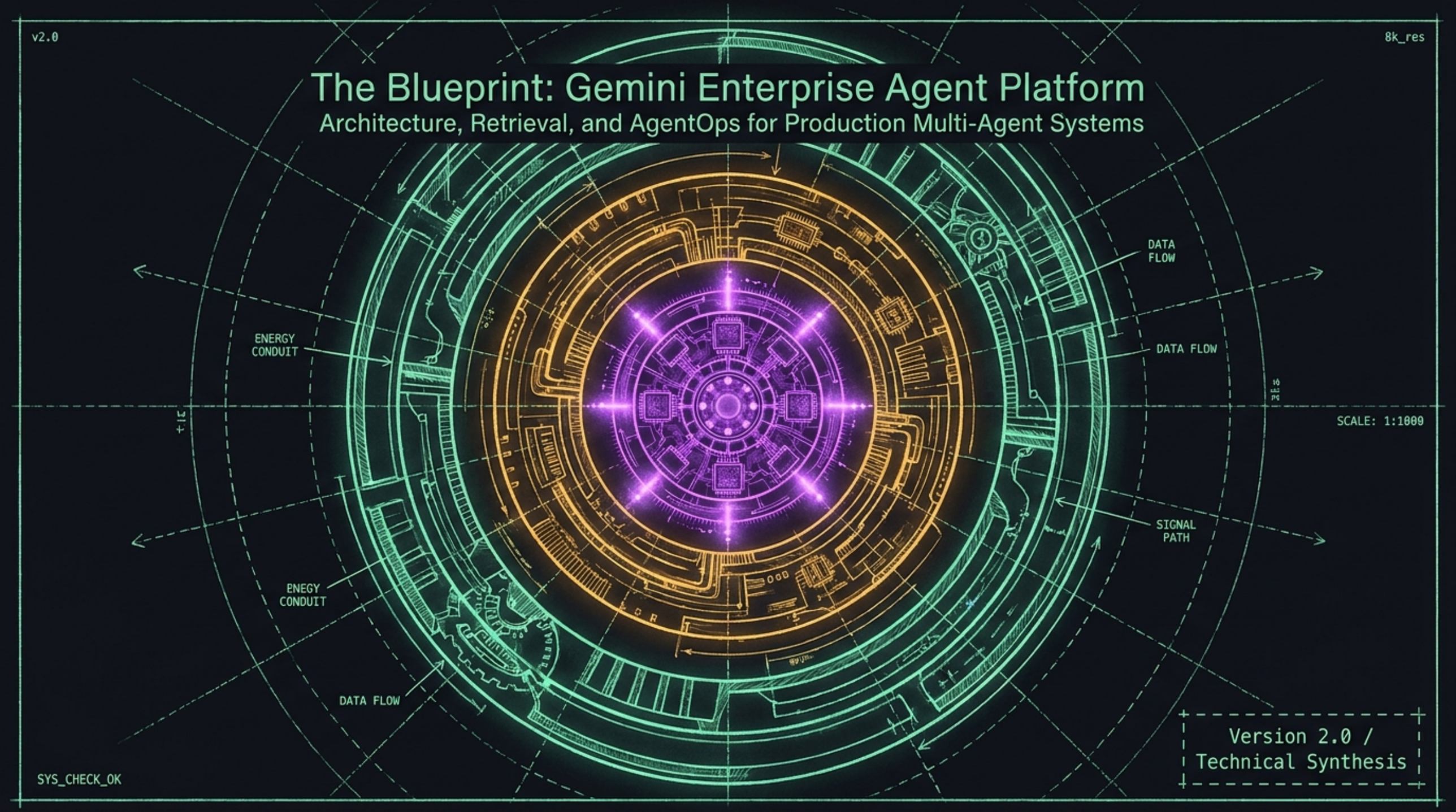
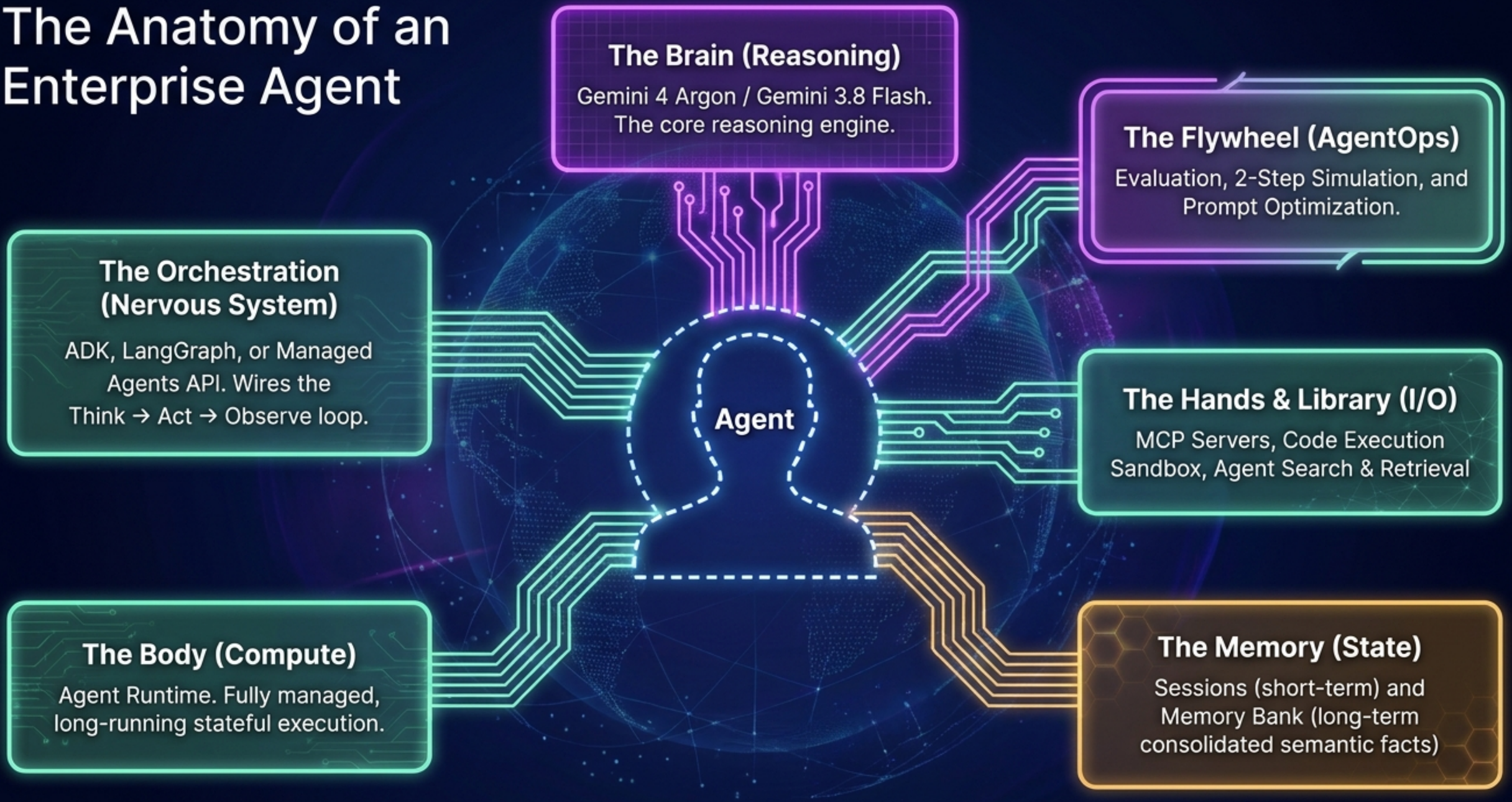
