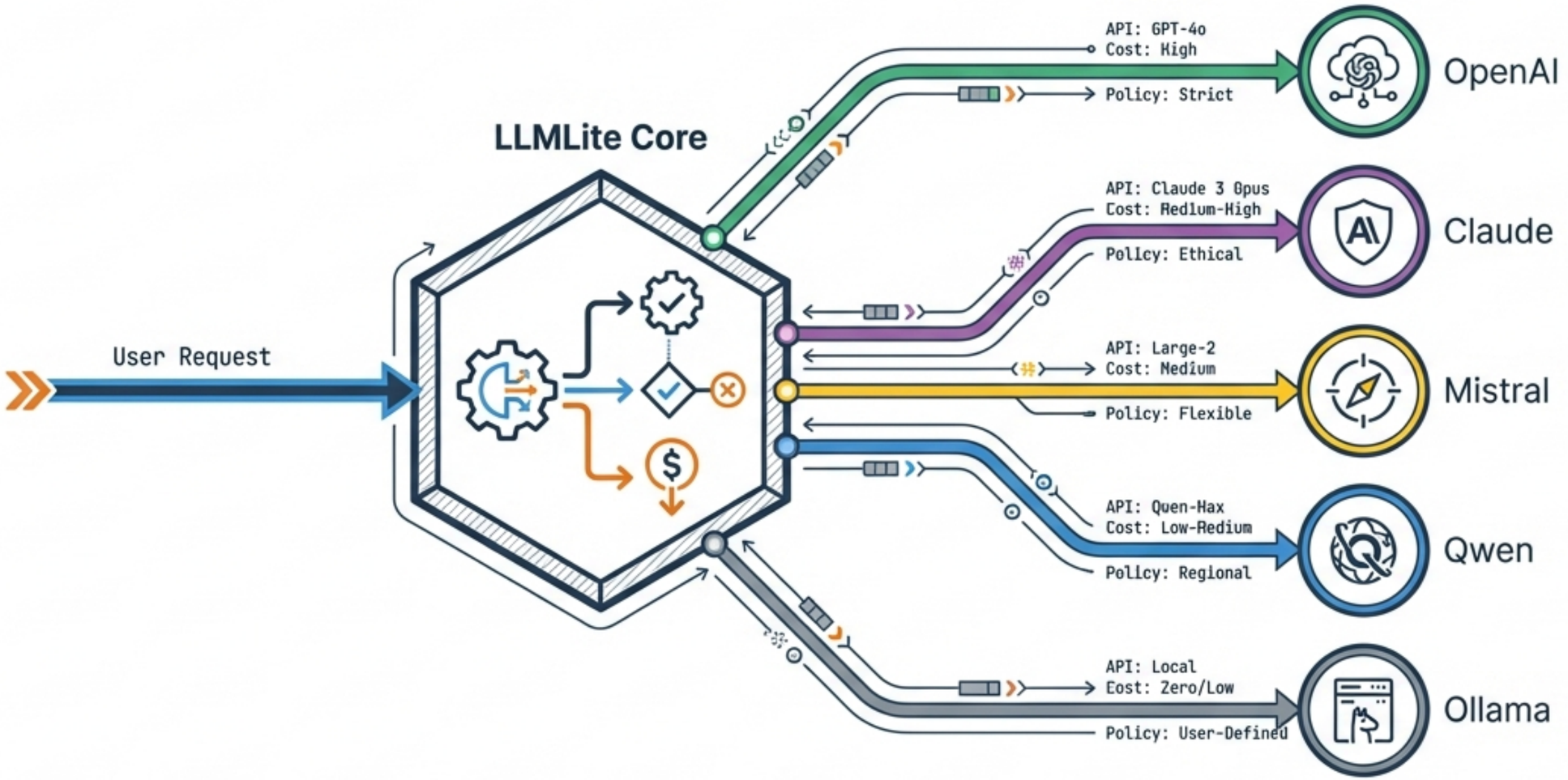
LLMLite: The Universal Model Gateway

An intent-driven architecture for cost-aware, policy-compliant, and vendor-agnostic Agentic Systems.

