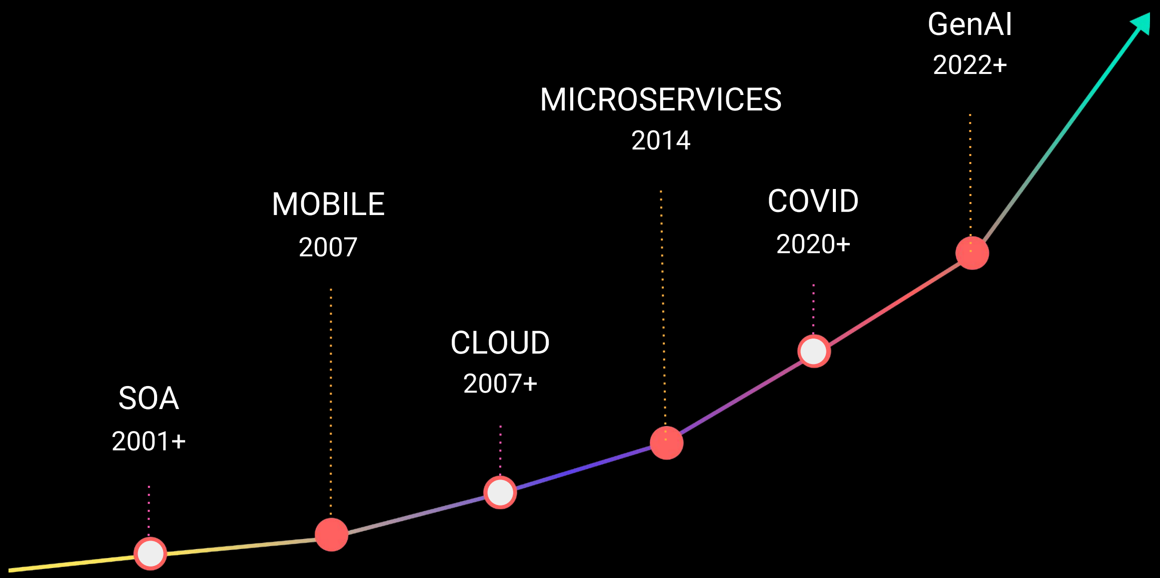
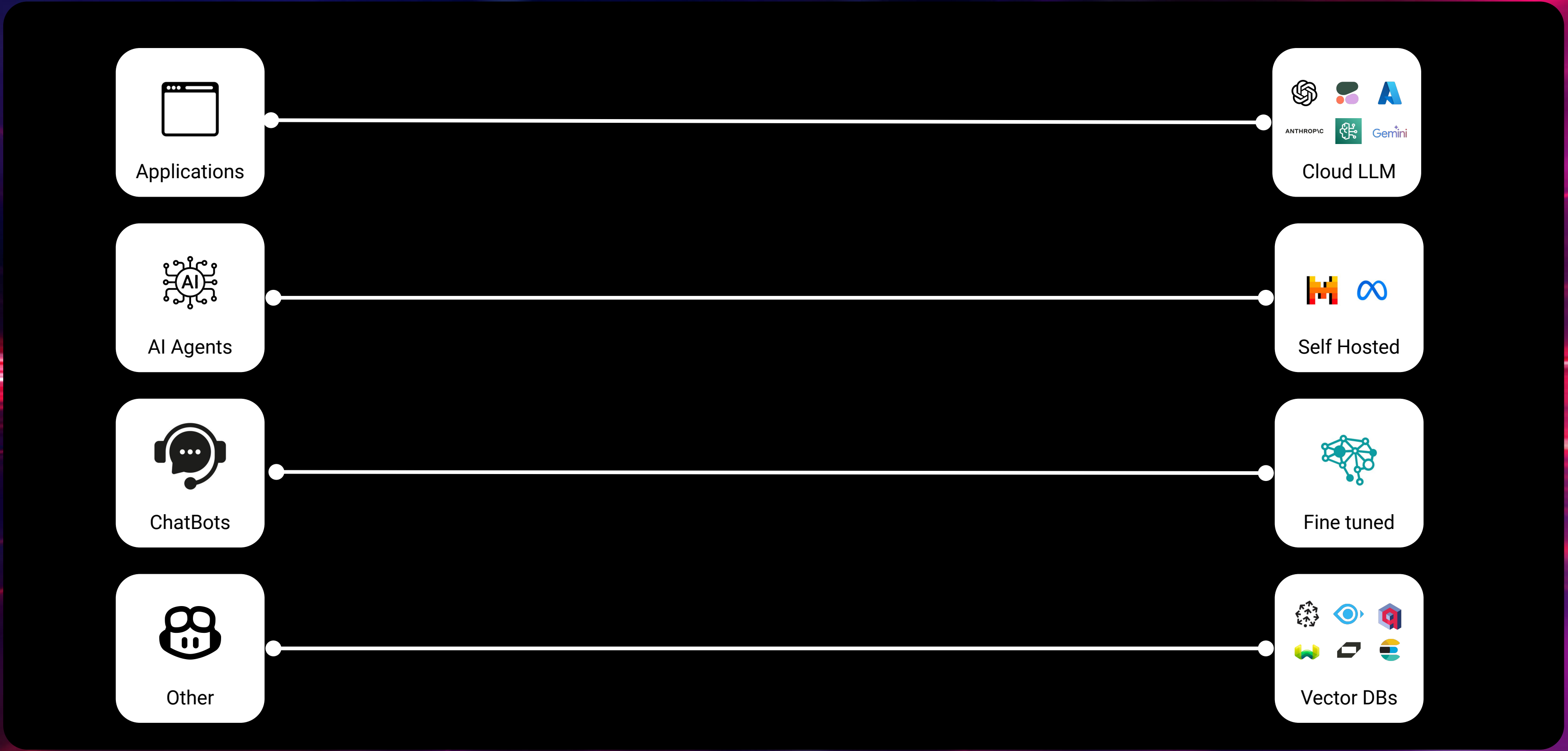
Overview

