Reducing generative AI token usage with intelligent prompt routing

Client
A global safety science company providing testing, inspection, and certification services

Company Size
14,000+ employees

Location
Global

A global safety science company needed to control the rising cost of running generative AI across a large-scale data extraction project. Marlabs deployed PromptRouter®, its proprietary routing platform, to analyze each request and send it to the most cost-effective model capable of a quality response, cutting token usage by roughly 30%.

Rather than sending every request to a single expensive model, PromptRouter® evaluates the complexity of each prompt and routes it to the model best suited for the task, reserving frontier intelligence for the work that truly needs it. Built-in guardrails keep every request within corporate and compliance guidelines.

Generative AI

AI Strategy

AI-Powered Analytics

Data Engineering

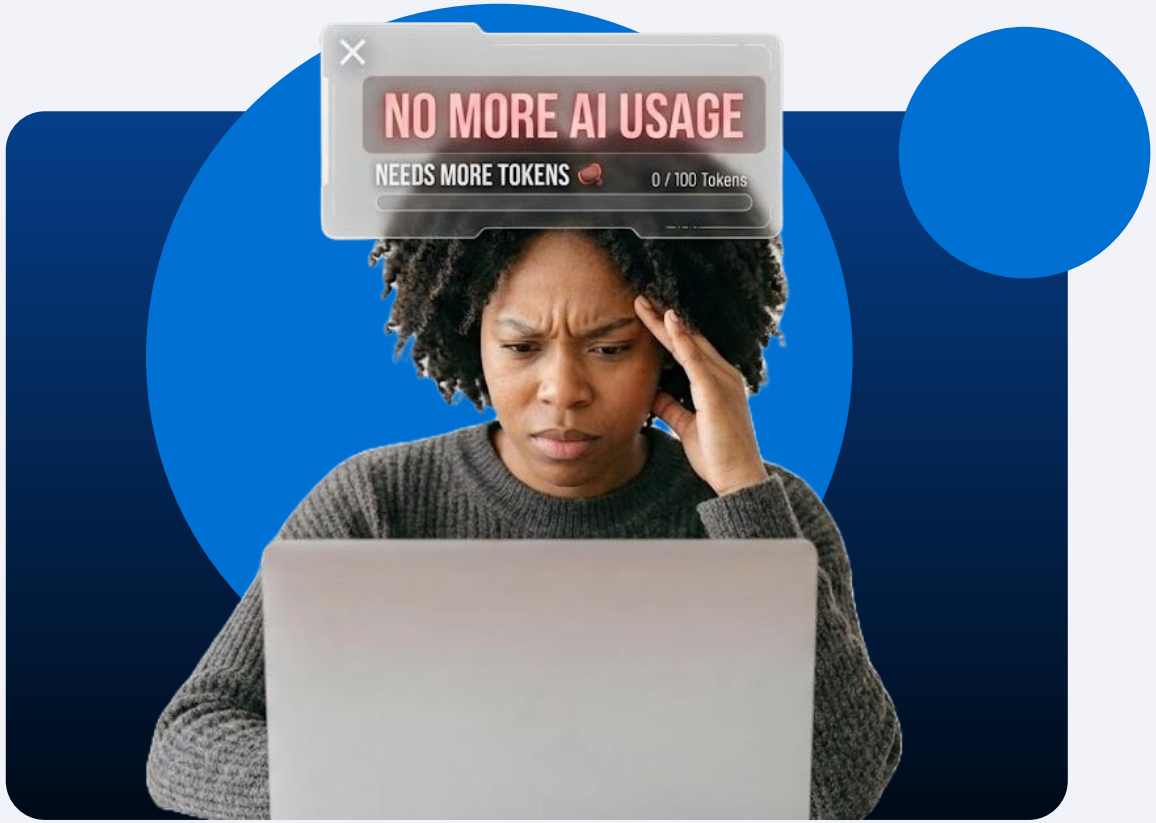
The Challenge: High costs and single-model habits limited generative AI at scale

Objective: Reduce the cost and token usage of generative AI on a large-scale data extraction project without sacrificing output quality.

Existing Issues: Sending every request to a single frontier model drove up token consumption and made generative AI expensive to run at scale.

Solution Needed: Deploy PromptRouter® to analyze each request and route it to the most appropriate, cost-effective model.

Outcome: Token usage fell by roughly 30% while maintaining response quality and governance.



Running generative AI for a large data extraction workload was expensive because every request, simple or complex, went to the same high-end model. The client needed to cut cost and token usage without compromising the quality of extracted results.

The Solution: Intelligent prompt routing that matches each request to the right model

Marlabs deployed PromptRouter®, its proprietary platform that analyzes each request's complexity and routes it to the most cost-effective capable model. The platform also enforces responsible-AI guardrails, keeping every request within corporate and compliance guidelines.

Phase 1: Use-Case Discovery & Assessment

We assessed the client's data extraction workload and identified where expensive frontier models were being used for simple requests.

Workstreams:

- Workload analysis
- Cost baseline assessment
- Model usage mapping

Phase 2: Prompt Routing Configuration

The team configured PromptRouter® to score each request's complexity and route it to the most appropriate, cost-effective model.

Workstreams:

- Complexity scoring setup
- Model selection rules
- Routing logic tuning

Phase 3: Governance & Security Integration

Marlabs embedded responsible-AI guardrails so every request stayed within the client's corporate and compliance guidelines.

Workstreams:

- Guardrail configuration
- Policy enforcement
- Prompt interception

Phase 4: Optimization & Validation

Our team validated routing accuracy and tuned thresholds to sustain response quality while lowering token consumption.

Workstreams:

- Response quality testing
- Threshold tuning
- Token usage tracking

Services and Technologies Used:

Services:

- Generative AI
- AI Strategy
- AI-Powered Analytics
- Data Engineering
- AI Readiness

Technologies:

- Marlabs PromptRouter®

The Results: Impact on the client organization

By routing each request to the right model, the engagement lowered the cost of generative AI while preserving quality and control. It also gave the organization a flexible, model-agnostic foundation for future AI adoption.

Lower Token Usage: Token usage on the data extraction project fell by roughly 30% with no loss in output quality.

Optimized AI Costs: Routing simpler requests to more affordable models cut the cost of running generative AI at scale.

Maintained Quality: Reserving frontier models for the most complex requests preserved the accuracy of extracted results.

Responsible AI Use: Built-in guardrails kept every request within the organization's corporate and compliance guidelines.

Model-Agnostic Flexibility: Model-agnostic routing positioned the organization to adopt newer, cheaper models as the landscape evolves.