



**Connect**

# Jak efektywnie budować chatboty i voiceboty

Na Platformie OpenShift AI

Jaroslav Stakun

[RHCA](#), Principal Solutions Architect

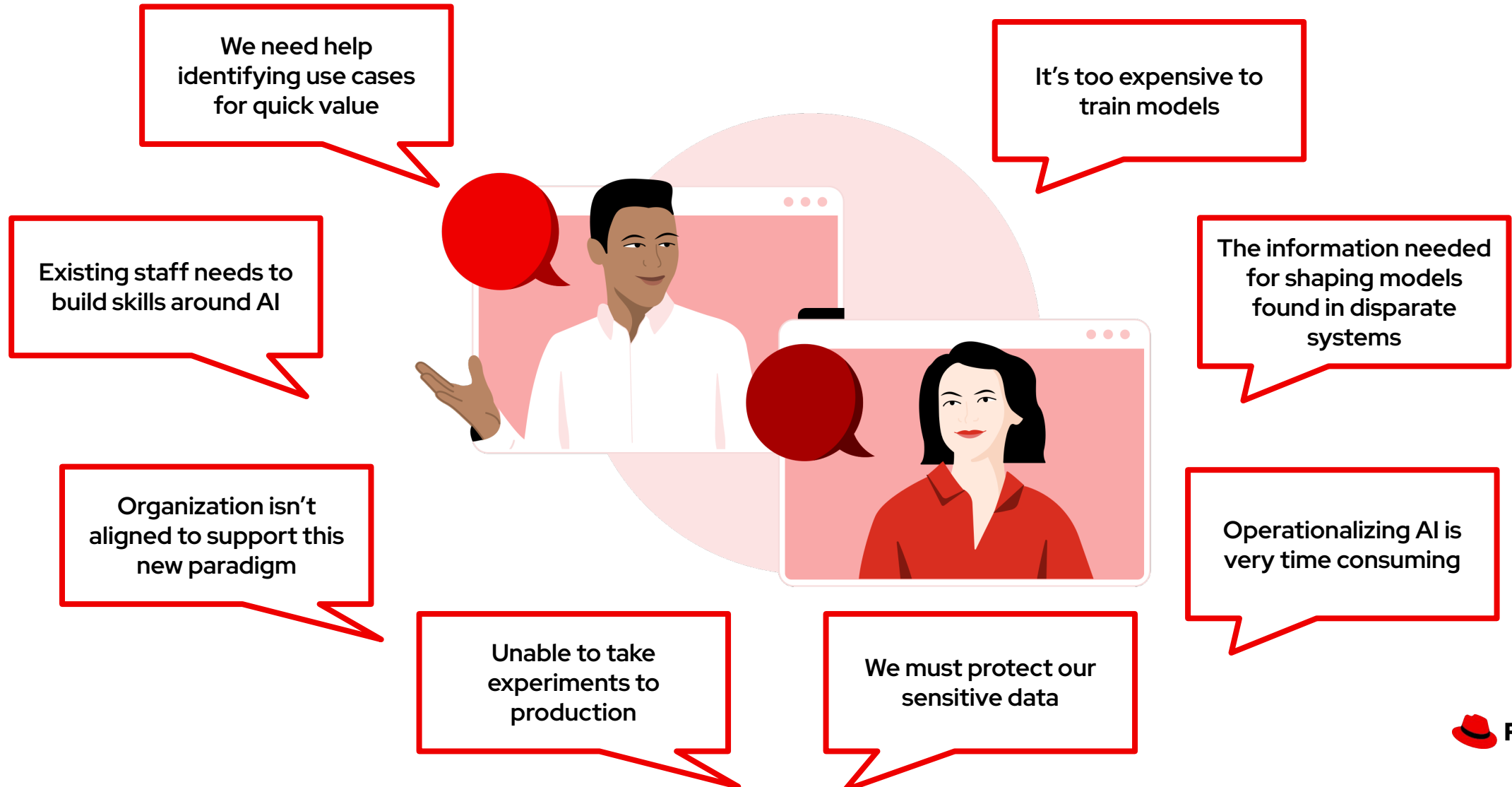
Red Hat CEE

jarek@redhat.com



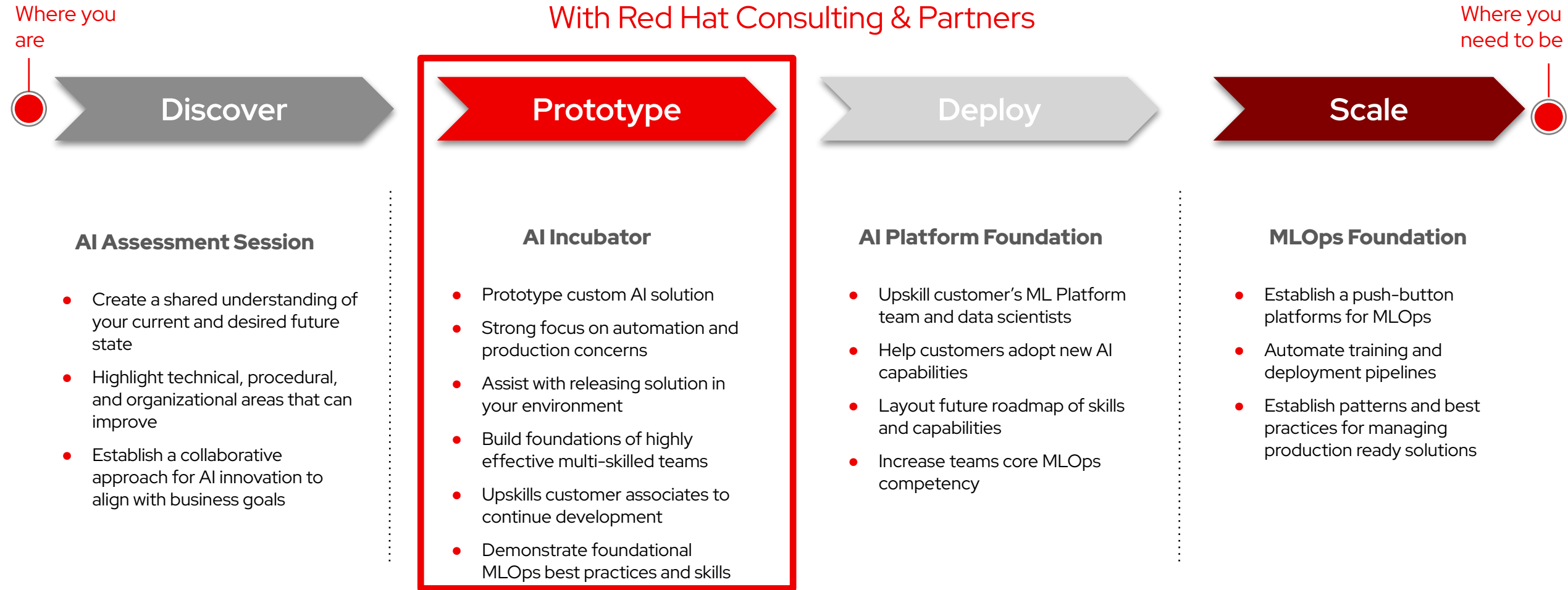
# Common customer challenges

What customers are telling us



# Red Hat AI Adoption Journey

With Red Hat Consulting & Partners



Training and Technical Account Management

# Sample use cases

Venture beyond LLM exploration and evaluation, to include:

## Models As A Service

Focus on offering multiple models, access control and metrics

## Enterprise chatbots/assistants/Document summarization

Document processing, retrieval and security

## Automation and model fine-tuning

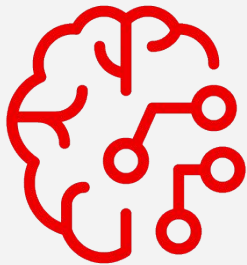
Synthetic data generation and model evaluation

## GenAI-powered application architectures

Semantic Routing, orchestration and agentic workflows

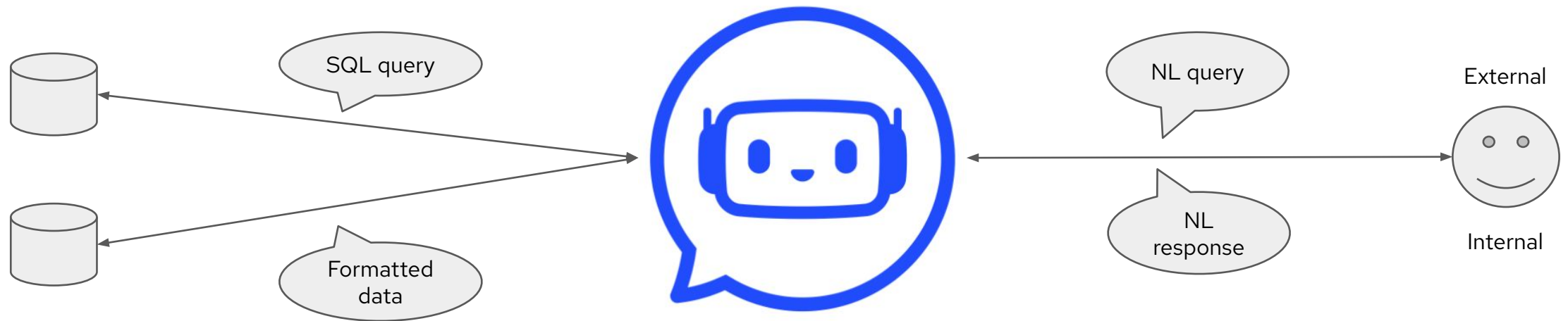
## Predictive AI SDLC automation

Automation of data engineering, streaming data processing, automatic model retraining