Kong AI Gateway

Accelerating the AI transformation

API adoption is driven by **digital demand**





Applications

AI Agents

ChatBots

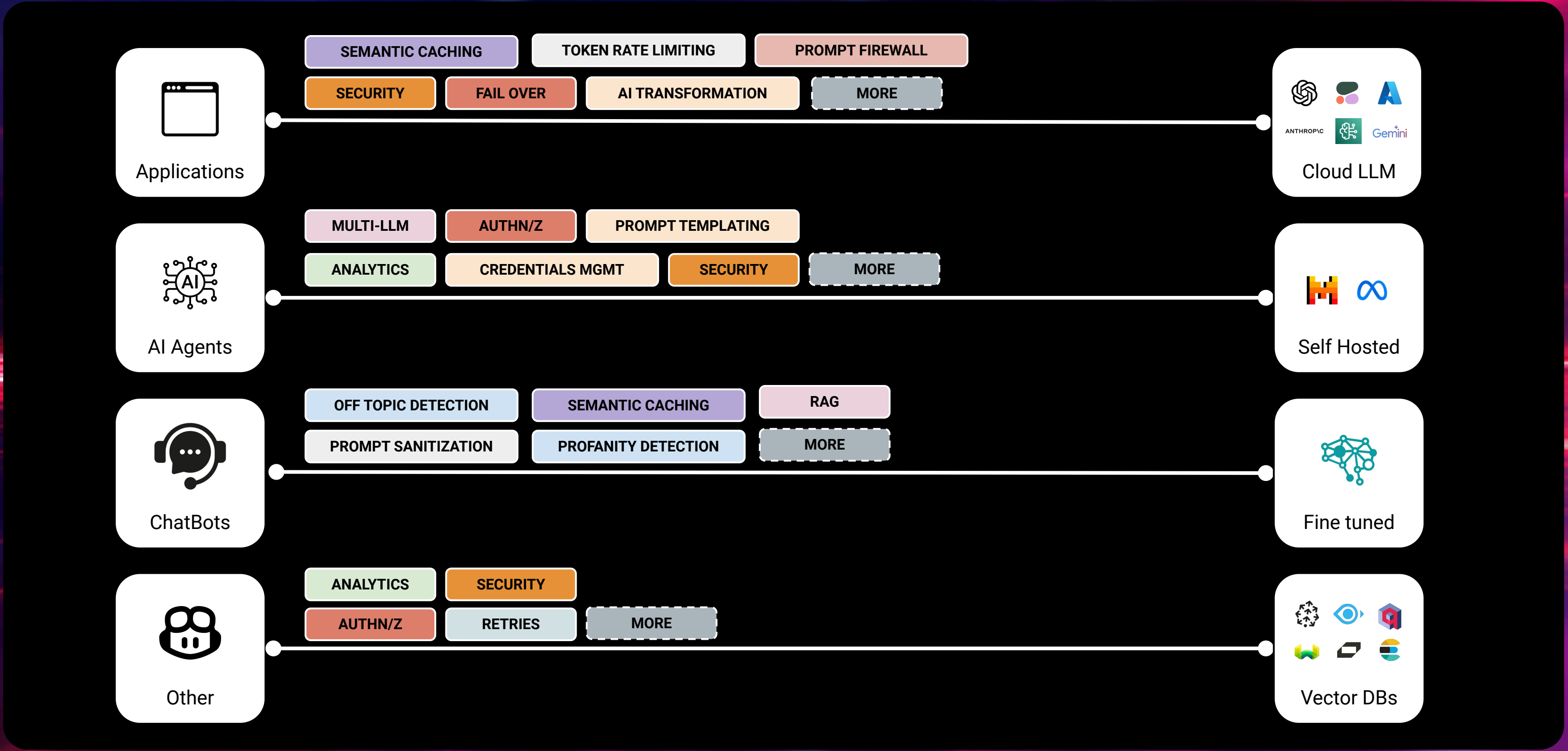
Other

Cloud LLM

Self Hosted

Fine tuned

Vector DBs



Applications

SEMANTIC CACHING
TOKEN RATE LIMITING
PROMPT FIREWALL
SECURITY
FAIL OVER
AI TRANSFORMATION
MORE

Cloud LLM

AI Agents

MULTI-LLM
AUTHN/Z
PROMPT TEMPLATING
ANALYTICS
CREDENTIALS MGMT
SECURITY
MORE

Self Hosted

ChatBots

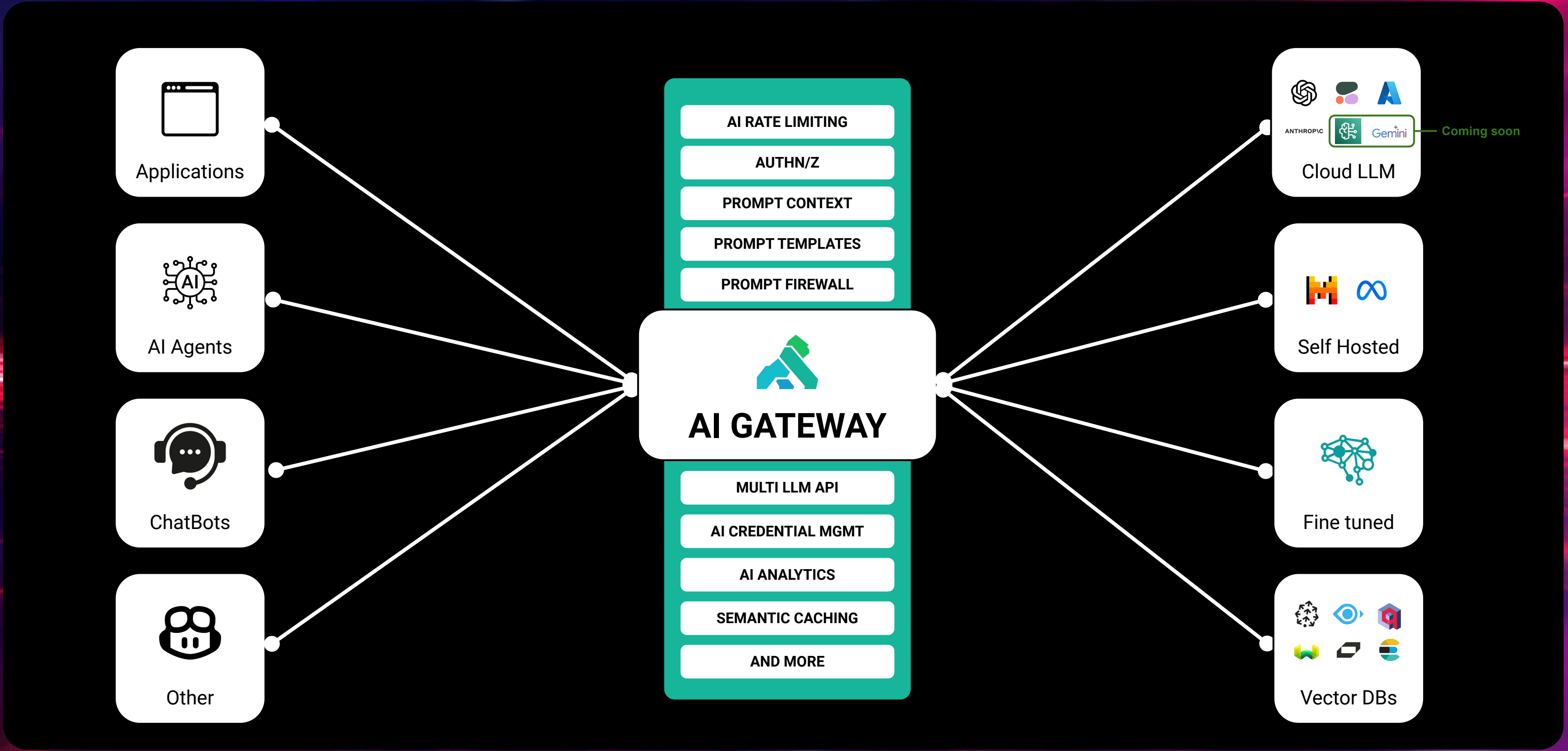
OFF TOPIC DETECTION
SEMANTIC CACHING
RAG
PROMPT SANITIZATION
PROFANITY DETECTION
MORE

Fine tuned

Other

ANALYTICS
SECURITY
AUTHN/Z
RETRIES
MORE

Vector DBs



Applications

AI Agents

ChatBots

Other

AI RATE LIMITING

AUTHN/Z

PROMPT CONTEXT

PROMPT TEMPLATES

PROMPT FIREWALL

AI GATEWAY

MULTI LLM API

AI CREDENTIAL MGMT

AI ANALYTICS

SEMANTIC CACHING

AND MORE

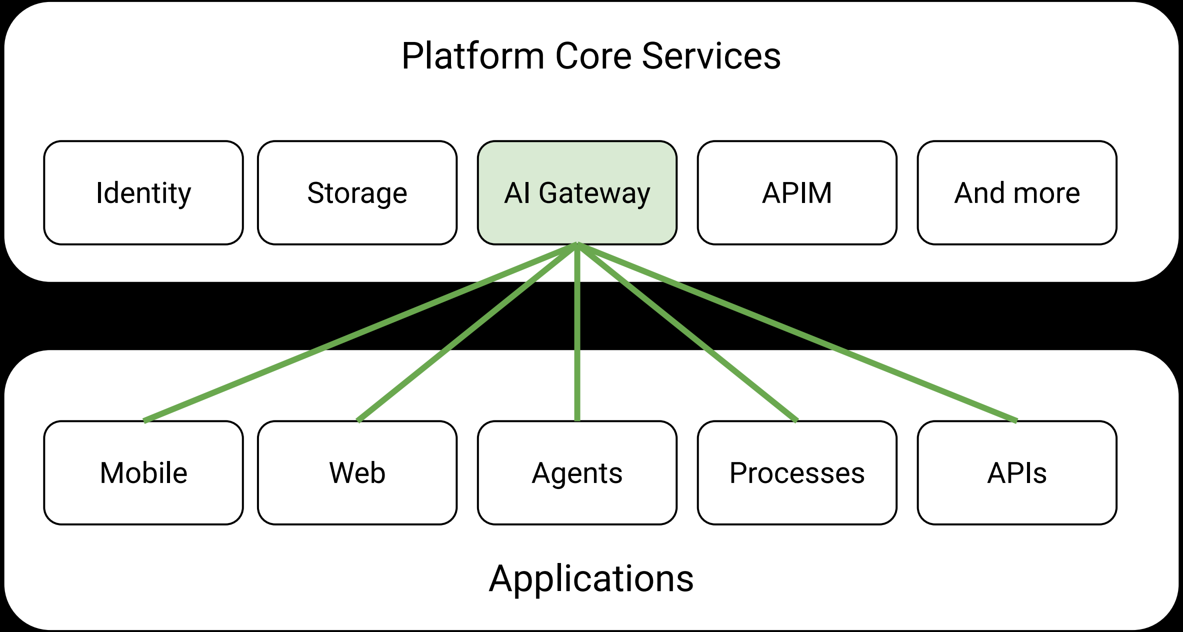
Cloud LLM

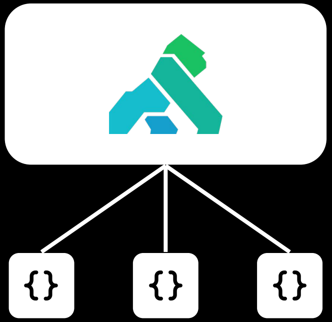
Coming soon

Self Hosted

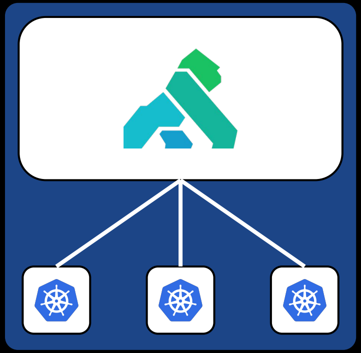
Fine tuned

Vector DBs

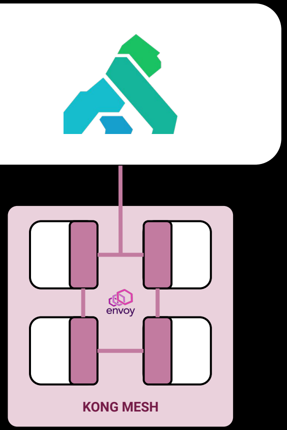

Platform Core Services
Identity
Storage
AI Gateway
APIM
And more
Mobile
Web
Agents
Processes
APIs
Applications



