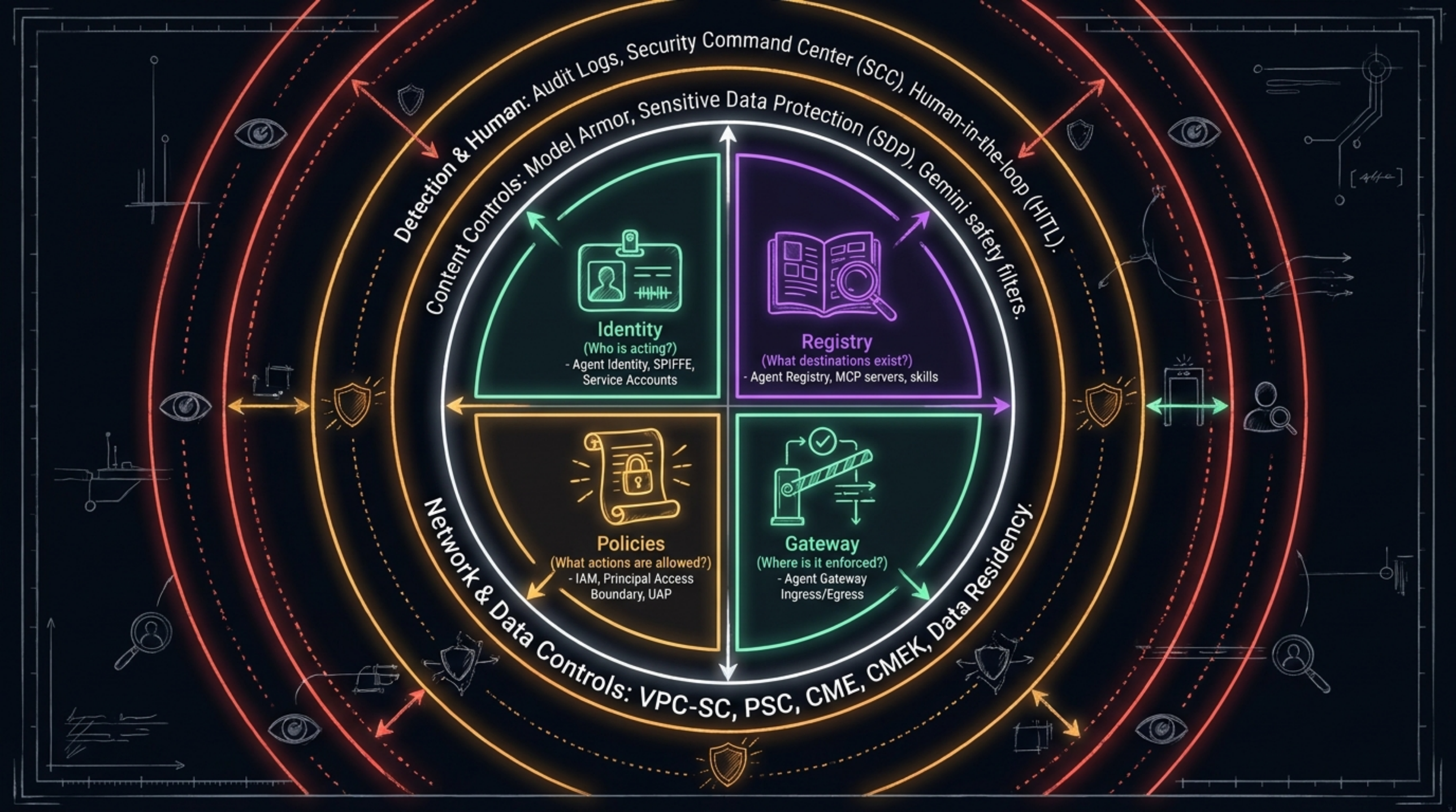


Securing Enterprise AI Agents on Google Cloud

A Master Architectural Blueprint for Identity, Governance, and Defense-in-Depth.

Target: Gemini Enterprise Agent Platform.

Focus: Agent Identity, Auth Manager, PAB, Agent Gateway, Model Armor, and VPC-SC.

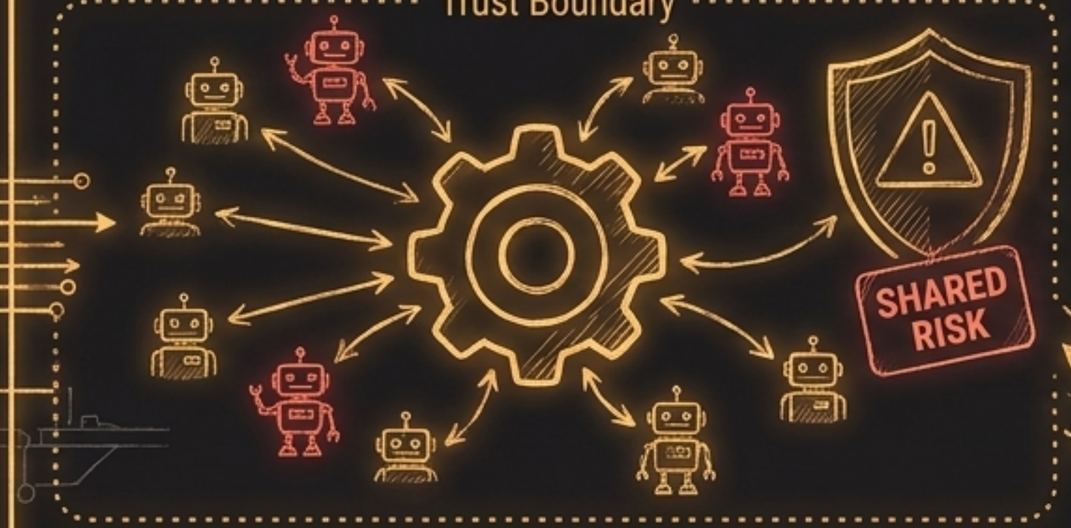

Detection & Human: Audit Logs, Security Command Center (SCC), Human-in-the-loop (HITL).
Content Controls: Model Armor, Sensitive Data Protection (SDP), Gemini safety filters.
Identity
(Who is acting?)
- Agent Identity, SPIFFE, Service Accounts
Registry
(What destinations exist?)
- Agent Registry, MCP servers, skills
Policies
(What actions are allowed?)
- IAM, Principal Access Boundary, UAP
Gateway
(Where is it enforced?)
- Agent Gateway Ingress/Egress
Network & Data Controls: VPC-SC, PSC, CME, CMEK, Data Residency.

