PyCon China 2025

张晋涛 - Kong Inc.

2025 年 9 月 20 日

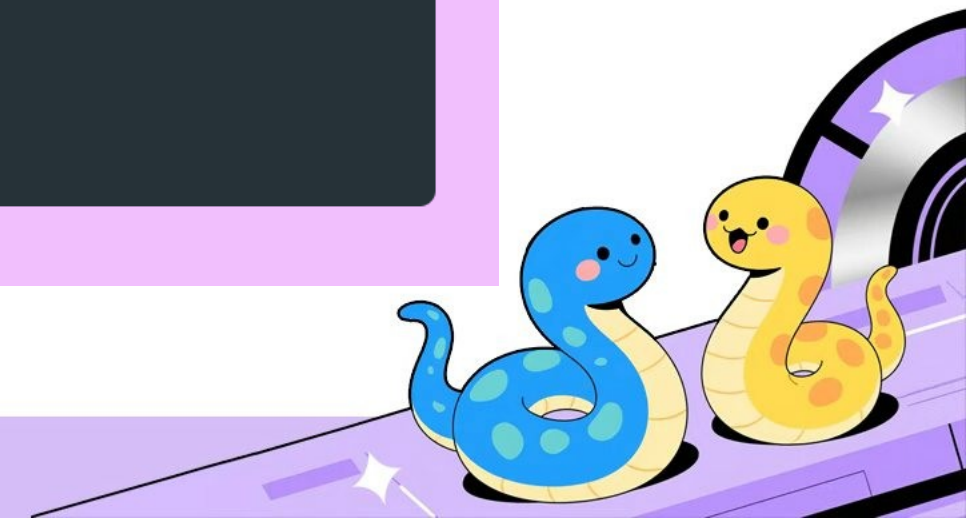
AI Gateway
在 AI 应用中是否有价值

个人介绍

- Kong Inc.
- Microsoft MVP
- CNCF Ambassador
- GitHub: https://github.com/tao12345666333
- X: https://x.com/zhangjintao9020

AI 应用

```
curl https://api.openai.com/v1/responses \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -d '{
    "model": "gpt-5",
    "input": "Write a short bedtime story about a unicorn."
  }'
```

