## Conclusione su BAFIT

Dopo svariati test, feedback e risorse continue questi sono i risultati ottenuti dalla AI.

Pero prima di svolgere i test sul modello allenato sono avvenuti i seguenti cambiamenti:

#### 1) cambiamento al prompt template

```
# — prompt template

PROMPT_TEMPLATE = """

Answer the question based only on the following context. Be thorough, detailed, and include all relevant informat

Do a long and complete answer with every information that the user asks for

{context}

---

Question: {question}
```

Ho deciso di fornire all'Al un prompt template migliore cercando di toglierli possibili limitiazione nella sua risposta

#### 2) cambiamento della funzione \_context

```
def _context(self, q: str, k: int = 10): lusage
   docs = vector_db.max_marginal_relevance_search(self._strip(q), fetch_k=20, k=k, lambda_mult=0.3)
   context_chunks = []
   for d in docs:
        source = d.metadata.get("source", "sconosciuto")
        context_chunks.append(f"[Fonte: {source}]\n{d.page_content}")
   return "\n\n---\n\n".join(context_chunks)
```

Ho deciso di modificare la funzione context in modo che l'ai utilizzasse più file per rispondere alle domande fatte.

Il cambiamento effettivo è avvenuto al valore di "k" il quale indica il numero di file utilizzati per rispondere alla domanda.

Precedentemente era **4** il suo valore ma è stato modificato a **10** in modo che l'Al potesse rispondere estraendo informazioni da più file.

#### 3) cambiamento scrittura del feedback

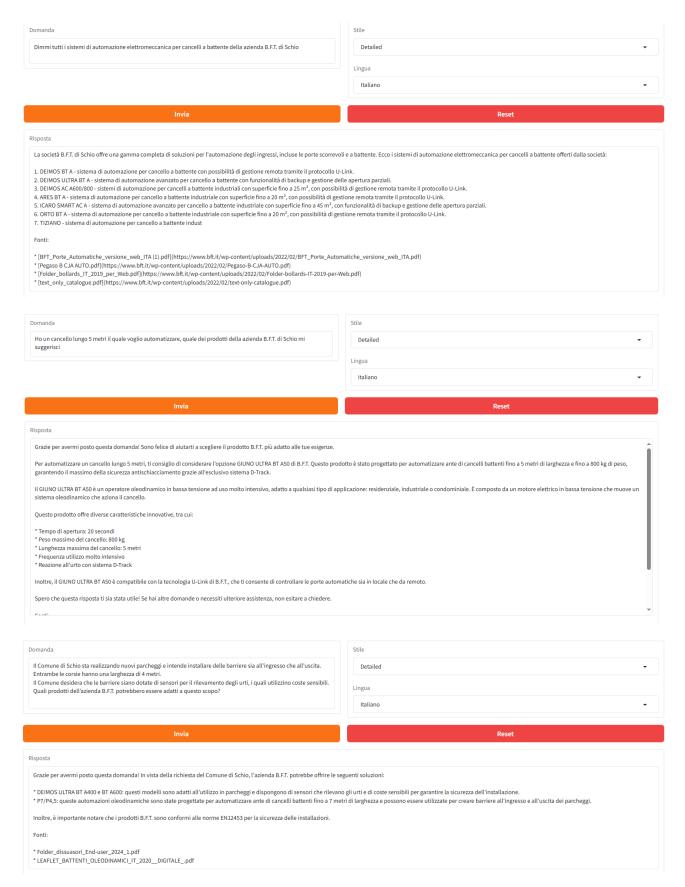
Dopo vari test, si può notare che nella scrittura dei feedback all'interno dei file PDF molte frasi vengano tagliate oppure oltrepassino il margine della pagina, compromettendo così la leggibilità.

Questo problema era dovuto al fatto che precedentemente era utilizzato per la scrittuta un metodo che non prevedeva l'andare a capo riga.

Per risolvere il problema è bastato sostituire il metodo precedente con "Paragraph".

#### Dopo questi cambiamenti questi sono i risultati:

Domanda  Dimmi le specifiche tecniche del VIRGO BT B20	Stile  Detailed   Lingua  Italiano
Invia	Reset
II VIRGO BT B20 è un operatore elettromeccanico irreversibile a braccio articolato per cancelli a battente fino a 200 kg di peso e 2 metri di lunghezza. Ecco le specifiche tecniche del prodotto:  * Modello: VIRGO BT B20 SLAVE  * Codice prodotto: P3930136 00002  * Tipo di motore: Operatore elettromeccanico irreversibile a braccio articolato  * Alimentazione: 24 V  II VIRGO BT B20 è un prodotto ideale per installazioni residenziali che richiedono la gestione di cancelli con ante di grandi dimensioni. Il sistema di leve articolate permette di risolvere facilmente i problemi legati ai pilastri di grandi dimensioni.  Fonti:  [VIRGO BT B.pdf][https://www.bft-automation.com/docs/NRGO_BT_B.pdf] [Specifiche tecniche.pdf][https://www.bft-automation.com/docs/Specifiche_tecniche.pdf] [Specifiche tecniche.pdf][https://www.bft-automation.com/docs/Specifiche_tecniche.pdf]	



Come si può notare questi cambiamenti hanno miglioratò le risposta dell'Al ma comunque le risposte risultano tutt'ora essere incomplete oppure prive di informazioni fondamentali

Per questo ho deciso di provare altri modelli AI per comprendere se questi problemi sono causati da una carenza di informazioni o da limitazioni imposte dai modelli.

