



Sviluppo di Agenti IA

Corso Base

AGENDA

- Introduzione al corso
- Presentazione degli strumenti di supporto alla formazione
- Concetti generali sulle applicazioni agentiche
 - LLM
 - Prompt engineering
 - Embeddings
 - Tool
 - Agenti
- Applicazioni RAG
- Processo di sviluppo di Assistenti IA
 - Implementazione del vectorstore
 - Implementazione di un ChatBot RAG stateless
 - Implementazione di un ChatBot RAG conversazionale
 - Implementazione di un assistente agentico
- Scenari di deployment in produzione

Strumenti di supporto alla R&D e alla Formazione



LangChain

Motivazioni alla base della scelta degli strumenti

Perché aziende come la **Fastal**, che hanno una consolidata esperienza di sviluppo sui framework **PHP**, **Java** e **Java Enterprise** usano il **Python** per le attività di ricerca e sviluppo sull'**IA**?



Motivazioni alla base della scelta degli strumenti

Python è diventato il linguaggio di elezione per l'IA

il vero motivo è molto semplice

i **gruppi di lavoro** che usavano **Python** per realizzare framework e API per i nuovi sistemi IA hanno raggiunto i risultati di progetto molto prima degli altri.

Motivazioni alla base della scelta degli strumenti

Python è:

- interpretato
- fortemente interattivo
- intuitivo e lineare nella sintassi
- facile da leggere
- molto espressivo (poche istruzioni per fare molto)

Per leggere il Python non serve conoscere il Python

```
public class HelloWorld {  
    public static void main(String[] args){  
        System.out.println("Hello, world!");  
    }  
}
```

Se sai leggere
questo...

... non avrai
problemi a
leggere questo

```
print('Hello, world!')
```

Motivazioni alla base della scelta degli strumenti

In virtù di queste qualità del Python:

- Sono stati sviluppati molti strumenti, framework e librerie per l'IA
- Sono disponibili framework di astrazione che facilitano la ricerca e lo sviluppo di nuove applicazioni robuste
- I progetti di applicazioni agentiche costano meno

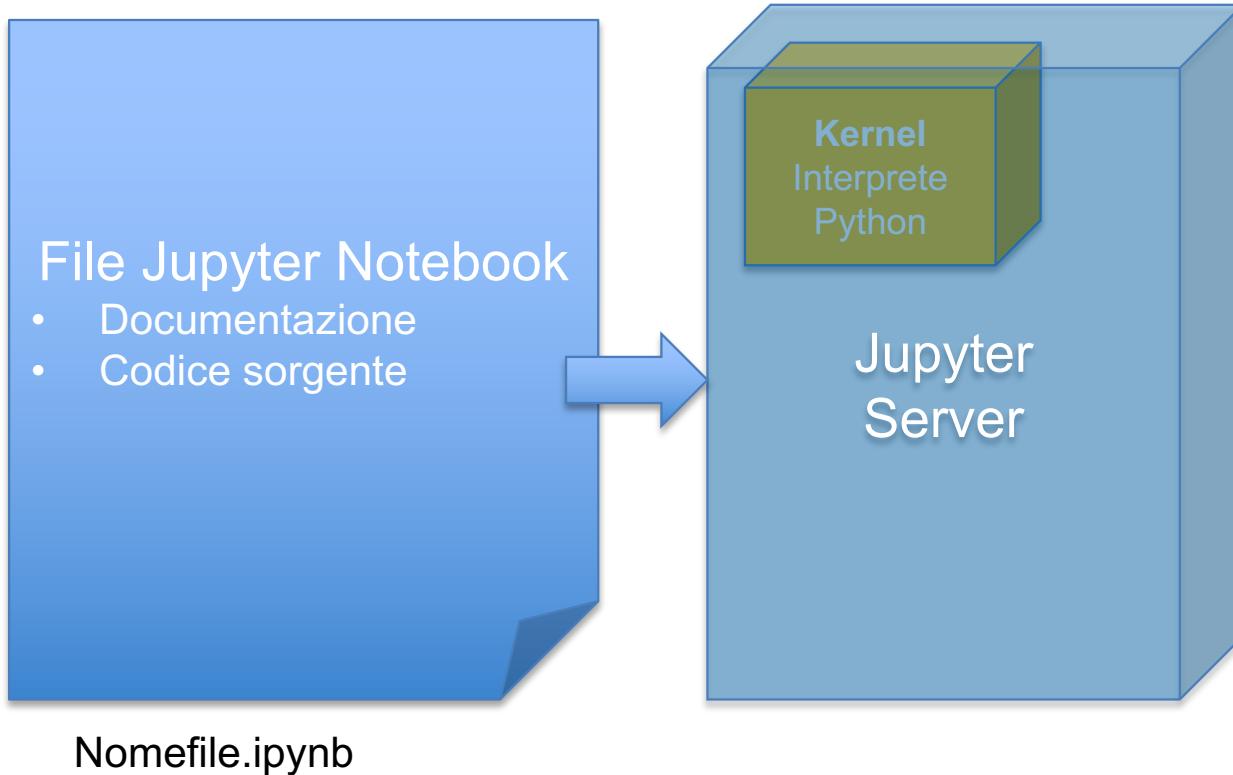
Strumento di condivisione del processo di sviluppo

