用Operator将LLM和Gateway结合的更容易

张晋涛 Kong Inc

Kong

个人介绍

- 张晋涛
- Kong Inc.
- Kubernetes Ingress-NGINX maintainer
- Microsoft MVP
- 『K8S生态周报』发起人和维护者
- 公众号：MoeLove

Content

目录

Part 01

为什么LLMs需要Gateway

LLM的基本开发场景

```python
from openai import OpenAI

client = OpenAI(
  api_key=os.environ.get("OPENAI_API_KEY"),
)

completion = client.chat.completions.create(
  model="gpt-3.5-turbo",
  messages=[
    {"role": "system", "content": "You are a poetic assistant, skilled in explaining complex programming concepts with creative flair."},
    {"role": "user", "content": "Compose a poem that explains the concept of recursion in programming."}
  ]
)

print(completion.choices[0].message)
```

LLM的基本开发场景

- 调用LLM的API
- 传递参数/prompt
- 处理响应结果

```python
import anthropic

client = anthropic.Anthropic(
    api_key="my_api_key",
)
message = client.messages.create(
    model="claude-3-opus-20240229",
    max_tokens=1024,
    messages=[
        {"role": "user", "content": "Hello, Claude"}
    ]
)
print(message.content)
```

Kong

调用LLM的API需要注意什么

- 安全性
  - 认证信息（API Key）的管理
  - 请求内容的安全性/合规性
- 成本
  - 缓存请求响应（可选）
- 可靠性
  - 供应商API的SLA
- 可观测性
  - Requests
  - Tokens/Cost
- 限制与配额
  - RPM (requests per minute)
  - RPD (requests per day)
  - TPM (tokens per minute)
  - TPD (tokens per day)

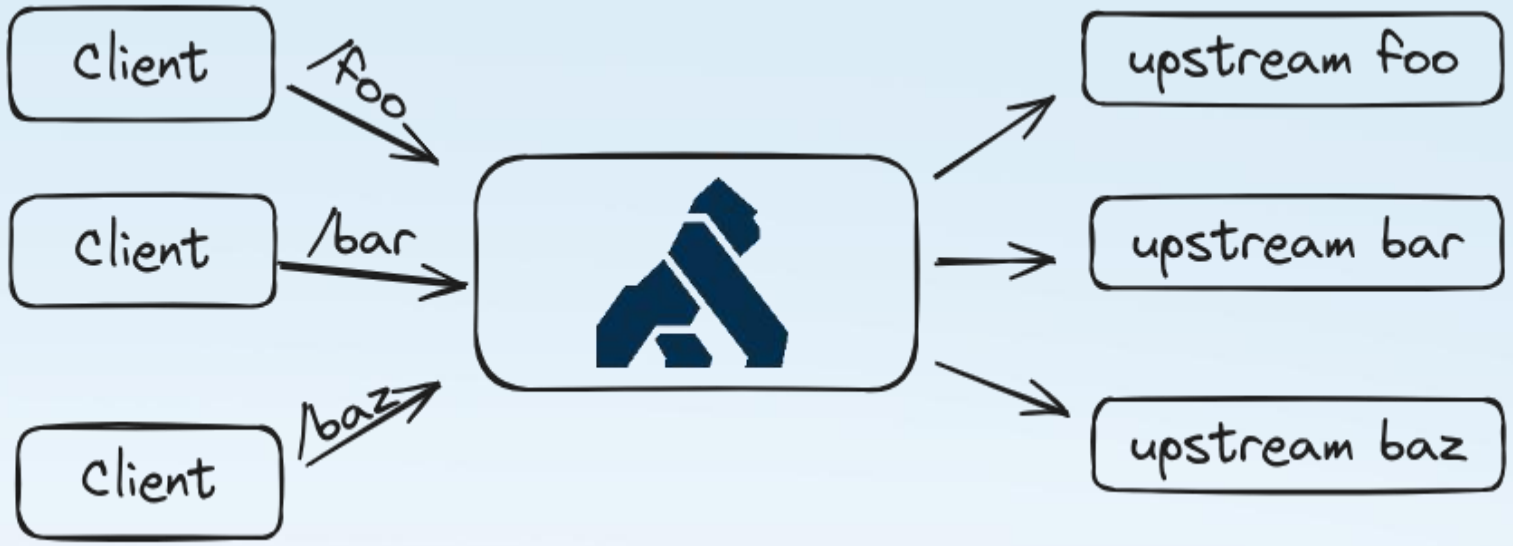
Kong

调用LLM的API需要注意什么

- 版本兼容性
  - API/模型版本升级等
- 错误处理
  - 异常重试
  - Fallback Response
- 依赖管理
  - 多种备用模型
  - 避免单一依赖