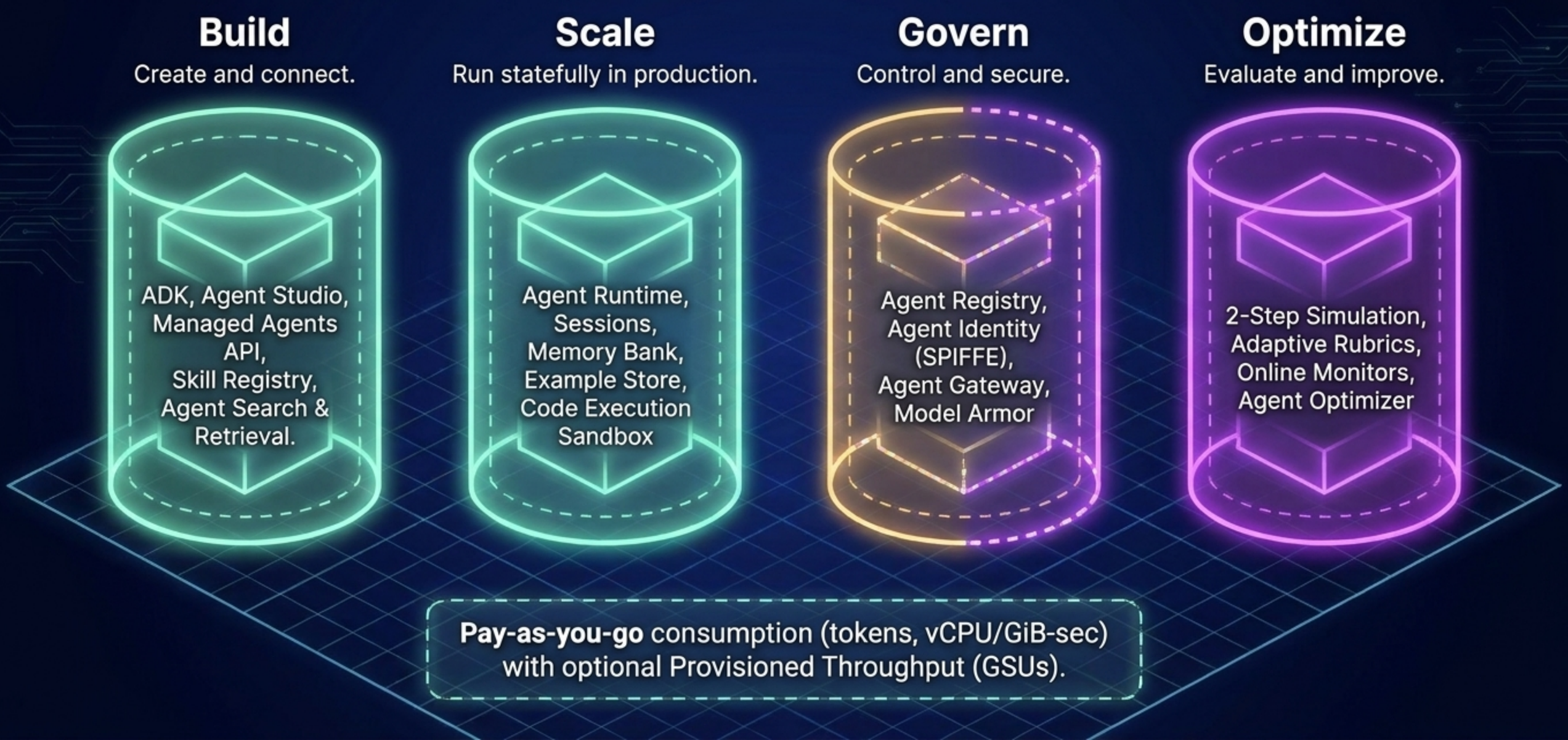

