Optimizing AI Economics through Intelligent Routing

A technical breakdown of the Cost-Control Smart Model Router—reducing API costs by up to 99.6% without sacrificing response quality.

Python 3.13 Compatible

FastAPI + Streamlit Architecture

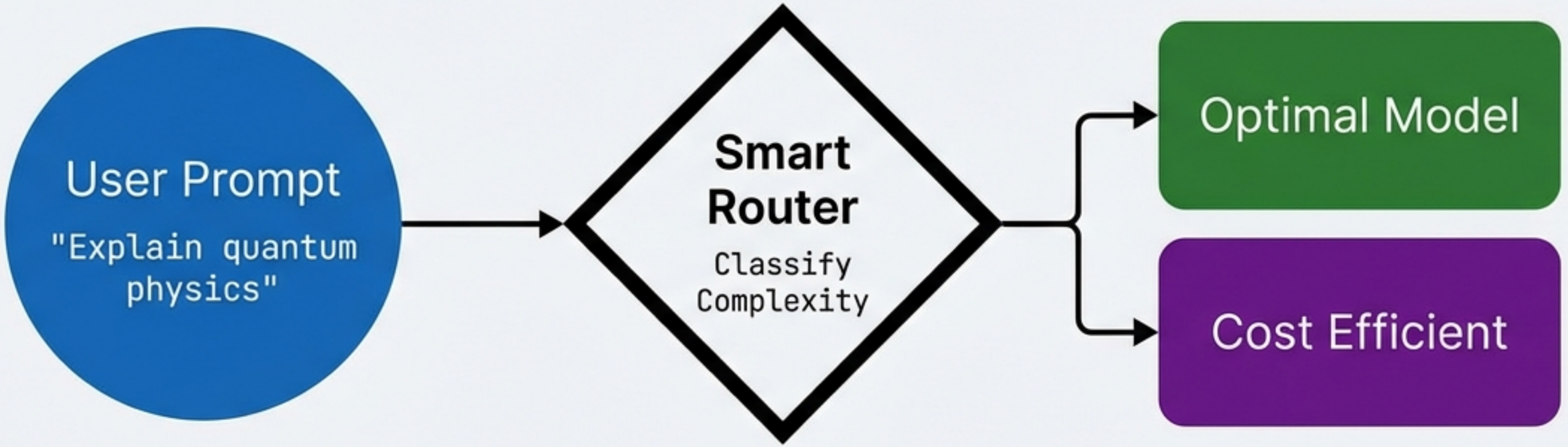
MIT License

Stop Treating Every User Prompt Like a PhD Thesis

Most applications default to a “One-Size-Fits-All” strategy, utilizing expensive models like GPT-4o for trivial tasks. This inefficiency scales linearly with user growth, resulting in massive waste on simple queries like greetings or factual lookups.

Scenario based on 1,000 requests/day at average $0.03/request.

The Solution: A Pre-Inference Decision Layer



Core Value Proposition: Reduce API costs by up to 70-99% while maintaining response quality.

A Full-Stack System for Intelligent Traffic Control

A Python-based routing system integrating backend logic with frontend analytics.

Smart Routing

Automatically classifies prompts and routes to optimal models based on difficulty.

Cost Optimization

Dynamic selection of cheaper models for simple tasks to minimize burn rate.

Multi-Provider Support

Native integration for OpenAI (GPT-4o), Google (Gemini), and local models.

Real-time Analytics

Track savings, latency, and token usage via a built-in Streamlit dashboard.