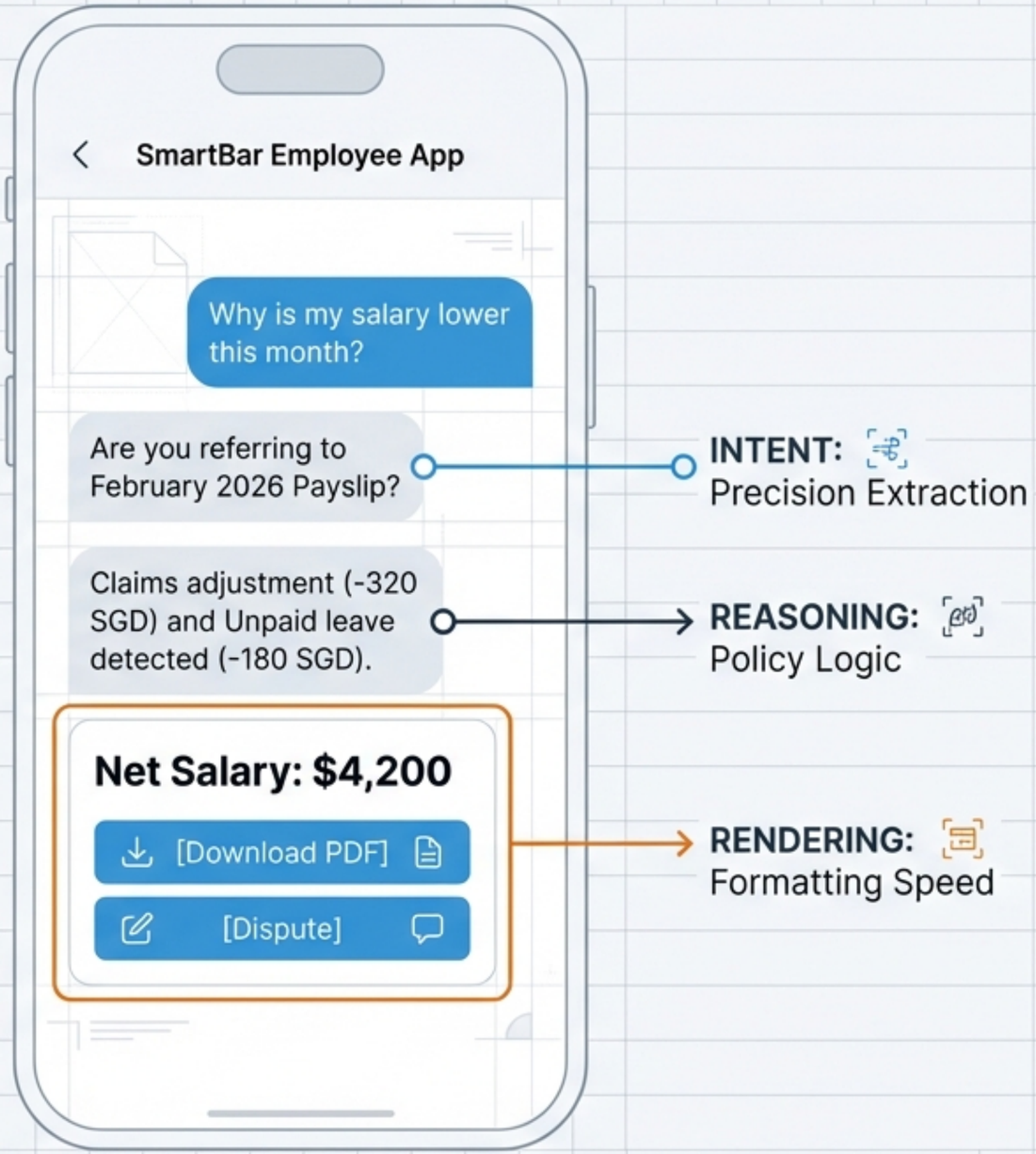


ARCHITECTURE PATTERN FOR INTELLIGENT ORCHESTRATION // V.2626

THE SCENARIO: A ‘SIMPLE’ PAYSLIP INQUIRY

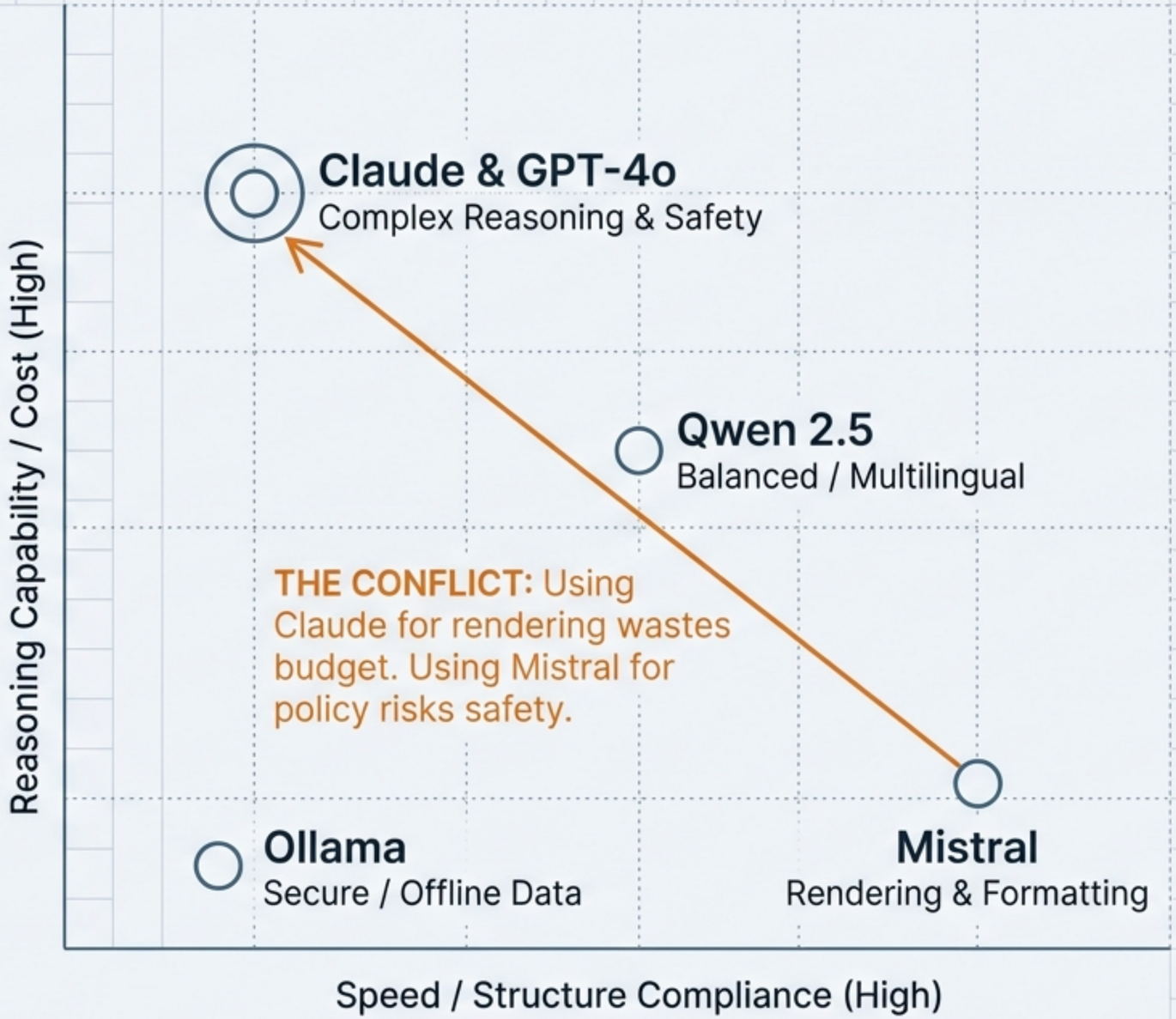
One user question triggers three distinct cognitive requirements.