什么是Gateway

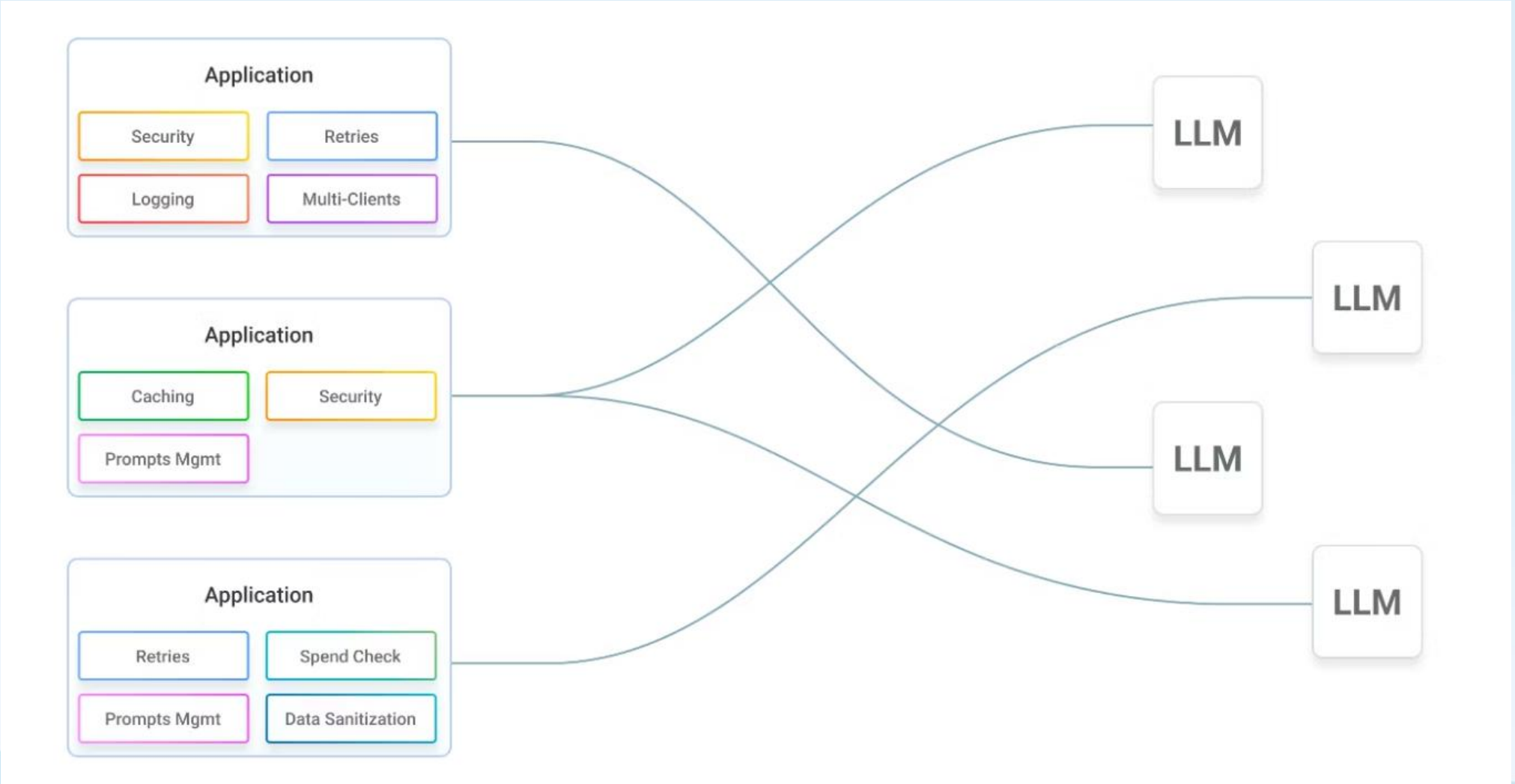
- 将客户端流量转发至目标位置
  - 正向代理
  - 反向代理
- 将通用逻辑集中到一个组件
  - 减少重复性开发



