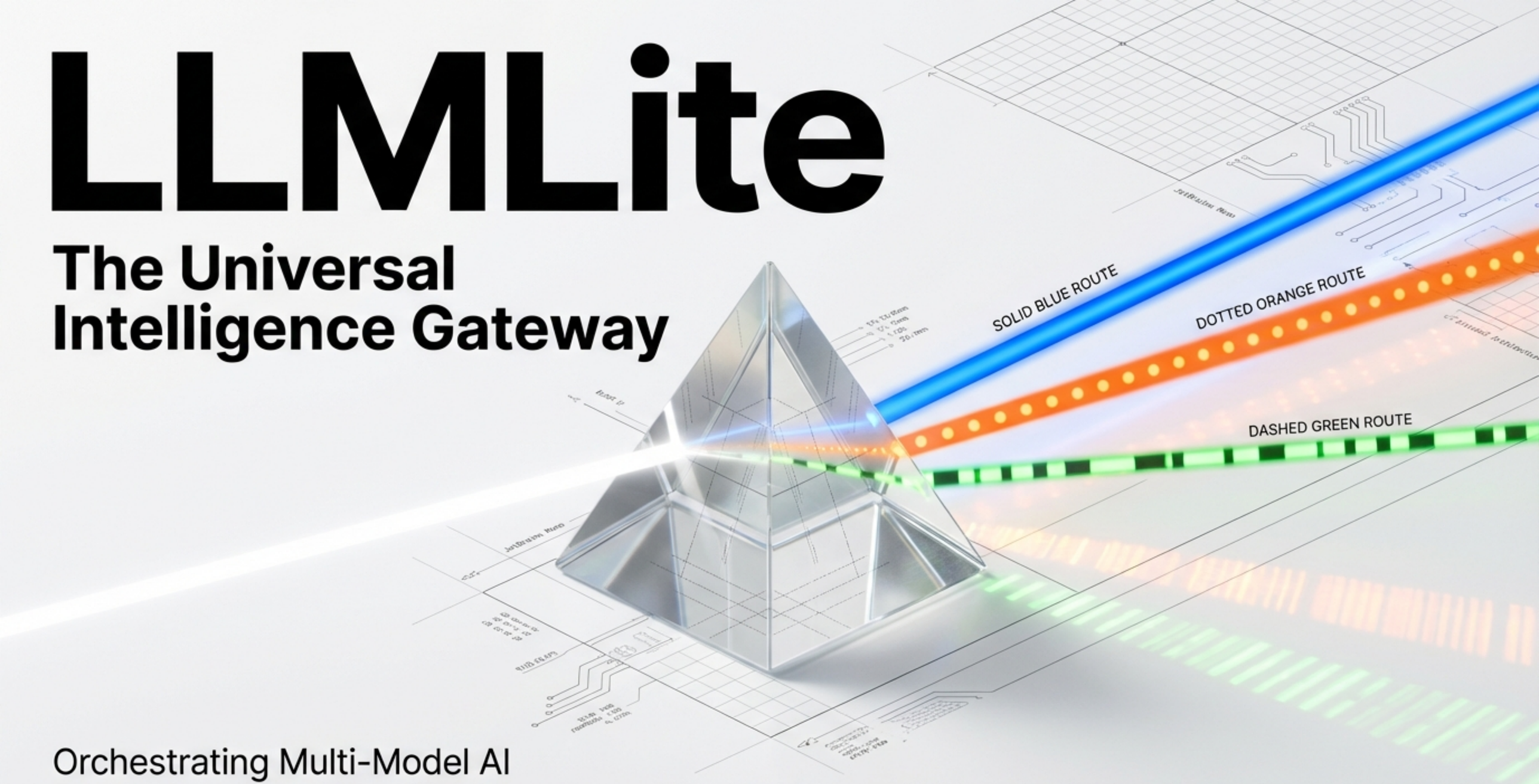
LLMLite

The Universal Intelligence Gateway



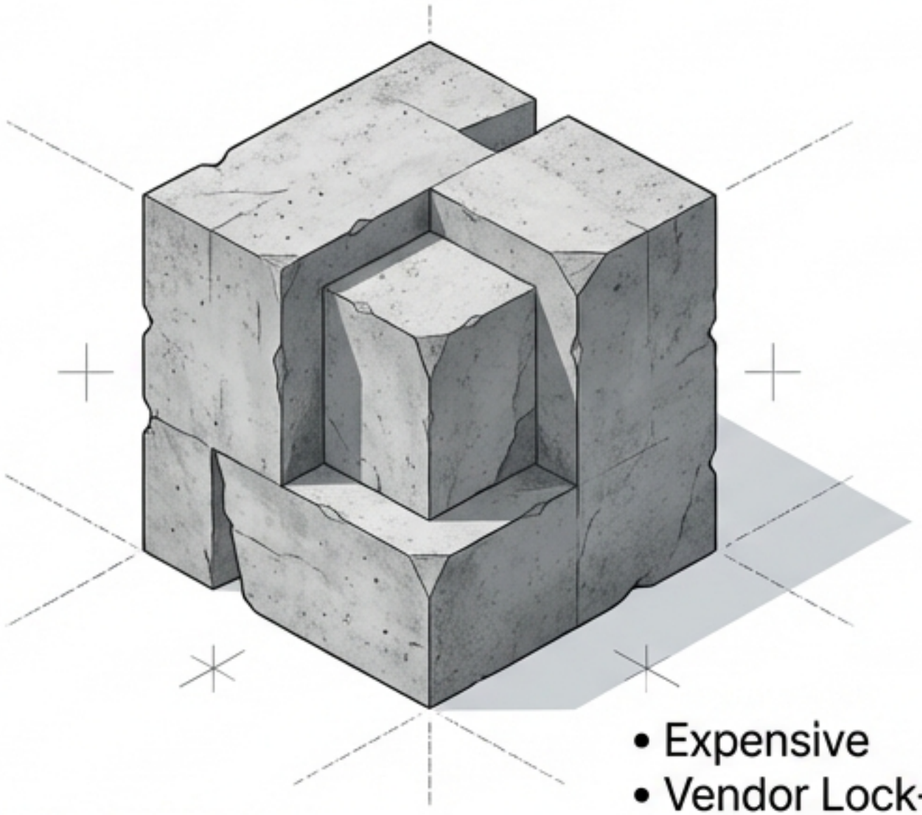
Orchestrating Multi-Model AI for the Employee Experience