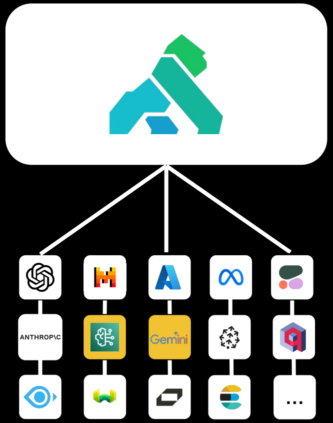
API GATEWAY



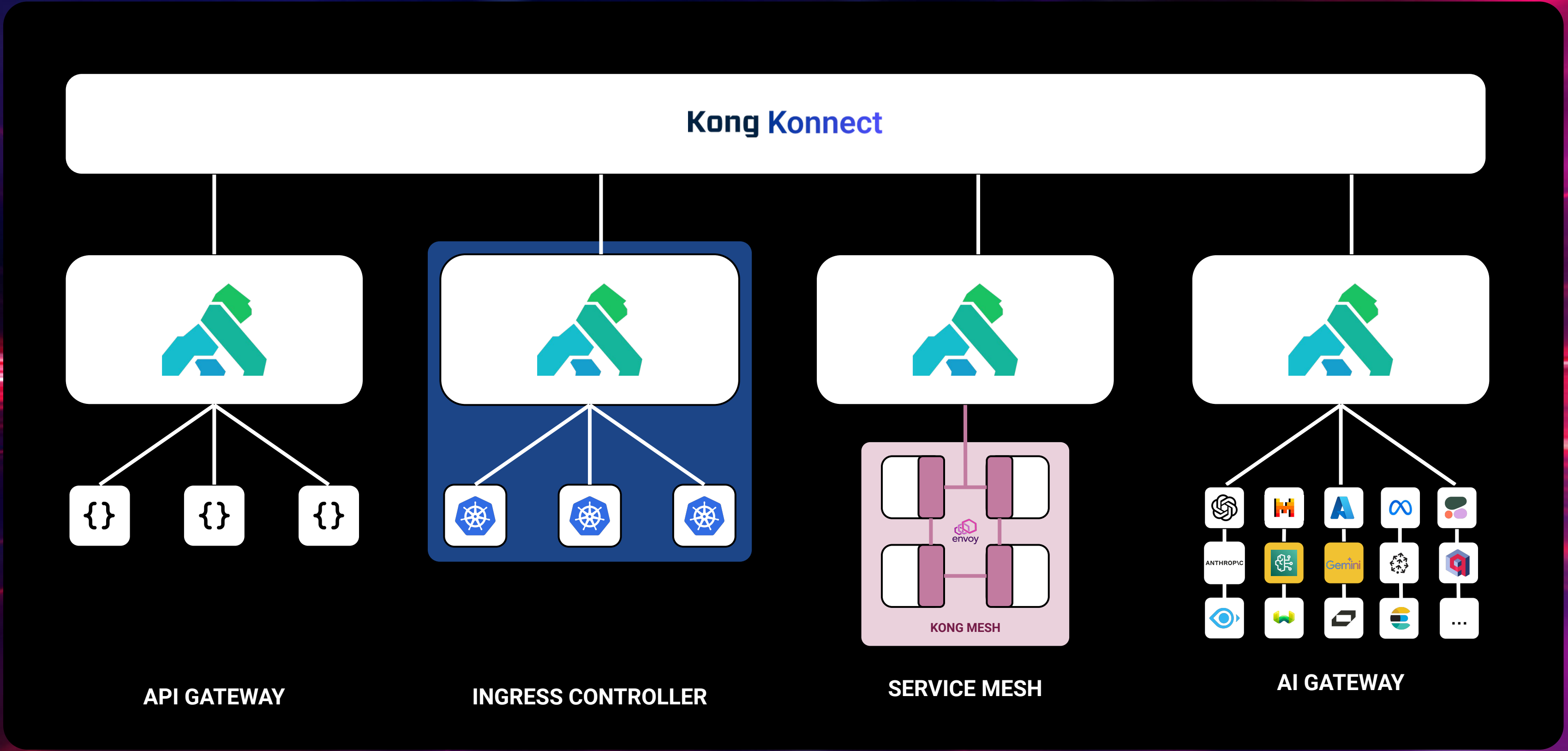
INGRESS CONTROLLER



SERVICE MESH



AI GATEWAY



Kong Konnect

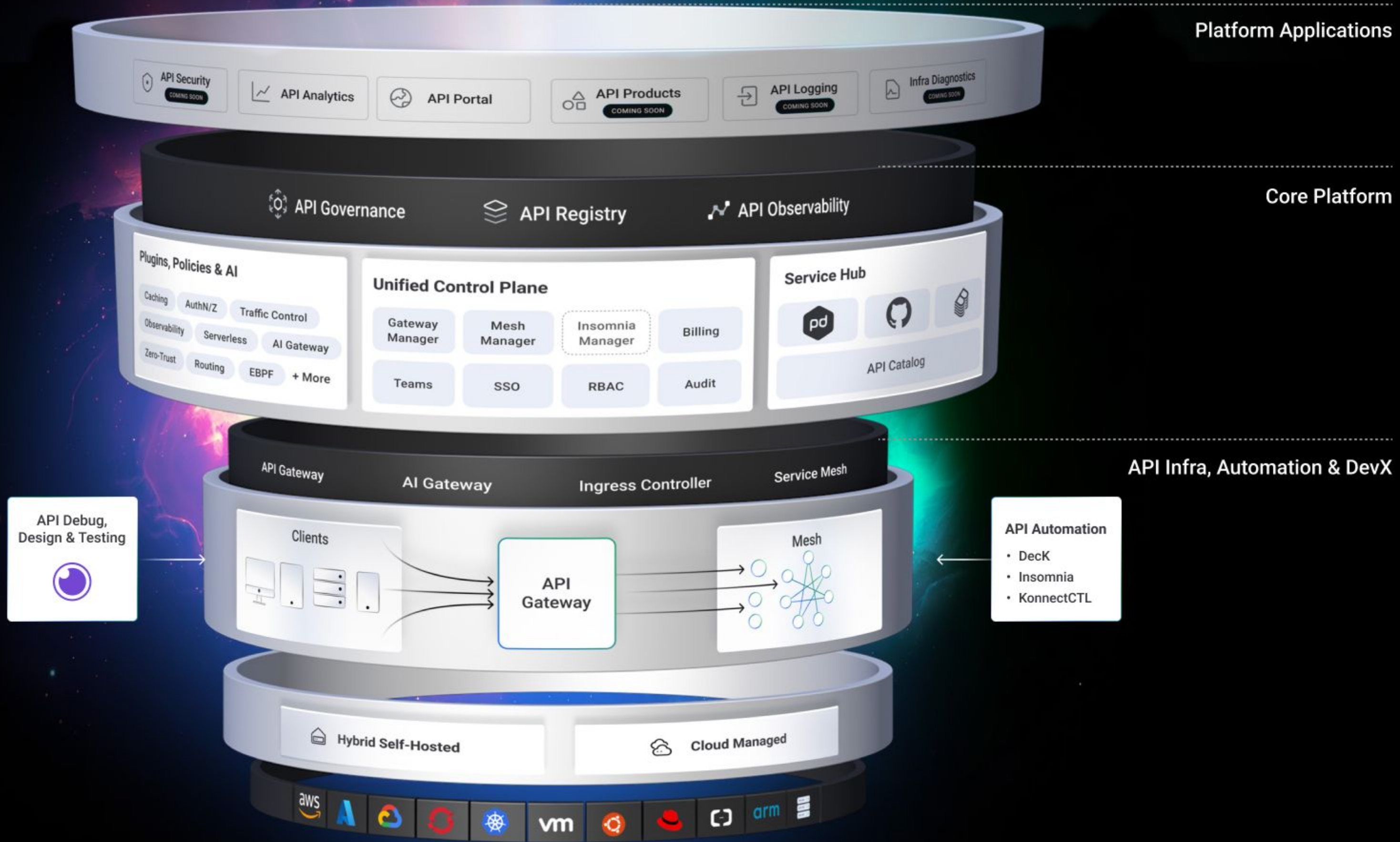
API GATEWAY

INGRESS CONTROLLER

SERVICE MESH

AI GATEWAY

Kong Konnect



Platform Applications

API Security COMING SOON

API Analytics

API Portal

API Products COMING SOON

API Logging COMING SOON

Infra Diagnostics COMING SOON

Core Platform

API Governance

API Registry

API Observability

Plugins, Policies & AI

Caching

AuthN/Z

Traffic Control

Observability

Serverless

AI Gateway

Zero-Trust

Routing

EBPF

+ More

Unified Control Plane

Gateway Manager

Mesh Manager

Insomnia Manager

Billing

Teams

SSO

RBAC

Audit

Service Hub

API Catalog

API Infra, Automation & DevX

API Gateway

AI Gateway

Ingress Controller

Service Mesh

API Debug, Design & Testing

Clients

API Gateway

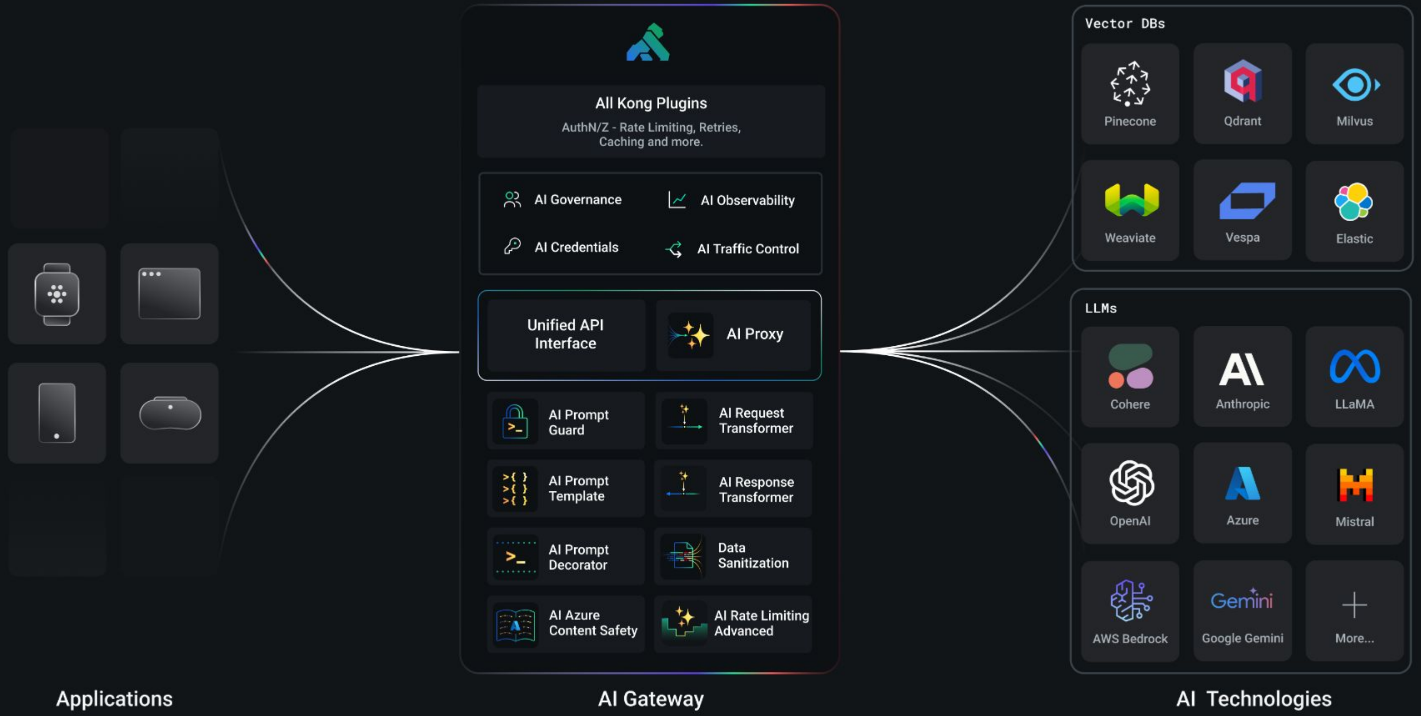
Mesh

API Automation

- DecK
- Insomnia
- KonnectCTL

Hybrid Self-Hosted

Cloud Managed



All Kong Plugins

AuthN/Z - Rate Limiting, Retries, Caching and more.

