



tinexta
defence

**Retrieval Augmented
Generation (RAG) per
la consultazione locale
e sicura di documenti:
una pipeline per la standardizzazione
dei dati e lo scoring semantico**

#TinextaDefenceBusiness

AI4Cyber

AI4Cyber studio, è il nuovo spazio di ricerca applicata di Tinexta Defence, dedicato alla ricerca e applicazione dell'**Intelligenza Artificiale in ambito cybersecurity**.

Al suo interno, diversi specialisti con competenze ed esperienze eterogenee collaborano per affrontare le sfide poste dal panorama delle minacce informatiche, accompagnare i clienti nell'adozione dell'AI nei propri processi aziendali e condividere le attività di ricerca con la comunità tecnico-scientifica.

Il team cura l'**intero ciclo di sviluppo dei sistemi basati su AI**: dalla raccolta e dal preprocessing dei dati, all'addestramento e validazione dei modelli, fino al deployment in produzione.

Le soluzioni proposte si fondano su tecnologie avanzate di **Machine Learning, Deep Learning e Large Language Models**, e possono essere personalizzate in base alle specifiche esigenze del cliente.

L'AI Team è costantemente impegnato in attività di ricerca e sperimentazione, con un'attenzione particolare agli **aspetti etici**, alla **trasparenza** e alla **tutela della privacy**.

L'obiettivo è proporre soluzioni innovative che siano in linea con i principi di **equità, inclusività e rispetto dei diritti fondamentali**.

Summary

Abstract	04
Introduzione	05
Architettura del sistema	06
Scoring Module	08
Modello e Metriche utilizzate	09
Caso d'uso	10
Validazione FASE 1	11
Validazione FASE 2	12
Validazione FASE 3	14
Sviluppi futuri	16
Bibliografia	17

This document is protected by copyright laws and contains material proprietary to the Tinexta Defence. It or any components may not be reproduced, republished, distributed, transmitted, displayed, broadcast or otherwise exploited in any manner without the express prior written permission of Tinexta Defence. The receipt or possession of this document does not convey any rights to reproduce, disclose, or distribute its contents, or to manufacture, use, or sell anything that it may describe, in whole or in part.

Abstract

Negli ultimi anni, l'evoluzione delle tecniche di Natural Language Processing (NLP) e dei Large Language Models (LLM) ha aperto nuove prospettive nell'automazione di processi informativi complessi, in particolare per quanto riguarda l'analisi e l'elaborazione di dati testuali non strutturati.

In questo contesto, una delle principali sfide consiste nella trasformazione di documenti eterogenei, caratterizzati da elevata variabilità linguistica e semantica, in rappresentazioni strutturate e formalmente utilizzabili dai sistemi informatici.

Il presente studio propone un approccio basato sull'utilizzo degli LLM come strumenti di strutturazione semantica dei dati, attraverso una pipeline che consente l'acquisizione automatica di documenti testuali e la loro normalizzazione in formato strutturato.

Il sistema integra modelli linguistici e tecniche di embedding multilingua per il confronto semantico tra contenuti, supportato da un meccanismo di scoring pesato per la valutazione della rilevanza.

L'architettura è progettata per operare in ambienti locali e controllati, garantendo trasparenza del processo, controllo dell'elaborazione e riproducibilità dei risultati con particolare attenzione alla natura sensibile dei dati trattati. La validazione sperimentale è stata condotta tramite metriche personalizzate derivate da concetti di precision e recall, al fine di misurare la perdita di informazione e la generazione di contenuti non supportati durante la fase di strutturazione.

I risultati mostrano come l'impiego degli LLM per la trasformazione di dati non strutturati in rappresentazioni formali consenta di ottenere basi informative affidabili, utilizzabili come fondamento per sistemi intelligenti e processi automatizzati basati su conoscenza strutturata.

