

RAG SYSTEM EVALUATION

**Benchmarking di Modelli, Hardware e Parametri per
Retrieval-Augmented Generation**

01

Contesto e
Obiettivo

02

Fondamenti
Teorici

03

Architettura del
Sistema

PANORAMICA

04

Metodologia
Sperimentale

05

Esperimenti e
Analisi

06

Conclusioni e
Sviluppi Futuri

Contesto e Obiettivo



Scenario e Contesto

Crescente interesse per i sistemi di domanda e risposta automatica (Q&A) basati su documenti, sia in ambito accademico che industriale.

Necessità di rispondere a domande in linguaggio naturale sfruttando grandi basi documentali.

L'approccio Retrieval-Augmented Generation (RAG) combina la generazione del linguaggio naturale con il recupero mirato delle informazioni.

01

Obiettivo

Sviluppare un sistema di Question Answering (Q&A) su documenti tramite una pipeline RAG
(Retrieval-Augmented Generation)

Rispondere a domande in linguaggio naturale sfruttando il contenuto dei documenti forniti

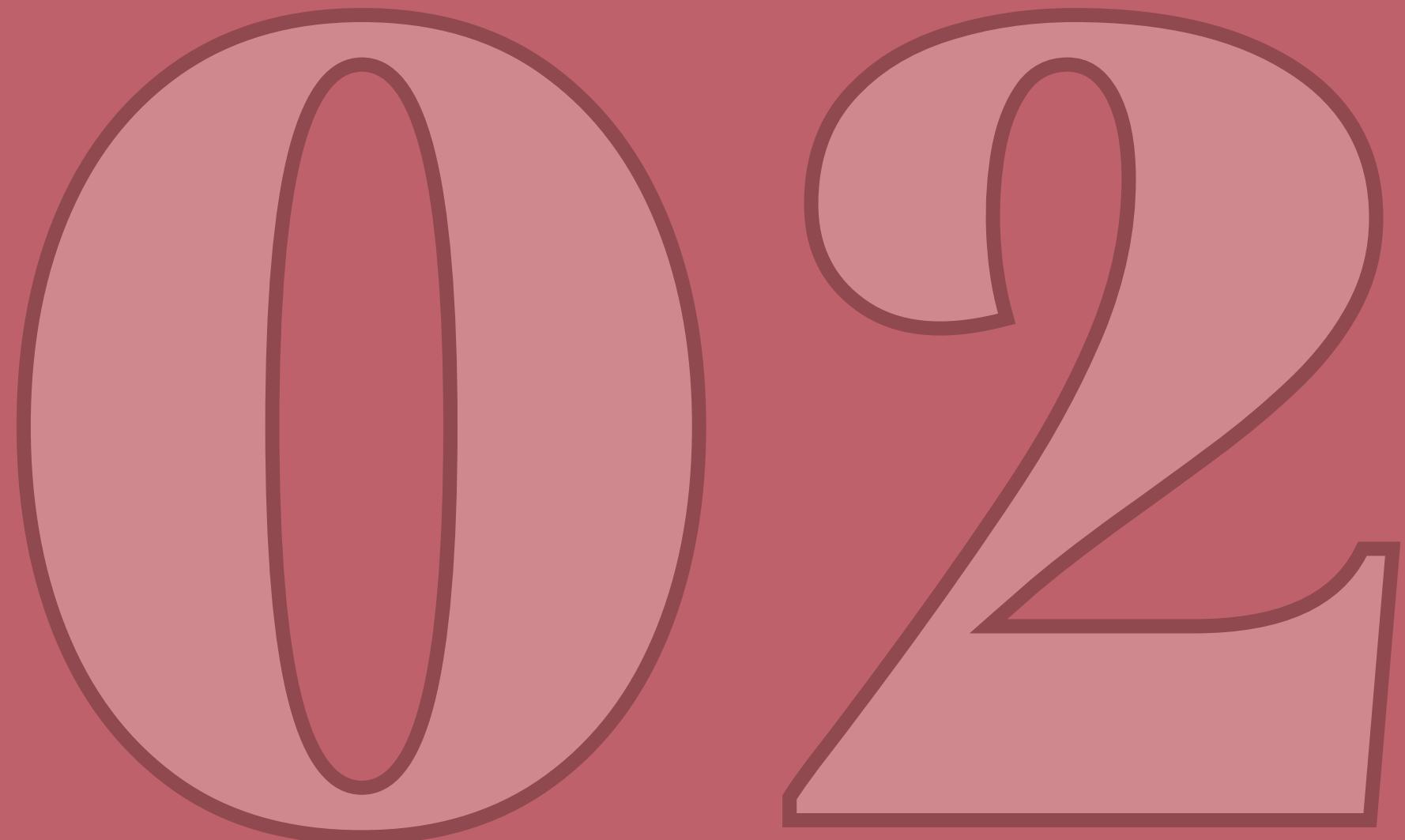
Costruire un benchmark di valutazione per misurare le prestazioni del sistema.

Fase di recupero dei documenti rilevanti (retrieval)