Internal Architecture Brief

NotebookLM

The One-Size-Fits-None Problem

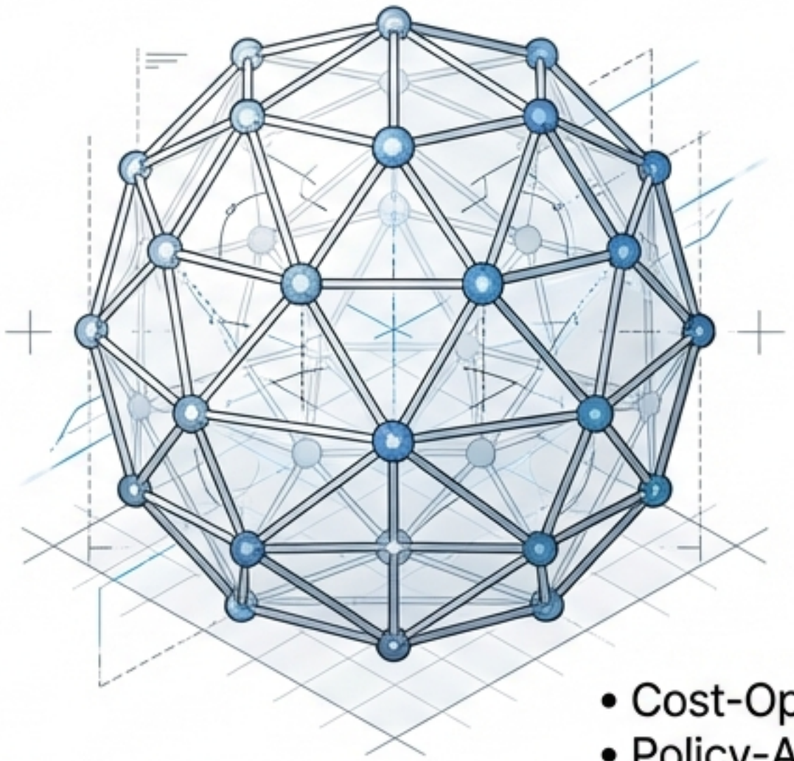
The Old Way



Hardcoded Monolith

- Expensive
- Vendor Lock-in
- High Latency

The LLMLite Way



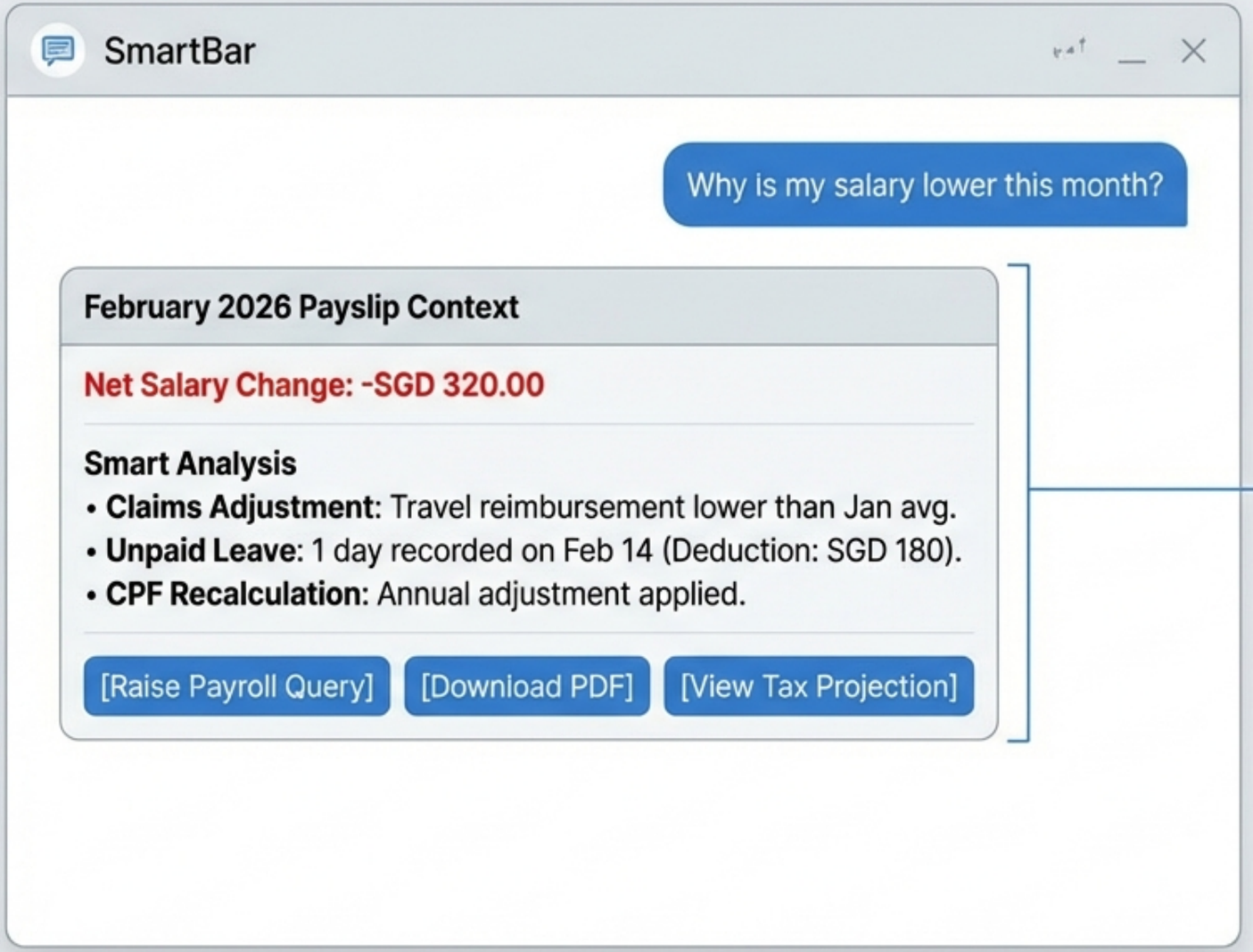
Dynamic Orchestration

- Cost-Optimized
- Policy-Aware
- Model-Agnostic