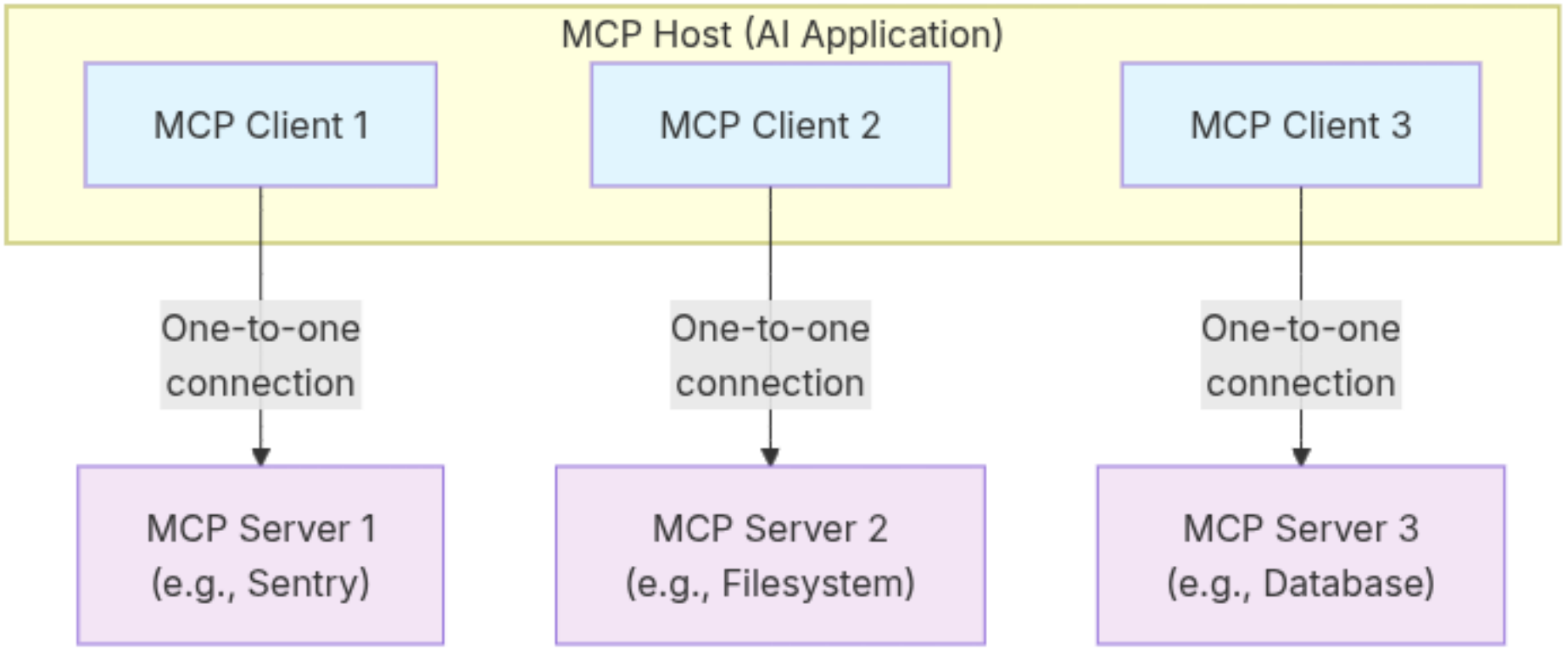


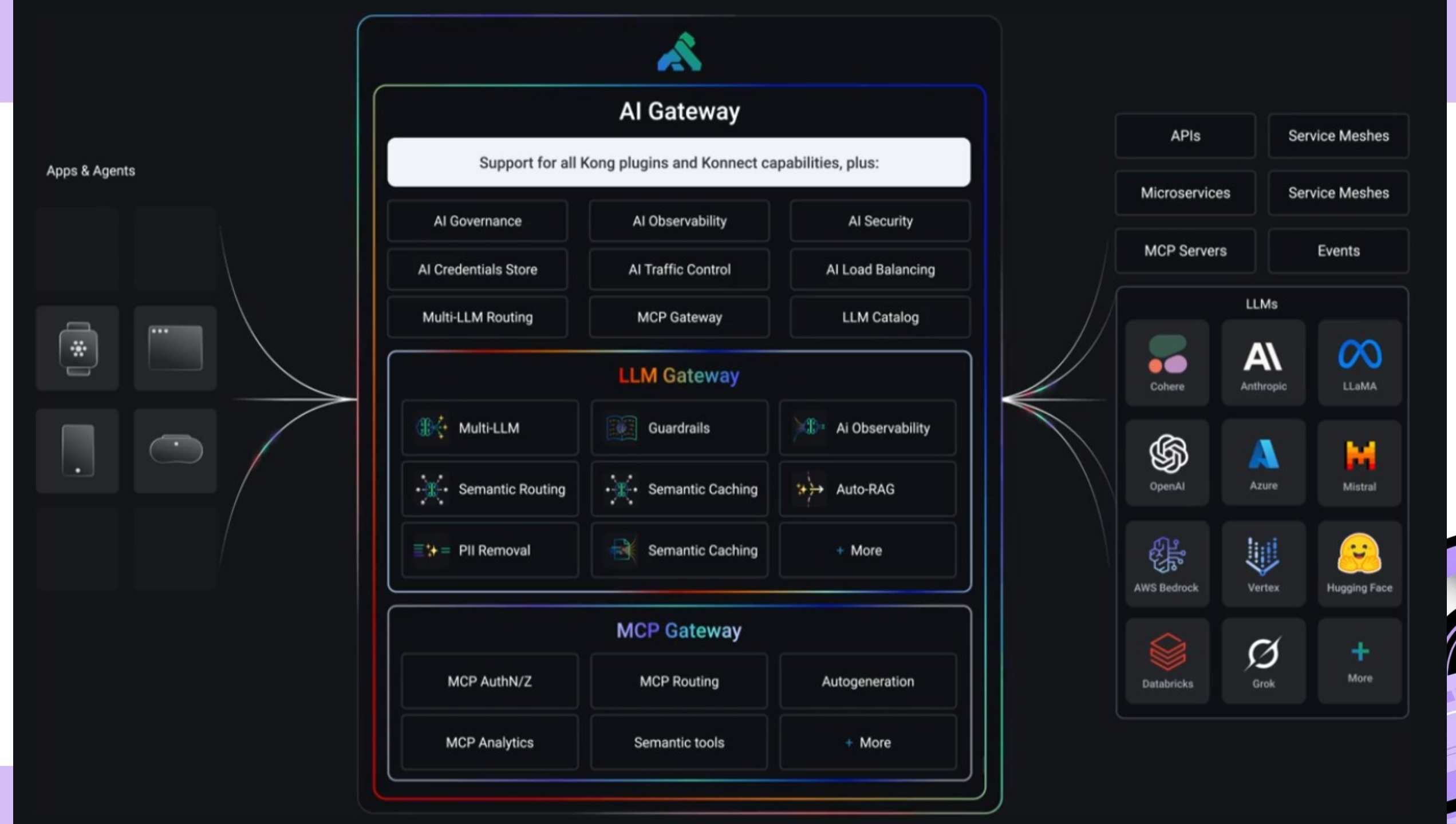
AI 应用

```python
from openai import OpenAI
client = OpenAI()

response = client.responses.create(
    model="gpt-5",
    input="Write a short bedtime story about a unicorn."
)

print(response.output_text)
```

