GLOO | AI GATEWAY

Your gateway to AI innovation

Secure, observe, and control your AI applications with Gloo AI Gateway, the leading cloud-native AI gateway built on Envoy Proxy.

Why Gloo AI Gateway

Developers integrating LLMs with their applications will require self-service and AI-enabled platforms. Gloo AI Gateway provides modern platform engineering tools and solutions to deliver AI-enabled applications to production.

Secure & control

Protect applications, models, and data from inappropriate access. Ensure safe use of AI with governance controls, auditability, and visibility into consumption.

Cloud-native & open source

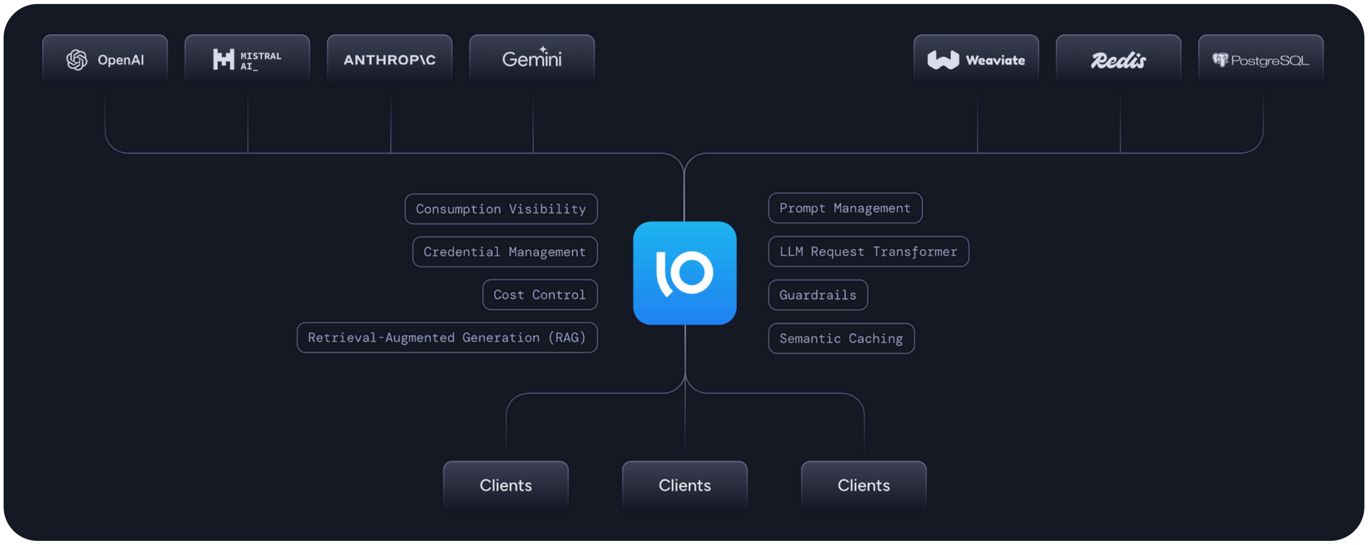
Cloud-native design – built on Envoy Proxy and Kubernetes Gateway API – scales to any environment with support for automation, configuration as code, and self-service with guardrails across private and public cloud environments.

Fast track development

Eliminate development friction, boilerplate code, and avoidable errors in applications consuming LLM APIs across multiple providers and use cases.

Expand & scale

Leverage advanced AI integration patterns for data augmentation and integration with cloud-native gateway capabilities to support high-volume, zero-downtime AI connectivity.



Gloo AI Gateway Capabilities

Unified access point for consumption of LLM APIs

- Provide an API proxy to serve as access point for LLM APIs
- Single endpoint regardless of LLM locality
- Simplify LLM access for consumers
- Centralized control, visibility, and governance

Credential management

- Gateway manages LLM provider keys in protected store
- API Keys created by gateway can map to multiple providers, simplifying client development across LLMs
- Leverage advanced authN/Z in gateway for controlling access via JWT, OPA, etc.
- Centralized point of key management (tracking, revocation, refresh)

Consumption control and visibility

- Rate limit requests to LLM APIs
- Set provider and client-specific consumption limits (e.g. max token)
- Track usage by client across multiple LLM providers with access logging for cost control and chargeback

Prompt management

- Prompt governance - screen for unwanted text in prompts and reject
- Prompt context - prepend or append additional prompt instructions for consistency across requests
- Prompt templating library - constrain prompt interface to known prompts that accept variable substitution
- Reject or transform inappropriate or sensitive response content

Model performance

- Retrieval Augmented Generation (RAG) enrichment of prompt requests with relevant context
- Semantic caching of prompt responses to allow immediate response to semantically equivalent queries
- Integration with third-party vector databases to leverage context enrichment and content caching via embeddings

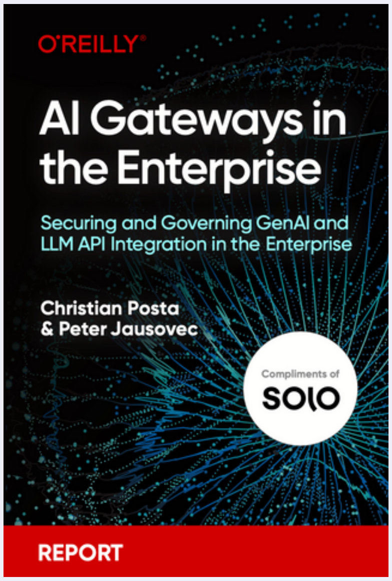
Supercharge cloud-native AI applications with Gloo AI Gateway. Learn more at solo.io/products/gloo-ai-gateway.

"As enterprises accelerate their adoption of LLMs, the challenges of model selection, governance, security, cost control, and observability become critical.

Gloo AI Gateway can help organizations address these challenges, enabling quick experimentation and model adoption, enforcing guardrails, managing costs through quota enforcement and failover, and gaining deep visibility into LLM usage."

Christian Posta,
Global Field CTO

Read Christian's latest book:



O'REILLY

AI Gateways in the Enterprise

Securing and Governing GenAI and LLM API Integration in the Enterprise

Christian Posta & Peter Jausovec

Compliments of SOLO

REPORT

Download