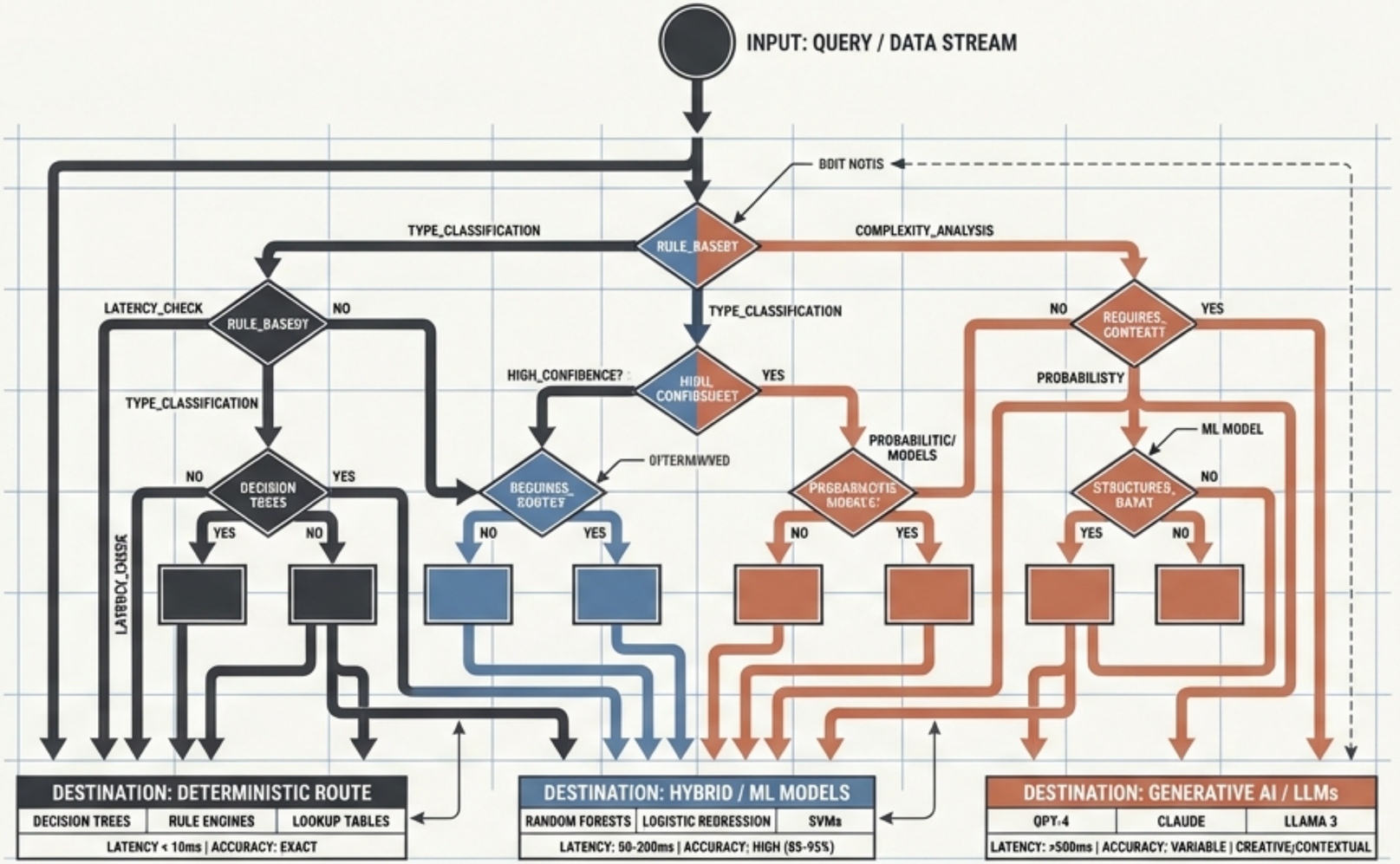


The AI Routing Matrix: Decision Models vs. LLMs

A practical guide to the new landscape of AI models, how they work, and exactly when to use each in your software architecture.

