



# ICT Training Center

**Il tuo partner per la Formazione e la Trasformazione digitale della tua azienda**



## Note



# SPRING AI

## GENERATIVE ARTIFICIAL INTELLIGENCE CON JAVA

Simone Scannapieco

Corso avanzato per Venis S.p.A, Venezia, Italia

Novembre 2025

## Note

- ➡ **Simple prompting:** utente costruisce il contesto manualmente
    - ⚠ LLM genera una risposta in base alla conoscenza su cui è stato addestrato (*knowledge cut-off*)
    - ⚠ Impensabile pre-addestrare LLM su qualsiasi sotto-dominio specifico
    - ⚠ Quando informazione contestuale non è sufficiente, LLM tende a generare risposte non veritieri (*hallucination*)
  - ➡ **Prompt stuffing:** contesto utente arricchito dinamicamente (es. *chat memory*)
    - ⚠ Prompt potenzialmente eccessivi in grandezza
    - ⚠ Problematiche di *context window*
    - ⚠ Costi esorbitanti con LLM commerciali
    - ⚠ Tempi di risposta proibitivi con LLM *open source*

## Note

# **RETRIEVAL AUGMENTED GENERATION**

## NATURA DINAMICA DEI CONTENUTI DIGITALI

10:11

ChatGPT >

Quanti nuovi articoli vengono generati su Wikipedia ogni giorno?

Il numero di nuovi articoli creati su Wikipedia ogni giorno varia, ma in media vengono creati circa **500-1.000 nuovi articoli** al giorno, considerando l'intera piattaforma in tutte le lingue. Tuttavia, il numero può oscillare a seconda delle lingue e della comunità attiva in quel momento.

La versione in inglese di Wikipedia è solitamente quella con la crescita più alta, ma ci sono molte altre lingue in cui la creazione di nuovi articoli avviene quotidianamente. Inoltre, non tutti gli articoli creati vengono mantenuti, poiché alcuni possono essere eliminati per non rispettare le linee guida della piattaforma.

Questa cifra varia anche in base a fattori come eventi speciali, iniziative di scrittura e la disponibilità di nuovi contributori.

10:13

ChatGPT >

E quante pagine vengono modificate o generate ogni giorno nell'intero World Wide Web?

Non esiste un dato esatto e univoco sul numero di pagine web che vengono modificate o generate ogni giorno su tutto il World Wide Web, poiché questa informazione varia costantemente ed è distribuita tra miliardi di siti web gestiti in modo decentralizzato. Tuttavia, possiamo fare alcune stime basate su ricerche e analisi precedenti:

- 1. Numero totale di pagine web:** Il numero complessivo di pagine web è in costante crescita e si stima che vi siano più di 1,9 miliardi di siti web attivi. Ogni sito può contenere da poche a milioni di pagine. Nel 2021, Google aveva indicizzato oltre 130 trilioni di pagine web.
- 2. Aggiornamenti giornalieri:** Secondo alcune stime, il numero significativo di pagine web ↓ che aggiornato

10:13

ChatGPT >

giornali online, e-commerce e piattaforme di contenuti, aggiornano continuamente i propri contenuti. Per esempio, piattaforme come Facebook, Twitter e YouTube vedono milioni di nuovi post e video caricati ogni giorno.

**4. Influenza degli algoritmi:** Molti siti generano pagine dinamicamente in risposta alle richieste degli utenti (ad esempio, motori di ricerca, e-commerce, social media). Queste pagine dinamiche possono non esistere fino a quando non vengono richieste e vengono create "al volo", il che rende difficile quantificare esattamente quante nuove pagine appaiono quotidianamente.

In sintesi, il numero di pagine modificate o generate quotidianamente sul web potrebbe essere nell'ordine di **decine di milioni** o addirittura più, ma non esiste un numero esatto facilmente calcolabile a causa della natura decentralizzata e dinamica di internet.

