
RAG Time!
mônade
Fabrizio Mele @ Working Software 2024 - Brescia

mònade

whoami

Fabrizio Mele - fabrizio@monade.io

sviluppatore @ mònade

podcaster a tempo perso

mônade

Agenda

1. LLM 101
2. L’idea del RAG
3. Un RAG in azione
4. Tocca a voi!

1. Di cosa stiamo parlando?

Basi Basi Basi: LLM

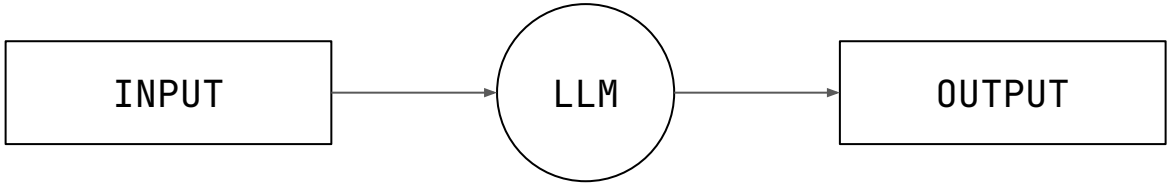
LLM sta per Large Language Model.

Large → addestrato su quantità di dati nell'ordine di grandezza dei TB, parametri centinaia di miliardi

Language Model → che modellizza il linguaggio naturale

mônade

Basi Basi Basi: LLM



La risposta alla domanda fondamentale sulla vita l’universo e tutto quanto è

La risposta alla domanda fondamentale sulla vita l’universo e tutto quanto è **42.**

PREDITTORE STATISTICO

mônade

Prompt Engineering? Si mangia?

Prompt Engineering → dare un’istruzione strutturata ad un LLM in modo da fargli eseguire un compito preciso in maniera predeterminata.

Il prompt engineering è permesso dall’in-context learning, ovvero l’abilità dei modelli di “imparare” dall’input fornito.

mônade

Zero-Shot Prompting

È la tecnica della “domanda secca”. Il risultato dipende dalla struttura del prompt e dalla conoscenza interna del modello.

Input:
Classificare il testo in negativo, neutro o positivo.
Testo: Il gatto ama il caldo
Classificazione:

Output:
Positivo

mônade

Few-Shot Prompting

Diamo degli esempi e abilitiamo l’apprendimento contestuale:

Input:

Caso di esempio

Il "gvnn" è un piccolo animale peloso originario della Tanzania. Un esempio di frase che utilizza la parola gvnn è:
Eravamo in viaggio in Africa e abbiamo visto questi simpaticissimi gvnn.

Caso reale, completato dal LLM

Fare un "agvzz" significa saltare su e giù molto velocemente.
Un esempio di frase che usa la parola agvzz è:

Output:

Quando sono felice faccio agvzz in mezzo alla strada.

mônade

Chain of Thought (CoT)

Consiste nell’indurre ad un ragionamento passo per passo. Si usa quando sono necessari ragionamenti complessi che richiedono dei passaggi.

Input:
La somma dei numeri dispari di questo gruppo corrisponde ad un numero pari: 4, 8, 9, 15, 12, 2, 1.
R: Sommando tutti i numeri dispari (9, 15, 1) si ottiene 25. La risposta è Falsa.

La somma dei numeri dispari di questo gruppo è un numero pari: 17, 10, 19, 4, 8, 12, 24.
R: La somma di tutti i numeri dispari (17, 19) risulta 36. La risposta è Corretta.

La somma somma dei numeri dispari di questo gruppo è un numero pari: 4,10,23,17,21.

Output:
La somma di tutti i numeri dispari (23,17,21) risulta 61. La risposta è Falsa.

mônade

Auto-CoT

Funziona anche dare una “spintarella” ad un LLM se non si hanno tanti esempi da fare.

Input:
D: Un giocoliere può giocolare con 16 palle. Metà delle palle sono palline da golf, e metà delle palline da golf sono blu. Quante palline blu ci sono?
R: Ragioniamo passo per passo:

Output:

La metà delle 16 palle è composta da 8 palline da golf e da 8 palline blu.

La metà di 8 è 4.