File Jupyter Notebook

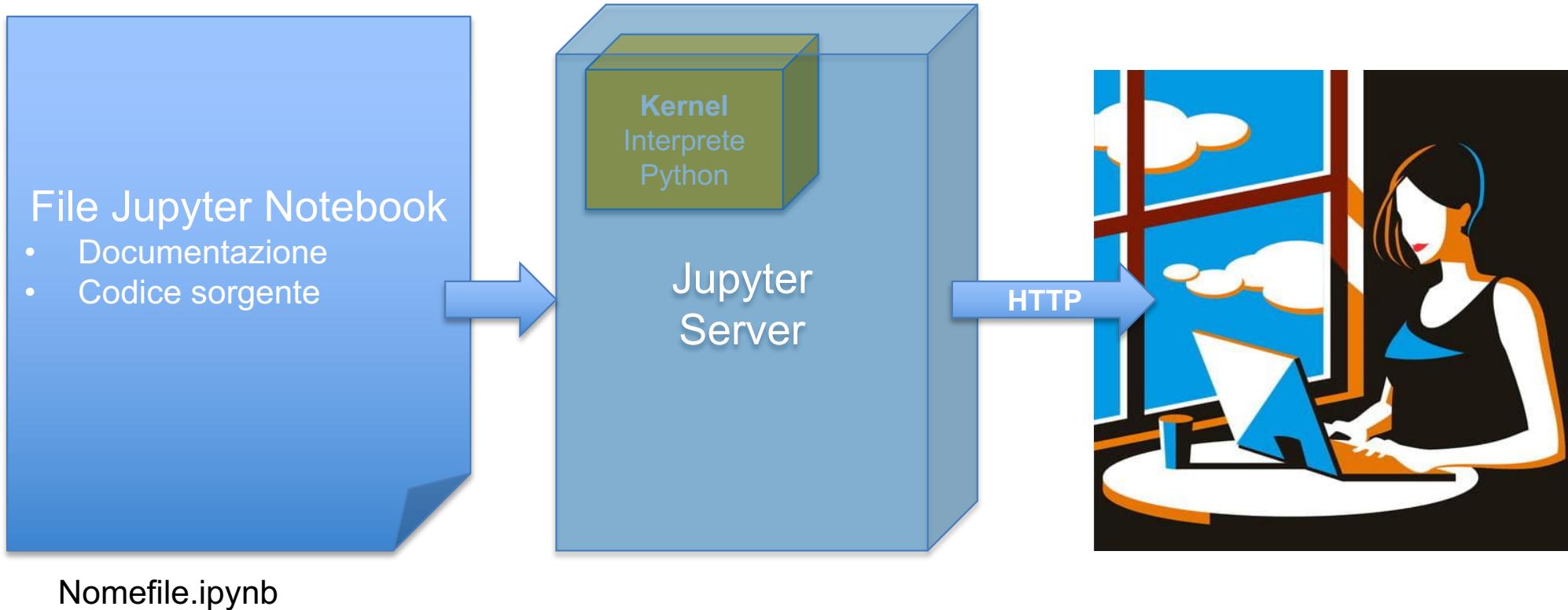
- Documentazione
- Codice sorgente

NomeFile.ipynb

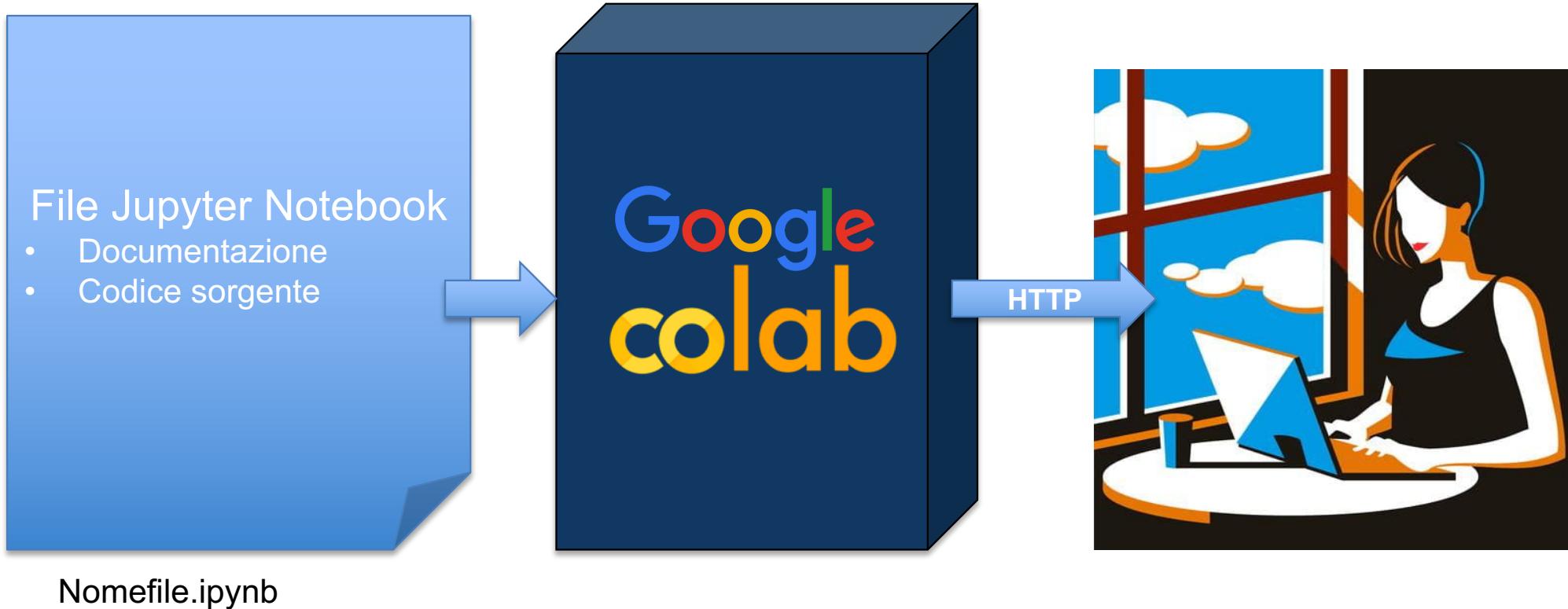
Jupyter Server



Jupyter Server



Jupyter Server