- Pre-addestramento costante per le LLM non è efficiente
  - Come fare con **fattoidi** (es. cariche politiche, popolazione per stato, . . . )?!

## Note

# **RETRIEVAL AUGMENTED GENERATION**

## RIVOLUZIONE NEL MONDO DELL'ADDESTRAMENTO LLM



Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Namen Goyal, Heinrich Kuttler, Mike Lewis, Wen-Tau Yin, Tim Rocktaschel, Sebastian Riedel, Douwe Kiela. *Retrieval-Augmented Generation of Knowledge-Intensive NLP Tasks*.

ArXiv, 2021.



11:43 4G

ChatGPT > ⚙️

Spiegami brevemente cosa è RAG

🕒 RAG (Retrieval-Augmented Generation) è una tecnica avanzata nell'ambito dell'intelligenza artificiale e del Natural Language Processing (NLP). Combina due approcci principali:

1. **Retrieval (Recupero):** Viene utilizzato un sistema di ricerca per estrarre informazioni pertinenti da una grande base di dati o documenti, che potrebbero essere rilevanti per una determinata query o richiesta.

2. **Generation (Generazione):** Dopo aver recuperato queste informazioni, viene utilizzato un modello di generazione di linguaggio naturale (come GPT) per sintetizzare e presentare una risposta ben strutturata basata sia sulle informazioni recuperate che sulle proprie capacità generative.

In pratica, RAG consente ai modelli di intelligenza artificiale di combinare il potere della ricerca documentale con la capacità di generare testi coerenti e dettagliati, risultando in risposte più accurate e informate, soprattutto in contesti dove è necessario un supporto fattuale preciso.

- Nuova frontiera del *prompt engineering*
  - Riduce i costi di addestramento LLM
  - Mantiene costantemente aggiornati i sistemi LLM
  - Costi di integrazione contenuti

## Note

- 1 Utente fornisce il *prompt*
  - 2 Il sistema effettua una ricerca mirata su un *data source* di informazione documentale
  - 3 Il risultato della ricerca viene aggiunta al contesto del *prompt*

**⚠️ Ricerca di tipo *semantica***

**⚠️ Solo il contesto rilevante alla richiesta aggiunto al *prompt* (a differenza di *chat memory*...)**

## Note

- ➔ Basi di dati creati per CRUD di informazione **numerica multi-dimensionale**
  - ➔ Ogni elemento nel DB è la rappresentazione numerica del suo significato (*embedding*)
    - ➔ Testi
    - ➔ Immagini
    - ➔ Video
    - ➔ Audio
    - ➔ ...
  - ➔ Evoluzioni dei DB ottimizzati per ricerca su *keyword* (es. Elastic-Search)
  - ⚠ Non più limitati alla ricerca per parola esatta/simile!

## Note

## → Data source native

- ➔ Qdrant
  - ➔ ChromaDB
  - ➔ Neo4j
  - ➔ ...