Threat Scenario vs. Primary Control

Threat Scenario		Primary Control
Shared service account key leaked.		Agent Identity (per-agent, non-impersonable).
Agent must act on user's Jira data with consent.		3-Legged OAuth (via Auth Manager).
Agent reaches unapproved org projects despite narrow intent.		Principal Access Boundary (PAB).
Agent calls unapproved external host.		Agent Gateway + Agent Registry (Default-deny UAP).
Jailbreak / indirect prompt injection.		Model Armor (Input templates / Semantic Governance).
Data exfiltration from Google APIs to outside project.		VPC Service Controls (VPC-SC).

Identity Evolution: Legacy vs. Modern Agent Identity

Service Agents (Legacy/Default)



- Service Agents and overloaded resource.
  roles/aiplatform.reasoningEngineServiceAgent (Shared)
- Shared by every agent in the project.
- Fails the principle of least privilege.
- Can be impersonated. Key-management risks.

Agent Identity (GA since April 2026)



- Per-agent, strongly attested SPIFFE-based principal.
- Tied specifically to the agent's lifecycle.
- Cannot be impersonated. No long-lived keys.
- Tokens bound to X.509 cert (24h, auto-rotated).
- Context-Aware Access enforces mTLS + DPoP by default.

principal://agents.global.org-ORG_ID.../reasoningEngines/AGENT_ID



AI Agent

Path 1: User-Delegated (3-Legged OAuth)

- **Agent acts as:** The end user (with consent).
- **Mechanism:** Auth manager coordinates 3LO.
- **Use case:** Per-user data in SaaS (Jira, GitHub, Salesforce).
- **Audit Trail:** Logs record both the agent and the user.

Path 2: Agent's Own Authority (2-Legged OAuth / APIs)

- **Agent acts as**: Itself.
- **Mechanism**: Agent Identity to GCP APIs; OIDC / 2LO / API keys to external services.
- **Use case**: Shared back-end services, batch jobs, reference data.

Anti-Pattern: Using one 2LO token for every user's request creates a confused deputy vulnerability.

Auth Manager & Token Lifecycle (Centralized Credential Vault)

3-Legged OAuth

Who consents: End user (revocable).

Stored: User's OAuth tokens.

Refresh: Handled by Auth Manager.

2-Legged OAuth

Who consents: None (agent authorized).

Stored: OAuth client ID/secret.

Refresh: Handled by Auth Manager.

API Key

Who consents: None.

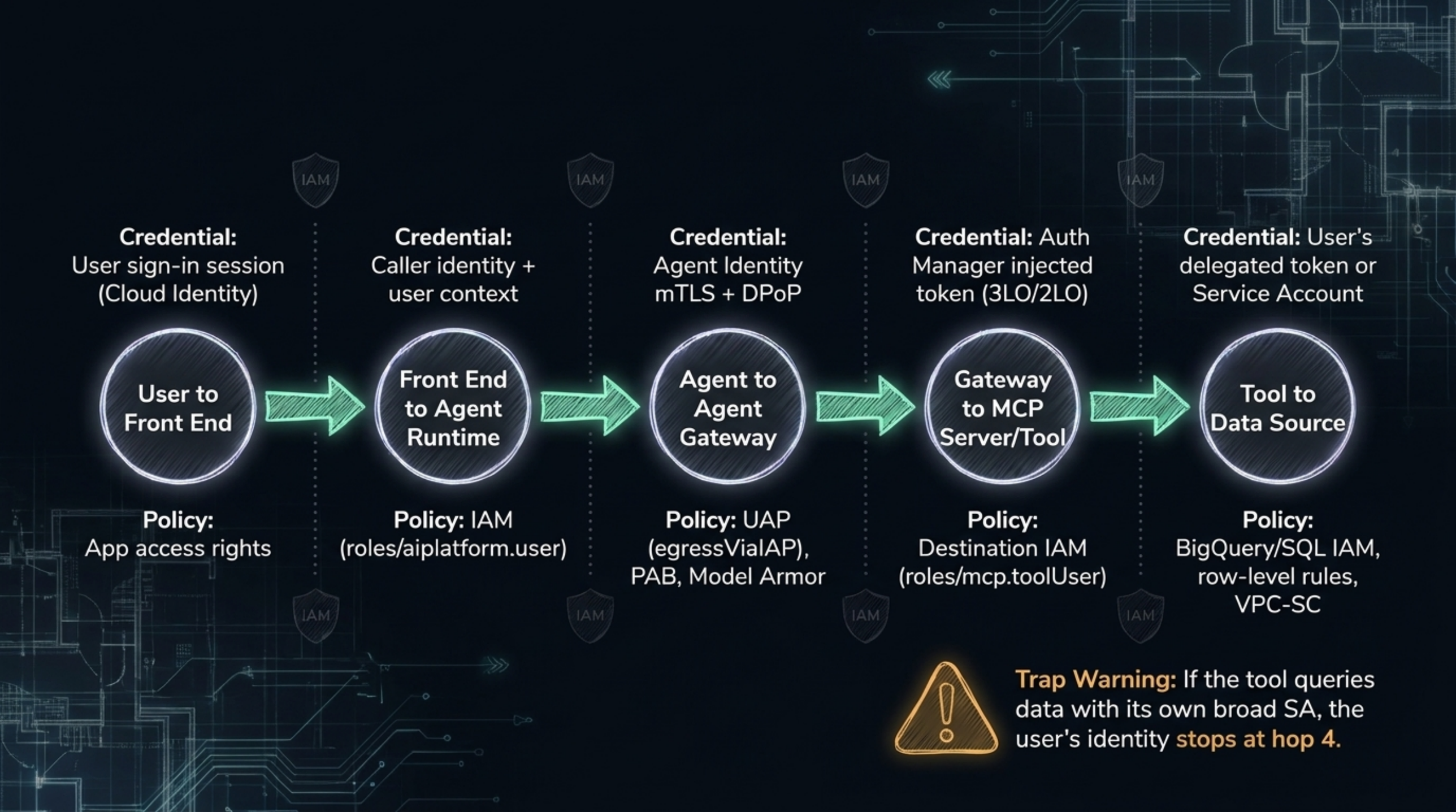
Stored: Static Key.

Refresh: Manual rotation required.