# Project scope

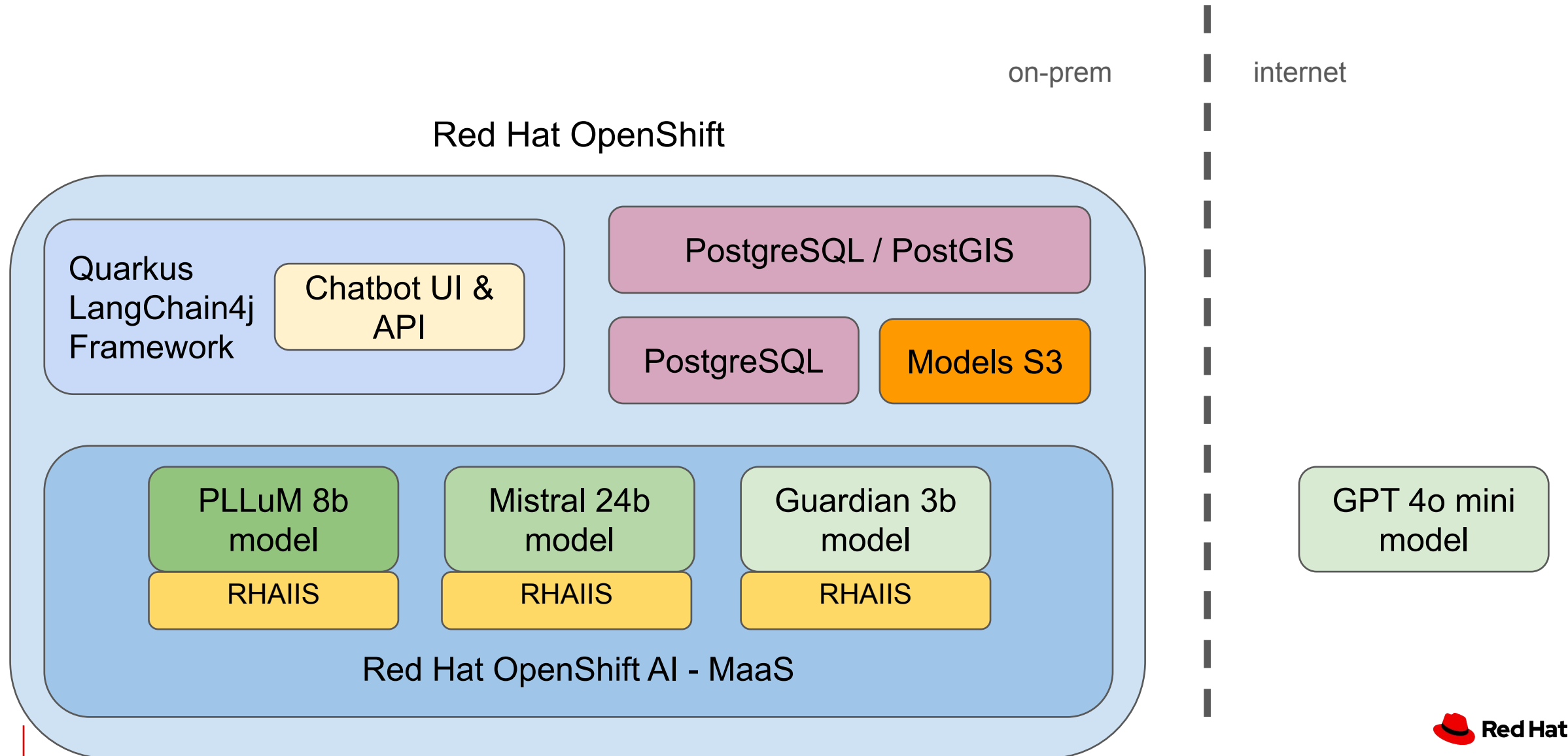
Implement conversational chatbot and voice assistant to query data stored in relational databases (including gis)



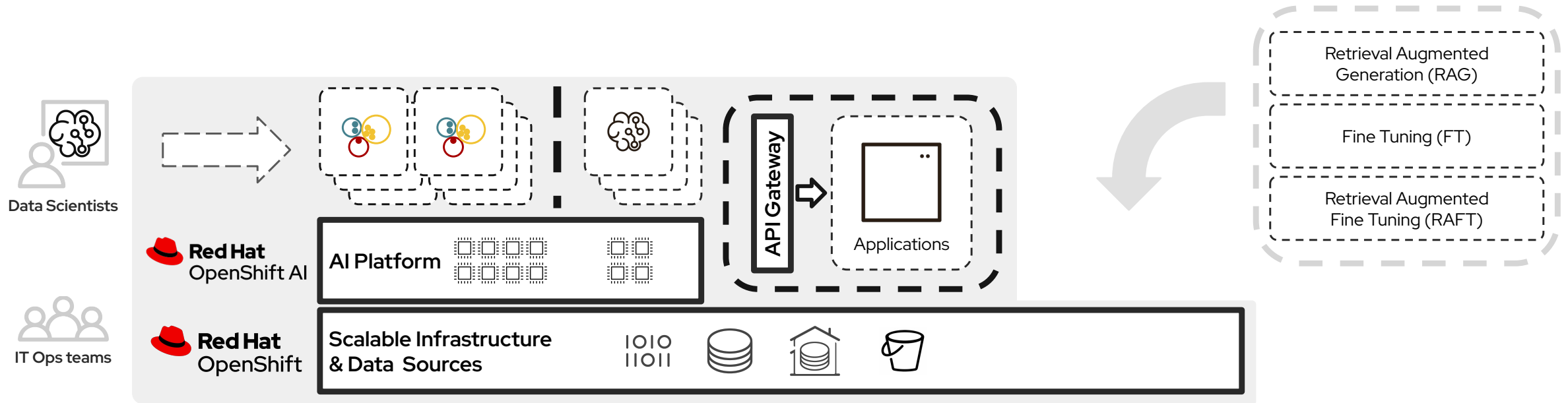
# Use case 1: Chatbot



# Architecture



# Model as a Service (MaaS) on OpenShift AI





# Example MaaS LLMs configuration

Parasol's MaaS

Apps and API Keys

Statistics

API Docs

Usage Examples

Configuration Information

Messages

Settings


Github


★30 v19

## Model Servers configuration

Model name	Model size	Quantization	Max tokens	GPU type	VRAM consumed	Function calling?
DeepSeek-R1-Distill-Qwen-14B-W4A16	14B	W4A16	40k	L40S	42 GB	Yes
Granite-3.3-8B-Instruct	8B	No	26k	A10G	21 GB	Yes
Granite-8B-Lab	8B	No	6144	L40	22 GB	No
Granite-3-Guardian-2B	2B	No	6048	T4	13 GB	No
Granite-Vision-3.2-2B	2B	No	16384	L40	21 GB	Yes
Llama-3.2-3B-Instruct	3B	No	120k	L40	21 GB	Yes
Llama-4-Scout-17B-16E-Instruct	109B	W4A16	400k	4xL40S	4x43 GB	Yes
Mistral-Small-3.1-24B-Instruct	24B	W8A8	90k	L40S	42 GB	Yes
Nomic-embed-text-v1.5 (4 workers in parallel)	137M	No	8192	T4	15 GB	N/A
Microsoft Phi-4	14B	No	16384	L40S	42 GB	No
Qwen2.5-VL-7B-Instruct	7B	FP8-Dynamic	100k	A10G	19 GB	No

Not an official Red Hat service. For Red Hat associate internal demo purposes only, provided 'as-is' without support or SLA.  
The intended purpose is to test connectivity of Red Hat products to models that customers may use. The models are provided for this limited purpose.