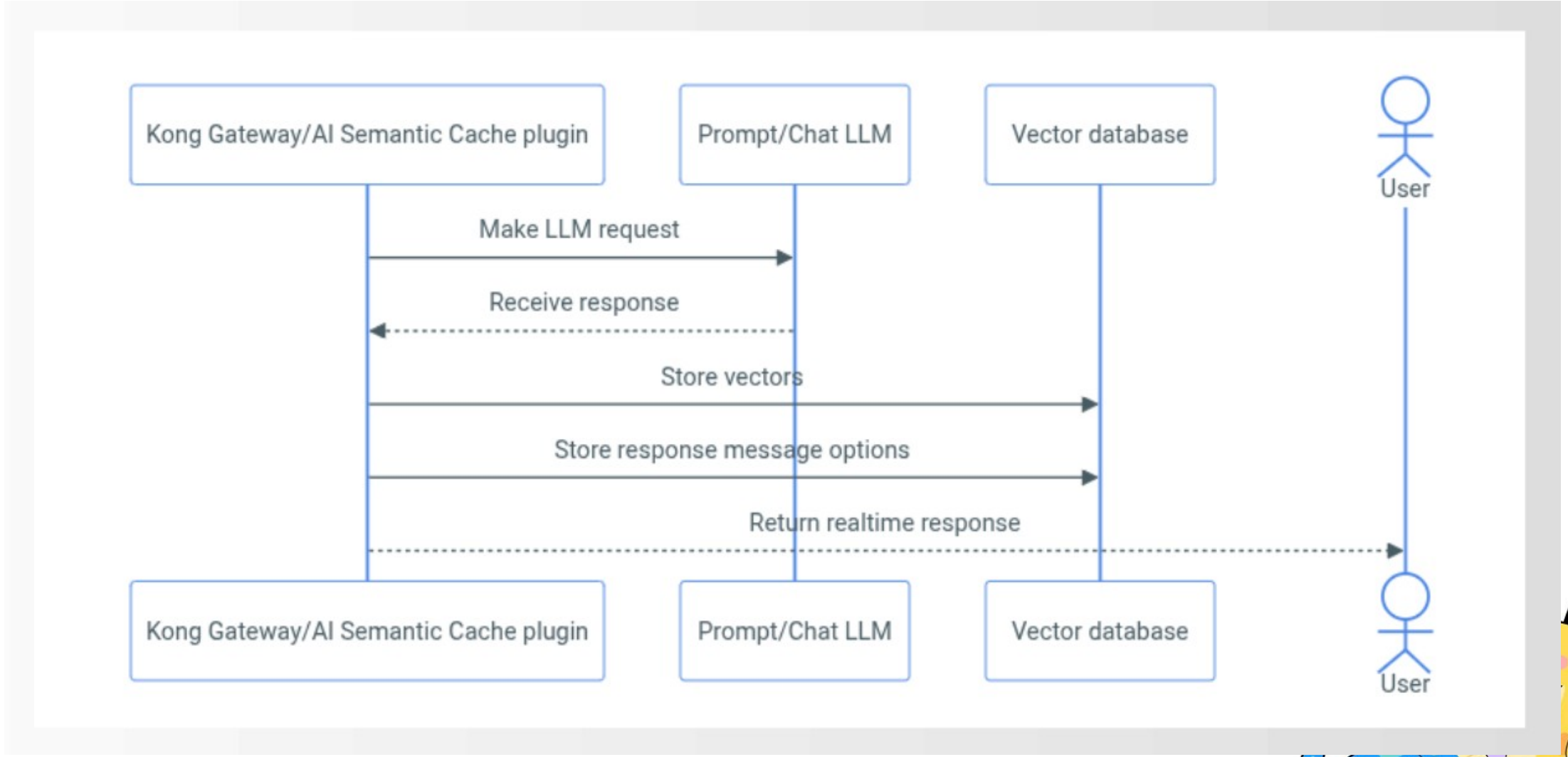
```python
    def run(self) -> int:
        conversation: List[Message] = []
        print("Chat with GPT (use 'ctrl-c' to quit)")

        while True:
            user_input, ok = self.get_user_message()
            if not ok:
                break

            # Append user message to the conversation
            conversation.append(
                {"role": "user", "content": [{"type": "text", "text": user_input}]}
            )

            # Inference
            assistant_text = self.run_inference(conversation)

            # Append assistant reply to the conversation
            conversation.append(
                {"role": "assistant", "content": [{"type": "text", "text": assistant_text}]}
            )

            # Print assistant's text (yellow)
            if assistant_text:
                print(f"\u001b[93mGPT\u001b[0m: {assistant_text}")

        return 0


def main() -> int:
    try:
        client = OpenAI()  # Reads OPENAI_API_KEY from environment
        agent = Agent(client, model="gpt-5", max_tokens=1024)
        return agent.run()
    except Exception as e:
        print(f"Error: {e}", file=sys.stderr)
        return 1
```

