Università degli Studi di Milano Bicocca

Dipartimento di Informatica

Corso di Laurea Magistrale

Architettura RAG su dominio ESG

Massimo Trippetta - 869286

ANNO ACCADEMICO 2024/2025

Indice

Elenco delle figure

Introduzione

Nel contesto attuale, l'informazione relativa alla sostenibilità e agli aspetti ambientali, sociali e di governance (ESG) sta assumendo un ruolo sempre più centrale, sia in ambito finanziario che nella valutazione delle imprese e delle politiche pubbliche [2]. Tuttavia, la maggior parte di questi dati ESG è contenuta in documenti testuali non strutturati e non finanziari, rendendo complessa la loro consultazione e interpretazione automatica.

L'obiettivo di questo progetto è sviluppare un sistema intelligente in grado di generare risposte accurate e coerenti a partire da documenti ESG, utilizzando tecniche di Retrieval-Augmented Generation (RAG) [6]. In particolare, il progetto si propone di integrare un modello linguistico (LLM) con una base documentale vettorializzata e arricchita attraverso un Knowledge Graph (KG) [5], per migliorare la capacità del sistema di comprendere e contestualizzare le richieste dell'utente.

Dopo una fase iniziale di raccolta e pre-processing dei dati, basata su dataset ESG scaricato dalla World Bank [11], il contenuto è stato esplorato, ripulito e suddiviso in chunk testuali, successivamente trasformati in vettori tramite tecniche di embedding [10]. I vettori così ottenuti sono stati memorizzati in un database, consentendo interrogazioni efficienti e contestuali.

Successivamente, è stata costruita l'architettura RAG del sistema, utilizzando il modello linguistico open-source Phi-2 [8], e, successivamente, il modello Phi-3.5. Il framework RAG è stato arricchito con l'integrazione di un KG per favorire un'espansione delle query e una maggiore precisione nelle risposte generate.

Infine, sono state sperimentate diverse tecniche di prompting, come il Direct Retrieval, il Chain-of-Thought [13] e l'approccio Fallback Hybrid, valutando qualitativamente e quantitativamente i risultati ottenuti tramite metriche dedicate, analisi della latenza e dei costi computazionali.

1. Analisi dei requisiti e obiettivi

1.1 Definizione dello scopo

Questo progetto è stato realizzato come parte dell'esame di *Large Scale Data Management*, con l'obiettivo di progettare un sistema in grado di generare risposte intelligenti a domande relative a tematiche ESG (Environmental, Social and Governance). L'attenzione è rivolta principalmente a contenuti testuali non finanziari, presenti in report istituzionali, dataset open data e documentazione ESG.
L'approccio si basa sull'integrazione di tecniche di Natural Language Processing (NLP), architetture di Retrieval-Augmented Generation (RAG) e supporto semantico tramite Knowledge Graph, per ottenere risposte pertinenti, accurate e interpretabili.

1.1.1 Domini coinvolti

Il progetto si colloca all'intersezione di diversi ambiti tematici:

- **Finanza sostenibile**: utilizzo di criteri ESG per orientare decisioni di investimento e valutazione di rischio.
- **Tassonomia ESG**: classificazione di attività economiche sostenibili, come definito da enti normativi (es. Unione Europea).
- **Reportistica non finanziaria**: documenti che descrivono impatti ambientali, sociali e di governance prodotti da aziende o enti pubblici.

1.1.2 Requisiti funzionali

- **Accesso e parsing dei dati ESG**: i dati sono stati raccolti dal portale World Bank Data Catalog[1], contenente indicatori ESG globali per il periodo 2010-2023.

[1] `https://datacatalog.worldbank.org/search/dataset/0037651`

- **Pre-processing e tokenizzazione**: ogni riga del dataset è stata trasformata in testo naturale, tokenizzata con `nltk` e `transformers`, e suddivisa in chunk di dimensione controllata tramite `AutoTokenizer` di HuggingFace.

- **Indicizzazione vettoriale**: i chunk testuali sono stati convertiti in rappresentazioni vettoriali e memorizzati in un vector store basato su `faiss`.

- **Architettura RAG**: la fase di retrieval è seguita da generazione tramite modelli linguistici (LLM), in particolare Phi-2 e Phi-3.5.

- **Integrazione con KG**: tramite query SPARQL, vengono recuperati concetti correlati semanticamente alla query utente, per migliorare il contesto del prompt.

1.1.3 Requisiti non funzionali

- **Performance**: il sistema deve rispondere in tempi più brevi possibile.
- **Scalabilità**: l'architettura è progettata per supportare una crescita progressiva del knowledge base.

Di seguito, è riportato uno schema che riassume il flusso dati del sistema:

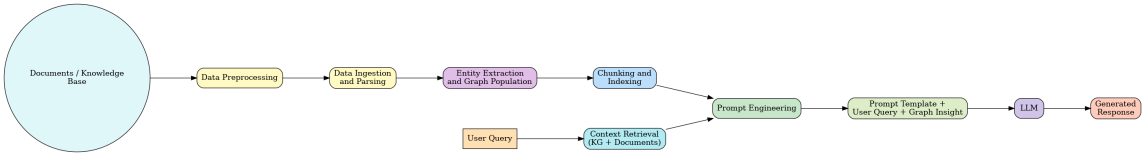

Figura 1.1: Architettura del sistema RAG

2. Accesso e parsing dei dati ESG

2.1 Accesso e parsing dati ESG

2.1.1 Download di un dataset ESG dal World Bank ESG Data Portal

Per alimentare il sistema RAG con contenuti reali e strutturati, è stato selezionato un dataset ESG dal *World Bank ESG Data Portal*[1]. Il dataset contiene 73 indicatori ESG per oltre 200 nazioni, distribuiti lungo un orizzonte temporale che va dal 2010 al 2023.
Ogni riga del dataset rappresenta una combinazione di indicatore ESG, paese e serie storica annuale. Le colonne temporali (es. `2010 [YR2010]`, `2011 [YR2011]`, ecc.) contengono i valori numerici per ciascun anno, mentre le colonne di identificazione (`Series Name`, `Country Name`, ecc.) descrivono l'indicatore e il contesto geografico.

