Open APIs for Open Minds

Enhancing Tourism with Smart Technologies

Integrating FIWARE and Large Language Models

David Campo (Ph.D. in Computational Intelligence | Sr FIWARE Expert & evangelist)
Javier Conde (Ph.D. in Telecommunication Engineering | Data Engineering, Professor)

FIWARE

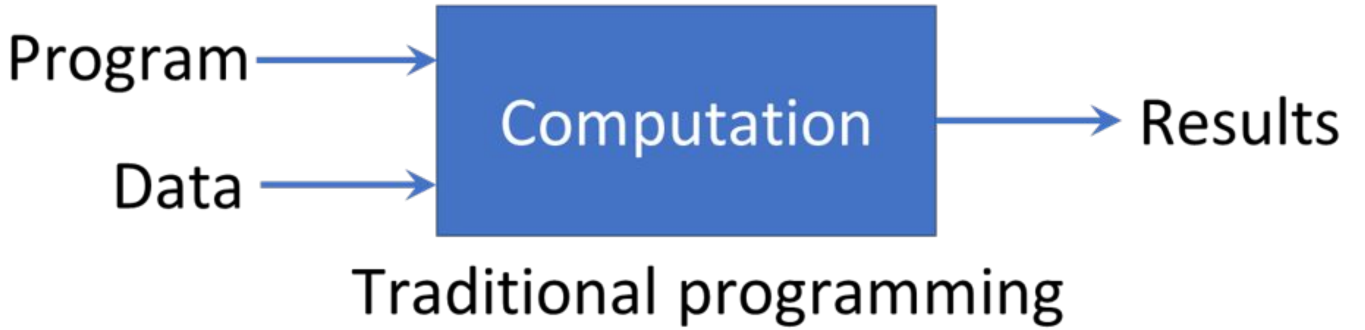
Agenda

1. Introduction to Generative AI and LLM
2. Retrieval Augmented Generation (RAG) and FIWARE
3. Implementation in FIWARE
4. DEMO: Naples Tourism use case
5. Summary

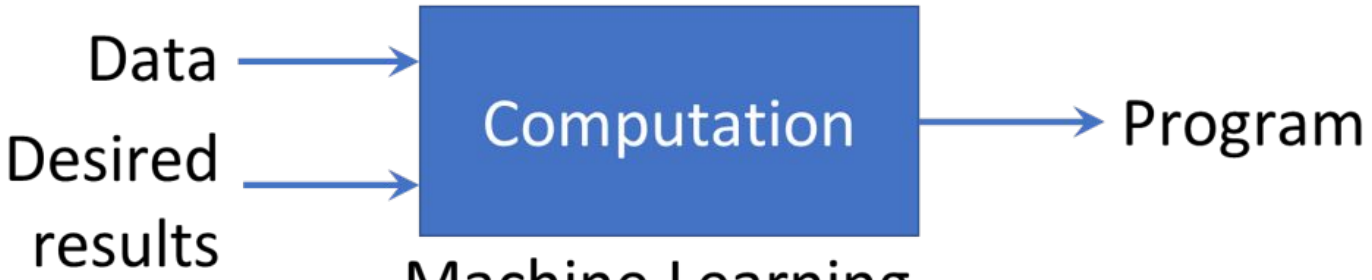
Agenda

1. **Introduction to Generative AI and LLM**
2. Retrieval Augmented Generation (RAG) and FIWARE
3. Implementation in FIWARE
4. DEMO: Naples Tourism use case
5. Summary