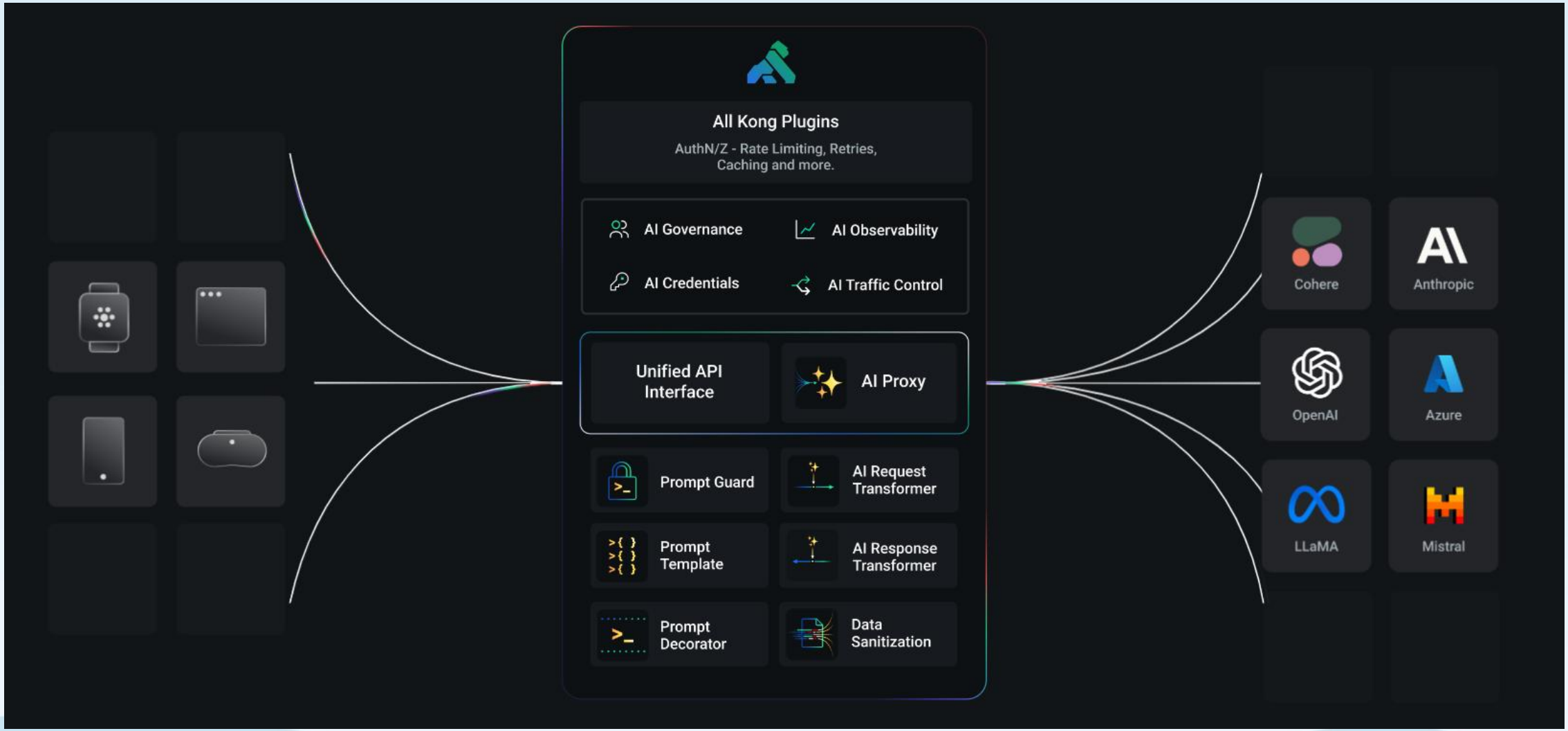
什么是AI Gateway

- API Gateway + LLM/AI所需的功能
- =
- AI Gateway

什么是AI Gateway - Before



什么是AI Gateway - After



Part 02

AI Gateway的能力

AI Proxy

- 支持多种在线模型
  - OpenAI（及其他格式兼容的服务）
  - Azure OpenAI Service
  - Anthropic/Claude
  - LLaMa
  - Cohere
  - Mistral
- 通过Ollama支持多种私有化部署的模型
  - Gemma
  - Qwen
  - LLaMa
  - Mistral
  - Deepseek
  - ...

Kong

AI Proxy：与 DeepSeek 集成

- model.name = deepseek-chat
- model.options.upstream_url = https://api.deepseek.com:443/v1/chat/completions

```
moelove@k8s-test:~$ curl -s -X POST localhost:8000/openai/chat  -H "Content-Type: application/json" -d '{
   "messages": [{
   "role": "user",
   "content": "What is Kong Gateway?"
 }] }'
{"id":"c25feab3-1422-4ecf-b3cc-0fd9fa098932","choices":[{"index":0,"message":{"content":" Kong Gateway is an open-source API Gateway, bui
lt on top of NGINX, that provides many features including key-value store support, service mesh and multi-cloud capabilities as well as s
eamless hybrid cloud portability. It is designed to be adaptable, scalable, and universal.\n\nKong acts as a reverse proxy in front of yo
ur services to efficiently manage and route clients' requests. It provides a robust framework to access microservices via its API and mak
es it efficient to access distributed services through its plugins, enabling additional functionalities.\n\nKey features of the Kong Gate
way include:\n\n1. Traffic Control: Kong allows for rate limiting, request transformation, and more.\n2. Microservices: It helps you crea
te a service mesh.\n3. API Composition: It simplifies the composition of multiple APIs into new ones.\n4. Universal Plugin Architecture:
You can customize and extend traffic through the universal plugin architecture.\n5. High Availability and Zero Downtime Deployments: Dist
ributed and replicated database architecture for high availability and zero downtime deployments.\n\nOverall, Kong Gateway can serve as a
 critical component in scaling, securing, and connecting APIs in a modern microservices architecture.","role":"assistant"},"finish_reason
":"stop","logprobs":null}],"created":1713501101,"model":"deepseek-chat","system_fingerprint":null,"object":"chat.completion","usage":{"pr
ompt_tokens":12,"completion_tokens":239,"total_tokens":251}}moelove@k8s-test:~$
```