2.1.2 Esplorazione iniziale e pulizia del dataset

La prima fase di pre-processing ha riguardato l'identificazione e la gestione dei valori mancanti. I valori assenti erano rappresentati nel file come `..` e sono stati sostituiti con `NaN` per una corretta gestione statistica e computazionale.
Successivamente, le colonne corrispondenti agli anni sono state convertite in formato numerico (float) per consentire analisi quantitative. Questa trasformazione ha permesso di calcolare rapidamente la percentuale di valori mancanti per ciascun anno. È emerso che i dati sono abbastanza completi fino al 2020, mentre il tasso di assenza cresce in modo significativo dal 2021 in poi, raggiungendo oltre il 97% nel 2023. Per questa ragione, si è scelto di includere nella fase successiva solo i dati dal 2010 al 2020.

[1] `https://datacatalog.worldbank.org/search/dataset/0037651`

Anno	Percentuale di valori mancanti (%)
2010	24.8
2011	25.9
2012	25.2
2013	24.2
2014	23.8
2015	24.5
2016	27.9
2017	29.1
2018	28.3
2019	30.2
2020	34.3
2021	55.6
2022	77.3
2023	97.4

Tabella 2.1: Percentuale di valori mancanti per anno nel dataset ESG

A livello descrittivo, il dataset comprende:

- **73 indicatori ESG unici**.
- **205 nazioni rappresentate**.
- Una media del 25–30% di dati mancanti per anno tra il 2010 e il 2020.

Dopo la pulizia, è stata effettuata una trasformazione da formato *wide* a *long*, utile per alimentare il vector store. In questa trasformazione, le colonne temporali sono state "fuse" in due sole colonne: una per l'anno e una per il valore, rendendo ogni riga un'osservazione univoca (indicatore, paese, anno, valore). I valori nulli sono stati eliminati per mantenere la coerenza del dataset.

Per consentire analisi regionali, ad ogni paese è stato associato il continente di appartenenza utilizzando una funzione di mapping(tramite la libreria pycountry_convert). Questo ha

permesso di ottenere una suddivisione completa per macro-area geografica, utile in fase di query semantica o di visualizzazione.

I dati risultanti presentano la seguente distribuzione per continente:

- **Africa**: 31.500 osservazioni
- **Asia**: 28.676 osservazioni
- **Europa**: 26.639 osservazioni
- **Nord America**: 13.113 osservazioni
- **Sud America**: 7.797 osservazioni
- **Oceania**: 6.939 osservazioni
- **Unknown**: 2.109 osservazioni (es. territori non codificabili)

Questa fase ha permesso di ottenere un dataset coerente, pulito e già pronto per le successive fasi di tokenizzazione e embedding.

2.1.3 Tokenizzazione e suddivisione in chunk

Dopo la fase di pulizia, si è passati alla trasformazione dei dati ESG in formato testuale utile per l'indicizzazione semantica. Per ogni riga del dataset in formato long è stato generato un testo descrittivo in inglese, che sintetizza l'informazione chiave in una frase, includendo l'indicatore, il valore, il paese, il continente e l'anno. Ad esempio:

> *"In 2019, the indicator 'People using safely managed sanitation services (% of population)' in Finland (Europe) had a value of 89.53."*

Successivamente, si è diviso i testi in *chunk* di dimensioni controllate, per evitare il superamento dei limiti di token dei modelli linguistici. È stato utilizzato il tokenizer `all-MiniLM-L6-v2` di HuggingFace. Ogni frase è stata analizzata e, se necessaria, suddivisa in più unità testuali in modo che ogni chunk contenesse al massimo 128 token.

L'intero processo ha generato una nuova colonna contenente le liste di chunk, poi "esplose" riga per riga tramite `pandas.explode()`, ottenendo così un record univoco per ogni chunk valido. I chunk vuoti o nulli sono stati scartati, e il DataFrame è stato infine reindicizzato. Il numero totale di chunk ottenuti è pari a **116.773**.

2.1.4 Embedding dei chunk

La fase successiva ha riguardato la trasformazione dei chunk testuali in rappresentazioni vettoriali dense (embedding), utilizzando lo stesso modello precedente(`all-MiniLM-L6-v2`), un modello leggero con circa 22 milioni di parametri, che genera embedding a 384 dimensioni e offre buone prestazioni in task di semantic search con un basso impatto computazionale.
L'intero set di chunk è stato elaborato in mini batch da 64 elementi. Il risultato è stato una matrice NumPy di forma `(116773, 384)`, contenente gli embedding vettoriali normalizzati, pronti per l'indicizzazione.

2.1.5 Salvataggio su Database

Per consentire ricerche semantiche sui chunk ESG, è stato creato un indice vettoriale basato su FAISS (*Facebook AI Similarity Search*). Si è scelta la distanza L2 (distanza euclidea), particolarmente adatta al tipo di embedding generati da MiniLM.
L'indice è stato popolato con i 116.773 vettori generati. È stato poi eseguito un test di retrieval semantico per verificarne il funzionamento, interrogando l'indice con una domanda tipo ("Access to clean cooking energy in African countries in 2015") e recuperando i 5 chunk più simili, che si sono rivelati coerenti e rilevanti rispetto alla query.
Per garantire riutilizzabilità del sistema, sono stati salvati tre oggetti distinti:

- **esgchunksmetadata.csv**: contiene i chunk testuali con i relativi metadati (indicatore, paese, anno, continente).
- **esgembeddings.npy**: matrice NumPy degli embedding a 384 dimensioni.
- **esgfaiss.index**: file binario dell'indice FAISS per query di similarità.

3. Costruzione dell'architettura RAG

L'architettura RAG (Retrieval-Augmented Generation) rappresenta una delle evoluzioni più efficaci dell'NLP moderno per rispondere a domande complesse sfruttando documenti esterni. Essa combina un modulo di recupero dei documenti (retriever) con un modulo generativo (LLM) che utilizza il contesto recuperato per formulare risposte coerenti, accurate e fondate su dati. In questo progetto, l'architettura è stata estesa con l'integrazione di un Knowledge Graph per migliorare la semantica del contesto e la precisione delle risposte. Come sottolineato in [12], il design del prompt e la qualità del contesto sono elementi centrali nel costruire un buon sistema RAG.

