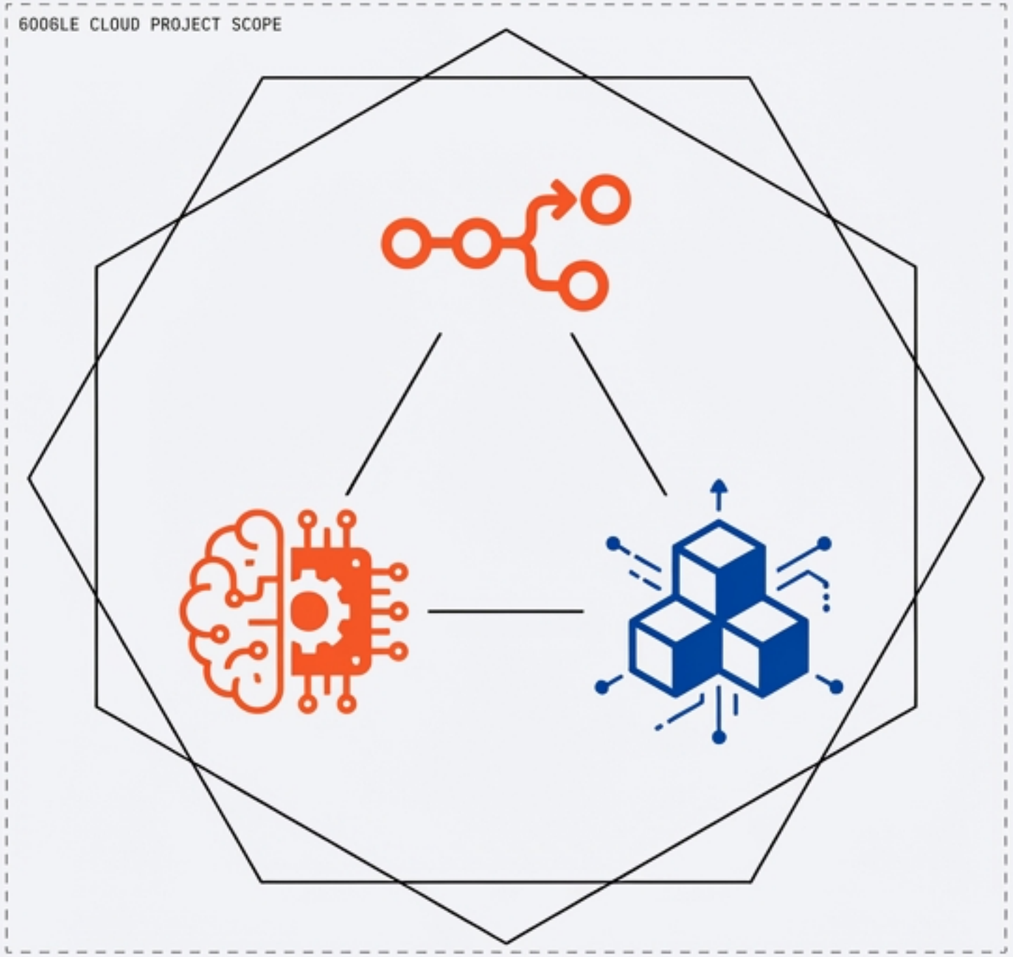


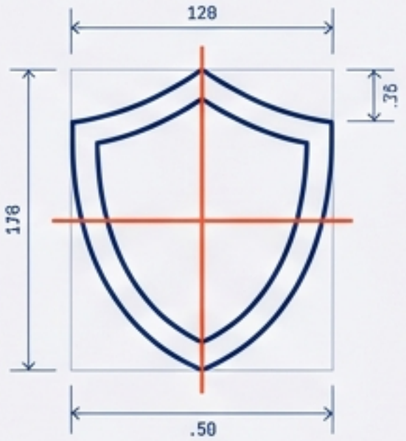
N8N AI on Google Cloud Platform: The Sovereign Stack

A supercharged deployment pre-configured for the era of Generative AI.

