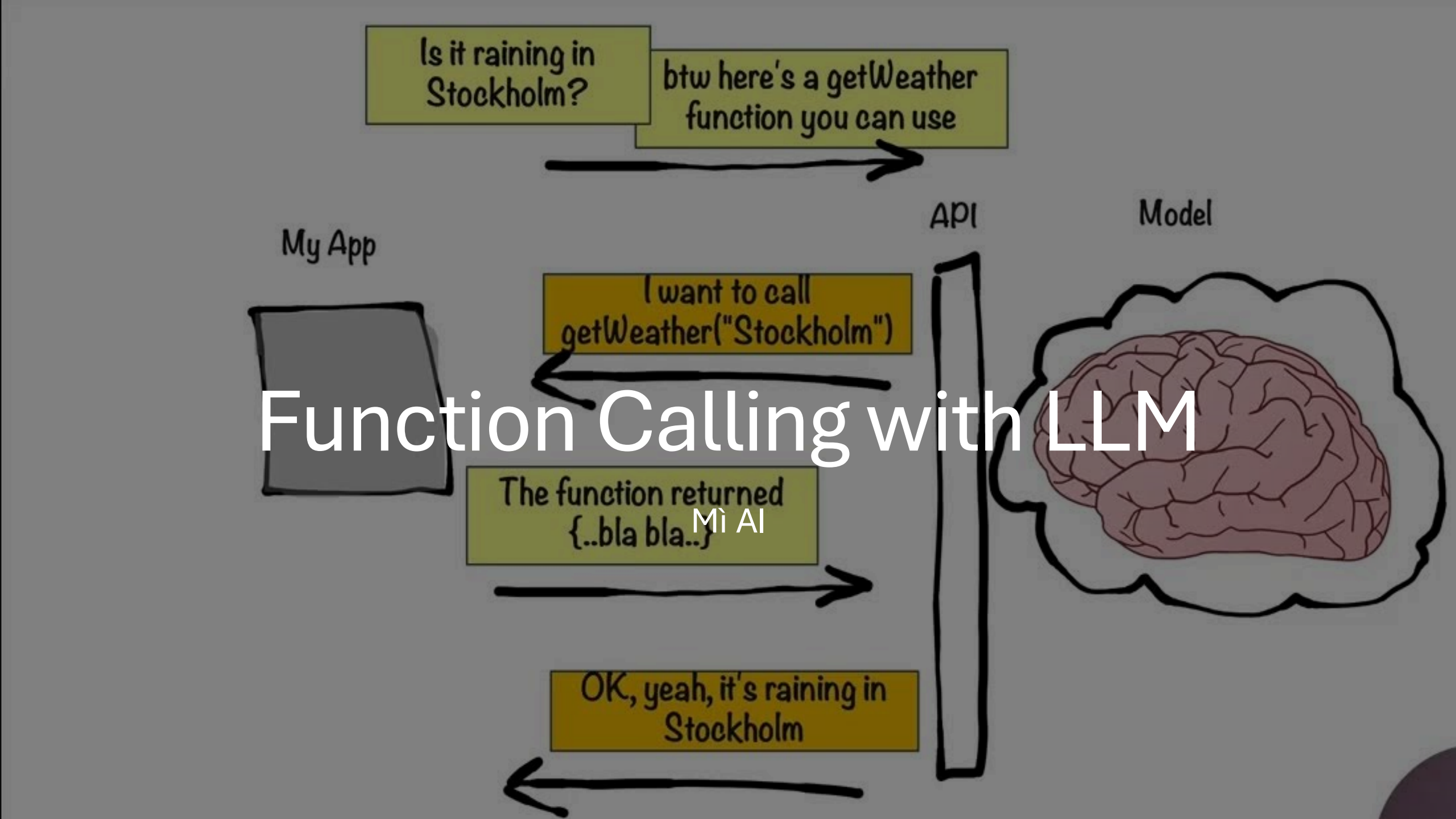
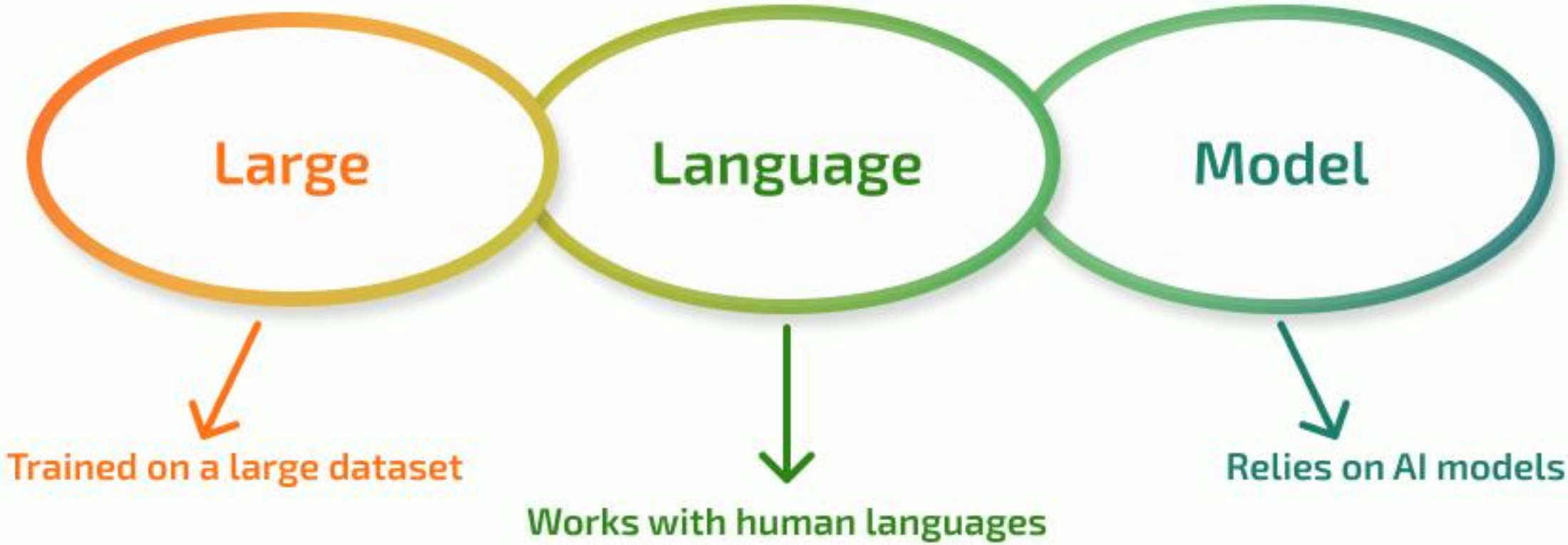