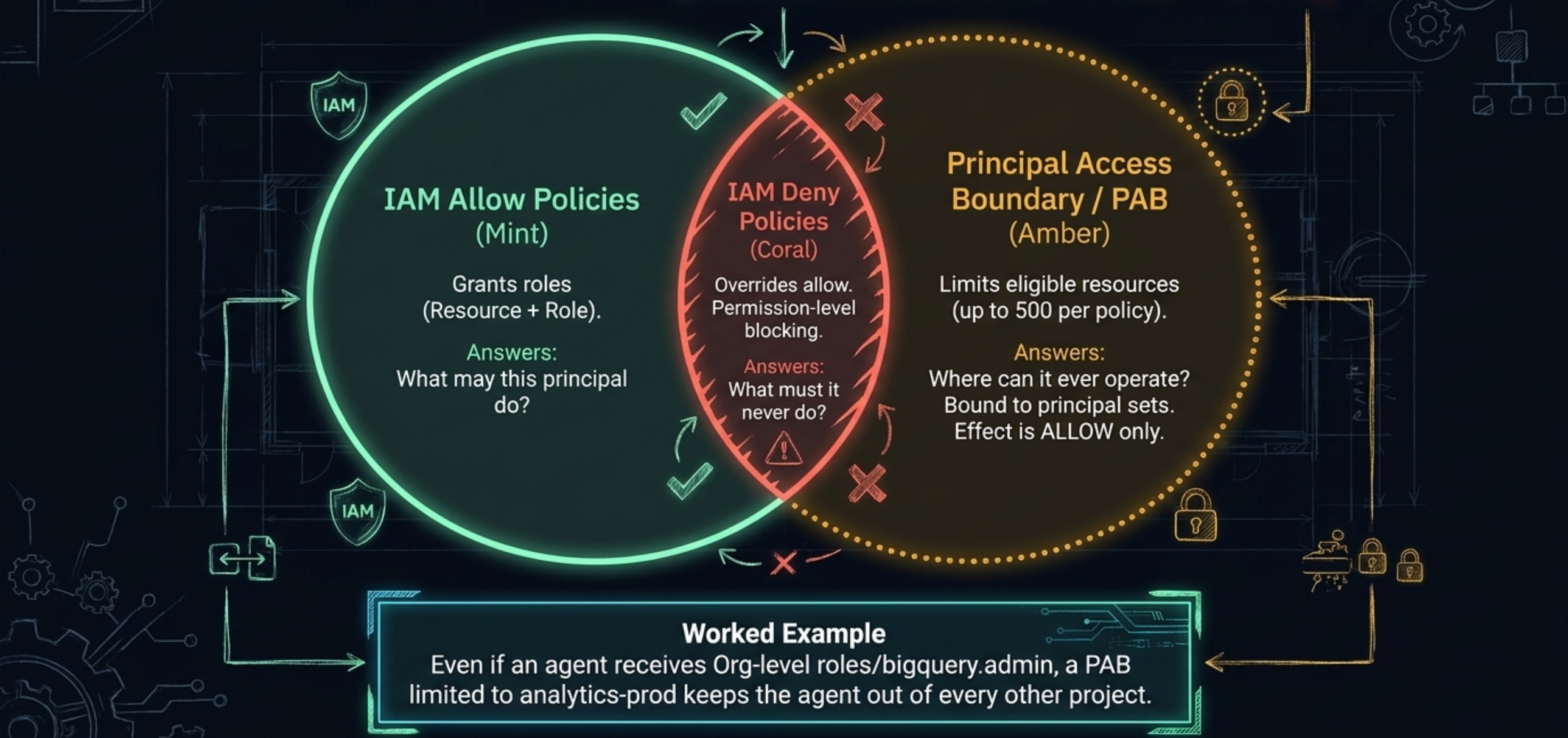
Token Exchange (RFC 8693)

Swaps external OIDC/SAML/K8s credentials for Google credentials via Workload Identity Federation (cross-cloud preview).

Key Mechanic: ADK integration injects headers (Authorization, X-Goog-Api-Key) directly into tool calls. Credentials never live in agent code or environment variables.

