

PRODUCT UPDATE

AI Gateway: New Features

A summary of what shipped in the latest AI Gateway release

Extensibility via Policy Templates

- The AI gateway now offers the same extensible policy framework as the Zuplo API Gateway
- Teams/apps can inherit policy templates and add custom code to the request chain
- Unlocks a wide range of custom use cases (custom routing, DLP, guardrails etc.)

JWT Authentication

- The AI gateway now supports JWT auth, along with the default API key auth
- The gateway can take different actions based on JWT claim values (budget, routing, authz)

Per-User Budgets

- Set spend/usage budgets scoped to an individual user, not just an app or team
- Typically a user sub, user ID or session ID would be forwarded via header
- **COMING SOON** ability to natively define unique budgets for specific users in the portal UI

Multi-Model Bundling

- A single gateway endpoint can front multiple models/providers
- Support for fallback/backup models for resilience and cost-based routing
- **COMING SOON** smart routing based on semantic prompt match (eg. simple prompts can be routed to cheaper models)

Support for New Models

- Native support for additional models including Bedrock Mantle (Vertex and Azure are coming soon)
- Custom providers are supported as well, as long as they are OpenAI-compatible
- <https://zuplo.com/docs/ai-gateway/providers>

Expanded Analytics & Observability

- **Team/app cost, token, and request rollups:** hierarchical visualization of spend and usage for the entire organization, individual teams, and apps/agents
- **Model usage dashboards with drilldowns:** breakdown by cost, tokens, and requests by model
- **Cache hit reporting** and other usage-pattern reporting
- **Detailed request-level logging and audit trail:** customizable per-request logs showing status code, route, model, latency, client geo/network info, tools called, the user prompt, etc.