Machine Learning

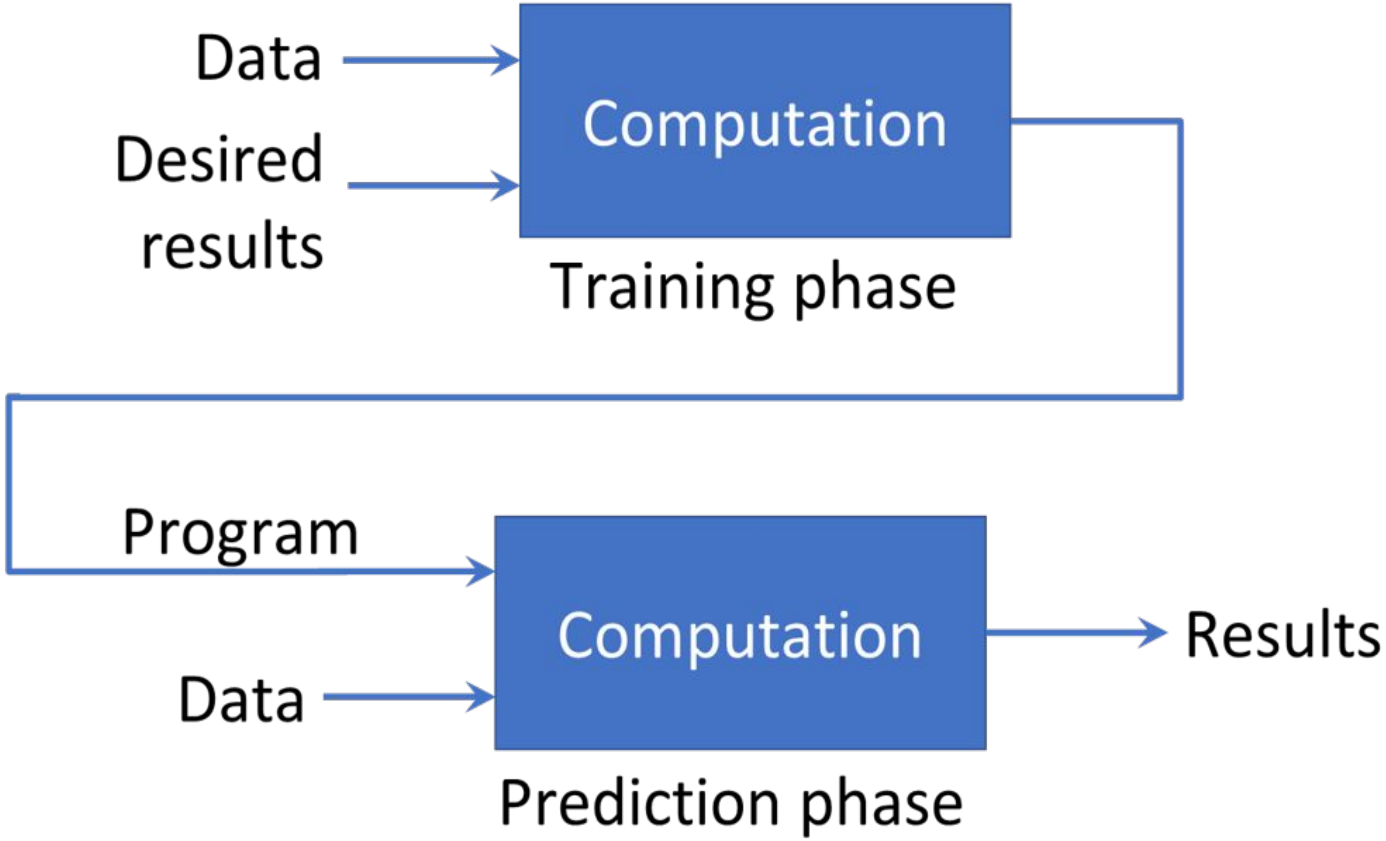


Traditional programming

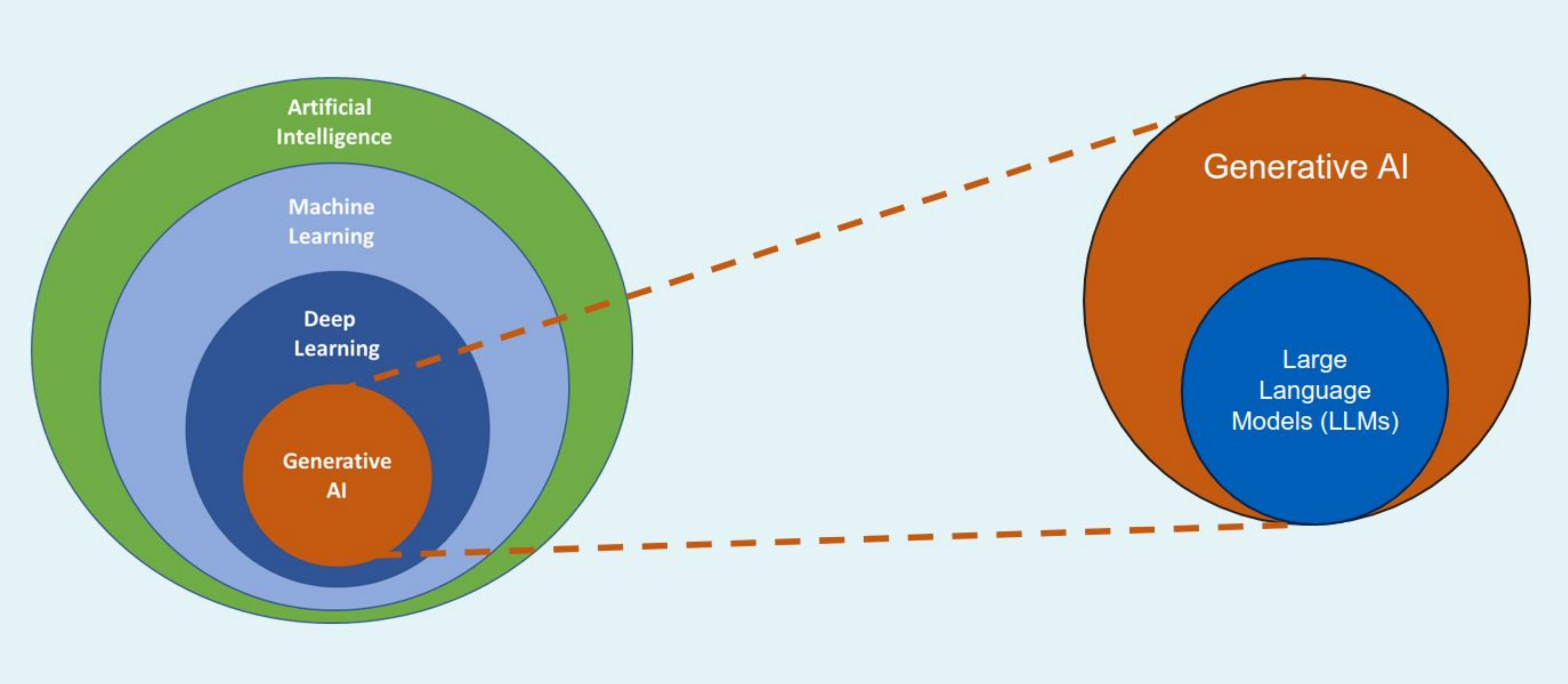


Machine Learning

Machine Learning



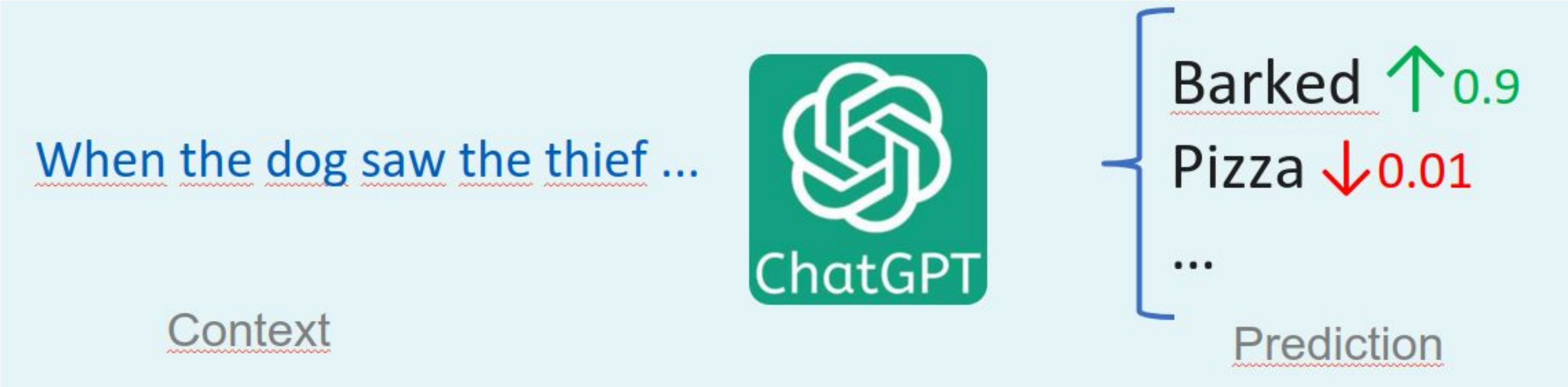
Generative AI



Large Language Models (LLM)

- From large amounts of text we train a model that learns to predict the following Word (in reality they predict tokens not words)

- Language models assume that the text can be predicted as a probabilistic system in which the next word depends on the previous ones



Large Language Models (LLM)

Once we have this model, we can complete sentences using the model to obtain each word

When the dog saw the thief barked ...