AI Governance

AI Observability

AI Credentials

AI Traffic Control

Unified API Interface

AI Proxy

AI Prompt Guard

AI Request Transformer

AI Prompt Template

AI Response Transformer

AI Prompt Decorator

Data Sanitization

AI Azure Content Safety

AI Rate Limiting Advanced

Vector DBs

Pinecone

Qdrant

Milvus

Weaviate

Vespa

Elastic

LLMs

Cohere

Anthropic

LLaMA

OpenAI

Azure

Mistral

AWS Bedrock

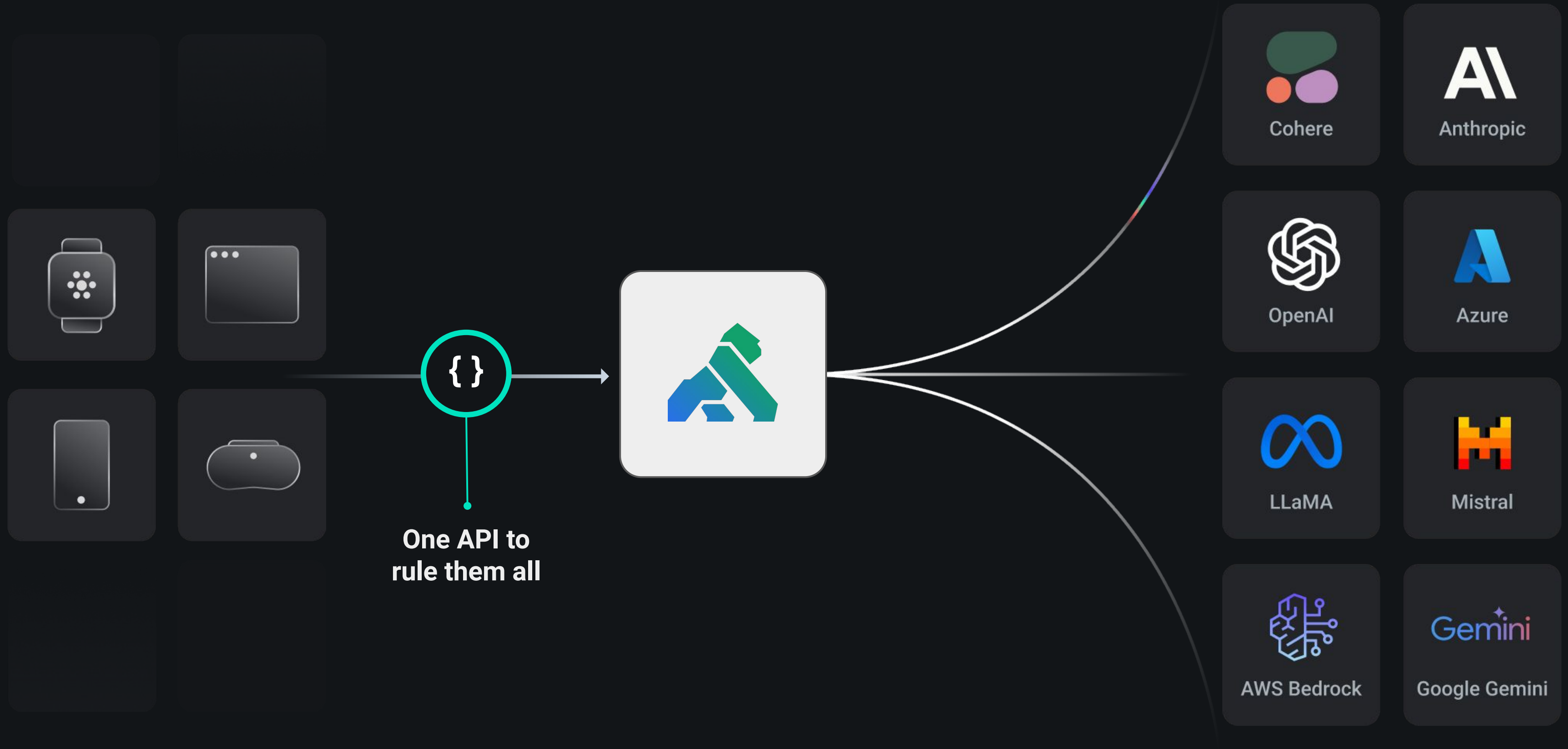
Google Gemini

More...