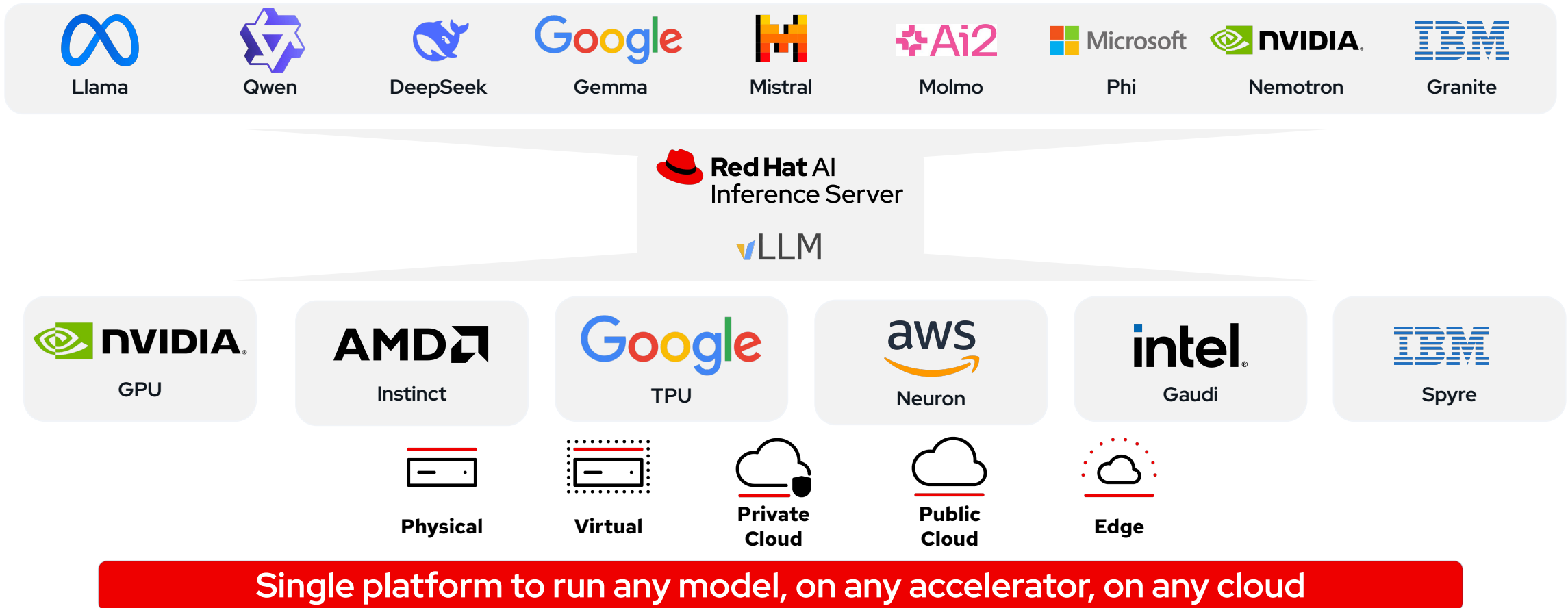
© 2025 PARASOL'S MAAS, Powered by  3scale

 **Red Hat**

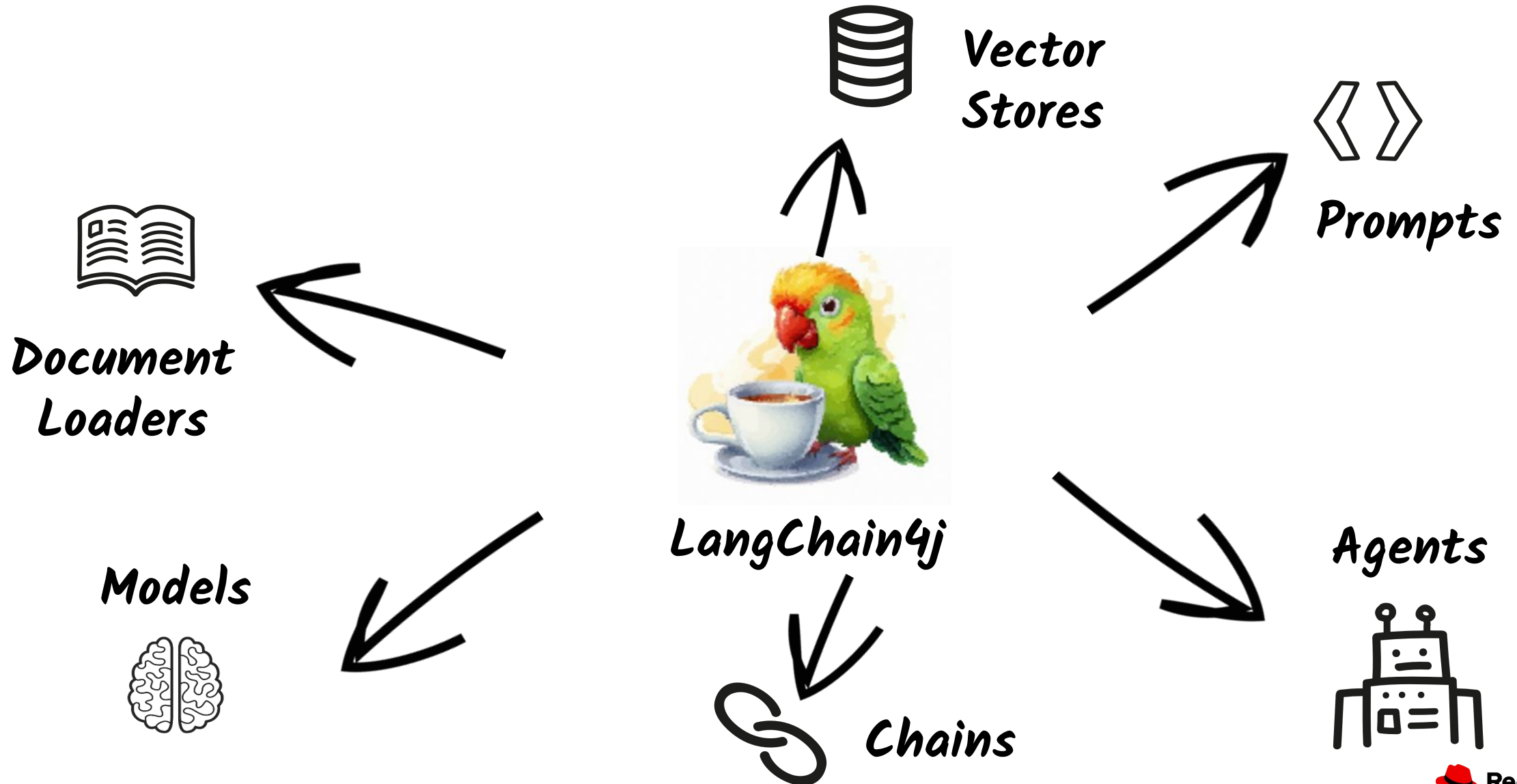


# Red Hat AI Inference Server

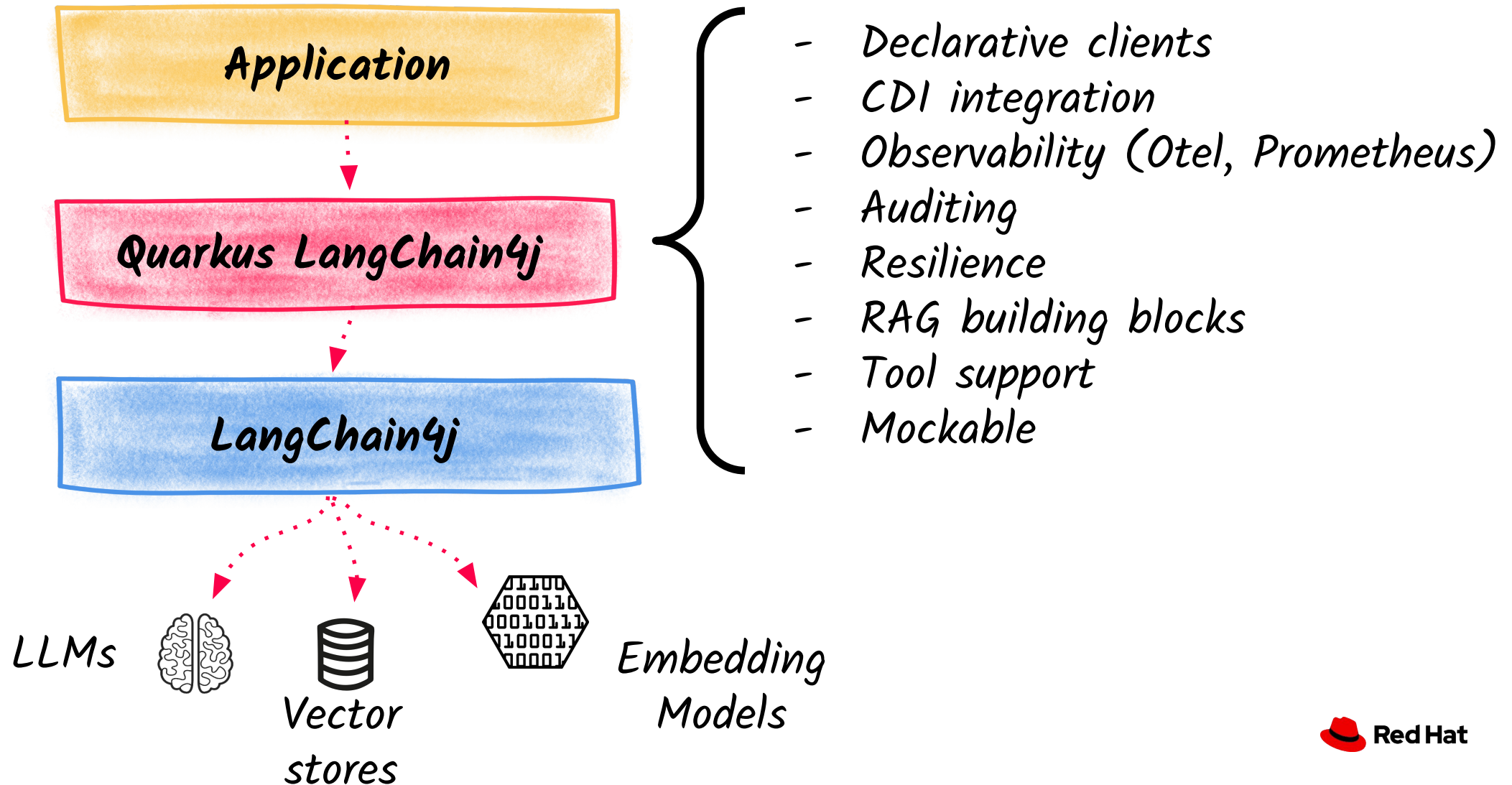
vLLM connects model creators to accelerated hardware providers