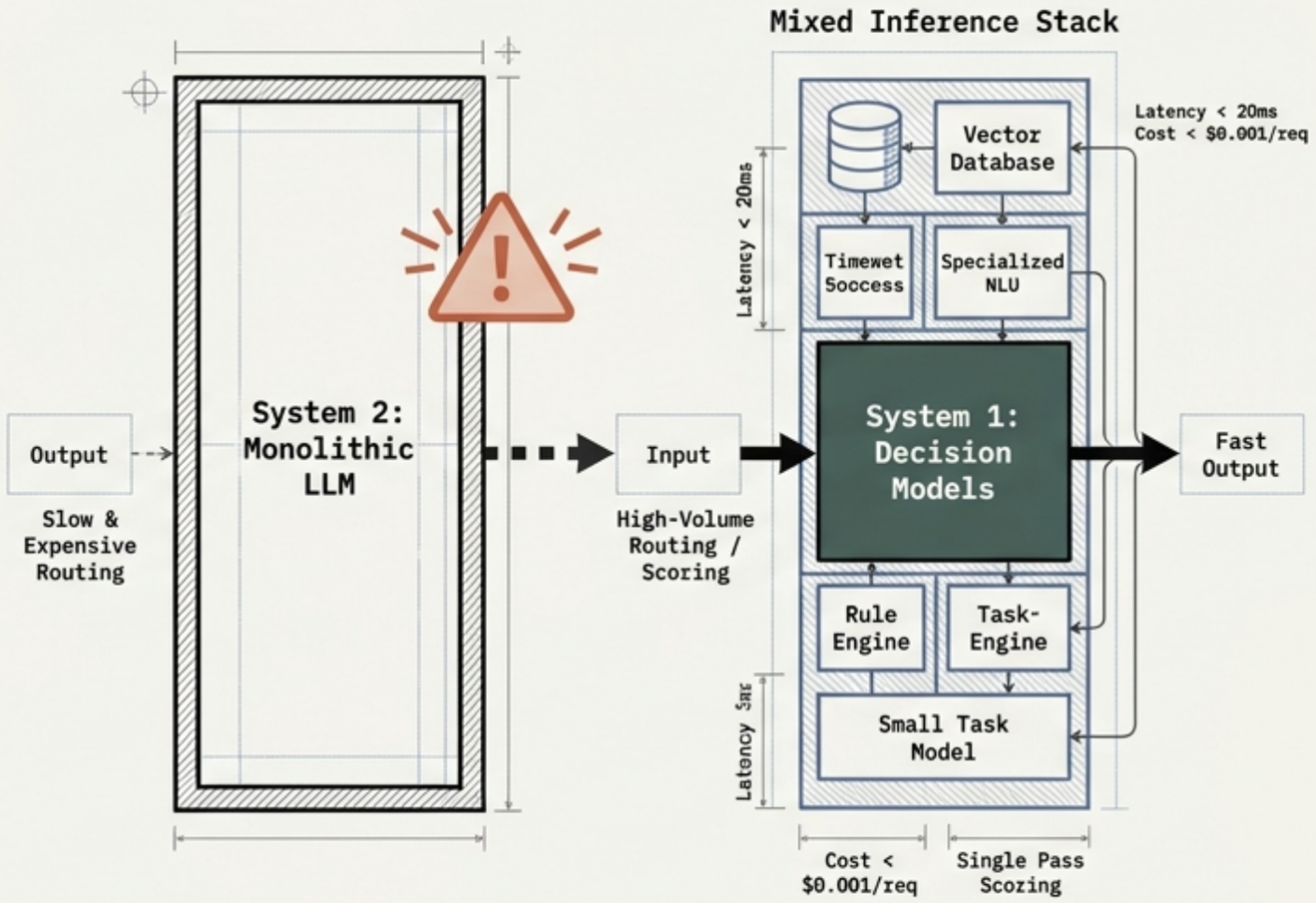
Gemini Notebook

The Shift to System 1 AI Architecture

The era of using one massive LLM for every task is over.

Large Language Models (System 2) are incredible for open-ended reasoning, but they are too slow and expensive for high-volume routing.

A new category of System 1 AI—Decision Models—drastically cuts costs and latency by scoring fixed answers in a single pass instead of generating text.

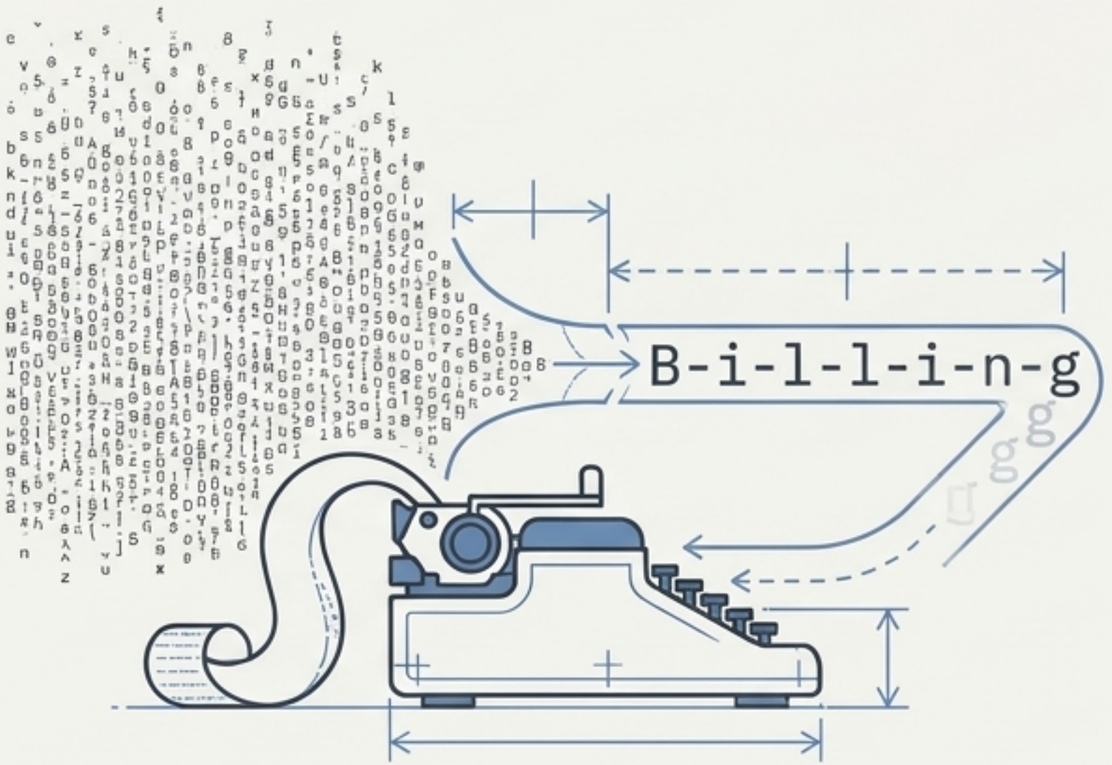


The key to modern AI architecture is mixed inference: matching the geometry of the task to the mechanical structure of the model.

Gemini Notebook

The Mechanical Flaw: Using LLMs for Judgments

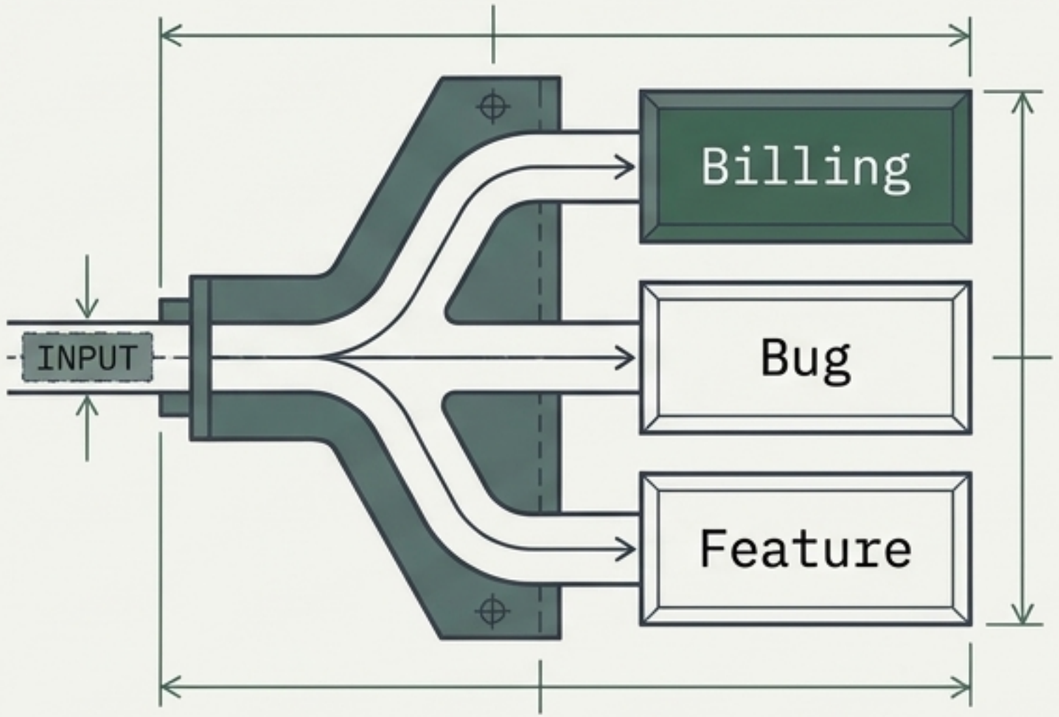
Generative LLM



Token by Token

Every LLM call generates its answer one token at a time from its entire vocabulary—even for a simple yes/no classification.