3.0.1 Integrazione di Phi-2 come LLM

Il modello linguistico scelto per la generazione è `Phi-2`, sviluppato da Microsoft [7]. Si tratta di un modello in grado di produrre risposte coerenti e comprensibili, senza necessità di accesso a infrastrutture cloud esterne. Il modello è stato caricato tramite Hugging Face Transformers, sfruttando la pipeline di `text-generation` e configurato per generare risposte fino a 256 token. Successivamente verrà provato anche l'utilizzo di Phi-3.5 Mini per la generazione: compatto, potente, ottimo su molti benchmark, con supporto per contesti lunghi. Supera modelli più grandi in benchmark [4].
Il suo ruolo è quello di generare risposte guidate da prompt costruiti a partire dal contesto ESG recuperato dai chunk indicizzati. L'utilizzo in locale consente un maggiore controllo, riduzione dei costi e adattabilità del sistema.

3.0.2 Framework RAG

Il sistema RAG è stato implementato utilizzando LangChain, che permette di orchestrare la combinazione tra modelli generativi, meccanismi di retrieval e strutture di prompt dinamici. Come modulo di embedding è stato utilizzato `all-MiniLM-L6-v2`, mentre l'indicizzazione vettoriale è stata effettuata con FAISS. Anche in questo caso è stato testato

anche un altro modello, ovvero bge-small-en-v1.5, il quale risulta Compatto (33M parametri), perfetto per ambienti a risorse limitate, ottimizzato per il retrieval, è infatti uno dei migliori modelli leggeri su task di retrieval nel benchmark MTEB [3].
Ogni chunk di testo ESG è stato trasformato in un oggetto `Document`, arricchito con metadati geografici e temporali, utile per il recupero contestuale. Il retriever è stato connesso all'LLM tramite un wrapper LangChain, e alimentato da un prompt template specializzato per contesto ESG.
Il primo test ha mostrato la capacità del sistema di recuperare i chunk rilevanti e sintetizzarli in una risposta chiara e focalizzata. L'integrazione con LangChain permette anche il tracciamento dei documenti usati come fonte.

3.0.3 KG Integration

Il Knowledge Graph (KG) è stato introdotto per rafforzare la comprensione semantica e strutturare la conoscenza. Le informazioni ESG, sono state trasformate in triple RDF del tipo:

(entità soggetto) – (relazione) – (entità oggetto)

Ad esempio:

(Africa) – (hasAverageAccessToCleanCooking) – (32.4%)

Il grafo è stato costruito con `rdflib` a partire dai chunk ESG e salvato in formato Turtle. Ogni osservazione ESG è rappresentata come un nodo RDF con tipo `ESGObservation` e relazioni per paese, indicatore, valore, anno e continente. Il grafo risultante contiene oltre 700.000 triple.

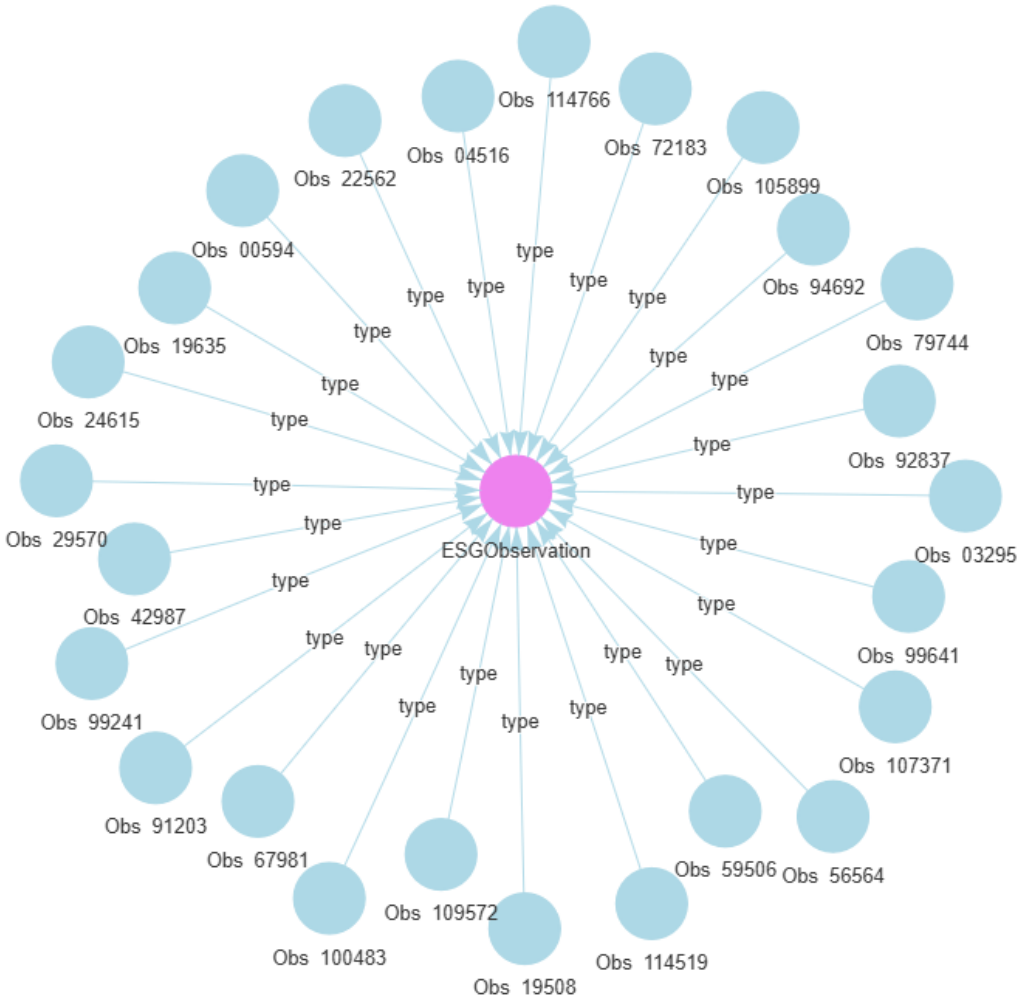

Figura 3.1: Visualizzazione semplificata del Knowledge Graph ESG

Una volta costruito il grafo, sono state effettuate interrogazioni SPARQL per estrarre informazioni specifiche, ad esempio tutte le osservazioni ESG in Africa nel 2015. Questo ha permesso di verificare il funzionamento del lavoro svolto.

