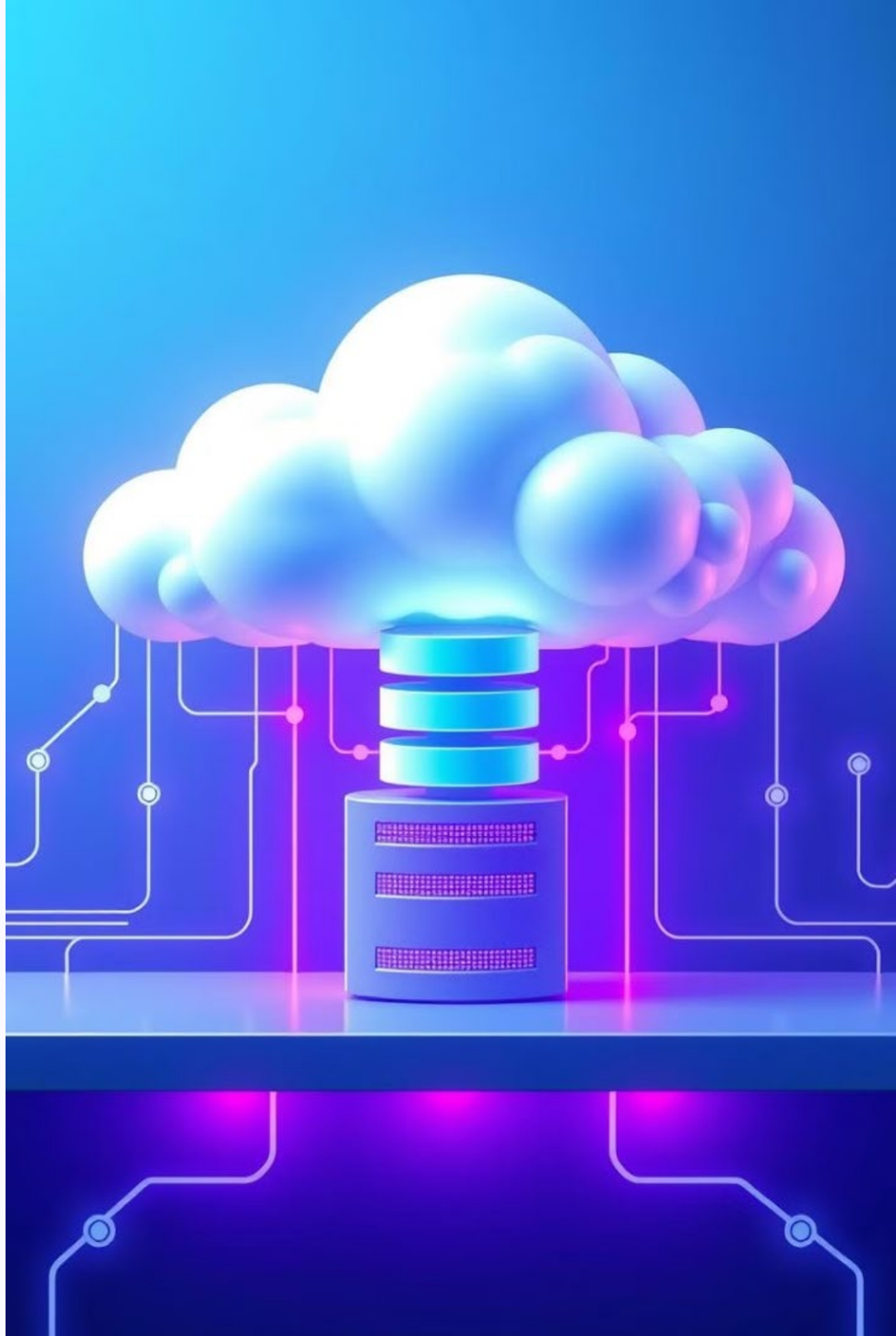
# From API Gateway to AI Gateway!

