EUROCC ITALY
HPC Competence Center

VECTORIA:

una soluzione AI privata e sostenibile in azienda

Autori: Luca Babetto, Leonardo Baroncelli, Carolina Berucci, Eric Pascolo, Andrea Proia

Sommario

IA aziendale: quando far firmare un patto di non concorrenza al modello

Nel contesto aziendale contemporaneo, l'intelligenza artificiale è divenuta un elemento centrale, non solo come strumento tecnologico ma come autentico collaboratore. Un algoritmo AI può affiancare i colleghi umani, acquisendo competenze tramite l'interazione e le loro esperienze. Integrato in diverse funzioni aziendali, esso supporta le attività quotidiane fino a ricoprire ruoli complessi, analizzando dati e apprendendo autonomamente per contribuire alla gestione delle informazioni e all'ottimizzazione dei processi.

L'affiancamento con il personale umano consente all'algoritmo AI di comprendere le dinamiche aziendali e di acquisire conoscenze specifiche del settore, migliorando costantemente la propria efficienza. La continua interazione con il team umano permette all'AI di comprendere meglio i contesti lavorativi e di adattarsi rapidamente alle mutevoli esigenze del mercato. L'AI non si limita a svolgere compiti operativi, ma supporta anche le decisioni strategiche, fornendo analisi dettagliate e previsioni basate su dati concreti. La sua presenza assicura un flusso continuo di informazioni aggiornate e rilevanti, fondamentali per mantenere la competitività dell'azienda.

Perdere le informazioni acquisite dall'AI equivale a perdere risorse umane preziose. Ogni dato raccolto e ogni esperienza assimilata rappresentano un valore aggiunto che deve essere preservato per evitare di favorire la concorrenza. Analogamente a un collega umano, l'AI possiede un'expertise che deve essere protetta e valorizzata.

Inoltre, il modello AI non si limita a replicare l'expertise di un singolo individuo, ma raccoglie e integra le conoscenze trasmesse da più persone attraverso le interazioni quotidiane. Questo lo rende un contenitore di competenze diversificate e interconnesse, che attraversano vari settori e funzioni aziendali. La perdita delle informazioni accumulate dall'AI rappresenta quindi una sfida considerevole, poiché comporta la dissipazione di un patrimonio di conoscenze collettive, frutto della collaborazione e dell'esperienza combinata di diversi membri del team. Proteggere queste informazioni non è solo una questione di sicurezza, ma di conservazione di un vantaggio competitivo derivante dall'integrazione di molteplici prospettive e competenze all'interno di un'unica entità digitale.

La dispersione delle informazioni equivale a cedere un vantaggio competitivo agli altri attori del mercato. È quindi cruciale implementare sistemi di protezione dei dati che garantiscano la riservatezza e la sicurezza delle informazioni gestite dall'AI, salvaguardando così il capitale aziendale.

Quando si utilizzano strumenti AI basati su API online, la privacy dei dati può essere compromessa, poiché le informazioni elaborate potrebbero fuoriuscire verso server esterni, mettendo a rischio la riservatezza aziendale. Invece, soluzioni come Vectoria, che implementano sistemi AI completamente on premise, offrono una maggiore sicurezza. Questo approccio assicura che tutte le operazioni di reperimento delle informazioni e generazione delle risposte avvengano all'interno dell'infrastruttura privata aziendale, eliminando il rischio di esposizione di dati sensibili a terzi. La protezione delle informazioni diventa quindi un vantaggio competitivo, garantendo che il patrimonio di conoscenza accumulato dall'AI rimanga confinato e protetto, preservando l'integrità e la sicurezza dei dati aziendali.