Come mostrato in [9], l'integrazione tra Knowledge Graph e RAG è particolarmente potente in scenari in cui:

- Il retrieval tradizionale fallisce o restituisce contesto ambiguo.
- La domanda richiede vincoli logici (es. "top N", "filtro per anno o area").
- Si vuole costruire un contesto controllato e verificabile.

La pipeline finale del sistema è quindi ibrida: unisce documenti testuali recuperati via FAISS e conoscenza strutturata in forma di triple RDF. Queste vengono formattate nel prompt finale sotto la sezione ”Context from knowledge graph”, a supporto della generazione.

Infine è stata aggiunta l'estrazione automatica di entità (paese, anno, continente, indicatore) dalla domanda tramite parsing ibrido, e un sistema di stima dinamica della dimensione del contesto da recuperare (`estimate context size`). Per ottimizzare l'efficienza e la rilevanza del sistema RAG, è utile stimare dinamicamente la quantità di contesto da includere (chunk testuali e triple RDF) in base al tipo di domanda posta dall'utente. Il metodo estimate context size fornisce una strategia semplice e per regolare il numero di elementi informativi da recuperare, in funzione di parole chiave rilevate nel testo della query. Ecco come sono state scelte:

Tabella 3.1: Criteri euristici per la stima della dimensione del contesto nella pipeline RAG

Categoria della domanda	**Parole chiave rilevate**	**Contesto stimato**
Temporale (trend, medie)	`"average"`, `"trend"`, `"from"`, `"between"`, `"range"`	50
Distribuzione/Estremi	`"maximum"`, `"minimum"`, `"max"`, `"min"`, `"lowest"`, `"highest"`, `"top"`	30
Fattuale diretta	(nessuna parola chiave specifica)	5

4. Tecniche di Prompting

Le tecniche di prompting rappresentano una componente cruciale nei sistemi Retrieval-Augmented Generation (RAG), in quanto determinano come il modello linguistico interpreta e sfrutta il contesto fornito per rispondere a una domanda. Un prompt ben progettato può migliorare drasticamente l'accuratezza, la coerenza e la trasparenza delle risposte generate[1].

In questo capitolo, esploreremo tre tecniche fondamentali di prompting applicate al nostro sistema ESG-RAG:

Direct Retrieval

Chain-of-Thought

Fallback Hybrid

Successivamente, introdurremo una quarta variante ottimizzata del prompt, basata su Direct Retrieval, adattata per massimizzare la qualità delle risposte usando phi3.5 mini e bge come modello di embedding.

Le tre tecniche principali sono tratte e ispirate dall'articolo [12].

4.0.1 Direct Retrieval

Questa tecnica si basa esclusivamente sui documenti recuperati dal retriever. Risulta ideale per domande fattuali e valutazioni basate su evidenze certe.

```
 You are a helpful assistant specialized in ESG indicators.  Use
ONLY the information from the provided documents to answer the
question concisely and informatively.  Do NOT use external knowledge
or the knowledge graph.
Context from documents:
Question:
Answer:
```

4.0.2 Chain-of-Thought

Questa tecnica guida il modello attraverso una catena di ragionamento esplicito, passo dopo passo. Risulta utile per domande analitiche o con più passaggi logici.

```
You are a data analyst specialized in ESG data. Use the following
context to answer the question step-by-step concisely and informatively.
First extract relevant facts, then explain your reasoning, and
finally present the answer.
Context from documents:
Context from knowledge graph:
Question:
Answer:
```

4.0.3 Fallback Hybrid

Questa tecnica combina retrieval e conoscenza del modello, è adatta per demo interattive o domande con informazioni parziali.

```
You are a helpful assistant specialized in ESG indicators and
sustainability data. Use the following context to answer the
question concisely and informatively. If they do NOT contain
sufficient information, you may rely on your own knowledge, but
clearly state that you are doing so.
Context from documents:
Context from knowledge graph:
Question:
Answer:
```

4.0.4 Prompt Ottimizzato (Phi-3.5 + BGE)

Il seguente prompt è una versione raffinata di *Direct Retrieval*, testato utilizzando il modello `Phi-3.5 Mini` e l'embedding model `bge-small-en-v1.5`. L'obiettivo è migliorare la qualità delle risposte.
Le principali differenze introdotte sono:

1. Linguaggio più diretto e vincolante (*DO NOT speculate*)
2. Inclusione di esempi strutturati in stile ICL (In-Context Learning)
3. Filtro esplicito per risposte fuori dominio (*I can't answer based on the provided data*)

```
 You are a concise assistant specialized in ESG indicators and
sustainability data.
ONLY use the information from the provided context to answer the
user's question.  DO NOT speculate, expand, predict, or explain
unless explicitly asked.  If the answer is not in the context,
say "I can't answer based on the provided data."
Example | Environment (E)
Context:  In 2017, the indicator 'CO2 emissions (metric tons per
capita)' in Vietnam was 2.1.  Question:  What were the CO2 emissions
per capita in Vietnam in 2017?  Answer:  The CO2 emission in Vietnam
in 2017 were 2.1 metric tons per capita
Example | Social (S)
Context:  In 2012, the indicator 'Access to basic sanitation services
(% of population)' in Uganda was 18.3%.  Question:  What was the
sanitation access rate in Uganda in 2012?  Answer:  The sanitation
access rate in Uganda in 2012 was 18.3%
Example | Governance (G)
Context:  In 2019, the indicator 'Voice and Accountability:  Estimate'
in Hungary was -0.37.  Question:  What was the voice and accountability
```

```
score for Hungary in 2019? Answer: 'Voice and Accountability:
Estimate' in Hungary in 2019 was -0.37
Example | Direct
Context: In 2020, the indicator 'Renewable electricity output
(% of total electricity output)' in Canada was 67.1%. Question:
What was the share of renewable electricity in Canada in 2020?
Answer: The share of renewable electricity in Canada in 2020
was 67.1%
Example | Chain-of-Thought (CoT)
Context: In 2010, South Korea had a forest area of 63.4%. In
2015, it was 63.8%, and in 2020, it was 64.1%. Question: How
did forest area change in South Korea between 2010 and 2020? Answer:
It increased gradually from 63.4% in 2010 to 64.1% in 2020.
Example | Fallback
Context: In 2015, the data includes access to electricity and
CO2 emissions for Bolivia. Question: What are the projected
ESG risks in Bolivia by 2035? Answer: I can't answer based on
the provided data.
Example | Out-of-scope
Context: No data is available on interplanetary development.
Question: What is the average ESG score of the Moon in 2020?
Answer: I can't answer based on the provided data.
Now answer:
Context from documents:
Context from knowledge graph:
Question:
Answer:
```

Nel prossimo capitolo classificheremo e testeremo i diversi tipi di domande che abbiamo inserito in quest'ultimo prompt per confrontare le performance delle tecniche di prompting presentate.