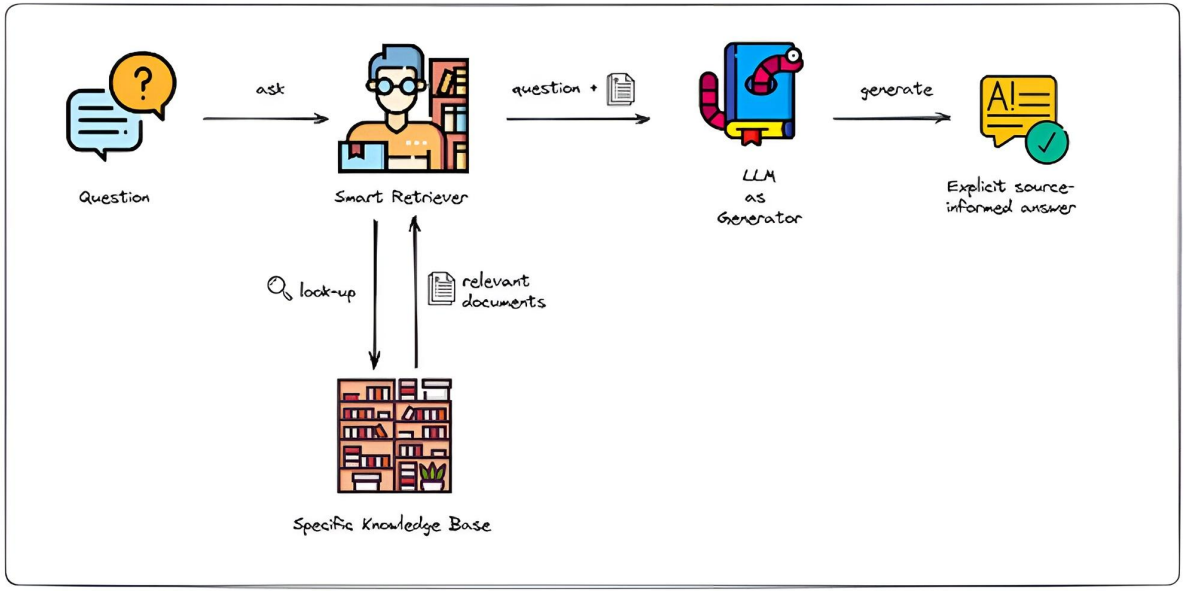
Ci sono quattro palline blu.

2. L’idea del RAG

mônade

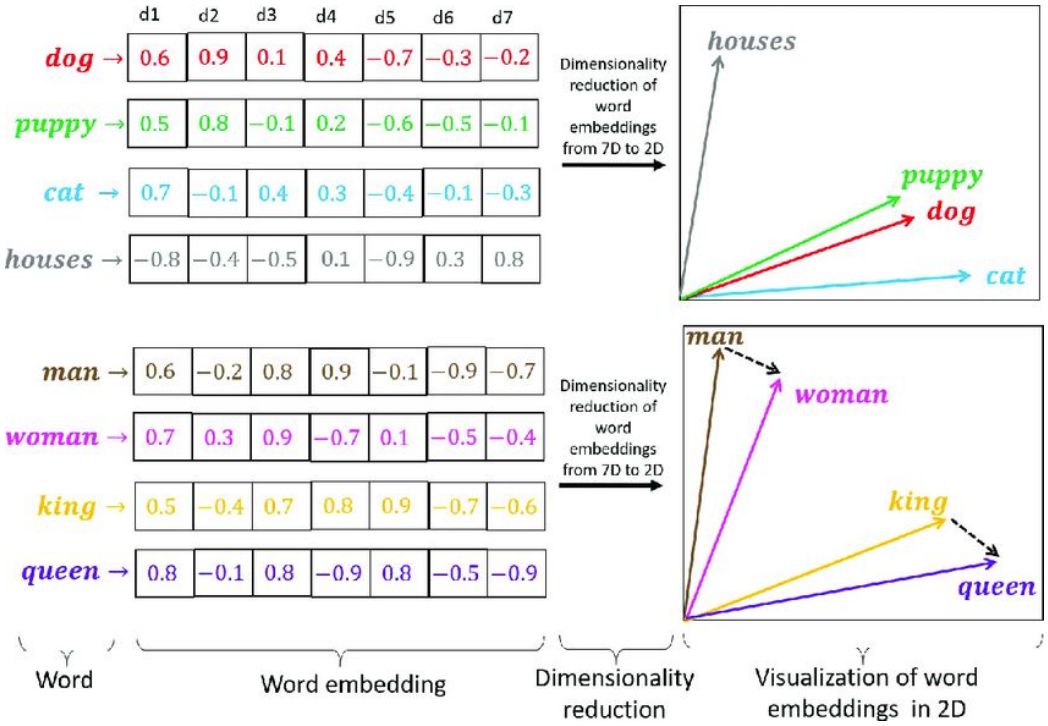
Retrieval Augmented Generation

Sfruttando l’in-context learning possiamo fare in modo di far ragionare un LLM sul contesto che gli forniamo.



