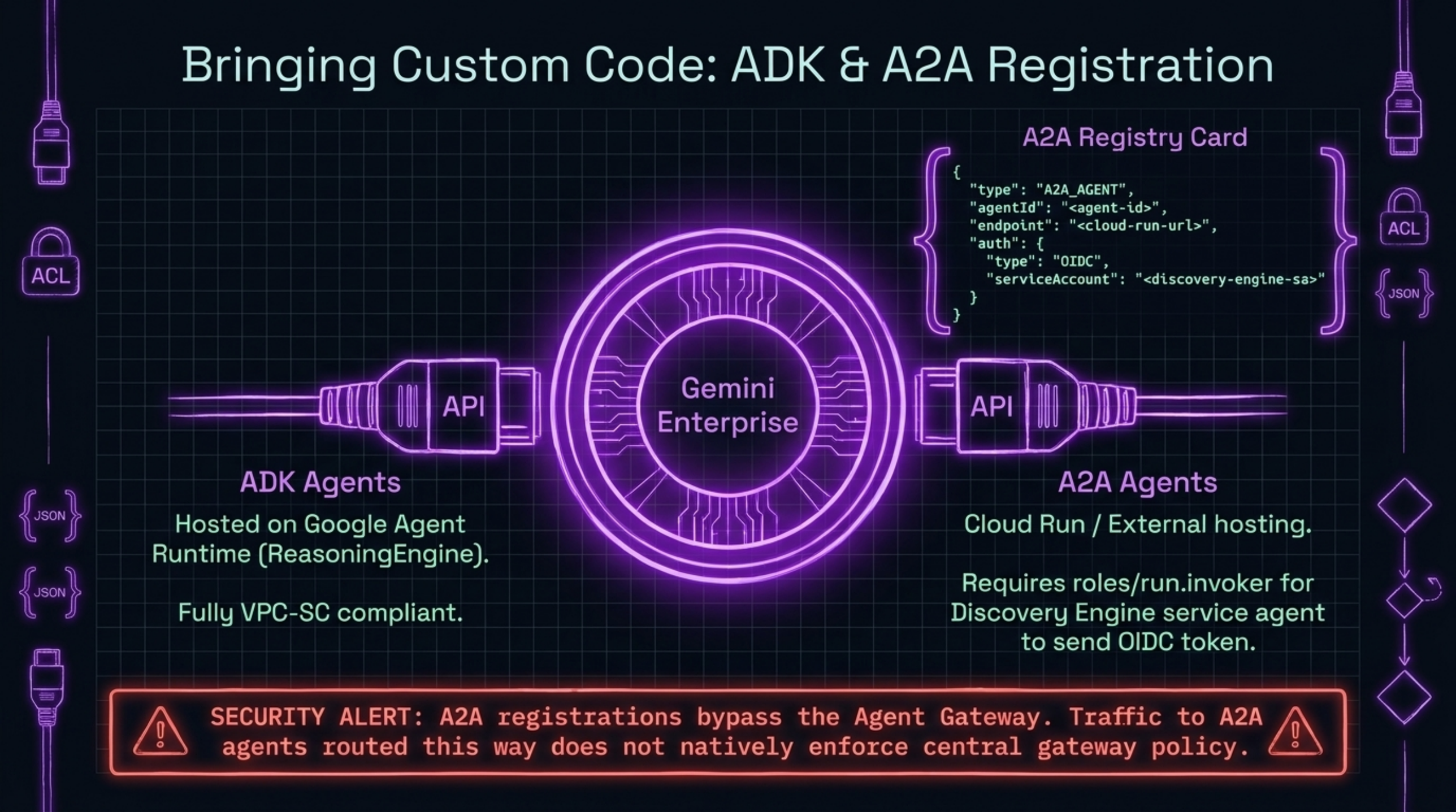
Mandatory architectural rule: Must be placed before any mutative connector actions (e.g., Salesforce updates).

Execution State: Paused in ‘Needs Review’ status inside the global Runs tab.

Quotas, Connectors & Custom MCP Traps



Bringing Custom Code: ADK & A2A Registration



Layered Governance & Model Armor



Trap Note: Console Model Armor screens the native assistant ONLY (`INSPECT_AND_BLOCK`). Custom agents must implement guardrails in code.