An enterprise-grade solution for building intelligent agents, chatbots, and document analysis workflows.

Leveraging state-of-the-art Local LLMs and Vector Databases, hosted securely in your own private cloud environment.

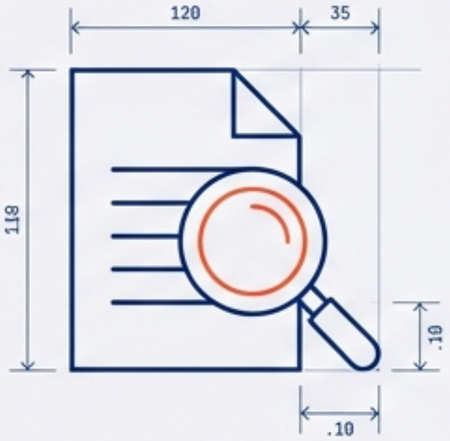
BUILT FOR PRIVACY, AGILITY, AND CONTROL



PRIVACY FIRST AI

Run Large Language Models like Llama 3 locally. Your sensitive data never leaves your cloud project.

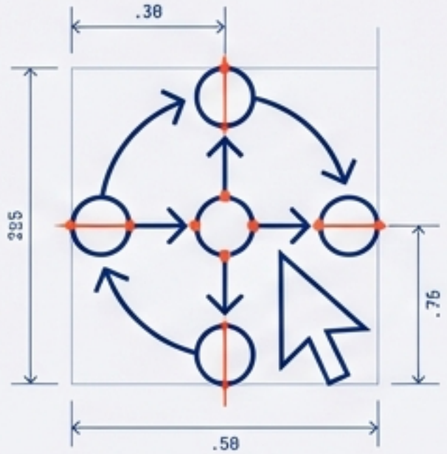
[SPEC-001: DATA SOVEREIGNTY]



RAG READY

Ready for Retrieval Augmented Generation immediately. Includes Qdrant to let AI read and understand company documents.

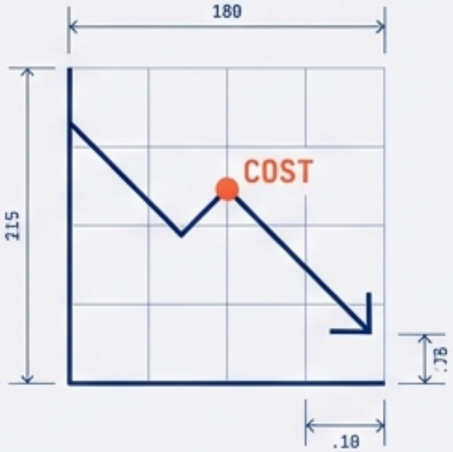
[SPEC-002: KNOWLEDGE RETRIEVAL]



NO-CODE BUILDING

Leverage n8n's drag-and-drop interface to construct complex AI chains without a team of ML engineers.

[SPEC-003: WORKFLOW AUTOMATION]