LLM e RAG: due modi differenti di apprendere

Attualmente, per svolgere task complessi si utilizzano modelli come i Large Language Models (LLM). In questo paragrafo, faremo il confronto tra l'uso base di questi modelli e una tecnica che ne aumenta le potenzialità, chiamata Retrieval-Augmented Generation (RAG). I Large Language Models (LLM) sono modelli di intelligenza artificiale che elaborano informazioni e generano contenuti. Pre-addestrati su grandi quantità di dati testuali, comprendono e producono linguaggio naturale con alta precisione. I sistemi di Retrieval-Augmented Generation (RAG) superano i limiti degli LLM recuperando informazioni specifiche, offrendo risposte precise e contestualizzate. Questo le rende ideali per applicazioni aziendali che richiedono accuratezza nella gestione dei dati.

Large Language Models (LLMs): l'intelligenza generativa tradizionale

I Large Language Models (LLMs) sono sistemi di intelligenza artificiale pre-addestrati su enormi quantità di dati eterogenei. Questa fase di addestramento consente loro di acquisire un'incredibile capacità di modellare il linguaggio naturale e generare risposte che risultano coerenti e linguisticamente sofisticate. Tuttavia, la loro competenza si basa su correlazioni statistiche tra i dati visti durante l'addestramento, il che comporta alcune limitazioni significative:

- **Dipendenza dai dati addestrati**: Gli LLM non possono accedere a informazioni nuove o non presenti nel set di dati originale. Ciò implica che un modello pre-addestrato possiede una conoscenza temporalmente limitata e non può conoscere informazioni riservate di aziende o privati.

- **Propensione agli errori "allucinatori"**: Essendo gli LLMs modelli probabilistici, essi possono generare risposte inesatte o non basate su fatti reali, generando un fenomeno di "allucinazione".
- **Scalabilità della conoscenza**: per aumentare la precisione e lo sviluppo di capacità espressive più sofisticate, sono richiesti dataset addestrativi sempre più ampi ed articolati, con conseguenti costi computazionali elevati.

Retrieval-Augmented Generation (RAG): l'ibrido tra generazione e ricerca

Le pipeline RAG mitigano le limitazioni degli LLM adottando un approccio modulare, in cui la generazione delle risposte è affiancata da un sistema di recupero delle informazioni. Questa metodologia si articola in due fasi principali:

1. **Retrieval**: I contenuti rilevanti vengono ricercati in un archivio strutturato, solitamente rappresentato da un database vettoriale. Esso archivia i documenti sotto forma di vettori multidimensionali, a seguito di una suddivisione in unità semantiche chiamate "chunk".
2. **Generation**: I dati recuperati vengono integrati nel contesto del modello generativo, il quale li utilizza per produrre risposte accurate e aggiornate.

Grazie a questa architettura, le pipeline RAG offrono diversi vantaggi:

- **Accesso a informazioni aggiornate**: Consentono l'integrazione di dati nuovi o specifici di dominio senza dover riaddestrare il modello generativo, consentendo quindi di ridurre ampiamente i costi relativi al training.
- **Personalizzazione**: Possono essere adattate a domini verticali e applicazioni aziendali. Idealmente, si potrebbe fornire alla pipeline di RAG l'intera documentazione aziendale da poter consultare come un'enciclopedia per poter fornire risposte specifiche e ben contestualizzate.
- **Privacy dei dati**: La gestione delle informazioni può essere confinata all'interno di infrastrutture sicure, garantendo la protezione di dati e informazioni sensibili.

In sintesi, mentre gli LLM rappresentano un approccio monolitico e generalista, le pipeline RAG abilitano un sistema flessibile, capace di combinare la potenza generativa degli LLM con l'accuratezza della ricerca mirata, adattandosi meglio a scenari dinamici e specifici.

Soluzioni AI : necessiatiamo di enormi risorse di calcolo per svilupparle?

La risposta, sorprendentemente, varia enormemente: si può andare da un comune computer portatile fino a richiedere la potenza di calcolo di un supercomputer.

Il cuore della questione risiede nei **componenti** che formano il sistema RAG, in particolare nei **modelli di intelligenza artificiale** che esso utilizza per le sue due fasi principali: il recupero dell'informazione (Retrieval) e la generazione della risposta (Generation).