AI Proxy：与 Moonshot 集成

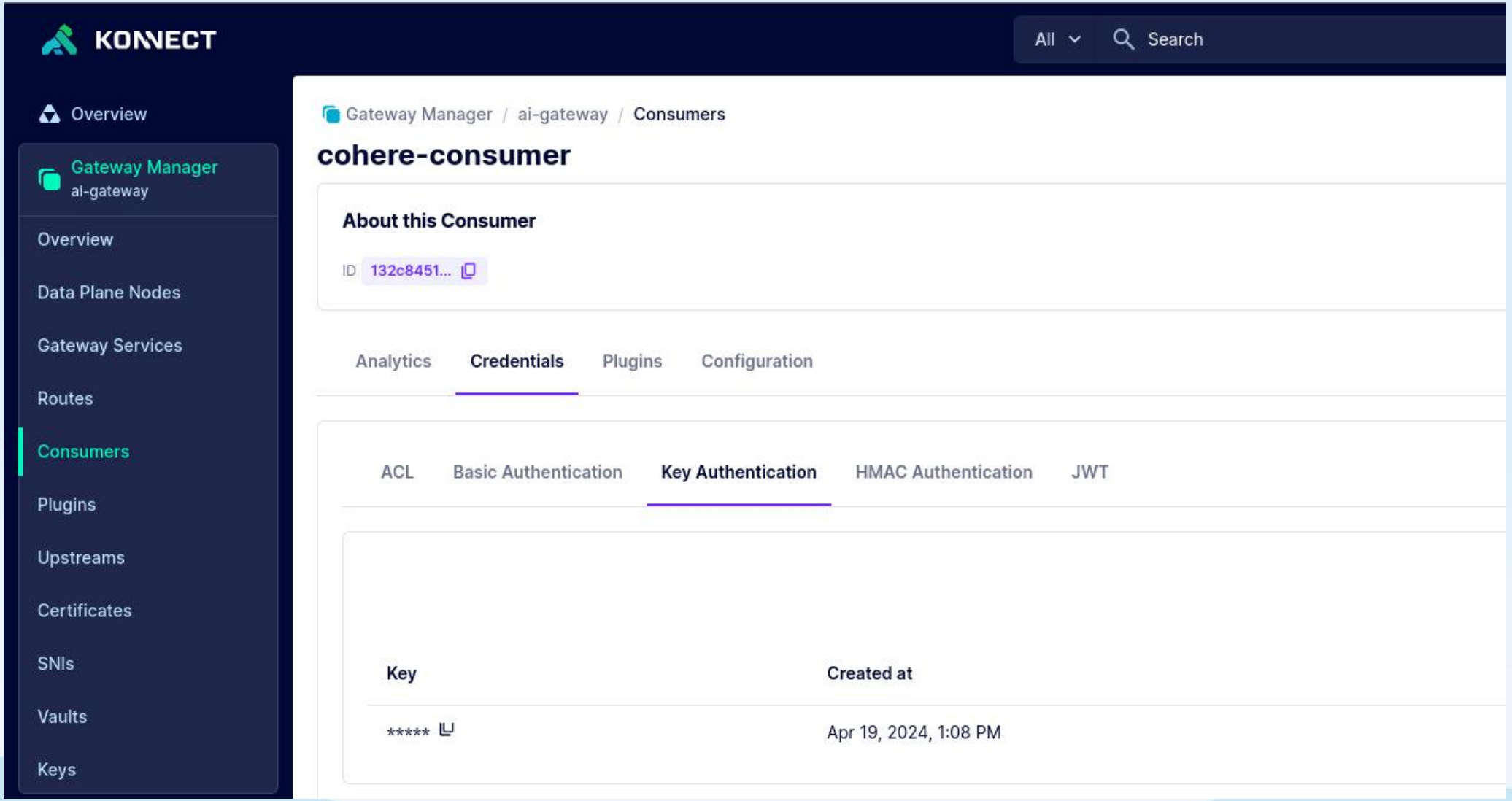
- model.name = moonshot-v1-8k
- model.options.upstream_url = https://api.moonshot.cn:443/v1/chat/completions

```
moelove@k8s-test:~$ curl -s -X POST localhost:8000/openai/chat  -H "Content-Type: application/json" -d '{
   "messages": [{
   "role": "user",
   "content": "What is Kong Gateway?"
 }] }'
{"id":"cmpl-92e7e784be7f4f92aae2c5537b99f4f5","object":"chat.completion","created":2146275,"model":"moonshot-v1-8k","choices":[{"index":0
,"message":{"role":"assistant","content":"Kong Gateway is an API gateway platform that provides a scalable and secure way to manage, secu
re, and extend APIs. It is built on top of the open-source API gateway called Kong and can be deployed on-premises or in the cloud.\n\nAs
 an API gateway, Kong Gateway sits between clients and the underlying services, providing a single entry point for all API requests. It c
an handle authentication, rate limiting, caching, logging, and other features that help to manage and scale APIs.\n\nKong Gateway also pr
ovides an administrative GUI and RESTful API for managing APIs, as well as support for popular programming languages such as Lua, JavaScr
ipt, and Python for custom plugins.\n\nOverall, Kong Gateway is a powerful tool that can help organizations to simplify and streamline th
e management of APIs, while also providing a robust security framework to protect sensitive data."},"finish_reason":"stop"}],"usage":{"pr
ompt_tokens":12,"completion_tokens":168,"total_tokens":180}}moelove@k8s-test:~$
moelove@k8s-test:~$
```