Relying on a single model creates a “Goldilocks” failure: premium models are too expensive for simple formatting, while cheaper models are unsafe for complex policy reasoning.

NotebookLM

The Scenario: Employee Payroll Inquiry

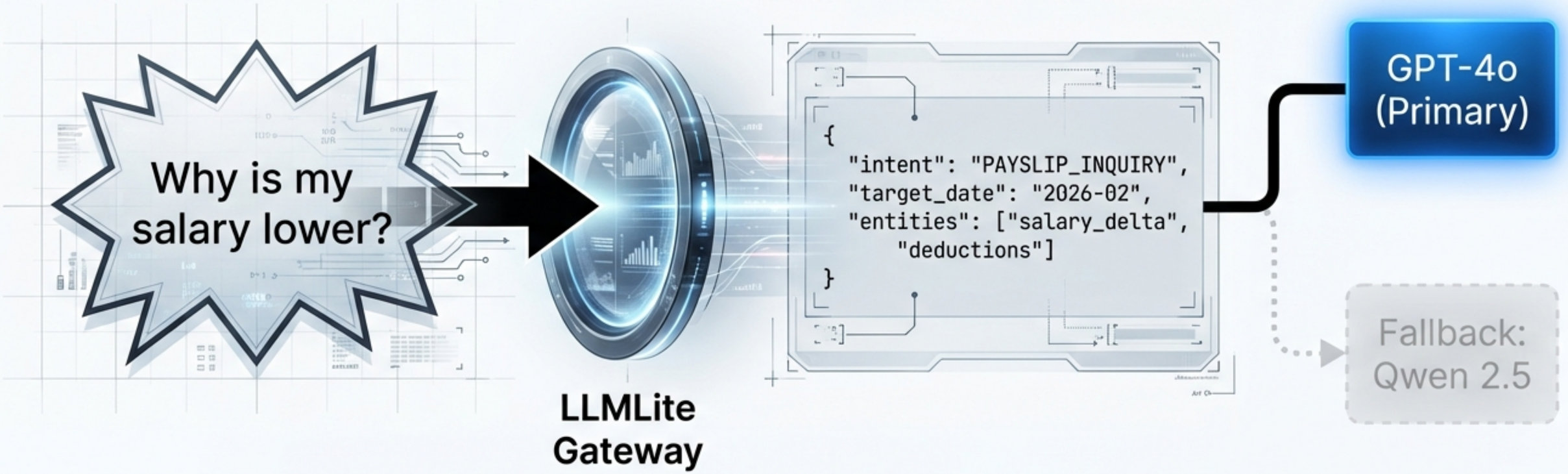


This seamless response requires three distinct types of intelligence working in harmony.