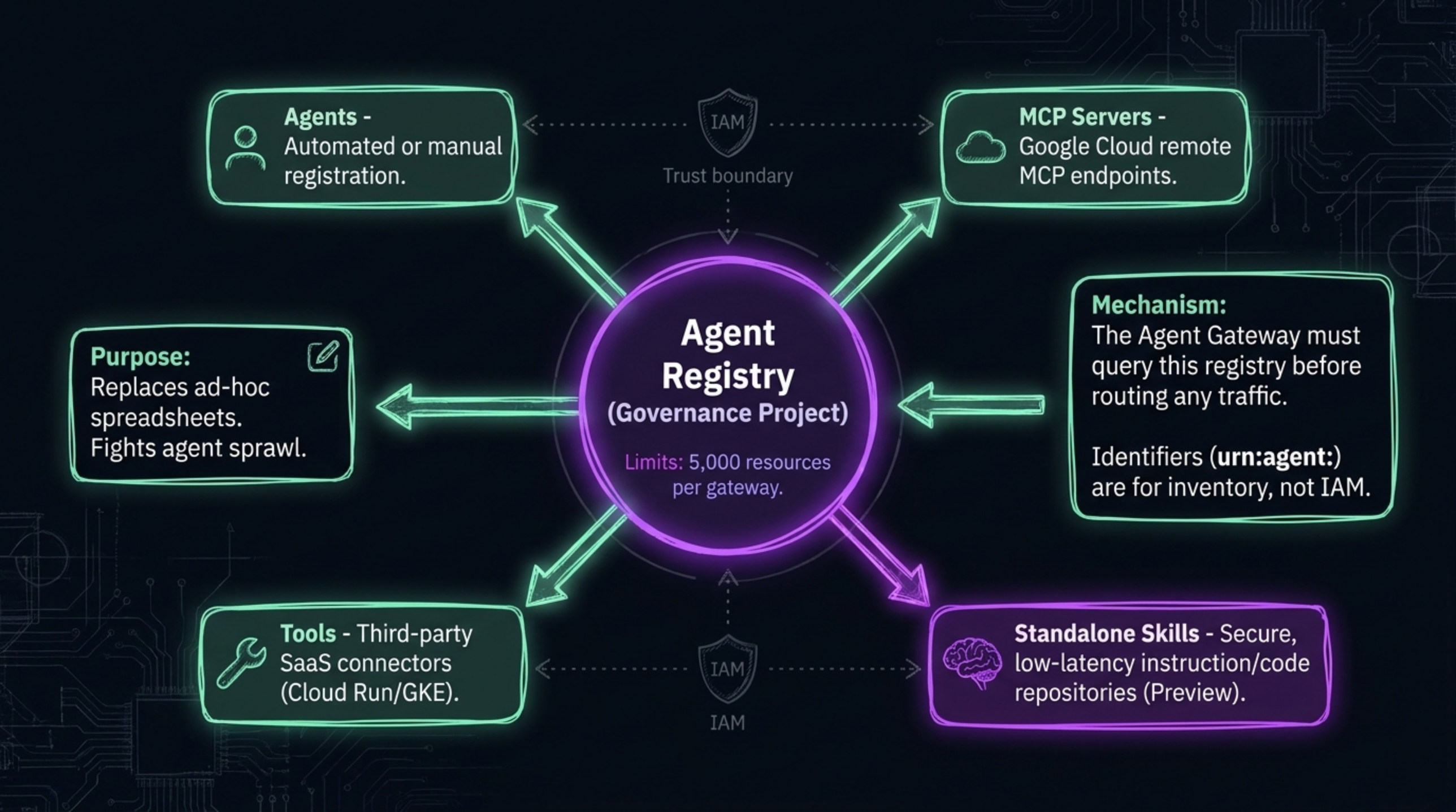

IAM
Credential: User sign-in session (Cloud Identity)
User to Front End
Policy: App access rights
Credential: Caller identity + user context
Front End to Agent Runtime
Policy: IAM (roles/aiplatform.user)
Credential: Agent Identity mTLS + DPoP
Agent to Agent Gateway
Policy: UAP (egressViaIAP), PAB, Model Armor
Credential: Auth Manager injected token (3LO/2LO)
Gateway to MCP Server/Tool
Policy: Destination IAM (roles/mcp.toolUser)
Credential: User's delegated token or Service Account
Tool to Data Source
Policy: BigQuery/SQL IAM, row-level rules, VPC-SC
Trap Warning: If the tool queries data with its own broad SA, the user's identity stops at hop 4.

Effective Access = (Allow intersect PAB) - Deny

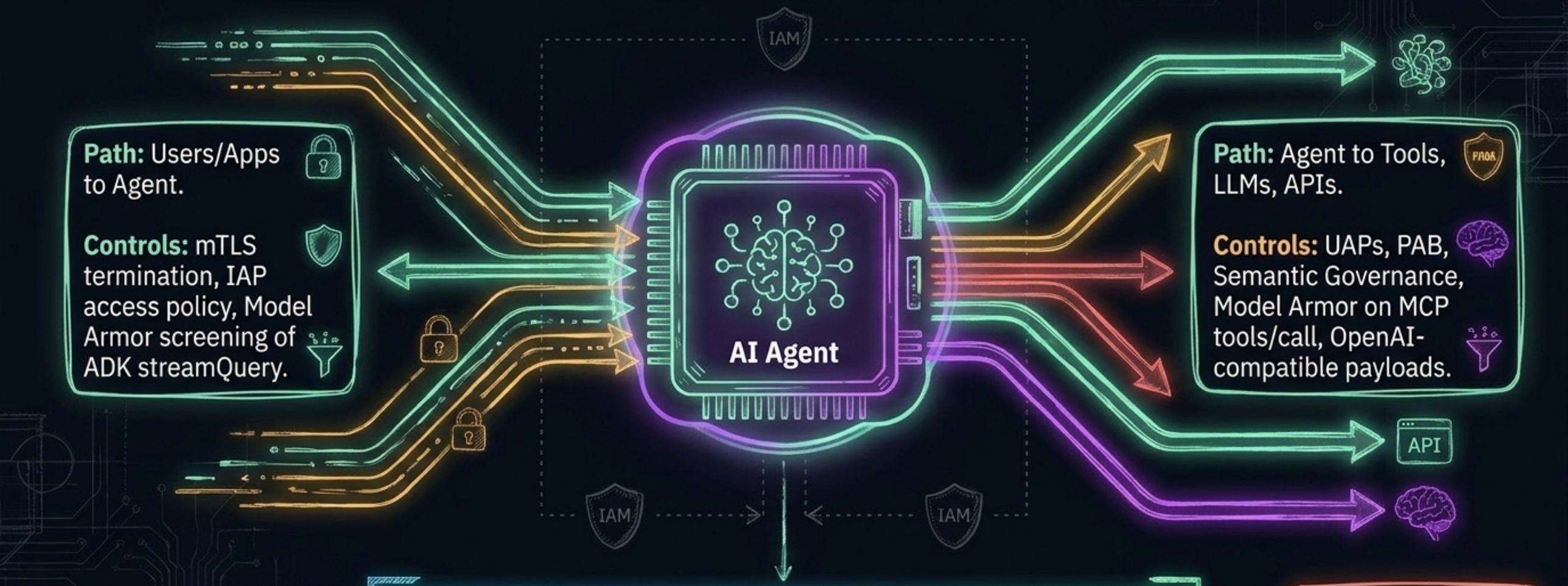


Worked Example

Even if an agent receives Org-level roles/bigquery.admin, a PAB limited to analytics-prod keeps the agent out of every other project.


Agents - Automated or manual registration.
IAM
Trust boundary
MCP Servers - Google Cloud remote MCP endpoints.
Purpose: Replaces ad-hoc spreadsheets. Fights agent sprawl.
Agent Registry (Governance Project)
Limits: 5,000 resources per gateway.
Mechanism: The Agent Gateway must query this registry before routing any traffic.
Identifiers (urn:agent:) are for inventory, not IAM.
Tools - Third-party SaaS connectors (Cloud Run/GKE).
IAM
IAM
Standalone Skills - Secure, low-latency instruction/code repositories (Preview).

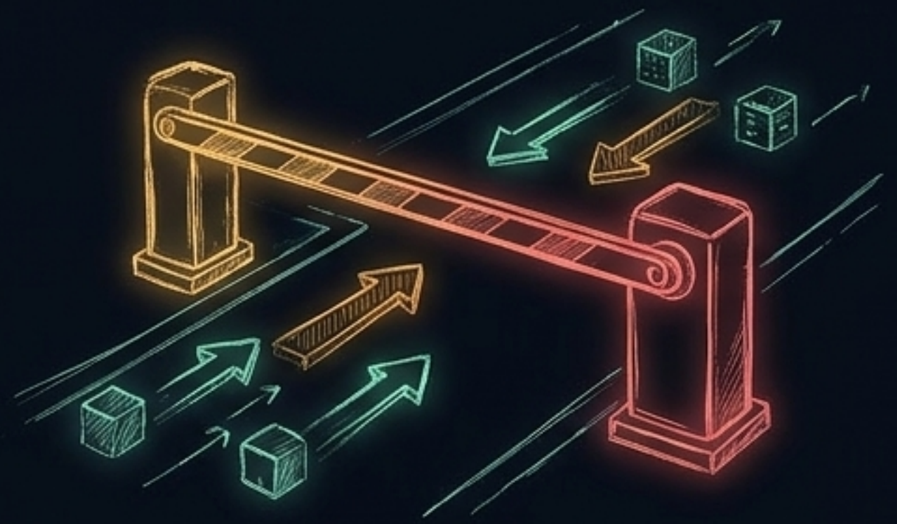
AI Agent Architecture: Cross-Section & Flow Control



Gateway Mechanics: Mediates MCP, A2A, REST, gRPC.
Extracts JSON-RPC tool names for policy.
Emits OpenTelemetry traces (Otel) to Cloud Logging.