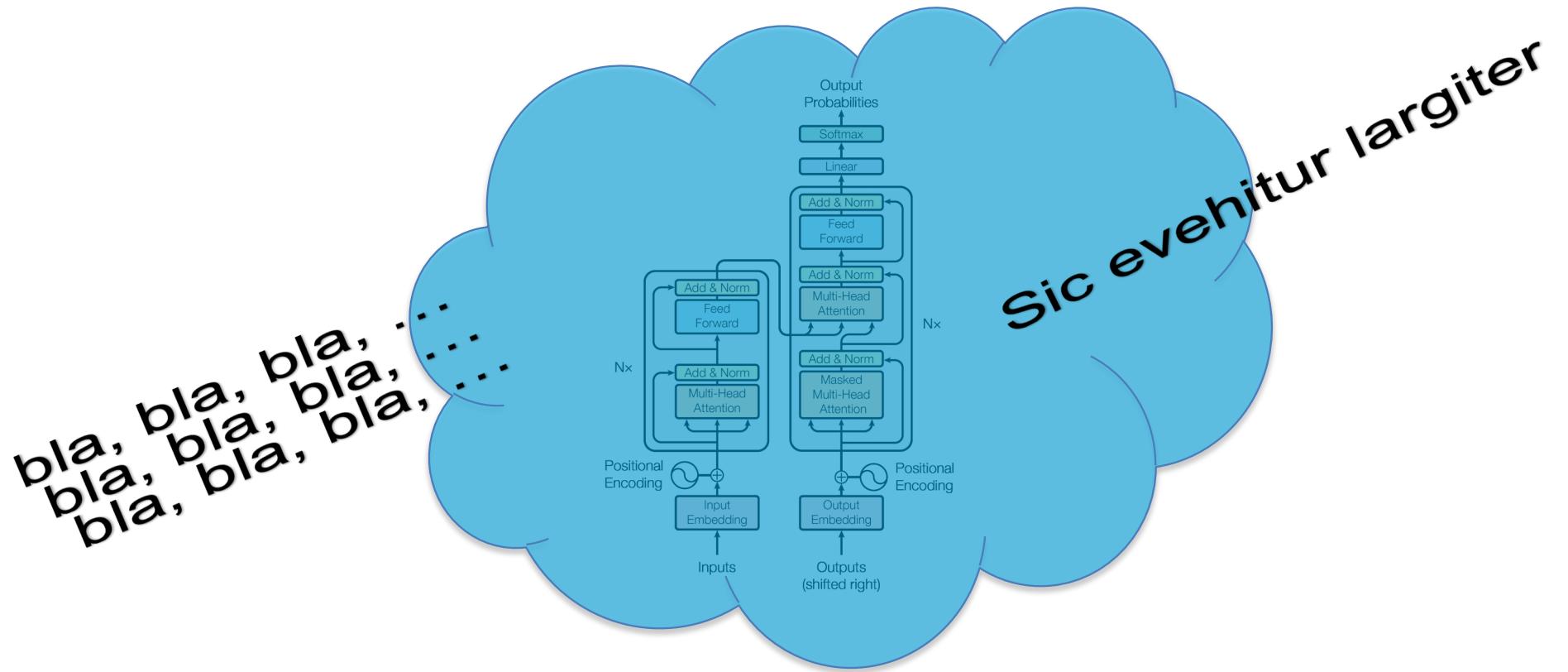


Accesso a Google Colab

<https://colab.research.google.com/>

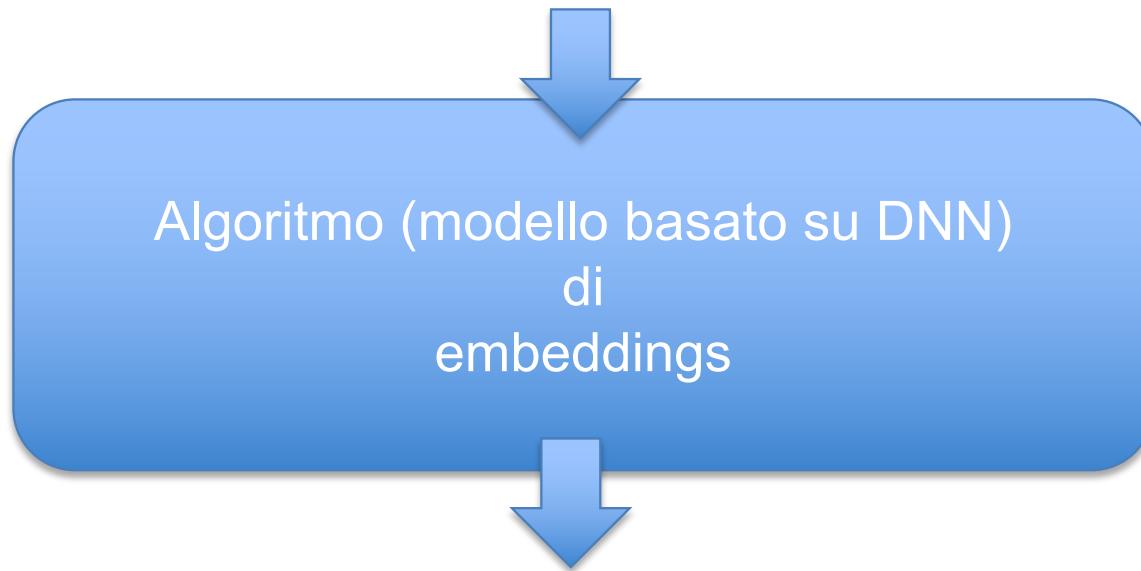


Large Language Model



Embeddings

Che tempo fa oggi a Parigi?