NotebookLM

Step 1: The Interpretation

Task Profile: INTENT_EXTRACTION

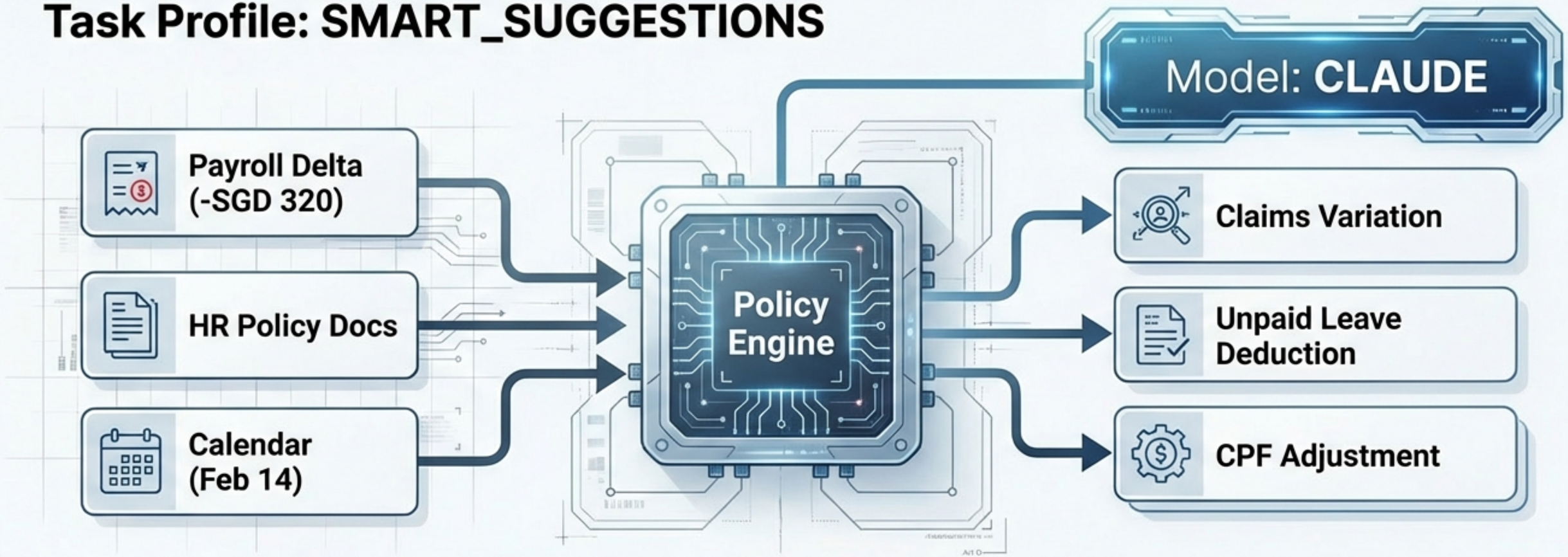


The system must parse messy user input into strict JSON context without being misled by conversational nuances. We route to GPT-4o for high precision and zero hallucination.

NotebookLM

Step 2: The Brain (Smart Suggestions)

Task Profile: SMART_SUGGESTIONS



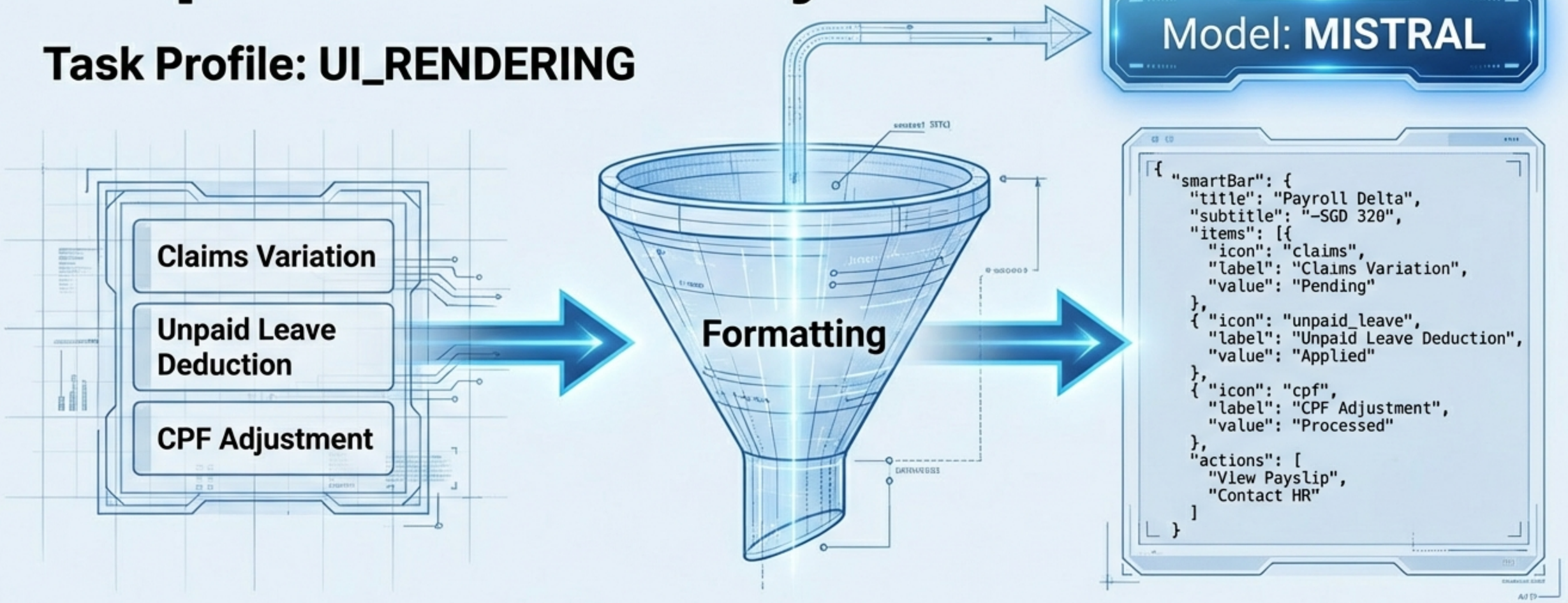
This is an HR-Sensitive Flow. It requires deep reasoning quality and strict adherence to policy. We route to Claude for safety first.