1. **Il "Ricercatore" (Modello di Retrieval):** La prima fase consiste nel trovare le informazioni pertinenti all'interno di una vasta base di conoscenza (documenti, database, siti web...). Qui entra in gioco un modello specializzato nel "capire" la domanda dell'utente e nel cercare i pezzi di testo più rilevanti.
    - **Scenario "Leggero":** Se si usa un modello di ricerca relativamente semplice, magari basato su parole chiave o su "embeddings" (rappresentazioni numeriche del significato) di dimensioni contenute, le risorse richieste sono moderate. Potrebbe bastare una buona CPU e una quantità ragionevole di memoria RAM.
    - **Scenario "Pesante":** Se invece si impiega un modello di retrieval molto sofisticato, capace di comprendere sfumature complesse del linguaggio e di analizzare semanticamente enormi quantità di dati con embeddings ad alta dimensionalità, le richieste computazionali aumentano. Serviranno processori più potenti (spesso GPU) e molta più memoria per gestire gli indici e i calcoli necessari.
2. **Lo "Scrittore" (Modello Generativo - LLM):** Una volta trovate le informazioni rilevanti, queste vengono passate a un modello linguistico di grandi dimensioni (Large Language Model - LLM), il quale ha il compito di utilizzarle per formulare una risposta coerente e completa. È qui che le differenze possono diventare enormi.
    - **Scenario "Leggero":** Si possono utilizzare LLM più piccoli, con un numero relativamente basso di "parametri" (le variabili interne che il modello apprende durante l'addestramento, nell'ordine di pochi miliardi). Questi modelli, pur essendo capaci, richiedono meno memoria (RAM e VRAM delle GPU) e potenza di calcolo. Un sistema RAG basato su un LLM compatto potrebbe girare efficacemente su un laptop di fascia alta o una workstation potente.
    - **Scenario "Pesante":** Se l'applicazione richiede la massima qualità, creatività e capacità di ragionamento complesso, si opterà per LLM allo stato dell'arte, modelli colossali con centinaia di miliardi o addirittura trilioni di parametri. Questi giganti richiedono enormi quantità di memoria (decine o centinaia di Gigabyte di VRAM) e una potenza di calcolo

parallela massiccia, fornita da cluster di GPU avanzate. Caricare ed eseguire inferenze su questi modelli è un compito che spesso richiede infrastrutture dedicate, server potenti o, nei casi più estremi, l'accesso a supercomputer.

Combinando modelli di retrieval e generazione di diversa complessità e dimensione, è possibile costruire sistemi RAG adatti a specifiche esigenze e budget, esistono anche casi intermedi rispetto a quelli elencati sopra. Si può partire con una configurazione "leggera" su hardware accessibile per compiti specifici, per poi scalare verso soluzioni molto più potenti – e costose in termini di risorse – quando sono richieste prestazioni e capacità di altissimo livello. La scelta dipende dall'equilibrio desiderato tra costo, velocità e qualità della risposta finale.

Vectoria: un workflow RAG per un LLM rispettoso della privacy

Vectoria è una soluzione innovativa sviluppata nell'ambito del progetto europeo EuroCC2, ideata per rispondere alla crescente esigenza di aumentare consapevolezza, competenza e sviluppo tecnologico tramite sistemi di intelligenza artificiale avanzati. Progettata specificamente per l'uso aziendale, Vectoria consente l'accesso rapido e sicuro a documentazione interna attraverso interfacce conversazionali di tipo chatbot. Gli obiettivi strategici di Vectoria risultano essere:

- **Centralizzazione della conoscenza aziendale**: La dispersione della documentazione su più repository viene risolta consolidando i dati in una struttura unica e accessibile facilmente dal modello generativo. Questa operazione viene svolta automaticamente dalla libreria software che implementa la pipeline RAG ed è facilmente configurabile per soddisfare differenti necessità.
- **Precisione nella ricerca**: Le tecniche di retrieval su database vettoriale tramite metriche di similarità consentono di individuare le informazioni più rilevanti rispetto alle query degli utenti. Queste informazioni permettono al LLM di rispondere in maniera puntuale, evitando "allucinazioni" o risposte non rilevanti.
- **Rispetto della privacy**: Tutte le operazioni avvengono all'interno dell'infrastruttura privata aziendale, evitando quindi la fuoriuscita di informazioni

private verso servizi di IA generativa esterni. Si ha quindi totale garanzia di privacy e si evita inoltre di dover sottoscrivere e gestire un abbonamento verso i suddetti servizi esterni.

