tetrate

Tetrate Agent Router Service

The Hidden Cost of GenAI: Complexity

As GenAI adoption accelerates, so does the complexity of managing multiple models, APIs, and infrastructure. Most developers still connect to large language models (LLMs) using single-vendor SDKs or brittle scripts, resulting in runaway costs, inconsistent performance, and uncontrolled shadow AI. With new providers and specialized models emerging rapidly, teams are struggling to keep up. They need a reliable way to connect to the right model, at the right time—without rebuilding infrastructure every quarter.

Real-Time GenAI Routing for Safe, Fast, and Profitable AI

Tetrate Agent Router Service simplifies GenAI integration by letting developers define their model preferences based on cost, latency, provider, or compliance, and automating query routing accordingly. Tetrate Agent Router Service (TARS) is a developer's shortest path to models anywhere. It is a managed service built on Envoy AI Gateway that reduces cost, increases resilience, and eliminates operational burden. Tetrate Agent Router Service provides:

- **Preference-Based Model Routing & Connectivity:** route requests dynamically to the best model based on your defined priorities like cost, performance, or policy requirements. Use your own API keys or Tetrate's to connect across clouds and providers.
- **Automatic Failover & Cost-Aware Execution:** Improve uptime and reduce token burn with intelligent fallback and price-aware routing.
- **Fully Managed Service:** get started instantly with zero infrastructure to maintain; supports isolated tenancy and on-prem data planes for sensitive workloads.

Use Cases

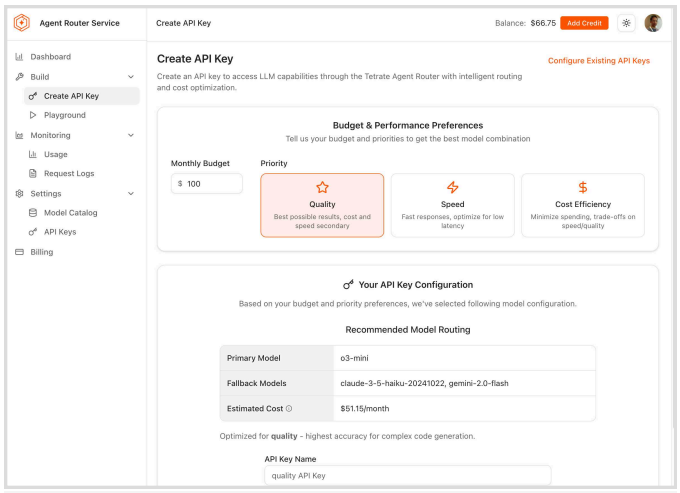
Use Case	Solution
Code Generation	Reliable, optimized code suggestions with reduced hallucinations and cost-aware model selection
Chatbots	Low-latency, cost-efficient conversations using responsive model routing and active failover
Agents	Seamless multi-model orchestration with secure access and unified policy enforcement

Why Tetrate Agent Router Service

Faster Than Building In-House – Eliminate infrastructure overhead and accelerate time to value with a fully managed, enterprise-grade service run by the experts behind Envoy.

Smarter Than Single-Provider Solutions – Avoid lock-in and reduce costs with dynamic model routing, multi-provider support, and automatic failover.

Stronger Than Other AI Gateways – Built on Envoy for proven reliability, with integrated governance and full lifecycle management tailored for enterprise needs.



Explore Tetrate Agent Router Service

Define Once, Route Intelligently Set model preferences for each use case—like cost, latency, or compliance—and let the service handle the routing. Get the best of every model, without operational headaches or vendor lock-in.

Full Visibility, Zero Infrastructure Monitor usage and costs in real time across all providers. Centralize audit logs, enforce policies, and manage spend from a single control plane. No infrastructure to scale or secure.

Built for Enterprise-Grade Resilience Avoid disruptions with built-in failover and latency-aware routing. The service automatically detects and reroutes around outages or performance degradation.

Rapid Experimentation Iterate faster with prompt playground and A/B testing tools. Compare model responses, fine-tune prompts, and improve application quality at speed.

Built on Envoy, Managed by Experts Leverage the same tech powering global-scale service meshes. Tetrate's team of Envoy maintainers runs the platform, so your team doesn't have to.

Ready to help accelerate the enterprise's path to fast, safe, and profitable AI?

We are on a mission to transform application networking and security for the modern, multi-cloud era.

Visit us at tetrate.io or email us at info@tetrate.io

tetrate