Traffic Layer

Agent Gateway (`REQUEST_AUTHZ`) and Model Armor.

Agent Layer

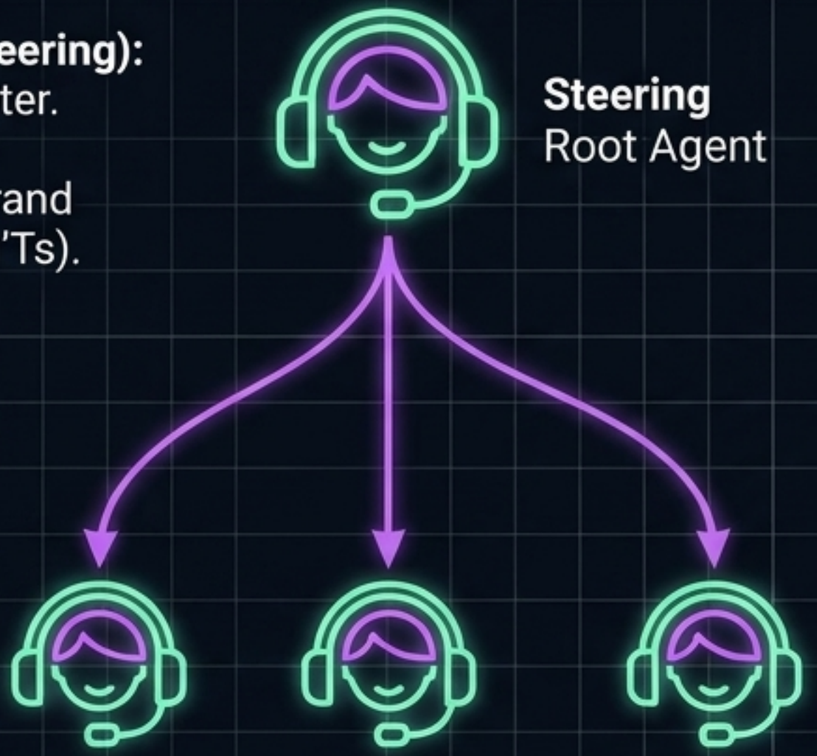
Agent Gatevsioned connectors, and Model Armor.

Network Layer

VPC Service Controls protecting Gemini Enterprise, Agent Runtime, and data sources.

CX Agent Studio: Customer-Facing Architecture

Root Agent (Steering):
The central router. Inherits global instructions (brand tone, DOs/DON'Ts).



Sub-Agents:
Modular, conditional components specialized by task.

Instruction Engineering

Built on ADK.

Uses explicit XML tags for structure:
<role>, <primary_goal>, <taskflow>, <step>, <action>

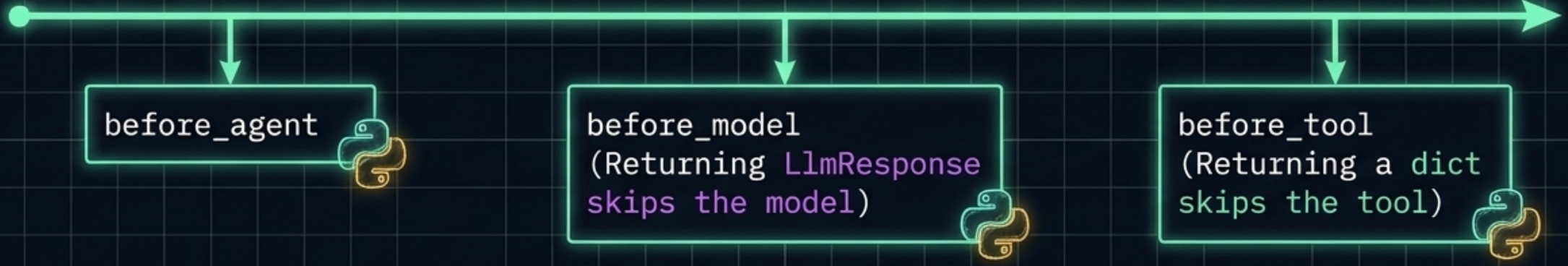
Tool referencing syntax:
{@TOOL: name} and {@AGENT: name}

Rule: Use inline few-shot examples sparingly; rely on clear XML taskflow first.