5. Testing e valutazione

In questo capitolo andremo a vedere la fase finale del progetto: il testing e la valutazione dell'architettura RAG sviluppata.
Per misurare l'efficacia del sistema, è stato costruito un benchmark composto da 35 domande, progettate per coprire un ampio spettro di situazioni: domande di tipo ESG (Environmental, Social, Governance), domande fuori ambito, domande multiple, contraddittorie e domande calibrate sulle diverse tecniche di prompting sperimentate (Direct, Chain-of-Thought, Fallback Hybrid).
Le risposte ottenute saranno valutate sia manualmente, in termini di correttezza e coerenza, sia utomaticamente tramite il punteggio ROUGE-L, che misura la sovrapposizione di sottostringhe comuni tra risposta attesa e generata, utile per valutare la similarità testuale.

5.0.1 Testing qualitativo – Domande manuali costruite

La costruzione del benchmark manuale ha lo scopo di testare in profondità la robustezza e l'affidabilità del sistema. Di seguito riportiamo la logica di costruzione delle domande:

Tipo di domanda	**Obiettivo**	**Note aggiuntive**
E, S, G (5 ciascuno)	Coprire le 3 dimensioni ESG	Variazione di temporalità, paese vs. continente e tipologia metrica.
Out-of-scope ESG (5)	Testare fallback / allucinazioni	Domande con risposte assenti nel dataset per verificare come gestisce domande fuori ambito.
Tecnica-specifica (5 per tecnica)	Verificare il comportamento sotto ciascun prompting	Include: 1 semplice, 1 comparativa, 1 temporale, 1 aggregata, 1 ambigua per Direct, CoT e Fallback.

Tabella 5.1: Strategia di costruzione del benchmark

Tutte le domande sono accompagnate da una risposta attesa, ottenuta tramite analisi diretta dei dati o, nel caso di domande fuori ambito, marcate esplicitamente come “No correct answer”.

Nella Tabella successiva è riportato il benchmark completo, composto da 35 domande suddivise per categoria e tipologia.

Question	**Category**	**Purpose**	**Expected Answer**
What was the percentage of access to clean fuels in Nigeria in 2015?	E	benchmark	The percentage of access to clean fuel in Nigeria in 2015 was 5.6
Which country in Asia had the highest CO2 emissions in 2012?	E	benchmark	Qatar had the highest CO2 emissions in Asia in 2012
How did renewable energy usage change in Brazil between 2010 and 2020?	E	benchmark	Between 2010 and 2020, renewable energy consumption in Brazil initially decreased from 46.81% in 2010 to a low of 41.71% in 2014. After that, it steadily increased, reaching 50.05% in 2020.
What is the average air pollution level in South Africa in 2018?	E	benchmark	The average air pollution level in South Africa in 2018 was 28.754018

Question	Category	Purpose	Expected Answer
Which European country had the lowest forest coverage in 2016?	E	benchmark	Monaco had the lowest forest coverage in Europe in 2016.
What was the literacy rate in India in 2011?	S	benchmark	The literacy rate in India in 2011 was 69.302559
Which African country had the highest access to clean water in 2019?	S	benchmark	Chad had the highest access to clean water in Africa in 2019.
How did the female labor force participation evolve in Mexico from 2010 to 2020?	S	benchmark	Overall, the female labor force participation in Mexico improved over the decade, increasing by about 3 percentage points, with minor fluctuations.
What is the child mortality rate in France in 2015?	S	benchmark	The child mortality rate in France in 2015 was 4.2
Which country in South America had the best school enrollment rate in 2014?	S	benchmark	Colombia had the best school enrollment rate in South America in 2014
What was the Corruption Perception Index score of Kenya in 2017?	G	benchmark	The Corruption Perception Index score of Kenya in 2017 was -0.975331

Question	Category	Purpose	Expected Answer
Which European country had the highest level of government effectiveness in 2016?	G	benchmark	Switzerland had the highest level of government effectiveness in Europe in 2016.
How did political stability vary in Nigeria from 2010 to 2020?	G	benchmark	Between 2010 and 2020, Nigeria experienced persistent low levels of political stability. The indicator fluctuated around -2.0, with the lowest point in 2010 (-2.21) and a slight improvement observed in 2016 (-1.87) and 2020 (-1.89). However, overall, the country's political environment remained unstable throughout the decade.
What is the rule of law indicator for Colombia in 2015?	G	benchmark	The rule of law indicator for Colombia in 2015 is -0.246755
Which Asian country had the lowest voice and accountability score in 2012?	G	benchmark	The Asian country with the lowest voice and accountability score in 2012 Turkmenistan
What is the percentage of clean fuel usage on Mars in 2020?	Out-of-scope	fallback	No correct answer

Question	Category	Purpose	Expected Answer
What are the ESG scores of Atlantis in 2015?	Out-of-scope	fallback	No correct answer
How did the unicorn population impact clean energy adoption?	Out-of-scope	fallback	No correct answer
What was the water quality in Narnia in 2016?	Out-of-scope	fallback	No correct answer
What is the gender equality index in Wakanda for 2017?	Out-of-scope	fallback	No correct answer
Provide the clean fuel access data for Kenya in 2010.	Direct	prompt-test	The clean fuel access data for Kenya in 2010 is 7.1
Give the renewable energy share in Germany in 2015.	Direct	prompt-test	The renewable energy share in Germany in 2015 is 14.55
How strong was the air pollution in China in 2012?	Direct	prompt-test	The air pollution rate in China in 2012 was 58.905136
List the CO2 emission levels of France for 2011.	Direct	prompt-test	The CO2 emission levels of France for 2011 is 5.13 metric tons per capita.
State the forest area percentage in Indonesia in 2018.	Direct	prompt-test	The forest area percentage in Indonesia in 2018 was 49.716818