When the dog saw the thief barked to ...

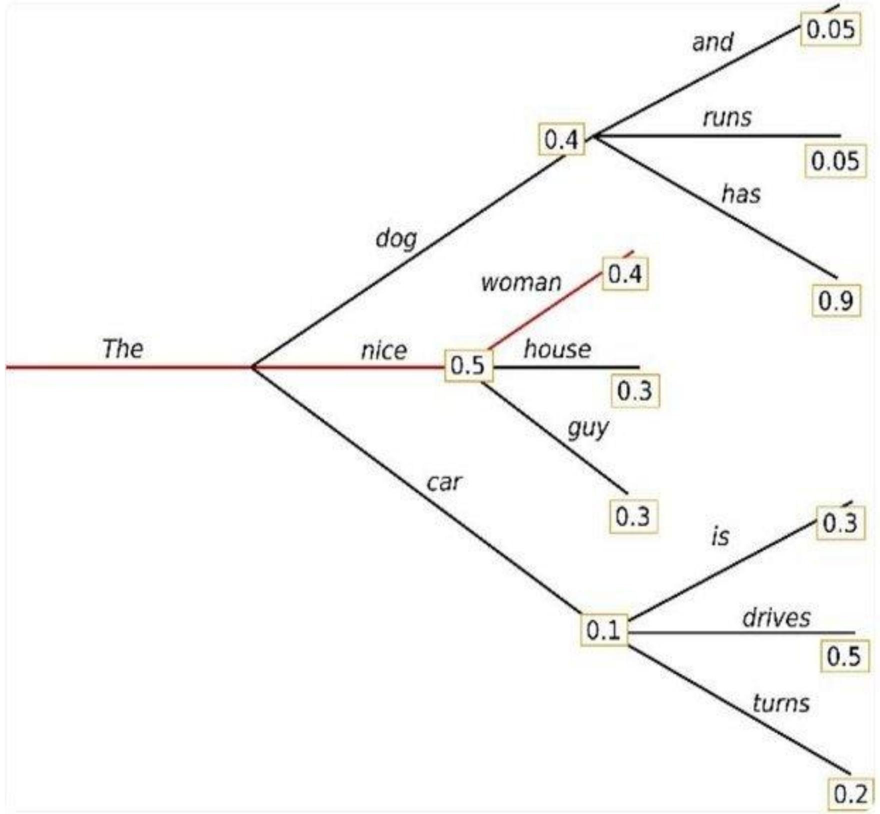
When the dog saw the thief barked to scare ...

Large Language Models (LLM)

Once we have this model, we can complete sentences using the model to obtain each word

As it is based on the probabilities of the next Word (token), there are a lot of possible paths depending on how the model decides:

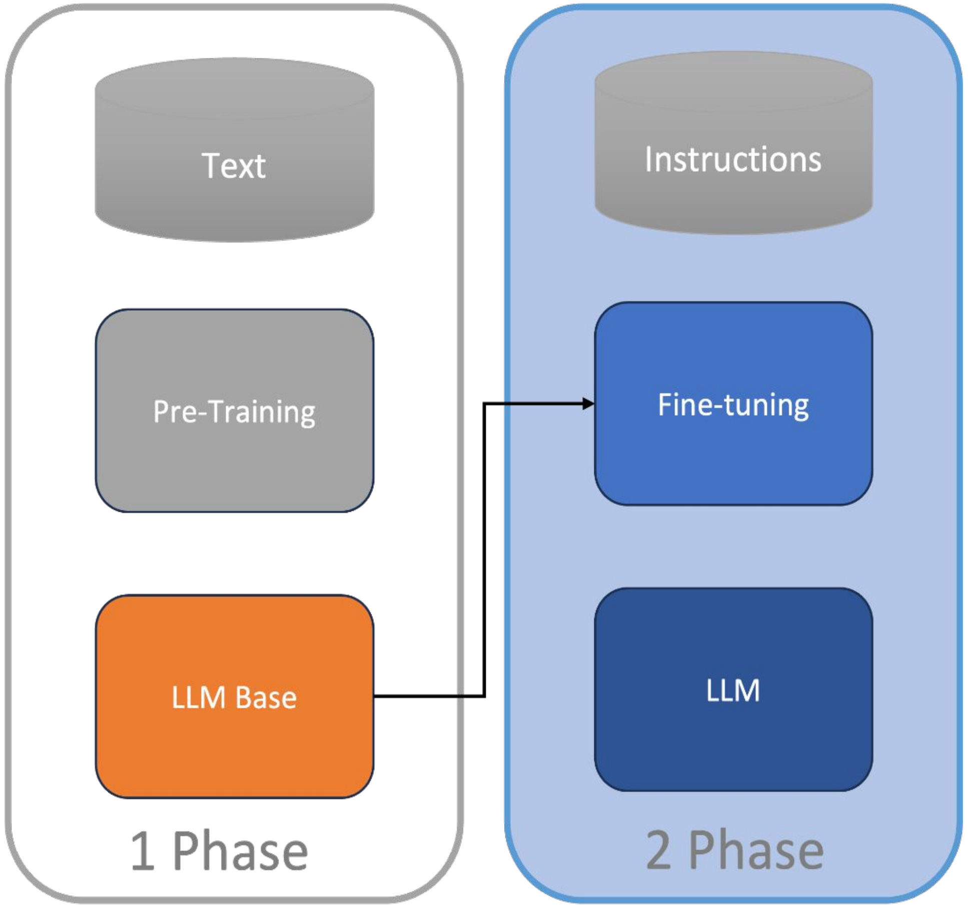
- Path more probable -> deterministic (always the same result)
- Roll a die to choose the next Word (token) based on their probabilities