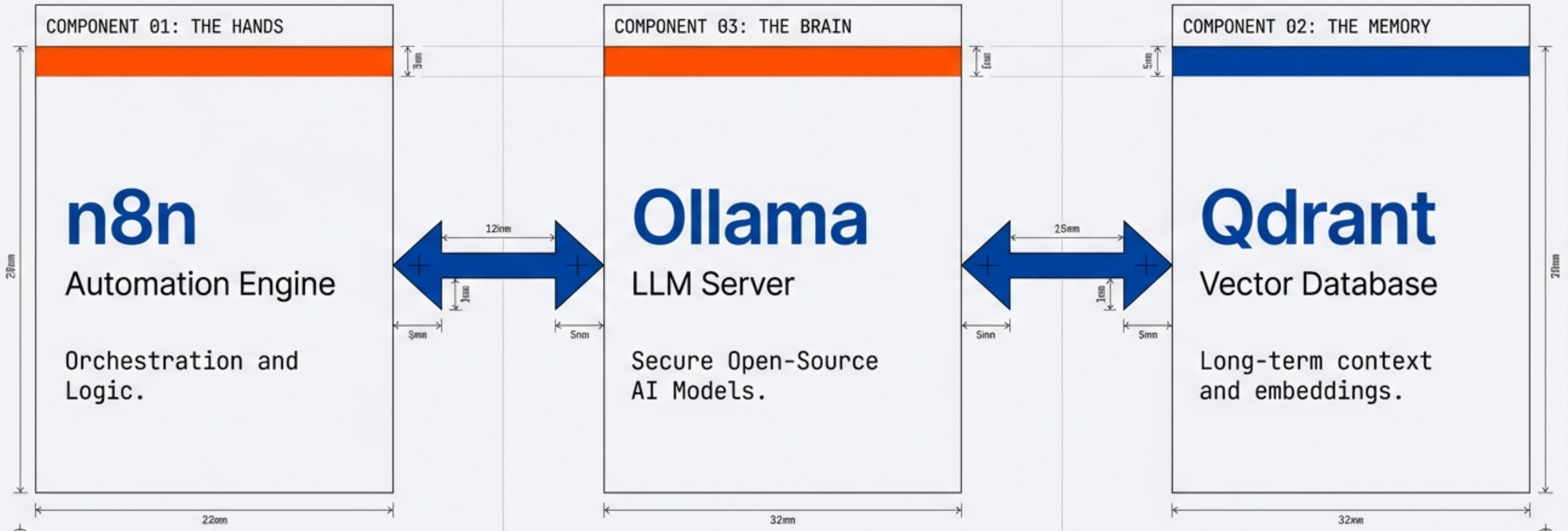


COST CONTROL

Eliminate unpredictable API costs from public providers by running your own models on fixed infrastructure.

[SPEC-004: OPERATIONAL EFFICIENCY]

THE ANATOMY OF A PRIVATE AI ECOSYSTEM



System Status: The RAD module connects all three components automatically. Configuration is "out of the box".

LOCAL INTELLIGENCE VIA OLLAMA

CORE ROLE: Deploys Ollama on Cloud Run to serve open-source AI models.

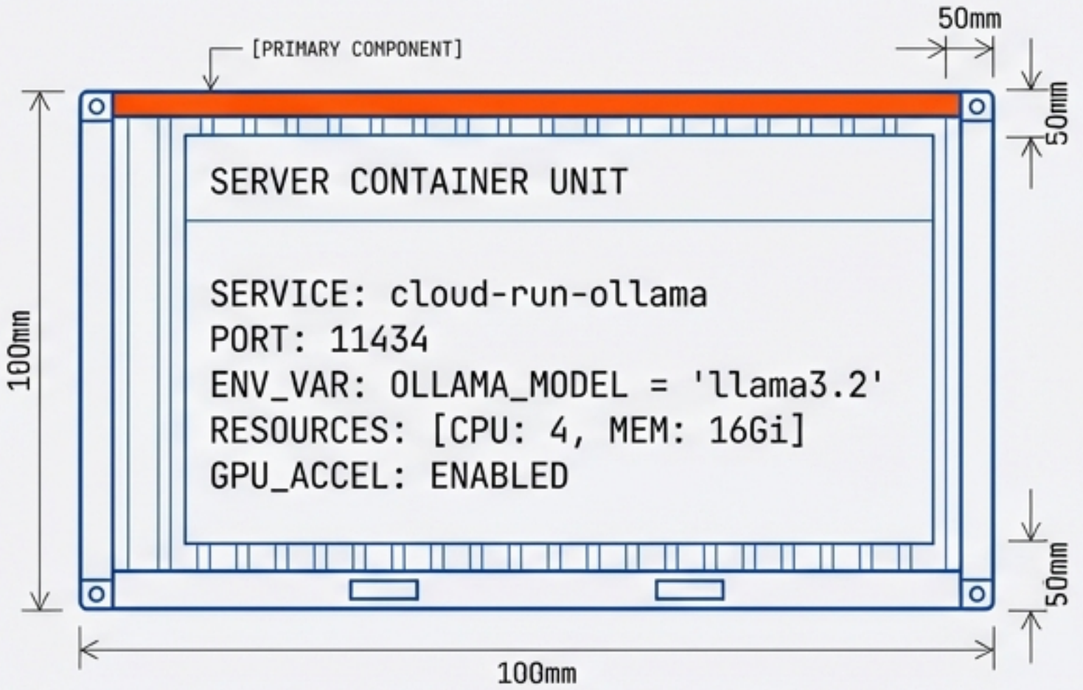
[SPEC-881: OLLAMA DEPLOY]