### Test Modello DeepSeek

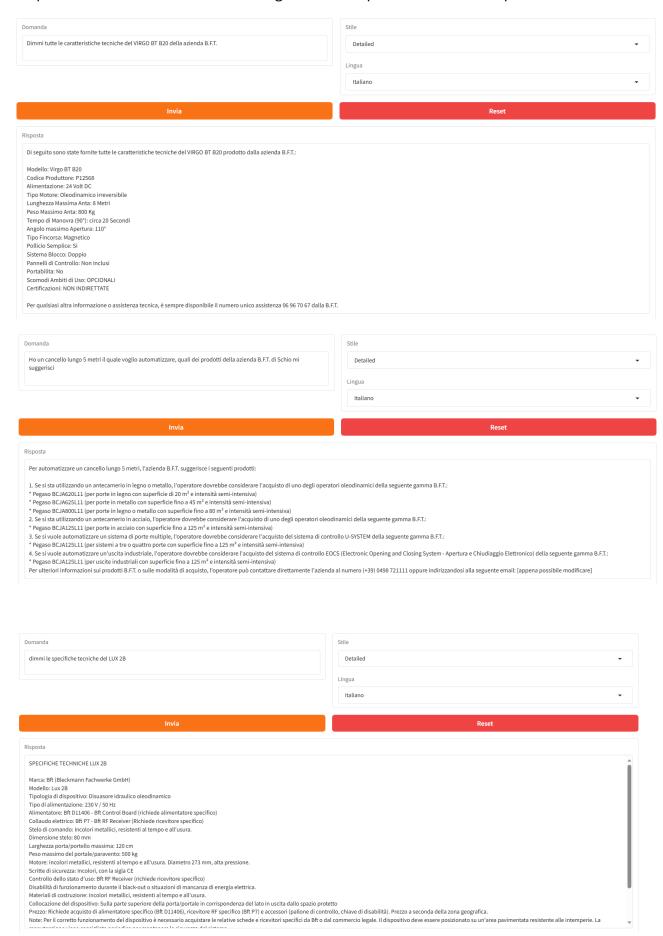
Ho deciso di installare il modello di deepseek-llm:7b.

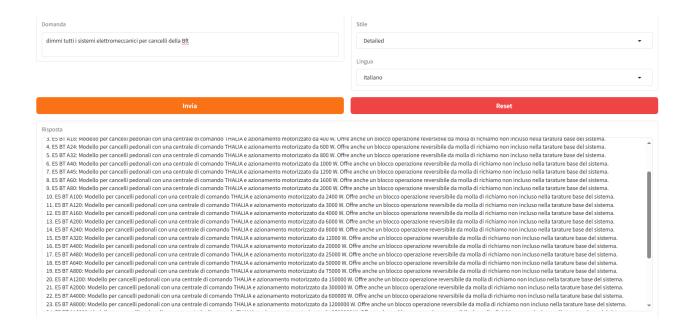
Per poter cambiare modello ho dovuto prima installare il modello in ollama e poi cambiare il file nel programma nominato "config.py"

Ho eseguito il comando: "ollama pull deepseek-llm:7b" per installare il modello su Ollama

Cambio il nome del modello sul file config.py

#### Dopo aver installato il modello ho eseguito dei test per vedere le sue risposte:





Come si può notare le risposte risultano essere molto pù lunghe e complete di informazioni rispetto a quelle precedenti, però i tempi di attesa per la risposta sono triplicati.

Le risposte risultano essere per lo più corrette ma dopo ulteriori prove deepseek risulta fare fatica a svolgere le traduzioni e spesso risponde con informazioni non tradotte in italiano.

Da certe risposte si capisce che uno dei possibili problemi nella qualità delle risposte potrebbe essere dovuto a come esse vengano divise e salvate dal modello

Per risolvere questo problema svolgo un cambiamento nel codice del programma "populate\_database" andando ad aumentare la dimensione di ciascun chunk

Questo aumenta la grandezza di divisioni e diminuisce il numero di chunk e ciò potrebbe impedire che parti di informazioni vengano perse durante il retrieval.

Prima di runnare il programma bisogna svuotare il database e caricare i nuovi chunk.