v2.0
8k_res
The Blueprint: Gemini Enterprise Agent Platform
Architecture, Retrieval, and AgentOps for Production Multi-Agent Systems
DATA FLOW
ENERGY CONDUIT
DATA FLOW
SCALE: 1:1000
SIGNAL PATH
ENEGY CONDUIT
DATA FLOW
SYS_CHECK_OK
Version 2.0 / Technical Synthesis

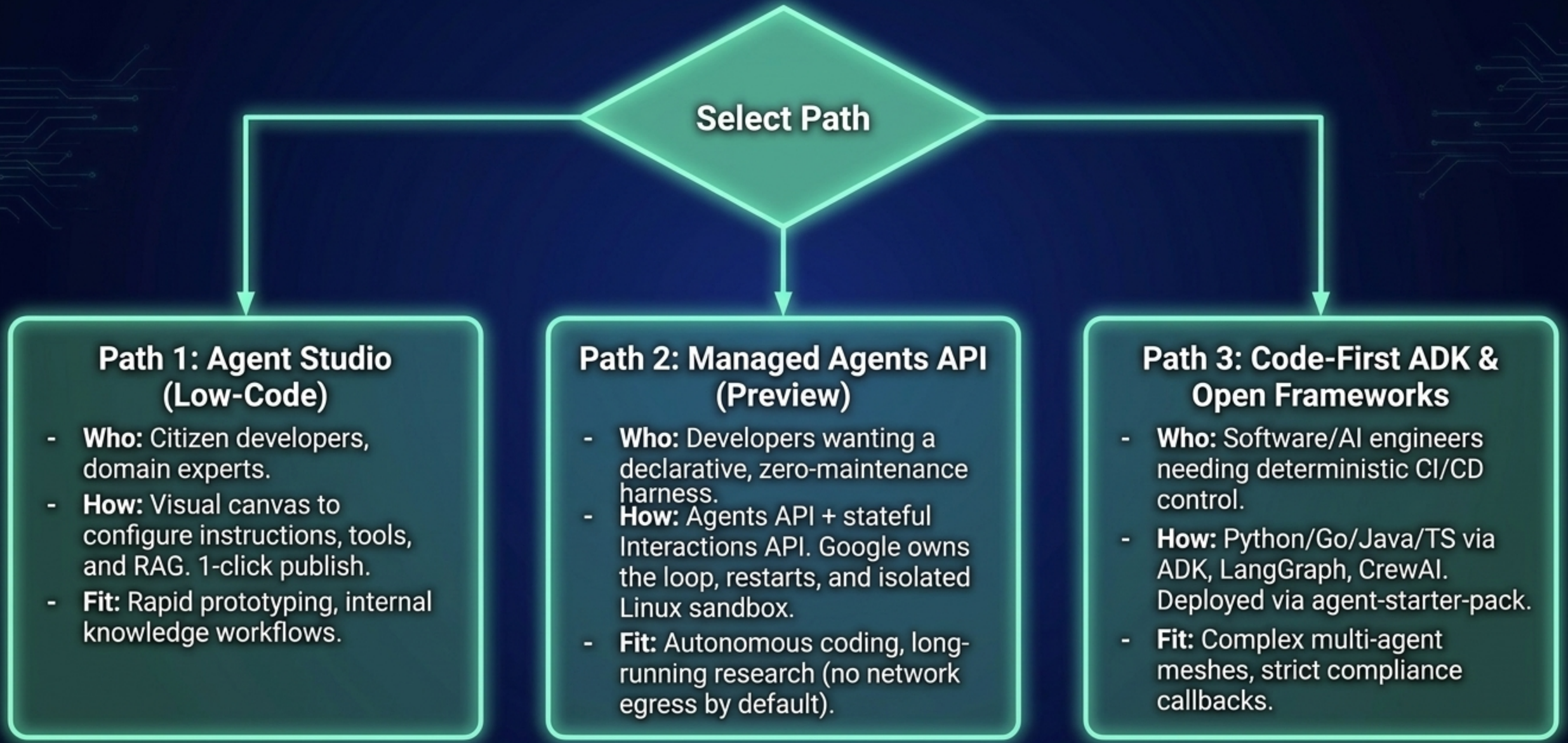
The Anatomy of an Enterprise Agent



The Four Pillars of Agent Platform



Choosing Your Orchestration Path



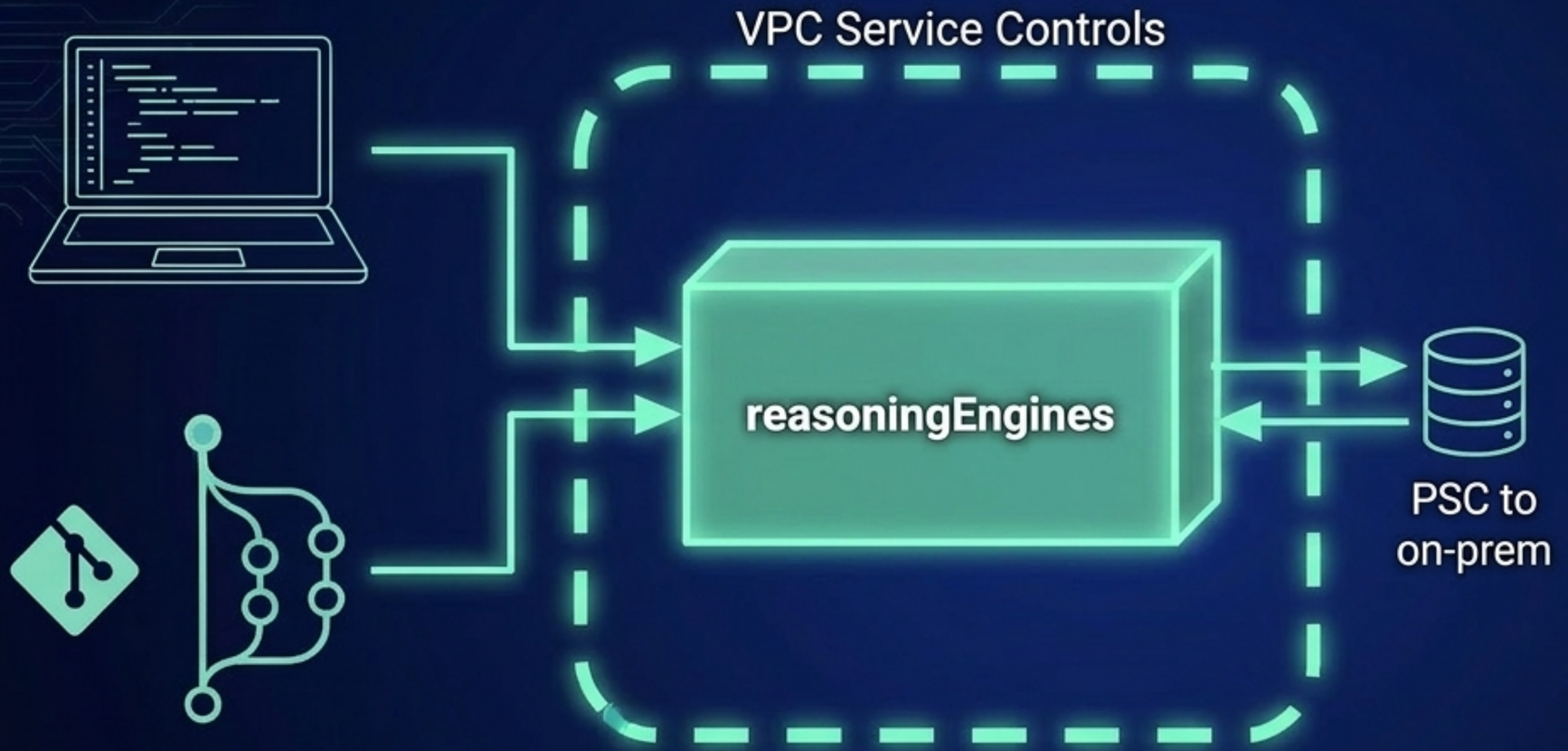
The Compute Matrix: Hosting the Stateful Loop