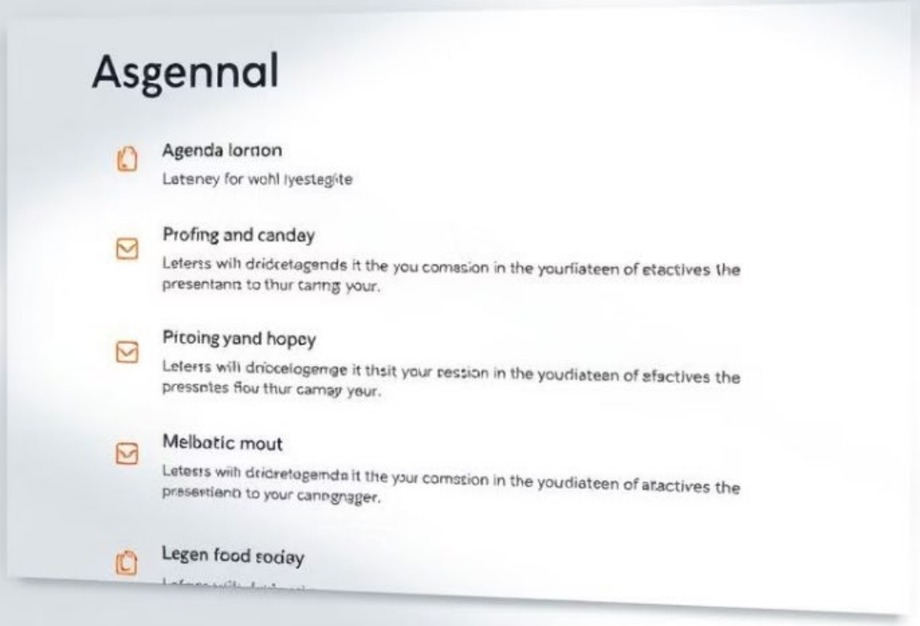


June 2026

by Dan Erez

# Agenda Overview

- The API Gateway
- Enter AI Applications
- The AI Gateway



# About Me

1. ## Lead Architect

   25 years of software development experience.
   Mother tongue: Java

2. ## Expertise

   Specializing in Enterprise & Cloud Architecture,
   Nowadays focusing in bringing AI tools to the SDLC.