Per svuotare il database eseguo il programma "ingest.py"

Il comando effettivo è "ingest.py --reset"

```
C:\Users\IslamT\Documents\Sviluppo Python\AI Python>ingest.py --reset
C:\Users\IslamT\Documents\Sviluppo Python\AI Python\ingest.py:16: LangChainDeprecationWarning: Importing OpenAIEmbedding
s from langchain.embeddings is deprecated. Please replace deprecated imports:

>> from langchain.embeddings import OpenAIEmbeddings

with new imports of:

>> from langchain_community.embeddings import OpenAIEmbeddings

You can use the langchain cli to **automatically** upgrade many imports. Please see documentation here <a href="https://python.langchain.com/docs/versions/v0_2/">https://python.langchain.com/docs/versions/v0_2/</a>

from langchain.embeddings.openai import OpenAIEmbeddings # Puoi sostituire con il tuo embedding

Using CPU. Note: This module is much faster with a GPU.

D Database Chroma svuotato.
```

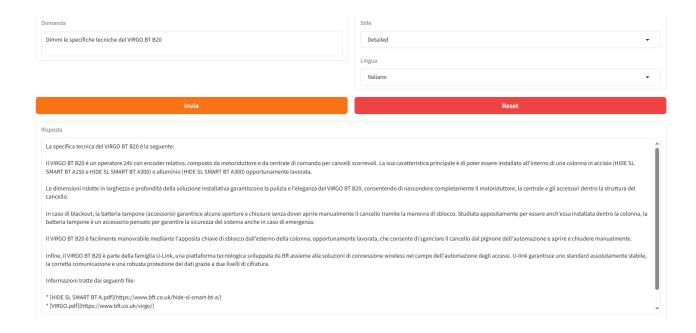
Questo andrà a svuotare il Database Chroma e ci permetterà di ricaricare nuovi chunk dato che sennò il codice lo impedirebbe dato che impedisce che dati uguali vengano caricati più volte.

# Per svolgere i test dopo questo cambiamento ho deciso di rinstallare il modello di llama3:8b

Per ricaricare i chunk nuovi all'interno del database è sufficiente eseguire nuovamente il programma "populate\_database.py"

Ecco i risultati delle domande dopo il cambiamento della grandezza dei chunk e del loro overlap:

Domanda	Stile
Dimmi le specifiche tecniche del PHOBOS	Detailed
	Lingua
	Italiano
Invia	Reset
Risposta	
La richiesta di informazioni sulle specifiche tecniche del PHOBOS è un tema molto importante per la nostra azienda, il B.F.T. Ecco le informazioni che ho reperito dalle nostre fonti:  Il PHOBOS AC B è un operatore elettromeccanico reversibile per cancelletti oscillanti da 230 V, progettato per pesi fino a 400 kg e lunghezze fino a 2,5 m. Il prodotto è caratterizzato da una progettazione innovativa Hi-Tech che ne garantisce la stabilità è l'affidabilità.  Le specifiche tecniche del PHOBOS AC B sono le seguenti:  *Potenza: 230 V  *Frequenza: 50/60 Hz  *Peso massilmo: 400 kg *Lunghezza massilma: 2,5 m  *Velocità di apertura: 30'/sec  *Velocità di chiusura: 20'/sec  *Potere di frenatura: 500 N	
II PHOBOS AC B è dotato di un sistema di controllo elettronico avanzato che garantisce la gestione precisa del movimento del cancelletto. Inoltre, il prodotto è equipaggiato con una serie di accessori e accessori opzionali per soddisfare le esigenze specifiche dei nostri clienti.	
Per quanto riguarda la manutenzione, il PHOBOS AC B presenta una struttura facile da mantenere che consente un agevole acceso alle parti interne del prodotto. Inoltre, il sistema di controllo elettronico è progettato per visualizzare in tempo reale lo stato del cancelletto, facilitando la diagnosi e la manutenzione.	



Risposte che predentemente venivano lasciare incomplete ora contengono più informazioni e risultano anche essere più lunghe ma tutt'ora il modello ha varie imprecisioni che però non risultano essere causate da una carenza di informazioni ma dal modello in se.

## Conclusione Finale

Dopo vari test e cambiamenti si è notato che il modello locale non risulta essere in grado di competere con quello in cloud dato che esso risulta avere una quantità di informazioni nettamente superiore a quella locale ma anche di avere una capacità di calcolo superiore.

Si pensava inizialmente che le imprecisioni nelle risposte del modello fossero causate da carenze di informazioni o di allenamento e controllo ma dopo dei test si è compreso che molte risposte risultano essere incomplete a causa di limitazioni di calcolo dei modelli testati.

Un'altro modo per poter migliorare le risposte sarebbe quello di strutturare meglio le informazioni fornite all'Al in modo da semplificare la comprensioni dei dati e impedire errori nelle risposte causati da incomprensioni.

In sintesi per poter migliorare BAFIT bisognerebbe utilizzare un modello con capacità superiori e utilizzare anche hardware capace di poter sostenere la potenza di calcolo.

Nonostante siano avvenuti questi cambiamenti e miglioramenti i risultati delle risposte continuano ad essere incomplete e insoddisfacenti.