Auto-Discovery

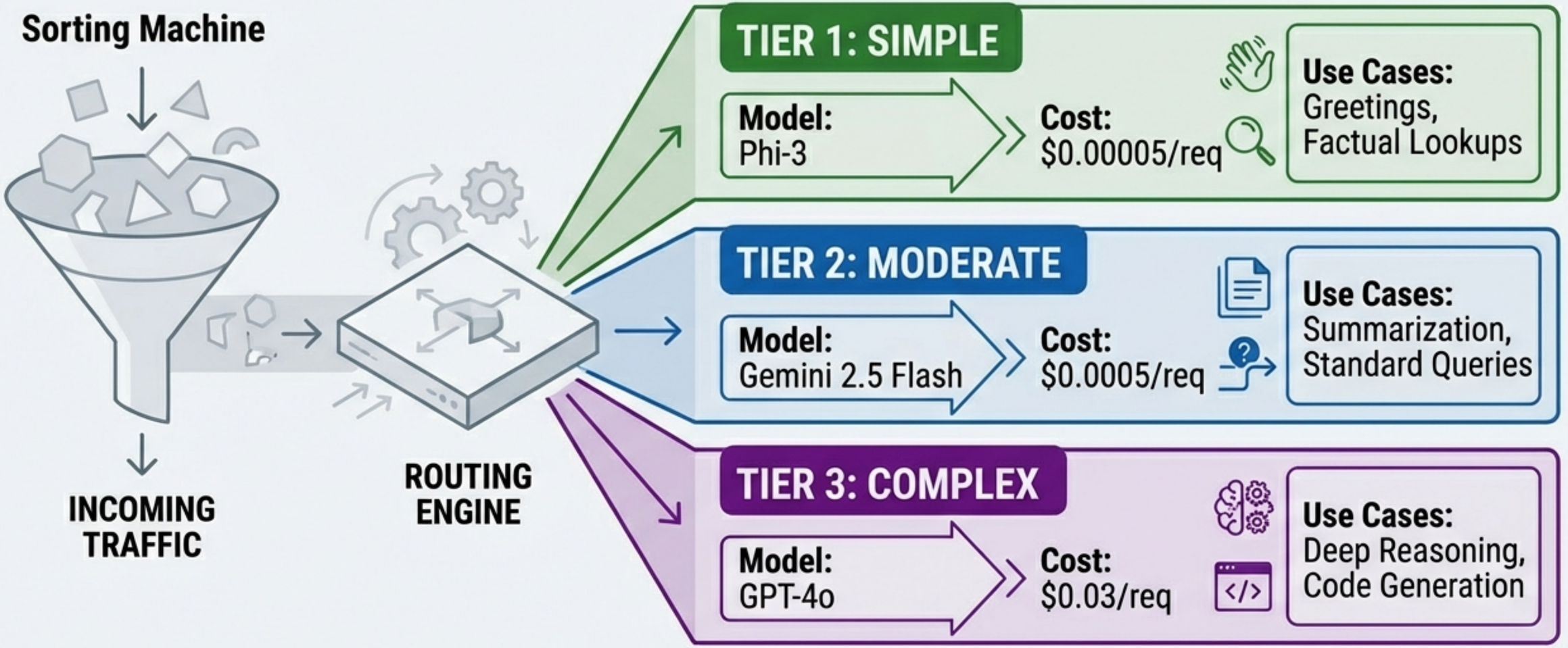
System automatically detects and configures available API keys and models.

Secure & Scalable

Built for enterprise deployment with secure key management and scalable architecture.

The Tiered Routing Logic

The router sorts incoming traffic into three distinct buckets based on complexity analysis.



Flexible Classification Engines

Administrators can toggle the classification mechanism via the environment configuration.

Rule-Based Mode

Mechanism: Keyword matching and heuristics.
Benefit: Zero latency, zero cost.

LLM-Based Mode

Mechanism: Lightweight model semantic analysis.
Benefit: High nuance and accuracy for ambiguity.

.env

```
# Classifier Configuration
CLASSIFIER_TYPE=rules
# Options: 'rules' or 'llm'
```

The Economics of Smart Routing

By routing 90% of traffic to cheaper models, costs collapse while utility remains constant.