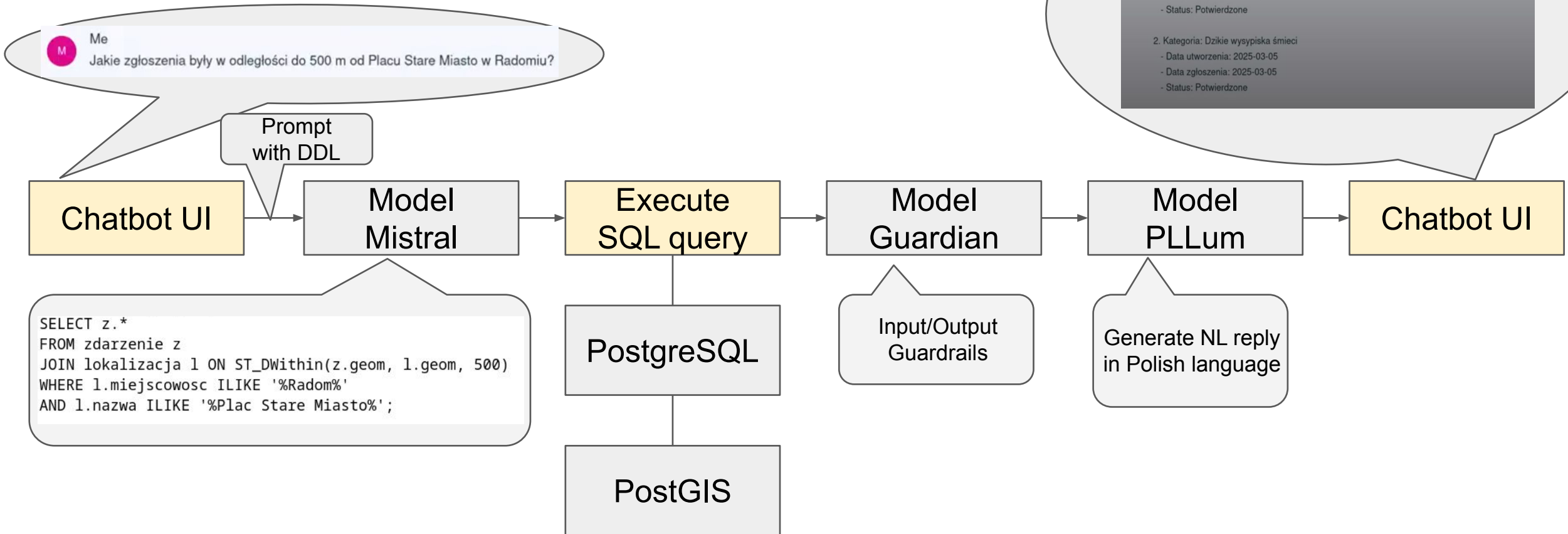
# LangChain4j Framework



# Quarkus LangChain4j



# Text2SQL flow





A007

Jestem Twoim Asystentem. Jak mogę Ci pomóc?



Me

Jakie zgłoszenia były w odległości do 300 m od ul 25 Czerwca 25 w Radomiu?



A007

W odległości do 300 m od ul. 25 Czerwca 25 w Radomiu zarejestrowano następujące zgłoszenia:

1. **Kategoria:** Nieprawidłowe parkowanie  
**Data zgłoszenia:** 2025-01-23  
**Status:** Niepotwierdzone
2. **Kategoria:** Nieprawidłowe parkowanie  
**Data zgłoszenia:** 2025-01-16  
**Status:** Niepotwierdzone
3. **Kategoria:** Spożywanie alkoholu w miejscach niedozwolonych  
**Data zgłoszenia:** Brak daty  
**Status:** Potwierdzone (wyeliminowane)



A007

Jestem Twoim Asystentem. Jak mogę Ci pomóc?



Me

Ile jest zgłoszeń w poszczególne dni tygodnia



A007

Oto liczba zgłoszeń w poszczególne dni tygodnia:

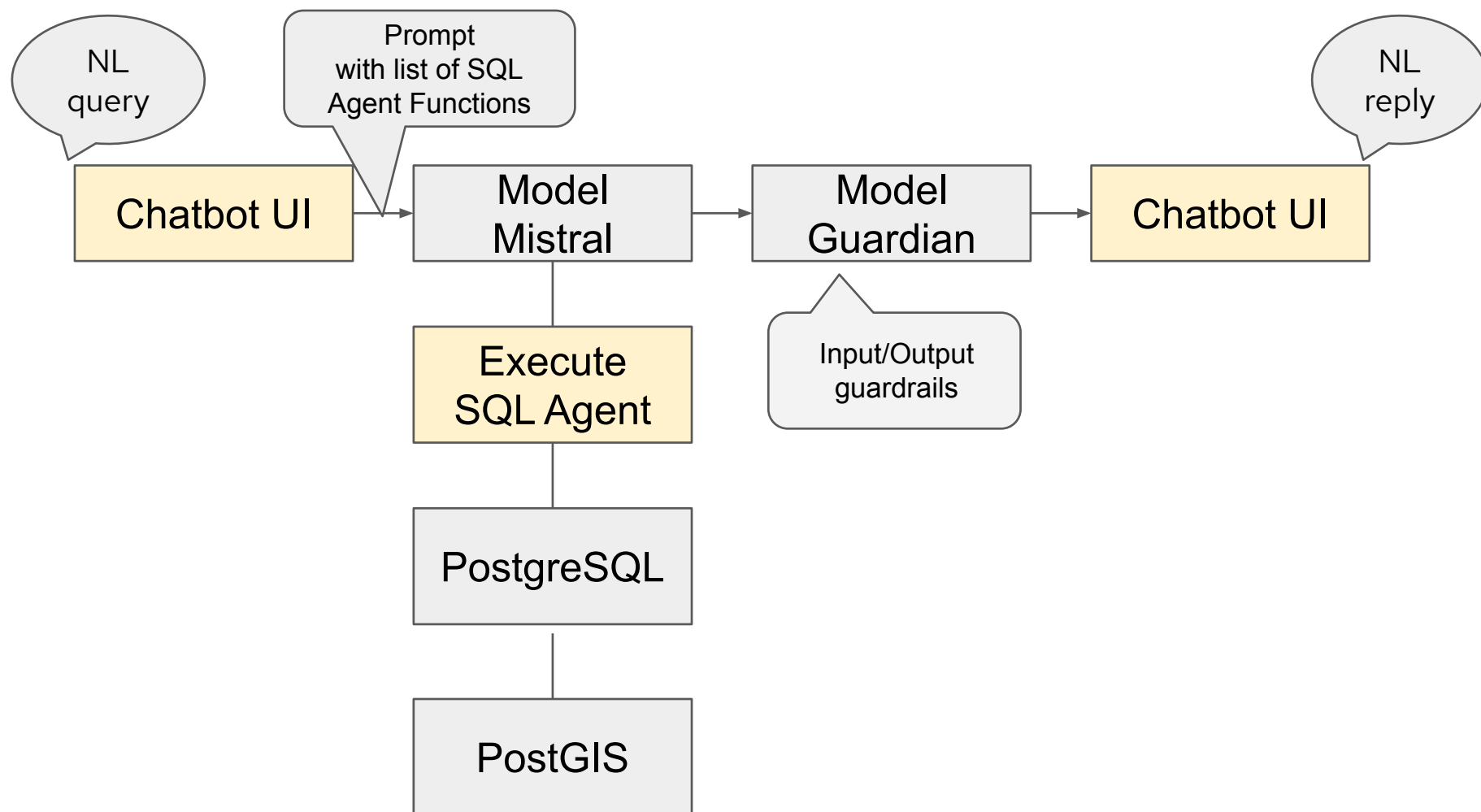
- Poniedziałek: 38 zgłoszeń
- Wtorek: 42 zgłoszenia
- Środa: 31 zgłoszeń
- Czwartek: 44 zgłoszenia
- Piątek: 37 zgłoszeń
- Sobota: 37 zgłoszeń
- Niedziela: 29 zgłoszeń