Autori:

- Simone Parrella: Analista AI4Cyber
- Simona Sorgente: Team Leader AI4Cyber

Introduzione

Con la crescente digitalizzazione dei processi informativi è inevitabilmente aumentata la mole di documenti testuali, eterogenei e con strutture variabili. Infatti, documentazioni tecniche, Curriculum Vitae, report ed e-mail rappresentano una componente centrale in vari contesti applicativi moderni.

Tuttavia, tali informazioni risultano difficilmente sfruttabili dai sistemi automatici, poiché la loro natura non strutturata rende complessa l'estrazione dei dati, la comparazione semantica e l'integrazione in pipeline decisionali. Le tradizionali tecniche basate su keyword matching o su semplici rappresentazioni vettoriali, pur offrendo buone prestazioni in scenari controllati, mostrano limiti evidenti in presenza di ambiguità semantica, sinonimia, informazioni implicite e variazioni linguistiche.

Negli ultimi anni, i Large Language Models hanno dimostrato una notevole capacità di comprensione del linguaggio naturale, aprendo nuove possibilità non solo per la generazione di testo, ma anche per la trasformazione semantica dei contenuti. In particolare, emerge l'opportunità di utilizzare tali modelli non come semplici motori di risposta, ma come strumenti di strutturazione della conoscenza, capaci di convertire documenti complessi in rappresentazioni formali utilizzabili dai sistemi informatici.

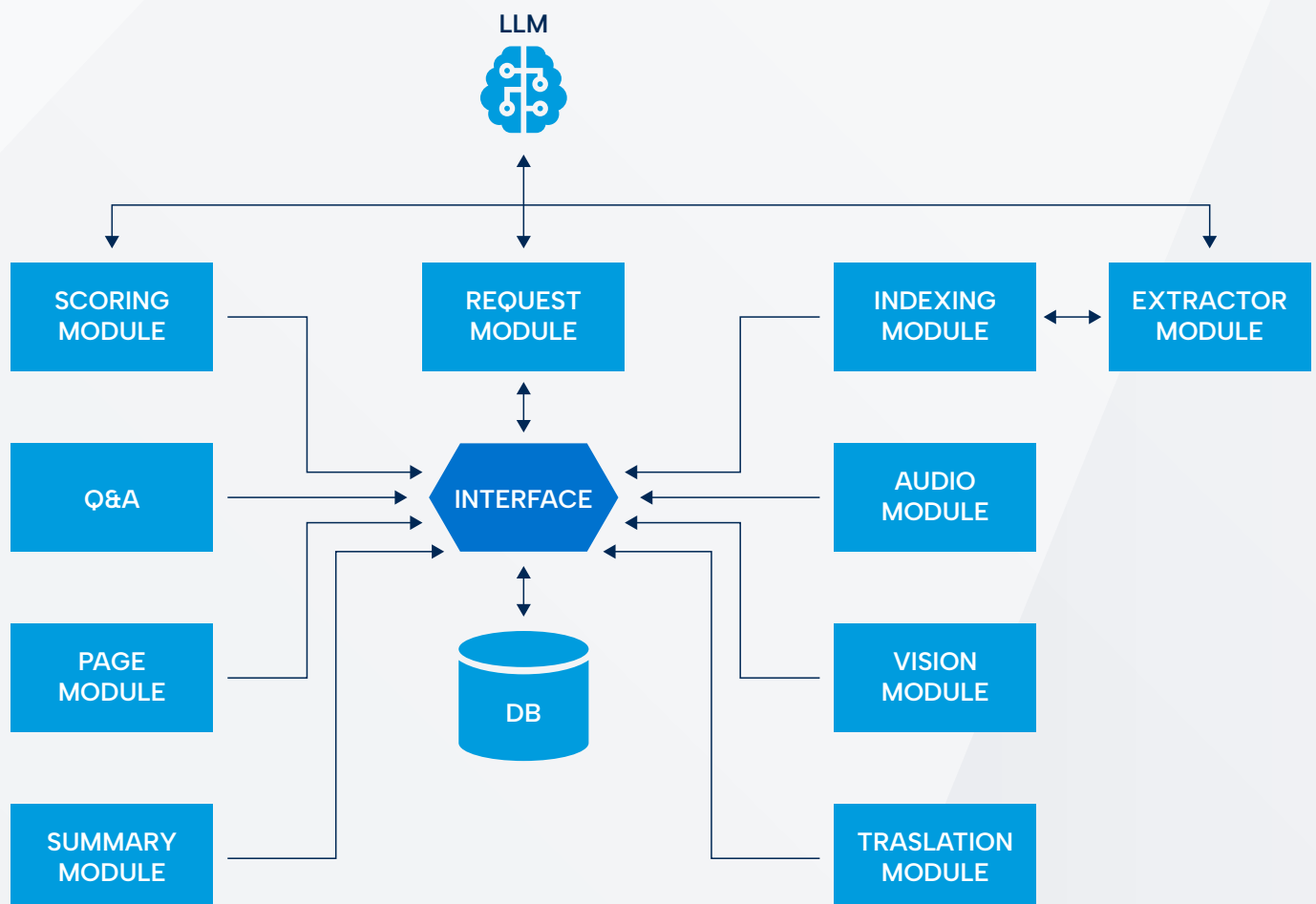
In questo contesto, il presente studio si propone di investigare un approccio basato sull'impiego degli LLM per la normalizzazione e strutturazione semantica di documenti testuali, integrando rappresentazioni distribuite e meccanismi di valutazione della similarità, al fine di supportare processi decisionali fondati su dati testuali affidabili e formalmente organizzati.