NotebookLM

Step 3: The Delivery

Task Profile: UI_RENDERING

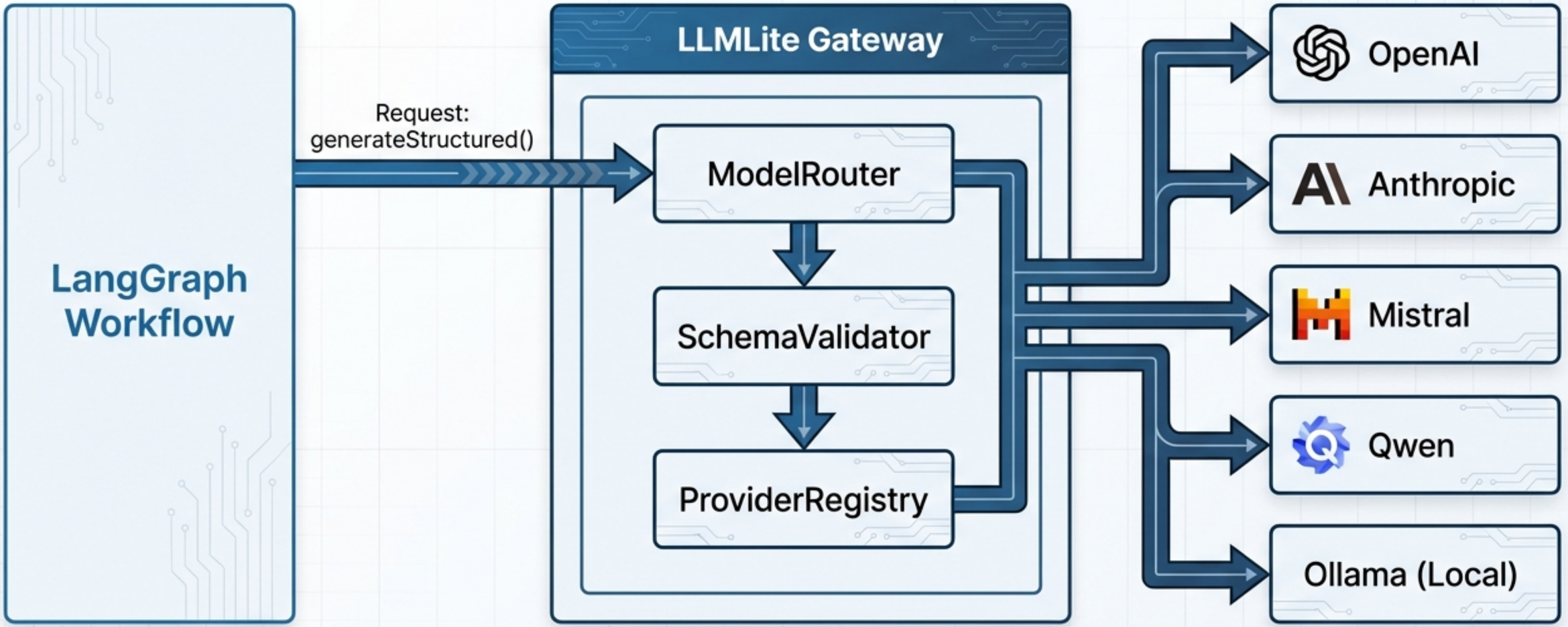
Model: **MISTRAL**



The reasoning is already done. This task is pure formatting—converting insights into the rigid JSON schema required by the front-end. We route to Mistral for speed and cost-efficiency.

NotebookLM

Architecture: The Universal Intelligence Gateway



The Application never speaks to the Model directly. It speaks to the Gateway.