Chatbots

- Additional training (i.e., fine-tuning) to learn how to solve instructions (Q-A)

- The data provided for this new training is pairs of instructions with your solution

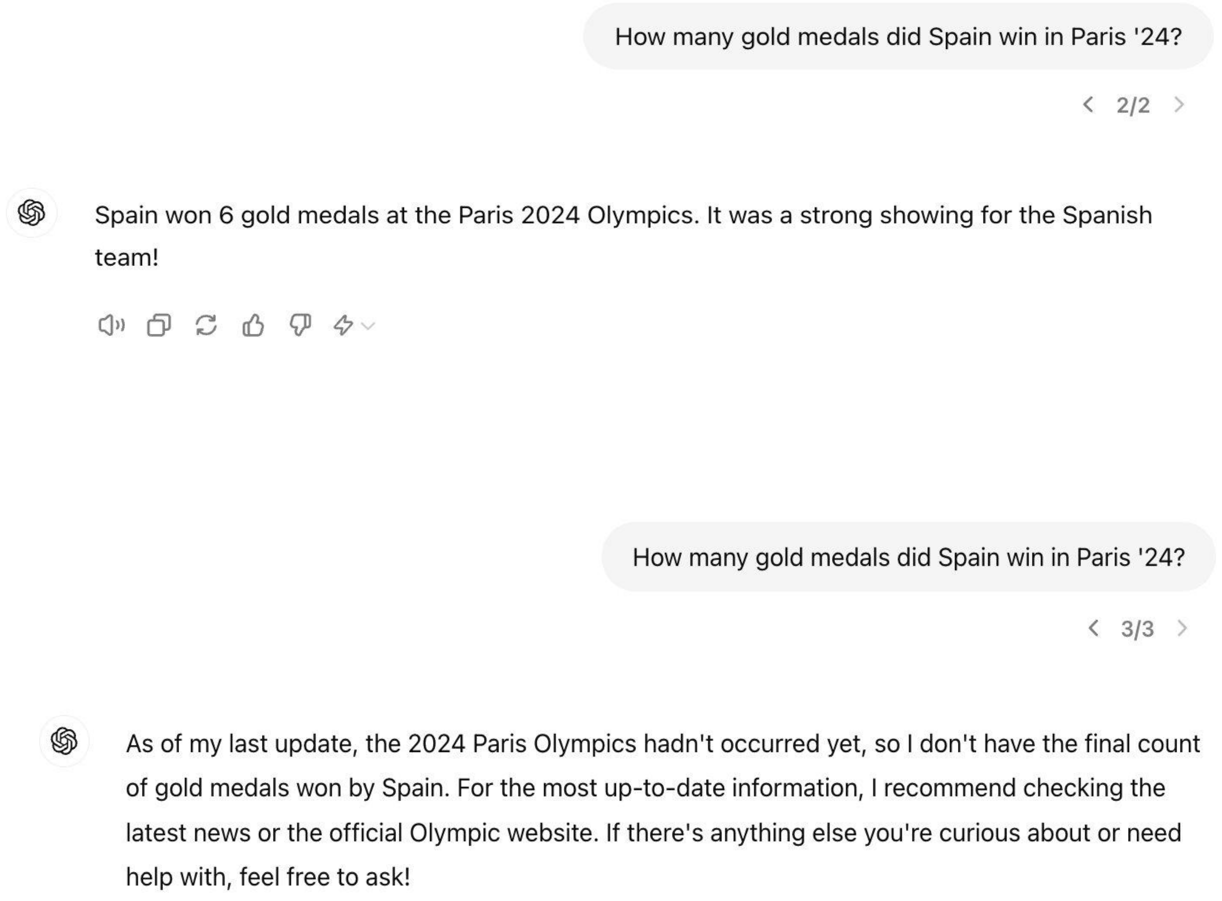


Limitations: Context Window

- Amount of information (tokens) that the model is able to process (remember) in a single session (conversation)
- LLMs are able to answer very complex questions because they have been trained with a lot of information
- But they are unable to remember what you said to them a few hours ago, or minutes ago…
- The Context Window Is Getting Bigger and Bigger (1 h of video, 700 k words)

Limitations of LLMs

- Hallucinations
- Traceability
- Information obsolescence
- Lack of expertise in some areas



How many gold medals did Spain win in Paris '24?

Spain won 6 gold medals at the Paris 2024 Olympics. It was a strong showing for the Spanish team!

How many gold medals did Spain win in Paris '24?