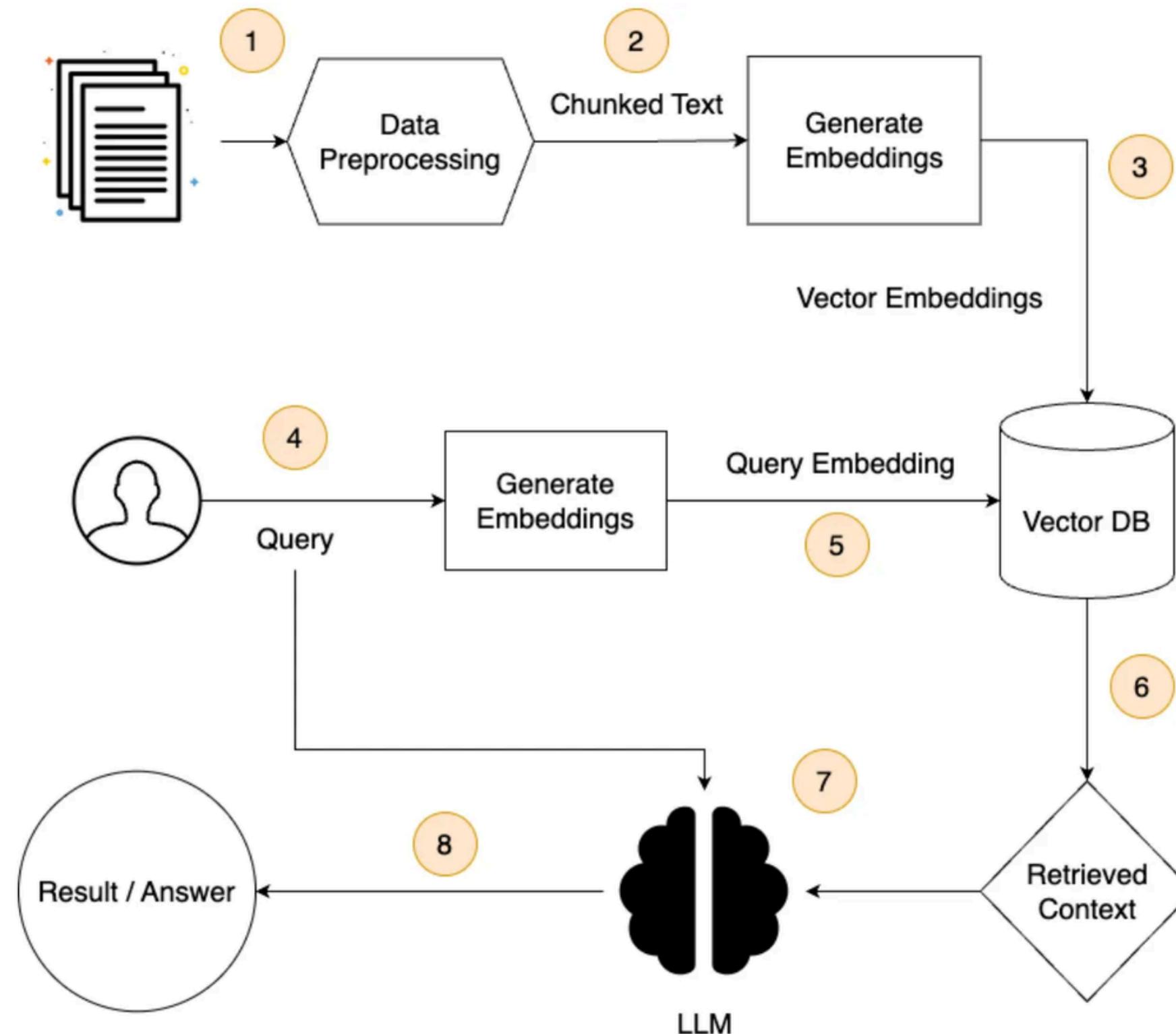
Fase di generazione delle risposte (generation)

01

Fondamenti Teorici



Pipeline RAG



02

Valutazione RAG

Due livelli di valutazione:

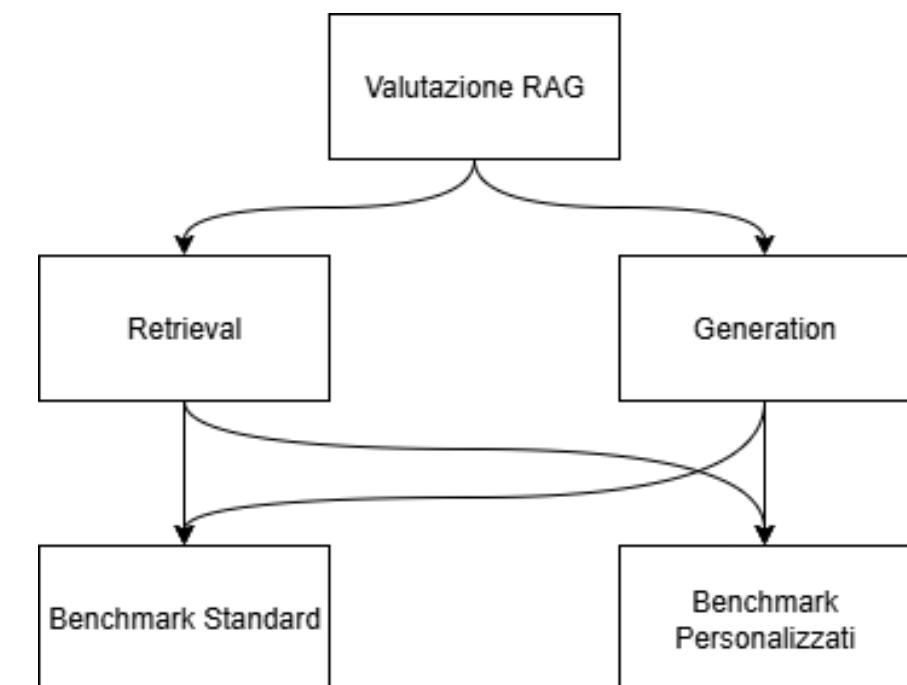
- Retrieval: capacità di recuperare testi utili alla risposta
- Generation: qualità e aderenza semantica della risposta generata

Benchmark standard:

- Raccolte predefinite di dataset per valutazioni uniformi
- Permettono confronto con lo stato dell'arte