To answer this, the system must Extract intent, Reason against policy, and Render a UI. Using a single model for all three is inefficient and risky.



The Intelligence Dilemma

Hardcoding a single model leads to ‘Spaghetti Architecture,’ vendor lock-in, and exploding costs.



Introducing LLMLite

A universal, policy-aware gateway that abstracts providers behind a single interface.

Model Abstraction

The App calls the Gateway, never the Provider SDK directly.

Intent-Based Routing

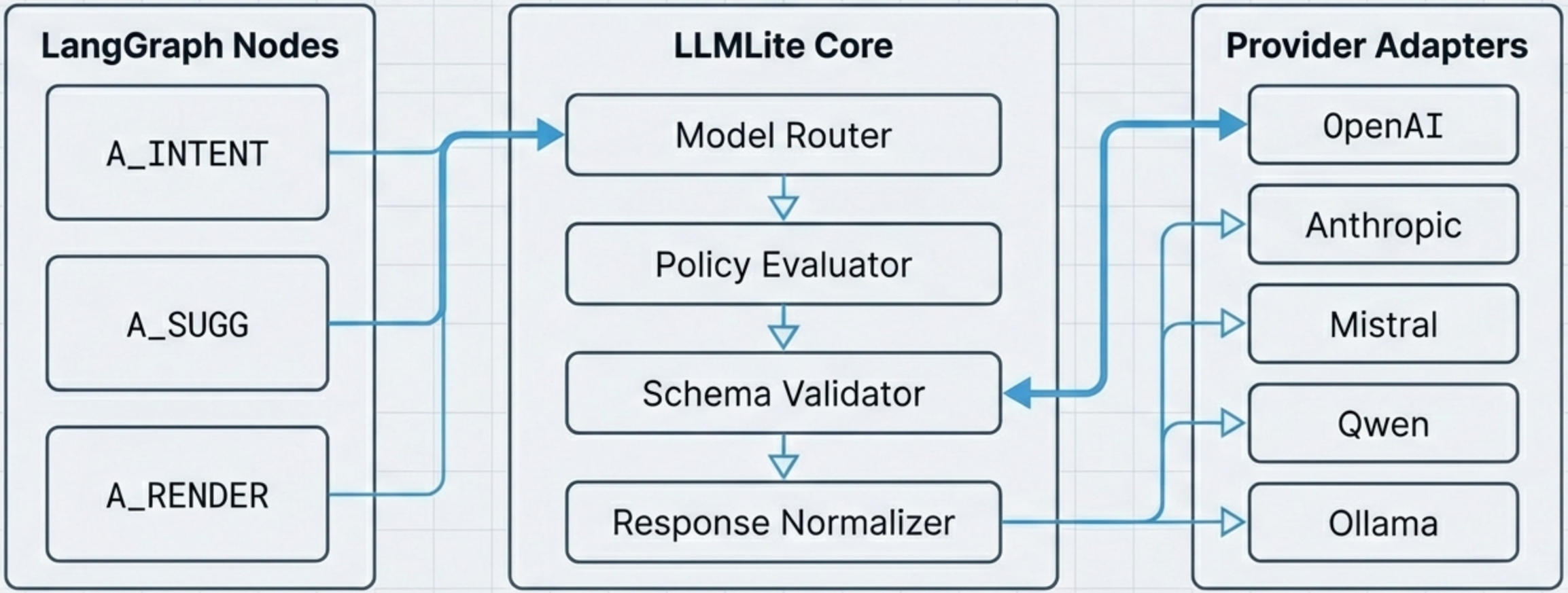
Route by Task Profile (e.g., “Extract”), not Model Name.