HARDWARE SPECS: Supports GPU acceleration where available, or CPU execution for smaller optimized models.

[SPEC-862: OLLAMA DEPLOY]

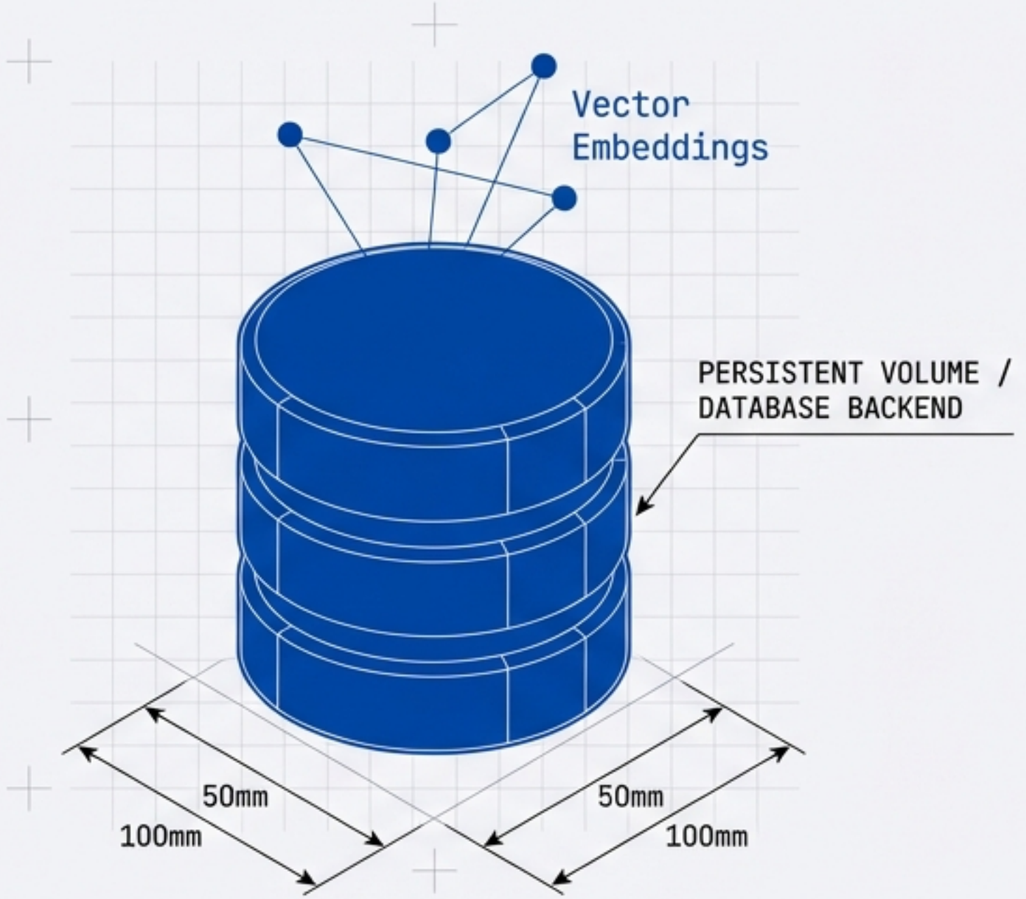
CONFIGURATION: Uses the 'ollama_model' variable to specify exact model pulls (e.g., llama3.2) at startup.