Architettura del sistema

Il sistema si inserisce all'interno di una pipeline di elaborazione documentale già esistente [1], sfruttandone i meccanismi di acquisizione dati già consolidati.

Tuttavia, la fase di indicizzazione dei documenti non fa ricorso a rappresentazioni vettoriali tradizionali, ma adotta un processo di elaborazione testuale avanzata che consente la normalizzazione dei contenuti in un formato standard facilmente accessibile da qualsiasi processo applicativo. La novità rispetto al sistema preesistente risiede quindi nella strategia di normalizzazione impiegata, che permette di trasformare i documenti in strutture dati uniformi e facilmente utilizzabili nelle fasi successive di elaborazione.

Affinché il sistema produca risultati soddisfacenti, è necessario disporre di LLM in grado di gestire contesti estesi e dotati di una solida capacità di interpretazione semantica, così da cogliere anche informazioni non esplicitamente dichiarate ed evitare la perdita di elementi rilevanti ai fini della selezione.



L'architettura mostrata nella figura sopra rappresenta la più recente evoluzione del sistema "Retrieval-Augmented-Generation (RAG) per la consultazione locale e sicura di documenti"[1]. Rispetto alla versione precedente, il sistema introduce due componenti principali: un database centralizzato e un modulo di scoring per la valutazione semantica dei contenuti.

Il sistema è organizzato secondo un'architettura modulare, in cui l'interfaccia funge da punto di accesso tra i vari moduli applicativi e il livello di persistenza dei dati, permettendo una gestione unificata dei flussi informativi e delle richieste dell'utente.

Di seguito si riporta una breve descrizione dei moduli già presenti nella versione originaria, le cui funzionalità sono state approfondite in [1][2]:

- Request Module: responsabile della gestione delle richieste dell'utente e dell'interazione con il modello linguistico selezionato.
- Indexing Module: deputato all'indicizzazione dei contenuti testuali e alla loro preparazione per le successive fasi di elaborazione.
- Extractor Module: utilizzato per l'estrazione dei contenuti informativi dai documenti sorgente.
- Q&A Module: consente l'interrogazione dei documenti indicizzati tramite domande in linguaggio naturale.
- Summary Module: dedicato alla generazione automatica di riassunti dei contenuti.
- Page Module: permette la ricerca mirata su specifiche porzioni o pagine dei documenti.
- Audio Module, Vision Module e Translation Module: moduli specializzati rispettivamente per l'elaborazione di contenuti audio, immagini e traduzioni multilingua.