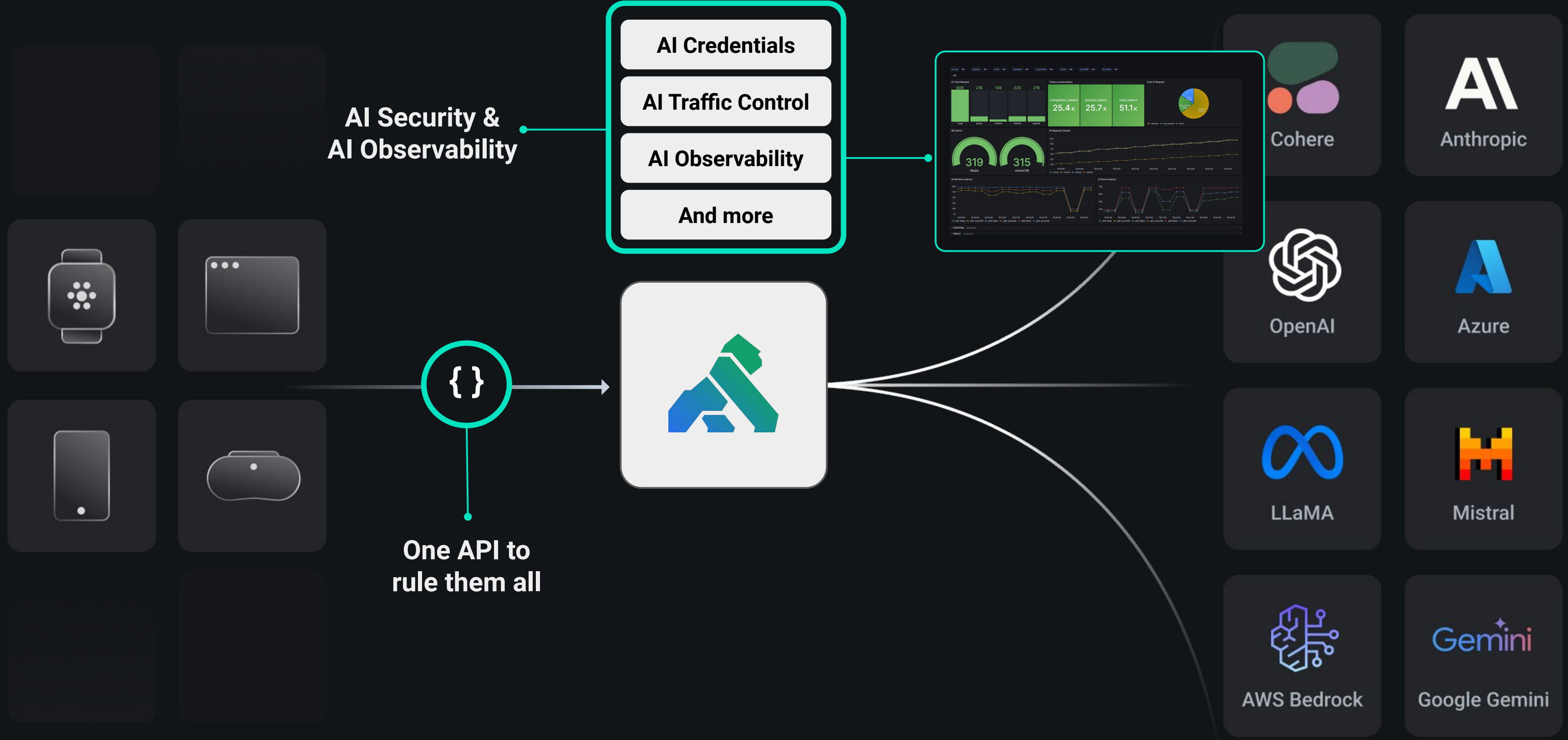
Applications

AI Gateway

AI Technologies



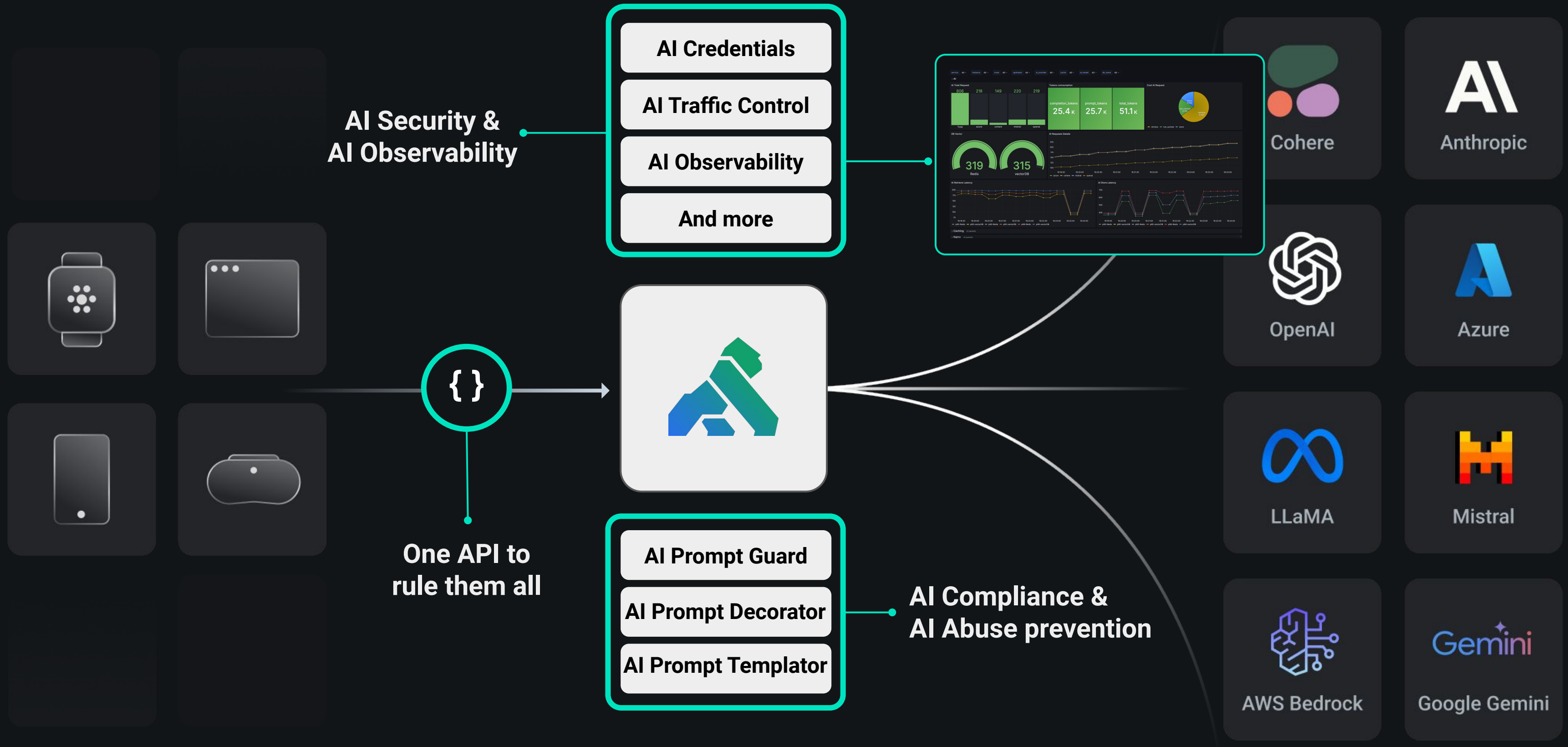
One API to
rule them all



AI Security &
AI Observability

- AI Credentials
- AI Traffic Control
- AI Observability
- And more

One API to
rule them all



AI Security &
AI Observability

AI Credentials

AI Traffic Control

AI Observability

And more

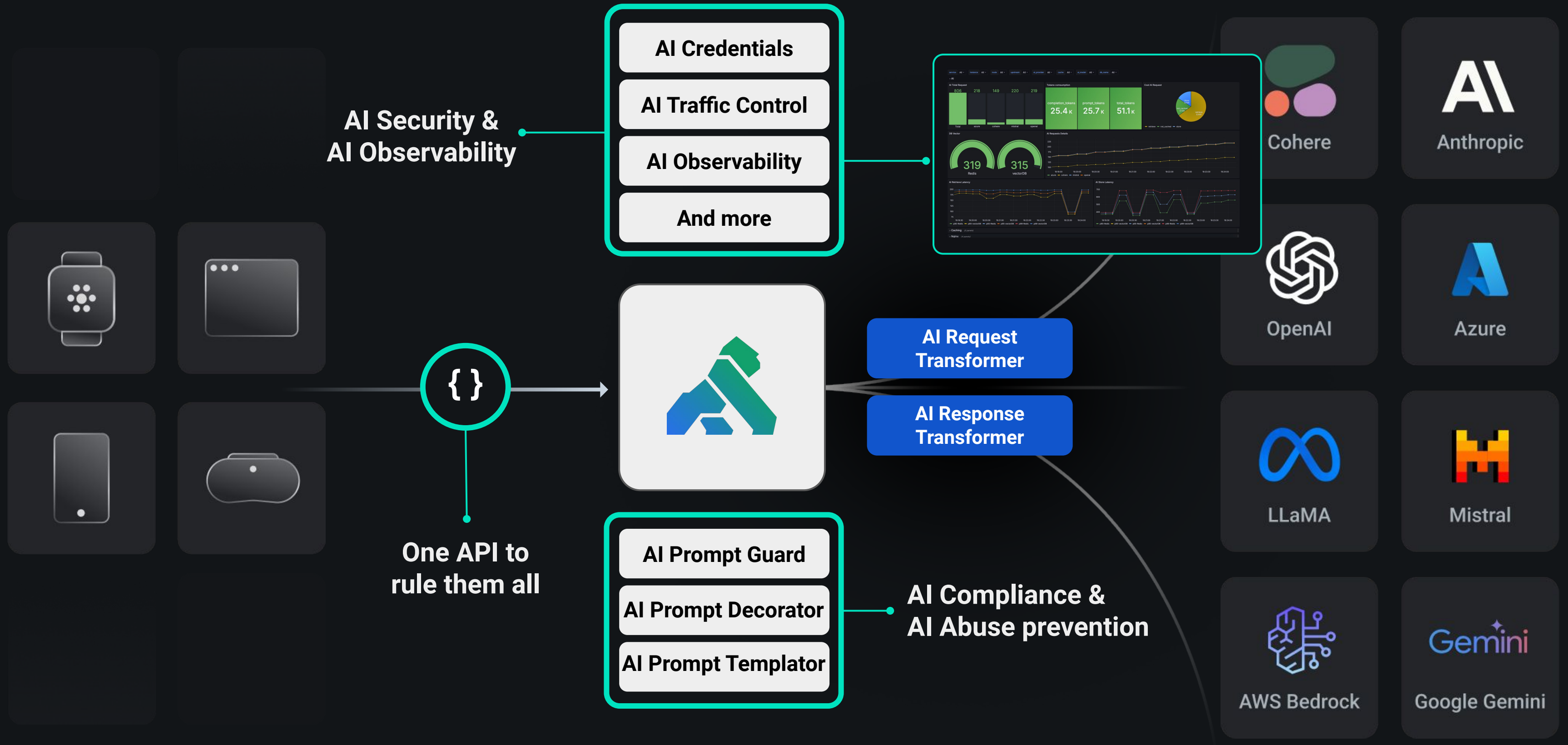
One API to
rule them all

AI Prompt Guard

AI Prompt Decorator

AI Prompt Templator

AI Compliance &
AI Abuse prevention



AI Security &
AI Observability

- AI Credentials
- AI Traffic Control
- AI Observability
- And more

{ }

One API to
rule them all

AI Request Transformer

AI Response Transformer

- AI Prompt Guard
- AI Prompt Decorator
- AI Prompt Templator

AI Compliance &
AI Abuse prevention

Cohere

Anthropic

OpenAI

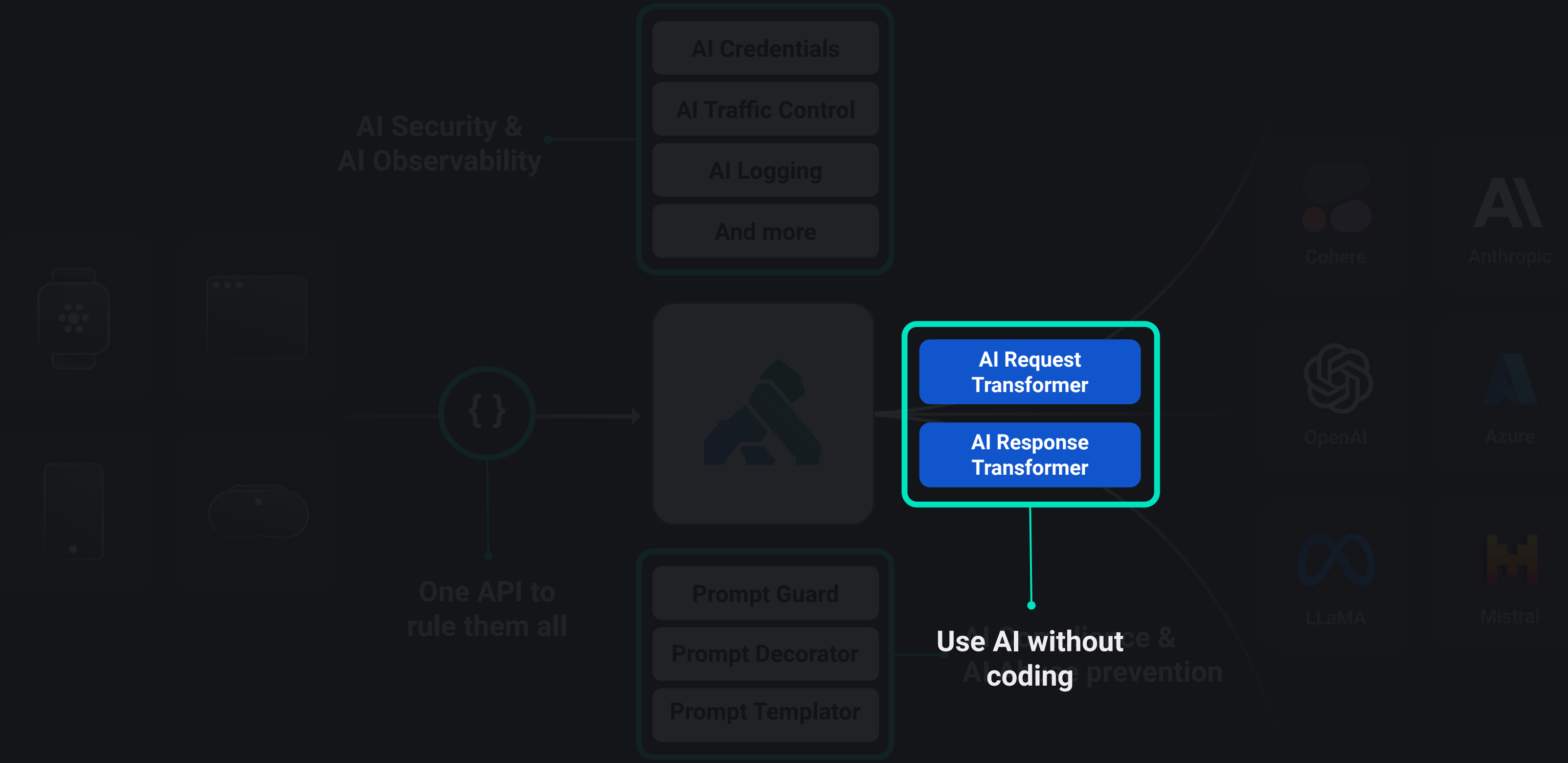
Azure

LLaMA

Mistral