NotebookLM

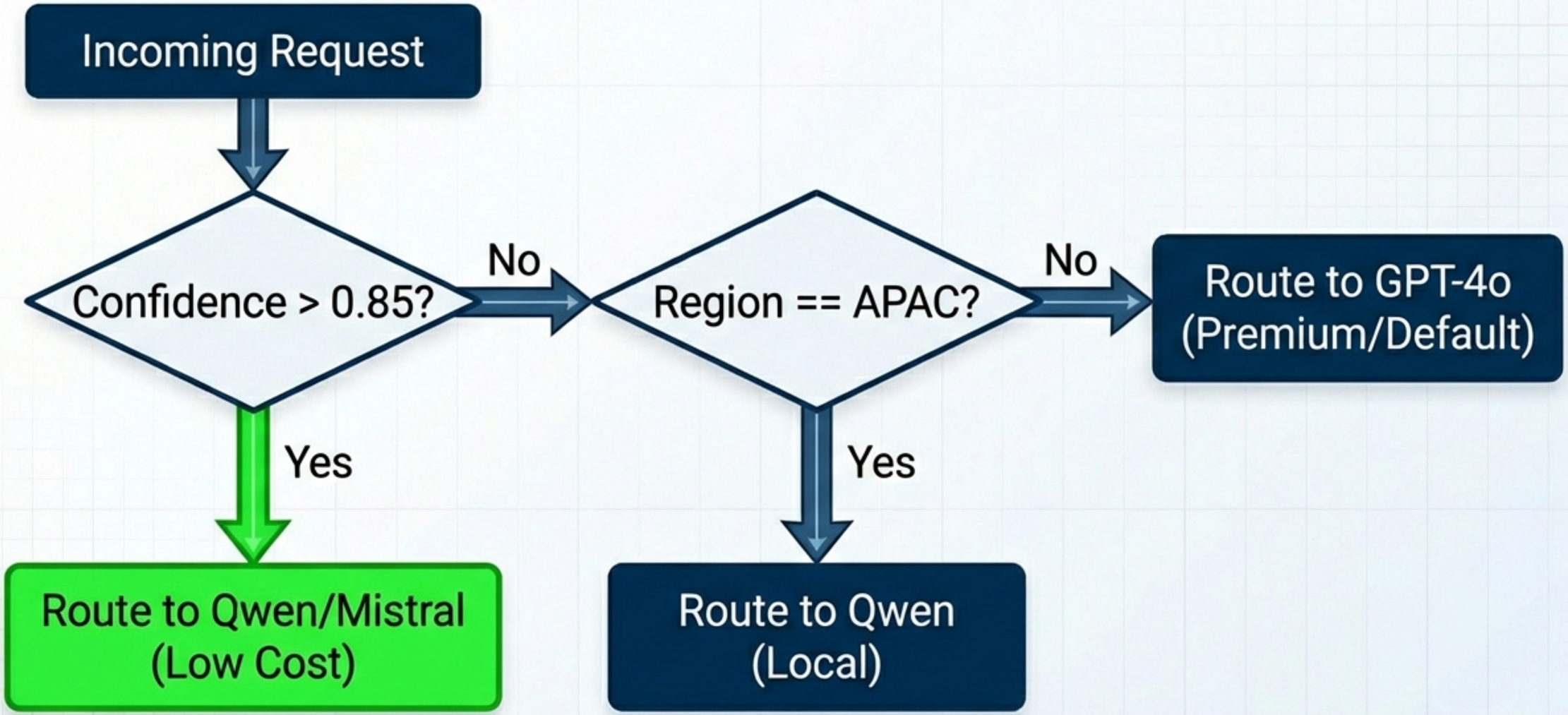
The Capability Matrix

Task Profile	Routing Logic	Best Model
INTENT_EXTRACTION	Precision Required	GPT-4o
SMART_SUGGESTIONS	Safety & Policy	Claude
UI_RENDERING	Speed & Cost	Mistral
MULTILINGUAL (APAC)	Localization	Qwen 2.5
HIGH SENSITIVITY	Data Sovereignty	Ollama (Local)

Routing is based on the Task Profile, not the provider name.