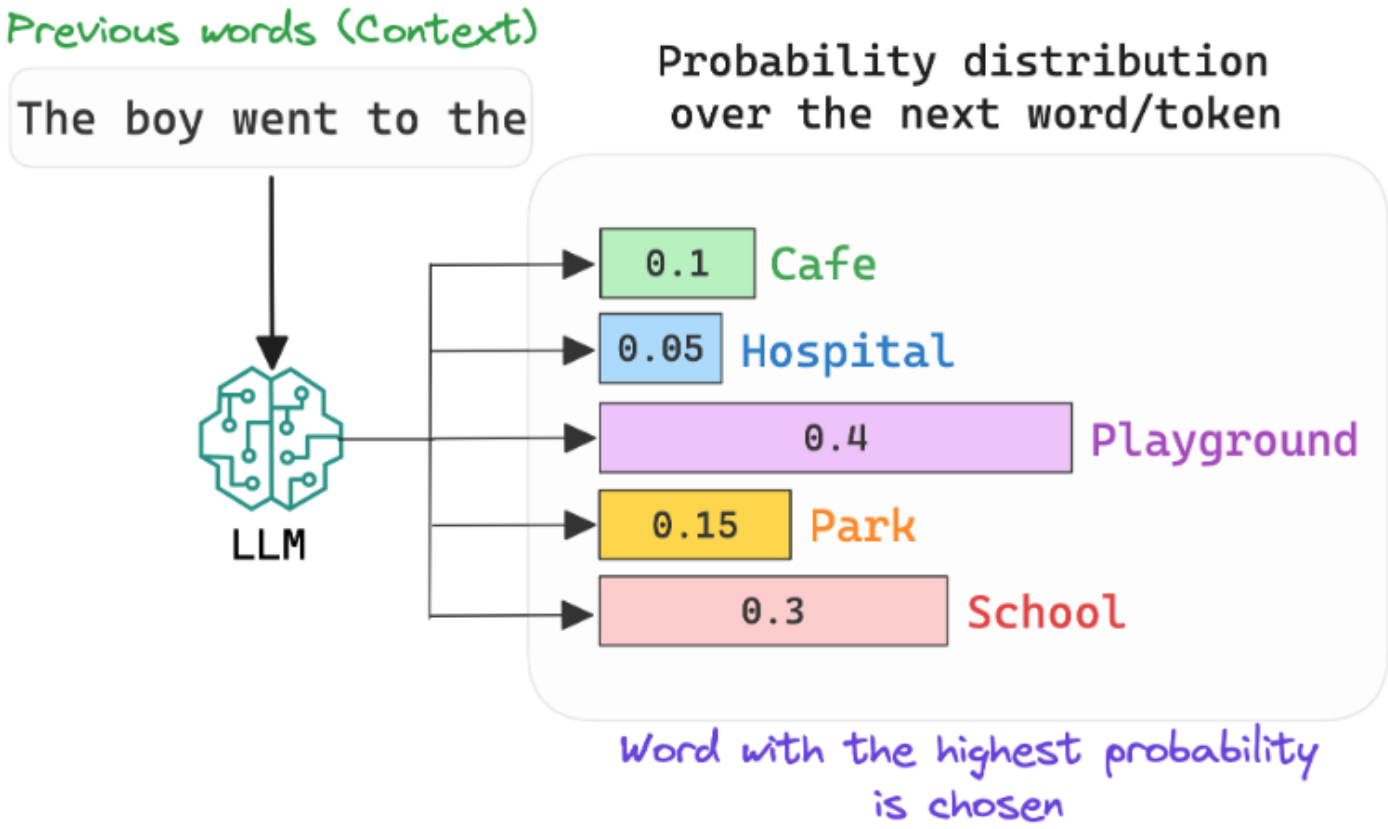