Agent Runtime	Cloud Run	GKE (Standard / Autopilot)
Best for: Python/ADK agents + minimal ops.	**Best for:** Polyglot webhooks, Streamable HTTP MCP servers.	**Best for:** Self-hosting Gemma 4 on GPUs, gVisor strict sandboxing.
Scale: Up to 1,000 instances (100 in VPC-SC). Multi-day async.	**Scale:** Scale-to-zero up to 1,000 instances. 60min HTTP / 24h Jobs.	**Scale:** Full HPA/VPA control.
State: Native Sessions, Memory Bank, Code Execution.	**State:** Ephemeral RAM/disk. Must externalize.	**State:** StatefulSets, persistent DBs.
Trade-off: No custom GPUs inside container.	**Trade-off:** In-process memory vanishes on scale-in.	**Trade-off:** Highest operational/K8s cost.



Trap Note: Cloud Run RAM Trap: In-process dicts vanish unpredictably. Always externalize to Sessions or Redis!

Agent Runtime: The Dedicated Stateful Harness



Core Capabilities: Fully managed, sub-second cold starts, cryptographic Agent Identity (spiffe://), native ADK integration.

Five Deployment Methods:

1. Developer Connect (GitOps)
2. Source files (Inline local dir)
3. Dockerfile (Polyglot/system packages)
4. Container Image (Artifact Registry - lowest latency)
5. SDK Object (In-memory pickle for notebooks)

Security & Networking: Enforces IAM and VPC Service Controls. Private Service Connect (PSC) network attachments for on-prem egress.

Trap Note: Reserved Env Vars: Never override GOOGLE_CLOUD_AGENT_ENGINE* or PORT. Never attach custom GPUs directly here (use PSC to GKE instead).

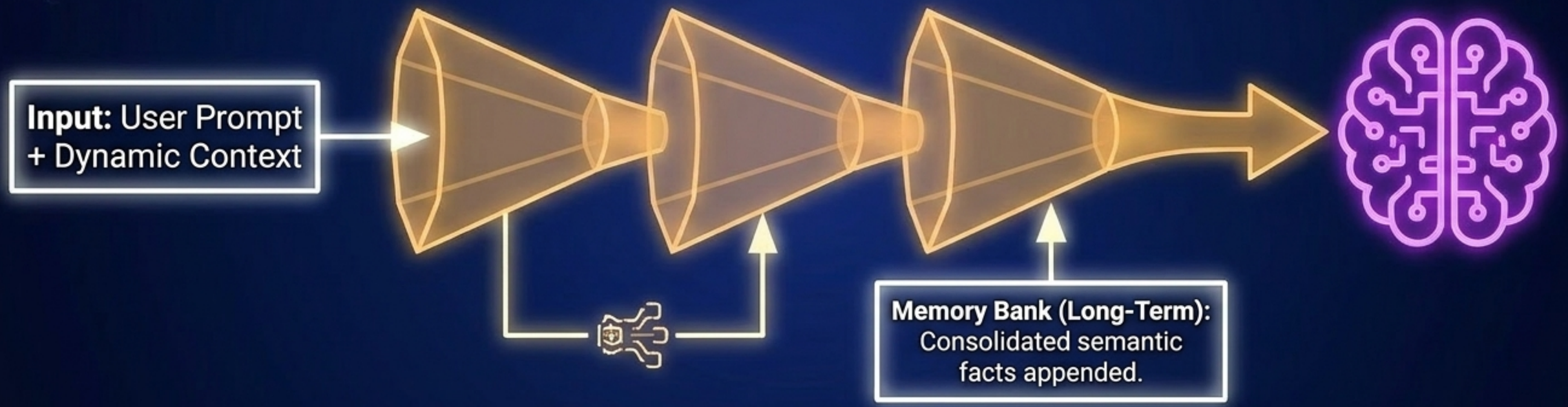
Context Engineering: Shaping the Context Window

The Goal: Prevent context rot and latency spikes by avoiding raw log stuffing.

Prefix Caching (Tier 1): Static system instructions and tool schemas.

Sessions (Short-Term): Append-only log of Events + mutable State. Compacted via EventsCompactionConfig.