Embedding

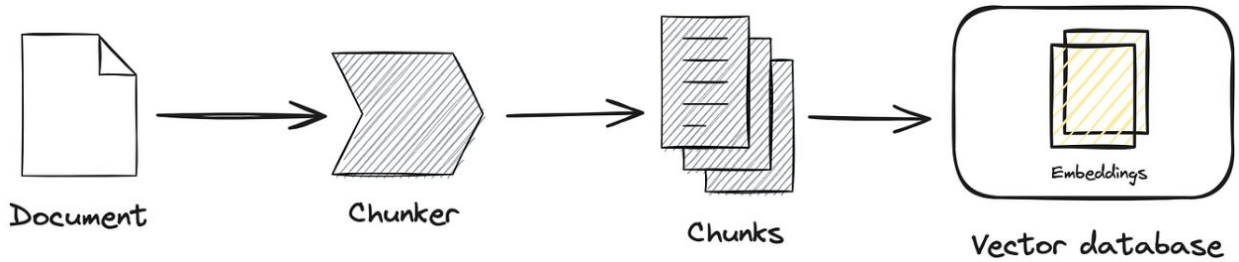
Gli embedding sono dei vettori che rappresentano la posizione di un oggetto in uno spazio n-dimensionale (tipicamente 500-1000).



mônade

Pre-elaborazione: ingestion

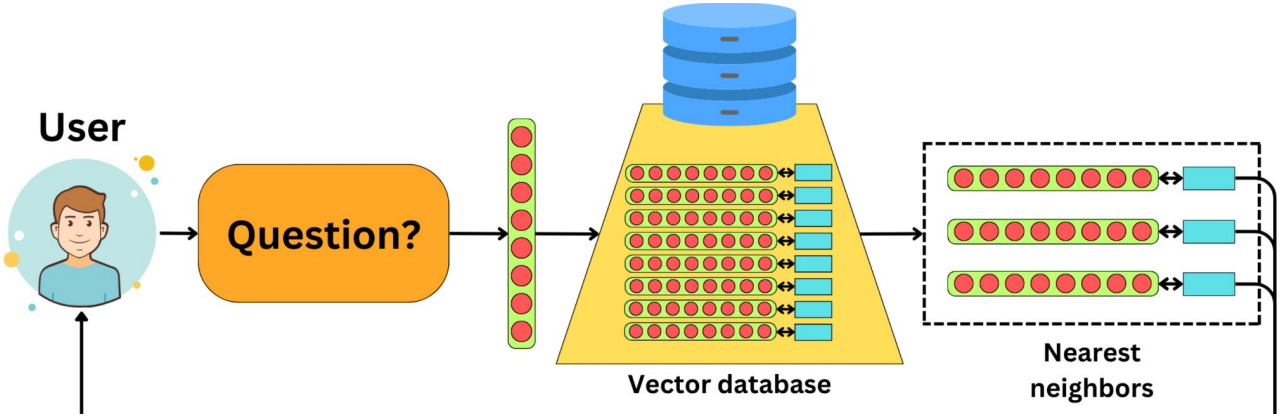
Popoliamo il database con informazioni indicizzate secondo il loro **embedding vector**.



mônade

Step 1: Retrieve

Utilizzando l’embedding della query facciamo una ricerca sul vector store e ci prendiamo i vettori più simili.



mônade

Step 2: Generate

Con query dell'utente e contesto aggiuntivo estratto dal vector store possiamo generare una risposta:

```
Il tuo compito è rispondere alle domande dell'utente
utilizzando un contesto specifico su cui basare la tua
risposta.

Ecco il contesto:
{{ context }}

Ecco la domanda:
{{ query }}

La tua risposta:
```

3. Un RAG in azione

4. Tocca a voi

https://github.com/melefabrizio/rag-time-workshop

Scompattate il file .env.shared.zip.

Idee:

- Migliorare il prompt
- Migliorare la qualità del retrieve
- Estrarre le fonti
- Implementare contromisure per il jailbreaking

mônade

Feedback