[SPEC-882: OLLAMA CONFIG]



TECHNICAL NOTE:
Monitor resource limits (memory, cpu) closely. LLMs are resource-intensive processes.

LONG-TERM CONTEXT VIA QDRANT



- **CORE ROLE:** Deploys the Qdrant container as a dedicated Vector
- **FUNCTION:** Stores high-dimensional vectors (embeddings) to provide AI context beyond the immediate prompt.
- **PERSISTENCE:** System utilizes persistent storage volumes so memory is retained across restarts.
- **INTEGRATION:** Automatically configured as a credentialed node within n8n. No manual setup required.

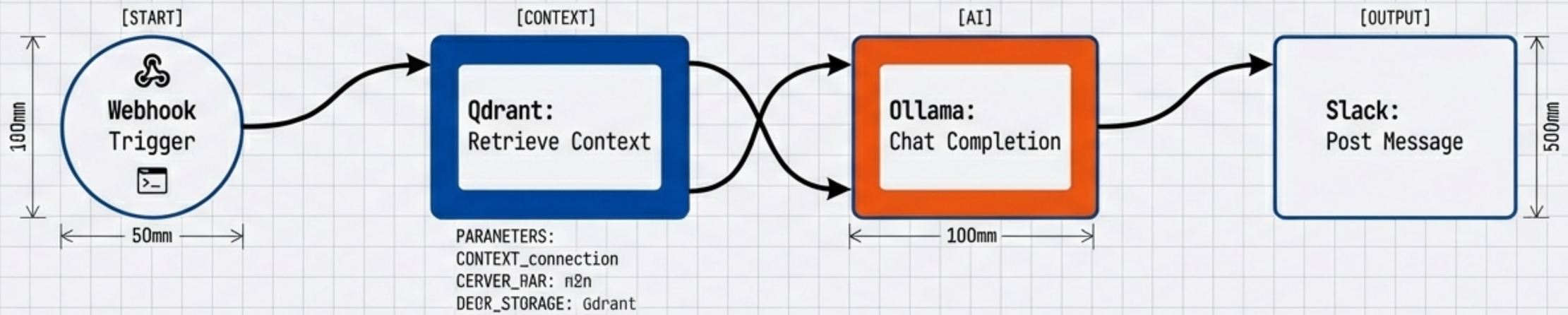
System Status: The RAD module connects all three components automatically. Configuration is 'out of the box'.

ORCHESTRATION WITHOUT CODE

CORE ROLE: n8n acts as the Automation Engine, dictating when to consult memory (Qdrant) when to query intelligence (Ollama). Replacing complex code with a visual drag-and-drop builder.