WITHOUT ROUTER

$900.00 / month

100% Traffic to GPT-4o

WITH SMART ROUTER

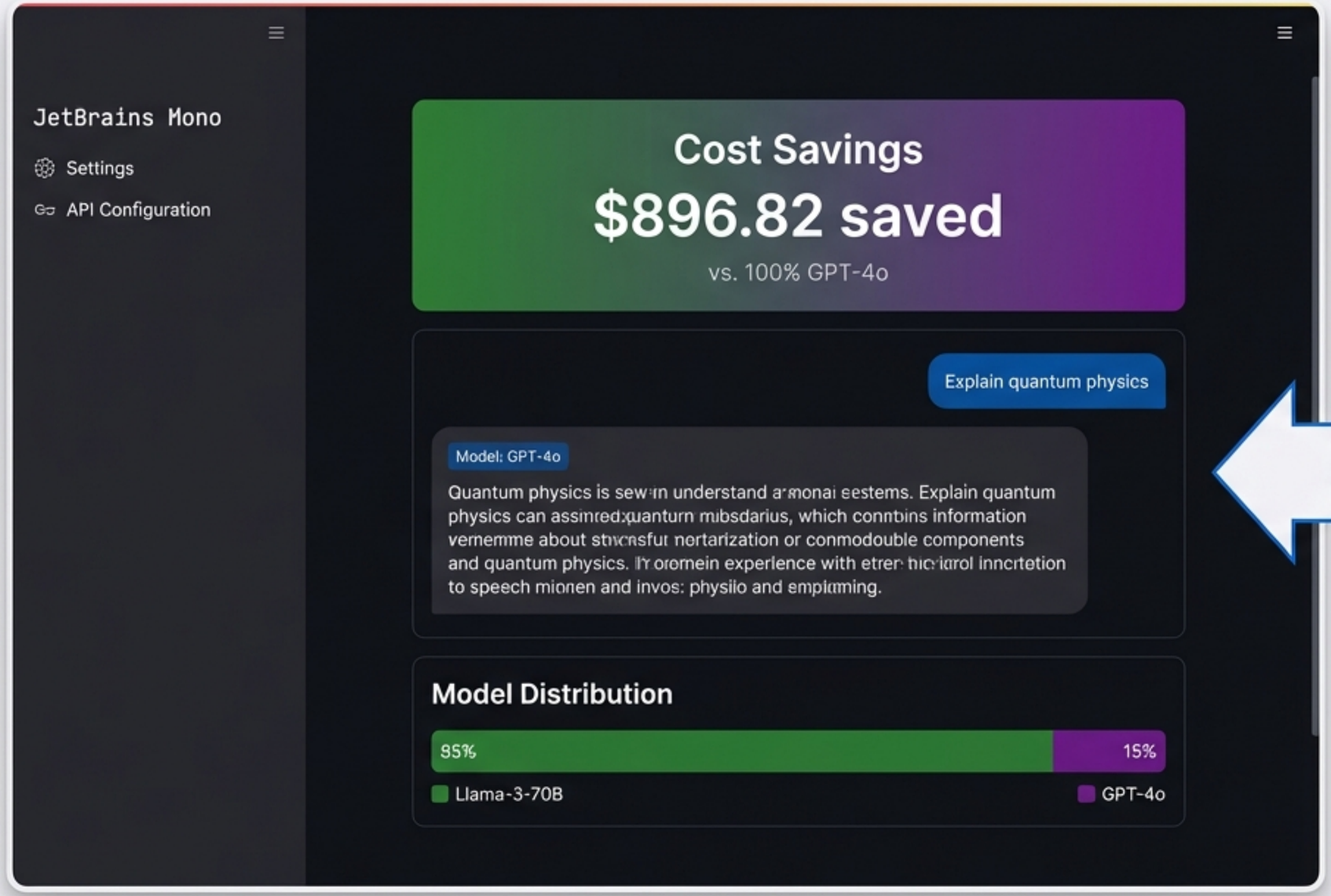
$3.18 / month

Optimization Breakdown:

- 600 Simple Requests (Phi-3): $0.03
- 300 Moderate Requests (Gemini): $0.15
- 100 Complex Requests (GPT-4o): $3.00