[1.7 -0.3, 12.0, , 0.27, -2.14]

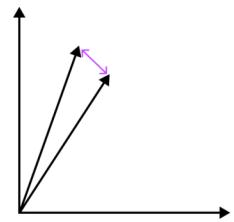
Vettore di embeddings a n dimensioni

Embeddings

Similarity Metrics for Vector Search | Zilliz

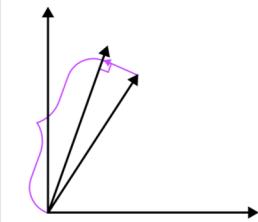
Euclidean Distance

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$



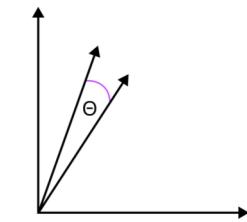
Inner Product

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i$$



Cosine Similarity

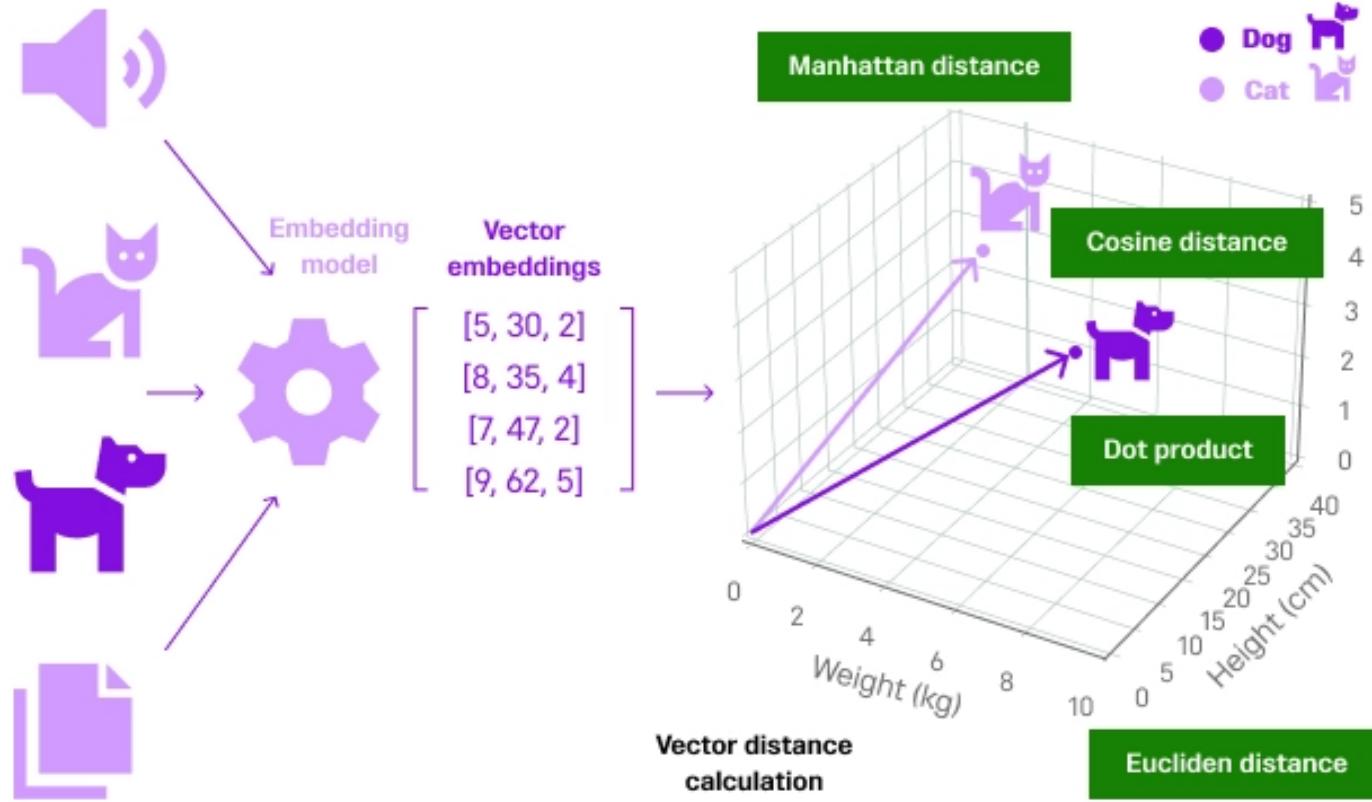
$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$



Distanza matematica tra vettori di embeddings
=

Correlazione semantica tra i testi

Esempio di similarity



Schema logico base LLM

Che tempo fa oggi a Parigi?

Layer 1

Richiesta informazioni metereologiche Luogo specifico Parigi

Layer 2

Richiesta conoscenza eventi tempo reale Meteorologia richiede dati esterni
Luogo specifico Parigi

Layer 3

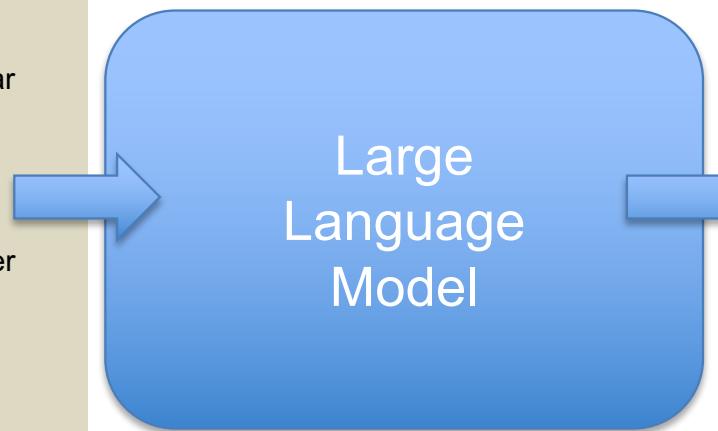
Mancanza conoscenza Risposta di cortesia Meteo Parigi

Layer n

Scusa, ma non ho modo di conoscere il meteo di Parigi.