Per quanto riguarda le novità introdotte, il database centralizzato svolge il ruolo di componente di persistenza, memorizzando sia i dati forniti dall'utente sia i metadati generati durante la fase di indicizzazione.

Il secondo elemento innovativo è il modulo di scoring analizzato nel dettaglio nei prossimi paragrafi.

Scoring Module

Lo Scoring Module ha l'obiettivo di valutare e confrontare documenti di qualsiasi tipo rispetto a una richiesta dell'utente, al fine di fornire uno score quantitativo che rappresenti la corrispondenza tra le informazioni contenute nei documenti e le richieste indicate. In altre parole, il modulo supporta la selezione e il ranking dei documenti più rilevanti rispetto a una determinata esigenza informativa.

Il funzionamento dello Scoring Module si articola nelle seguenti tre fasi: elaborazione dei documenti, elaborazione della richiesta e scoring. Ognuna di queste fasi, a sua volta, è articolata come segue:

FASE 1 – Elaborazione dei documenti:

- Caricamento dei documenti tramite l'interfaccia
- Selezione della modalità d'indicizzazione "custom" e avvio del processo
- Estrazione dei contenuti
- Elaborazione e validazione del contenuto
- Salvataggio sul database

FASE 2 – Elaborazione richiesta:

- Impostazione del peso associati a ciascun campo di ricerca, al fine di attribuire maggiore o minore importanza ai diversi campi
- Inoltro della richiesta dell'utente in linguaggio naturale
- Elaborazione della richiesta
- Validazione del contenuto

FASE 3 – Scoring:

- Confronto del contenuto della richiesta con ogni documento fornito in input
- Attribuzione di uno score mediante la media ponderata dei valori di similarità calcolati utilizzando i pesi assegnati nella fase precedente
- Stampa a schermo dei risultati

Modello e Metriche utilizzate

L'LLM impiegato per la realizzazione di questo nuovo modulo è GPT-OSS-20B [3], eseguito in locale tramite OLLAMA [4], scelto per il buon compromesso tra qualità delle generazioni e velocità di inferenza; l'esecuzione in locale consente inoltre di garantire la riservatezza dei documenti elaborati.

Considerata la natura fortemente personalizzata del sistema, non è possibile applicare direttamente le metriche standard utilizzate nei sistemi di NLP tradizionali. È quindi necessario adattare concetti provenienti dall'information extraction e dalla valutazione della qualità dei dati.

Le metriche adottate, derivanti dall'Error Rate, sono Information Loss e Hallucination Rate e sono rispettivamente il reciproco di precision e recall. In particolare, Information Loss corrisponde al False Negative Rate, che misura la percentuale di istanze positive reali non individuate dal modello, mentre Hallucination Rate corrisponde al False Discovery Rate, che rappresenta la percentuale di predizioni che risultano errate rispetto al totale delle predizioni effettuate.

- Error Rate: $Error\ Rate = \frac{\text{Numero di errori}}{\text{Numero totale di tentativi}}$
- Information Loss: $LOSS = \frac{\text{Numero di informazioni assenti}}{\text{Totale contenuti estratti}}$
- Hallucination Rate: $HALL = \frac{\text{Numero di allucinazioni trovate}}{\text{Totale contenuti estratti}}$

Caso d'uso

Un caso d'uso concreto del modulo proposto è rappresentato dal supporto ai processi di selezione del personale, in particolare nelle fasi di analisi e confronto tra Curriculum Vitae (CV) e requisiti richiesti per una specifica posizione lavorativa.