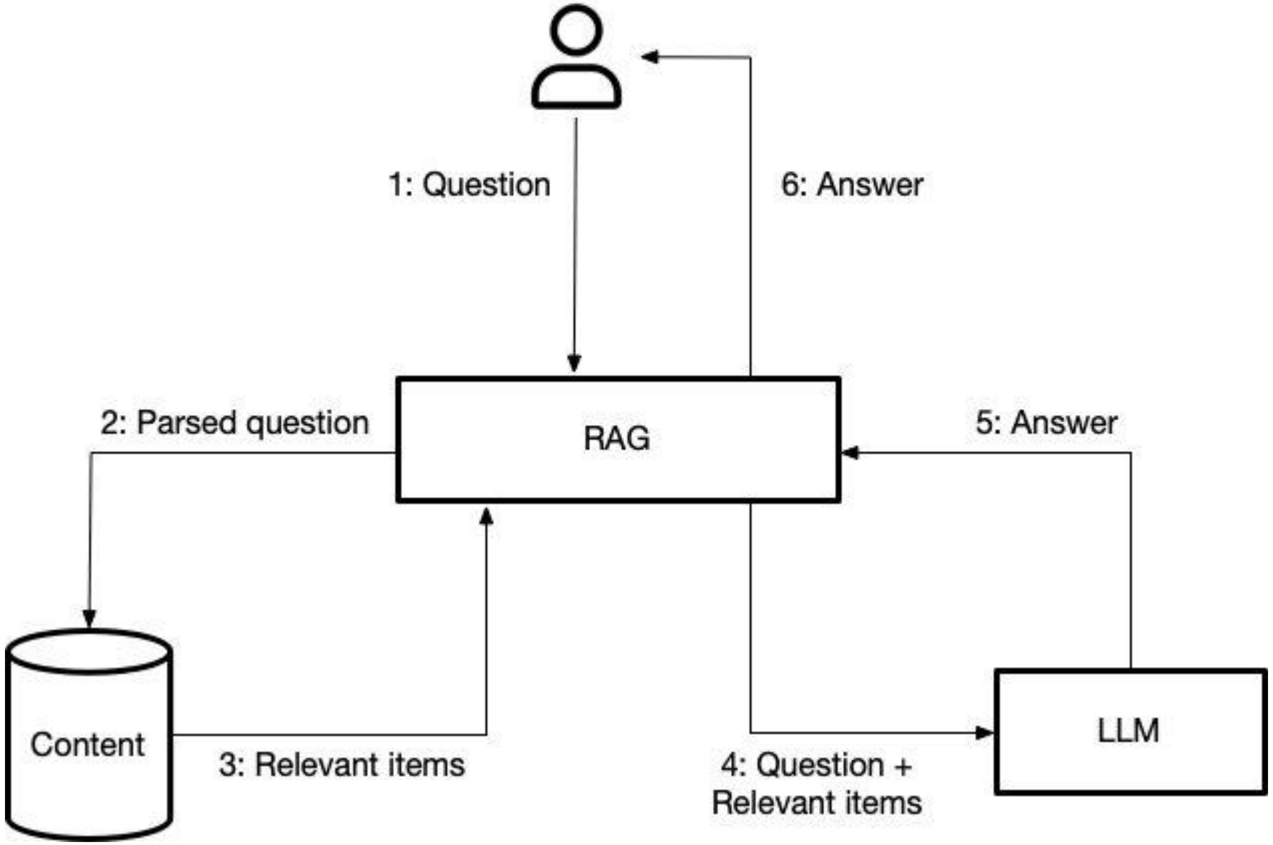
As of my last update, the 2024 Paris Olympics hadn't occurred yet, so I don't have the final count of gold medals won by Spain. For the most up-to-date information, I recommend checking the latest news or the official Olympic website. If there's anything else you're curious about or need help with, feel free to ask!

Limitations: Context Window and Knowledge of the LLM

To solve these problems there are alternatives

- Fine-tuning: retrain-your model with additional data (not valid for real-time data, and not easy to obtain good results)
- Retrieval Augmented Generation architectures: the model receives the information from an external system and answers the question with that information received

Retrieval Augmented Generation (RAG)

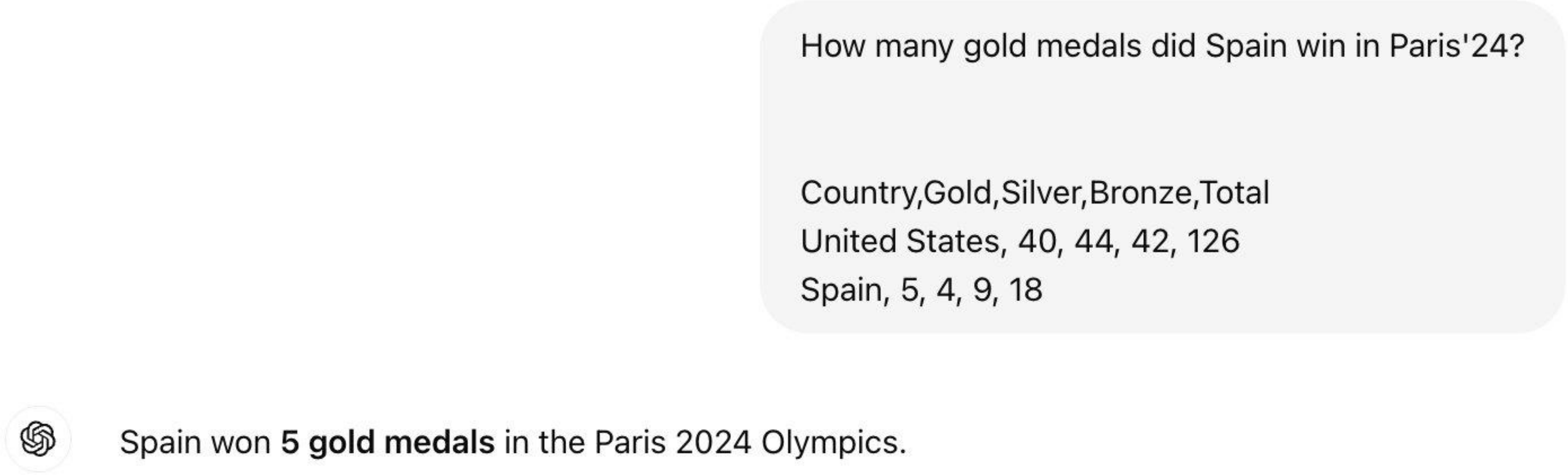


Agenda

1. Introduction to Generative AI and LLM
2. **Retrieval Augmented Generation (RAG) and FIWARE**
3. Implementation in FIWARE
4. DEMO: Naples Tourism use case
5. Summary

Retrieval Augmented Generation (RAG)

Example



How many gold medals did Spain win in Paris'24?

Country,Gold,Silver,Bronze,Total
United States, 40, 44, 42, 126
Spain, 5, 4, 9, 18

Spain won **5 gold medals** in the Paris 2024 Olympics.

Retrieval Augmented Generation (RAG)