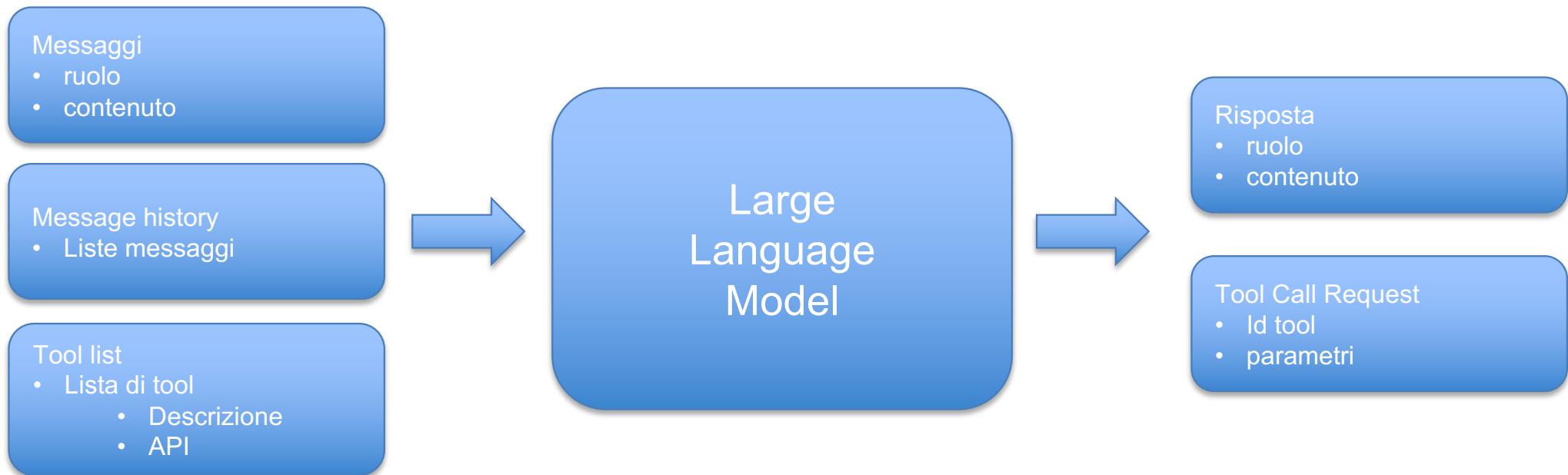
LLM State Of the ART - LLM SOA o ChatLLM

```
{ "input": { "user_query": "Trova l'ultimo articolo scientifico sulla fusione nucleare e genera un riassunto.",  
          "tools": [ { "name": "browser",  
                      "action": { "type": "search",  
                                 "query": "latest scientific article on nuclear fusion" } },  
          { "name": "text_summarizer",  
            "action": {  
                "type": "summarize",  
                "input_text": "Full article text from browser tool" } },  
          "context": {  
              "user_info": {  
                  "name": "Chirale",  
                  "preferences": {  
                      "language": "Italian",  
                      "summary_length": "short" } },  
              "conversation_history": [ { "query": "Spiegami come funziona la fusione nucleare",  
                                         "response": "La fusione nucleare è un..." } ] } } }
```



```
{"response": { "text": "Ho trovato un recente articolo ....",  
             "tools_output": [ { "name": "browser",  
                               "action": "search",  
                               "result": {  
                                   "title": "Recent advancements in nuclear fusion",  
                                   "url": "https://www.scientificjournal.com/nuclear-fusion-2024",  
                                   "summary": "This article discusses ...." } } ],  
             "name": "text_summarizer",  
             "action": "summarize",  
             "result": { "original_text": "The full article...",  
                        "summary": "The article highlights ...." },  
             "context_update": {  
                 "conversation_history": [ { "query": "Trova l'ultimo articolo scientifico sulla fusione nucleare e genera un riassunto.",  
                                             "response": "Ho trovato un recente articolo..." } ] } } }
```

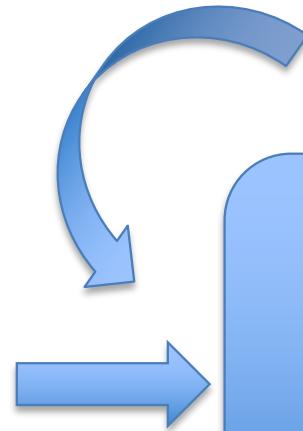
LLM State Of the ART - LLM SOA o ChatLLM



