Kong

Powering the API world

Kong AI Gateway

AI Gateway to run, secure, and govern traffic to LLMs, AI agents, and MCP servers

Large language models (LLMs) and AI agents don't operate in isolation—they rely on APIs to access data, trigger actions, and interact with other systems. Even with the recent introduction of the Model Context Protocol (MCP), APIs are still the underlying connective tissue that power agentic workflows, and as AI becomes increasingly autonomous, so does the need for secure, observable, and policy-driven connectivity.

Kong is uniquely positioned to support the AI wave by extending its proven API infrastructure to cover AI use cases through the Kong AI Gateway—built on the same core runtime as Kong Gateway. The Kong AI Gateway makes your GenAI projects production-ready by enabling secure, low-code integration with multiple LLMs, while abstracting away cross-cutting concerns like prompt management, PII sanitization, token rate limiting, traffic observability, and much more.

Whether you're deploying AI agents to automate business processes, copilots to enhance developer workflows, or chatbots to improve the customer experience, Kong helps you govern and scale AI usage responsibly across your entire organization.

Cohere | Anthropic | OpenAI | Azure | LLaMA | Mistral | AWS Bedrock | Google Vertex | Hugging Face

Kong AI Gateway helps organizations accelerate their AI transformation:

Centrally Manage Multi-LLMs

Govern LLM consumption across all popular AI providers, including OpenAI, Azure AI, AWS Bedrock, GCP Vertex, Databricks, and more. Centrally enforce policies across all AI traffic to safeguard sensitive data.

Keep AI Costs In Check

Save on AI costs by rate limiting token consumption per consumer, caching responses to redundant prompts, and automatically routing requests to the best model for the prompt.

Monitor Internal AI Consumption

Track LLM usage via pre-built dashboards and AI-specific analytics to make informed decisions around LLM exposure and AI project rollouts.

Why Kong AI Gateway

- **Multi-LLM Consumption**
  Write application code only once and leverage Kong's universal API or native SDKs to consume LLMs with centralized AI security, metrics, credentials management, and more. Unlock new use cases and optimize costs by seamlessly switching from one LLM provider to another without having to change your code base.

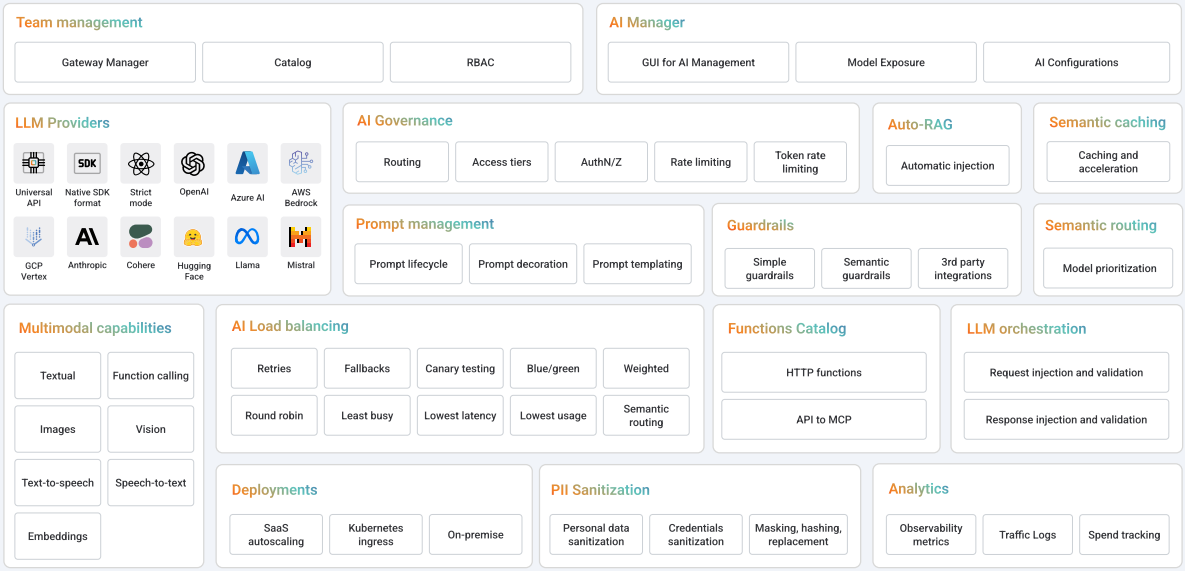
- **Prompt Management**
  Restrict what goes to and from an LLM with advanced prompt management. Block content categories from being prompted with Kong's context-aware semantic guardrails. Configure thresholds for different moderation categories and ensure that prompts sent from your organization meet requirements set by the business, regardless of the original request.

- **Automated RAG**
  Improve AI response accuracy and reduce hallucinations with RAG implemented at the gateway layer by default. The Kong AI Gateway will automatically enrich prompts with relevant context from your data store, without needing developer or AI agent intervention.

- **AI Traffic Control**
  Ensure high availability of AI traffic by load balancing requests across multiple instances of LLM models, with automated retry and fallback. Intelligently route to the most cost-effective model for the job with semantic routing and use semantic caching to efficiently return cached responses for queries that share the same underlying meaning.

- **Automated PII Sanitization**
  Enforce PII sanitization in a global manner across all AI traffic. Kong will automatically detect and sanitize 30+ categories of PII across 12 different languages – ensuring sensitive data never reaches the LLMs. Preserve the end-user experience by optionally reinserting the redacted PII back into the response path.

- **No-code AI Integrations**
  Use Kong's AI request and response transformers to introduce AI inside of your organization without needing to write a single line of code. Easily augment, enrich, or transform API traffic using any LLM provider that Kong supports.



Securely manage both APIs and AI in a consistent manner with a unified platform.

- **AI Access Control**
  Centrally manage multiple Kong AI Gateway deployments across distributed teams and environments. Define roles and groups with RBAC, authentication, authorization, and more to create tiers of access for AI consumption. Apply token rate limiting to enforce token quotas and keep AI costs in check.

- **Self-serve AI APIs**
  Provide a self-serve system of record for developers and AI agents to easily discover, understand, and subscribe to AI services and APIs with the Kong Developer Portal and Service Catalog. Build scorecards that can be attached to each AI API to measure how compliant the services are with your organizational best practices.

- **AI Analytics**
  Create dashboards natively in Kong Konnect Advanced Analytics or export to tools like Grafana, Splunk, or Prometheus for detailed AI usage insights. Monitor prompt and response token volumes, associate cost per token, and attribute spend to specific teams or consumers to enable a chargeback model.

- **Automation via Infrastructure as Code**
  Declaratively deploy and manage AI-powered API, enforce compliance policies, and streamline updates via the Kong admin API, deck CLI, Kubernetes operator, or the Terraform provider. Empower platform owners with the ability to reduce manual configuration and enforce consistent security and compliance across distributed teams and environments.

Kong

Global HQ
44 Montgomery Street, Suite 2920,
San Francisco, Ca, 94104
USA

Contact
sales@konghq.com
www.konghq.com