Decision Model

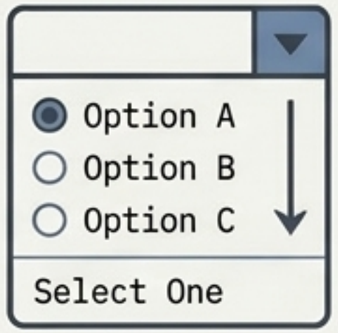


Single Forward Pass

Decision models fix the answer set before the forward pass, evaluating all declared options simultaneously. Hallucinating a format is structurally impossible.

Gemini Notebook

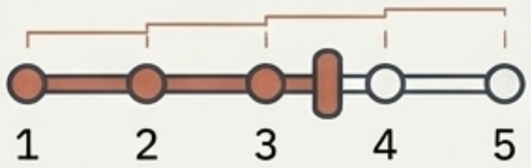
The Three Output Shapes of a Decision Model



Choice

Select one of your predefined options (e.g., routing a ticket to 'Biling', 'IT', or 'HR').

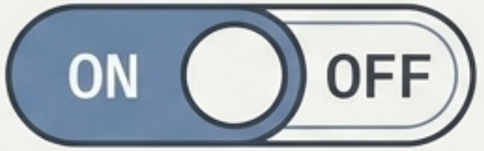
Returns the choice, a probability per option, and a confidence score.



Score

Rate an input against ordered levels in a rubric (e.g., ticket severity from 1 to 5).

Returns a probability-weighted average of the level indices.



Predicate

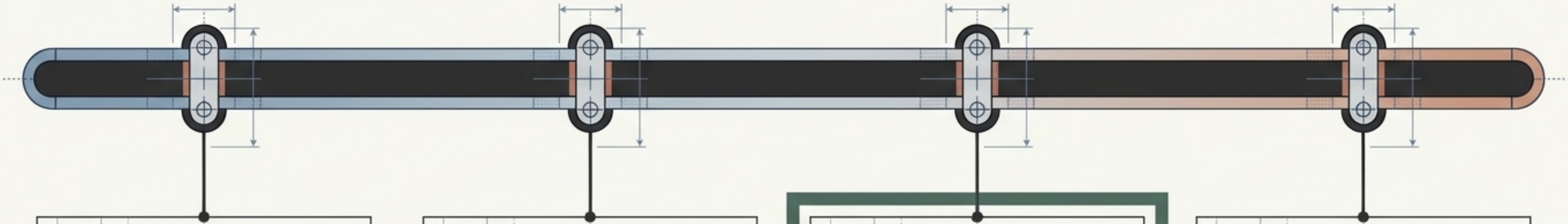
Estimate the probability that a specific condition is true (e.g., 'Is this a refund request?').

Returns a float from 0 to 1.

Gemini Notebook

The Enterprise AI Model Spectrum

General / Generative — **Narrow / Predictive**



Frontier LLMs

GPT-6 Luna, Claude Fable 5.1

Maximum flexibility, highest cost, sequential text generation.

Small Language Models (SLMs)

Mistral 7B, Llama-3-8B

Faster generalists, runs on-premise, still generative.

Decision Models

Jev, Clef, Laya

Semantic understanding, single-pass scoring, strictly bounded typed outputs.

Pre-Trained Classifiers

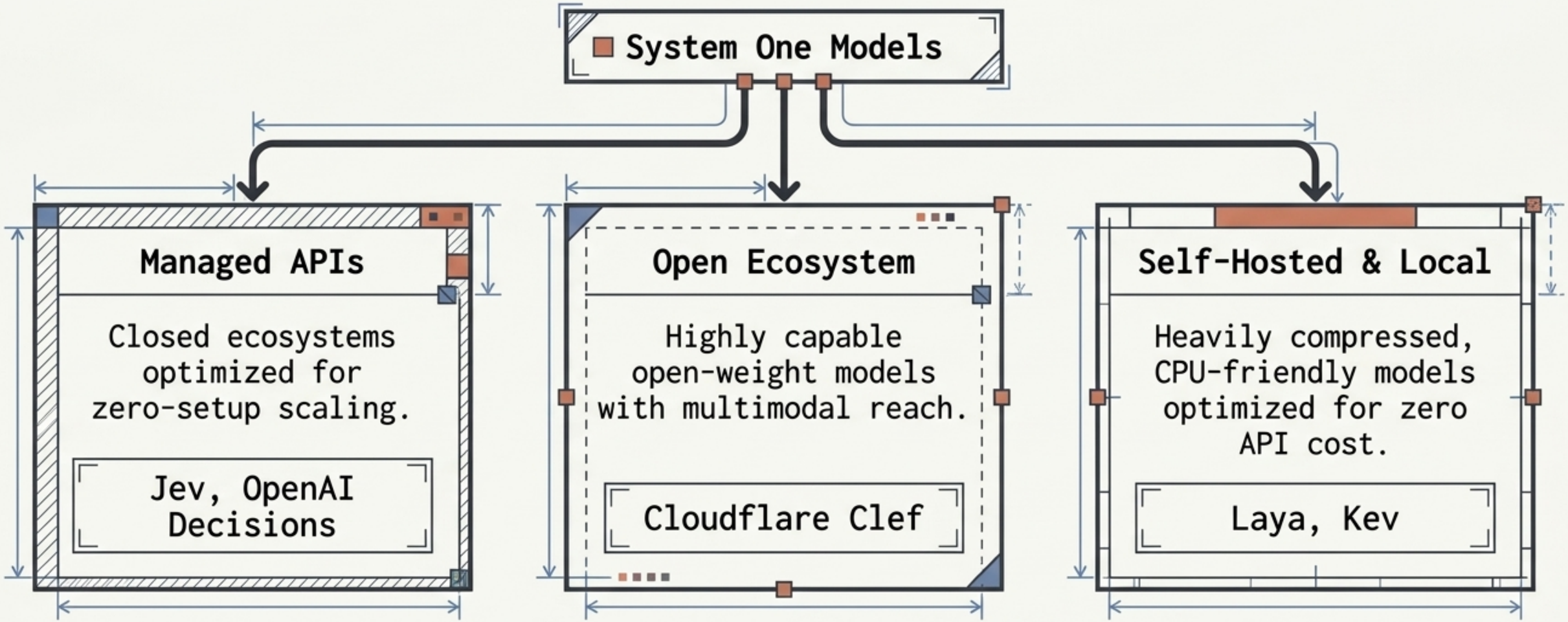
DeBERTa

Extreme low latency, rigid training data, singular fixed tasks.