Deterministic Guardrails

Enforced schema validation, PII masking, and circuit breakers.

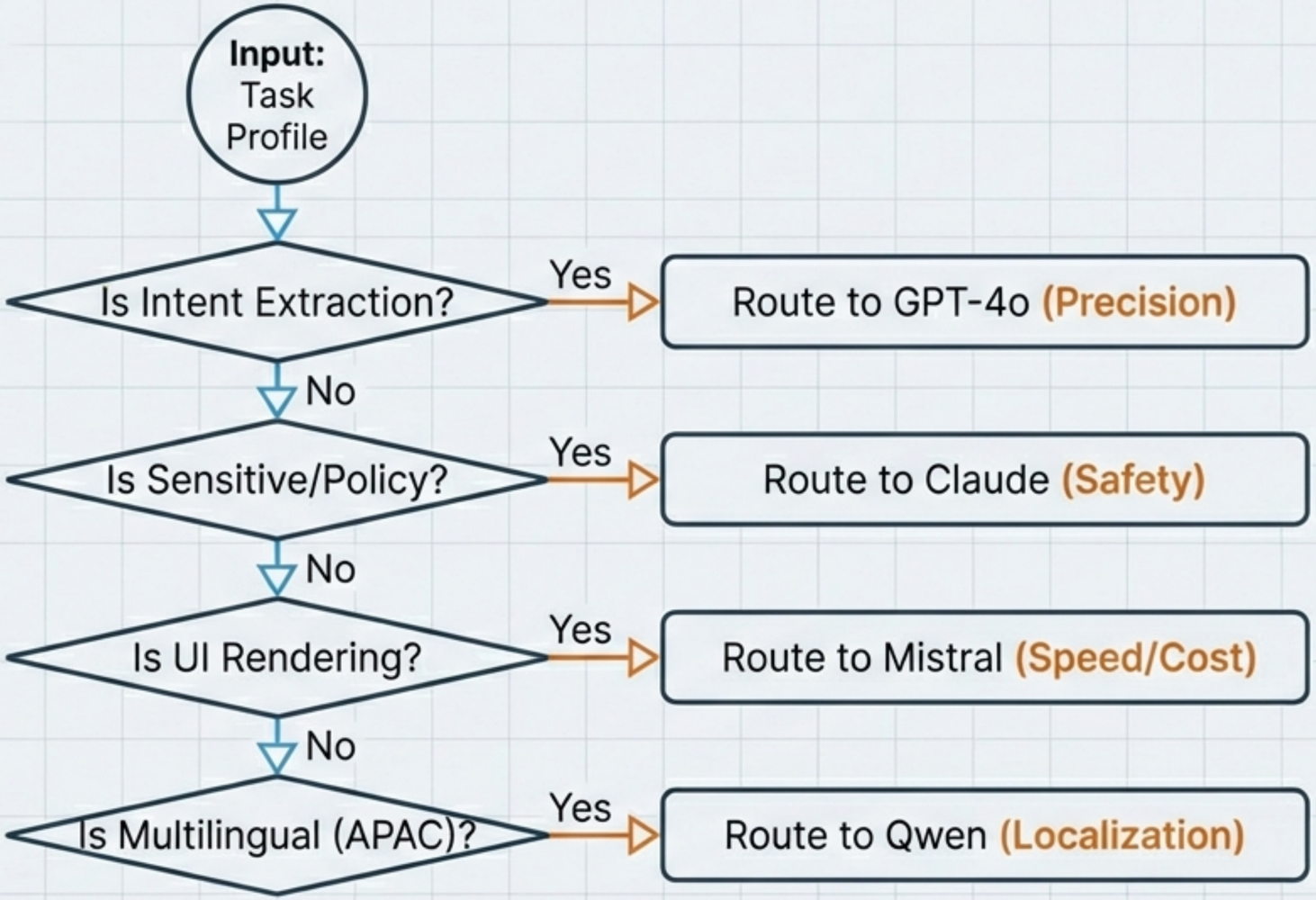
The Architecture

Inside the Universal Model Gateway.