Question	Category	Purpose	Expected Answer
Compare the clean water access between Nigeria and Ghana in 2015.	CoT	prompt-test	The clean water access in Nigeria (24.965881) is lower than Ghana (33.249600)
How did access to electricity improve in rural India from 2010 to 2020?	CoT	prompt-test	Between 2010 and 2020, access to electricity in India improved significantly. The share of the total population with access to electricity increased from 76.3% in 2010 to 96.5% in 2020. While this indicator refers to the total population and not specifically rural areas, the steady increase suggests that rural electrification also progressed substantially during this period.
Identify the country with the highest increase in literacy rate between 2012 and 2018.	CoT	prompt-test	The country with the highest increase in literacy rate between 2012 and 2018 is Bangladesh which increased by 16%, from 57.860748 in 2012 to 73.912201 in 2018.
Compare governance effectiveness in Brazil and Argentina in 2016.	CoT	prompt-test	Government effectiveness in Brazil (-0.248415) in 2016 was lower than in Argentina (0.219069)

Question	Category	Purpose	Expected Answer
Track the trend of forest coverage in Congo Rep. over the last decade.	CoT	prompt-test	Between 2010 and 2020, the forest area in the Republic of Congo gradually declined. The percentage of land area covered by forest decreased from 64.64% in 2010 to 64.26% in 2020, showing a slow but steady downward trend throughout the decade. Despite relatively high overall coverage, the data reveals a consistent reduction year over year, indicating ongoing pressure on forest resources.
What are the latest ESG trends in Africa for 2025?	Fallback	prompt-test	
Summarize the overall ESG performance of Southeast Asia.	Fallback	prompt-test	
How is clean energy adoption expected to evolve by 2030?	Fallback	prompt-test	
Give an overview of social development in Sub-Saharan Africa.	Fallback	prompt-test	
What future risks are associated with ESG factors in the Arctic?	Fallback	prompt-test	

Da notare il fatto che nelle ultime 5 domande non sia presente una risposta corretta, essendo domande futuristiche, sono basate su previsioni.

5.1 Testing quantitativo – Metriche

Per valutare le prestazioni del sistema sono stati condotti test quantitativi utilizzando due metodi principali:

- **Valutazione manuale binaria**: ogni risposta viene confrontata con quella attesa. Se il contenuto informativo è corretto (anche in caso di differenze grammaticali o di forma), viene assegnato il punteggio 1 (corretto); in caso contrario, 0 (sbagliato).
- **Valutazione automatica tramite ROUGE-L**: questa metrica misura l'overlap tra la risposta generata e quella attesa, considerando la Longest Common Subsequence (LCS). Sebbene utile per confronti automatici, non sempre è affidabile per valutare risposte corrette dal punto di vista semantico, poiché penalizza parafrasi o riformulazioni corrette.

Di seguito i risultati aggregati della valutazione binaria manuale:

Tabella 5.3: Accuratezza complessiva per metodo (valutazione manuale)

Metodo	**Risposte corrette (1)**	**Risposte errate (0)**	**Accuratezza (%)**
Direct Retrieval (DR)	17	18	48.57
Chain of Thought (CoT)	9	26	25.71
Fallback Hybrid	14	21	40.00
Prompt ottimizzato	28	7	**80.00**

Si evidenzia come la combinazione finale (Prompt ottimizzato) raggiunga un'accuratezza nettamente superiore agli altri approcci, dimostrando come, mostrare degli esempi nel prompt, insieme ad un modello leggermente migliore, porti a risultati nettamente superiori.

5.1.1 Accuratezza per categoria

Per ogni metodo sono state analizzate anche le percentuali di accuratezza per categoria (E, S, G, Direct, CoT, Fallback, Out-of-scope):

Tabella 5.4: Accuratezza per categoria – Direct Retrieval

Categoria	**Accuratezza (%)**
E	60.0
S	40.0
G	60.0
Direct	60.0
CoT	60.0
Fallback	40.0
Out-of-scope	20.0

Il metodo di *Direct Retrieval* si comporta in modo equilibrato tra le diverse categorie, ottenendo buoni risultati soprattutto nelle domande dirette e di governance (G). Tuttavia, mostra debolezza sui casi fuori dominio (Out-of-scope).

Tabella 5.5: Accuratezza per categoria – Chain of Thought

Categoria	**Accuratezza (%)**
E	20.0
S	20.0
G	20.0
Direct	60.0
CoT	20.0
Fallback	40.0
Out-of-scope	0.0

Il *Chain of Thought* ha mostrato performance inferiori nelle categorie ESG. Solo le domande dirette raggiungono una buona accuratezza. Il metodo non sembra efficace su quesiti out-of-scope.

Tabella 5.6: Accuratezza per categoria – Fallback Hybrid

Categoria	**Accuratezza (%)**
E	40.0
S	20.0
G	40.0
Direct	60.0
CoT	20.0
Fallback	80.0
Out-of-scope	20.0

La strategia *Fallback Hybrid* migliora la gestione delle domande di fallback (80%) e mantiene prestazioni accettabili per ESG, ma ha ancora margini di miglioramento nella componente sociale (S) e nella gestione del ragionamento (CoT).

Tabella 5.7: Accuratezza per categoria – Prompt ottimizzato

Categoria	**Accuratezza (%)**
E	60.0
S	80.0
G	80.0
Direct	**100.0**
CoT	40.0
Fallback	**100.0**
Out-of-scope	**100.0**

Il prompt finale ottimizzato rappresenta la configurazione più efficace, raggiungendo il 100% di accuratezza nelle domande dirette, di fallback e fuori dominio. Anche sulle categorie ESG mostra performance molto solide, confermando la bontà dell'approccio finale.