Function Calling with LLM

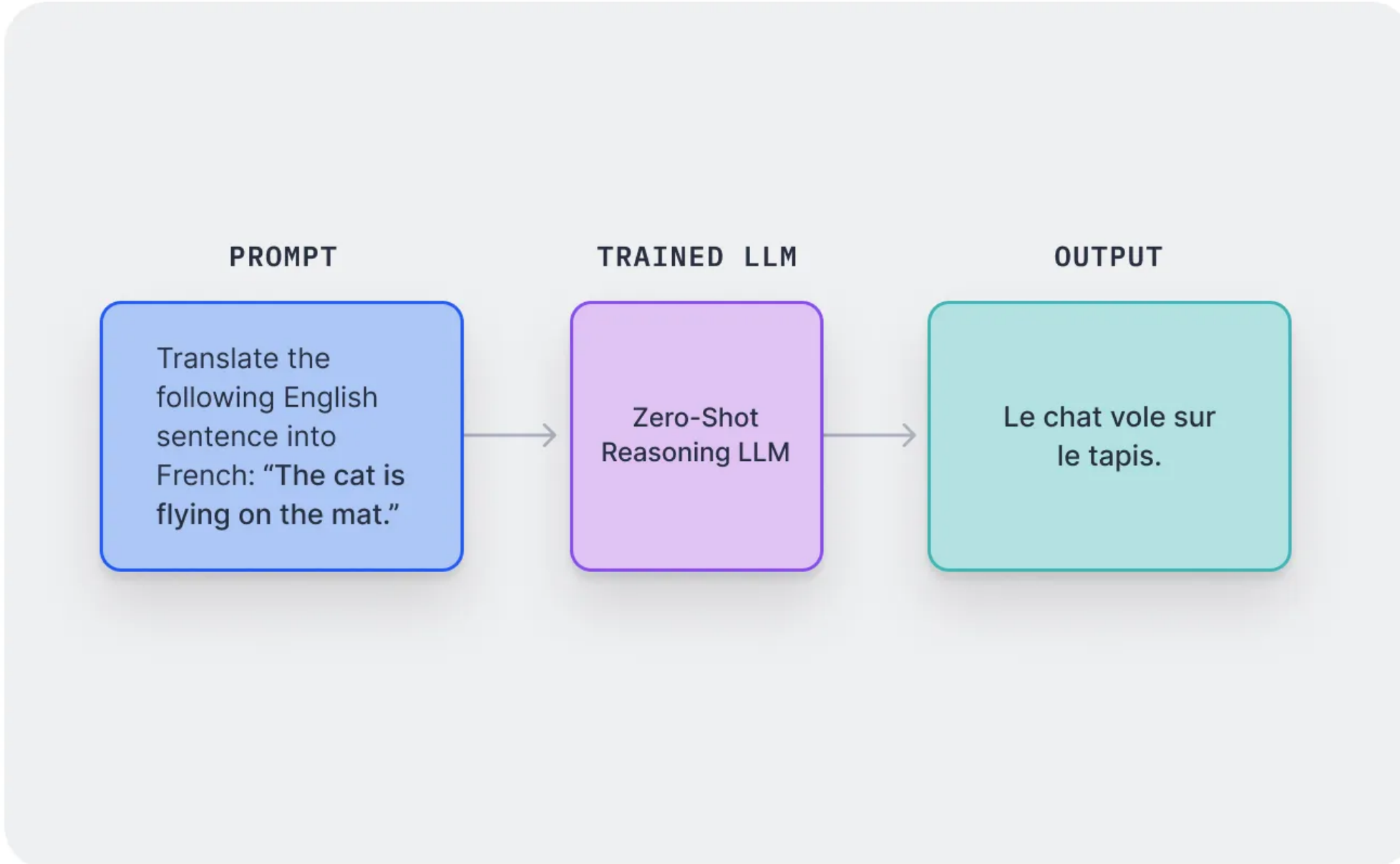
Mì AI

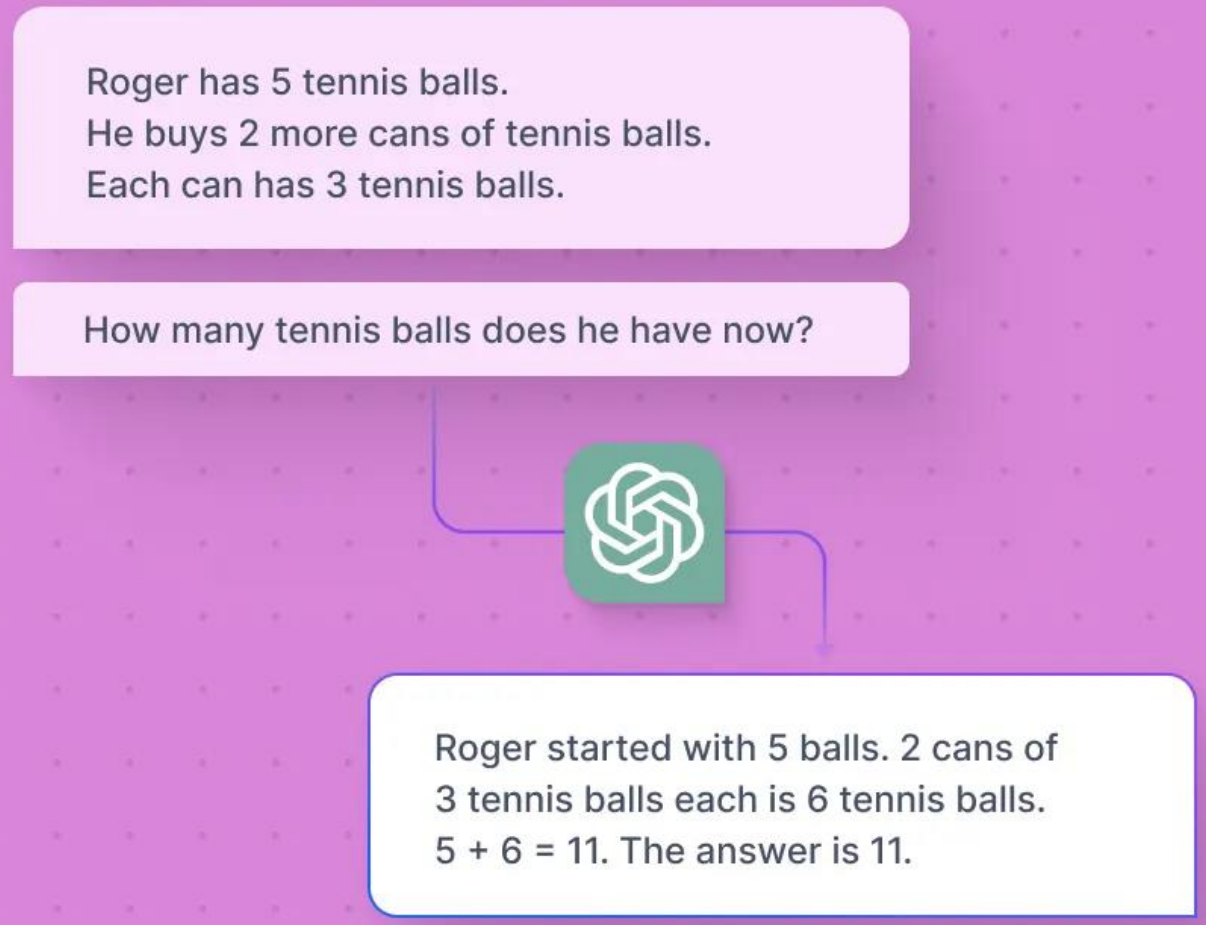


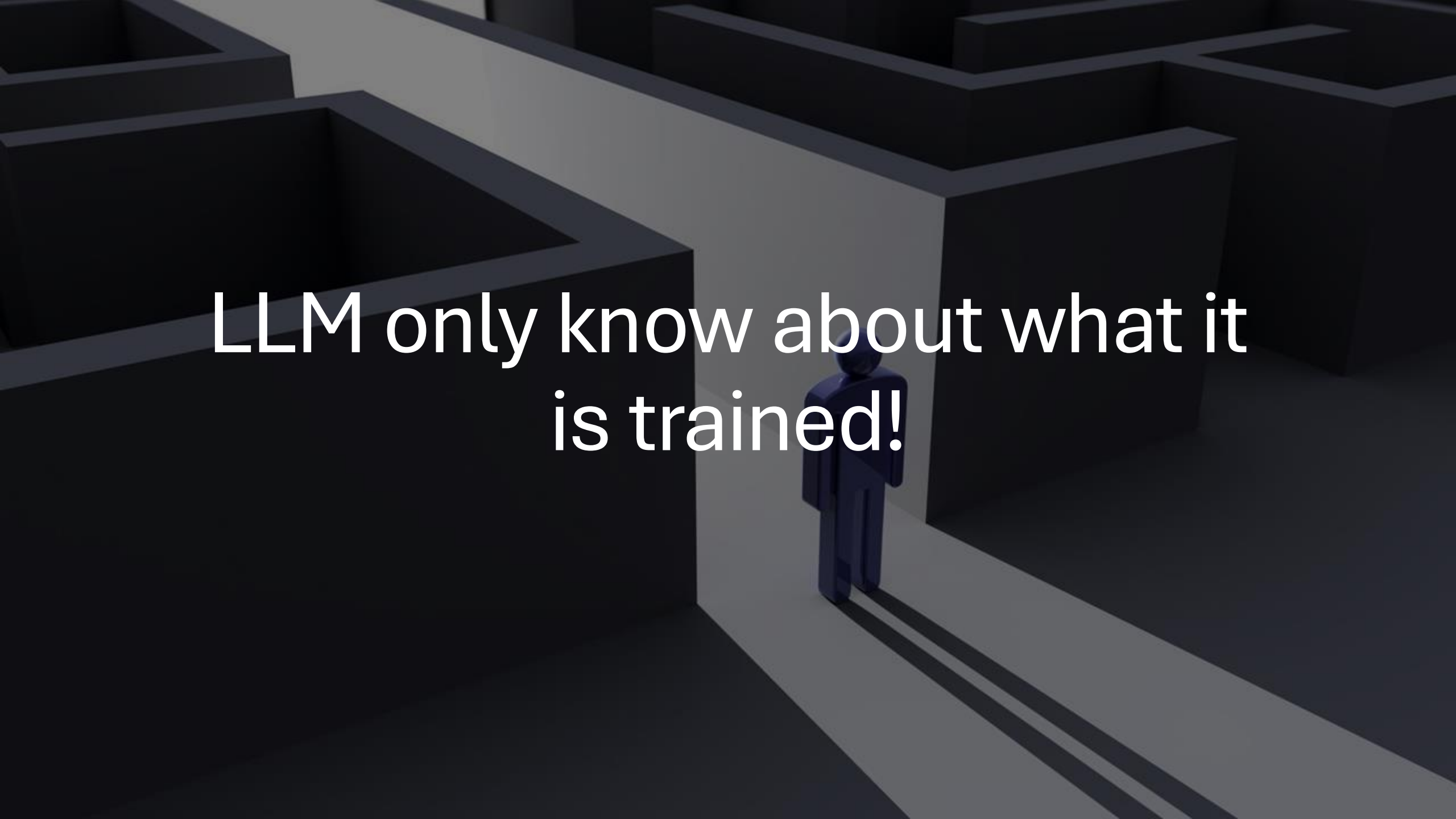
Next word prediction!



@akshay_pachaar


PROMPT
TRAINED LLM
OUTPUT
Translate the following English sentence into French: “The cat is flying on the mat.”
Zero-Shot Reasoning LLM
Le chat vole sur le tapis.


Roger has 5 tennis balls.