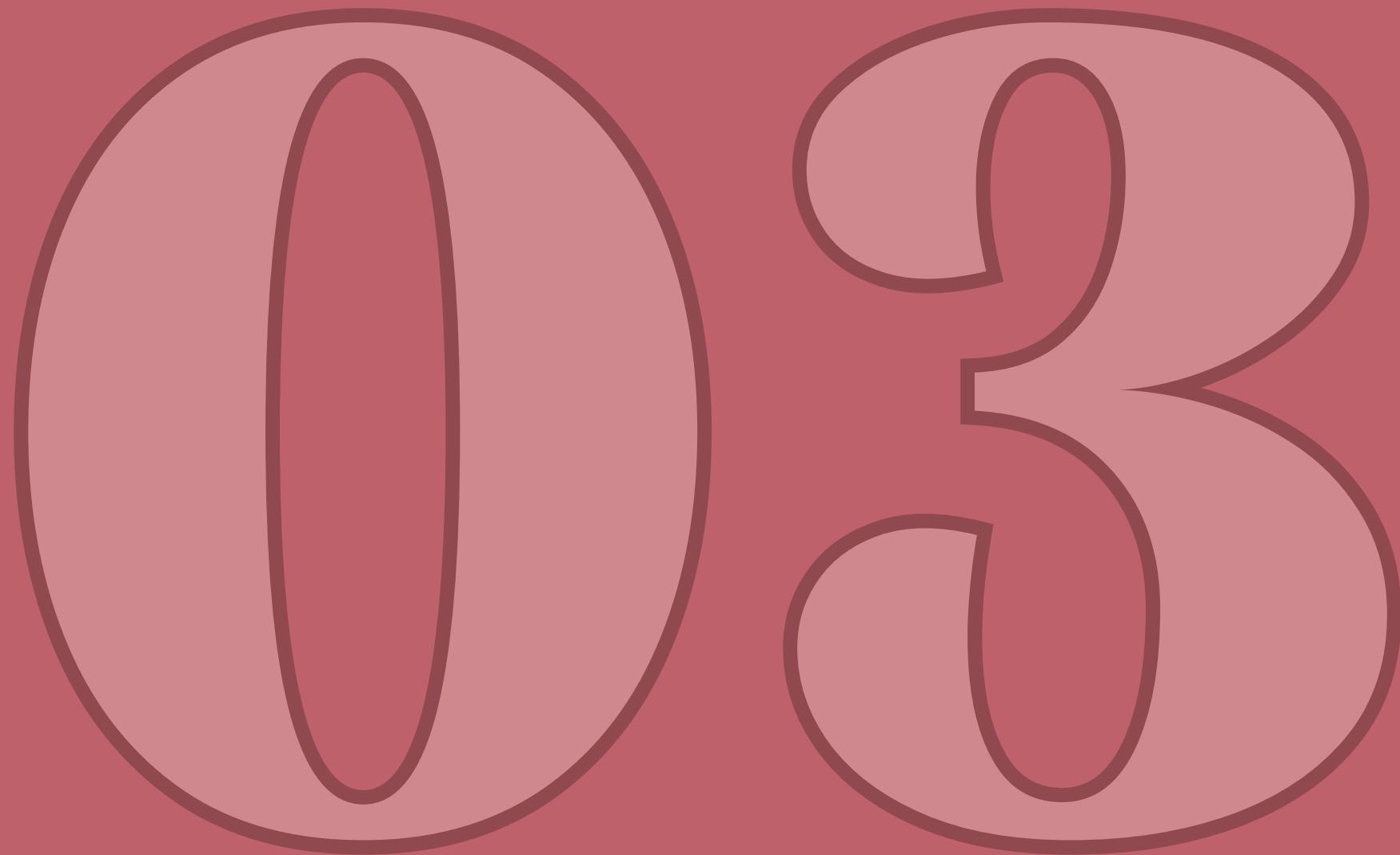
Benchmark personalizzati:

- Creati ad hoc per contesti specifici
- Simulano esigenze e dati reali del dominio applicativo



02

Architettura del Sistema



Due Pipeline Complementari

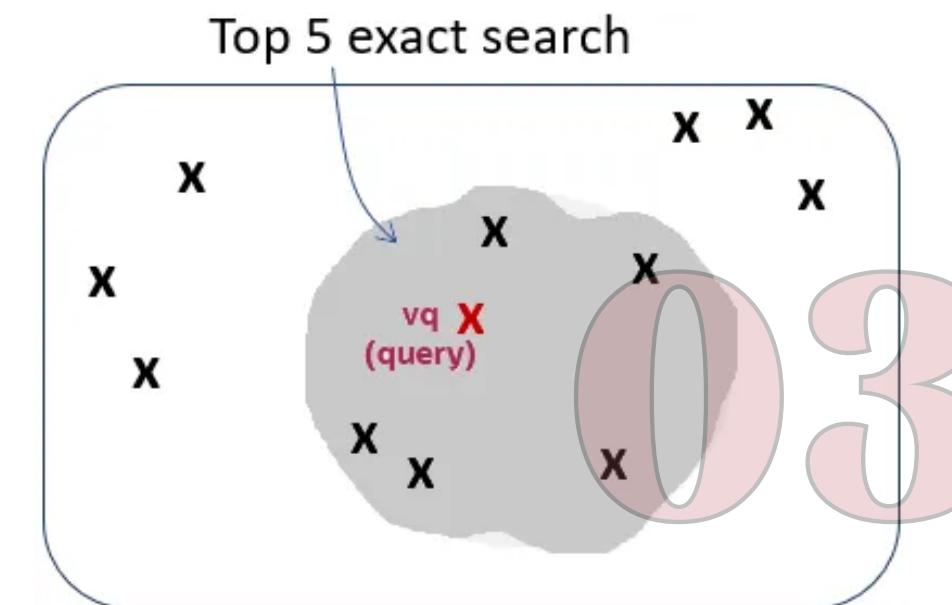
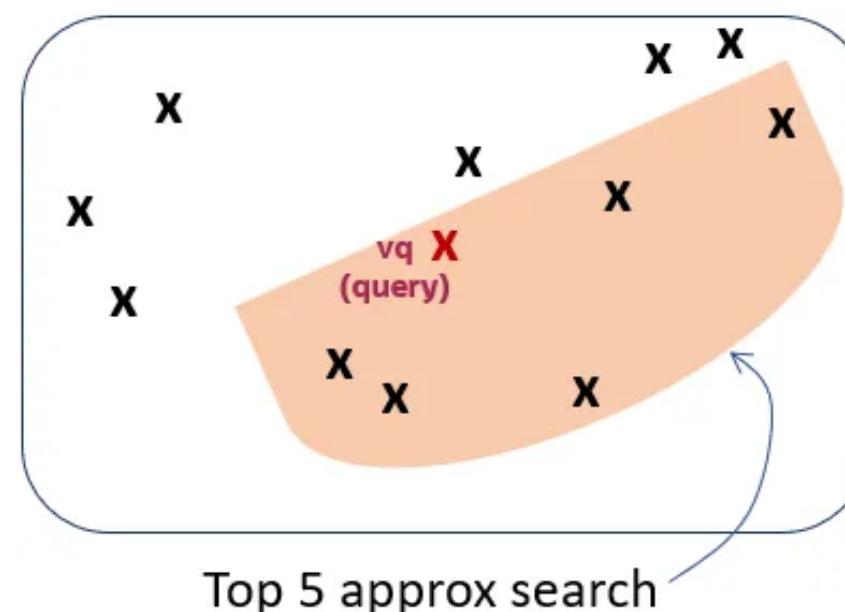
Pipeline Custom: flessibilità, controllo granulare, prompt engineering

Pipeline BEIR: framework standardizzato, confronto con stato dell'arte

Paradigmi diversi: similarity search vs exact search

```
docs_faiss = vectorstore.similarity_search(query, k=config['top_k'])
```

```
retriever = EvaluateRetrieval(dres(sbert, batch_size=config['batch_size']), score_function=config['score_function'])
```