Deployment Note:
Gemini Enterprise currently supports Egress only.

UAP Mechanics (IAM v3)



- Enforced by IAP (`iap.googleapis.com` (`iap.googleapis.cm/resources.egressViaIAP`)
- Operates on a strict **Default-Deny** basis.
- **Deny** overrides **allow**.

Common Expression Language (CEL) Conditions

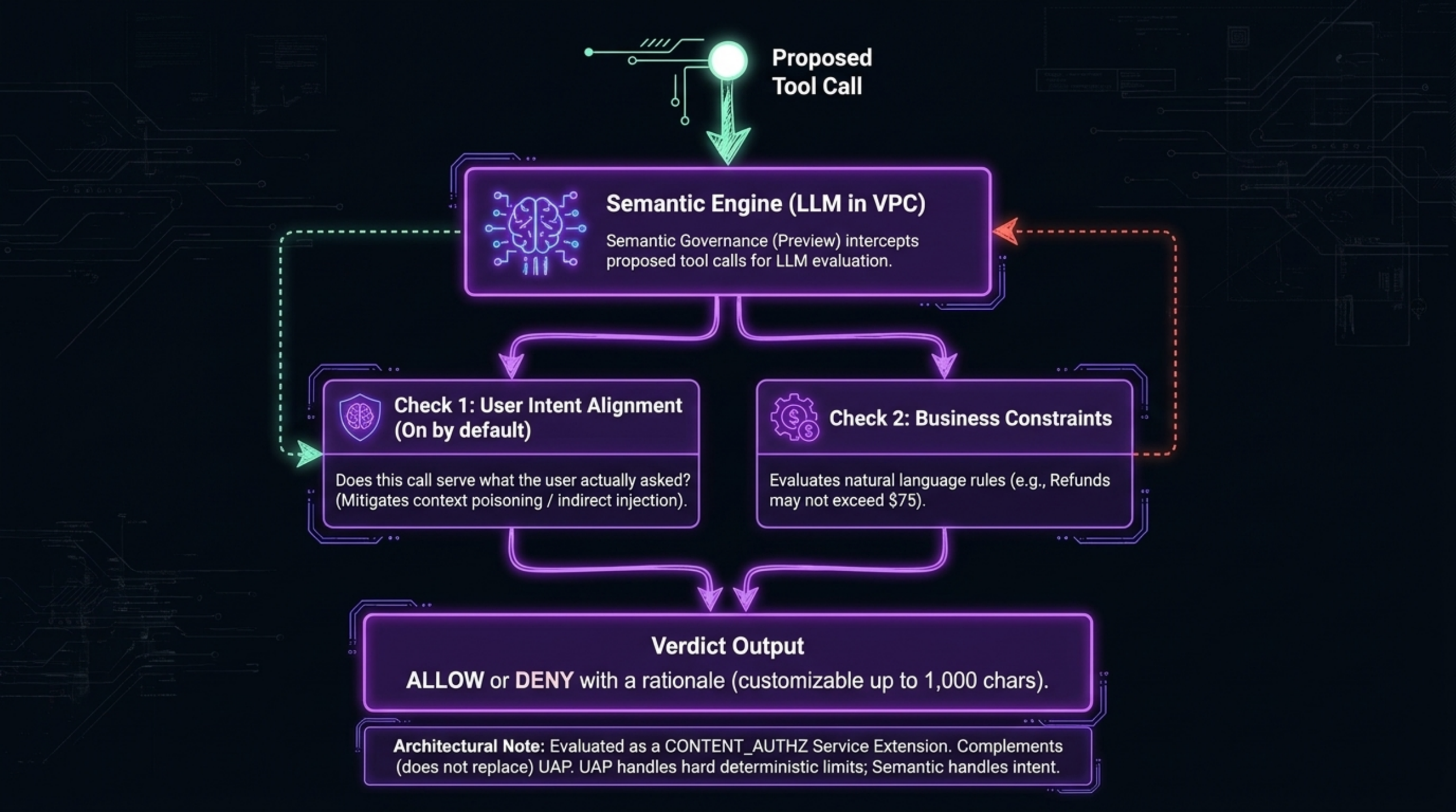
```
// Allow specific tools
destination.agent_registry.mcp_server.tool.name

// Guardrail flags
destination.agent_registry.mcp_server.tool.
    .annotations.destructive_hint == true

// Block unregistered
destination.unregistered.host
```

Rollout Strategy Box

Set iamEnforcementMode: DRY_RUN in IAP metadata. Logs would-be denials to **Audit Logs** without breaking the runtime.


Proposed
Tool Call
Semantic Engine (LLM in VPC)
Semantic Governance (Preview) intercepts proposed tool calls for LLM evaluation.
Check 1: User Intent Alignment (On by default)
Does this call serve what the user actually asked? (Mitigates context poisoning / indirect injection).
Check 2: Business Constraints
Evaluates natural language rules (e.g., Refunds may not exceed $75).
Verdict Output
ALLOW or DENY with a rationale (customizable up to 1,000 chars).
Architectural Note: Evaluated as a CONTENT_AUTHZ Service Extension. Complements (does not replace) UAP. UAP handles hard deterministic limits; Semantic handles intent.

Model Armor: 6 Detection Engines

1. Responsible AI (RAI)

Hate, harassment, explicit, dangerous. Default threshold is Medium and above.

2. Prompt Injection & Jailbreak

Attempts to override instructions. (Inputs <3 words return NO_MATCH_FOUND).

NO_MATCH_FOUND

3. Sensitive Data Protection (SDP)

Basic (6 fixed infoTypes) vs. Advanced (custom templates + de-identification).

4. Malicious URL

Phishing/malware link detection (first 256 URLs).