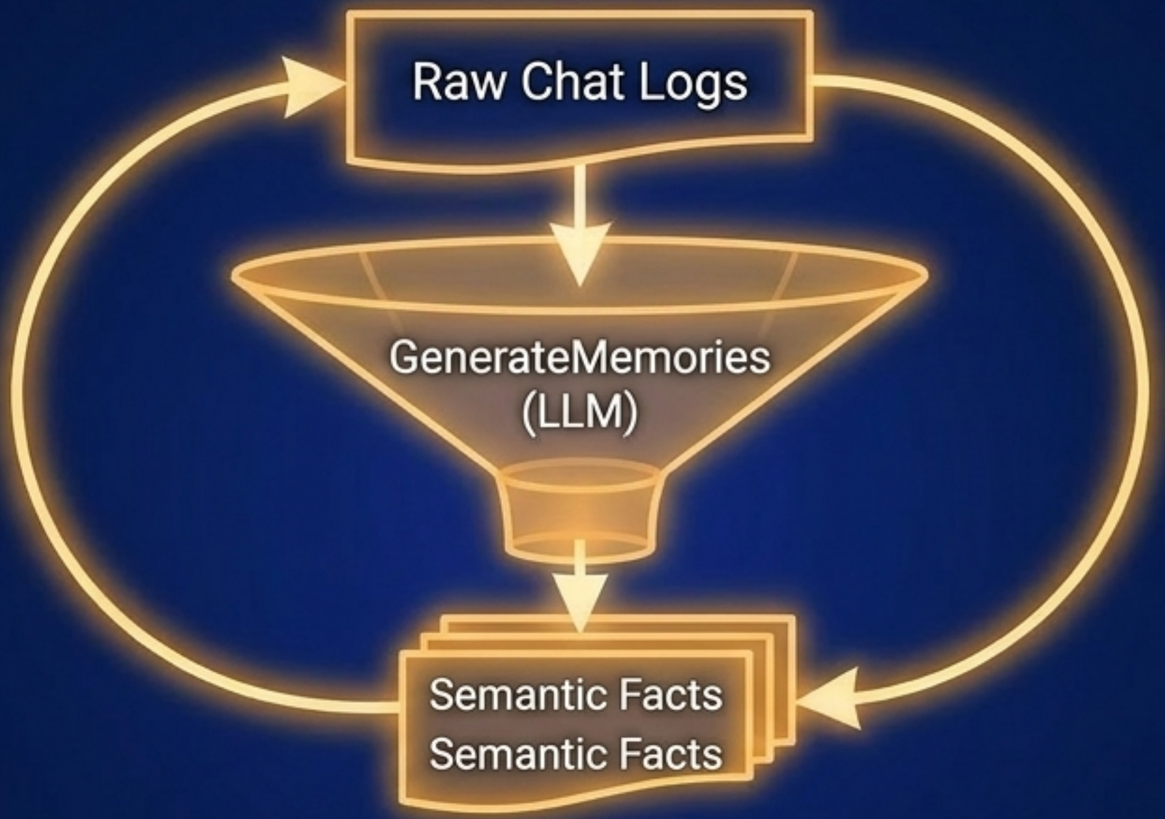
Example Store (Dynamic): Top-k few-shot trajectories injected at runtime.



Trap Note: Prefix-Cache Breaker: Placing dynamic timestamps or user IDs at the top of the system prompt invalidates the implicit cache every turn! Keep static schemas first.

Memory Bank: Consolidation & Episodic Memory

Asynchronous Extraction: GenerateMemories runs in the background (zero hot-path latency). Merges duplicates, updates stale preferences, resolves contradictions.



Explicit Writes: CreateMemory tool. Triggered when user explicitly says “Always remember X.”

Tenant Isolation: Every memory keyed by strict identity scope {user_id, app_name}.

Enterprise Governance: Enforced via Memory Topics (defining what to ignore), TTL expirations for GDPR, and CMEK.

Trap Note: Memory Poisoning: Untrusted external tool content can trick the extraction LLM into storing a malicious instruction. Block with Model Armor and strict Memory Topics.

The External State Matrix

Data Need	Service
Conversation Sessions (Events/State)	**Agent Platform Sessions**. Zero-ops, CMEK, feeds Memory Bank. (**Fallback:** Cloud SQL). 
Semantic Cache & Rate Limits	**Memorystore** (Redis/Valkey). Sub-millisecond HNSW/FLAT, RPM counters, Pub/Sub idempotency. 
Real-time UI Sync & HITL Queues	**Firestore**. Serverless, 99.999% consistency, subcollection turn logs. 
Transactional Tool I/O	**AlloyDB / Cloud SQL**. ACID workflows, pgvector next to live rows.
Blobs (PDFs, Audio, Sandbox I/O)	**Cloud Storage**. GcsArtifactService, Zonal Rapid buckets.



Trap Note: Asynchronous Audit: Memorystore uses async replication. Replica failovers lose in-flight writes. Never use Redis as a system-of-record audit ledger!

The Vector Ecosystem Matrix

Agent Search	RAG Engine	Agent Retrieval	Vector Search 1.0
Turnkey managed RAG. Native SharePoint/Drive connectors, hybrid search, inline citations, document-level ACLs. Max 10 data stores per request.	Code-first modular corpus orchestration. Document AI Layout Parser (20MB/500p limit).	Self-tuning collection store. JSON Data Objects + automatic managed embeddings + keyword RRF.	Billion-scale ScaNN index-as-a-service. BYO vectors. Bills 24/7 node-hours even at 0 QPS.