Configurazione

File JSON completo per controllo granulare di tutti i parametri

```
{  
    "dataset_split": "train[100:1000]",  
    "output_folder": "output/",  
    "run_name": "sample",  
    "embedding_model_name": "sentence-transformers/all-MiniLM-L6-v2",  
    "chunk_size": 500,  
    "chunk_overlap": 100,  
    "top_k": 5,  
    "k_values": [1, 3, 5, 10],  
    "num_valid_examples": 10,  
    "llm_model": "HuggingFaceTB/SmollM3-3B",  
    "temperature": 1.0,  
    "max_new_tokens": 1024,  
    "prompt_type": "basic",  
    "only_retrieve": false  
}
```

03

Preprocessing

Dataset Natural Questions (Google)

Selezione di 10 esempi validi per limiti computazionali

Rimozione markup HTML dai documenti Natural Questions

Suddivisione in chunk sovrapposti con parametri configurabili

Normalizzazione del testo per ottimizzare l'embedding

RecursiveCharacterTextSplitter con chunk_size e overlap



```
splitter = RecursiveCharacterTextSplitter(  
    chunk_size=config["chunk_size"],  
    chunk_overlap=config["chunk_overlap"]  
)
```

03

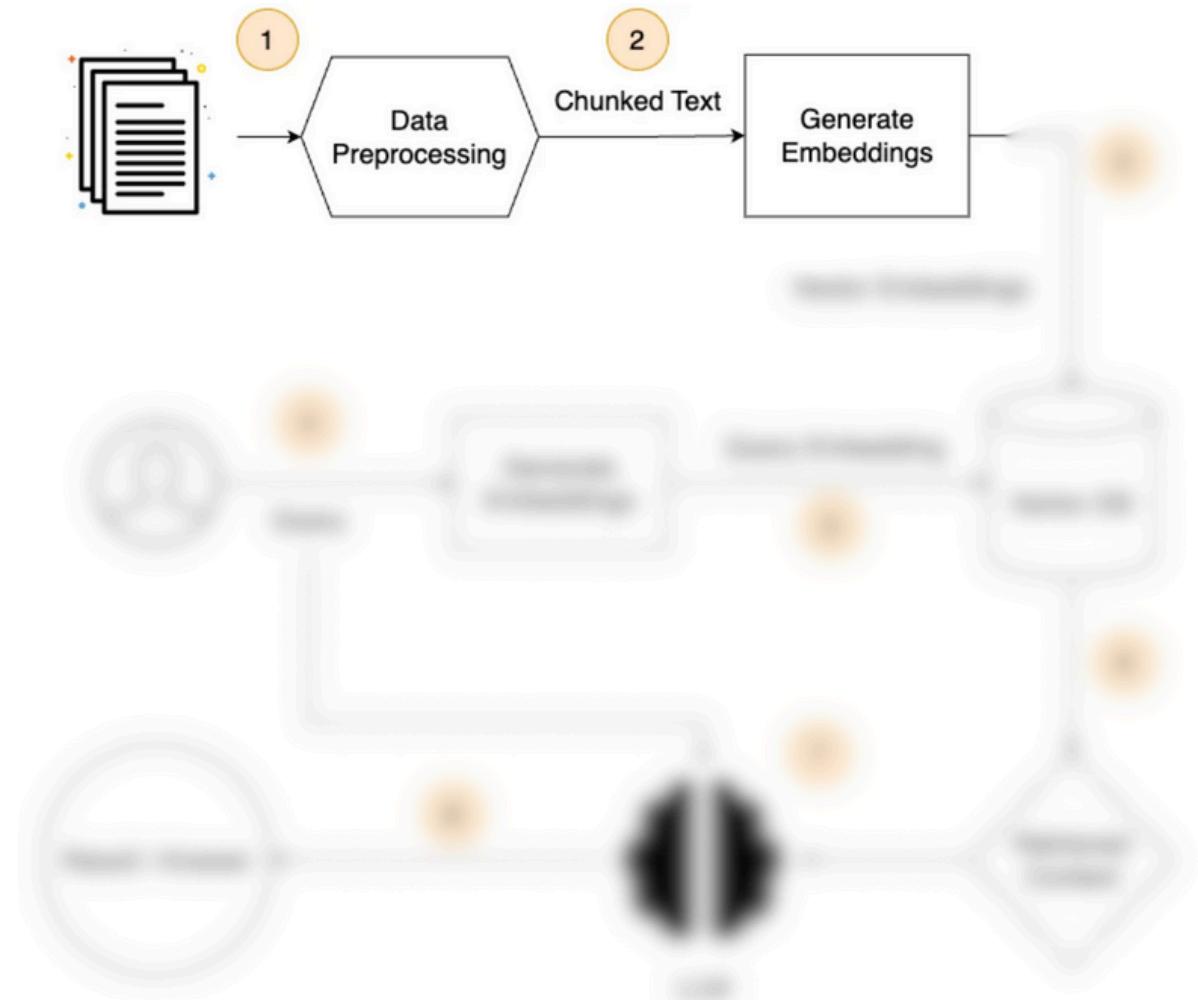
Conversione in Embedding

Modello SentenceTransformers per conversione vettoriale

Embedding configurabile (bge-base, e5-base, MiniLM)

Rappresentazione semantica di ogni chunk di testo

HuggingFaceEmbeddings con modello pre-addestrato



```
embedding_model = HuggingFaceEmbeddings(  
    model_name=config["embedding_model_name"]  
)
```