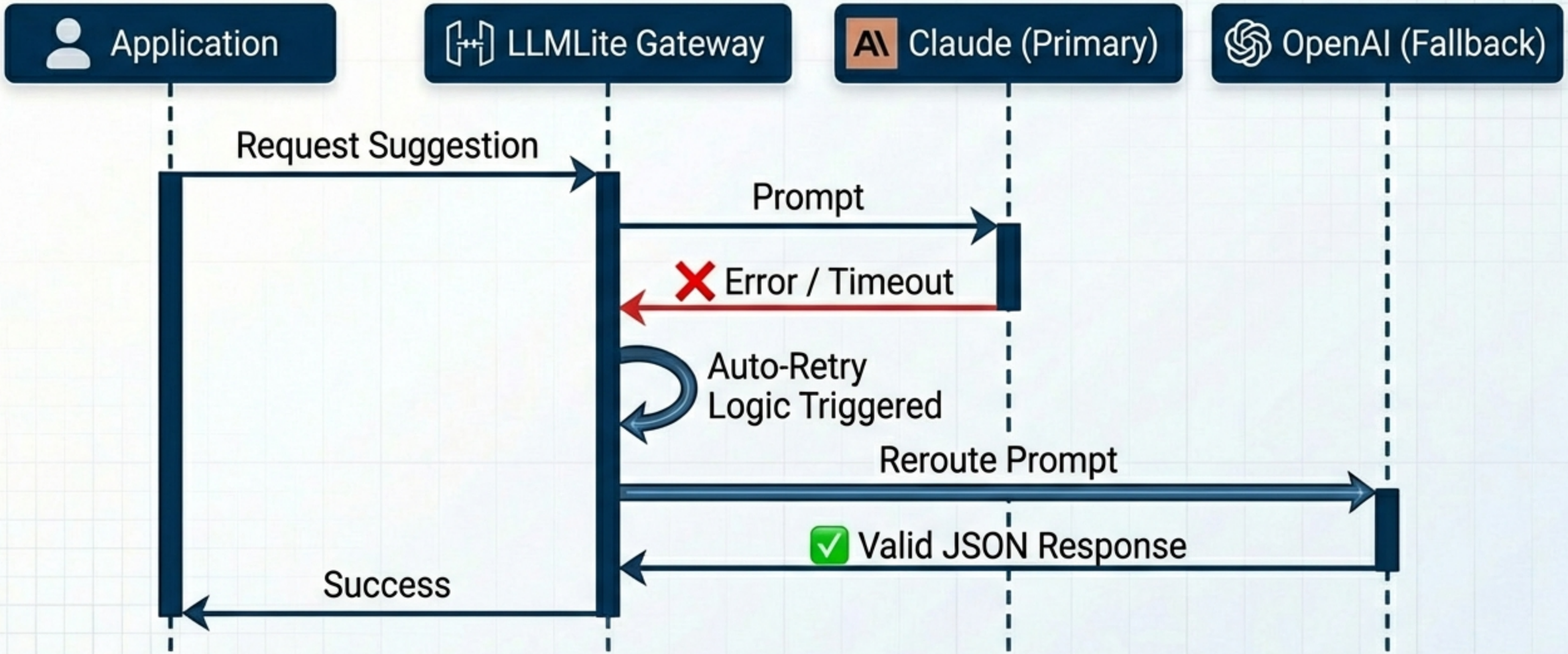
NotebookLM

Dynamic & Cost-Aware Routing



NotebookLM

Resilience: The Self-Healing Request

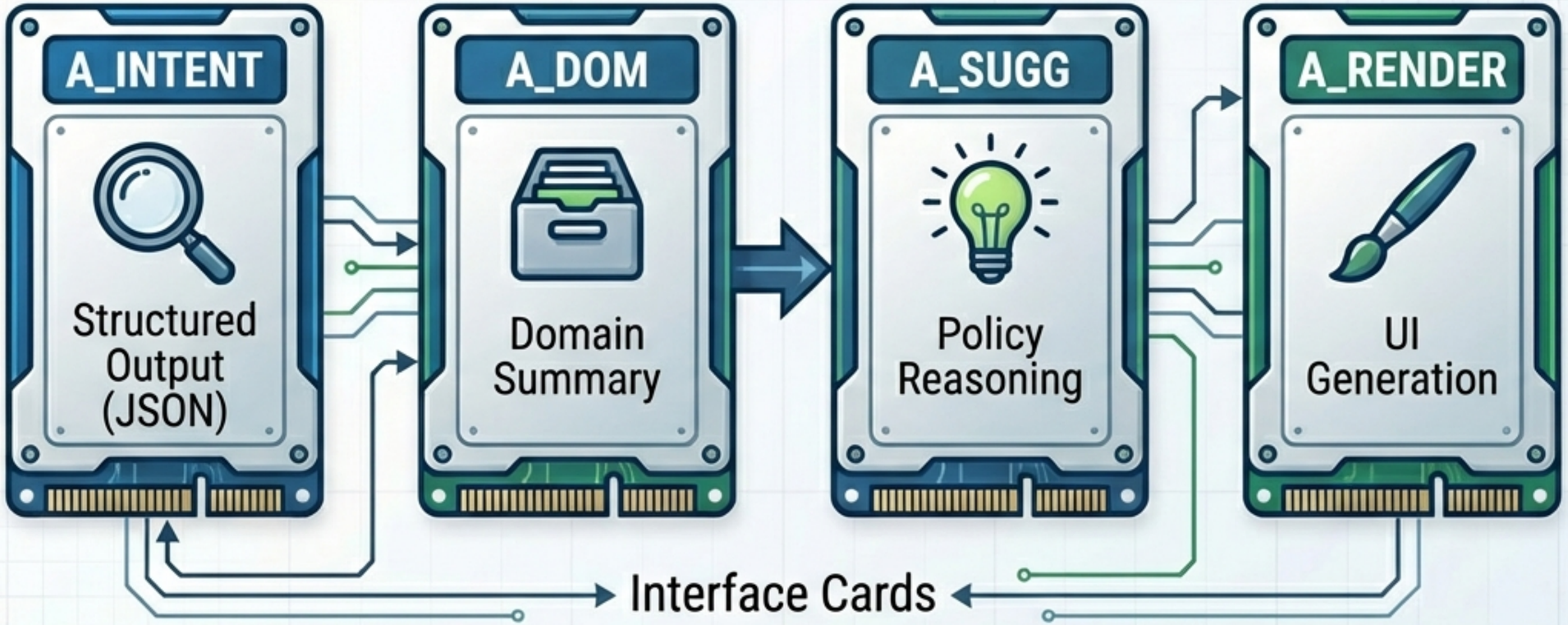


The user never sees the error. The system heals in real-time.