5.1.2 Valutazione ROUGE-L

Di seguito i valori medi di ROUGE-L calcolati sulle risposte, da specificare che, le risposte senza una risposta attesa, non sono state utilizzate per il calcolo della metrica:

Tabella 5.8: ROUGE-L medio per metodo

Metodo	**ROUGE-L medio**
Direct Retrieval (DR)	0.5787
Chain of Thought (CoT)	0.2423
Fallback Hybrid	0.3205
Final RAG	0.4898

Sebbene le metriche ROUGE forniscano un'indicazione automatica utile, non sempre corrispondono alla qualità semantica effettiva della risposta. Ad esempio, risposte corrette ma riformulate possono ottenere punteggi ROUGE bassi. Per questo motivo, utilizziamo ROUGE come metrica ausiliaria, affiancandola ad una valutazione umana.

5.2 Misurazione di latenza, costi, qualità

A questo punto andremo a misurare la latenza e il costo di ogni tipologia di prompt utilizzata, lo faremo in due modi:

- **Numero di token generati**: ci permette di capire quanti token utilizza in media ogni prompt per dare una risposta. Un risultato migliore si ottiene fornendo la risposta giusta con il numero minore di token, permettendo di ridurre i costi ed il tempo.
- **Tempo risposta (ms)**: ci permette di capire quale prompt spende più tempo per rispondere, ovviamente l'obiettivo è quello di ricevere risposte corrette nel minor tempo.

5.2.1 Numero di token generati

Andiamo a vedere, in media, quanti token ha generato ogni prompt:

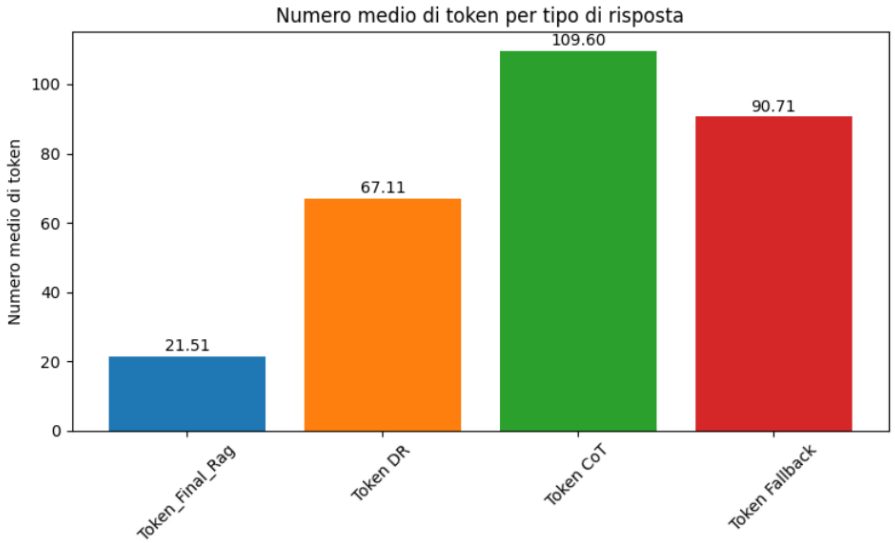

Figura 5.1: Numero medio di token generati per tipo di risposta

Il grafico in Figura 5.1 mostra il numero medio di token generati nei vari metodi di risposta. È evidente che il metodo **Chain of Thought (CoT)** genera risposte significativamente più lunghe, con una media di 109.60 token, seguito da **Fallback Hybrid** con 90.71 e **Direct Retrieval** con 67.11. Al contrario, il sistema **Final Rag**, che rappresenta la strategia di prompting ottimizzato, produce risposte molto più concise con una media di soli 21.51 token.

Questo risultato è in parte attribuibile alla configurazione di inferenza: per il metodo **Final Rag** è stato impostato un limite massimo di max_new_tokens = 64 per contenere la lunghezza della risposta, mentre gli altri metodi utilizzavano un limite più permissivo (128 token). Tuttavia, la sintesi efficace del Final Rag non si spiega solo con la limitazione tecnica: la struttura del prompt ottimizzato spinge il modello a rispondere in maniera più diretta e informativa, evitando divagazioni come avviene nel CoT.

Inoltre, le risposte CoT tendono a includere ragionamenti espliciti passo-passo, che naturalmente aumentano la lunghezza della generazione. Il Fallback Hybrid risente invece di una doppia concatenazione tra i risultati semantici e una possibile elaborazione aggiuntiva,

spiegando il numero di token relativamente elevato. Direct Retrieval si colloca nel mezzo, offrendo risposte dirette ma talvolta ridondanti.

Questo confronto evidenzia un trade-off rilevante tra *lunghezza* e *efficienza*: il metodo Final Rag, pur generando risposte molto più brevi, è anche quello con accuratezza più elevata, come discusso nelle sezioni precedenti.

5.2.2 Tempo risposte(ms)

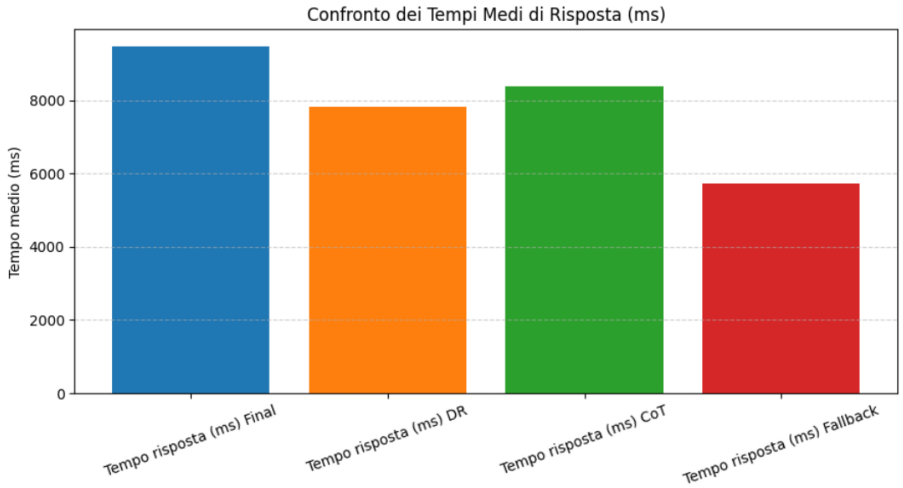

Figura 5.2: Confronto dei tempi medi di risposta (in millisecondi) per ogni metodo

Il grafico in Figura 5.2 confronta i tempi medi di risposta dei vari approcci utilizzati nel sistema RAG. I risultati evidenziano alcune differenze rilevanti:

Il metodo più lento risulta essere il **Final Rag**, con un tempo medio di risposta superiore a 9000 ms. Questo è attribuibile al fatto che, pur generando meno token rispetto agli altri (come visto nel grafico precedente), utilizza un prompt ottimizzato più articolato, che richiede una maggiore elaborazione semantica prima di generare la risposta. Inoltre utilizza due modelli più performanti e costosi come Phi 3.5 e Bgesmall1.5 per l'embedding.