MCP

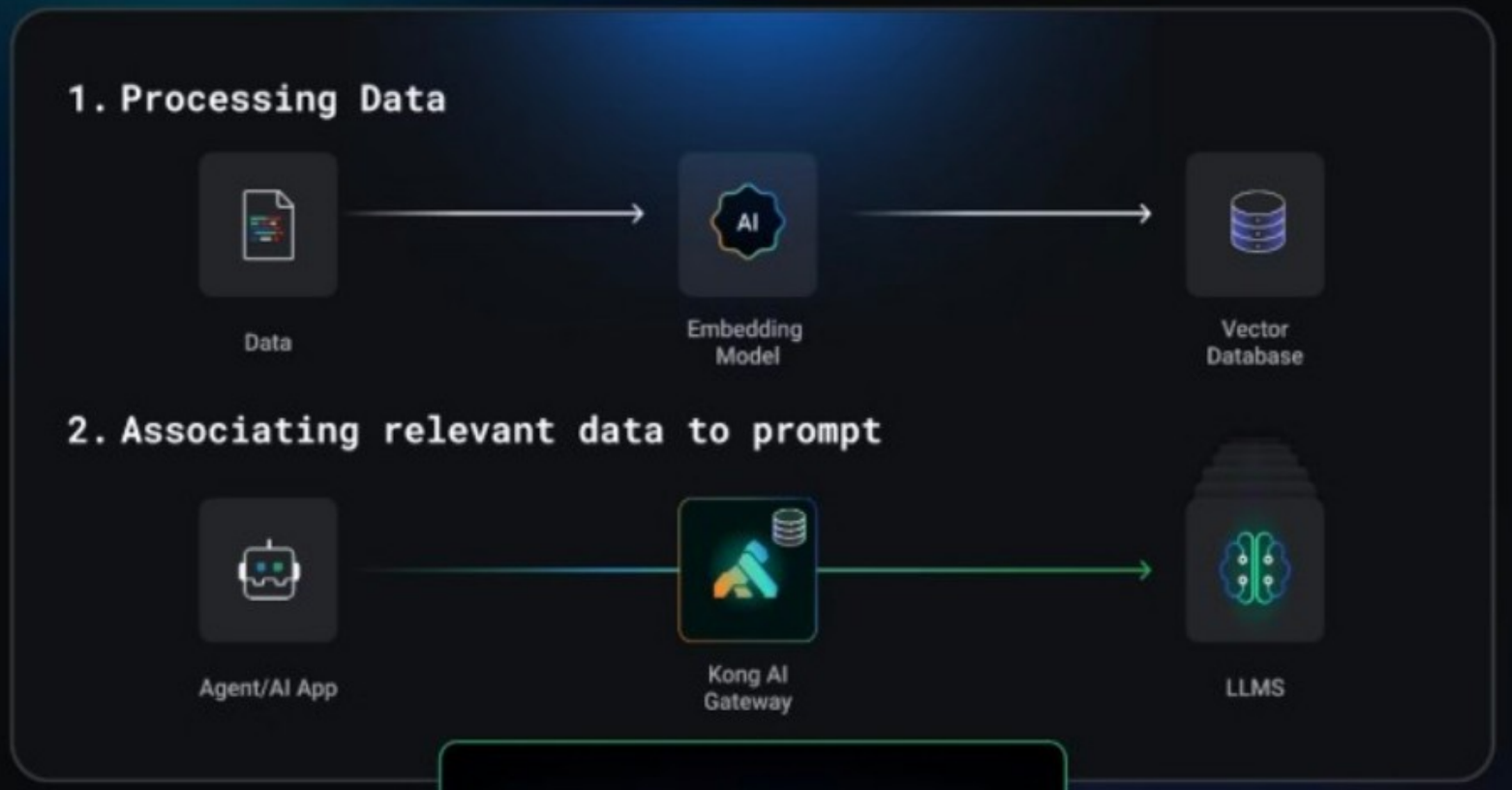

MCP Host (AI Application)
MCP Client 1
MCP Client 2
MCP Client 3
One-to-one connection
One-to-one connection
One-to-one connection
MCP Server 1 (e.g., Sentry)
MCP Server 2 (e.g., Filesystem)
MCP Server 3 (e.g., Database)


Apps & Agents
AI Gateway
Support for all Kong plugins and Konnect capabilities, plus:
AI Governance
AI Observability
AI Security
AI Credentials Store
AI Traffic Control
AI Load Balancing
Multi-LLM Routing
MCP Gateway
LLM Catalog
LLM Gateway
Multi-LLM
Guardrails
Ai Observability
Semantic Routing
Semantic Caching
Auto-RAG
PII Removal
Semantic Caching
+ More
MCP Gateway
MCP AuthN/Z
MCP Routing
Autogeneration
MCP Analytics
Semantic tools
+ More
APIs
Service Meshes
Microservices
Service Meshes
MCP Servers
Events
LLMs
Cohere
Anthropic
LLaMA
OpenAI
Azure
Mistral
AWS Bedrock
Vertex
Hugging Face
Databricks
Grok
More

Semantic Cache



Automated RAG with Kong



- Lower LLM hallucinations
- Improve dev productivity
- Grow secure

Thanks!