Optimizing generative AI use with Marlabs' PromptRouter®

Client
A leading Workday consulting firm

Company Size
500+ employees

Location
Headquartered in Chicago, IL

Featured Partners
aws

A technology consulting firm specializing in Workday support was experiencing rising costs and inconsistent quality in its use of large language models (LLMs). With various models available but no routing intelligence in place, high-cost APIs were being overused even for simple tasks, and responses often lacked organizational context.

To address this, Marlabs developed and deployed a solution in their proprietary, web-based PromptRouter® platform that dynamically routes AI prompts based on complexity and business need. This solution integrated the client's knowledge base to enhance accuracy and relevance while optimizing API usage. Through this platform, we cut costs by over 50% and improved AI-assisted productivity across the organization.

- Generative AI Platform Development
- Intelligent Prompt Routing
- Enterprise Data & Knowledge Integration
- Cost Optimization Strategy

The Challenge: Managing cost & quality in LLM-based AI operations

Objective: Improve the efficiency and relevance of AI responses while reducing overall LLM API costs.

Existing Issues: High-cost models were used indiscriminately, and AI outputs lacked integration with proprietary knowledge.

Solution Needed: A smart routing platform that dynamically matches prompts with the most suitable and cost-effective LLM.

Outcome: Substantial cost savings, better response quality, and more consistent AI support across departments.

The client faced mounting expenses from LLM API usage, which was driven by growing internal demand for intelligent responses and context-aware assistance. Without an automated routing mechanism, the organization struggled to align prompt complexity with the appropriate model. This led to unnecessary costs and inconsistent results.

The Solution: Deploying our AI-powered prompt routing platform

Marlabs designed and implemented the solution through our proprietary PromptRouter® platform, which evaluates each AI request in real-time and determines which model to use based on the prompt's complexity and business context. By integrating with the client's knowledge base, the system ensured both cost efficiency and high contextual relevance.

Phase 1: Prompt Analysis & Model Mapping

Our team defined complexity tiers and mapped them to appropriate LLM providers and pricing models.

Workstreams:

- Prompt complexity categorization
- LLM capability matrix
- Business rule alignment

Phase 2: Platform Architecture & Development

We built a scalable web-based platform with front-end and back-end components for routing logic.

Workstreams:

- Python & FastAPI services
- Jinja2 & JavaScript UI
- Routing engine logic

Phase 3: Knowledge Base Integration

The team connected internal content and documentation to enhance LLM responses with enterprise context.

Workstreams:

- Knowledge source indexing
- API connectors
- Context retrieval layer

Phase 4: Deployment, Monitoring, & Cost Tracking

We containerized and deployed the solution with dashboards for real-time monitoring and optimization.

Workstreams:

- Docker deployment
- ELK & Splunk logging
- Cost analytics dashboard

Services and Technologies Used:

Services:

- GenAI Platform Development
- Intelligent Prompt Routing
- Enterprise Data & Knowledge Base Integration
- Cost Optimization Strategy

Technologies:

- Programming & APIs: Python, FastAPI
- Frontend: Jinja2, JavaScript, CSS, and HTML
- Cloud and Database: AWS, PostgreSQL
- AI & ML Frameworks: LangChain, LangGraph, MLFlow
- Deployment, Monitoring & Logging: Docker, Splunk, ELK

The Results: Impact on the client organization

Marlabs' PromptRouter® platform enabled the client to take control of their LLM usage by intelligently matching prompts with the right model and leveraging internal data to deliver high-quality and organizationally contextualized responses. As a result, the client saw immediate cost savings and long-term operational efficiency—transforming their AI approach from reactive to strategic.

Over 50% Cost Reduction: This initiative cut LLM API expenses by routing simple prompts to low-cost models.

Consistent Response Quality: Marlabs standardized outputs by embedding domain-specific context.

Improved Contextual Accuracy: We delivered richer, more relevant responses using integrated knowledge sources.

Lower Prompt Engineering Burden: Our team minimized the need for manual tuning and AI troubleshooting.

Accelerated Workflow Turnaround: The new platform reduced wait times for AI-generated content across departments.

Strategic LLM Utilization: We created a repeatable framework for efficient and cost-conscious AI scaling.