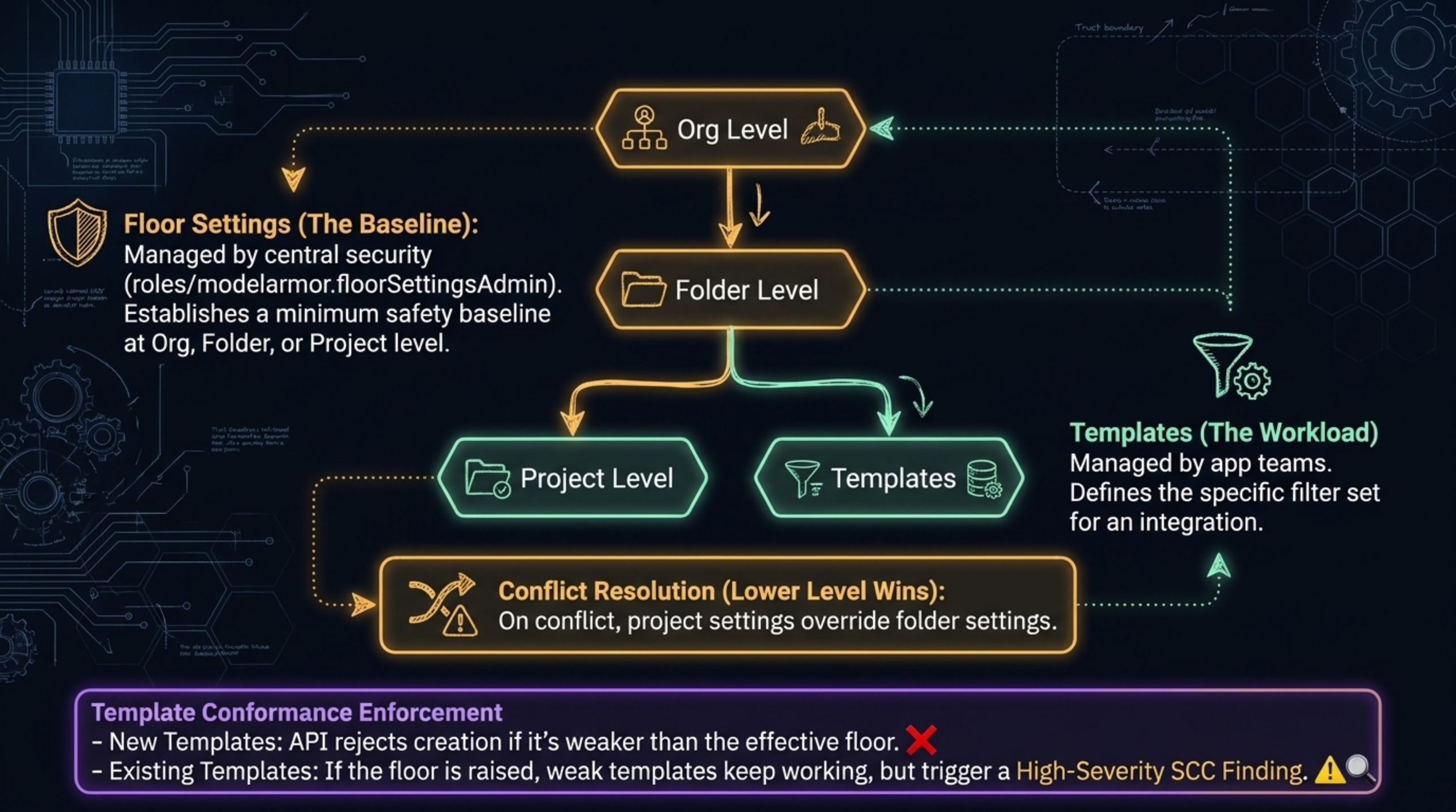
5. Document Screening

Text in PDF, CSV, Word, Excel. (Max 4MB).

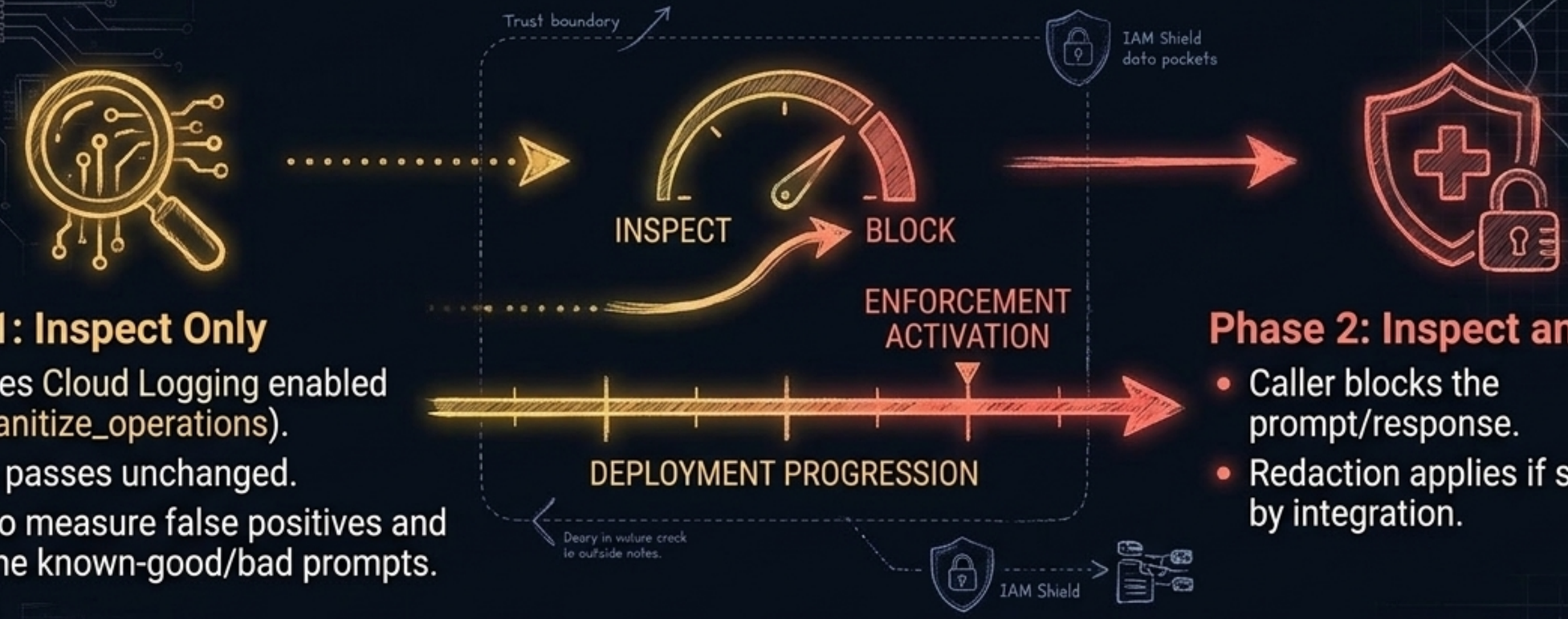
6. Image Screening (Preview)

OCR and visual content via Advanced SDP. (Max 4MB).

⚠ Limits Note: Screens single turns (no memory). Encrypted/Base64 text is not decoded.