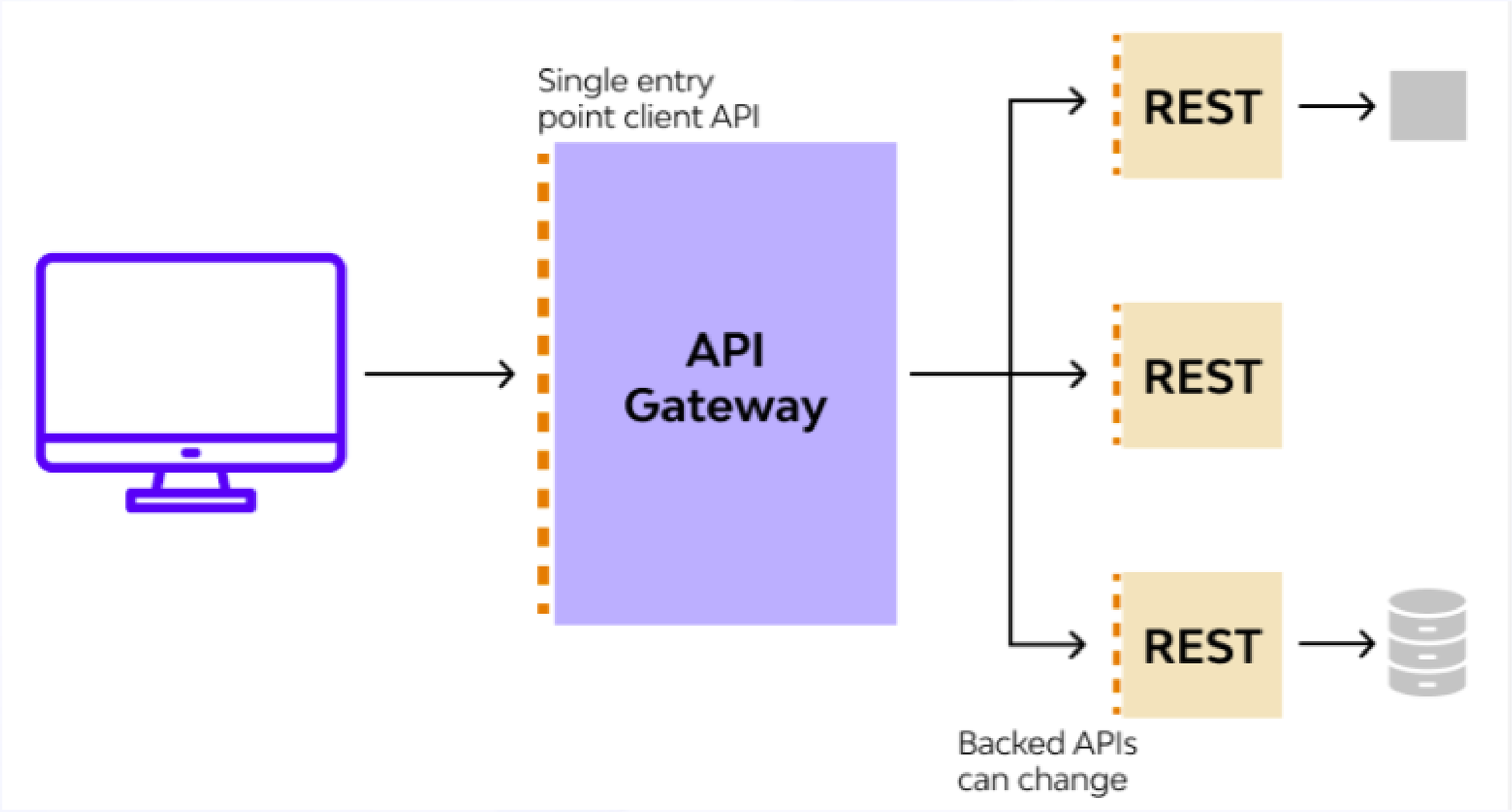
3. ## Ideas

   Speaking my mind at conferences worldwide.
   Also writing at Medium under the user dan.erez.

# The beginning…

- Once there was the monolith
- Problems, problems …
- Here come the micro services
- Problems, problems …

# The API Gateway



# The API Gateway

- Routing
- Load balancing
- Authentication
- Headers management
- Logging
- Much more!

# API Gateways out there

- Kong
- Nginx Plus
- Apigee
- IBM API Connect
- Cloud services (e.g. Amazon API Gateway)

