Integrating LLM models into your applications?

4 Reasons Why You *Need* an AI Gateway

First, what is an AI Gateway? An AI Gateway connects, secures and observes traffic from AI applications to backend AI model endpoints, like large language models (LLMs).

AI Gateways are purpose-built to deliver on the specific needs of AI applications, offering:

1. Model Diversity:

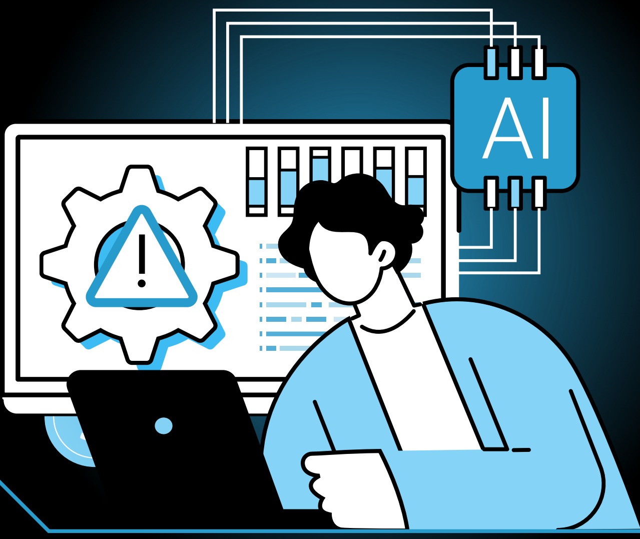
AI Gateways integrate across models and LLM APIs to support GenAI services. They enable reliable multi-LLM adoption and traffic operations across AI endpoints.

2. Cost Optimization:

AI Gateways incorporate consumption control mechanisms to help manage the costs often associated with AI workloads. This includes setting consumption limits and rate-limiting requests for services.

3. Observability:

AI Gateways need to provide detailed insights into LLM-specific metrics such as token usage, error rates, and quota management, to aid in optimizing LLM consumption, controlling costs and troubleshooting issues.



4. Robust Security:

LLM models potentially introduce new security threats and risks. An AI Gateway provides the control and security features to help organizations avoid breaches, privacy leaks, and unauthorized access across AI workloads.

Gloo AI Gateway is a cloud-native API gateway that eliminates friction, enabling innovation with confidence.

To learn more about Gloo AI Gateway, read the [Best Practices Guide: Using GenAI APIs with AI Gateways.]

solo.io