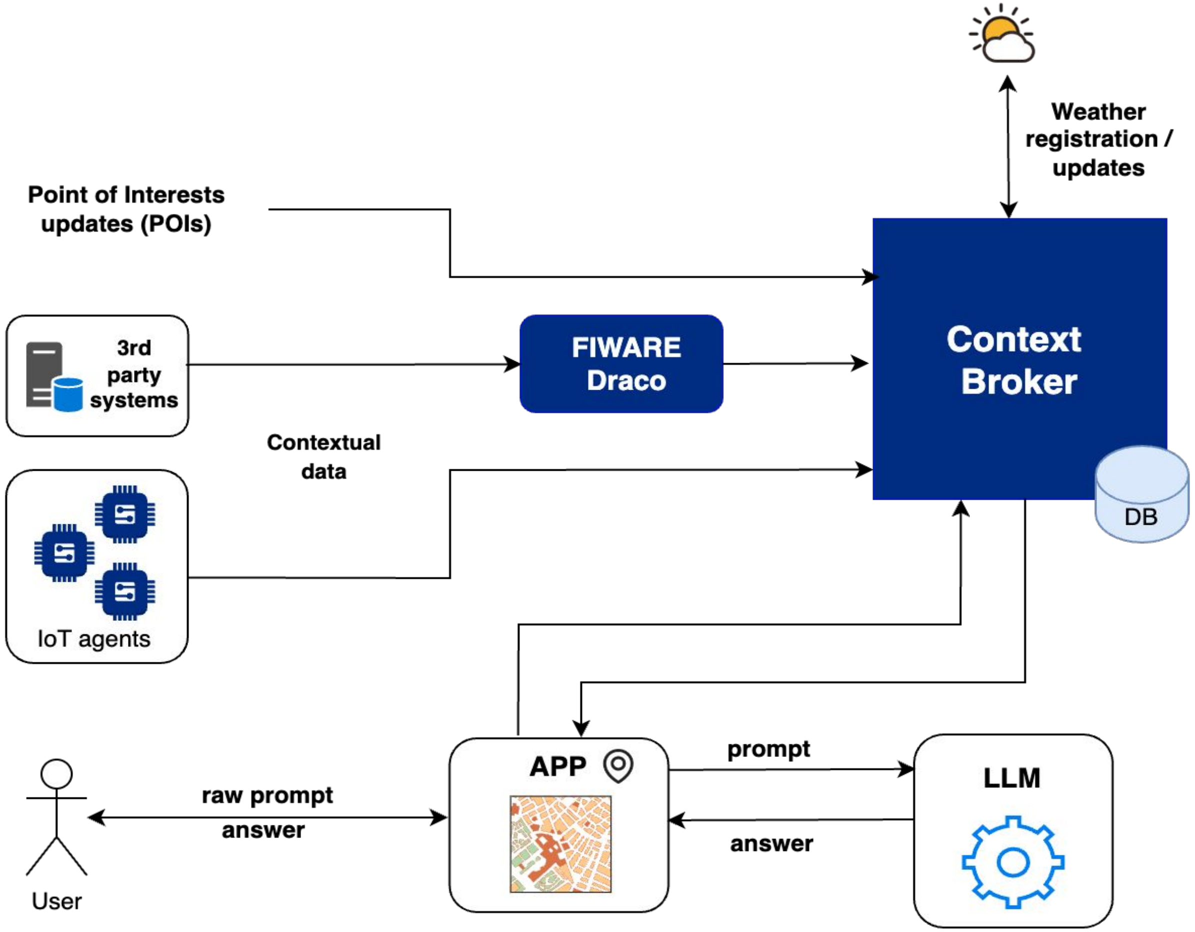
- Typically the information is stored in a vector database
  1. Information is divided in chunks
  2. These chunks are stored in vector databases
  3. When the user makes a question the most semantic similar chunks are retrieved
  4. The question + the chunks are passed to the LLM
  5. The LLM answers the question using only the information from the chunks

- But this is not the only option…

FIWARE and RAG

- We can use the Context Broker as the source of content
- Advantages:
    - Modeled in NGSI -> The LLMs have been trained with NGSI entities so they understand them
    - Up-to date information
    - Filter by geo-location, time, entity type, properties, etc.
    - Possibility to build NGSI queries using the LLM (not explored yet)

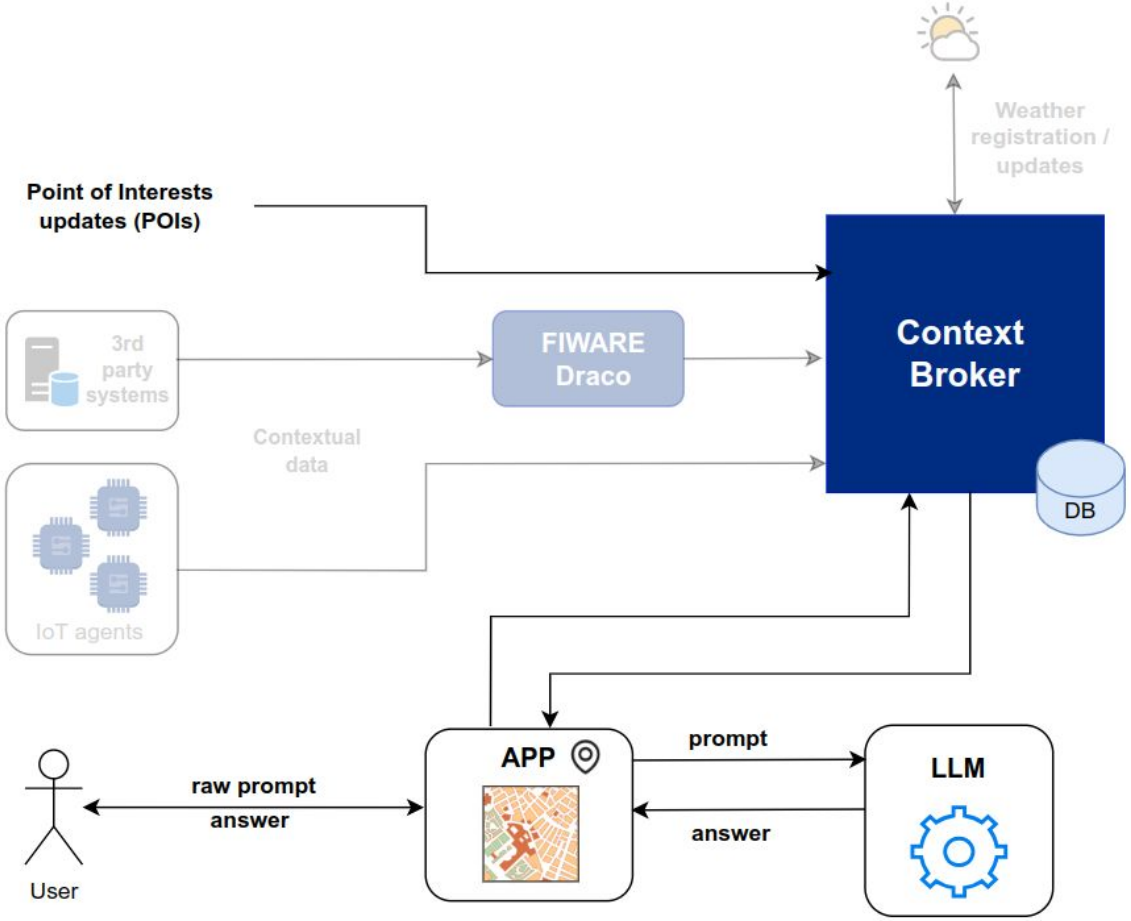
FIWARE in Real—Time Spatial RAG



Agenda

1. Introduction to Generative AI and LLM
2. Retrieval Augmented Generation (RAG) and FIWARE
3. **Implementation in FIWARE**
4. DEMO: Naples Tourism use case
5. Summary