AWS Bedrock

Google Gemini



AI Security &
AI Observability

AI Credentials

AI Traffic Control

AI Logging

And more

One API to
rule them all

Prompt Guard

Prompt Decorator

Prompt Templator

AI Request Transformer

AI Response Transformer

Use AI without
coding

AI Compliance &
AI Abuse prevention

Cohere

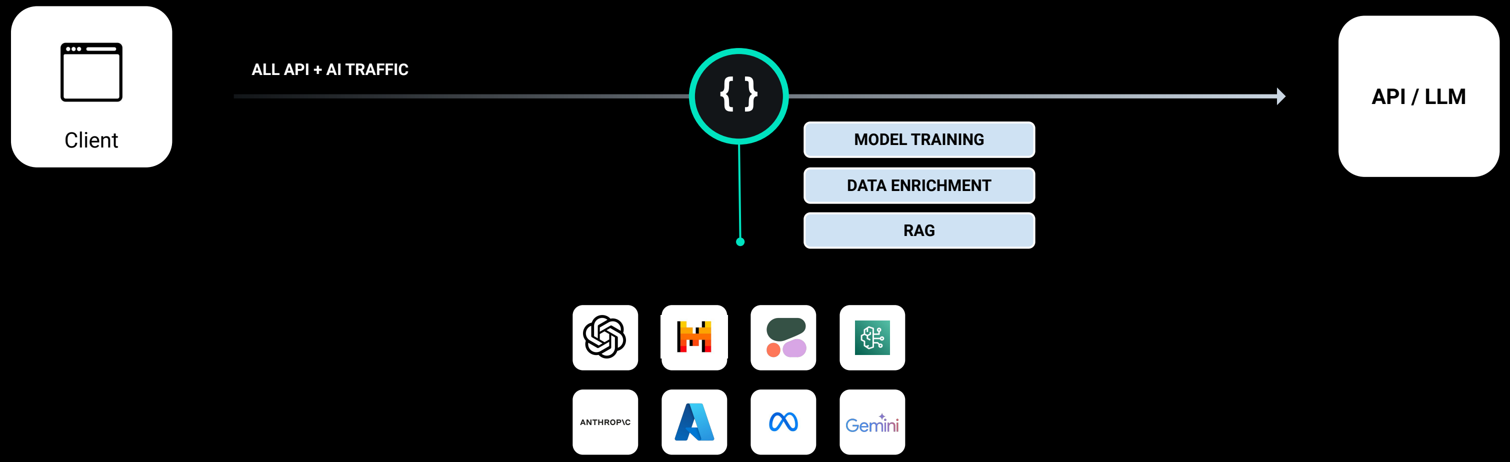
Anthropic

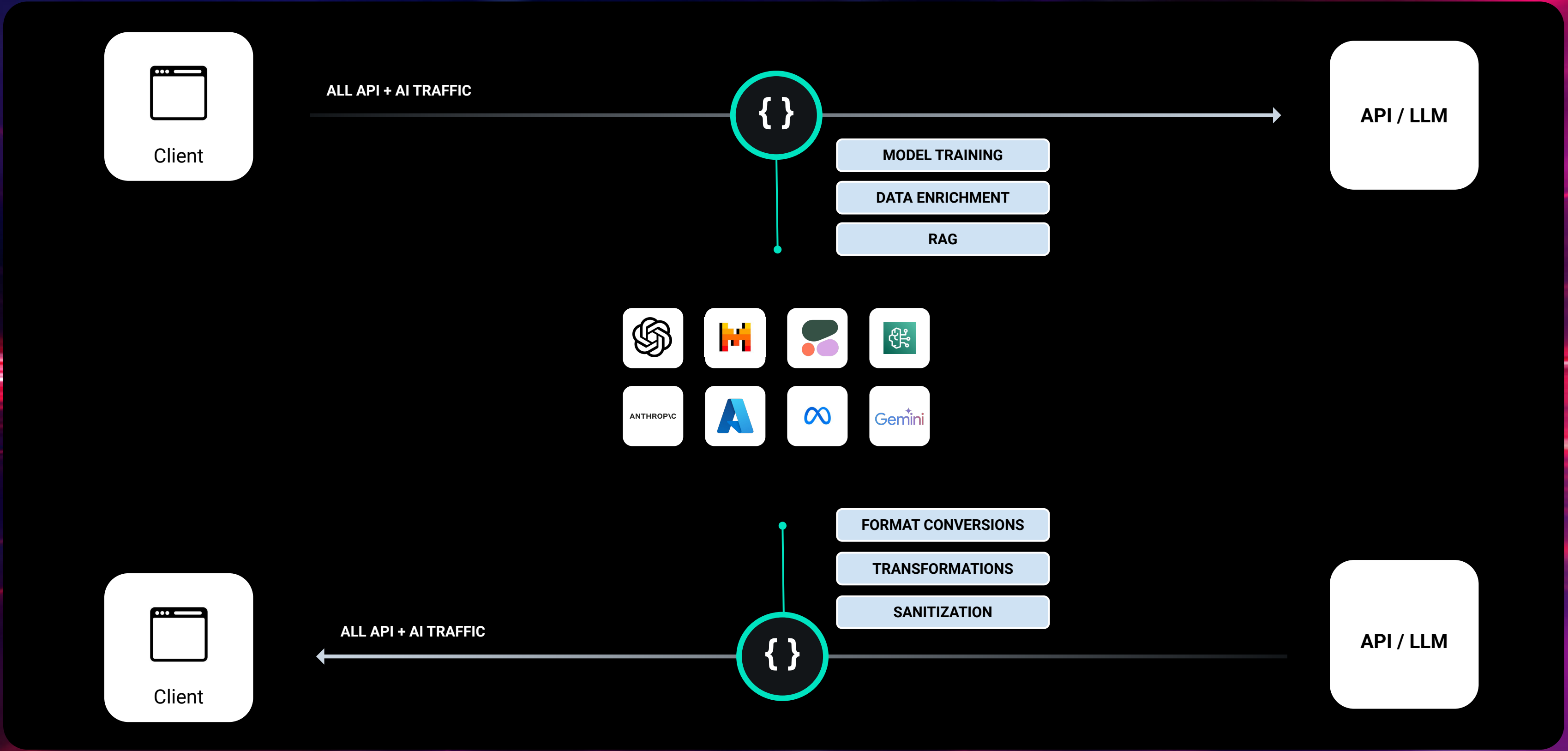
OpenAI

Azure

LLaMA

Mistral


Client
ALL API + AI TRAFFIC
{ }
MODEL TRAINING
DATA ENRICHMENT
RAG
API / LLM
ANTHROP\C
Gemini



Client

ALL API + AI TRAFFIC

API / LLM

MODEL TRAINING

DATA ENRICHMENT

RAG

FORMAT CONVERSIONS

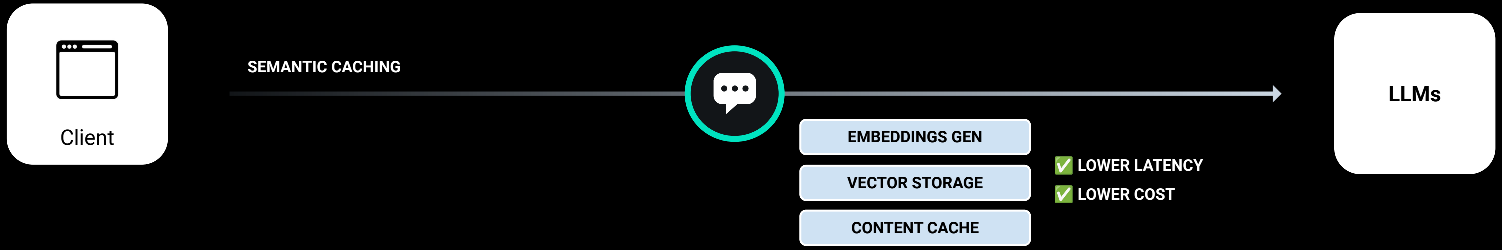
TRANSFORMATIONS

SANITIZATION

Client

ALL API + AI TRAFFIC

API / LLM



Client

SEMANTIC CACHING

EMBEDDINGS GEN

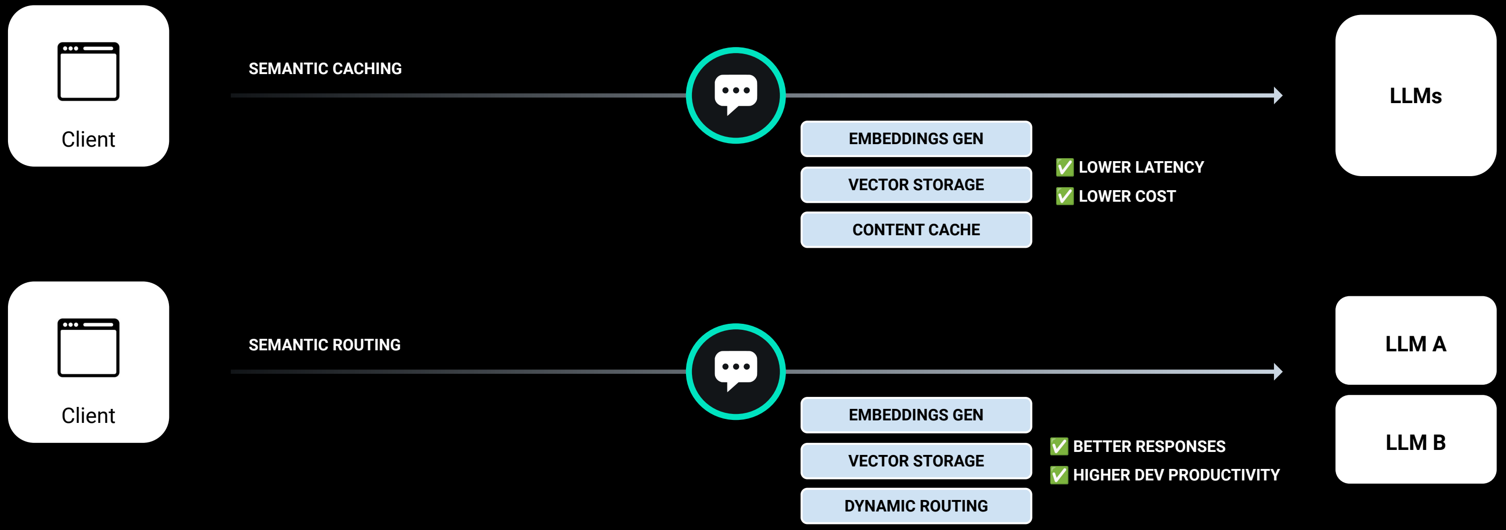
VECTOR STORAGE

CONTENT CACHE

✅ LOWER LATENCY

✅ LOWER COST

LLMs



Client

SEMANTIC CACHING