Gemini Notebook

The New Model Landscape

The rapid emergence of System One APIs and open-weight models in late 2026 has fragmented the decision space into three distinct tiers:



Gemini Notebook

Managed APIs: TypeSafe Jev & OpenAI Decisions



Jev (TypeSafe)

Status:	The Pioneer.
Mechanics:	Bounded text decisions via Reinforcement Learning for Calibrated Decisions (RLCD).
Economics:	$0.042 per 1M input tokens. $0 output token cost.
Context:	32k window limit. Text-only.



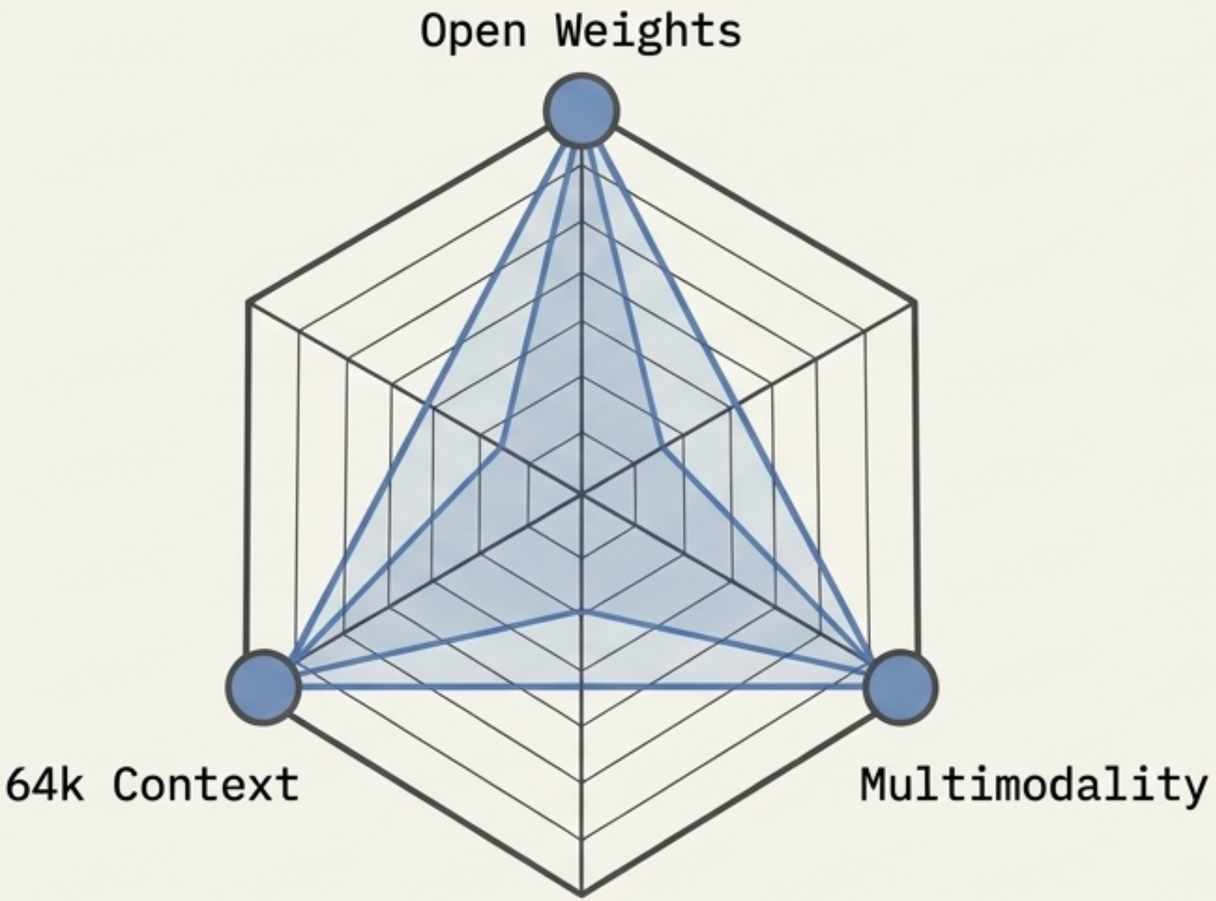
OpenAI Decisions API (Luna)

Status:	The Ecosystem Giant.
Mechanics:	Focuses GPT-6 Luna's intelligence strictly on predefined answers.
Performance:	10x faster than the standard Responses API.
Capabilities:	Supports both text and visual (image) state. $0.10 per 1M input tokens.

Gemini Notebook

The Open Multimodal Edge: Cloudflare Clef

Released under Apache 2.0 and built on a Qwen backbone, the Clef family offers high-capability decision scoring without API lock-in.



Clef

The primary model for subtle questions and long states (64k context window).