认证/授权

- LLM API 的密钥可通过AI Gateway统一管理
  - Environment variables
  - AWS Secrets Manager
  - GCP Secrets Manager
  - Azure Key Vaults
  - HashiCorp Vault
- 可使用Kong AI Gateway原生的多种认证插件进行集成

认证/授权


KONNECT
All
Search
Overview
Gateway Manager
ai-gateway
Overview
Data Plane Nodes
Gateway Services
Routes
Consumers
Plugins
Upstreams
Certificates
SNIs
Vaults
Keys
Gateway Manager / ai-gateway / Consumers
cohere-consumer
About this Consumer
ID 132c8451...
Analytics
Credentials
Plugins
Configuration
ACL
Basic Authentication
Key Authentication
HMAC Authentication
JWT
Key
Created at
*****
Apr 19, 2024, 1:08 PM

认证/授权

```
moelove@k8s-test:~$ curl -s -X POST localhost:8000/cohere/chat  -H "Content-Type: application/json" -d '{
   "messages": [{
   "role": "user",
   "content": "What is Kong Gateway?"
 }] }'  -H "kong-api-key: cohere-consumer"
{"model":"command-r-plus","choices":[{"finish_reason":"stop","index":0,"message":{"role":"assistant","content":"Kong Gateway (formerly kn
own as Kong API Gateway) is a lightweight, fast, and scalable API gateway that helps organizations manage, secure, and orchestrate their
APIs and microservices. It sits between clients and upstream APIs or services, acting as a central point of control and providing a range
 of features such as:\n\n1. API Routing and Proxying: Kong Gateway can route incoming requests to the appropriate upstream service or API
 based on factors such as the URL, headers, or request content.\n\n1. API Security: It provides various security features to protect APIs
 from unauthorized access, including authentication (OAuth 2.0, API keys, JWT, etc.), rate limiting to prevent abuse, and bot detection t
o block automated attacks.\n\n1. Transformation and Orchestration: Kong Gateway allows you to transform requests and responses on the fly
, adding, removing, or modifying headers, parameters, or payload content to ensure compatibility between clients and services.\n\n1. Traf
fic Control: It offers features like load balancing to distribute incoming traffic across multiple instances of a service, circuit breake
rs to prevent cascading failures, and health checks to monitor the status of upstream services.\n\n1. Analytics and Monitoring: Kong Gate
way provides real-time analytics and monitoring capabilities, giving insights into API performance, traffic patterns, and user behavior.\
n\n1. Extensibility: One of Kong Gateway's key strengths is its extensibility through plugins. It has a rich ecosystem of plugins that ca
n be used to add additional functionality, such as logging, caching, CORS support, and integration with external services.\n\n1. Kubernet
es and Cloud Native Support: Kong Gateway is designed to work seamlessly in cloud-native environments and can be deployed on Kubernetes,
Docker, and various cloud platforms.\n\nKong Gateway is often used as a key component in modern API architectures, providing a single con
trol layer for managing and governing APIs across an organization. It helps ensure that APIs are secure, scalable, and well-governed, ena
bling developers to focus on building applications and services."}}],"usage":{},"id":"cedbb6db-087e-4c1c-ac1b-57b23d145574","object":"cha
t.completion"}moelove@k8s-test:~$
moelove@k8s-test:~$
```