CX Agent Studio: Tools & Callbacks

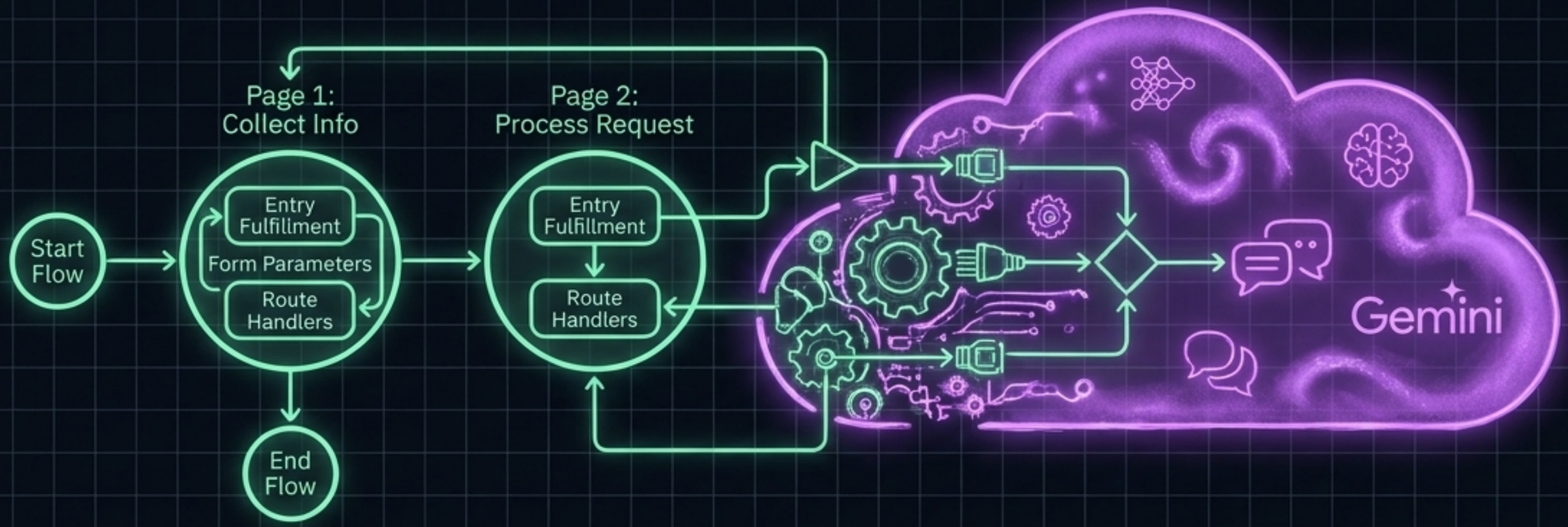
Synchronous Tools: **< 5s** (Data stores, OpenAPI, SaaS actions).

Asynchronous Tools: **~5-60s** (Agent keeps talking while tool processes).



Evaluation Suites (The 3 Metrics)		
Scenario: AI-simulated open-ended robustness.	**Simulation**: Ordered milestones for multi-step goals.	**Golden**: Regression testing with semantic match **(0-4 score)**.