03

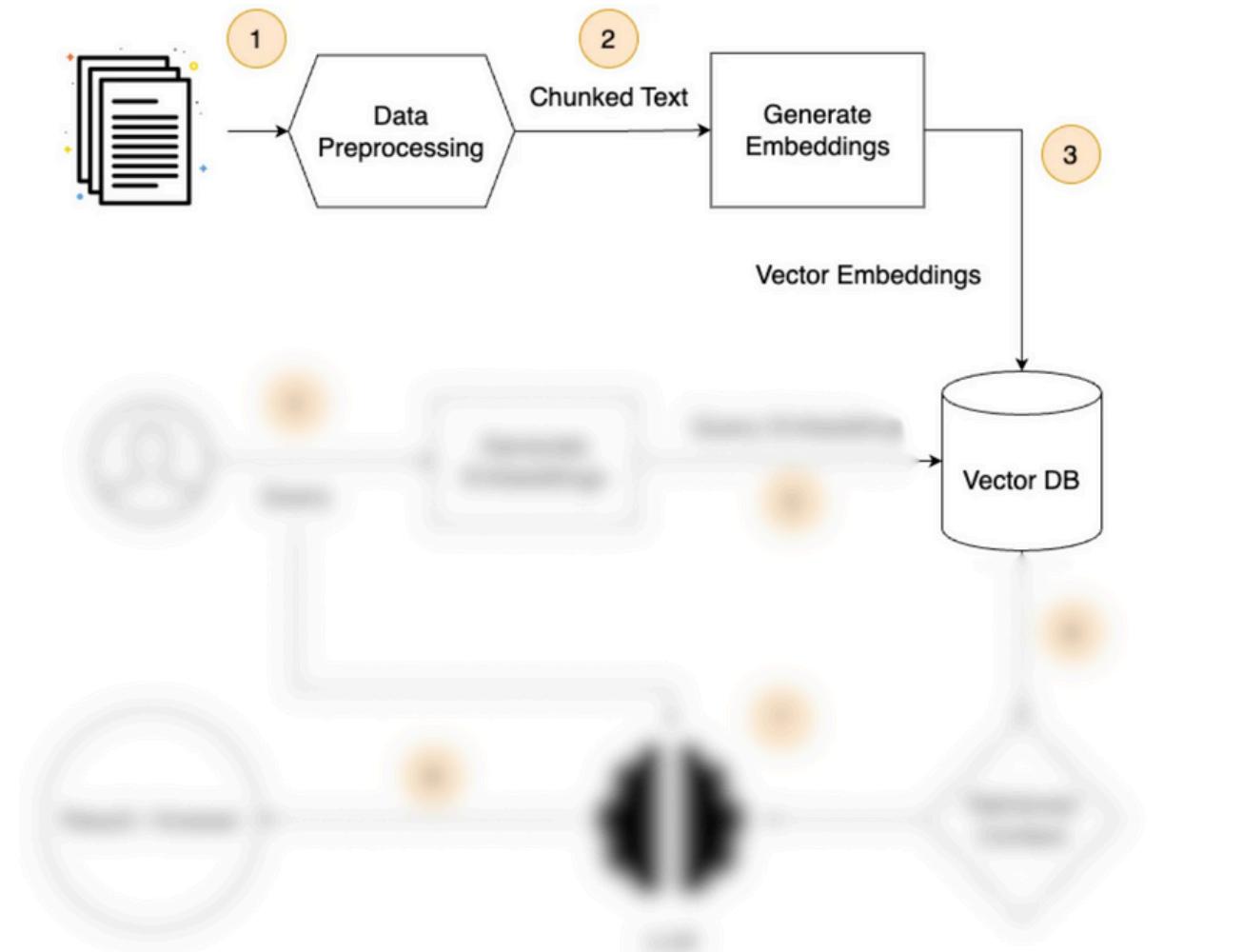
Indicizzazione nel Database Vettoriale

FAISS per storage e ricerca efficiente

Indicizzazione automatica di tutti gli embedding

Ottimizzazione per similarity search veloce

Database vettoriale per ricerca semantica



```
docs_metadata = [doc_id : { doc_text } for
vectorstore = FAISS.from_documents(split_docs, embedding_model)

docs_faiss = vectorstore.similarity_search(query,
k=config['top_k'])
```

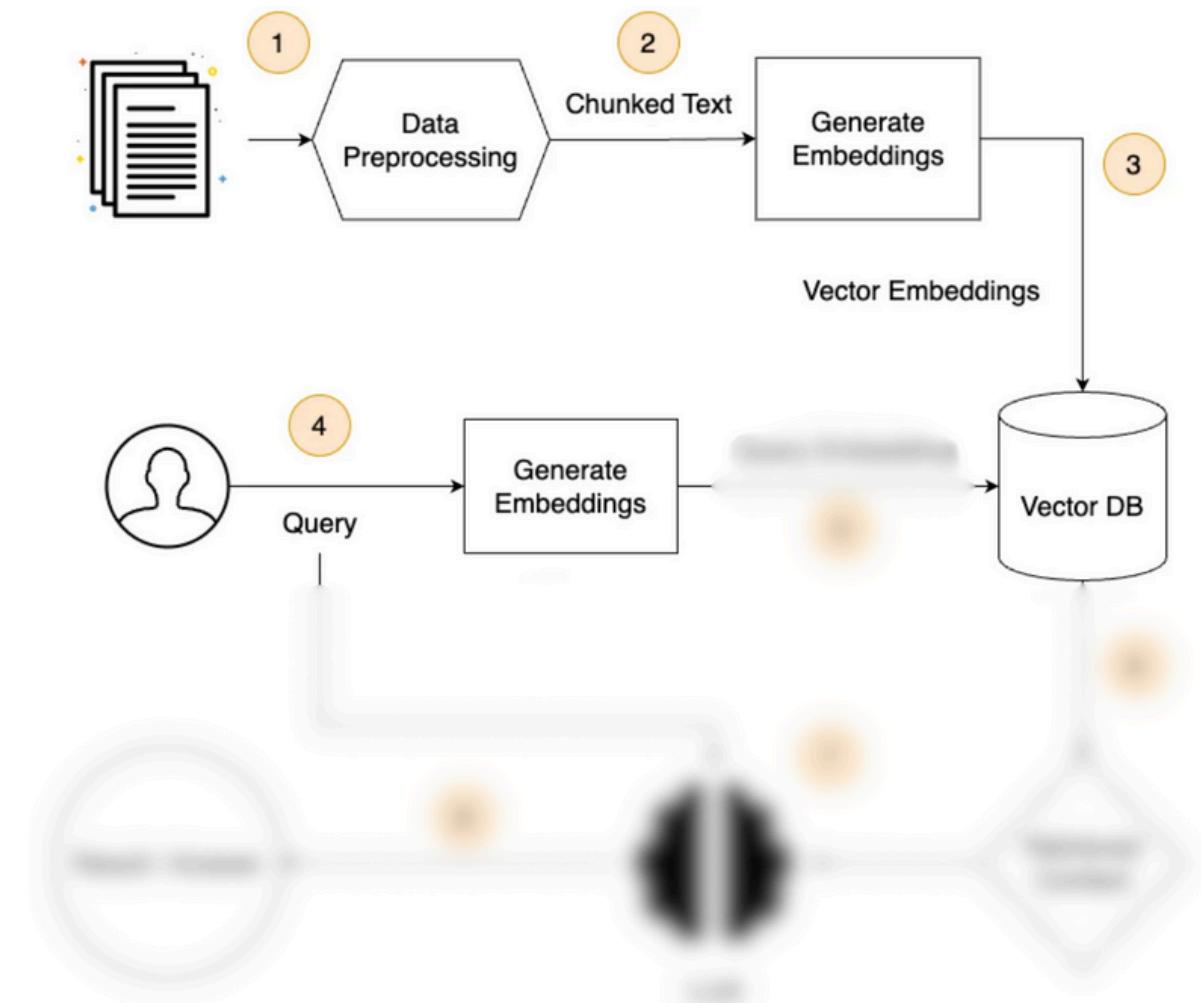
03

Conversione Query in Embedding

Stesso modello di embedding per coerenza semantica

Trasformazione della query in spazio vettoriale

Preparazione per similarity search



```
SentenceTransformer(config["embedding_model_name"])
```