Clef-flash

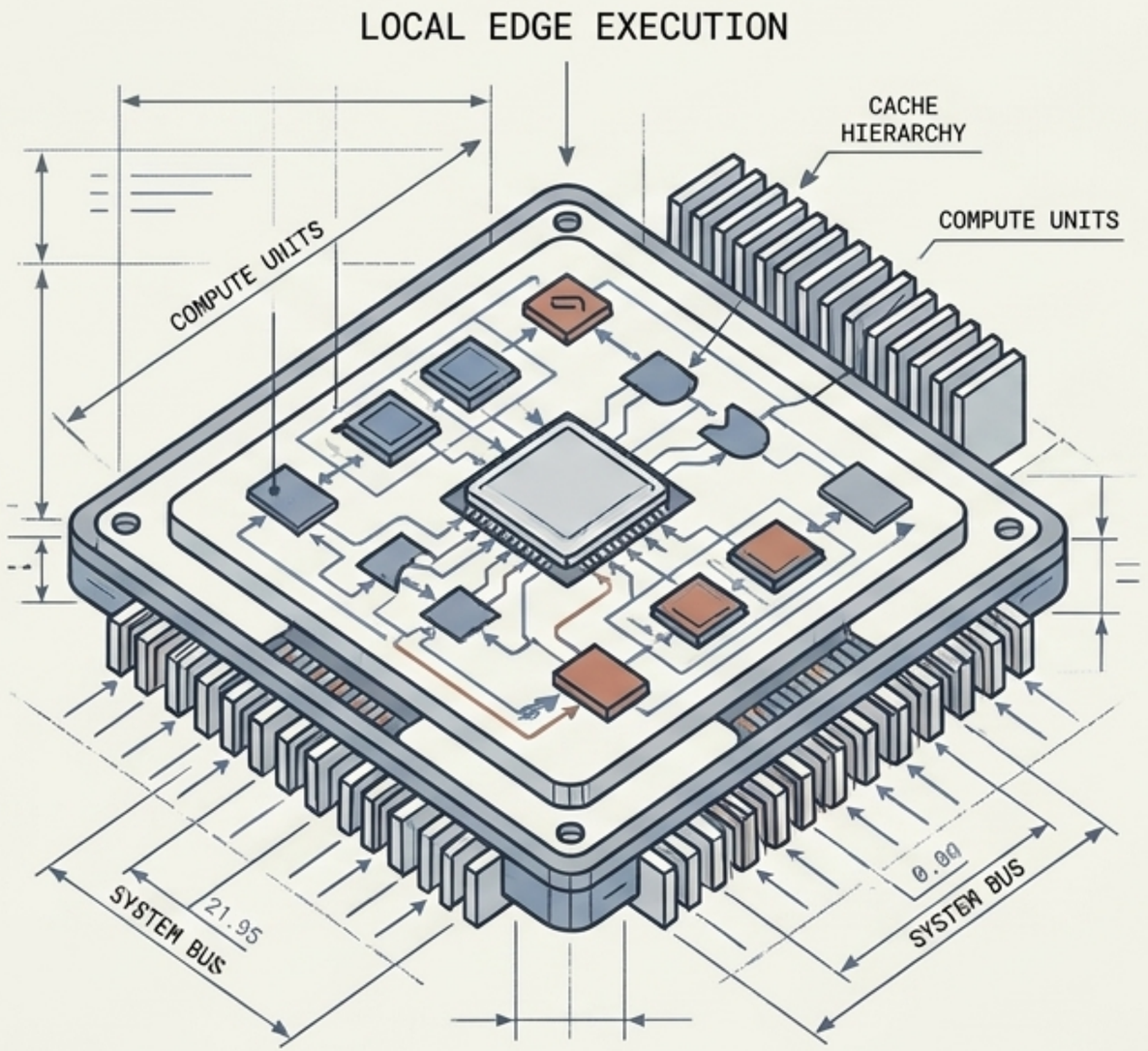
Optimized for extreme low latency on high-volume classification.

Clef-omni

Built on a Mixture-of-Experts architecture. Can ingest Audio (mp3/wav) and Video (mp4/webm) to score decisions on recordings.

Gemini Notebook

The Self-Hosted Innovators: Laya & Kev



Warning/Caution

Laya (Convai Innovations)

Specs: 421M parameters, runs on CPU, $0 token bill.

API

The Catch: Weak zero-shot performance (scores ~0.362 on benchmarks). Requires fine-tuning on a few thousand labeled examples to hit advertised 0.766 accuracy. Perfect for established, repetitive tasks.

Kev (TurinTech / Jared Palmer)

Specs: 4B/9B Qwen-based models compressed via Artemis GGUF (2.54 GB footprint).

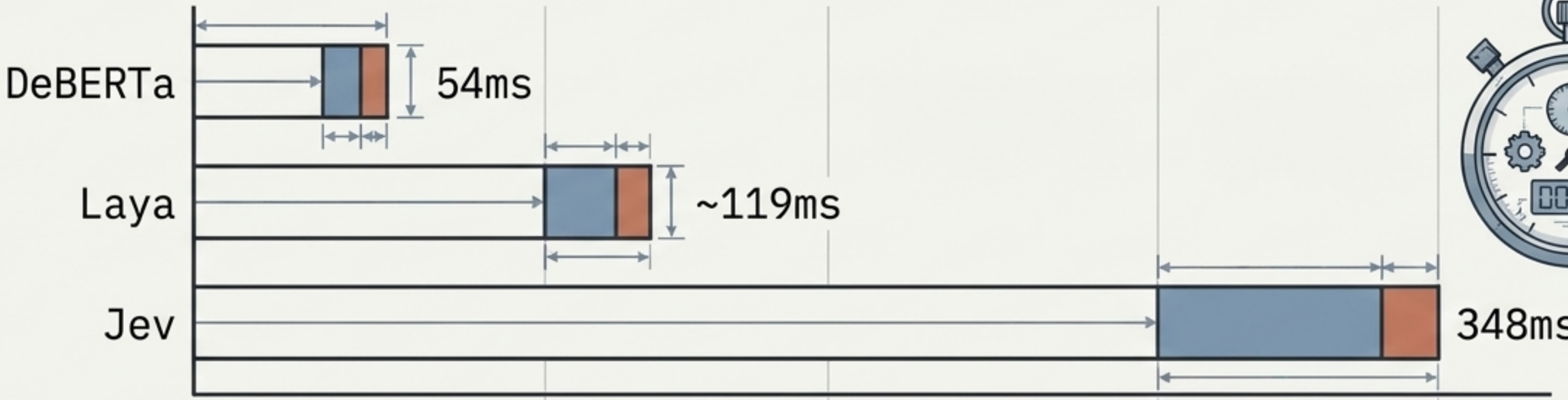
API

Advantage: Highly accurate locally. Jev-API compatible (/v1/systemone), allowing instant switching from managed APIs to local hardware.

Gemini Notebook

The Baseline Reality Check: Pre-Trained Classifiers

For established, static risks, bespoke CPU-scale classifiers remain undefeated.





The Data (Red Hat Benchmark)

1. On adversarial tasks like prompt-injection detection, Red Hat's deberta-v3-base-prompt-injection-v2 achieved 89.01% accuracy with a median latency of just 54.1ms.
2. By contrast, zero-shot decision models and LLM-as-a-judge approaches took 300ms+ for the same task.
3. Takeaway: Do not use a semantic decision model for a problem that already has a dedicated, pre-trained classifier.