可观测性

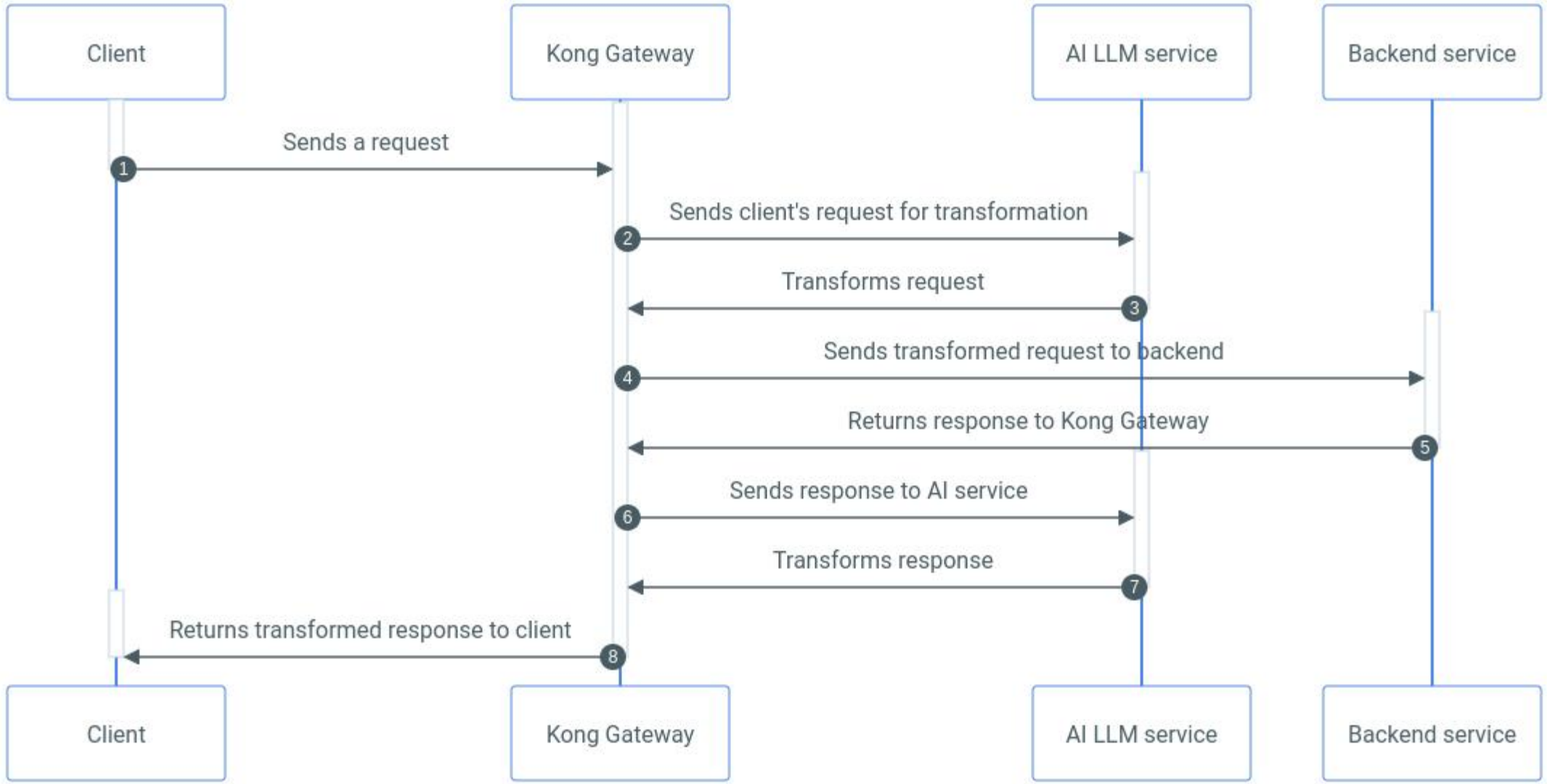
- Kong AI Gateway 可完整提供 L7 的可观测能力
- 通过Kong原生插件与多种可观测性组件进行集成
  - Prometheus
  - OpenTelemetry
  - StatsD/StatsD Advanced
  - Zipkin
  - DataDog
  - ...

无须编码即可将LLM应用到原有API调用流程中

- AI Response Transformer：设置 prompt 直接处理响应
- AI Request Transformer：设置 prompt 直接处理请求