Tipicamente, le organizzazioni si trovano a gestire un elevato numero di candidature eterogenee, redatte in formati diversi e con stili espressivi fortemente variabili. I CV contengono informazioni sia esplicite che implicite relative a competenze, esperienze e qualifiche, mentre le richieste dei recruiter sono spesso formulate in linguaggio naturale, con requisiti non sempre rigidamente strutturati.

Il problema principale risiede nella difficoltà di effettuare un confronto accurato e scalabile tra questi due insiemi di informazioni, evitando sia l'esclusione di profili potenzialmente idonei sia l'inclusione di candidature poco pertinenti. Approcci puramente basati su parole chiave o su regole rigide risultano infatti insufficienti a cogliere la reale corrispondenza semantica tra profili e requisiti.

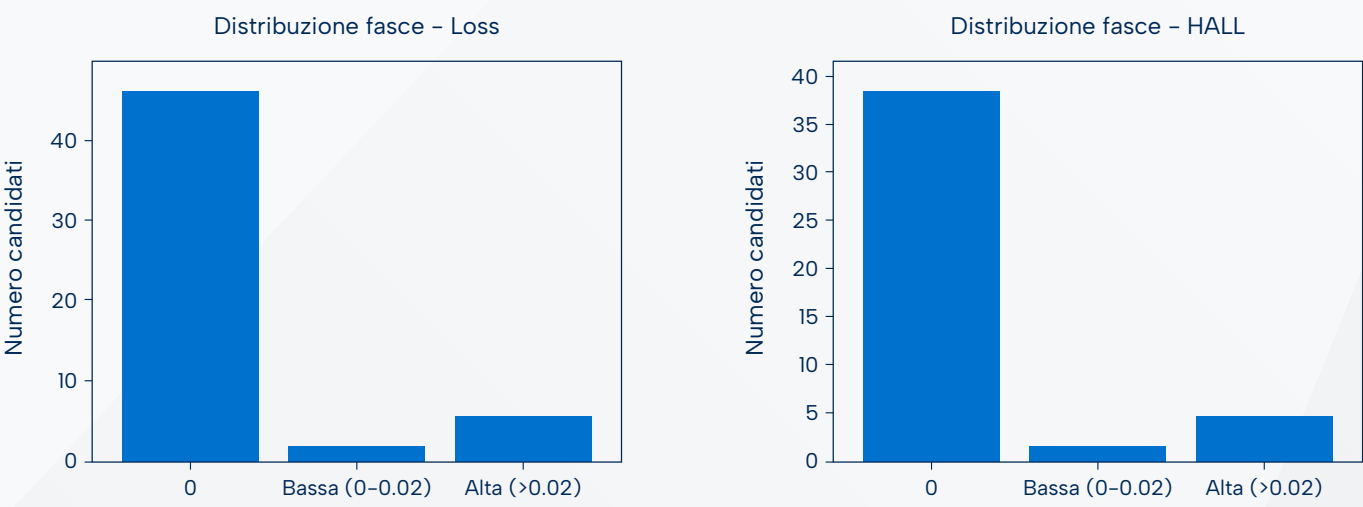
Il sistema descritto in questo studio affronta tale problematica attraverso una pipeline che consente l'acquisizione dei CV, la loro normalizzazione in una rappresentazione strutturata e il confronto semantico con le richieste dell'utente. L'integrazione di modelli linguistici, tecniche di similarità semantica e meccanismi di scoring pesato permette di valutare il grado di compatibilità tra candidato e richiesta del recruiter, tenendo conto sia delle competenze dichiarate sia di informazioni deducibili dal contesto.

In questo scenario, il sistema non si pone come strumento di sostituzione del decisore umano, ma come supporto alla selezione, con l'obiettivo di ridurre il carico manuale, aumentare la coerenza delle valutazioni e migliorare l'efficienza complessiva del processo.

Validazione FASE 1

In questa fase è stata valutata la capacità del sistema di standardizzare correttamente le informazioni fornite in input, a partire dai Curriculum Vitae, utilizzando le due tipologie di errore illustrate precedentemente.