FIWARE in Real—Time Spatial RAG: TOURISM USE CASE

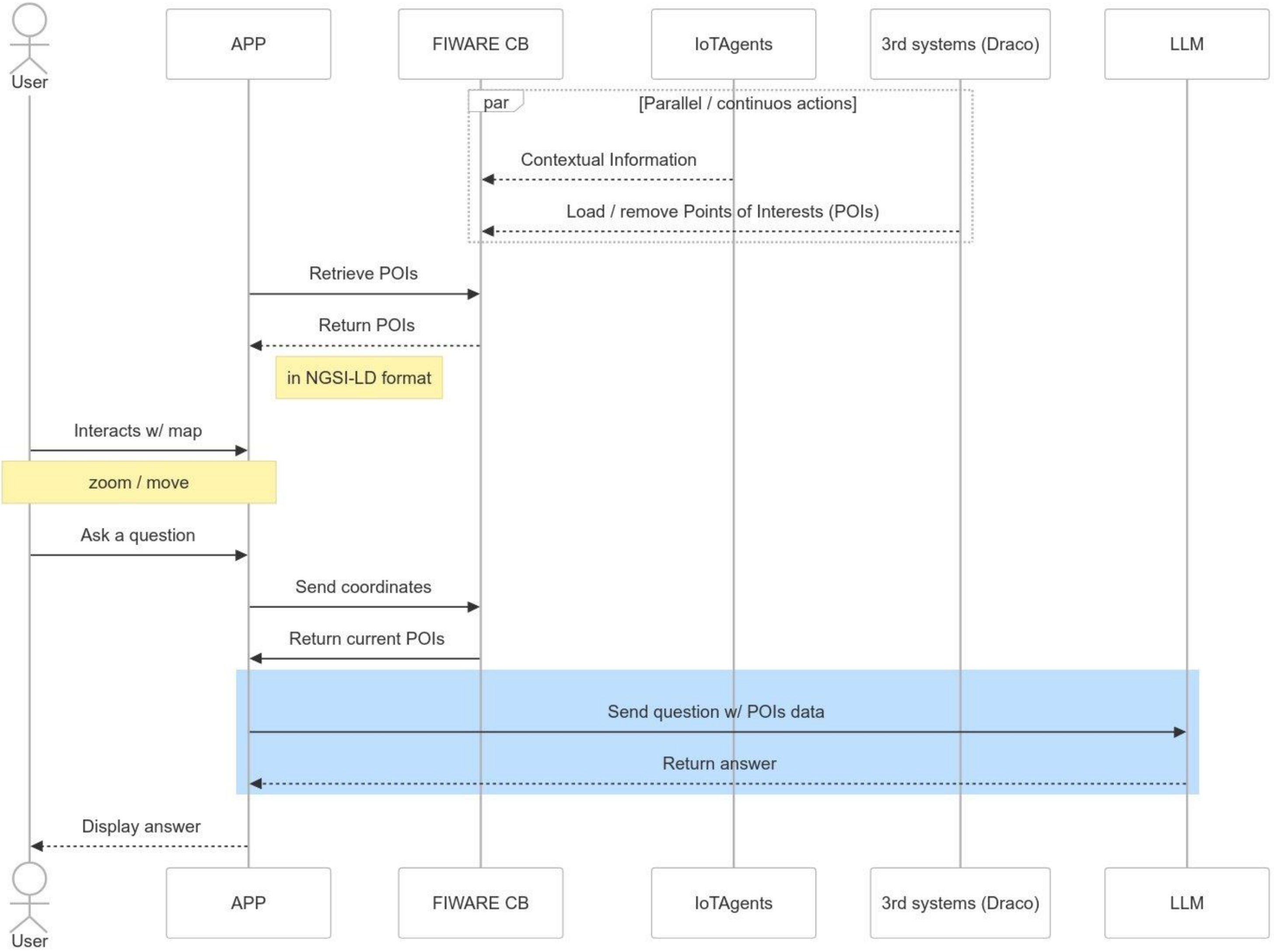


FIWARE in Real—Time Spatial RAG: TOURISM USE CASE. Entities

```
{
  "id": "urn:ngsi-ld:PoI:Naples:001",
  "type": "PoI",
  "title": {
    "type": "Property",
    "value": "Pompeii Archaeological Site"
  },
  "relevance": {
    "type": "Property",
    "value": 0
  },
  "image": {
    "type": "Property",
    "value": "pompei.jpg"
  },
  "location": {
    "type": "GeoProperty",
    "value": {
      "type": "Point",
      "coordinates": [
        14.487198,
        40.749351
      ]
    }
  },
```

```
"price": {
     "type": "Property",
     "value": "16€"
   },
   "description": {
     "type": "Property",
     "value": "Pompeii is a vast archaeological site that once
was a thriving Roman city. It was buried under meters of ash
after the catastrophic eruption of Mount Vesuvius in 79 A.D."
   },
   "capacity": {
     "type": "Property",
     "value": 30000
   },
   "occupancy": {
     "type": "Property",
     "value": 12000
   }
 }
```