TOTAL SAVINGS: 99.6%

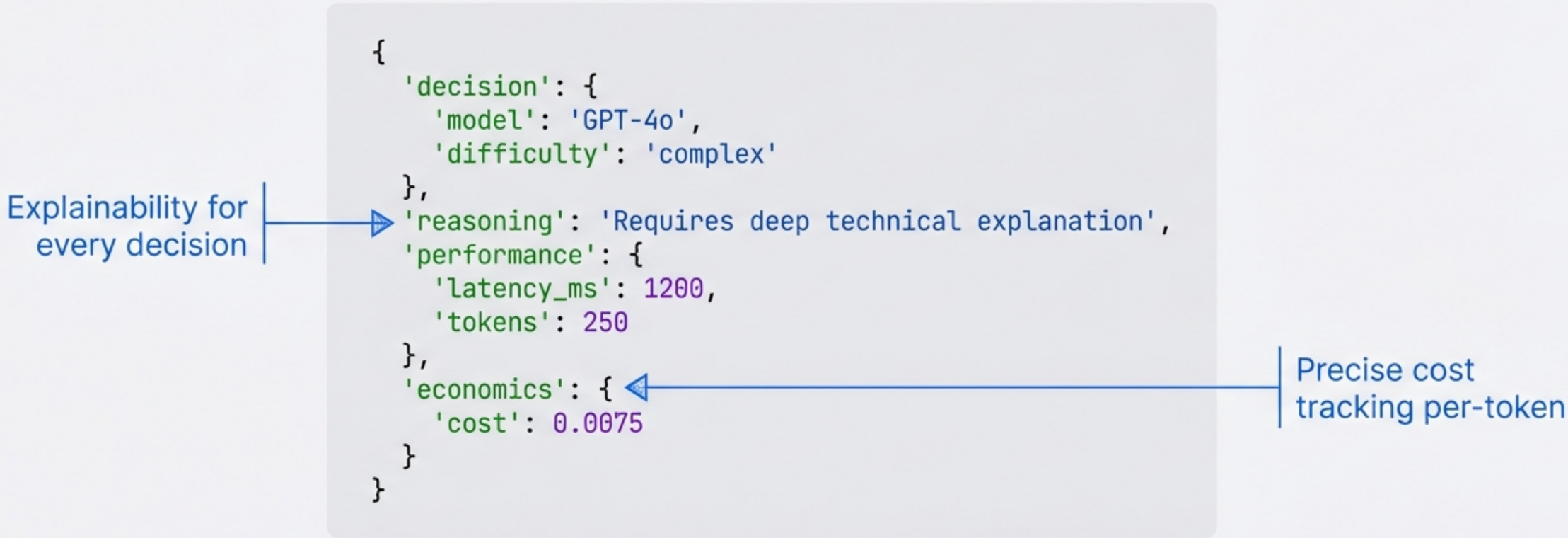
Real-Time Visibility via Streamlit



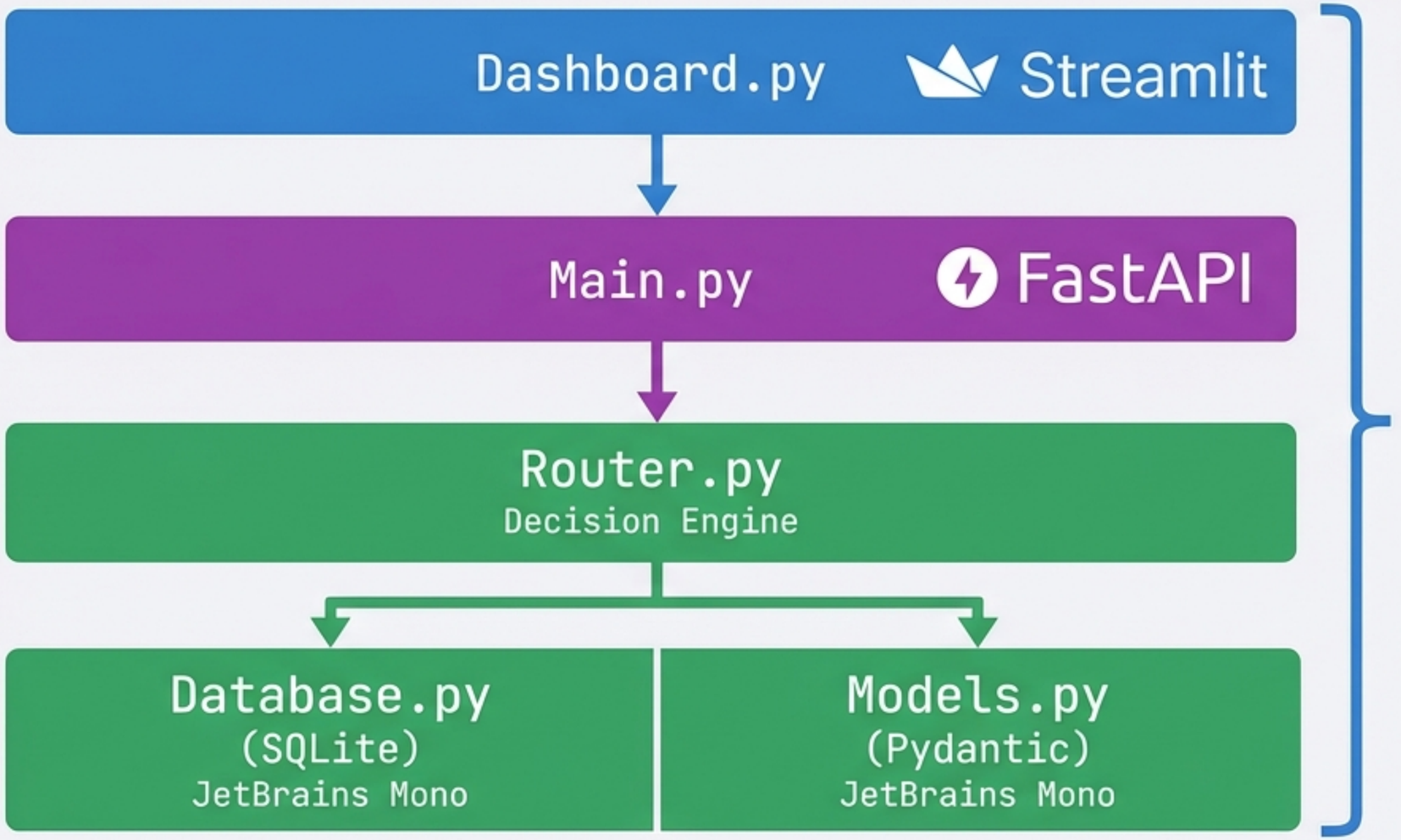
Engineering Editorial

Granular Analytics & Transparency

The system provides detailed metadata for every request, ensuring routing logic is fully auditable.