Gemini Notebook

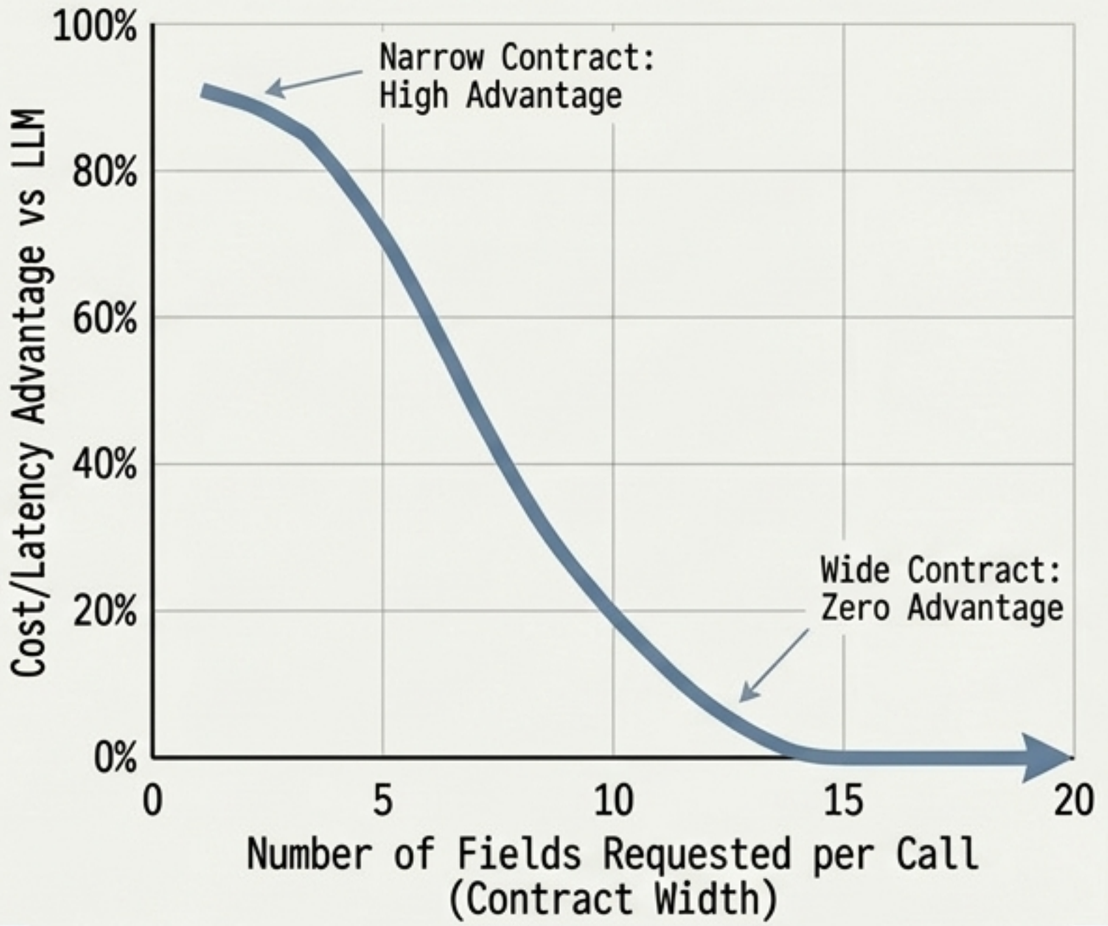
SYNTHESIS: Head-to-Head Comparison Matrix

Model	Paradigm	Typical Latency	Cost	Modality	Best Use Case
Jev	Managed API	Med Latency	$0.042/1M in	Text	High-volume routing
OpenAI Decisions	Managed API	Med Latency	$0.10/1M in	Text, Vision	Visual gating, agents
Clef Family	Open (Apache)	Variable	Free / Hosted	Text, Vision, Audio	No lock-in, Multimodal
Laya	Local CPU	Low Latency	$0 API	Text	Air-gapped, fine-tuned tasks
LLMs (Claude/GPT)	Managed API	High Latency	High Cost ($10+/1M)	Text, Vision	Generation, Reasoning
Pre-trained Classifiers	Local CPU	Extreme Low	$0 API	Text	Known static risks (e.g., Toxicity)

Gemini Notebook

Benchmark Insights: Where the Advantage Erodes

The Contract Width Threshold



Data Point 1: The UTS Benchmark

On narrow, 4-field contracts, Jev reduced API cost per correct decision by 59.7% to 80.9%. However, as the contract width (fields requested) grows, the advantage vanishes.

Data Point 2: The LLM-as-a-Judge Reality

On complex, adversarial guardrails (catching subtle manipulation), a well-tuned Qwen3.6-35B LLM-as-a-judge (312ms) actually beat Jev (348ms) on speed and accuracy.

Takeaway: The advantage is task-shaped. Decision models dominate narrow judgments but lose their edge on wide, multi-field extractions.

Gemini Notebook

Framework: When to Use a Decision Model vs. an LLM



The Task Geometry



Use a Decision Model When...

- The answer space is strictly bounded (Choice, Score, Yes/No).
- Call volume is exceptionally high (thousands of times a day).
- Latency is on the critical path (blocking a user-facing response).
- You need a calibrated probability threshold to automate an action.

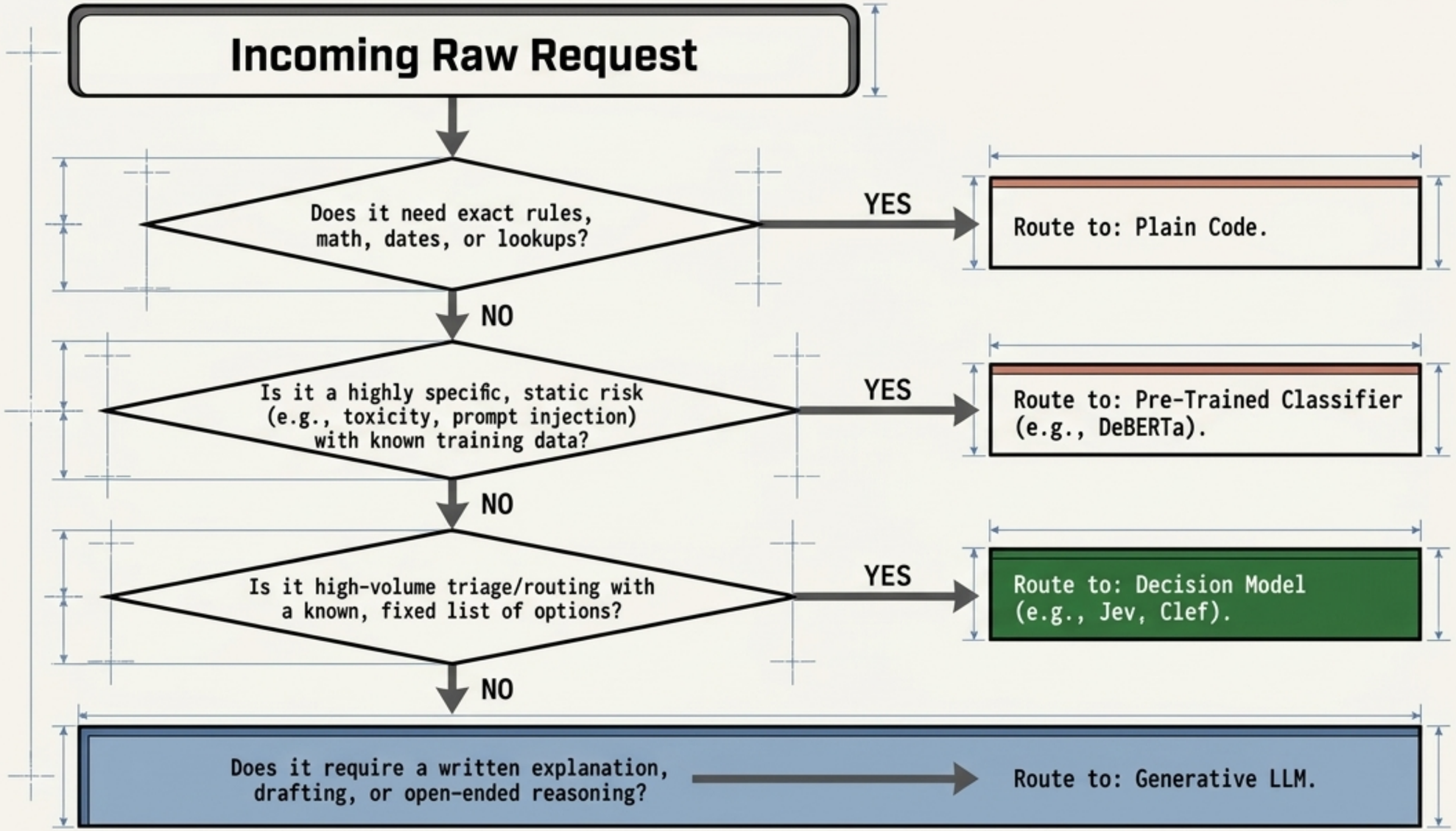
Use an LLM When...

