dgroove
Tecnologia emozionale.

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

Retrieval Augmented Generation

Note

Retrieval Augmented Generation
Criticità simple/stuff prompting

- *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 qualunque sotto-dominio specifico
  - Quando informazione contestuale non è sufficiente, LLM tende a generare risposte non veritiere (*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

- 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.

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

Note

Retrieval Augmented Generation
Come funziona

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

Retrieval Augmented Generation
Vector store

- 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* nativi
  - Qdrant
  - ChromaDB
  - Neo4j
  - …
- *Data source* SQL/noSQL con interazione *vector data storage*
  - MongoDB
  - PostgreSQL + pgvector
  - Redis + RediSearch
  - ElasticSearch
  - …

Note

Workflow del processo RAG
Approccio non generativo

Note

WORKFLOW DEL PROCESSO RAG
APPROCCIO GENERATIVO

Note

RAG non generativo: esempio
Ricerca semantica di film e serie

Note

RAG non generativo: esempio
Ricerca semantica di film e serie

Note

RAG non generativo: esempio
Ricerca semantica di film e serie

Note

RAG non generativo: esempio
Ricerca semantica di film e serie

Note

RAG non generativo: esempio
Ricerca semantica di film e serie

Note

RAG non generativo: esempio
Ricerca semantica di film e serie

Note

RAG non generativo: esempio
Ricerca semantica di film e serie

Note