Message

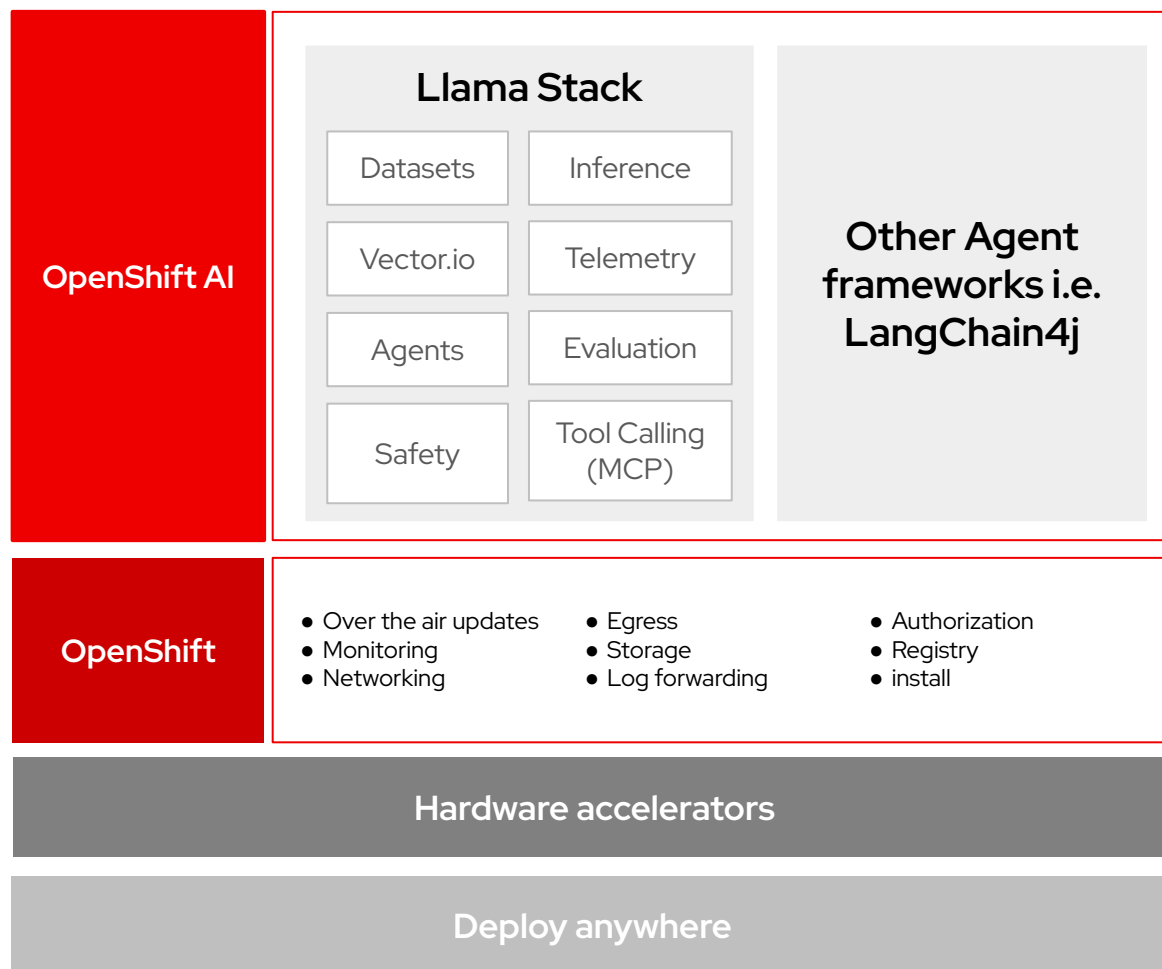
Send



# Architecture - Agentic flow



# A modular approach to building AI agents



## Red Hat AI allows to:

- ▶ Build agents using **Llama Stack's native capabilities and implementations**.
- ▶ **Bring compatible Llama Stack implementations** to OpenShift AI.
- ▶ **Use your own agent framework** and selectively incorporate Llama Stack APIs.
- ▶ **Build with Core Primitives** and manage your own agent framework as a standard workloads.

# How it works?

## AI Assistant



A007

Jestem Twoim Asystentem. Jak mogę Ci pomóc?



Me

Who is sales rep responsible for IBM. Please describe steps taken to answer the question.



A007

The sales representative responsible for IBM is János Illés.

Steps taken to answer:

1. I used the searchVendorRepresentative tool to look up the sales and presales representatives assigned to IBM.
2. The tool returned János Illés as the sales representative for IBM.
3. I provided his name as the final answer as requested.

Wiadomość

Wyślij

# Multi language

## AI Assistant



A007

Én vagyok az asszisztens. Hogyan segíthetek?



Me

Ki a Red Hat értékesítési képviselője?



A007

A Red Hat értékesítési képviselője Norbert Szabó, email címe: [norbert.szabo@arrow.com](mailto:norbert.szabo@arrow.com).



Me

Ki a Red Hat előértékesítője?



A007

A Red Hat előértékesítője István Kerekes, email címe: [istvan.kerekes@arrow.com](mailto:istvan.kerekes@arrow.com).

Üzenet

Elküld

# Chat memory

## AI Assistant

M Me  
Jacy dostawcy oferują rozwiązania backup?

A A007  
Dostawcy oferujący rozwiązania backup to:

- Commvault
- Veeam

M Me  
Szukam dostawców do projektu bezpieczeństwa sieci

A A007  
Dostawcy oferujący rozwiązania do projektu bezpieczeństwa sieci to:

- Fortinet
- Netwitness
- Trend Micro

M Me  
Kto jest opiekunem handlowym firmy Red Hat?

A A007  
Opiekunem handlowym firmy Red Hat jest Norbert Szabó.

M Me  
Jak mogę się z nim skontaktować?

A A007  
Możesz skontaktować się z Norbertem Szabó za pomocą e-maila: [norbert.szabo@arrow.com](mailto:norbert.szabo@arrow.com).

# Observability



Administrator

Home

Operators

Workloads

Serverless

Networking

Storage

Builds

Observe

Compute

User Management

Administration

You are logged in as a temporary administrative user. Update the [cluster OAuth configuration](#) to allow others to log in.

Project: ai-assistant

Pods

Create Pod

Name	Status	Ready	Restarts	Owner	Memory	CPU	Created
ai-assistant-66f45f4c4f-prccz	Running	1/1	1	ai-assistant-66f45f4c4f	551.1 MiB	0.000 cores	6 maj 2025, 11:32
postgis-db-0	Running	1/1	0	postgis-db	213.4 MiB	0.000 cores	6 maj 2025, 11:32



Administrator

Home

Operators

Workloads

Serverless

Networking

Storage

Builds

Observe

Compute

User Management





Administration

You are logged in as a temporary administrative user. Update the [cluster OAuth configuration](#) to allow others to log in.

Project: ai-labs

Pods

Create Pod

Name	Status	Ready	Restarts	Owner	Memory	CPU	Created
 pllum-predictor-00005-deployment-6d78d7487d-nj7t5	 Running	3/3	0	 pllum-predictor-00005-deployment-6d78d7487d	6,323.3 MiB	0.008 cores	 6 maj 2025, 13:27