- The output requires a written explanation, plan, or drafted text.
- The workflow requires a human-readable audit trace of the reasoning.
- The task involves open-ended problem solving or multi-step logic.
- The context width exceeds 64k tokens.

Gemini Notebook

The Workload Routing Decision Tree

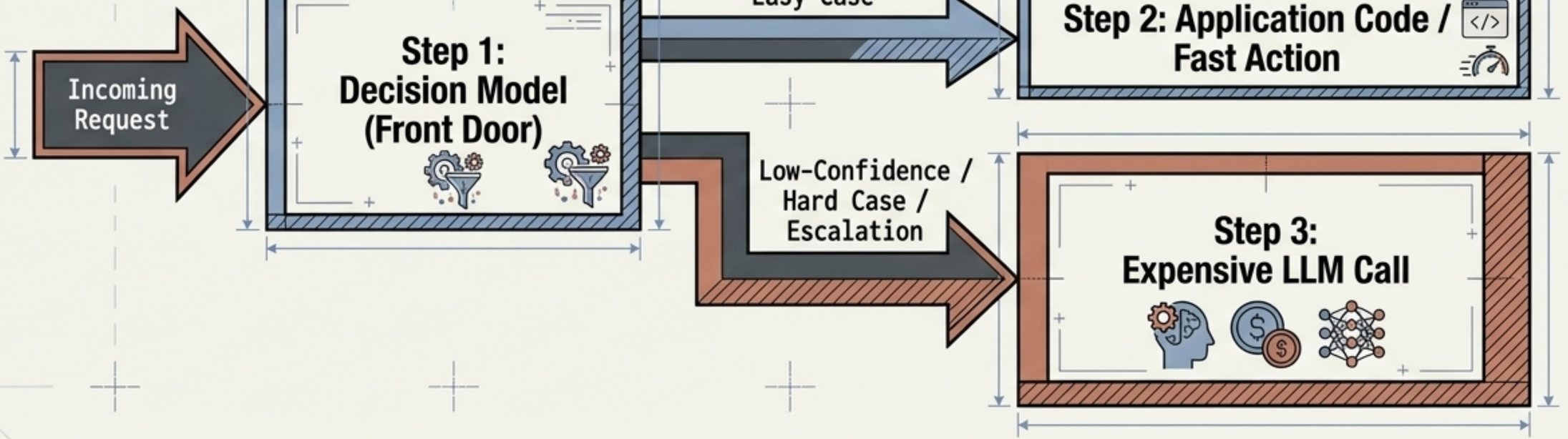
Walk the questions from top to bottom.