# Enter AI Applications

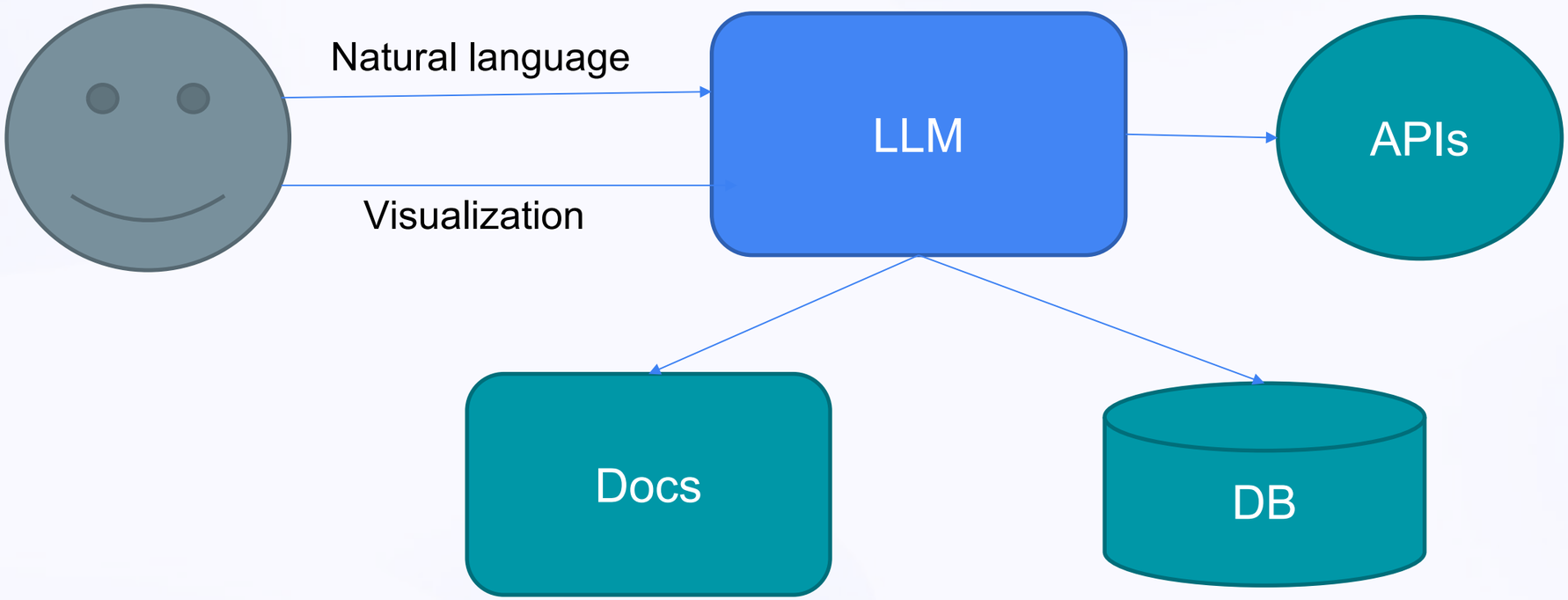
# Who needs the back end?

- You can get the business knowledge:
  - RAG
  - MCP, Tools, APIs
  - Unstructured data
  - Text 2SQL
- No more business + DB layers!