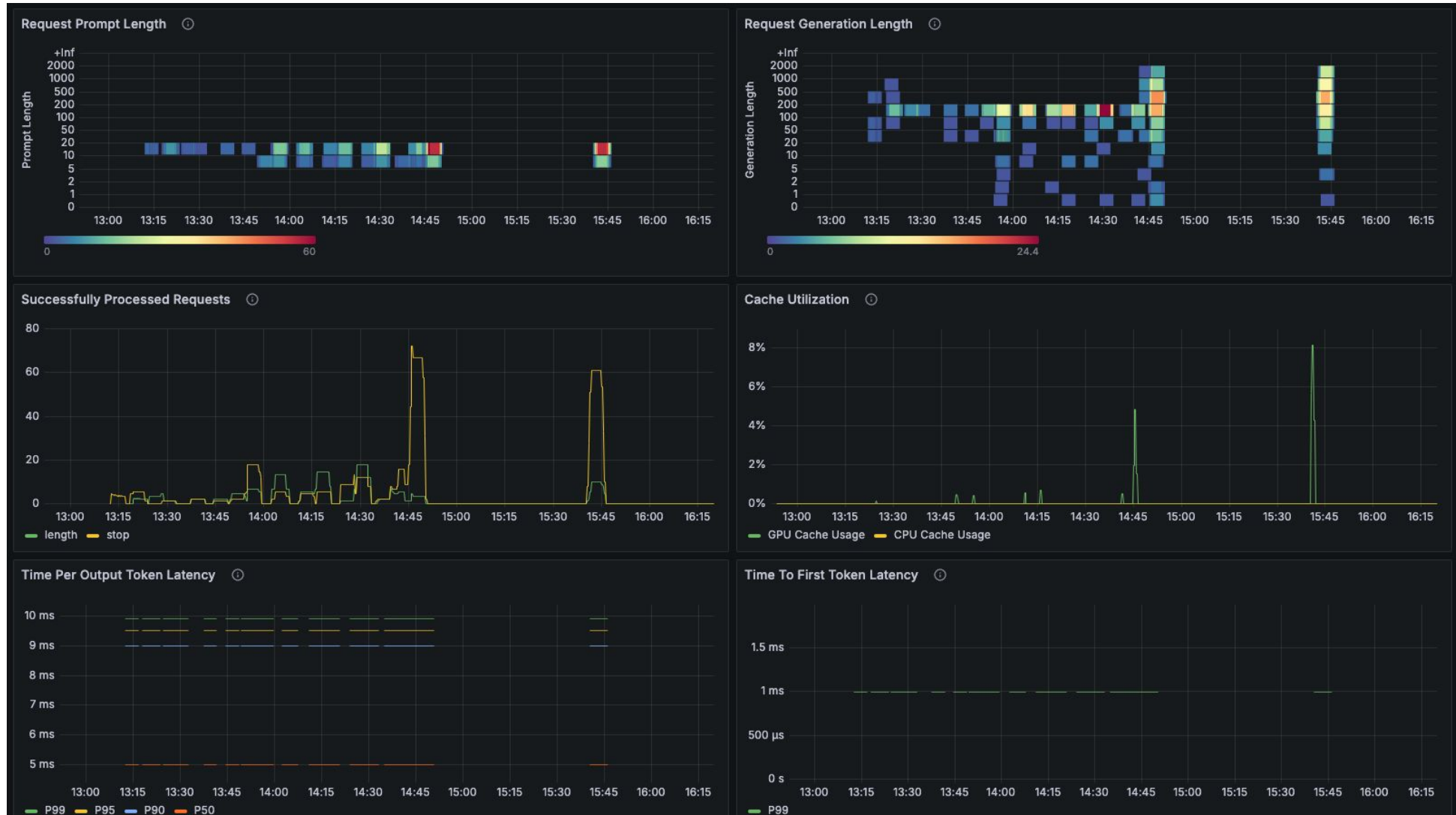




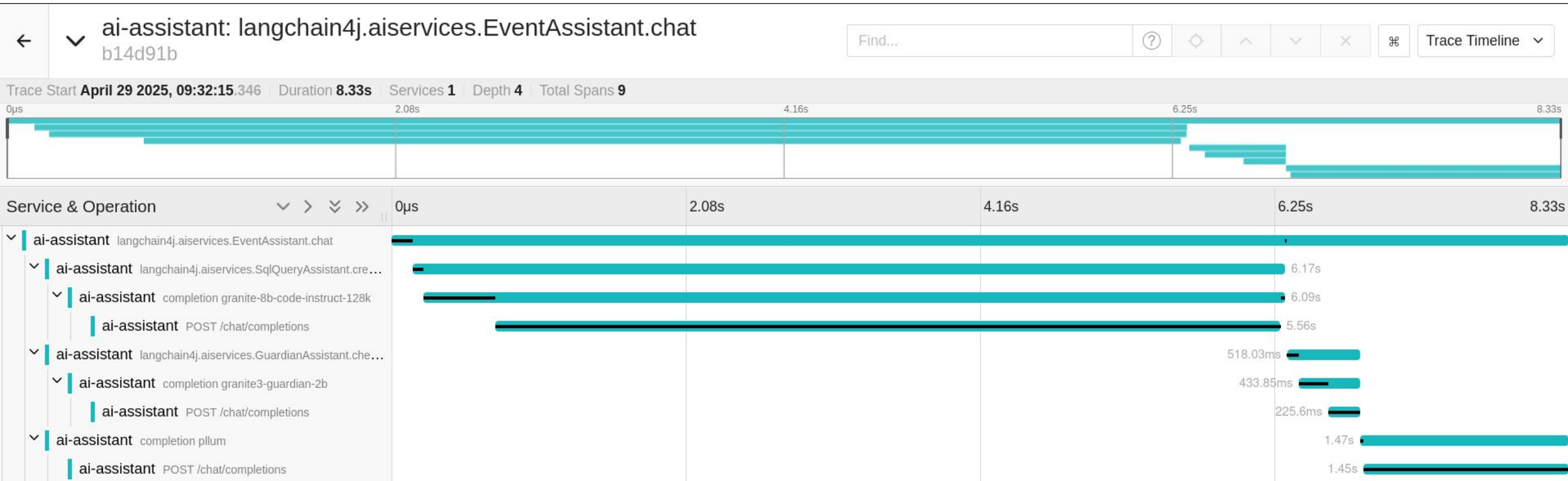
# Example model request log

```
2025-09-15 11:17:59,373 INFO [io.qua.lan.oape.com.OpenAiRestApi$OpenAiClientLogger] (vert.x-eventloop-thread-7) Request:
- method: POST
- url: https://api.openai.com/v1/chat/completions
- headers: [Accept: text/event-stream], [Authorization: Be...IA], [Content-Type: application/json], [User-Agent: langchain4j-openai], [content-length: 4755]
- body: {
  "model" : "gpt-4.1",
  "messages" : [ {
    "role" : "system",
    "content" : "You are an intelligent client assistant. Users will ask you questions about IT vendors, their solutions and sales and presales representatives from distributor named Arrow.\nYour task is to provide accurate and concise answers.\nMake sure to create answers according to following guidelines:\n- You have access to a collection of tools. You can use multiple tools simultaneously.\n- Complete your answer using the data obtained from the tools.\n- If you are asked about specific role i.e. sales or presales provide only name of that role person\n- If you are unable to access the tools to answer the user's question, state that the requested information is currently unavailable and that they can try again later.\n- Reply in the same language as the question was asked.\n- If you are asked about vendors answer only with vendor name\n- If there is any difference between requested vendor, sales or presales name and the name used in response please mention that\n"
  }, {
    "role" : "user",
    "content" : "How can you help me?"
  }, {
    "role" : "assistant",
    "content" : "I can help you with the following:\n\n- Provide information about IT vendors and their solutions available through Arrow.\n- Identify sales or presales representatives for specific vendors.\n- Find which vendors offer specific IT solutions (e.g., security, storage, cloud).\n- List which vendors are managed by a specific sales or presales representative.\n\nIf you have a specific vendor, solution, or Arrow contact in mind, just ask!"
  }, {
    "role" : "user",
    "content" : "Who is presales responsibel for Redhat?"
  }, {
    "role" : "assistant",
    "tool_calls" : [ {
      "id" : "call_fms8MZ3V01s1qXoAyyvXZspGh",
      "type" : "function",
      "function" : {
        "name" : "searchVendorRepresentative",
        "arguments" : "{\n  \"vendor\" : \"Redhat\"\n}"
      }
    }
  ]
}, {
  "role" : "tool",
  "tool_call_id" : "call_fms8MZ3V01s1qXoAyyvXZspGh",
  "content" : "[ {\n  \"id\" : 18,\n  \"vendorName\" : \"Red Hat\",\n  \"salesRepName\" : \"Norbert SzabÃ³\",\n  \"salesRepEmail\" : \"norbert.szabo@arrow.com\",\n  \"presalesRepName\" : \"IstvÃ¡n Kerekes\",\n  \"presalesRepEmail\" : \"istvan.kerekes@arrow.com\",\n  \"vendorTags\" : \"operating system,Enterprise Linux,Open Source Software,Cloud Computing,Virtualization,Containerization,Automation\",\n  \"confidenceScore\" : 0.5\n } ]"
```