Il test set comprende 50 CV contenenti informazioni eterogenee, tra cui competenze professionali, esperienze lavorative, formazione e altri elementi rilevanti per valutare la completezza e la qualità dei dati estratti.



%	Information Loss	Hallucination Rate
0	44	39
0>V<2	1	5
>2	5	6

La tabella e i grafici riportano il numero di casi osservati per ciascuna categoria di errori, suddivisi per fasce di severità. I risultati mostrano che nella quasi totalità dei casi il sistema non introduce errori significativi e mantiene tutte le informazioni correttamente.

Solo in una minima parte dei casi si osservano perdite o interpretazioni errate, con valori di Information Loss e Hallucination Rate compresi tra l'1,4% e il 5%, evidenziando come gli errori di maggiore severità siano rari e limitati a pochi episodi.

Validazione FASE 2

Anche in questo caso sono state utilizzate le metriche citate nel paragrafo precedente ma, a differenza della validazione della FASE 1, il modulo è stato testato sull’elaborazione di 20 richieste fatte dall’ utente con complessità crescente (facile, media, difficile, molto difficile).

La complessità delle richieste è stata assegnata sulla base di 3 criteri principali:

- Numero di vincoli presenti nella frase
- Tipologia di vincoli (tecnici, accademici e soft skill)
- Il grado di ambiguità semantica dei requisiti

Richieste con pochi vincoli, prevalentemente tecnici e facilmente individuabili sono state classificate come facili. All’aumentare dei requisiti, della loro eterogeneità e del livello di astrazione semantica, la richiesta è stata classificata come media, difficile o molto difficile.

Di seguito si riportano le richieste eseguite con il relativo grado di complessità:

Richiesta	Complessità
Cerco un candidato con almeno 2 anni di esperienza in sviluppo software e conoscenza di Python.	Facile
Mi serve un candidato che sappia usare SQL.	Facile
Cerco un profilo junior con laurea in informatica o ingegneria informatica.	Facile
Il candidato deve avere un B1 in inglese.	Facile
È richiesta la patente di guida.	Facile
Cerco uno sviluppatore backend con almeno 3 anni di esperienza in Java e Spring.	Media
Il candidato ideale ha competenze in Python, Pandas e analisi dei dati.	Media

È richiesta esperienza lavorativa in ambito sviluppo web e conoscenza di JavaScript.	Media
Cerco un profilo con laurea magistrale e almeno una certificazione tecnica.	Media
Il candidato deve aver lavorato in team e avere buone capacità di problem solving.	Media
Cerco un ingegnere informatico con almeno 5 anni di esperienza, competenze in cloud computing e certificazione AWS.	Difficile
Il candidato deve aver partecipato a progetti di machine learning e data science.	Difficile
Richiesta esperienza in ambito cybersecurity e conoscenza delle normative di compliance.	Difficile
Cerco una figura con esperienza come project manager in progetti software complessi.	Difficile
Il candidato deve conoscere italiano madrelingua e inglese almeno livello B2.	Difficile
Cerco un candidato con più di 30 anni, di nazionalità italiana, con patente di guida, almeno 5 anni di esperienza come sviluppatore software e competenze avanzate in Python e Java.	Molto difficile
Richiedo competenze tecniche in SQL, Docker e Kubernetes, oltre a soft skills come teamwork e leadership.	Molto difficile
Il candidato deve aver lavorato in aziende medio-grandi e aver gestito progetti complessi.	Molto difficile
Cerco un profilo senior con esperienza internazionale, conoscenza di più lingue e certificazioni professionali.	Molto difficile
Il candidato ideale ha esperienza in progetti di intelligenza artificiale, ha pubblicato articoli scientifici ed è in possesso di un master in AI.	Molto difficile

Il sistema ha fornito risposte corrette per tutte le richieste testate. Di conseguenza, l'aumento della complessità delle richieste non ha avuto alcun impatto sulla qualità delle risposte.