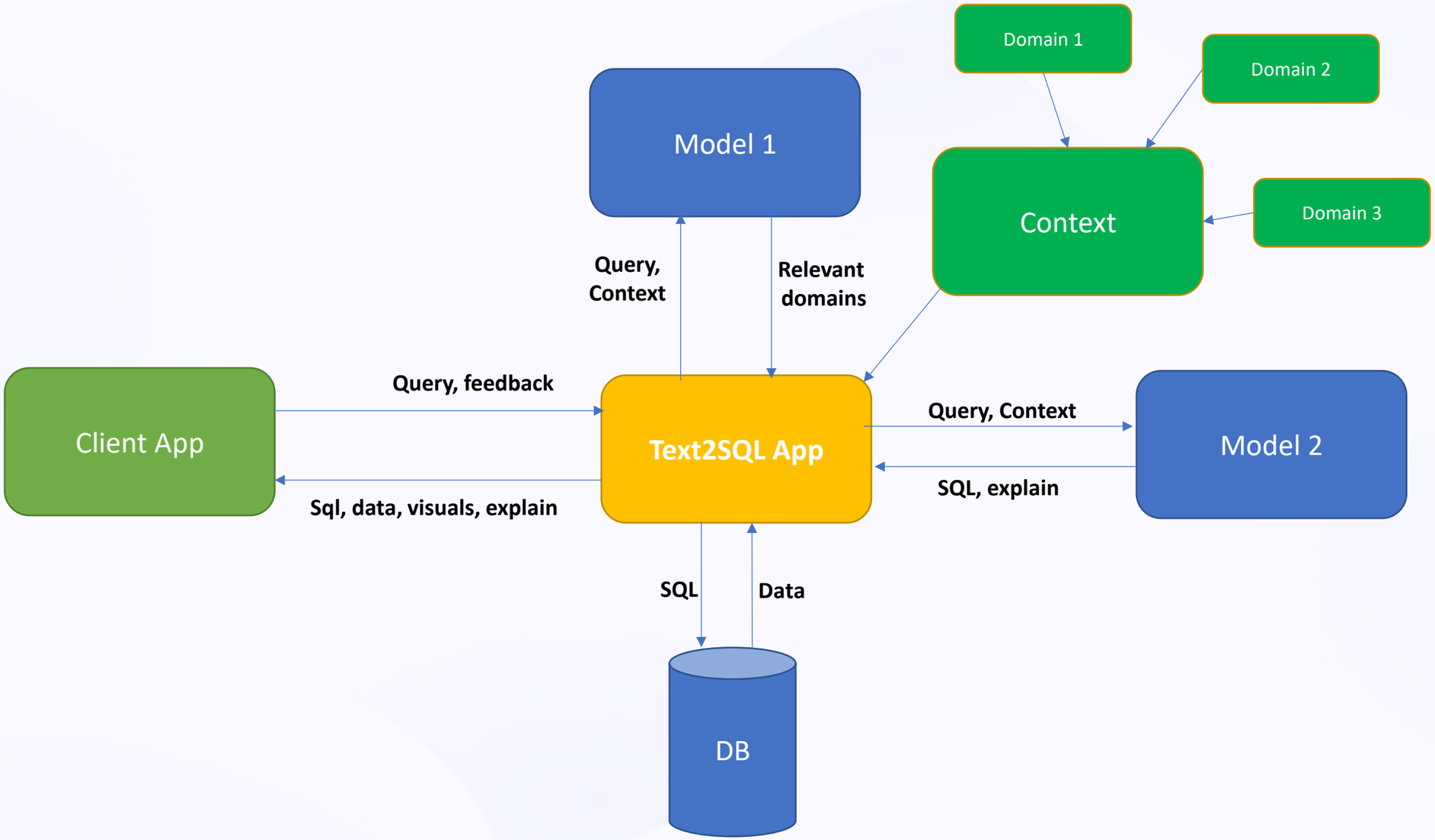
# Who needs the front end?

- Dashboards and pre-made UI are so 2025…
- MCP-Apps, A2UI & other protocols
- The death of the web applications

# The application of the future (present?)



# Text2SQL


Domain 1
Domain 2
Domain 3
Model 1
Context
Query,
Context
Relevant
domains
Client App
Query, feedback
Sql, data, visuals, explain
Text2SQL App
Query, Context
SQL, explain
Model 2
SQL
Data
DB