JETBRAINS NONO

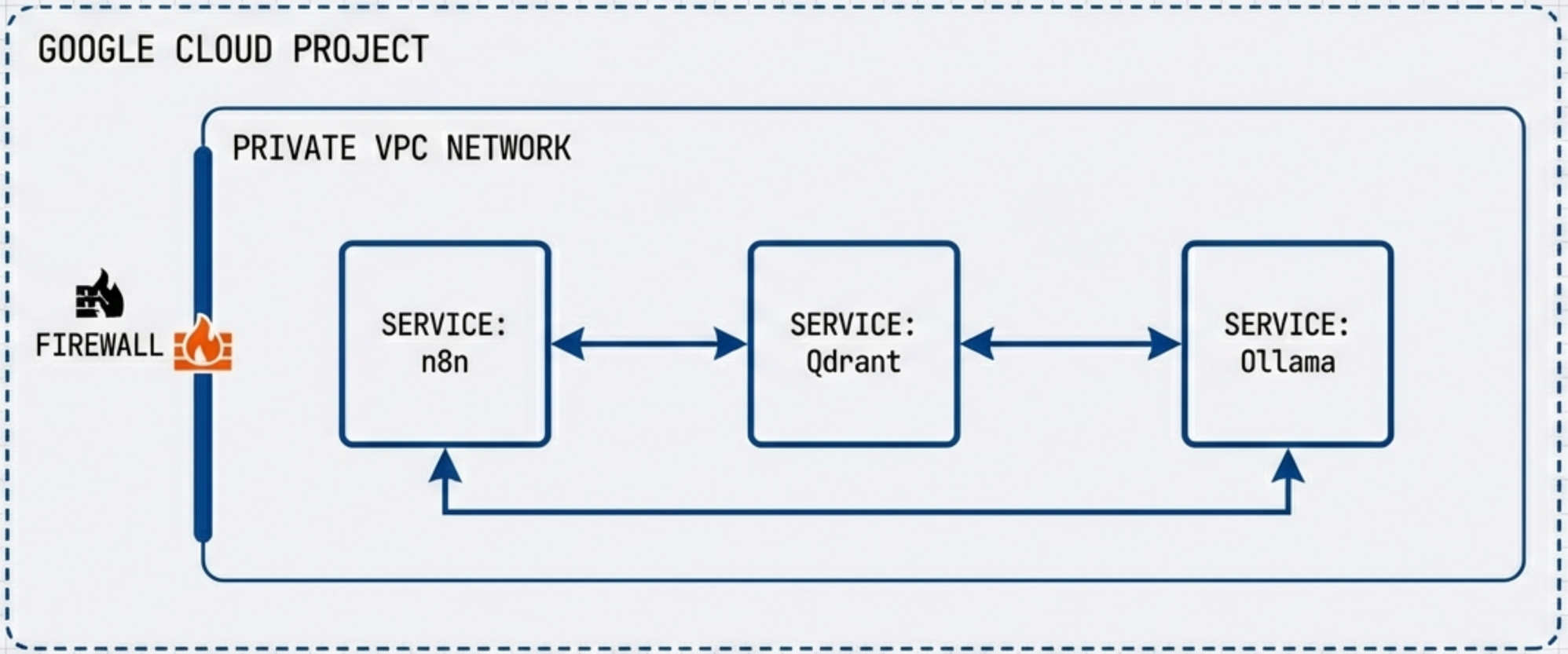
EXAMPLE: INTELLIGENT AGENT PIPELINE



JETBRAINS MONO

System Status: The RAD module connects all three components automatically. Configuration is 'out of the box'.