Validazione FASE 3

Per validare, infine, il corretto funzionamento del sistema di scoring implementato, sono stati selezionati 5 CV “pilota” da un corpus complessivo di circa 300, caratterizzati da specifiche proprietà; a partire dalle informazioni contenute in questi 5 CV sono state formulate diverse richieste ad hoc, mirate alla ricerca di determinati requisiti presenti nei Curriculum Vitae scelti.

Lo scopo di questi test è verificare il corretto funzionamento di tutti i componenti e valutare l’effettiva efficacia del sistema nel suo complesso. Per ciascuna richiesta, il sistema restituisce una lista di CV ordinata in modo decrescente in base allo score di similarità calcolato tra la richiesta e i CV totali. Quindi, nella tabella seguente, che riporta i risultati dei test, con il termine match si intende la posizione del CV “pilota” corrispondente alle competenze richieste all’interno della lista restituita dal sistema.

Richiesta	Esito
Mi serve un candidato che abbia esperienza nella consulenza informatica, conoscenze di C, C++, C# e db SQL in particolare MySQL e PostGres.	Primo match
Candidato con almeno 3 anni di esperienza, un dottorato in informatica, che conosca Matlab, R e Python. Deve avere competenze in materia di GDPR, AI Act. Deve essere italiano e conoscere un po’ di inglese.	Primo match
Mi serve un avvocato specialista in compliance, deve avere una conoscenza ottima dell’inglese e deve avere più di 3 anni di esperienza. Inoltre, deve aver avuto esperienza in azienda come consulente di rischio e compliance.	Dodicesimo match
Mi serve un candidato che abbia una laurea in intelligenza artificiale. Deve conoscere C, Java e Python, inoltre deve essere italiano e deve avere un’ottima padronanza dell’inglese. Deve saper lavorare in team e aver pubblicato articoli inerenti all’intelligenza artificiale.	Primo match

Cerco un Project Manager IT con 4+ anni di esperienza in telecomunicazioni, delivery e assurance. Deve aver lavorato su reti FTTH, smart grid e coordinamento gruppi tecnici. Richieste competenze in networking, Linux, cloud Azure, Docker e Kubernetes, oltre a IT governance e cybersecurity.	Primo match
---	-------------

Come è possibile osservare, il sistema funziona correttamente nell'individuare i CV più rilevanti per una data richiesta. In particolare, nella maggior parte dei casi il CV atteso risulta posizionato al primo posto, seguito da tutti gli altri con score minore; tuttavia, emerge un caso in cui esso viene collocato al dodicesimo posto del ranking.

È stata dunque effettuata un'analisi qualitativa in cui il CV è stato isolato e analizzato manualmente per verificare un eventuale malfunzionamento del sistema. Questa operazione ha fatto emergere la perdita di un'informazione nella fase di indicizzazione del CV, che ha compromesso l'esito della ricerca. Infatti, ripetendo la ricerca in cui si bypassa tale campo specifico, il CV è stato individuato correttamente come primo match.

Questo risultato evidenzia come, in presenza di imperfezioni nel processo di indicizzazione, anche una minima perdita informativa possa influenzare significativamente il posizionamento dei CV nell'output fornito. Per questo motivo, il sistema deve essere interpretato come strumento di supporto decisionale, finalizzato a facilitare e velocizzare il processo di selezione, piuttosto che come un meccanismo di sostituzione del giudizio umano.

Di seguito vengono mostrati due esempi di output prodotti dal sistema, relativi alla fase di interrogazione, che illustrano il primo match restituito in base allo score di similarità calcolato.

Si precisa che il valore "Score" non è particolarmente alto in quanto è il risultato della media ponderata di tutti i campi con il relativo peso, che in questo caso è lo stesso per ogni parametro (cioè nessun requisito richiesto ha più importanza degli altri).

RICHIESTA:

Cerco un Project Manager IT con 4+ anni di esperienza in telecomunicazioni, delivery e assurance.