# Prepare to pay…

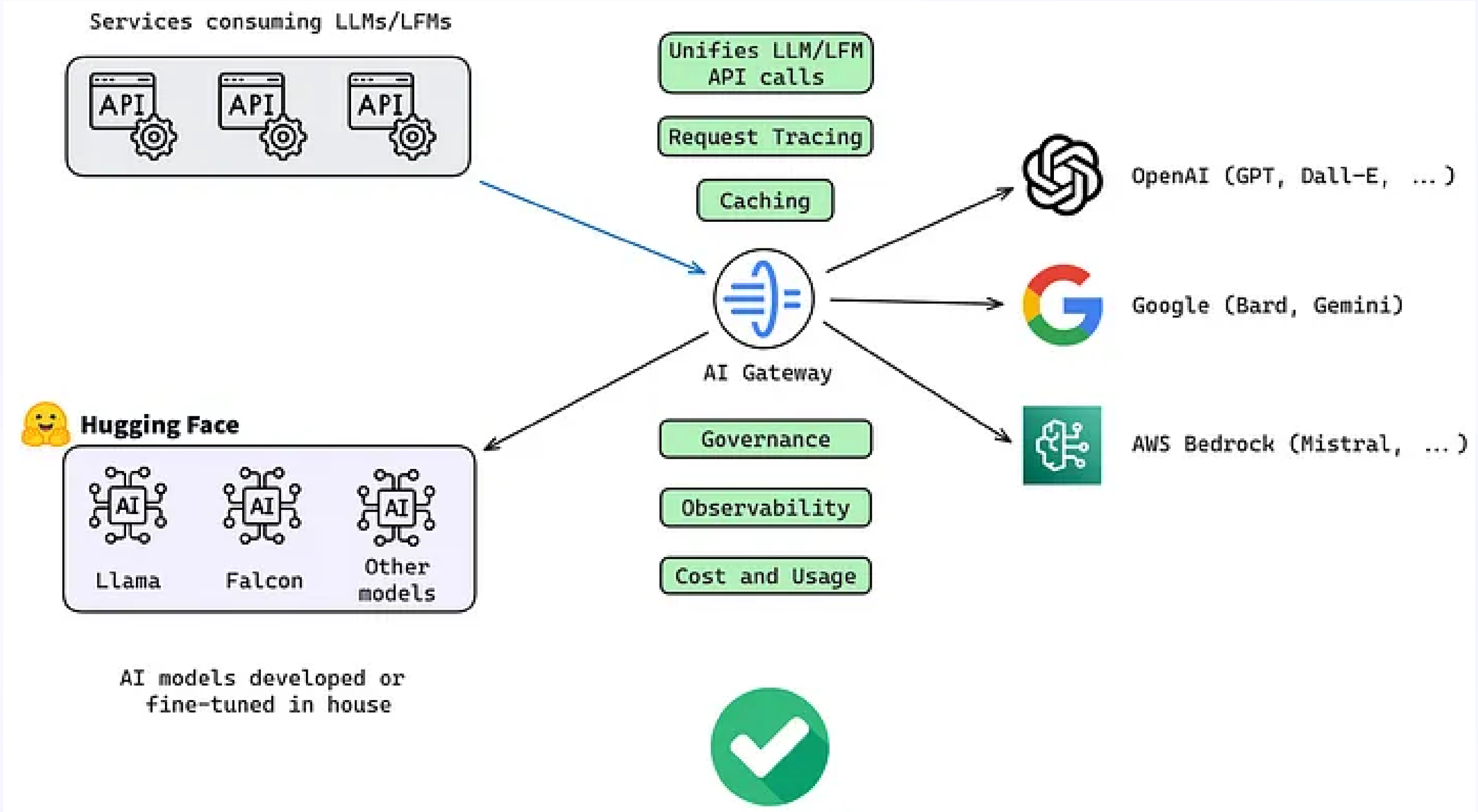
# Money Saving by model routing

- Opus 4.6 costs 5$ / Mtok , 25$ output
- Haiku 3 costs 0.25$ / Mtok , 1.25$ output
- 10K users, 50 requests/day, 10KT In, 1K out, half are simple
- Saving : 16875$ per day!!!

# New Apps –New Gateway

# AI Gateway

- Route to models
- Track quotas
- Apply security and guardrails
- Semantic caching
- Format questions/answers
- Can also be an API gateway!
- Much more …


Services consuming LLMs/LFMs
API
API
API
Unifies LLM/LFM API calls
Request Tracing
Caching
AI Gateway
Governance
Observability
Cost and Usage
OpenAI (GPT, Dall-E, ...)
Google (Bard, Gemini)
AWS Bedrock (Mistral, ...)
Hugging Face
AI
AI
AI
Llama
Falcon
Other models
AI models developed or fine-tuned in house

# What ’s out there?

- Open Router
- Portkey AI Gateway
- Cloudflare
- LiteLLM (open source)
- LocalAI
- Many more …

# Portkey AI Gateway

- Universal API
- Fallbacks
- Cache (Simple and Semantic)
- Retries
- Load Balancing

# Building your own

- Free!
- For example, in Java/Spring Boot:
  - Built on top of existing Gateway
  - Customizable for your needs
- Why not?

# Key Takeaways

- You'll need an AI gateway
- Lots of commercial options
- Buidling your own

# Thank You "

**Too shy to ask here?**
dan.erez@gmail.com
Feel free to consult!