CLOUD RUN ARCHITECTURE & TOPOLOGY

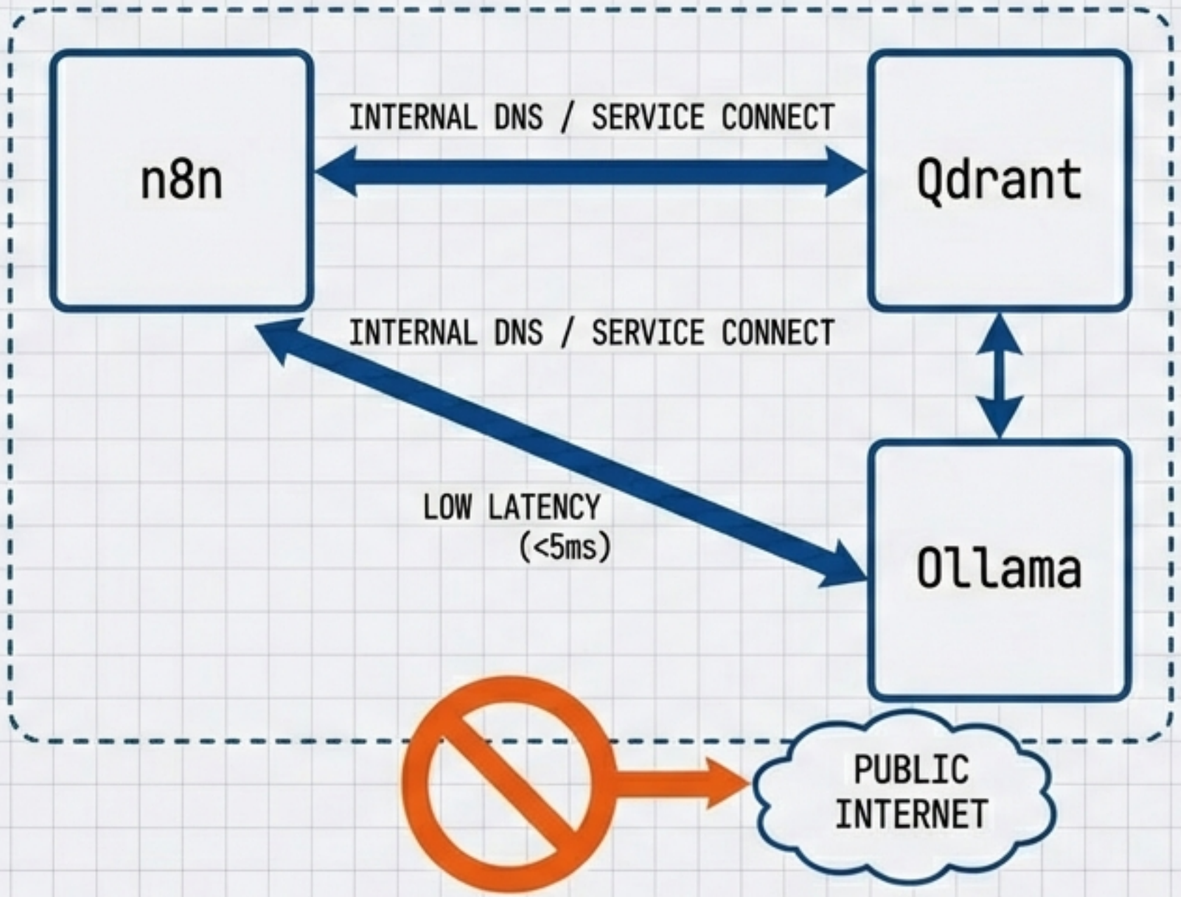


Service Relationship: While deployed as distinct Cloud Run services, they function as a cohesive unit within a private enclave.

JET8841K5 NOKD

System Status: The RAD module connects all three components automatically. Configuration is 'out of the box'.

ZERO-TRUST NETWORKING



Mechanism: Uses internal VPC DNS or Service Connect to facilitate communication.

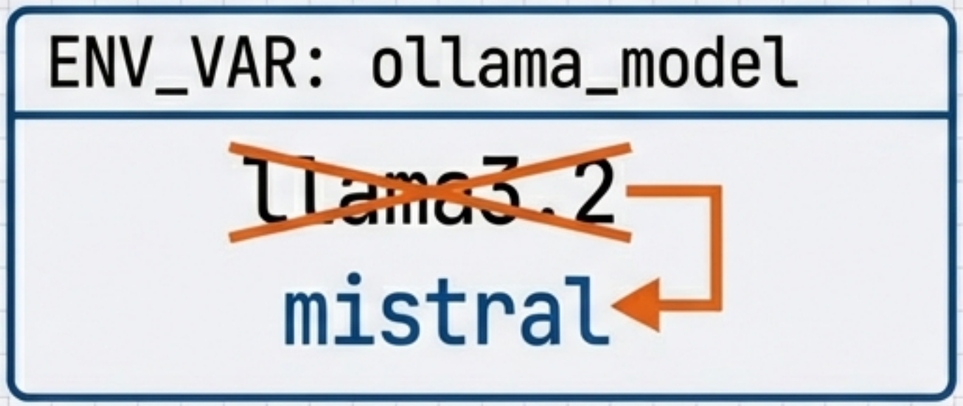
Security Guarantee: The system allows n8n to talk to Qdrant and Ollama without public internet exposure.

Takeaway: Your infrastructure provides the barrier; no data exits to third-party API providers.

System Status: The RAD module connects all three components automatically. Configuration is 'out of the box'.