Gemini Notebook

Architectural Blueprint: Mixed Inference

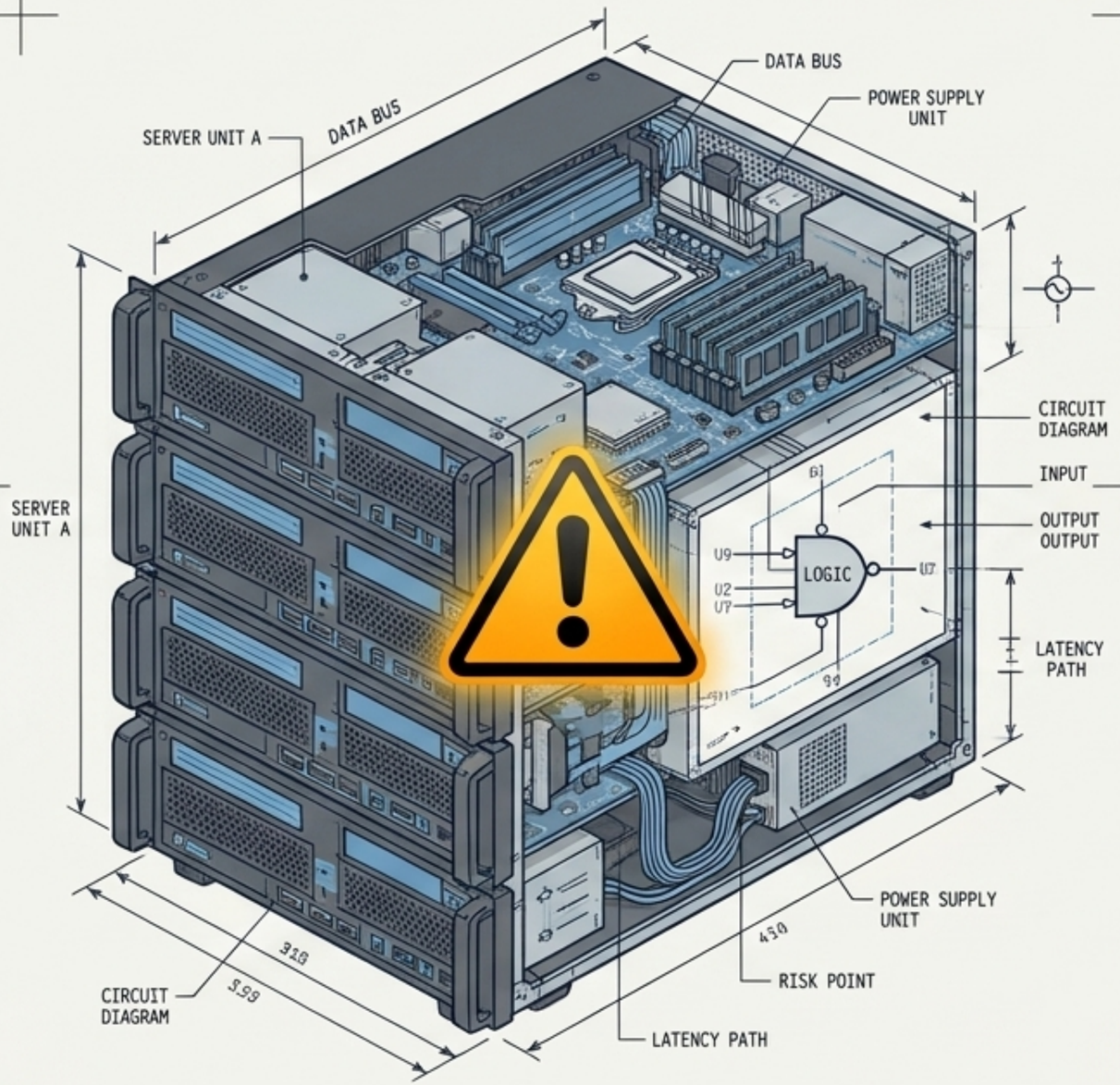
The Triage-and-Escalate Pattern



The most effective enterprise pattern is mixed inference. A fast decision model (like Clef-flash or Jev) sits at the front door to score, filter, and handle the bulk of standard requests. This preserves expensive, high-latency LLM calls strictly for edge cases, reasoning, and complex escalations.

Gemini Notebook

The Shadow AI Risk: Governance in the Decision Layer



Because decision models are incredibly cheap and fast (sub-100ms), engineers can easily wire them into critical paths in a weekend, completely undocumented.

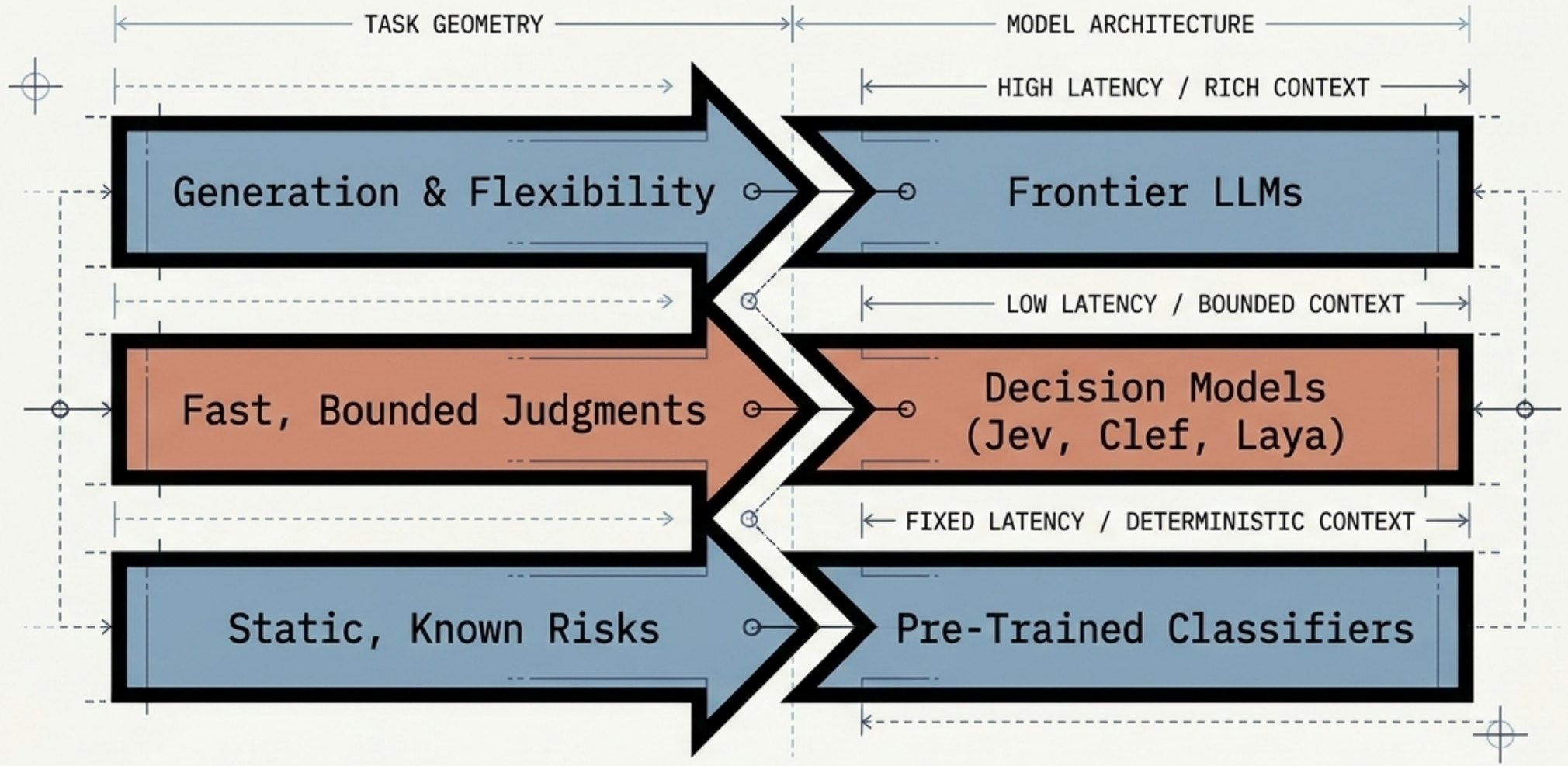
The Security Rule:

The model handles fuzzy meaning; the system retains authority. A model should never grant permissions or bypass authorization.

Always maintain deterministic rules (allow-lists, scopes, role checks) alongside model probabilities.

Gemini Notebook

Synthesis: Match the Model to the Task Geometry



The era of the monolith is over. Optimize your AI architecture not by finding the smartest single model, but by matching the mechanical structure of the model to the exact geometry of your task.

Gemini Notebook