Dialogflow CX Flows & State Handlers



JSON

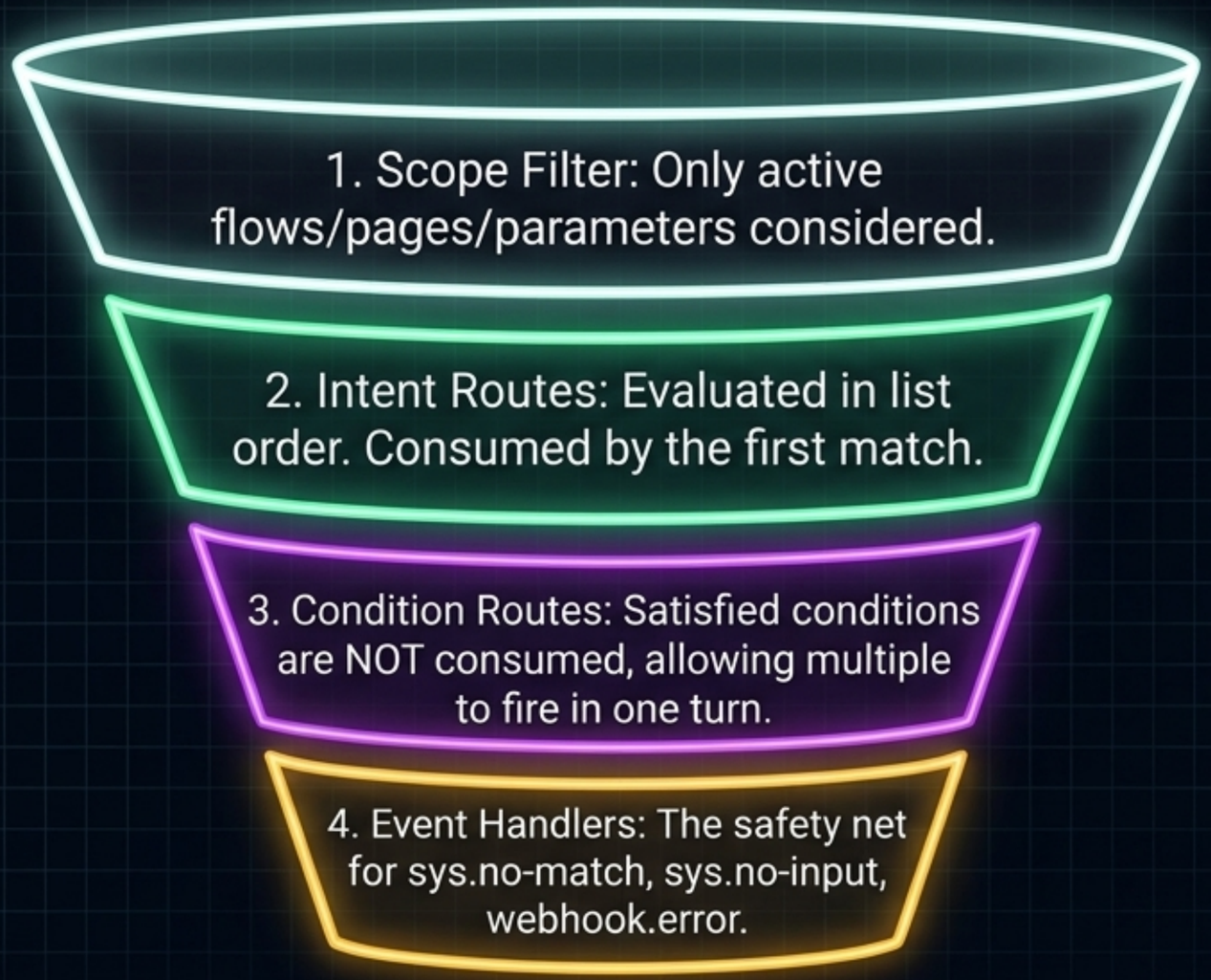
Deterministic Hierarchy: Flows -> Pages -> State Handlers -> Form Parameters

One Active Page Rule:
A session sits on exactly one page, executing its entry fulfillment, evaluating form parameters, and checking routes.

Form Filling:
Parameters pre-fill from session state. Missing slots trigger reprompt handlers.
Page exit relies on condition **$page.params.status = 'FINAL'.**

sys.invalid-parameter:
A crucial event handler where server-side webhooks reject invalid LLM guesses (e.g., a hallucinated account number).

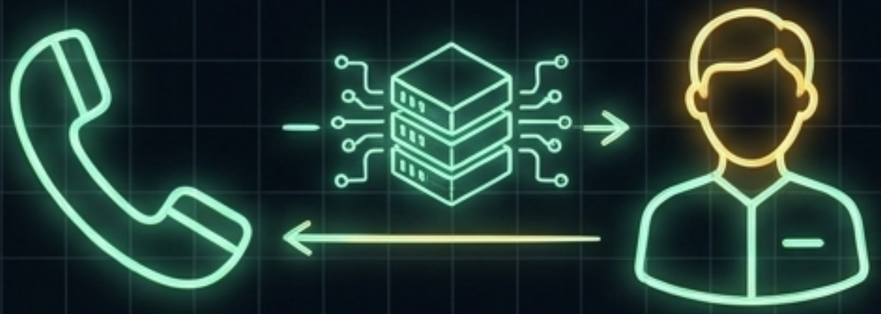
Form Filling & Evaluation Order



Stop Rule:

A handler with a transition target terminates evaluation; a handler without one queues its fulfillment and lets evaluation continue.