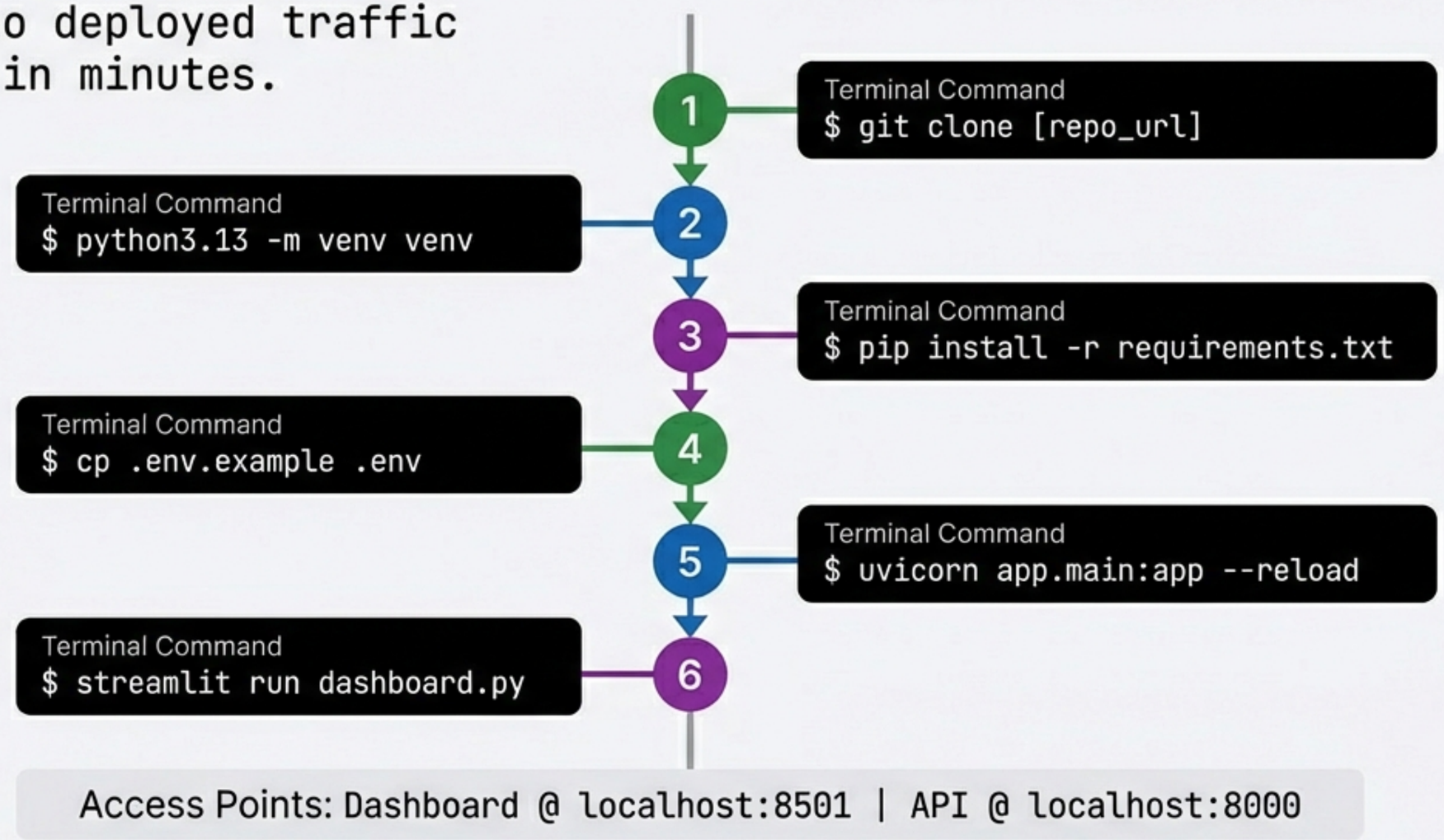
Built on a Modern Python Stack



Engineering Editorial

Deployment in Six Steps

From zero to deployed traffic controller in minutes.



Access Points: Dashboard @ localhost:8501 | API @ localhost:8000

Engineering Editorial

Designed for Extensibility

Adding new providers (e.g., Claude) is achieved through clean class inheritance.

1. Define Client (app/llm/providers.py)

```python
class ClaudeClient(LLMClient):
    async def generate(self, prompt, …):
        # Implementation
        return response, cost, tokens

```

Adding new providers...

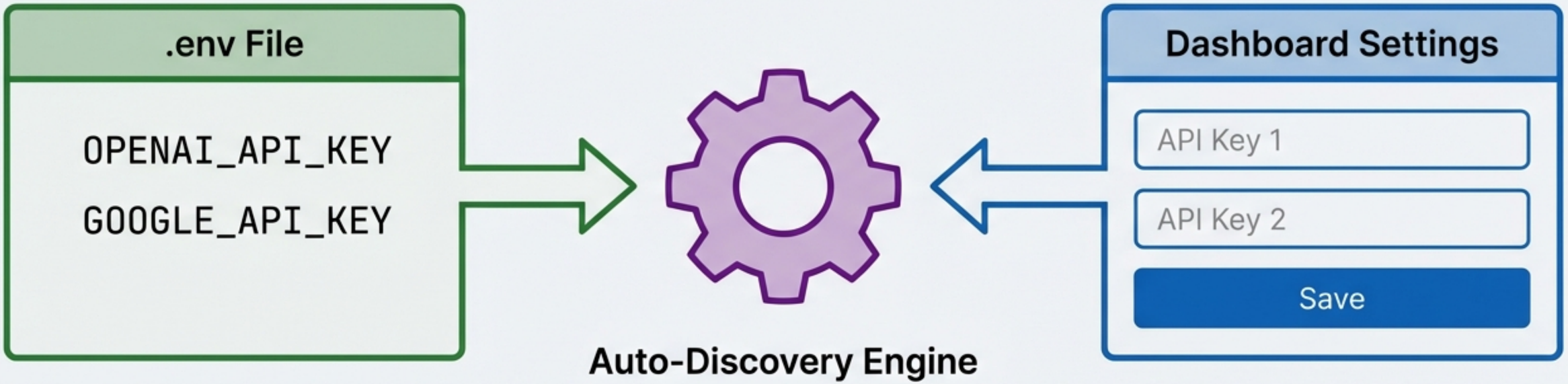
2. Register Model (app/router.py)

```
self.clients = {
    'simple': Phi3Client(),
    'moderate': GeminiClient(),
    'complex': ClaudeClient()  # Added
}
```