Il Workflow della pipeline RAG di Vectoria

Vectoria utilizza un'architettura modulare e flessibile, progettata per gestire tutte le fasi necessarie alla generazione di risposte contestualmente rilevanti. Ecco una descrizione dettagliata dei componenti chiave:

1. **Configurazione iniziale**: Il sistema è altamente personalizzabile attraverso file di configurazione che permettono di adattare Vectoria ad esigenze specifiche. La flessibilità di utilizzo diventa quindi un punto cardine, consentendo di specializzare il sistema a seconda delle necessità: possono essere processate diverse tipologie di documenti, modificati i modelli di encoding e generazione sottostanti, i parametri di retrieval, i prompt, ecc.
2. **Preprocessing documentale**: Il sistema consente il caricamento e la pulizia dei documenti, i quali vengono processati per eliminare elementi inutili che potrebbero "sporcare" il contenuto informativo del testo. Successivamente, avviene la suddivisione in "chunk": ogni documento viene frammentato in porzioni di testo gestibili, al fine di ottenere delle informazioni ben suddivise ed auto contenute, ottimizzando la granularità delle informazioni per la fase di ricerca.
3. **Generazione degli embedding**: Utilizzando modelli di encoding open-source pre-addestrati, i chunk vengono trasformati in vettori numerici che catturano il loro significato semantico. Questa operazione è necessaria al fine di poter sfruttare le metriche di similarità tra vettori, consentendo di confrontare le domande poste dagli utenti con le informazioni contenute nella documentazione a disposizione. Così facendo, si recuperano solo le più rilevanti per rispondere alla domanda presentata dall'utente.
4. **Archiviazione nel database vettoriale**: Gli embedding sono memorizzati utilizzando un database vettoriale, insieme a metadati che migliorano le capacità di filtraggio e ricerca. Un database vettoriale memorizza, gestisce e indicizza dati vettoriali ad alta dimensionalità. I dati sono memorizzati come array di numeri che vengono raggruppati in base alla similarità, consentendo query a bassa latenza. A differenza dei database relazionali tradizionali (con righe e colonne), i dati in un database vettoriale sono rappresentati da vettori con un numero fisso di dimensioni. Ogni dimensione del vettore denso corrisponde a una caratteristica latente o a un aspetto dei dati, ossia un attributo sottostante che non viene osservato direttamente, ma che viene dedotto dai dati attraverso modelli matematici o algoritmi.

5. **Inferenza e risposta**: La domanda (o “query”) dell’utente viene convertita anch’essa in un embedding e confrontata con gli embedding presenti nel database per identificare i chunk più simili. I chunk selezionati vengono utilizzati per costruire un prompt che, combinato con il modello generativo genera una risposta accurata e contestualizzata.

L’integrazione con i supercomputer HPC: Vectoria è “HPC-Ready”

Un elemento distintivo di Vectoria è la possibilità di utilizzare infrastrutture di High-Performance Computing (HPC) per potenziare l’intera pipeline. Questa scelta permette di aumentare la scalabilità del sistema, permettendo di gestire grandi volumi di dati ed utilizzare modelli generativi di dimensioni proibitive per una normale workstation. Non è richiesto alcun tipo di modifica al sistema a livello di sviluppo codice: essendo la libreria “HPC-ready”, è sufficiente utilizzare un file di configurazione specifico per l’utilizzo su HPC che si interfacci con lo scheduler appropriato. Inoltre, utilizzare Vectoria su un’infrastruttura ad alta capacità di calcolo, consente di aumentare la velocità del sistema, ottenendo una riduzione dei tempi di elaborazione sia durante la fase di retrieval che nella generazione delle risposte.