Il **Chain of Thought** ha tempi leggermente inferiori (circa 8400 ms), ma comunque più alti del Direct Retrieval, a causa della maggiore lunghezza delle risposte generate.

Il metodo **Direct Retrieval** si posiziona su un tempo medio di circa 7800 ms, evidenziando un buon compromesso tra efficienza e qualità.

Il più rapido è il metodo **Fallback Hybrid**, con un tempo medio intorno ai 5700 ms. Questo risultato sorprendente è dovuto al fatto che in molti casi il sistema seleziona una risposta già presente nei risultati di retrieval senza generare testi lunghi o complessi.
In sintesi, mentre il metodo *Final Rag* garantisce accuratezza elevata, il suo costo computazionale è il più alto. Il metodo *Fallback*, pur essendo veloce, presenta risultati qualitativamente inferiori. Tali osservazioni sono cruciali per bilanciare precisione ed efficienza in contesti reali.

Conclusioni

L'approccio basato su prompt ottimizzati (Final RAG) ha ottenuto i risultati migliori in termini di accuratezza, raggiungendo l'80% di risposte corrette, superando nettamente tutte le altre strategie testate (Direct Retrieval: 48.6%, CoT: 25.7%, Fallback: 40%).
Inoltre, il metodo ha mostrato una maggiore robustezza nelle domande fuori dominio (out-of-scope), evitando allucinazioni e proponendo risposte prudenti o nulle nei casi in cui i dati non erano presenti nei documenti.
Tuttavia, le metriche automatiche come **ROUGE-L** hanno evidenziato alcune limitazioni: ad esempio, il metodo Chain of Thought ha ottenuto valori bassi (0.24) pur generando talvolta risposte più ragionate, mentre il Final RAG ha ottenuto un valore medio pari a 0.49. Questo conferma che, nel contesto ESG, metriche come ROUGE non sono sempre affidabili per valutare la correttezza semantica, soprattutto in presenza di riformulazioni linguistiche o risposte sintetiche ma corrette.
Efficienza e complessità:
Dal punto di vista computazionale, il metodo Final RAG ha tempi di risposta medi più elevati (oltre 9000 ms), dovuti alla maggiore elaborazione semantica del prompt e ai modelli più pesanti, ma compensa in accuratezza e stabilità. In contrasto, il metodo Fallback è risultato il più veloce (circa 5700 ms) ma meno preciso.
La lunghezza media delle risposte ha seguito la stessa tendenza: il metodo Final RAG ha generato risposte più brevi e concise (21 token medi), mentre il metodo CoT ha prodotto le più lunghe (oltre 100 token), a dimostrazione del diverso stile generativo.
Considerazioni finali:
Nel complesso, l'approccio RAG si è dimostrato valido per il dominio ESG, offrendo un compromesso solido tra precisione, controllo e flessibilità. Il progetto ha evidenziato come l'integrazione tra retrieval e generazione, tramite l'aiuto dei Knowledge Graph, se ben bilanciata, può ridurre le allucinazioni, migliorare la pertinenza delle risposte e garantire un comportamento più affidabile anche in domini tecnici e sensibili.
Infine, questo lavoro ha fornito una base replicabile per la costruzione di benchmark ESG

personalizzati, sottolineando l'importanza di test qualitativi manuali oltre che metriche automatiche.

Bibliografia

[1] Oracle AI e Data Science Blog. *Enhancing RAG with Advanced Prompting*. https://blogs.oracle.com/ai-and-datascience/post/enhancing-rag-with-advanced-prompting. Accessed: 2025-06-23. 2024.

[2] European Commission. *Sustainable Finance and ESG*. Accessed: 2025-06-03. 2023. URL: https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/sustainable-finance_en.

[3] Hugging Face. *Hugging Face Open Embedding Leaderboard*. Accessed on June 2025. 2025. URL: https://huggingface.co/spaces/mteb/leaderboard.

[4] Hugging Face. *Hugging Face Open LLM Leaderboard*. Accessed on June 2025. 2025. URL: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/?params=0%2C4&official=true.

[5] Aidan Hogan et al. «Knowledge graphs». In: *ACM Computing Surveys (CSUR)* 54.4 (2021), pp. 1–37.

[6] Patrick Lewis et al. «Retrieval-augmented generation for knowledge-intensive NLP tasks». In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474.

[7] Microsoft. *Phi-2: Hugging Face Model Card*. Accessed on June 2025. 2023. URL: https://huggingface.co/microsoft/phi-2.

[8] Microsoft Research. *Phi-2: A small language model with high performance*. Accessed: 2025-06-03. 2024. URL: https://www.microsoft.com/en-us/research/blog/phi-2/.

[9] Omotolani Osemwegie. «Building an Advanced RAG Chatbot with Knowledge Graphs Using LlamaIndex, Neo4j and Llama 3». In: (2024). Accessed on June 2025. URL: https://medium.com/@omotolaniosems/building-an-advanced-rag-chatbot-with-knowledge-graphs-using-llamaindex-neo4j-and-llama-3-1e3d3b07ede3.

[10] Nils Reimers e Iryna Gurevych. ≪Sentence-BERT: Sentence embeddings using Siamese BERT-networks≫. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019, pp. 3982–3992.

[11] The World Bank. *ESG Data Portal*. Accessed: 2025-06-03. 2024. URL: https://esgdata.worldbank.org/.

[12] Ajay Verma. ≪The Art and Science of RAG: Mastering Prompt Templates and Contextual Understanding≫. In: (2024). Accessed on June 2025. URL: https://medium.com/@ajayverma23/the-art-and-science-of-rag-mastering-prompt-templates-and-contextual-understanding-a47961a57e27.

[13] Jason Wei et al. ≪Chain of thought prompting elicits reasoning in large language models≫. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 24824–24837.