03

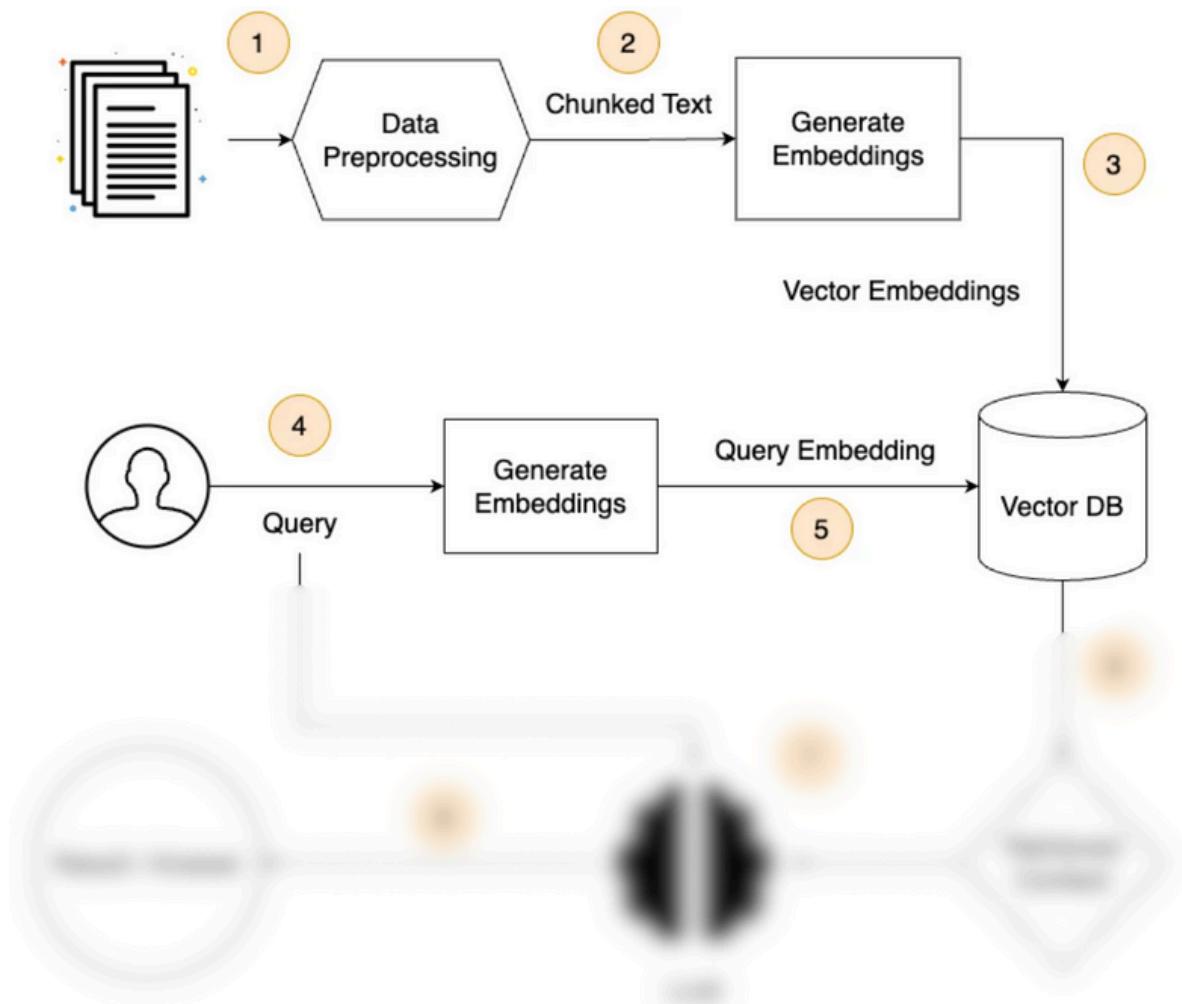
Ricerca Semantica

Similarity search tra query e documenti

Top-k configurabile

Recupero chunk più rilevanti per il contesto

FAISS similarity_search() con parametro k



```
def retrieve_local(query, vectorstore, k=config["top_k"]):  
    docs_faiss = vectorstore.similarity_search(query, k=k)  
    return [d.page_content for d in docs_faiss]
```