DYNAMIC MODEL MANAGEMENT

Future-proof your stack against the evolution of open-source models.



The Mechanism: The "ollama_model" environment variable.

Capability: Switch between different open-source models (Mistral, Gemma, Llama) without redeploying the entire infrastructure.

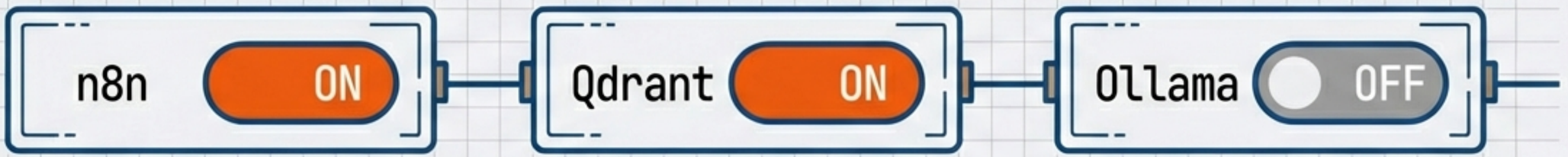
Benefit: Adaptability to new model releases instantly.

System Status: The RAD module connects all three components automatically. Configuration is 'out of the box'.

MODULAR FEATURE TOGGLES

JETBRAINS MONO

USE CASE: HYBRID ARCHITECTURE



Scenario: Use n8n and Qdrant for private memory, but route intelligence to OpenAI API via n8n nodes.

USE CASE: LEAN DEPLOYMENT

Scenario: Run pure automation workflows to save resources when AI components are not required.

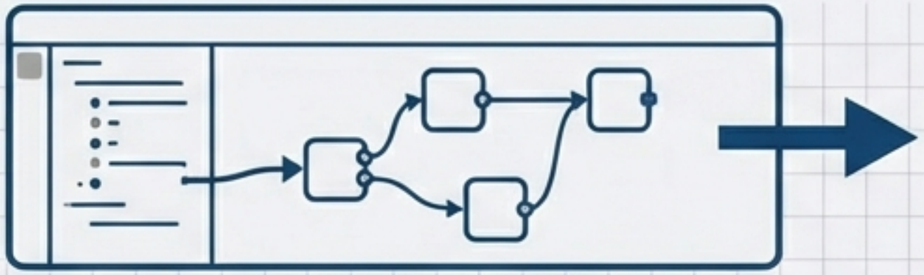
VARIABLES: enable_qdrant / enable_ollama

JET8041A3 NOKD

System Status: The RAD module connects all three components automatically. Configuration is 'out of the box'.

THE COMPLETE RAD AI PLATFORM

JETBRAINS MONO

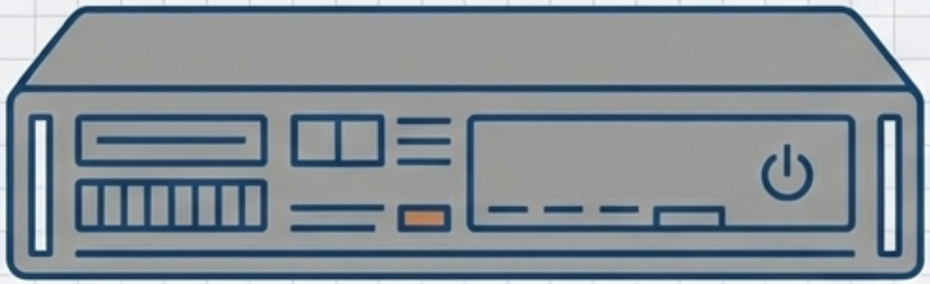


USER INTERFACE

No-Code Orchestration

THE TRINITY

Logic + Memory + Intelligence



INFRASTRUCTURE

Google Cloud Run / Private VPC / GPU & CPU

.JET8041A3 NOAD

A private, cost-controlled, and RAG-ready ecosystem for building the next generation of intelligent agents.