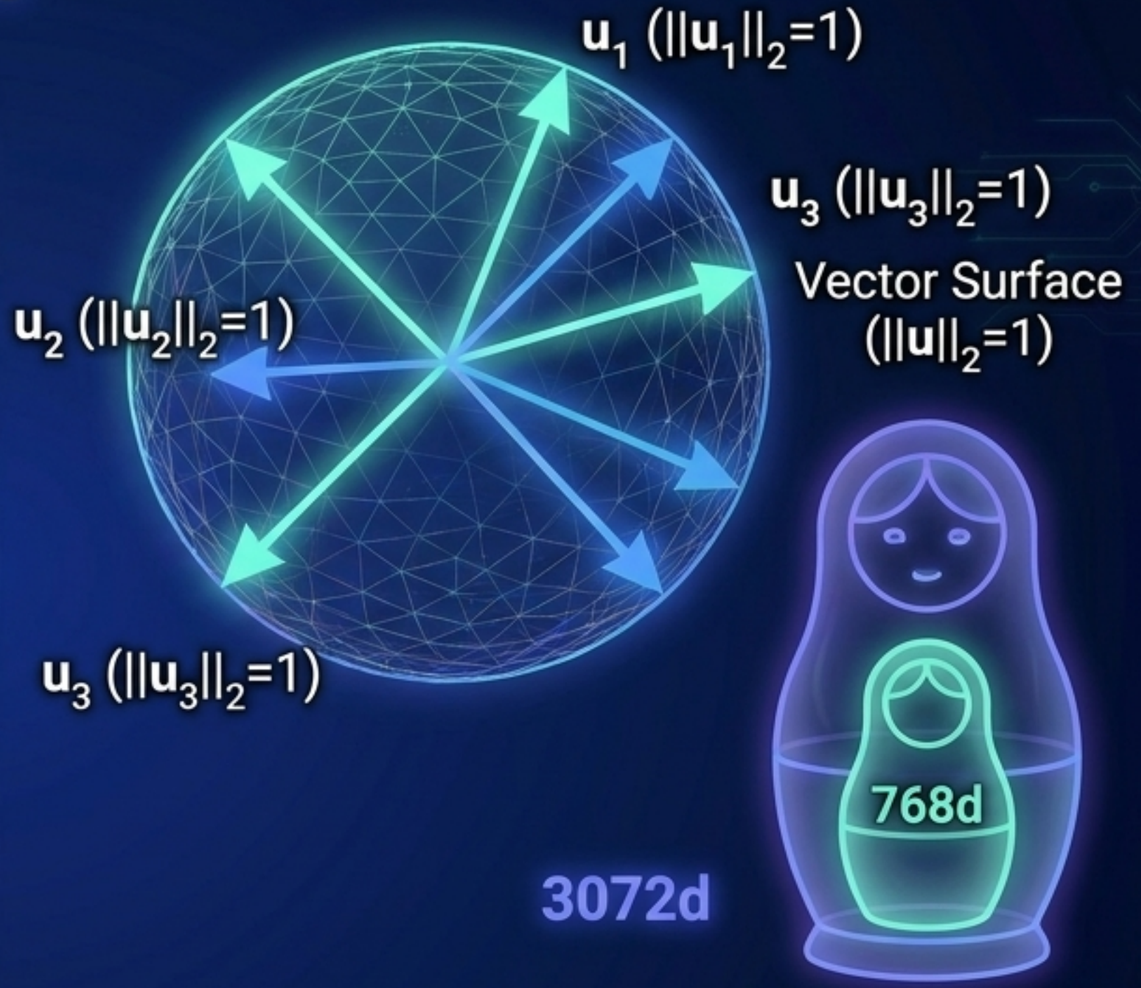
Trap Note: **Spanner vs Serverless Lock-in**: RAG Engine Serverless lacks CMEK. Spanner mode supports CMEK. These modes cannot be migrated in place without re-indexing.

Embedding Math & Metric Traps

The Models:

- text-embedding-005: 768d, 2k tokens. English/Code. 8 task_type values.
- gemini-embedding-001: 3072d, 2k tokens. Multilingual Matryoshka.
- gemini-embedding-2: 3072d, 8k tokens. Omni-modal (text/image/video).

The Normalization Law:
Under unit L2 normalization (||**u**||_2 = 1), cosine similarity exacts inner dot product. DOT_PRODUCT_DISTANCE is the fastest hardware SIMD path.