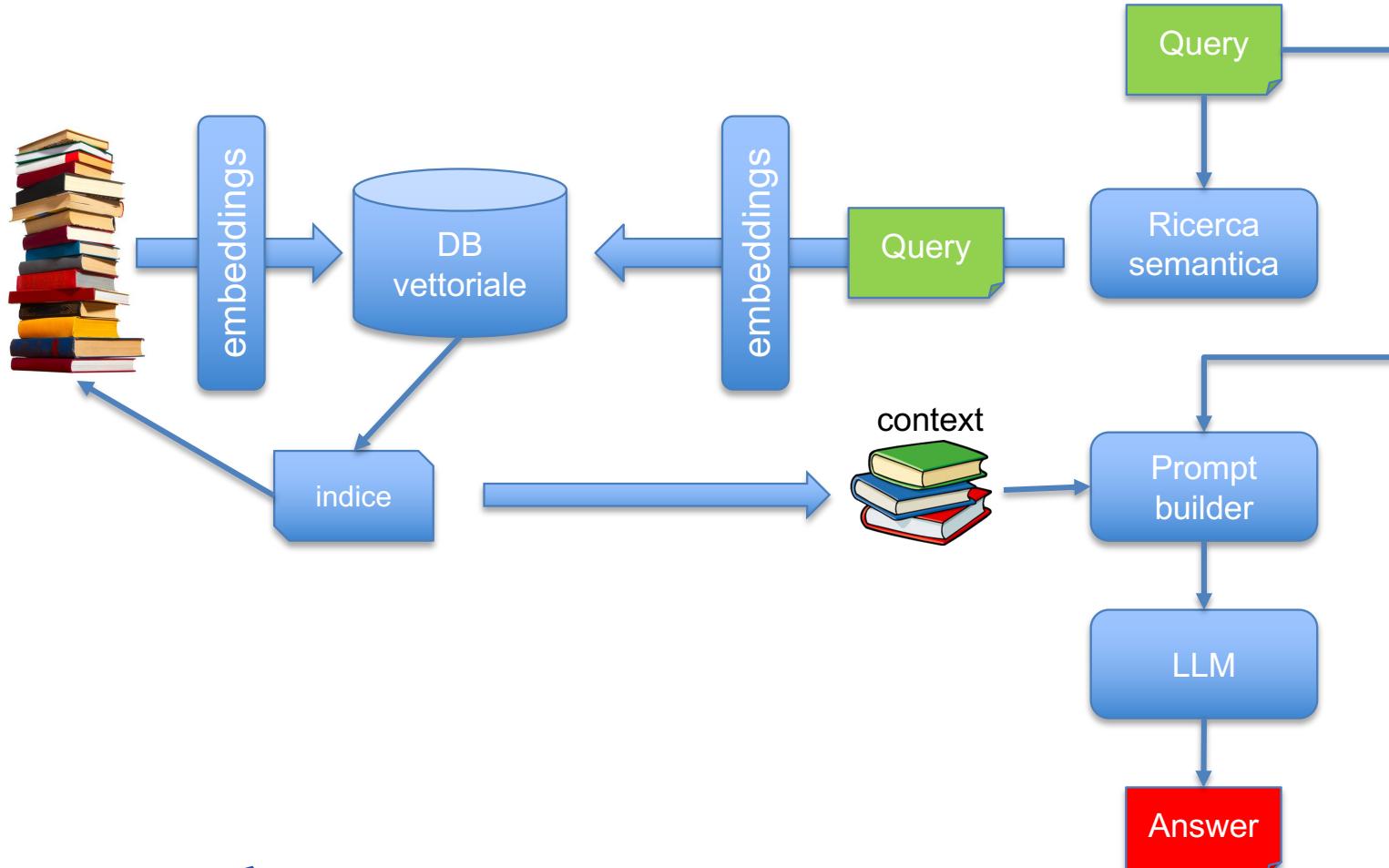
Come si usano i LLM nello sviluppo di applicazioni?

La realizzazione di un Assistente di tipo ChatBot specializzato su un compito specifico è un esempio di applicazione sviluppabile con i LLM

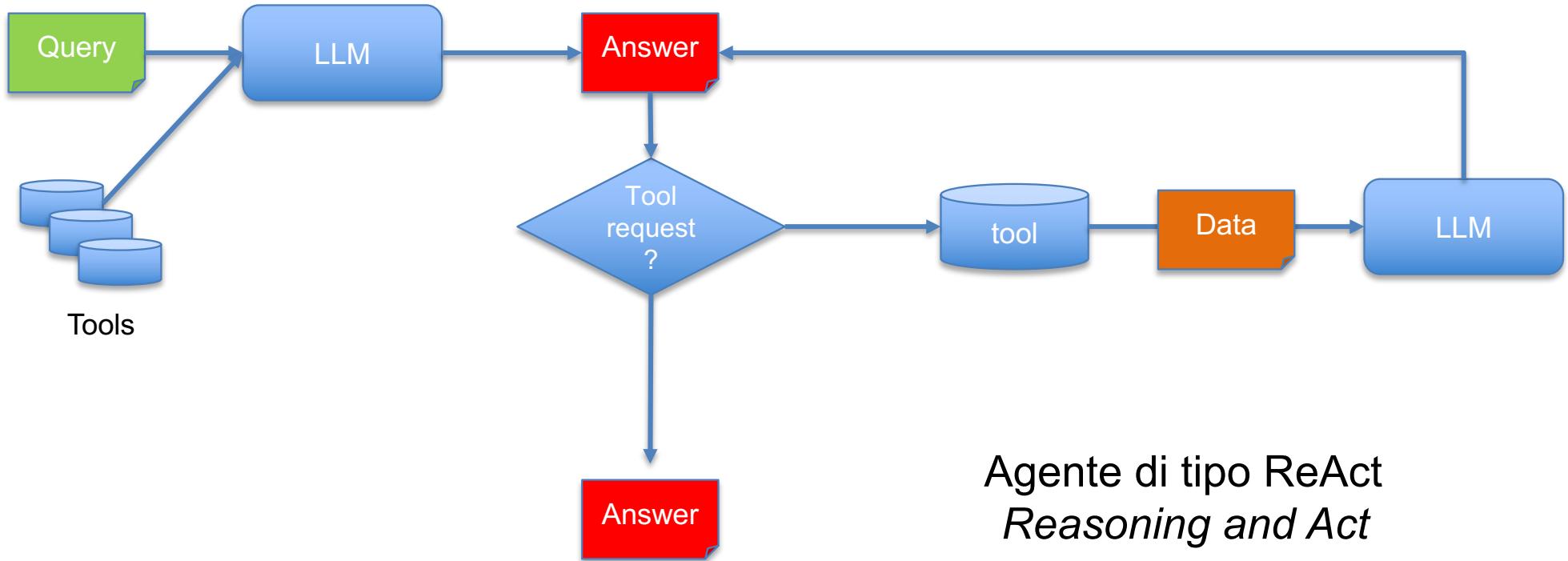
Specializzazione del LLM mediante training