NotebookLM

The Adapter Strategy

Standardized Profiles



We route based on abstract ‘Task Profiles’, effectively decoupling the application logic from the underlying model providers.

NotebookLM

Production Guardrails

JSON Schema Validation

PII Masking

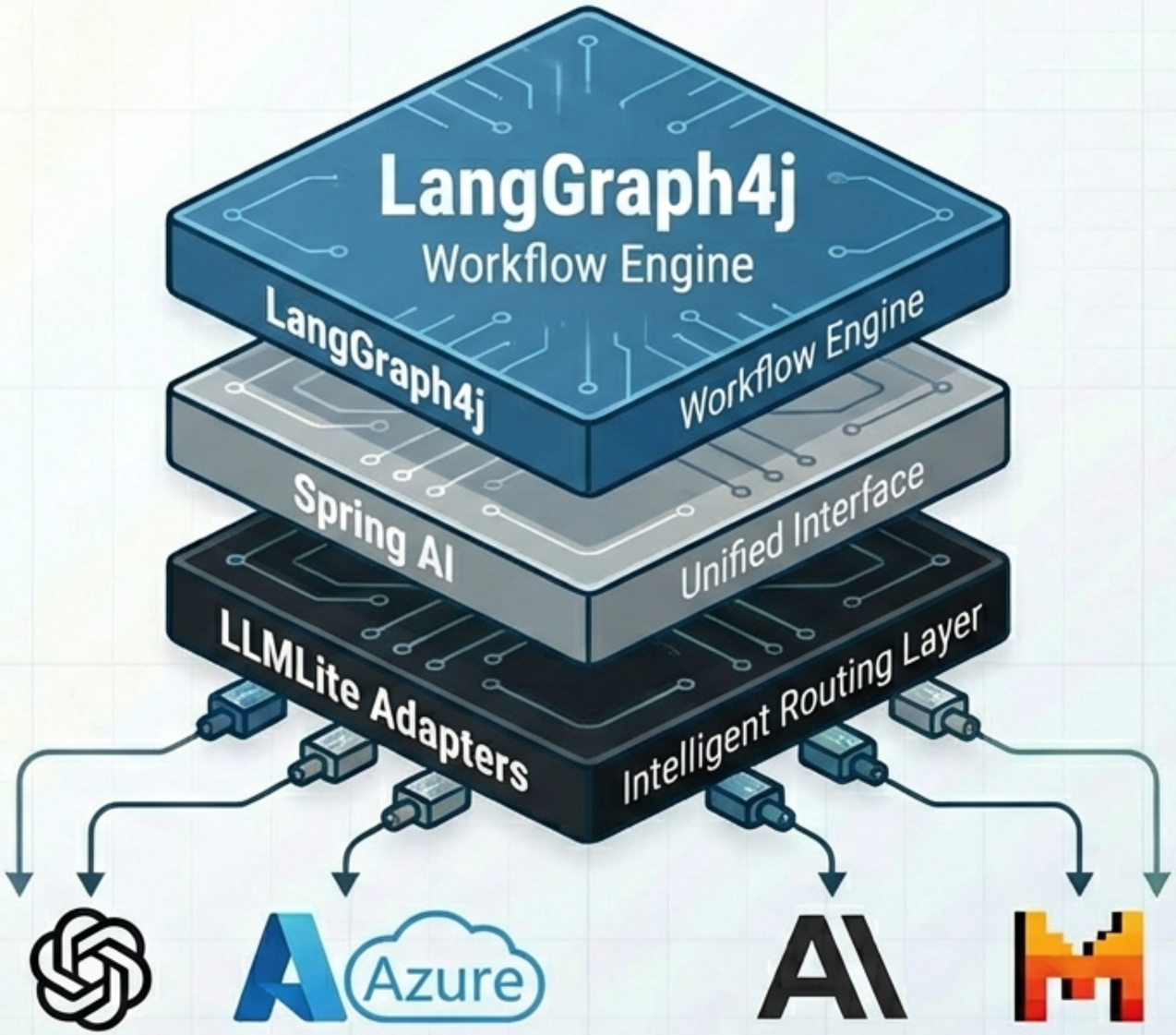
No-New-Facts Rule

Circuit Breakers

Ensuring enterprise readiness through strict validation, privacy enforcement, and cost controls.

NotebookLM

Implementation Stack



NotebookLM

The Enterprise Advantage

Vendor Agnosticism

Replace OpenAI tomorrow without rewriting code. Avoid lock-in.

Cost Control

Use cheap models for 80% of tasks; premium only for the the complex 20%.

Governance

Centralized policy enforcement, audit trails, and data sovereignty.

NotebookLM

The Universal Switchboard

“LLMLite dynamically assigns the right model to the right cognitive task—keeping governance, cost, and flexibility under control.”

NotebookLM