EMBEDDINGS GEN

VECTOR STORAGE

CONTENT CACHE

✅ LOWER LATENCY

✅ LOWER COST

LLMs

Client

SEMANTIC ROUTING

EMBEDDINGS GEN

VECTOR STORAGE

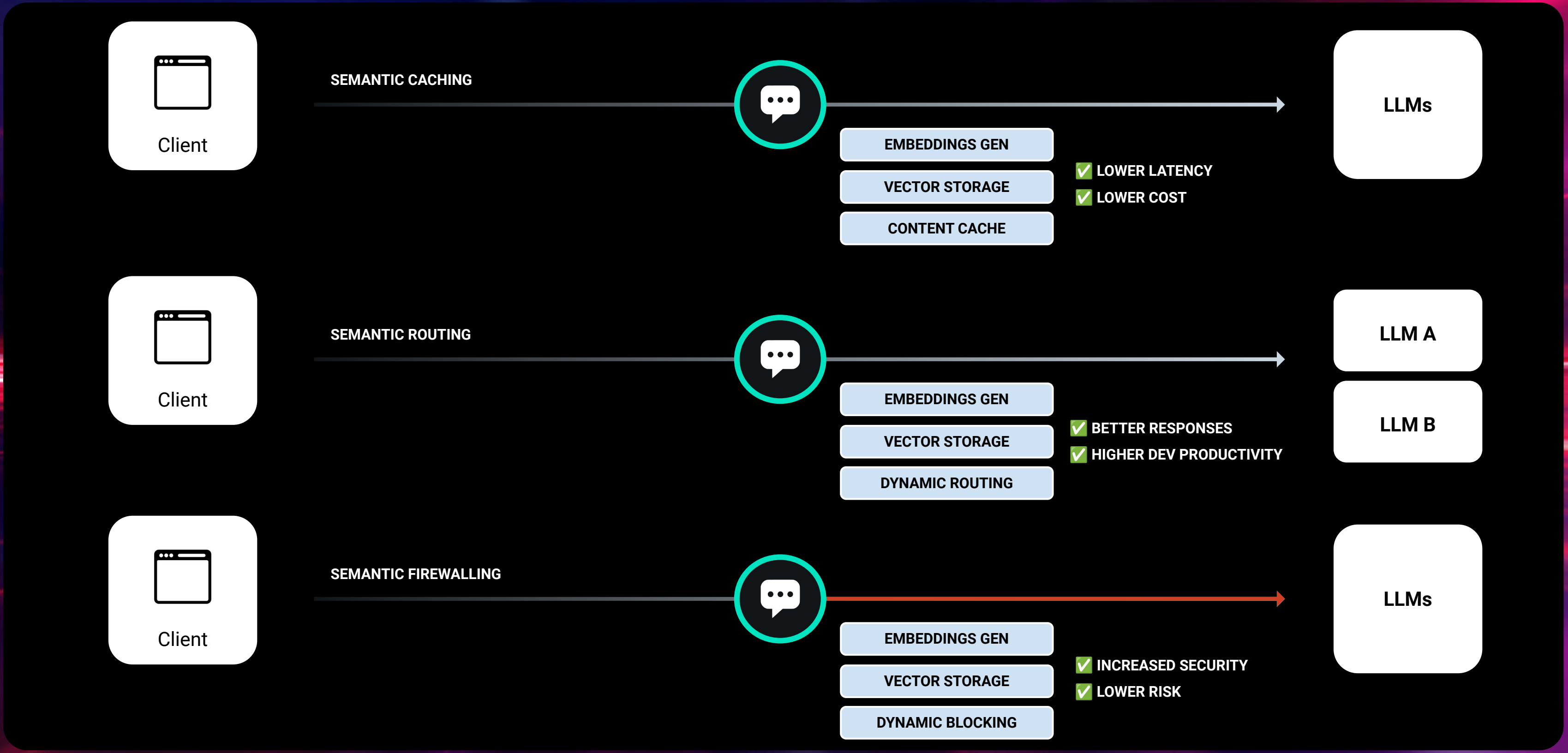
DYNAMIC ROUTING

✅ BETTER RESPONSES

✅ HIGHER DEV PRODUCTIVITY

LLM A

LLM B



Client

SEMANTIC CACHING

EMBEDDINGS GEN

VECTOR STORAGE

CONTENT CACHE

✓ LOWER LATENCY

✓ LOWER COST

LLMs

Client

SEMANTIC ROUTING

EMBEDDINGS GEN

VECTOR STORAGE

DYNAMIC ROUTING

✓ BETTER RESPONSES

✓ HIGHER DEV PRODUCTIVITY

LLM A

LLM B

Client

SEMANTIC FIREWALLING

EMBEDDINGS GEN

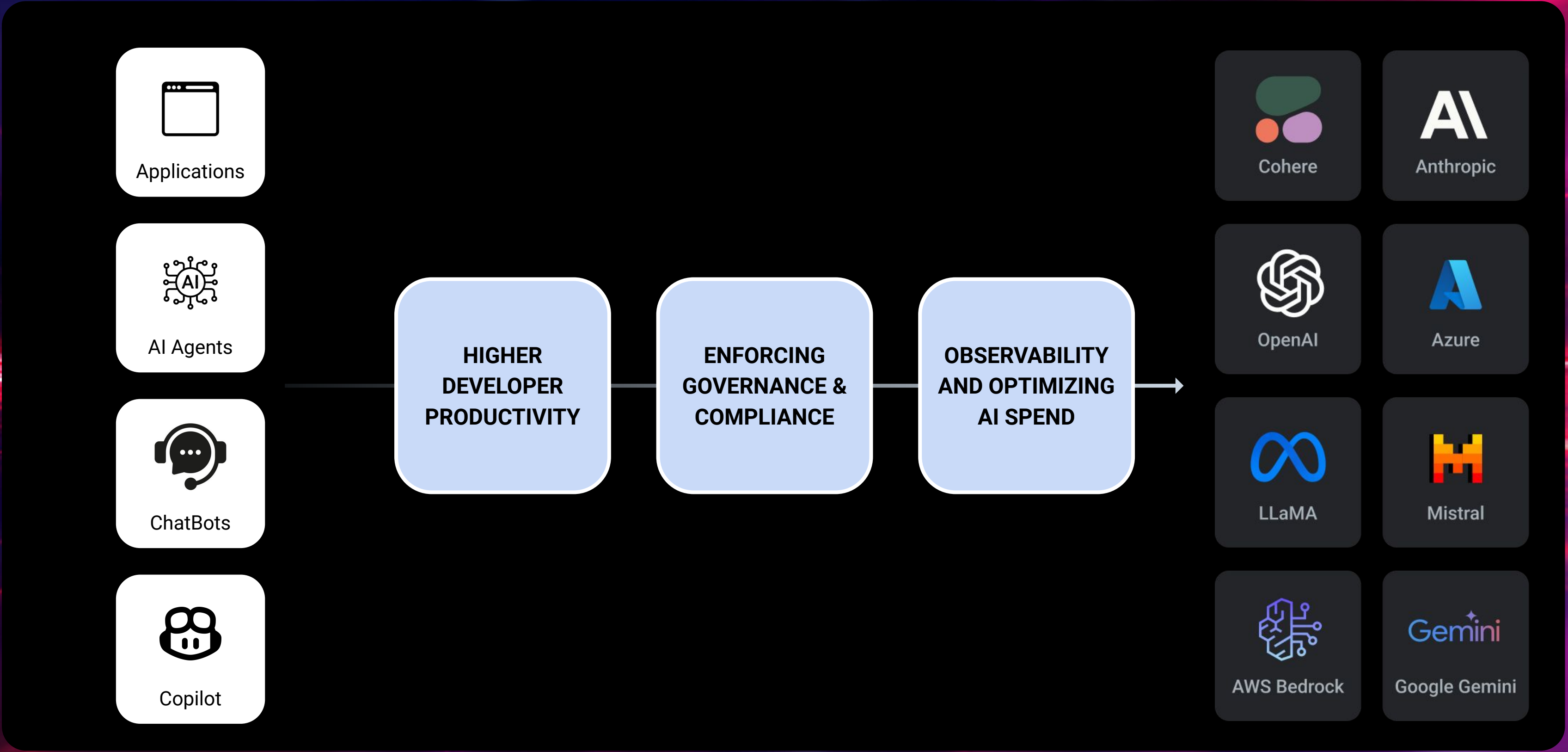
VECTOR STORAGE

DYNAMIC BLOCKING

✓ INCREASED SECURITY

✓ LOWER RISK

LLMs



Applications

AI Agents

ChatBots

Copilot

HIGHER DEVELOPER PRODUCTIVITY

ENFORCING GOVERNANCE & COMPLIANCE

OBSERVABILITY AND OPTIMIZING AI SPEND

Cohere

Anthropic

OpenAI

Azure

LLaMA

Mistral

AWS Bedrock

Google Gemini

May 2024 (3.7)