03

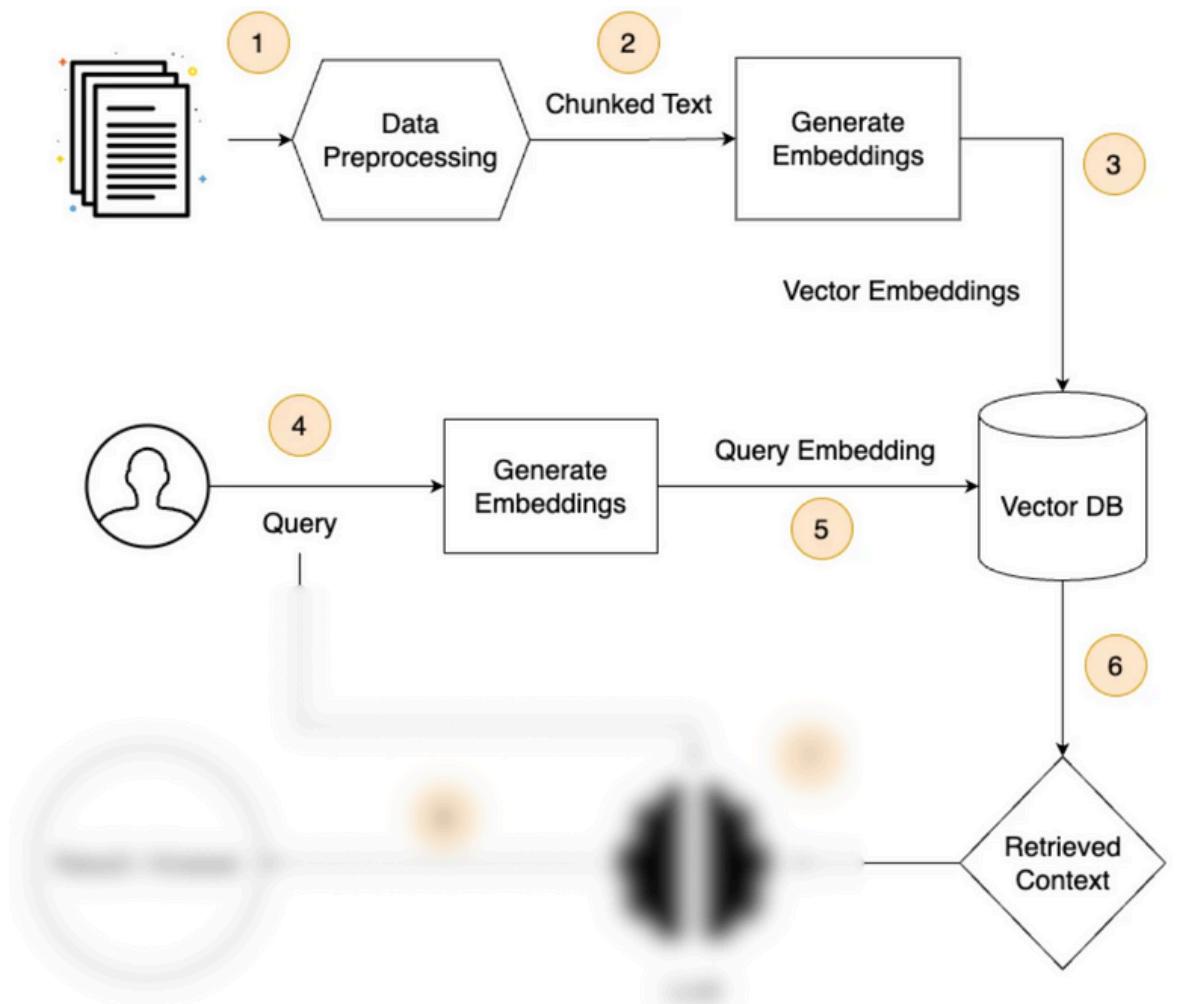
Recupero del Contesto

Aggregazione chunk recuperati in contesto unificato

Selezione automatica dei frammenti più simili

Preparazione per generazione

Concatenazione intelligente dei documenti



```
context = retrieve_local(query, vectorstore)  
prediction = ask(query, context)
```

03

Generazione della Risposta

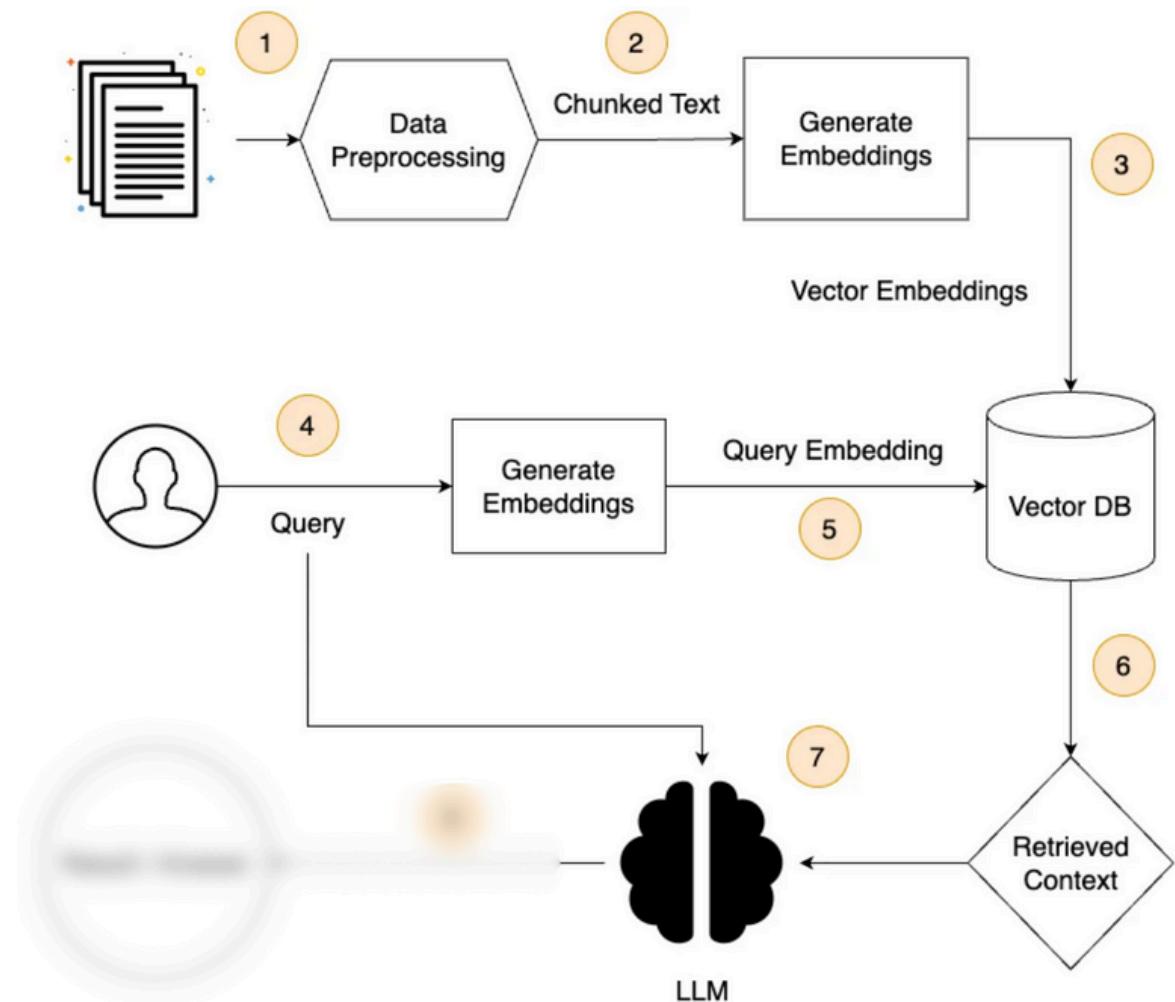
LLM HuggingFace (SmolLM3-3B) per generazione

7 tipi di prompt (basic, chain-of-thought, structured, etc.)