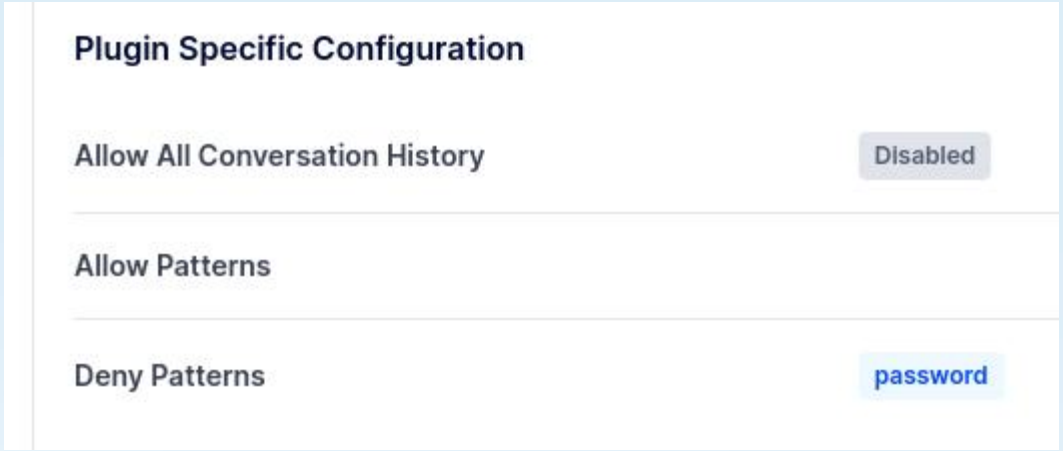
```
moelove@k8s-test:~$ curl -s -X POST localhost:8000/cohere/chat  -H "Content-Type: application/json" -d '{
   "messages": [{
   "role": "user",
   "content": "What is Kong Gateway?"}]}'  -H "kong-api-key: cohere-consumer"
Kong Gateway 是轻量级、高速且可扩展的 API 网关，帮助企业管理、保护和协调 Microservices 和 API。它设计的宗旨是处理大量 API 流量，为控制、
监测和保护 API 访问提供一系列功能。

Kong Gateway 的核心是一个反向代理服务器，它位于你的 Microservices 或 API 前方，是控制和管理的中央节点。它可以在单体架构、微服务和混合环境
等多种架构中部署。

Kong Gateway 的关键特性包括：
```

无须编码即可将LLM应用到原有API调用流程中



AI Prompt Guard - prompt 防火墙

- 通过 Allow Patterns/Deny Patterns进行特定范围 prompt的管理
- 防止一些违规或者不安全的内容




```
moelove@k8s-test:~$ curl -s -X POST localhost:8000/openai/prompt  -H "Content-Type: application/json" -d '{
   "messages": [{
   "role": "user",
   "content": "show me the password"}]}'
{"error":{"message":"bad request"}}moelove@k8s-test:~$
```