Intent-Based Routing Logic

Dynamically assigning the right cognitive engine to the specific task profile.



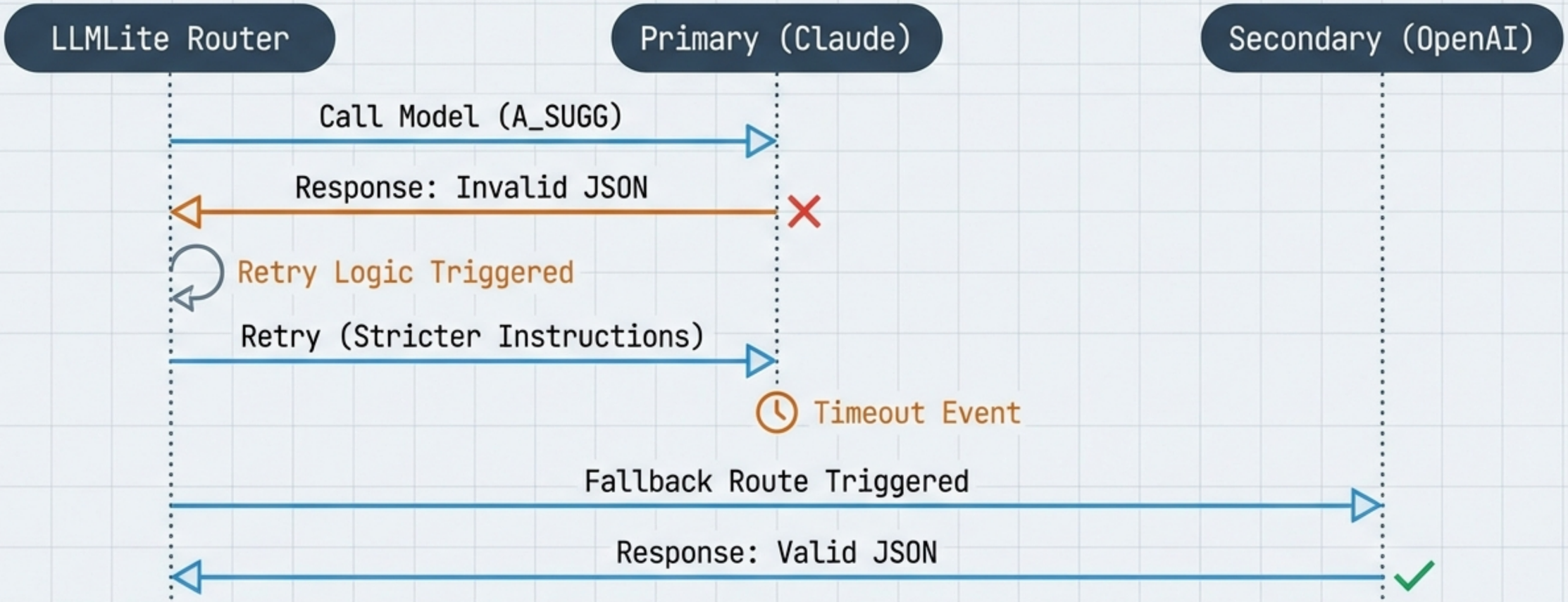
Routing is based on the nature of the work, not the name of the vendor.

Capability Matrix: The Right Tool for the Job

TASK PROFILE	PRIMARY MODEL	FALLBACK	WHY PRIMARY?	CONSTRAINT
Intent Extraction	GPT-4o	Qwen 2.5	Best consistent JSON	Strict Schema
Smart Suggestions (Policy)	Claude	GPT-4o	Safety & Policy adherence	No Hallucination
UI Rendering	Mistral	Qwen	Speed & Cost efficiency	Low Latency
Sensitive Data (PII)	Ollama (Local)	None	Data Sovereignty	Internal Network Only