Leonardo Spa – uso di Vectoria in ambito aziendale

Leonardo S.p.A. è una multinazionale che opera a livello globale nei settori della Difesa, dell'Aerospazio e della Sicurezza. L'azienda è da sempre impegnata nell'innovazione tecnologica, mirando a ottimizzare i propri processi aziendali e a migliorare l'efficienza operativa in tutti gli ambiti in cui è coinvolta.

Il Business Management System (BMS), è una piattaforma strategica che funge da punto centrale per l'archiviazione e la gestione della documentazione aziendale. Il BMS di Leonardo include anche un corpus documentale che descrive procedure aziendali, relative a diversi aspetti operativi, dai processi interni alla gestione della sicurezza e della qualità. Queste procedure sono fondamentali per garantire la coerenza e l’efficienza nelle attività quotidiane, ma la loro consultazione e fruizione richiedono spesso un impegno significativo, soprattutto per i nuovi assunti, che potrebbero non essere ancora del tutto familiari con le pratiche aziendali. In un contesto di digitalizzazione e trasformazione tecnologica, Leonardo ha cercato di rendere l'accesso e la comprensione di queste informazioni aziendali più agevole e immediato. Questo obiettivo è diventato ancora più rilevante con l'ingresso di nuovi talenti, che potrebbero trovarsi a dover affrontare un'enorme mole di documentazione, talvolta complessa e

difficile da navigare, ma anche per gli utenti esperti che necessitano di una consultazione più rapida ed efficiente.

Per affrontare questa sfida, l'azienda ha deciso di sperimentare una soluzione innovativa basata sull'uso di un chatbot RAG (Retriever-Augmented Generation). Il sistema RAG è stato progettato per ottimizzare l'accesso alle informazioni memorizzate nel BMS, utilizzando tecniche avanzate di intelligenza artificiale per recuperare e generare risposte altamente pertinenti e personalizzate in base alle domande specifiche degli utenti.

Per la prototipazione di questa soluzione, è stata utilizzata la libreria open-source Vectoria, che offre potenti funzionalità di recupero informazioni e generazione basata su modelli di linguaggio avanzati. L'intero sistema è stato ospitato sul supercomputer aziendale DaVinci-1, una risorsa fondamentale per garantire la velocità di elaborazione e l'affidabilità del sistema. La soluzione è stata testata con un sottoinsieme della documentazione aziendale, al fine di verificarne l'efficacia e la capacità di rispondere alle esigenze degli utenti. I primi risultati della sperimentazione sono stati molto promettenti. Il chatbot RAG è stato in grado di rispondere in modo rapido e accurato alle richieste degli utenti, rendendo più facile l'accesso alle informazioni e riducendo significativamente i tempi di ricerca.

Nel seguito si andrà a descrivere più nel dettaglio l'architettura operativa adottata per l'implementazione della soluzione RAG, con particolare attenzione all'utilizzo della libreria open-source Vectoria, che ha rappresentato il cuore tecnologico della sperimentazione.

Tutte le principali componenti del processo, dal parsing iniziale della documentazione, alla creazione dei chunk, fino alla fase di retrieval e reranking sono gestite nativamente da Vectoria, semplificando notevolmente la costruzione della pipeline e riducendo lo sforzo ingegneristico necessario all'integrazione di moduli esterni. La documentazione aziendale, su cui è stata condotta la sperimentazione, era fornita in formato .docx, formato che Vectoria è in grado di processare direttamente. Questo ha permesso un parsing robusto e strutturato, grazie alla capacità della libreria di interpretare la gerarchia interna del documento (titoli, paragrafi, sezioni), mantenendo così intatta la struttura logica del contenuto. Successivamente, Vectoria ha provveduto in maniera automatica alla segmentazione del testo in chunk di 512 caratteri, con un overlapping di 256 caratteri tra i segmenti. Questo approccio garantisce la continuità semantica e migliora l'accuratezza nelle fasi successive. Ogni chunk è stato inoltre arricchito con metadati significativi, anch'essi generati automaticamente dalla libreria, come il nome del documento e il numero del paragrafo di origine, migliorando la tracciabilità e