FIWARE in Real—Time Spatial RAG: TOURISM USE CASE. Interactions



Agenda

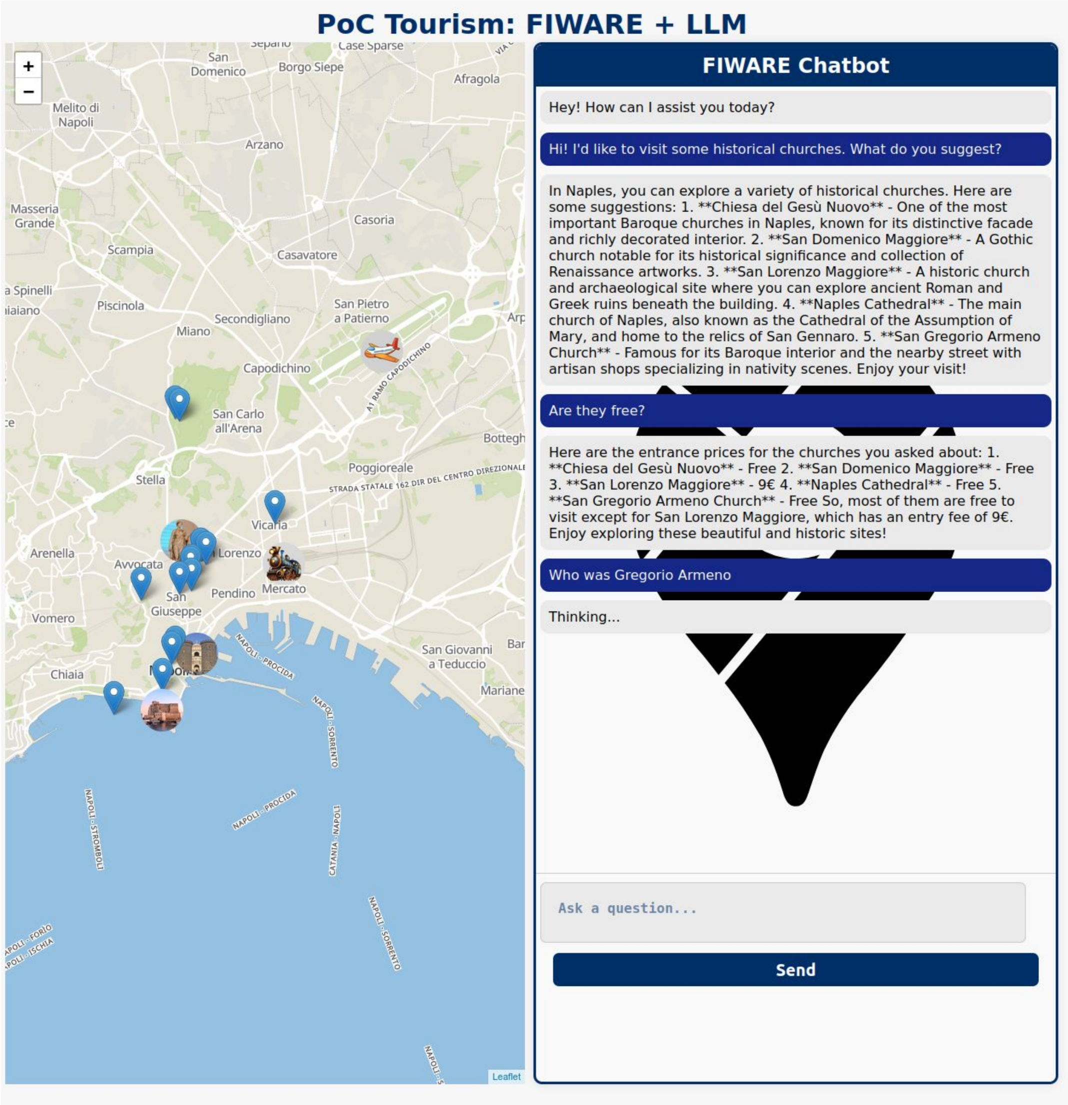
1. Introduction to Generative AI and LLM
2. Retrieval Augmented Generation (RAG) and FIWARE
3. Implementation in FIWARE
4. **DEMO: Naples Tourism use case**
5. Summary

Demo

DEMO: during the demo we will:

- Ask info about current place
- Reformulate to refine the inquiry
- Drill down the question
- Explore options
- Iterate over possible options
- Change language to ask more information

Demo: Interactive part



Hey! How can I assist you today?

Hi! I'd like to visit some historical churches. What do you suggest?

In Naples, you can explore a variety of historical churches. Here are some suggestions: 1. **Chiesa del Gesù Nuovo** - One of the most important Baroque churches in Naples, known for its distinctive facade and richly decorated interior. 2. **San Domenico Maggiore** - A Gothic church notable for its historical significance and collection of Renaissance artworks. 3. **San Lorenzo Maggiore** - A historic church and archaeological site where you can explore ancient Roman and Greek ruins beneath the building. 4. **Naples Cathedral** - The main church of Naples, also known as the Cathedral of the Assumption of Mary, and home to the relics of San Gennaro. 5. **San Gregorio Armeno Church** - Famous for its Baroque interior and the nearby street with artisan shops specializing in nativity scenes. Enjoy your visit!

Are they free?

Here are the entrance prices for the churches you asked about: 1. **Chiesa del Gesù Nuovo** - Free 2. **San Domenico Maggiore** - Free 3. **San Lorenzo Maggiore** - 9€ 4. **Naples Cathedral** - Free 5. **San Gregorio Armeno Church** - Free So, most of them are free to visit except for San Lorenzo Maggiore, which has an entry fee of 9€. Enjoy exploring these beautiful and historic sites!

Who was Gregorio Armeno

Thinking...

Agenda

1. Introduction to Generative AI and LLM
2. Retrieval Augmented Generation (RAG) and FIWARE
3. Implementation in FIWARE
4. DEMO: Naples Tourism use case
5. **Summary**

Summary

- **LLM** ⊂ GenAI ⊂ Deep Learning ⊂ Machine Learning ⊂ Artificial Intelligence
- **RAG** (Retrieval-Augmented Generation) in LLMs refers to a technique that combines large language models with external retrieval systems to fetch relevant documents or data, used to enhance the model's response generation
- We have used **FIWARE** to perform the **RAG** and support the **LLM** with contextual data from a city
- The **CB** is used as a central piece that handles the contextual data
- This sets up the basis for a general architecture that could be used to enhance the experience in tourism and could be the ground for other types of **powered-by-FIWARE** solutions that could be improved to interact with users.

Resources

- Link to the code and slides
  - https://github.com/dncampo/Poc_LLM_CB_FIWARE-Naples

Thank you!

http://fiware.org
Follow @FIWARE on Twitter

FIWARE