Org Level
Folder Level
Project Level
Templates
Floor Settings (The Baseline):
Managed by central security (roles/modelarmor.floorSettingsAdmin). Establishes a minimum safety baseline at Org, Folder, or Project level.
Templates (The Workload)
Managed by app teams. Defines the specific filter set for an integration.
Conflict Resolution (Lower Level Wins):
On conflict, project settings override folder settings.
Template Conformance Enforcement
- New Templates: API rejects creation if it's weaker than the effective floor. ❌
- Existing Templates: If the floor is raised, weak templates keep working, but trigger a High-Severity SCC Finding. ⚠️🔍

Model Armor: Deployment Phases & Enforcement



Phase 1: Inspect Only

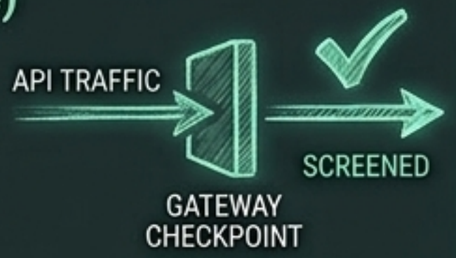
- Requires Cloud Logging enabled (log_sanitize_operations).
- Traffic passes unchanged.
- Used to measure false positives and baseline known-good/bad prompts.

Phase 2: Inspect and Block

- Caller blocks the prompt/response.
- Redaction applies if supported by integration.

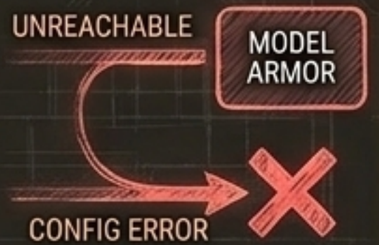
Inline Enforcement (Zero-Code)

Project floor settings can screen Gemini API and Google Cloud MCP traffic directly, without application code changes.



Fail-Open Design

If Model Armor is unreachable in a region, Gemini processes the request unscreened to maintain availability. Configuration errors under block mode, however, will block.



SDP Method Diagnostic Table: Properties & Capabilities

Trust boundary

Method	Reversible?	Format kept?
Replace with InfoType (e.g., [EMAIL_ADDRESS])	No	No
Redaction (Drops value entirely)	No	No
Deterministic Tokenization (CryptoDeterministicConfig): Maintains referential integrity (Customer A stays Customer A across multiple turns). Good for multi-to…	Yes (via KMS key)	No
Format-Preserving Encryption (FPE): Keeps exact length and alphabet (e.g., 16-digit card stays 16 digits). Slower, use only when schema validation requires it.	Yes	Yes
Date Shifting: Maintains intervals for clinical/financial timelines without exposing exact dates.	No	Yes (Still a valid date)

The Anti-Pattern: Prompt-Only Privacy (telling the system prompt never repeat card numbers) is NOT a control. Raw data still gets indexed and logged.



Boundary 1: Before Indexing (RAG)

Land files in a quarantine bucket -> Run **SDP Inspect** with De-identify action -> Write to Clean Bucket -> Index clean data.

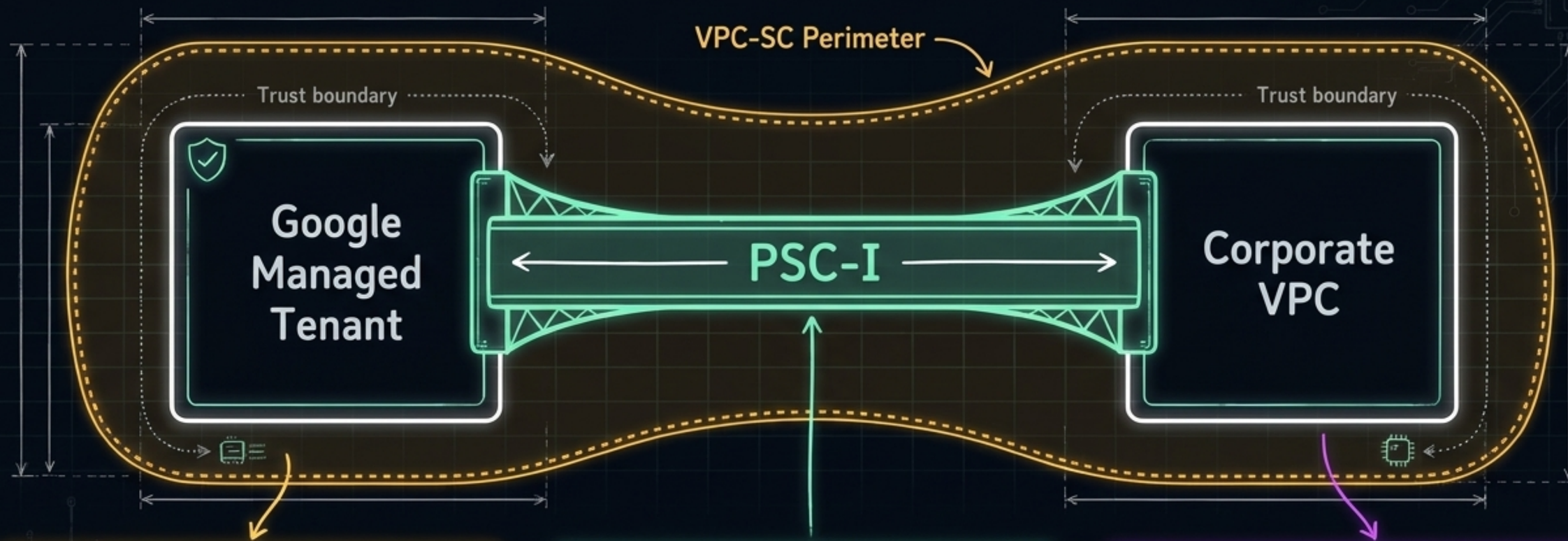
Boundary 2: Before Memory Bank Writes

Use **Advanced SDP** on the gateway or ADK callbacks to scrub SSNs before they become long-lived session memories.

Boundary 3: Log & Trace Export

Set ADK_CAPTURE_MESSAGE_CONTENT_IN_SPANS=false to keep PII out of OTEL span attributes. Restrict raw sanitize logs.

Physical Network Topology & Security Controls: VPC-SC, PSC, Encryption & Residency



VPC Service Controls (VPC-SC)

- Blocks public internet access. ✓
- ✓ Protects Google APIs. Prevents data exfiltration. ✓
- ✗ Violations appear as Policy Denied audit logs.

Private Service Connect (PSC)

- PSC Interface (PSC-I): Network attachment for private egress from the Agent to on-prem/VPC.
- ⚠ (Requires subnet /28). Cannot be changed without recreating the agent.

Encryption & Residency

- CMEK: Customer-managed encryption keys.
- ⚠ (Note: Preview models lack CMEK/VPC-SC support).
- Data Residency: Global endpoint has no residency guarantee. Use Regional endpoints to enforce ML processing jurisdiction.