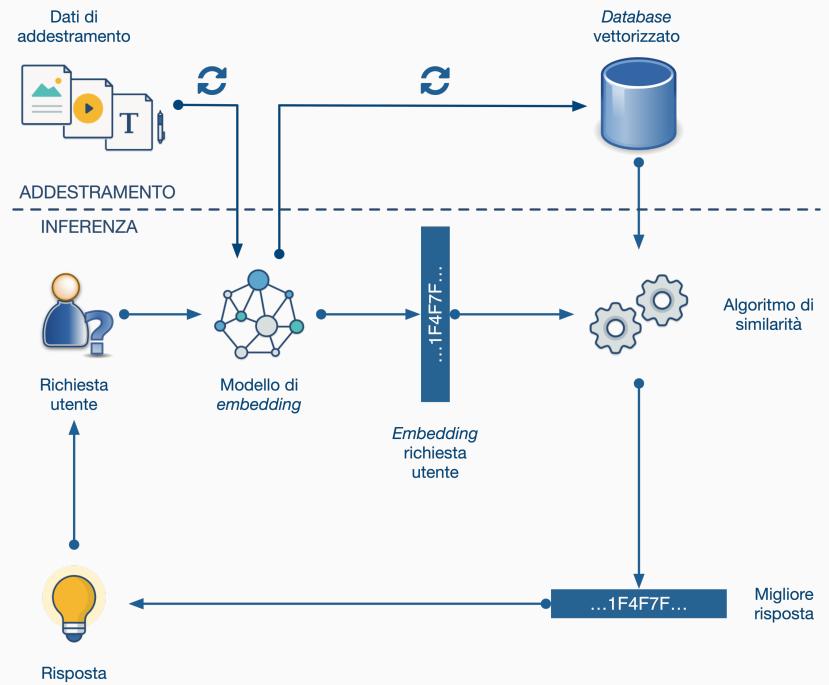
Data source SQL/noSQL con interazione vector data storage

- ➔ MongoDB
  - ➔ PostgreSQL + pgvector
  - ➔ Redis + RediSearch
  - ➔ ElasticSearch
  - ➔ ...