ChatBot stateless RAG



Agenti – il flusso dell'elaborazione dipende dal LLM

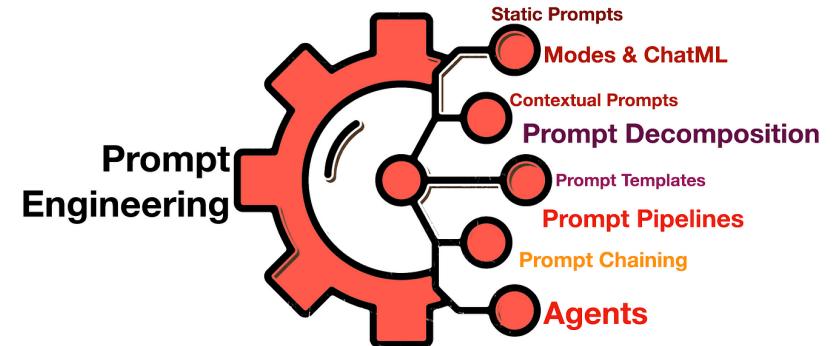


Criticità dello sviluppo di applicazioni basate su LLM



API complesse
Assenza di standard

Scenario in rapida evoluzione
Molteplicità di prodotti leader
Prestazioni in rapido miglioramento



LangChain Framework



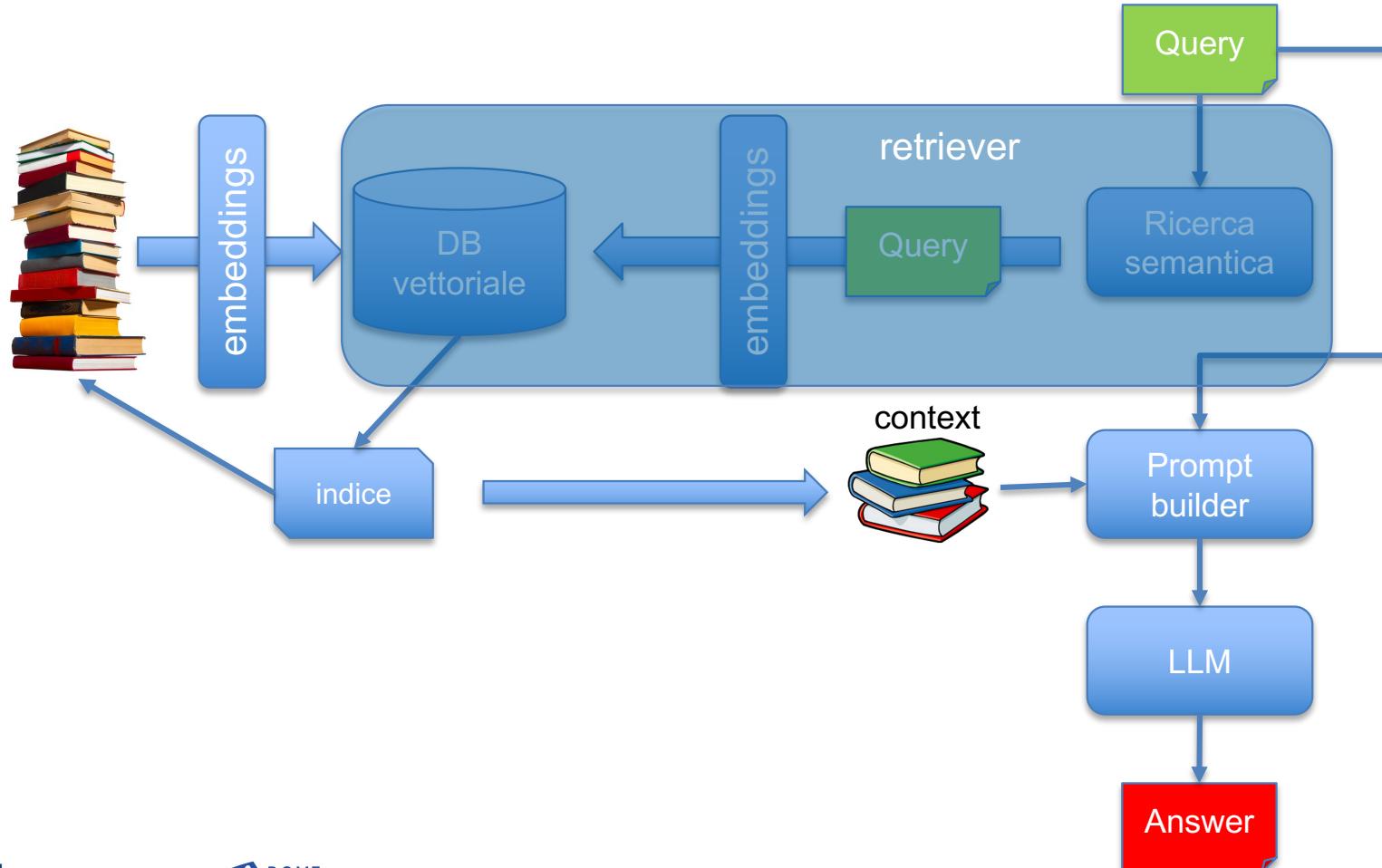
- Framework di astrazione con API di alto livello
- Indipendente dai provider LLM
- Integrazione con la quasi totalità dei provider
- Integrazione verso quasi tutti i DBMS vettoriali
- API di alto livello per realizzare «chain»
- API di alto livello per realizzare «workflow» (agenti)
- Ricca libreria di schemi applicativi «prebuild»