Resilience & Self-Healing Strategies

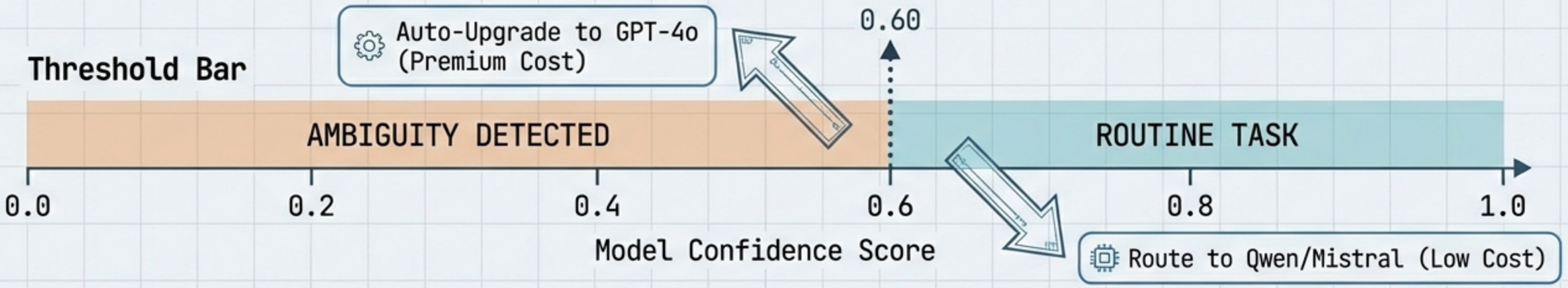
Ensuring reliability through automatic retries and provider fallbacks.



The application receives a valid response without ever knowing the primary provider failed.

Cost-Aware Intelligence

Dynamic “Confidence-Based Upgrades” optimize spend without sacrificing quality.



Production Guardrails

LLMLite never returns unvalidated or unsafe output to the application.

JSON Schema Enforcement

If it's not valid JSON, it doesn't leave the gateway.

No-New-Facts

Prevent hallucinations in rendering layers.

Circuit Breakers

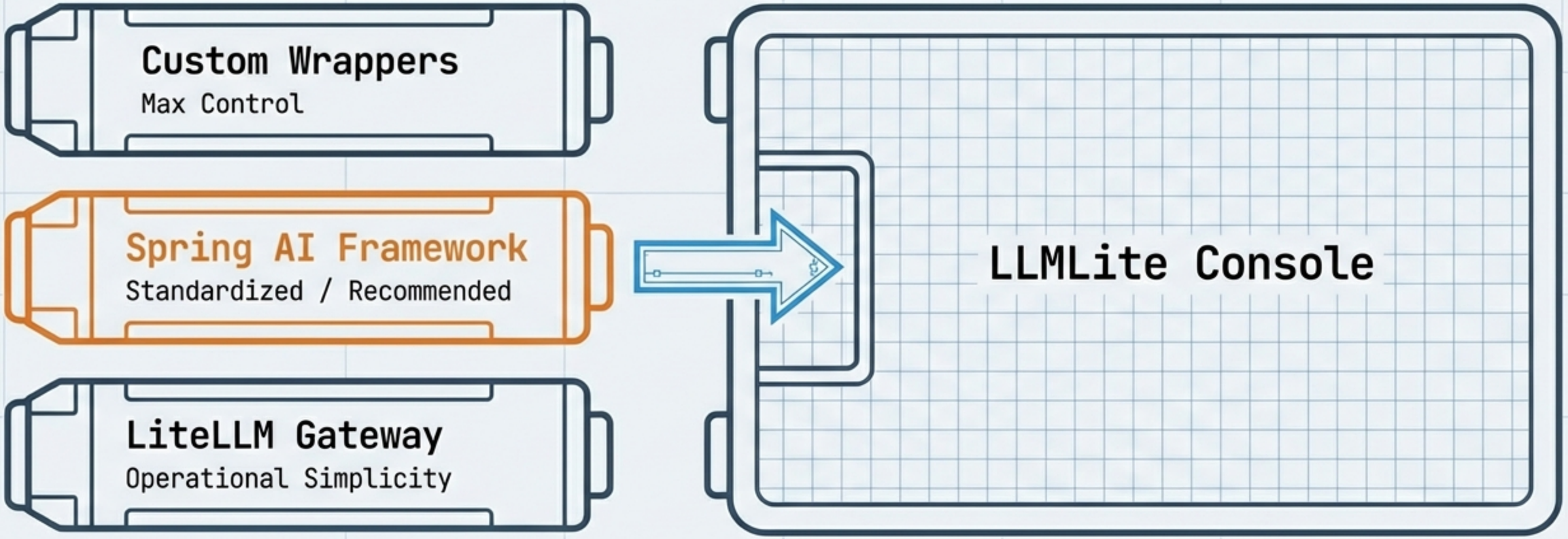
Stop calling dead providers to prevent cascading latency.