## Note

# WORKFLOW DEL PROCESSO RAG

## APPROCCIO GENERATIVO



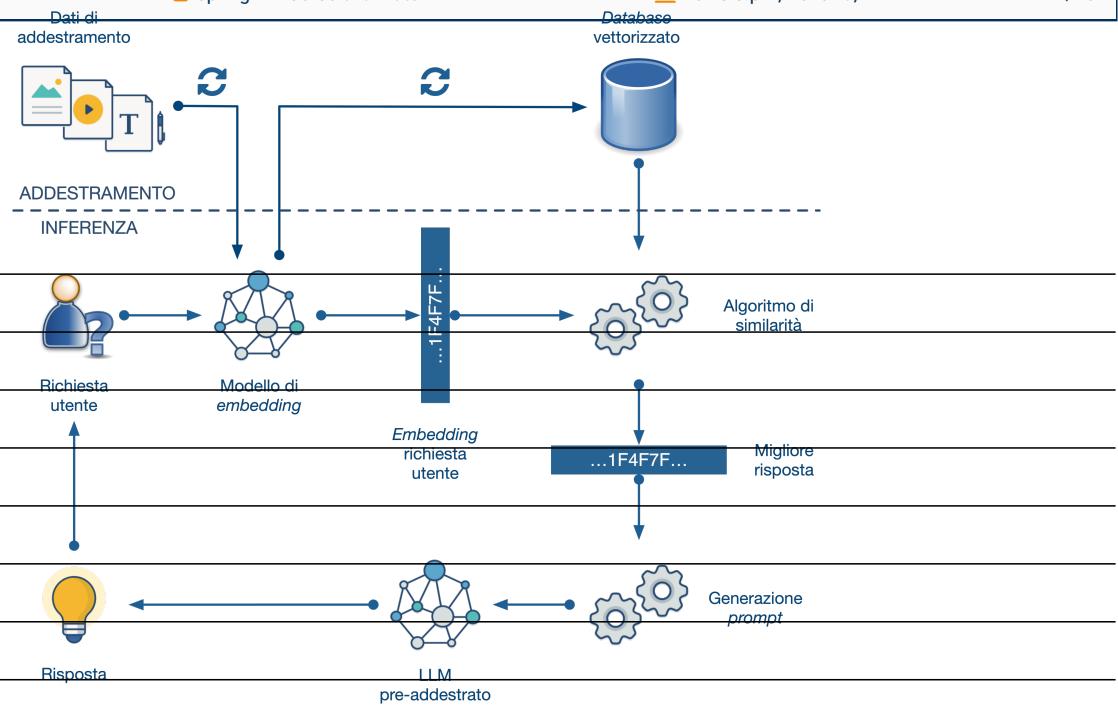
Simone Scannapieco

Spring AI - Corso avanzato

Venis S.p.A, Venezia, IT

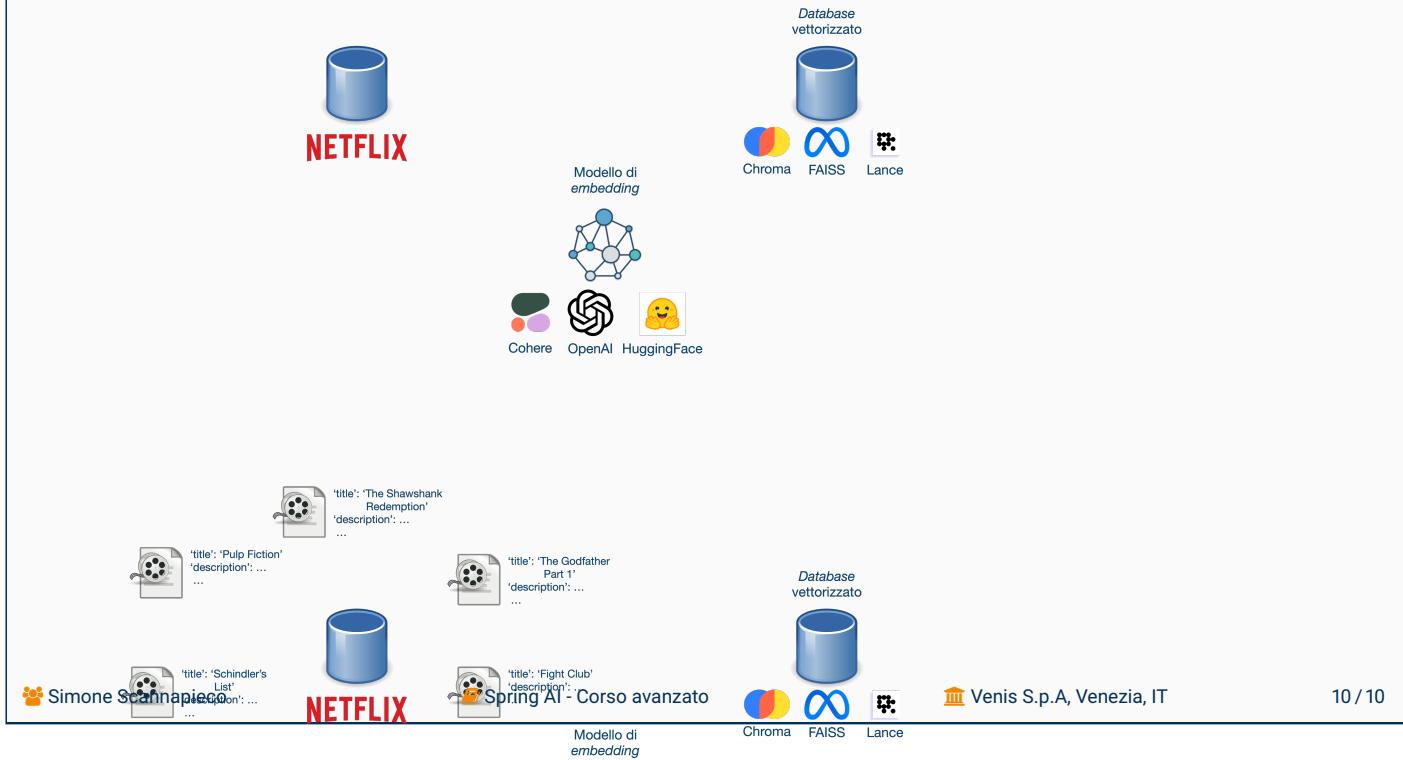
9 / 10

### Note



# RAG NON GENERATIVO: ESEMPIO

## RICERCA SEMANTICA DI FILM E SERIE



### Note