Human Handoff & Playbooks



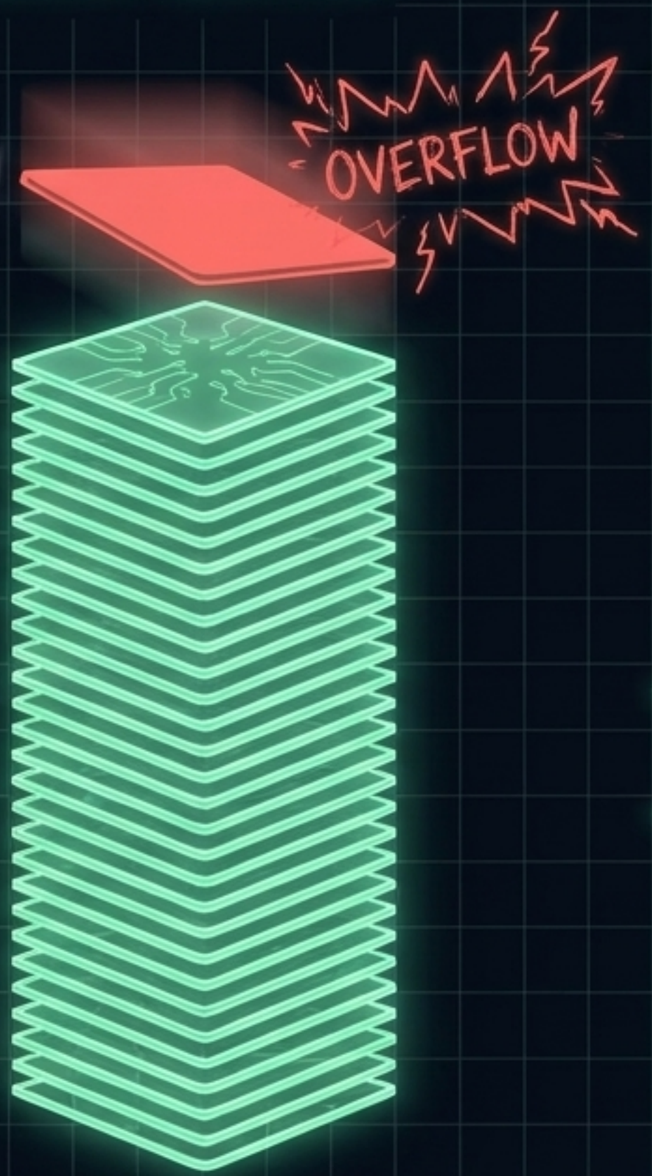
```
{
  "escalationMessage": "Transferring
you to billing...",
  "liveAgentHandoff": {
    "endSession": false,
    "sipRefer": true
  }
}
```

- The `end_session` system tool:
- Required on the root agent and any escalating sub-agent.
- Agent Assist Trap: Keep `liveAgentHandoff.endSession = False` to transfer the conversation to a human with Agent Assist context. Setting it to `True` drops the AI assist.
- Playbooks (Generative) vs Flows (Deterministic).

Symbolic Targets & The Flow Stack Limit

Symbolic Targets:

- `START_PAGE`, `CURRENT_PAGE`, `PREVIOUS_PAGE`.
- `END_FLOW` (Pops the active flow, returns to caller).
- `END_SESSION` (Clears session, restarts next input).



The Flow Stack Limit = 25

Consequence: Going deeper drops the base flows. An `END_FLOW` command will return the user somewhere unexpected, or end the session **entirely** if the stack empties. Keep sub-flow chains **shallow**.

The Low-Code vs Pro-Code Decision Matrix