He buys 2 more cans of tennis balls.
Each can has 3 tennis balls.
How many tennis balls does he have now?
Roger started with 5 balls. 2 cans of
3 tennis balls each is 6 tennis balls.
5 + 6 = 11. The answer is 11.


LLM only know about what it
is trained!

User Query

I wonder if there were any mentions of expansion plans in Asia last year?

Input Query Embedding
[0.334, 0.231, 0.729 …]

Document segment embedding 4
Document segment embedding 3
Document segment embedding 2
Document segment embedding 1
[0.185, 0.648, 0.127, 0.792 …]

Vector Database

Retriever

"Sleuth"

Prompt

Given the following context, answer the following question about Company A's filings with the Securities Exchange Commission (SEC)…

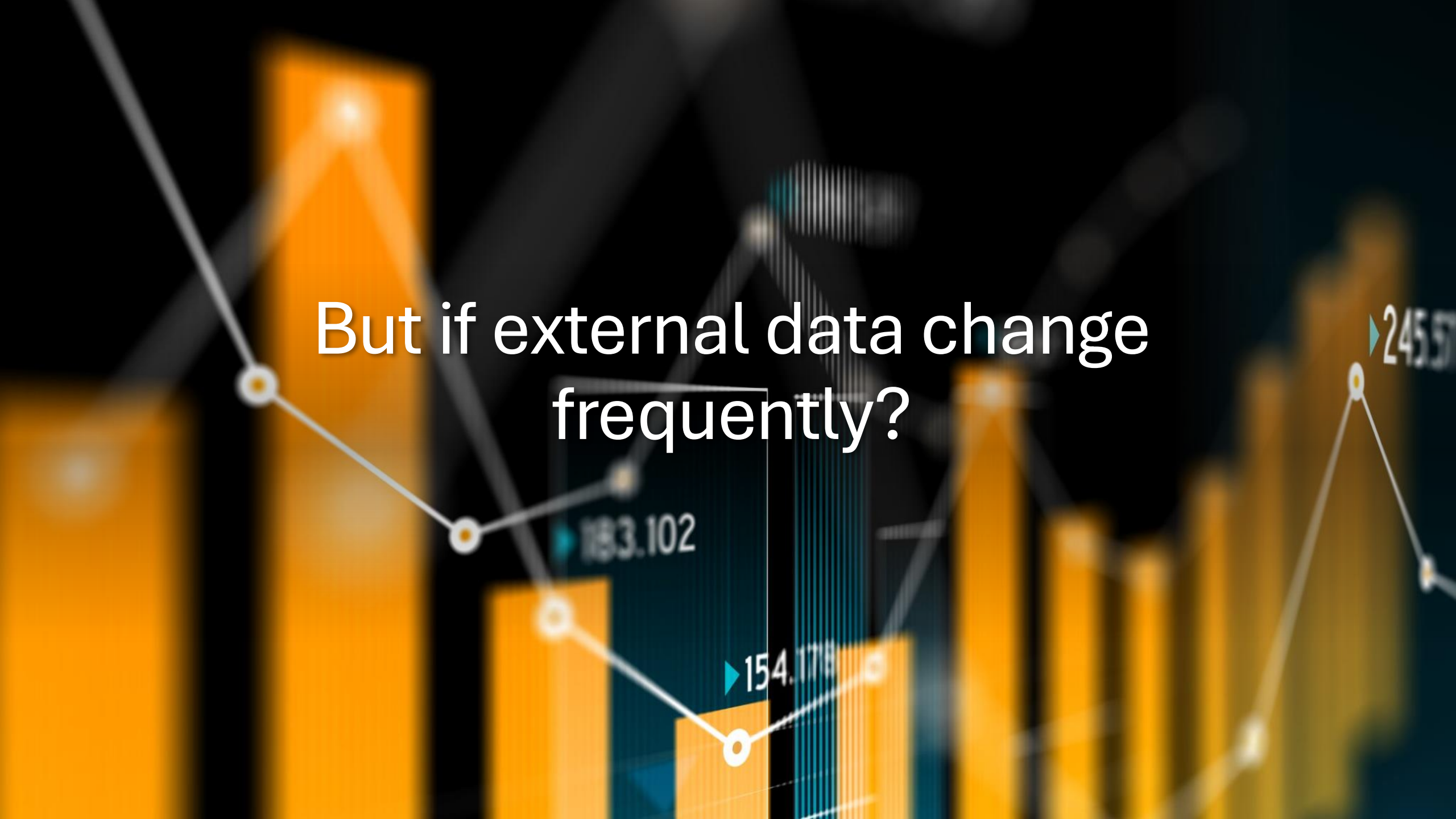
Context:

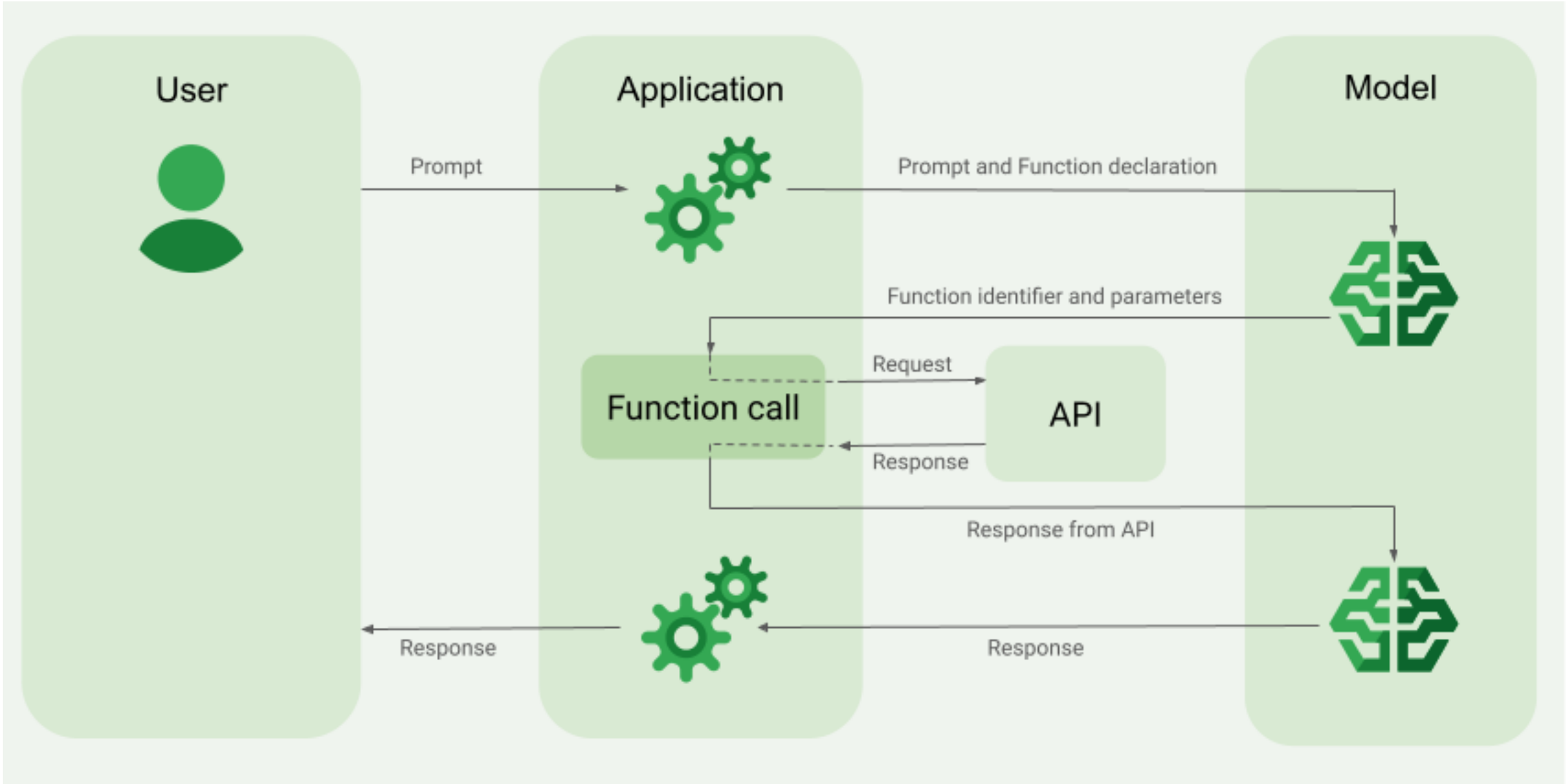
Question:

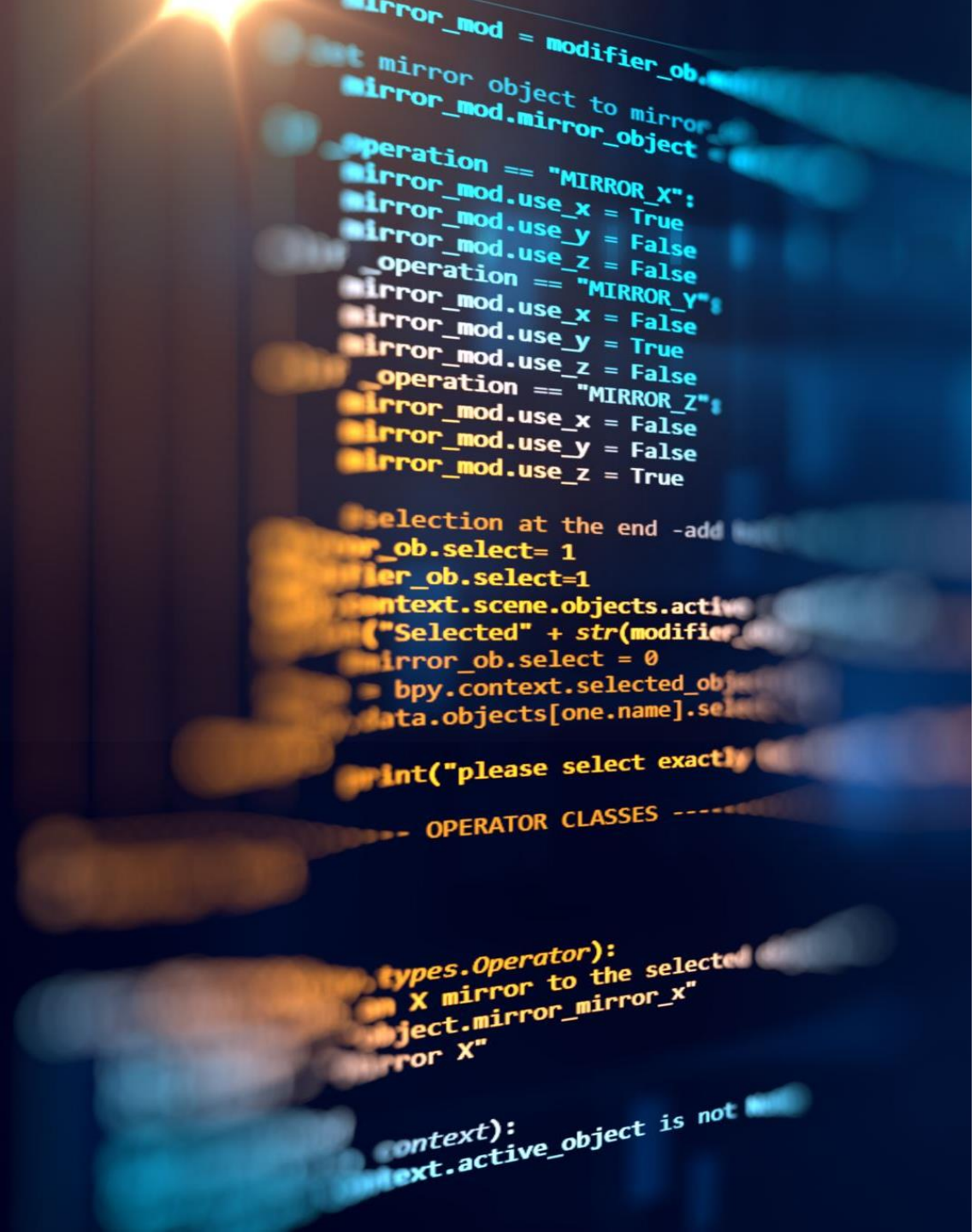
Text Generator (LLM) *"Storyteller"*

"Yes, in Company A's recent 10-K filing, they mentioned their strategic intent to expand their operations in Southeast Asia. The document outlines plans to establish three new retail outlets in Singapore, Bangkok, and Jakarta by the end of 2024. Additionally, the company is in preliminary discussions to form a joint venture with a local distributor in Malaysia to strengthen their market presence in the region."