Cloud Audit Logs

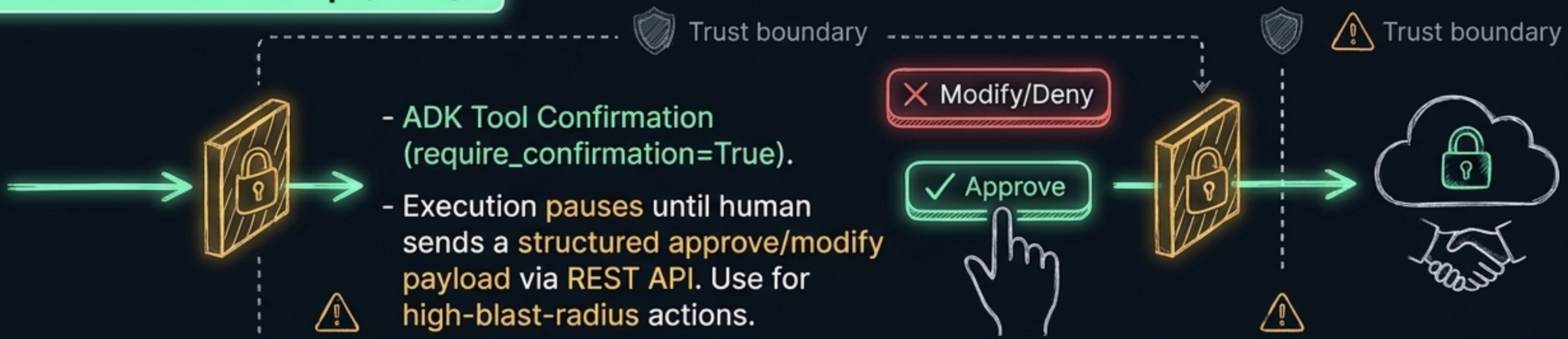
- **Admin Activity**: Always on (IAM, Metadata).
- **Data Access**: Must be manually enabled to see data reads/writes (model usage).



Security Command Center (AI Protection - Premium)

- Detects anomalies audit logs only record.
- Alerts on:
  - IAM Anomalous Grant to Agentic Identity
  - Agentic Identity Credential Used Outside GCP
  - Model Armor Floor Setting Violations.

Human-in-the-Loop (HITL)



- ADK Tool Confirmation (require_confirmation=True).
- Execution pauses until human sends a structured approve/modify payload via REST API. Use for high-blast-radius actions.

Identity & Credentials - Agent Identity (SPIFFE) + Auth Manager (3LO/2LO)

Boundaries - IAM Allow intersect PAB minus Deny

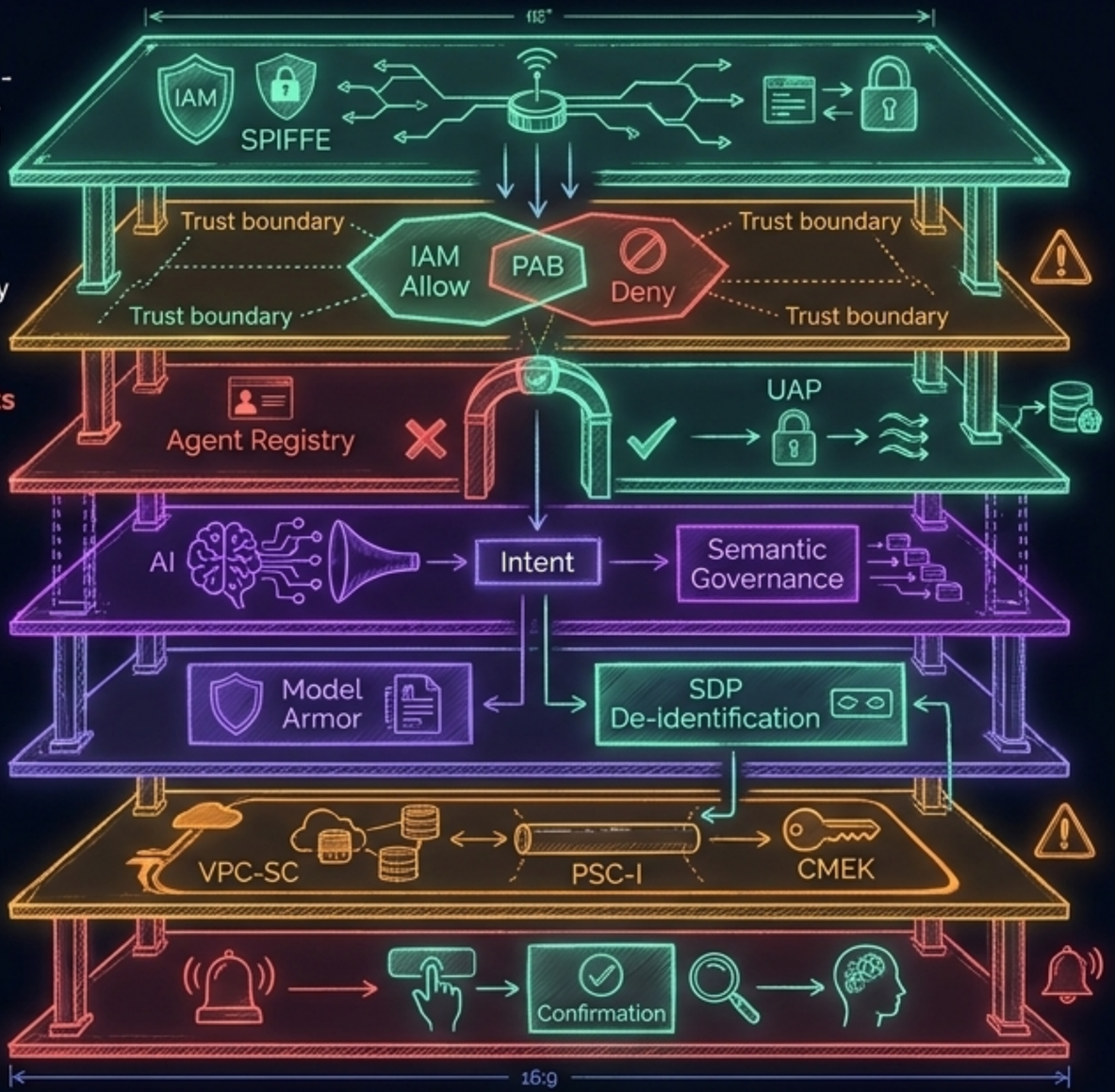
Gateway Enforcements - Agent Registry + Default-Deny UAPs

Intent - Semantic Governance intercepts

Content Safety - Model Armor Templates + SDP De-identification

Network & Data - VPC-SC + PSC-I + CMEK

Audit & Humans - SCC AI Protection + HITL Tool Confirmation



"Defense in Depth is not a checklist; it is an integrated, hop-by-hop topology. Secure the identity, limit the boundary, govern the intent, and inspect the content."