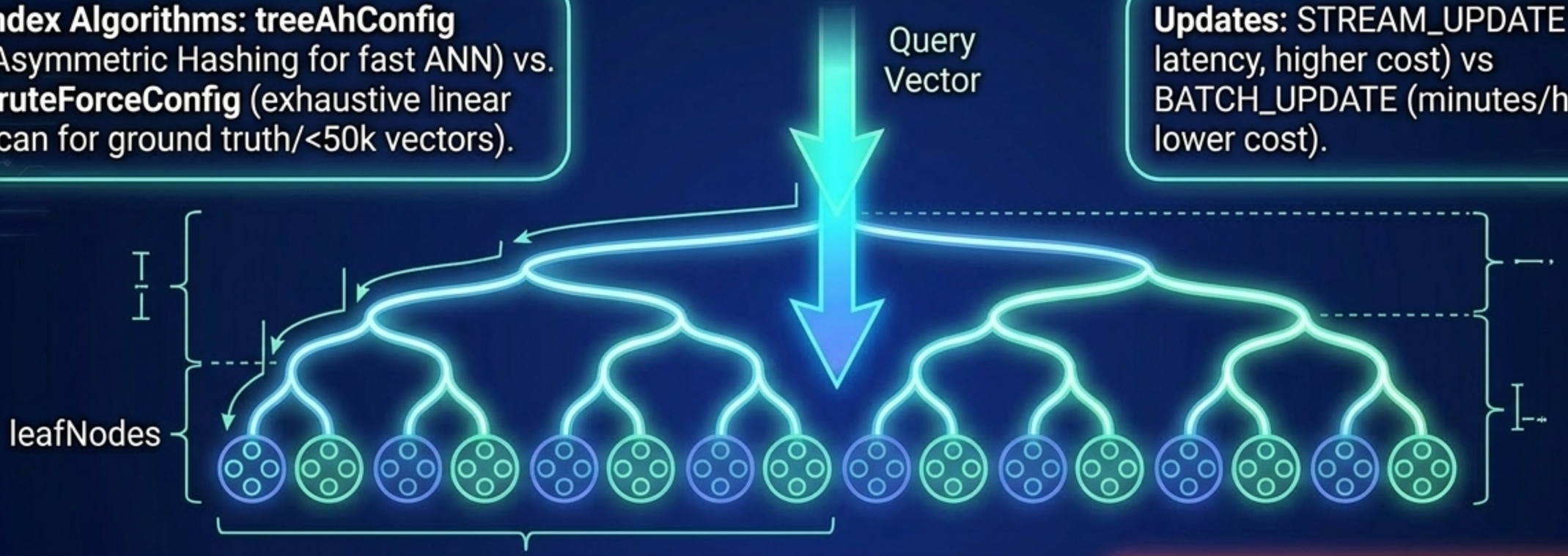


Trap Note: **The Matryoshka / Firestore Trap:** Slicing gemini-embedding-001 via output_dimensionality to fit Firestore's 2048d cap breaks unit length. Using raw DOT_PRODUCT biases results. You **MUST** L2 re-normalize the slice before indexing!

Tuning High-Scale Vector Search 1.0

Index Algorithms: treeAhConfig (Asymmetric Hashing for fast ANN) vs. **bruteForceConfig** (exhaustive linear scan for ground truth/<50k vectors).

Updates: STREAM_UPDATE (seconds latency, higher cost) vs BATCH_UPDATE (minutes/hours lag, lower cost).



Performance Dials:

- leafNodesToSearchPercent: Primary latency vs. recall dial.
- approximateNeighborsCount: Pre-reordering pool (keep strictly > k).
- crowding: Restricts max neighbors per document_id to prevent verbose-doc monopoly.

Trap Note: RAG Engine Distance Failure: RAG Engine uses cosine distance. If your backing Vector Search index is left at DOT_PRODUCT, runtime queries fail with FAILED_PRECONDITION.

The Hands: Trusted Runtime vs. Untrusted Sandbox

Agent Runtime (Trusted)

- Executes your engineered ADK/LangGraph code.
- Holds IAM/MCP credentials and secure VPC paths.

Code Execution Sandbox (Untrusted)

- Executes model-generated Python inside a hardened micro-container.
  - Sub-second provisioning, 100 MB file I/O limit.
  - Preserves stateful variables/files across turns for up to 14 days.
  - In Managed Agents API, runs with no outbound network access by default.



Trap Note: The **exec() Vulnerability**: Running LLM-generated Python via exec() inside the main agent container exposes service-account tokens and internal network endpoints to prompt injection. Always sandbox!

First-Party Google Cloud MCP Servers



Protocol: Auto-registered in the global Agent Registry. Supports stateless StreamableHTTP MCP spec (2026-07-28) with no Mcp-Session-Id.

The Catalogue:

- bigquery.googleapis.com/mcp: Schema inspection, dry-runs, SQL execution.
- container.googleapis.com/mcp: GKE pod diagnostics.
- compute.googleapis.com/mcp: VM inventory and lifecycle.
- developerknowledge.googleapis.com/mcp: Public developer docs grounding.

Trap Note: Regional Masking: Auto-registered 1P servers exist only in --location=global. Querying us-central1 returns NOT_FOUND.

Trap Note 2: VPC-SC Block: Developer Knowledge MCP indexes public docs and does not support the restricted VIP. Route egress via private.googleapis.com.

Enterprise Tool Integration Layers

1. SaaS & ERP (100+ Platforms)
- **Layer**: Integration Connectors + Application Integration (ApplicationIntegrationToolset).
- **Auth**: Auth Manager 3LO user-delegated consent.



2. Relational & Analytical Databases
- **Layer**: MCP Toolbox for Databases (ToolboxToolset).
- **Features**: Sub-50ms pooled queries across 30+ DBs, declarative *tools*.yaml, pre-built schemas.

3. Cataloged Microservices
- Layer: Apigee + API Hub (ApiHubToolset).
- Features: OpenAPI REST, token quotas, mTLS, inline Model Armor.

4. Polyglot Domain Logic
- Layer: Custom Remote MCP on Cloud Run (McpToolset).

Trap Note: Header Split: Cloud Run consumes Authorization for OIDC verification. Pass downstream user OAuth tokens in X-Serverless-Authorization or delegate to Agent Gateway.

Agent Registry & Skill Governance

Control Plane

Agent Registry: Central control-plane catalog for AgentCards, McpServers, and standalone skills.

- **Discovery vs Invocation:** roles/agentregistry.viewer grants discovery only. Invocation requires roles/iap.egressor + endpoint IAM.
- **Registration:** Manual registration of external MCPs requires uploading a 10 KB toolspec.json.

Semantic Governance interceptor

Skill Registry (Preview): SKILL.md (max 500k chars) + scripts/assets. Loaded dynamically at runtime.
Semantic Governance Policies: NLCs (Natural Language Constraints) that intercept load_skill or run_skill_script. Can deny skills if session history becomes tainted by external web inputs.



Trap Note: Privilege Escalation: Never grant agentregistry.admin to an agent identity. A compromised agent could alter tool definitions globally.

Transports, Triggers, and Timeouts

SSE (Server-Sent Events)		Unidirectional stream (max 60 min). Great for chat UI. Cannot support voice barge-in.
Gemini Live API		Bidirectional full-duplex WebSockets. Requires gemini-3.8-live. Enables VAD barge-in with session resumption handles. Not supported on Gemini 4 Argon.
Pub/Sub Push & Eventarc		Decoupled fan-out. 600s ack deadline.
Cloud Tasks		Rate-limited point-to-point dispatch (maxDispatchesPerSecond) to absorb bursts.
Cloud Workflows		Deterministic multi-day serverless sagas (1-year wait time).