Deve aver lavorato su reti FTTH, smart grid e coordinamento team tecnici. Richieste competenze in networking, Linux, cloud Azure, Docker e Kubernetes, oltre a IT governance e cybersecurity.

CANDIDATI:

— Score: 0.3064

```
{
  "patente" : 0
  "anni_esperienza" : 1
  "ruolo_corrente" : 0.7086
  "competenze_tecniche" : 0.7668
  "competenze_compliance" : 1
  "lingue" : 0
  "esperienze_lavorative" : 0.8137
  "progetti" : 0
  "articoli" : 0
  "formazione" : 0
  "certificazioni" : 0
  "soft_skills" : 0
  "nazionalita" : 0
  "data_di_nascita" : 0
  "nome_file" : "CV_.pdf"
```

RICHIESTA:

Candidato con almeno 3 anni di esperienza, un dottorato in informatica, che conosca matlab, R e Python. Deve avere competenze in materia di GDPR, AI Act. Deve essere italiano e conoscere un po' di inglese.

CANDIDATI:

— Score: 0.4099

```
{
  "patente" : 0
  "anni_esperienza" : 1
  "ruolo_corrente" : 0
  "competenze_tecniche" : 1
  "competenze_compliance" : 1
  "lingue" : 1
  "esperienze_lavorative" : 0
  "progetti" : 0
  "articoli" : 0
  "formazione" : 0.738
  "certificazioni" : 0
  "soft_skills" : 0
  "nazionalita" : 1
  "data_di_nascita" : 0
  "nome_file" : "CV_.pdf"
```

Sviluppi futuri

Lo studio presentato mette in luce il potenziale degli LLM non soltanto come strumenti di generazione testuale, ma soprattutto come componenti fondamentali per la strutturazione semantica di dati non strutturati, trasformando documenti eterogenei in rappresentazioni formali e computabili.

In questa prospettiva, l'approccio proposto può essere esteso a numerosi contesti applicativi in cui la disponibilità di dati strutturati costituisce un prerequisito essenziale per l'automazione di processi complessi, come ad esempio sistemi di question answering, piattaforme di knowledge management, motori di ricerca semantici e architetture basate su agenti intelligenti.

Un primo sviluppo futuro riguarda l'impiego di modelli LLM dotati di una finestra di contesto più ampia e di capacità di ragionamento più avanzate, al fine di migliorare ulteriormente la qualità della fase di strutturazione, riducendo la perdita di informazione e la generazione di contenuti non supportati. Modelli più performanti consentirebbero infatti di catturare in modo più efficace informazioni implicite e relazioni semantiche complesse presenti nei documenti sorgente.

Infine, l'introduzione di meccanismi di feedback umano (human-in-the-loop) permetterebbe di affinare progressivamente il processo di strutturazione, rendendo il sistema adattivo e migliorandone l'affidabilità nel tempo. Ciò rafforza l'idea degli LLM come strumenti centrali non solo per l'interazione linguistica, ma per la costruzione automatica di basi di conoscenza strutturate a partire da fonti testuali non strutturate.

Bibliografia

[1] Retrieval-Augmented-Generation (RAG) per la consultazione locale e sicura di documenti:

https://tinextadefence.it/wp-content/uploads/2025/07/Report_AI4Cyber.pdf

[2] Retrieval-Augmented-Generation (RAG) per la consultazione locale e sicura di documenti: miglioramenti e nuove funzionalità:

https://tinextadefence.it/wp-content/uploads/2025/08/Report_AI4Cyber.pdf

[3] GPT-OSS: <https://openai.com/it-IT/index/introducing-gpt-oss/>

[4] OLLAMA: <https://ollama.com/>



tinexta
defence

Next | Donexit | Foramil | Innodesi

Via Giacomo Peroni, 452 – 00131 Roma
tel. 06.45752720 – info@defencetech.it
www.tinextadefence.it

#TinextaDefenceBusiness