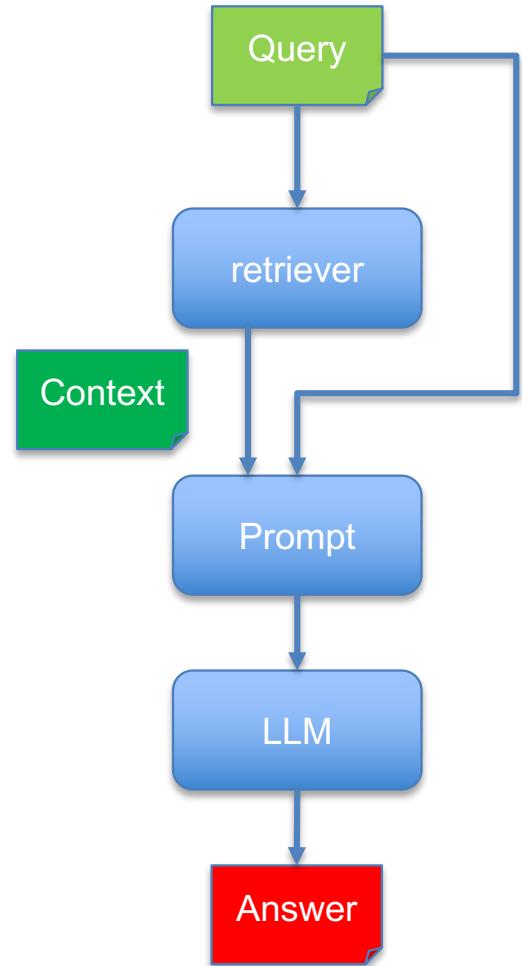
Andiamo al Notebook...



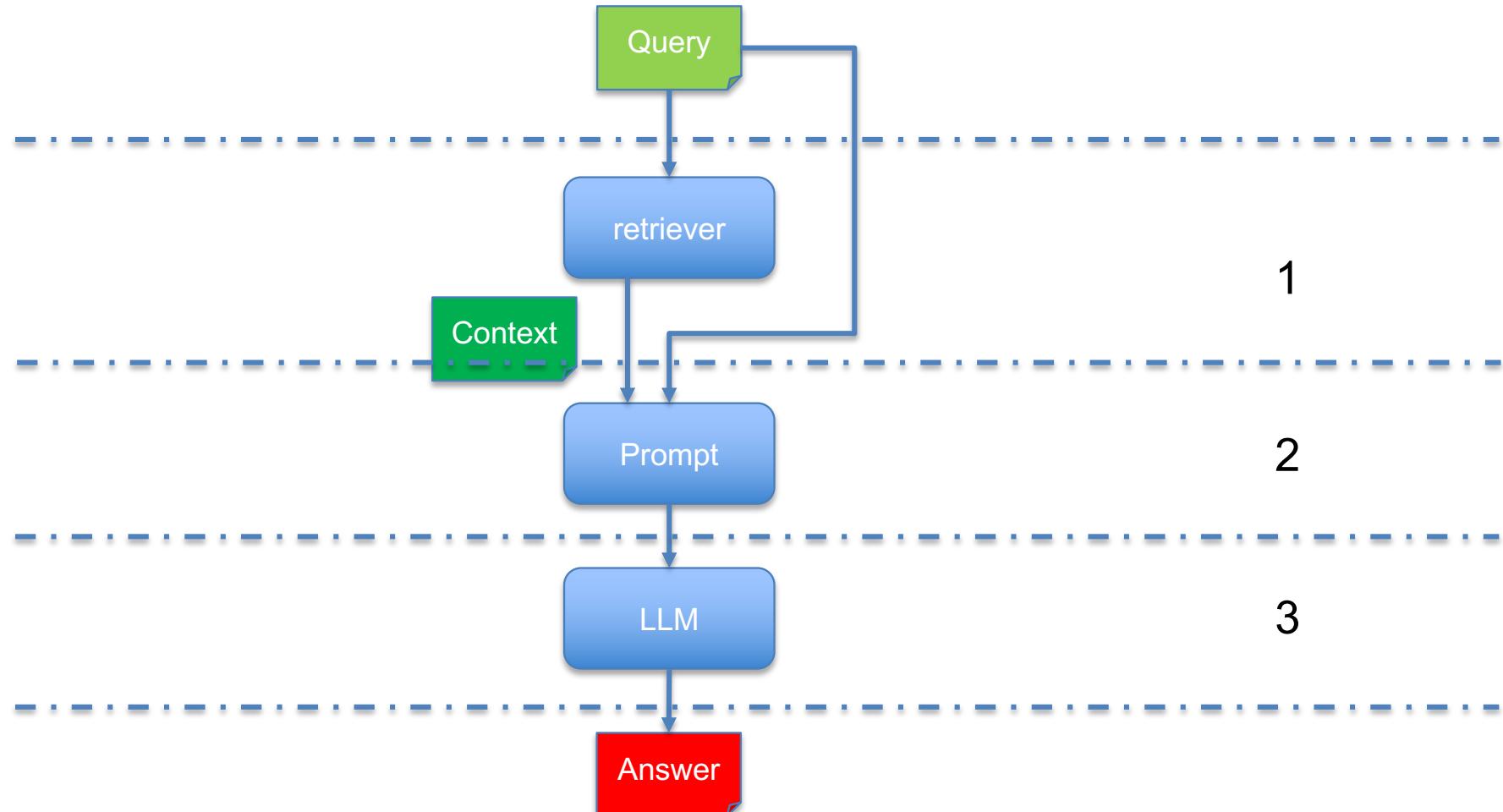
LangChain ChatBot stateless RAG



LangChain ChatBot stateless RAG



LangChain ChatBot stateless RAG



LangChain ChatBot stateless RAG