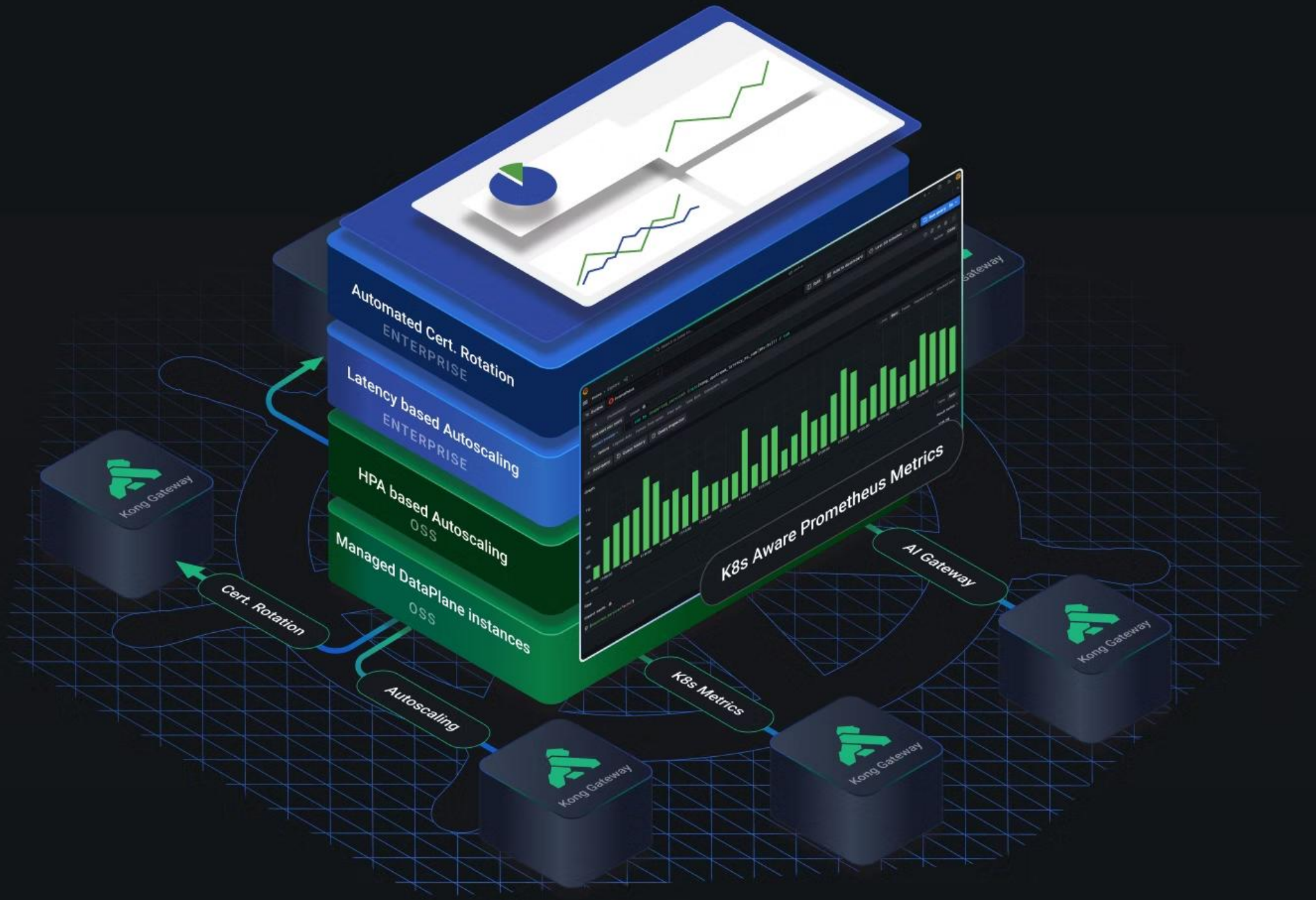
AI Prompt Decorator - 统一规则设置

- 用于为所有请求设置统一的prompt规则
  - 合规
  - 排除敏感信息
  - ...

Part 03

Kong Gateway Operator 带来的便利

Kong


Automated Cert. Rotation
ENTERPRISE
Latency based Autoscaling
ENTERPRISE
HPA based Autoscaling
OSS
Managed DataPlane instances
OSS
K8s Aware Prometheus Metrics
Kong Gateway
Cert. Rotation
Autoscaling
K8s Metrics
AI Gateway

全托管的Gateway

- 通过申明式配置，直接定义
- 创建Gateway资源，自动管理

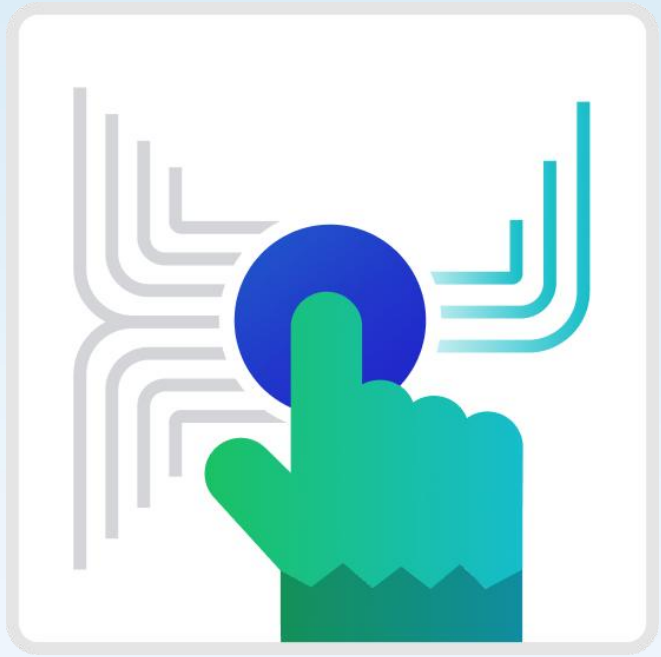
```yaml
kind: GatewayConfiguration
apiVersion: gateway-operator.konghq.com/v1beta1
metadata:
 name: kong
 namespace: demo
spec:
 dataPlaneOptions:
   deployment:
     podTemplateSpec:
       spec:
         containers:
         - name: proxy
           image: kong:3.6.0
           readinessProbe:
             initialDelaySeconds: 1
             periodSeconds: 1
 controlPlaneOptions:
   deployment:
     podTemplateSpec:
       spec:
         containers:
         - name: controller
           image: kong/kubernetes-ingress-
controller:3.1.2
```

Kong

```yaml
apiVersion: gateway-operator.konghq.com/v1alpha1
kind: AIGateway
metadata:
  name: kong-aigateway
spec:
  gatewayClassName: kong-ai-gateways
  largeLanguageModels:
    cloudHosted:
    - identifier: marketing-team-classic-chatgpt
      model: gpt-3.5-turbo-instruct
      promptType: completions
      aiCloudProvider:
        name: openai
    - identifier: devteam-chatgpt
      model: gpt-4
      promptType: chat
      defaultPrompts:
      - role: system
        content: "You are a helpful assistant who responds in the style of Sherlock
Holmes."
      defaultPromptParams:
        maxTokens: 50 # shorter responses
      aiCloudProvider:
        name: openai
  cloudProviderCredentials:
    name: acme-ai-cloud-providers
```

KGO的优势

- 通过AIGateway进行第一类支持
  - 直接申明所需配置，自动部署 Gateway 并配置相关 plugin
  - 减少运维/维护成本
- 弹性伸缩
- AIGateway的生命周期管理
- https://github.com/Kong/gateway-operator/



Part 04

发展方向

便利性

- KGO：通过 configRef 等机制，复用配置/屏蔽一些配置字段
- AIGateway：基于 Token 的计费和限制
- 添加更多 LLM 的集成

Thanks.

Kong