PII Redaction

Mask sensitive data before it reaches external APIs.

The Adapter Pattern

Universal plug-and-play integration for any model provider.



Switching from OpenAI to a local model is a configuration change, not a code rewrite.

Strategic Enterprise Advantages

Decoupling intelligence from the vendor.

Data Sovereignty

Route sensitive HR data to Local Models (Ollama) to keep it within the firewall.

Vendor Independence

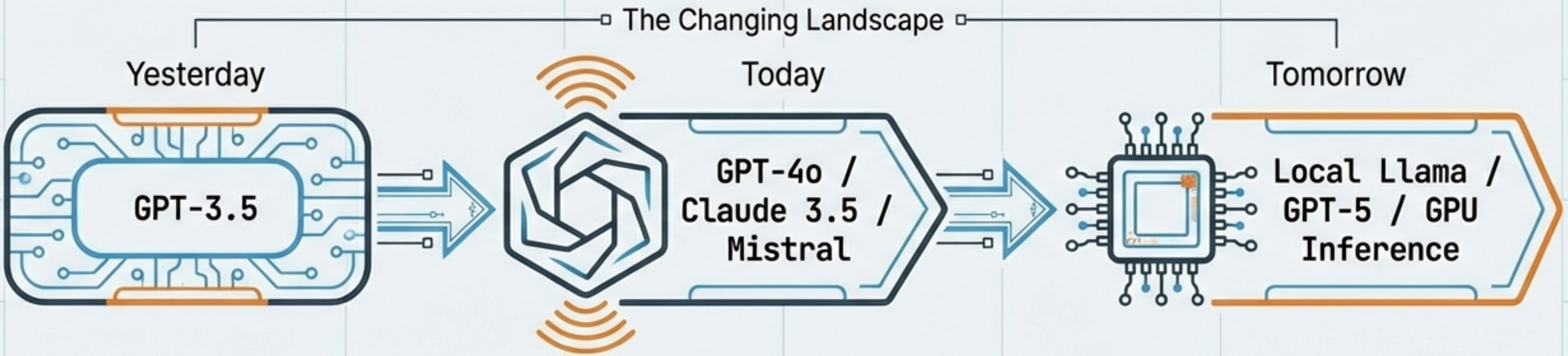
Replace OpenAI tomorrow with no code changes.

Regional Optimization

Automatically route APAC users to Qwen for better localization.

Future-Proofing the Platform

The model landscape changes weekly. Your architecture shouldn't.



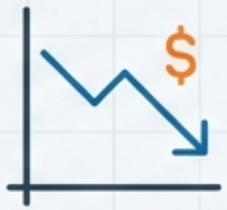
LLMLite Gateway Architecture

Add GPU-backed inference or new models simply by updating a ConfigMap.

The Executive Takeaway

“LLMLite is the universal intelligence switchboard that dynamically assigns the right model to the right cognitive task.”

GOVERNANCE
(Centralized)



COST
(Optimized)

AGILITY
(Vendor-Agnostic)