# Example model metrics



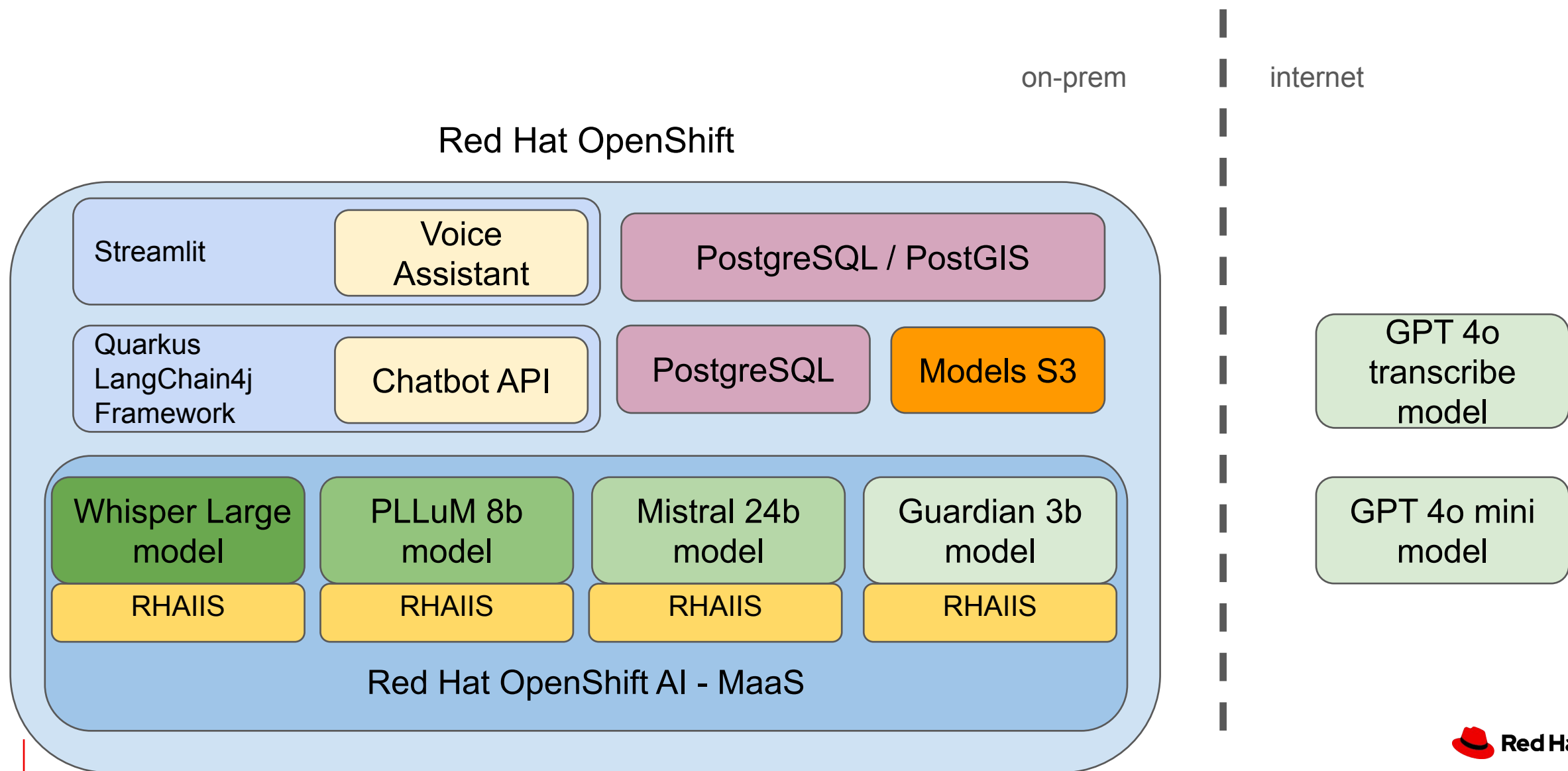
# Example Text2SQL flow trace



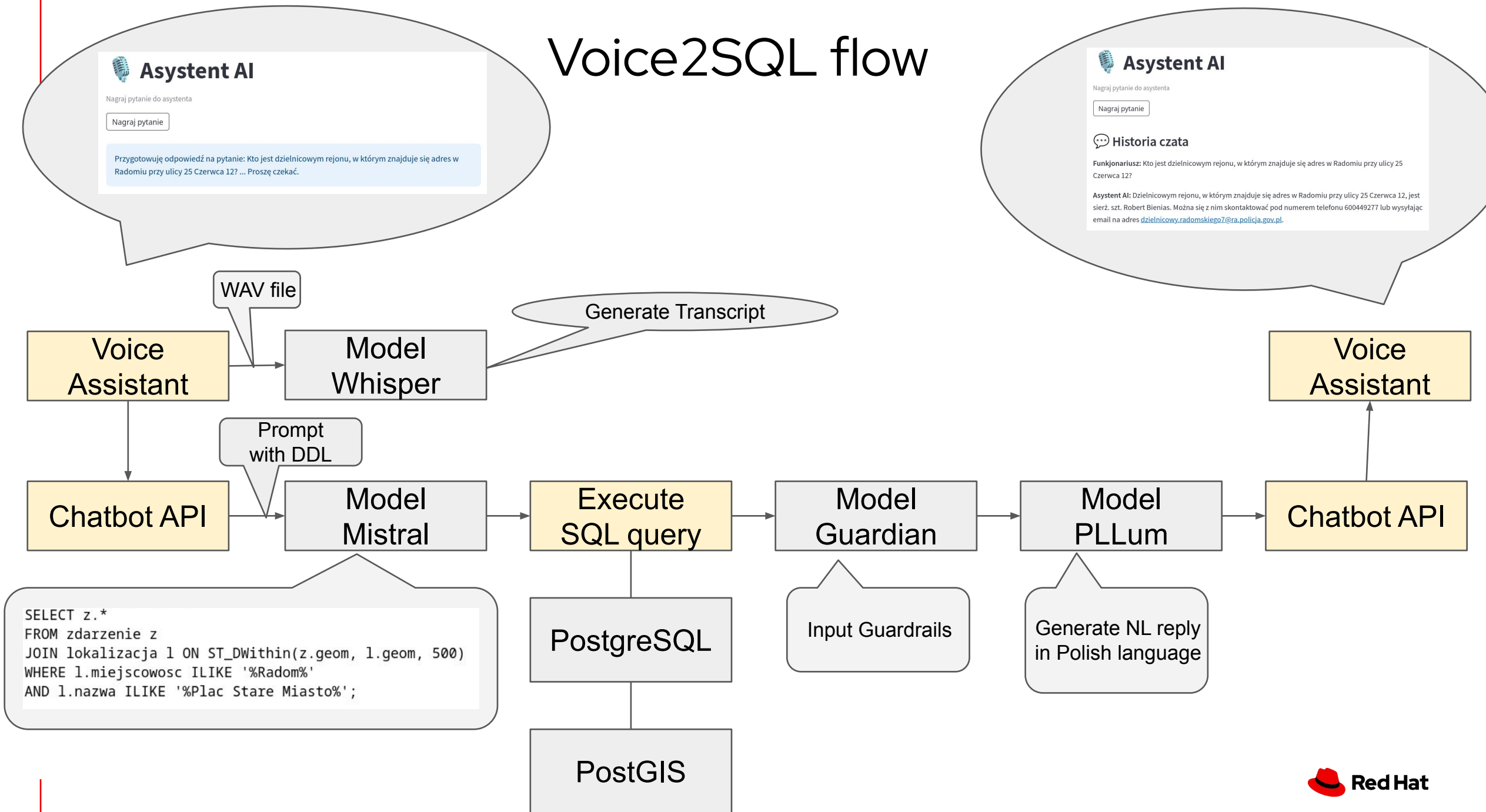
# Use case 2: Voice Assistant (Voice to Text)



# Architecture



# Voice2SQL flow





# Asystent Głosowy

Nagraj pytanie dotyczące dzielnicowych i jednostek policji o maksymalnej długości 20 sekund

Wpisz token i zatwierdź ENTER:

....



Rozpocznij nagranie



0:00 / 0:08



## Historia czata

**Funkcjonariusz:** Podaj adres jednostki Policji dzielnicowego rejonu, w którym znajduje się Plac Stare Miasto w Radomiu.

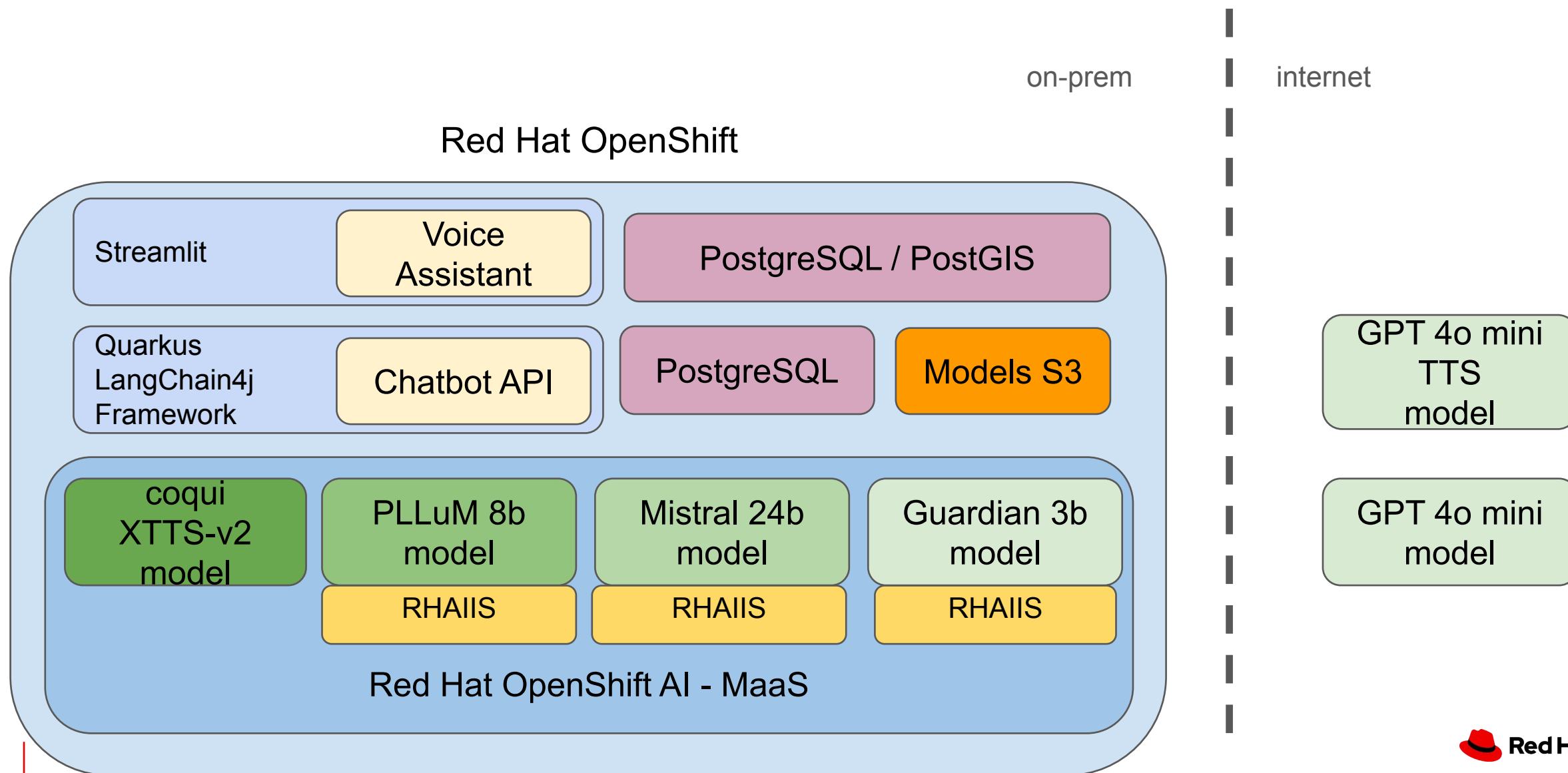
**Asystent AI:** Adres jednostki Policji dzielnicowego rejonu, w którym znajduje się Plac Stare Miasto w Radomiu, to: ul. 11-go Listopada 37/59, Radom, 26-600.