Trap Note: **Pub/Sub Duplicate Mutation:** Multi-step agents exceeding the 600s ack deadline trigger redelivery. Enforce atomic messageId idempotency checks in Memorystore + add a Dead-Letter Topic (DLT).

Agent Evaluation & Metric Hierarchy



Adaptive Rubrics: Autorater generates prompt-specific testable pass/fail criteria dynamically. Recommended default over static LLM judges.

Agent Outcome Metrics: FINAL_RESPONSE_QUALITY, GROUNDING (entailment against retrieved chunks), HALLUCINATION (supported/disputed span labels).

Trajectory Metrics: Compares tool-call sequence against a golden trajectory (trajectory_exact_match, trajectory_in_order_match).

Computation Metrics: Deterministic Python regex/schema validators (sub-ms latency, 0 LLM cost).

Trap Note: 'Right Answer, Wrong Path': FINAL_RESPONSE_QUALITY alone passes an agent that issues a refund without checking eligibility. Always mandate trajectory matching for compliance!

Simulation, Observability & Prompt Optimization

2-Step Agent Simulation:
- Step 1: Synthesize a hidden Conversation Plan (UserPersona, goals, violation_rubrics for red-teaming).



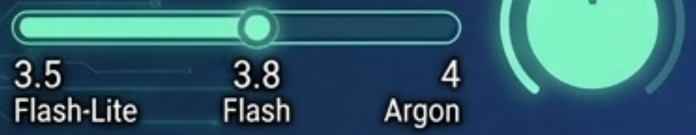
- Step 2: UserSimulator conducts live multi-turn execution, emitting OpenTelemetry traces.

Environment Simulation:
Injects deterministic tool faults (503s, 429s) without mutating production systems.

The Flywheel:
- Capture production traces (OTel) → generate_loss_clusters() to group root causes.
- De-identify via Sensitive Data Protection (SDP).
- adk optimize automatically mutates system instructions to fix loss clusters.
- Rule: Always split sets (Train vs Held-Out Validation) to prevent instruction overfitting.

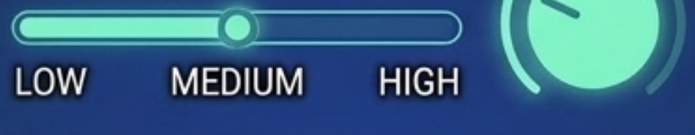
Cost Engineering, Quotas & FinOps

Tiered Routing



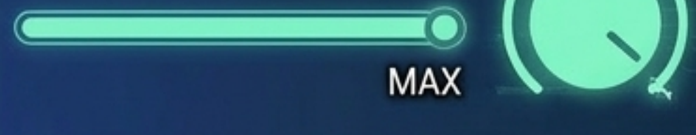
3.5 Flash-Lite (guardrails) → 3.8 Flash (tools) → 4 Argon (complex synthesis).

Thinking Level



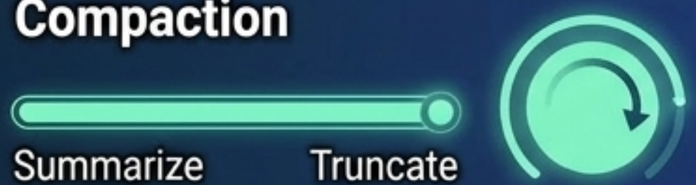
Default to LOW/MEDIUM; reserve HIGH for ambiguity.

Prefix Alignment



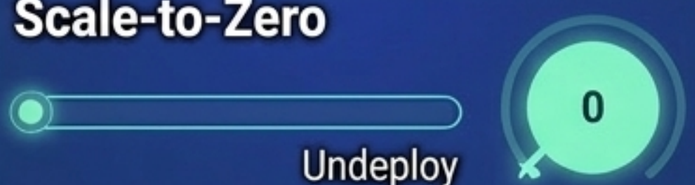
Byte-identical system prompts maximize implicit cache hits (95% discount).

Compaction



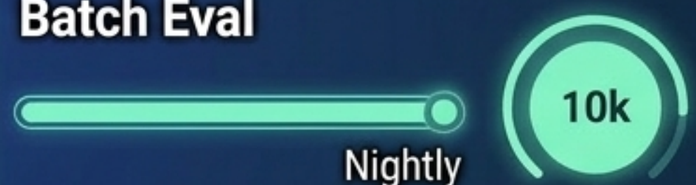
Summarize old events; truncate massive MCP JSON blobs.

Scale-to-Zero



Undeploy idle Vector Search 1.0 indexes (stops 24/7 billing).

Batch Eval



Run nightly 10k-case evaluations on Flex PayGo or Batch.



Trap Note: **Capacity vs Quota:** A 'Region temporarily at capacity' error during deploy means physical compute pool exhaustion, NOT a project quota limit. A quota ticket will not fix this; switch regions or reduce vCPU.

Synthesis: Enterprise Reference Architectures

1. Customer Support Agent

- **Runtime**: Agent Runtime.
- **State**: Sessions + Memory Bank (device history).
- **Knowledge**: Agent Search (PDFs).
- **Tools**: Application-IntegrationToolset (CRM).

2. Data Analysis Agent

- **Runtime**: Agent Runtime.
- **State**: Sessions + Memory Bank (analyst prefs).
- **Knowledge**: BigQuery MCP.
- **Tools**: Code Execution Sandbox (for Python charts) + Cloud Storage Artifacts.

3. High-QPS Public FAQ Agent

- **Runtime**: Cloud Run (autoscaled).
- **State**: Firestore subcollections (ephemeral).
- **Knowledge**: Agent Retrieval behind a Memorystore semantic cache.
- **Tools**: Memorystore IP rate limiters.

4. Regulated Industry Agent

- **Runtime**: Agent Runtime (via PSC) + Private GKE vLLM (VPC-SC).
- **State**: DatabaseSessionService (AlloyDB with PgBouncer, CMEK).
- **Knowledge**: AlloyDB scann vectors (row-level security).
- **Tools**: Custom Cloud Run MCP via Agent Gateway (IAP).