l'interazione con i risultati restituiti all'utente. Per la fase di embedding, è stato impiegato il modello BAAI/bge-m3, anch'esso integrabile nativamente all'interno di Vectoria. Questo modello è stato scelto per la sua efficacia nella rappresentazione semantica dei testi e per le sue ottime performance nei task di retrieval. La pipeline RAG così costruita prevede una prima fase di retrieval basata sulla cosine similarity tra gli embedding della query e quelli dei chunk indicizzati, seguita da una fase di reranking che consente di affinare la selezione dei contenuti più rilevanti. Anche il reranking è una feature direttamente supportata da Vectoria, che in questo caso ha utilizzato il modello BAAI/bge-reranker-v2-m3 per migliorare l'ordinamento dei documenti restituiti in funzione della query dell'utente. Per la generazione della risposta finale è stato adottato il modello LLaMA-3.2 70B, un language model di ultima generazione capace di produrre output coerenti, precisi e contestualmente rilevanti, sfruttando le informazioni recuperate dai documenti. Considerate le dimensioni dei modelli coinvolti, in particolare per la fase di generazione, l'intera pipeline è stata eseguita sul supercomputer aziendale DaVinci-1, sfruttando capacità di calcolo multi-GPU ad alte prestazioni. Questo ha garantito tempi di risposta competitivi e un'elevata affidabilità anche in fase di sperimentazione.

Questa esperienza ha dimostrato come l'adozione di tecnologie basate su RAG possa rappresentare una risorsa strategica per Leonardo Spa, non solo per semplificare la fruizione della documentazione aziendale, ma anche per supportare il processo di digitalizzazione e innovazione continua all'interno dell'azienda. L'applicazione di RAG nel contesto di Leonardo è un chiaro esempio di come le soluzioni di intelligenza artificiale possano essere sfruttate per migliorare l'efficienza operativa e ottimizzare la gestione delle conoscenze aziendali.

Vectoria: posso avere un chatbot privato nella mia PMI?

Grazie alla sua architettura flessibile, Vectoria può essere implementato anche su infrastrutture private aziendali quali workstation o server locali. Questo lo rende una soluzione ideale per piccole e medie imprese che desiderano un chatbot sicuro e completamente interno, eliminando la necessità di dipendere da servizi cloud/HPC esterni. Questa configurazione garantisce che tutte le operazioni, dall'elaborazione dei documenti alla generazione delle risposte, rimangano all'interno dell'ambiente aziendale, preservando la riservatezza dei dati e delle informazioni. Inoltre, il sistema è pensato per sfruttare risorse hardware standard e software open source, consentendo

alle aziende di implementare una soluzione avanzata di intelligenza artificiale senza dover investire in infrastrutture specializzate o licenze software.

L’installazione locale, è semplice e per eseguirla è necessario configurare un ambiente virtuale Python in cui vengono installati tutti i requisiti necessari. Successivamente, il sistema viene avviato tramite un playbook Ansible, uno strumento open-source che automatizza la configurazione dei sistemi, il deployment delle applicazioni e la gestione delle risorse IT. Ansible semplifica il processo di installazione, gestendo i dettagli tecnici, effettuando il download dei modelli AI che servono a Vectoria per funzionare e permettendo l'avvio del sistema senza dover configurare manualmente ogni componente. Per l'installazione remota, è richiesto l'accesso SSH al server ospitante. Anche in questo caso, la procedura prevede la creazione di un ambiente virtuale Python e l'uso di Ansible per installare e configurare il sistema. Questa configurazione garantisce un'installazione sicura e affidabile, mantenendo la riservatezza dei dati aziendali e sfruttando al meglio le potenzialità di Vectoria, specialmente se integrato con infrastrutture ad alta capacità di calcolo.