# Use case 3: Voice Assistant (Text to Speech)

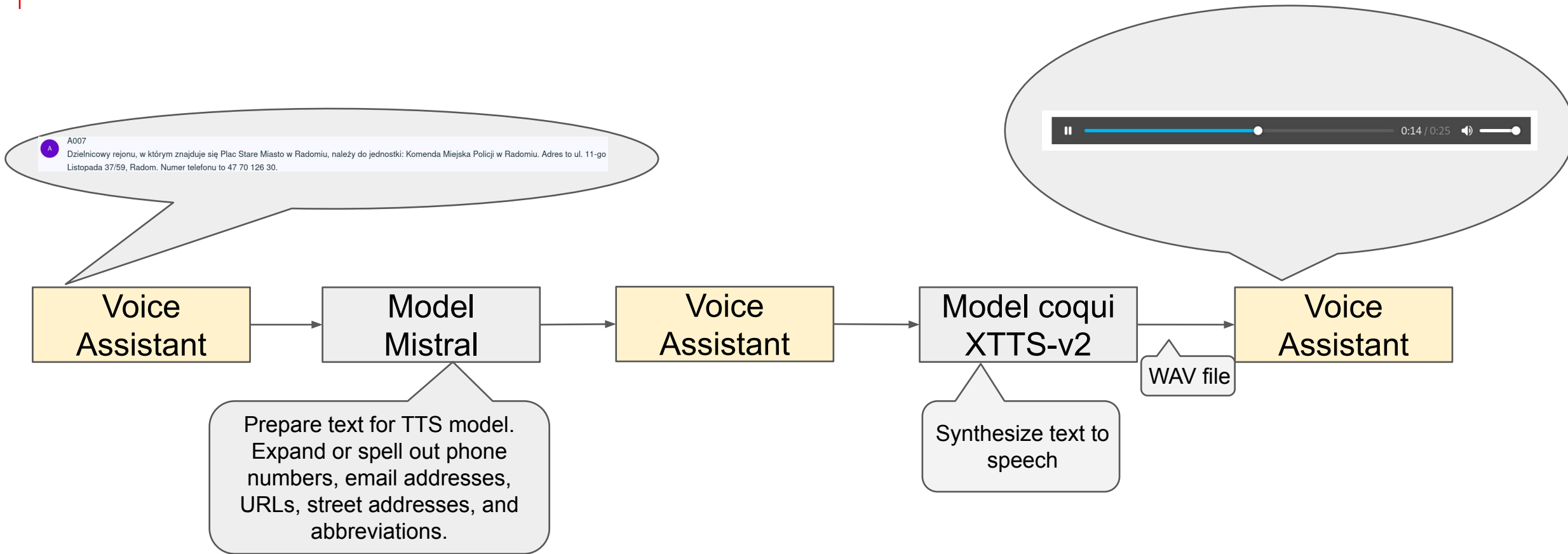




# Architecture



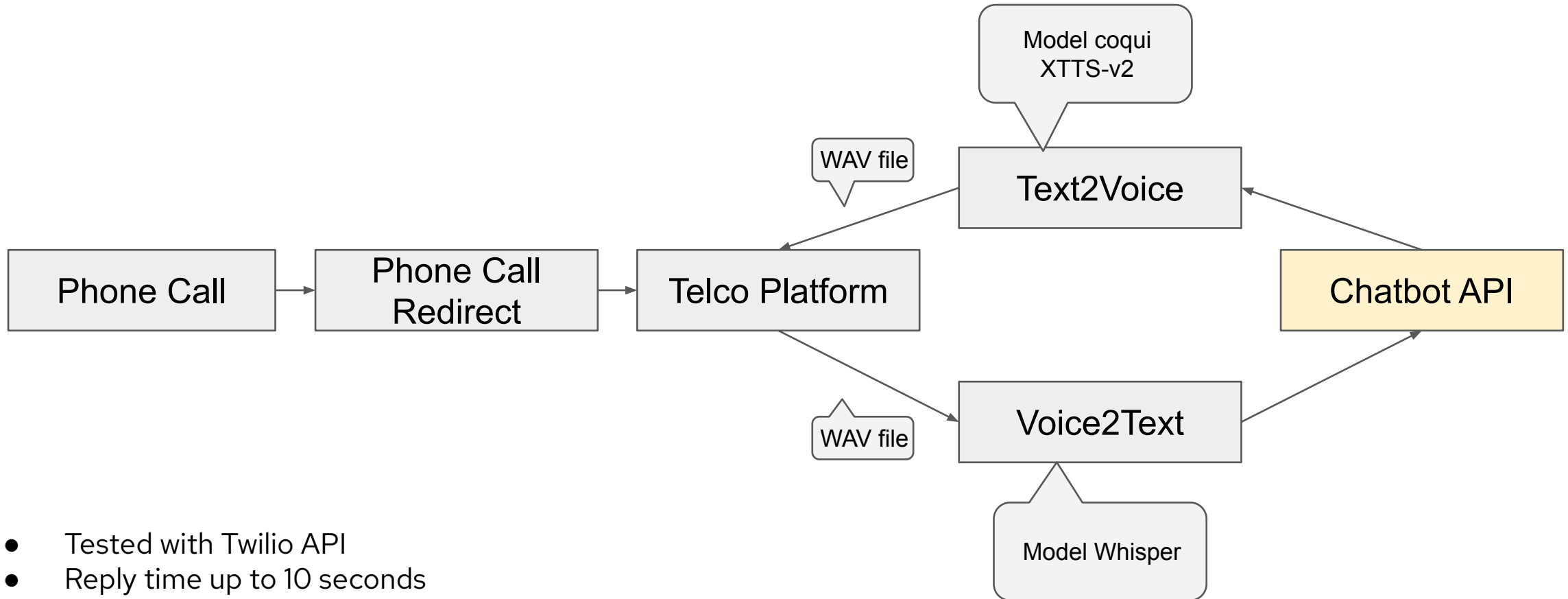
# Text2Voice flow



# Use case 4 prototype: Phone assistant



# Phone assistant flow



# Summary



# Solution Highlights

- Simplified RAG
  - No need to implement Vector DB / LAB alignment / Fine Tuning
  - Less infrastructure and configuration needed
- Easier to implement for the customer and for us
  - Only sql/api access to data or data dumps are needed
- Available small/medium LLM models can be used AS-IS (+90% correct answers)
  - Granite, PLLum, Mistral, Whisper, ...
  - Quantized models from [Red Hat AI HF repo](#)
- Tested on lower cost NVIDIA GPU
  - A10G 24 GB - [Ampere microarchitecture](#)
- Can be deployed to single secure hybrid cloud platform
  - Models and Apps together
  - OpenShift AI on top of OpenShift
  - Foundation for other AI projects

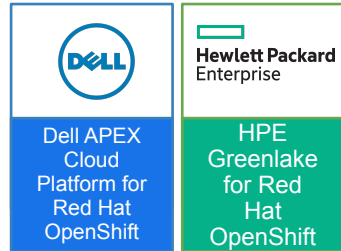
# Hybrid Cloud deployment options



On-Premises or Edge  
Customer Managed  
Application Platform



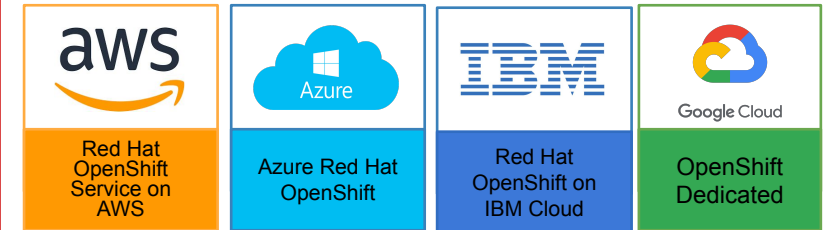
On-Premises  
Partner  
Provided



Public Cloud  
Customer  
Managed



Public Cloud  
Cloud Managed  
Application Platform



USAGE MODEL

PAYMENT MODEL



LOCATION

# Development Challenges -> Solutions

- Dealing with improper user prompt and possible model hallucinations -> PE, Guardrails
- Database result set controls
  - **Database Schema naming -> ORM naming**
  - **SQL Query correctness validation -> RAG (vector database with similar SQL queries)**
  - SQL Query syntax -> execute query, a few shots strategy
  - SQL Query optimization -> check result set size, a few shots strategy
  - Wrong number of returned results -> PE
  - Response strictness -> PE, Confidence score
- Streaming tokens vs waiting for all tokens received from the model
- User session chat memory vs No chat memory
  - Chain of models calls
- Connection / response timeouts
  - Models
  - Database





Connect

# Thank you



[linkedin.com/company/red-hat](https://linkedin.com/company/red-hat)



[facebook.com/redhatinc](https://facebook.com/redhatinc)



[youtube.com/user/RedHatVideos](https://youtube.com/user/RedHatVideos)



[twitter.com/RedHat](https://twitter.com/RedHat)