Output

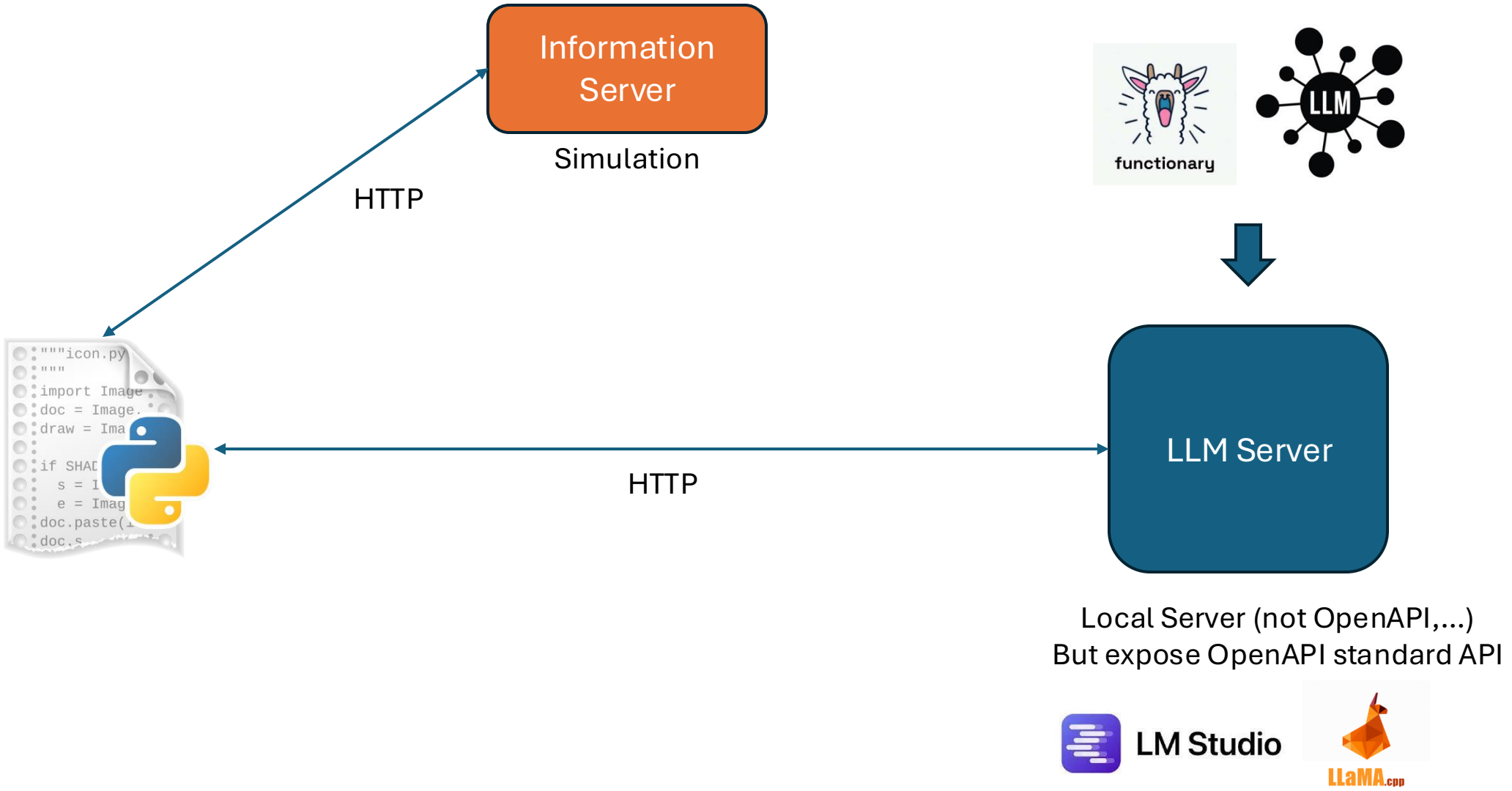

But if external data change frequently?


User
Application
Model
Prompt
Prompt and Function declaration
Function identifier and parameters
Request
Function call
API
Response
Response from API
Response
Response



Demo now!

- No chat GUI
- Just demo Function Calling


Information Server
Simulation
HTTP
functionary
LLM
"""icon.py
"""
import Image
doc = Image.
draw = Ima
if SHAD
s = I
e = Imag
doc.paste(
doc.s
HTTP
LLM Server
Local Server (not OpenAPI,...)
But expose OpenAPI standard API
LM Studio
LLaMA.cpp

Note

- Not all models support local function calling. Check Hugging Face, which hosts a variety of function calling models such as Llama-3 based or Phi3 based models.
- https://huggingface.co/meetkai/functionary-7b-v2-GGUF
- Server: https://github.com/ggerganov/llama.cpp
- python -m llama_cpp.server --model model/functionary-7b-v2.q8_0.gguf --chat_format functionary-v2 --hf_pretrained_model_name_or_path ./model