- STREAMING SUPPORT
- AI TOKEN RATE LIMITING
- AZURE CONTENT SAFETY
- URL-SOURCED ROUTING
- CLAUDE 2.1 MESSAGES
- OPENAI SDK COMPATIBILITY

June 2024 (3.7.1)

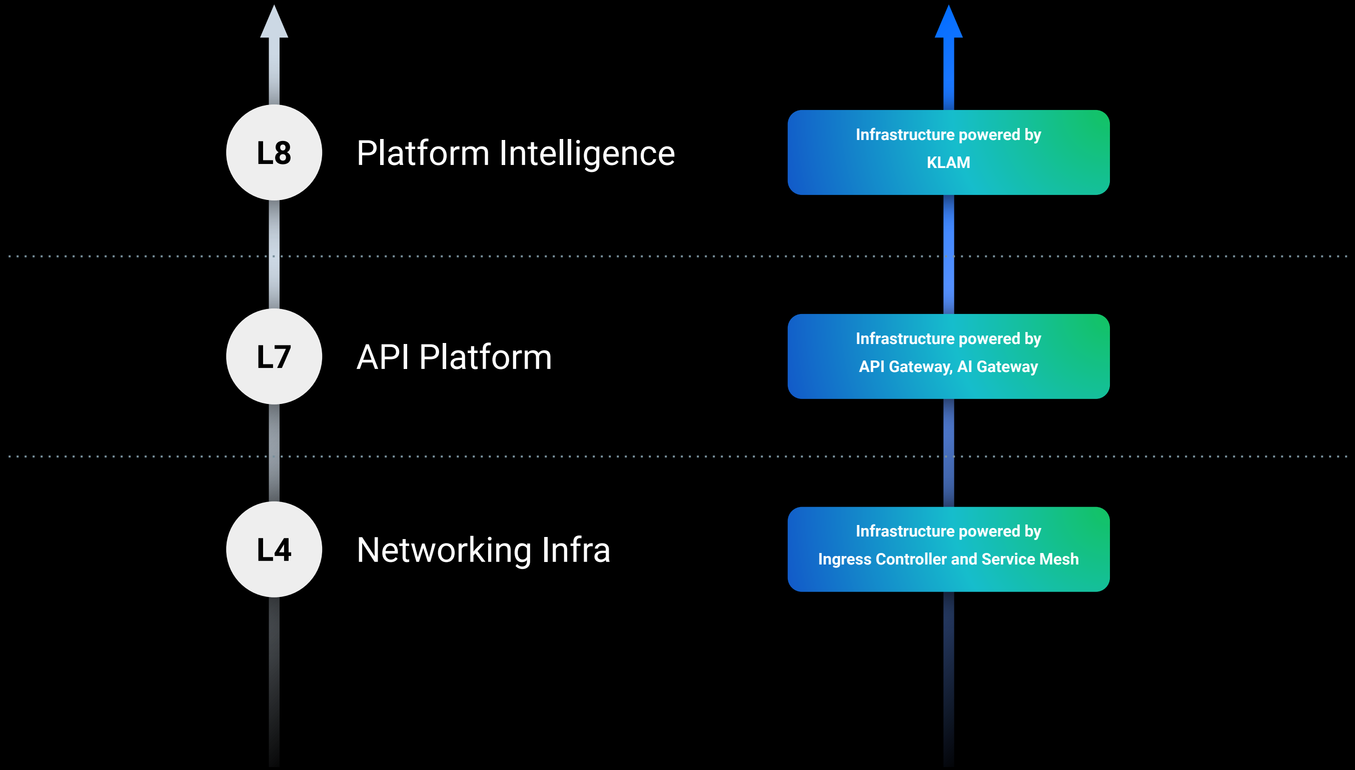
- GCP VERTEX LLM SUPPORT
- AWS BEDROCK LLM SUPPORT
- ADVANCED LOAD BALANCING

Q3-Q4 (3.8+)

- SEMANTIC CACHING
- SEMANTIC PROMPT GUARD
- SEMANTIC ROUTING

Semantic AI Plugins (Aug 2024)

- PROMPT JAILBREAKING SEC
- HALLUCINATION MITIG.
- PROMPT INJECTION SEC
- PROFANITY MITIG.
- AI TRAIN PLUGIN
- SEMANTIC AI RAG



L8 Platform Intelligence

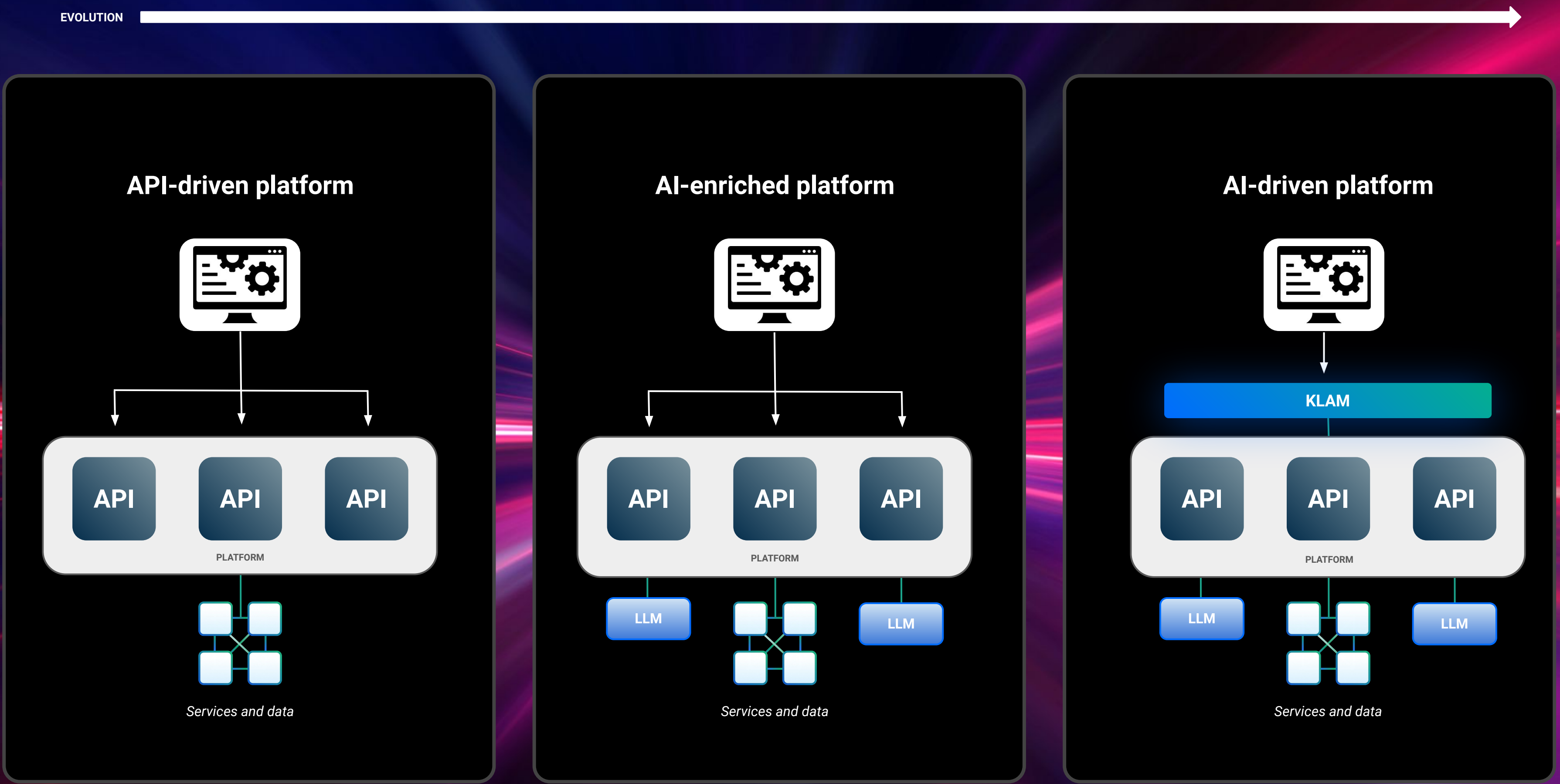
Infrastructure powered by KLAM

L7 API Platform

Infrastructure powered by API Gateway, AI Gateway

L4 Networking Infra

Infrastructure powered by Ingress Controller and Service Mesh


EVOLUTION
API-driven platform
API
API
API
PLATFORM
Services and data
AI-enriched platform
API
API
API
PLATFORM
LLM
LLM
Services and data
AI-driven platform
KLAM
API
API
API
PLATFORM
LLM
LLM
Services and data

Q&A

Kong AI Gateway

konghq.com