Engineering Editorial

Configuration & Auto-Discovery

The system automatically identifies available models based on provided keys.

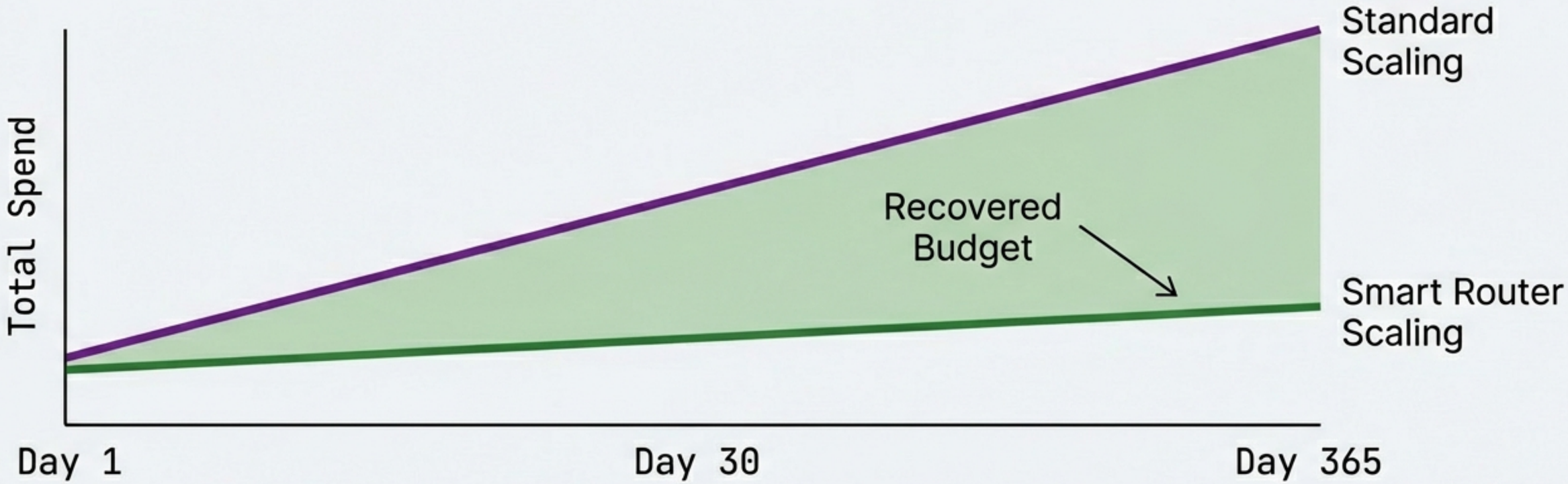


System detects capabilities and enables Smart Classification.

Engineering Editorial

Small Optimizations, Enterprise Impact

Per-request savings compound significantly over time.



“What gets measured gets managed. What gets routed gets saved.”

Route Smart, Build Faster

The Cost-Control Smart Model Router is open source and ready for contribution.

adi-7192 /

Cost-Control-Smart-Model-Router

MIT License Python Open Source

- Fork the repo to customize routing logic.
- Submit Pull Requests for new providers.
- Deploy your own traffic controller.

```
> git clone https://github.com/adi-7192/Cost-Control-Smart-Model-Router
```