Contesto + query come input per risposta informata

Prompt engineering configurabile

```
context = retrieve_local(query, vectorstore)  
prediction = ask(query, context)
```



03

Valutazione Completa

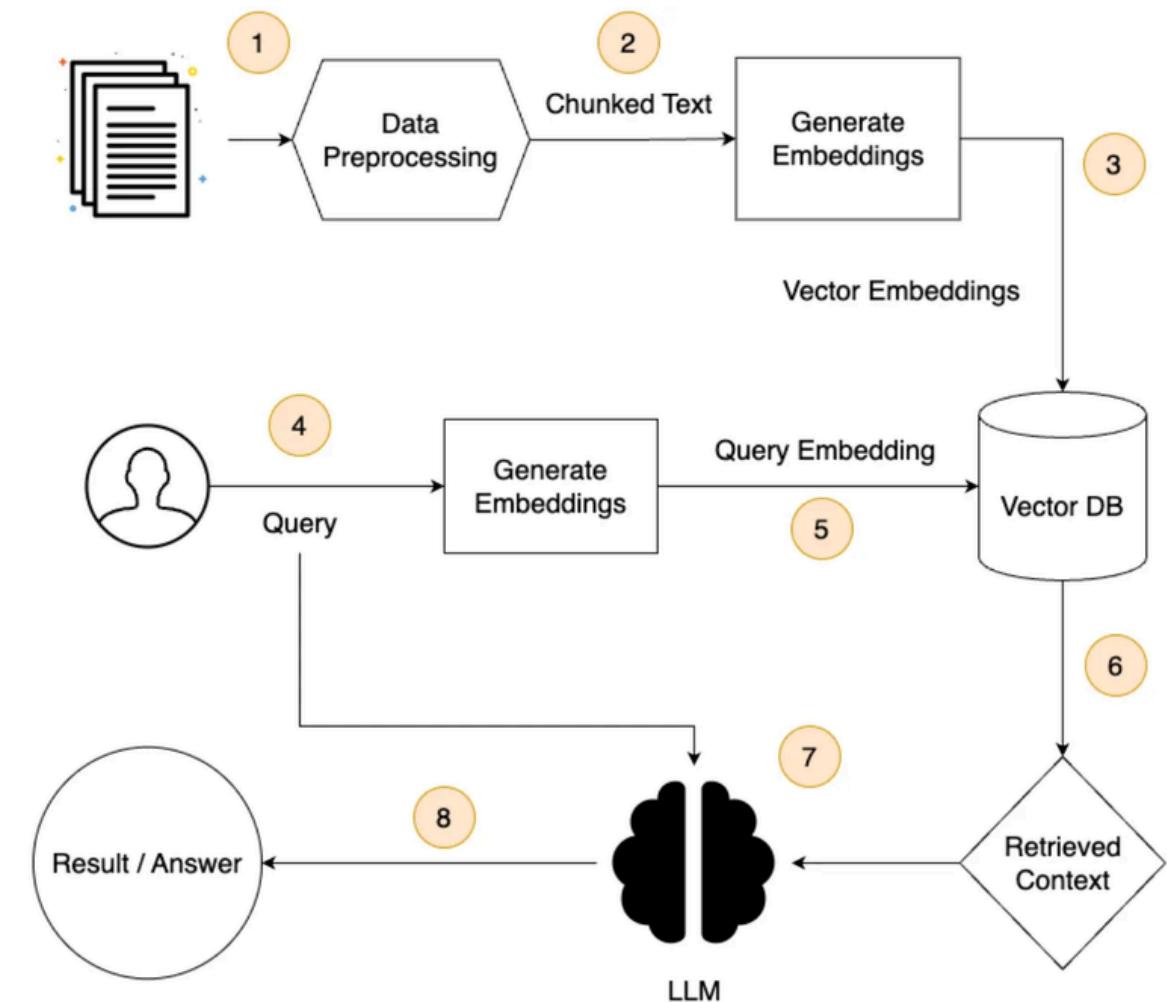
Metriche retrieval: Precision@k, Recall@k, MAP, nDCG

Metriche generazione: ROUGE-L, BLEU, Cosine Similarity

Valutazione semantica automatica condotta dallo stesso
LLM

Analisi completa di retrieval e generazione

```
else:  
    result = {  
        "index": i,  
        "query": query,  
        "gold_answer": golden_answer,  
        "golden_context": golden_context,  
        "relevant_ids": relevant_ids,  
        "retrieved_ids": retrieved_ids,  
        "precision_dict": precision_dict,  
        "recall_dict": recall_dict,  
        "map_dict": map_dict,  
        "ndcg_dict": ndcg_dict  
    }  
return result
```



03

Pipeline BEIR

Framework standardizzato per Information Retrieval

Sentence-BERT per embedding densi

Dataset eterogenei (corpus + queries + qrels)

Fasi: caricamento → embedding → retrieval → valutazione

Metriche standardizzate (NDCG, MAP, Recall, Precision)



Beir
Benchmarking IR

03

Metodologia Sperimentale



Metodologia: Analisi Univariata

Variazione univariata: un parametro alla volta

Mantenimento costanti tutti gli altri parametri

Modalità Analisi: retrieval-only vs retrieval+generazione

Tracciabilità: configurazioni JSON + output strutturati

Riproducibilità: notebook Python + file di output

04

Valutazione Retrieve

NDCG (Normalized Discounted Cumulative Gain): Valuta la qualità del ranking, premiando i documenti rilevanti nelle prime posizioni.

MAP (Mean Average Precision): Precisione media calcolata su tutte le posizioni e query.

Recall: Percentuale di documenti rilevanti effettivamente recuperati.

Precision: Percentuale di risultati restituiti che sono effettivamente rilevanti.

04

Valutazione Sintattica Generazione

ROUGE-L: Misura l'overlap in termini di sottosequenze più lunghe tra risposta generata e gold.

BLEU: Valuta la precisione n-gram, confrontando sequenze di parole tra generato e gold.

04

Valutazione Semantica Retrieve

Cosine Similarity: Similarità tra gli embedding della risposta generata e della risposta gold.

Valutazione semantica LLM: Il modello rilegge la risposta generata e fornisce un giudizio automatico di correttezza.

04

Esperimenti e Analisi

05

Esperimenti Pipeline Custom

Parametro	Valori Testati	Obiettivo
Modello di embedding	all-MiniLM-L6-v2, bge-base-en-v1.5, e5-base-v2, ecc.	Valutare l'impatto del modello di embedding sulla fase di retrieval
Chunk size e overlap	Chunk: 100, 200, 500, 1000, 2500; Overlap: 1/10, 1/5, 1/2, >1/2	Analizzare l'effetto della segmentazione del testo sulla qualità del contesto
Top-k	1, 3, 5, 10	Studiare la quantità ottimale di contesto da fornire al generatore
Temperature	0.0, 0.2, 0.5, 0.7, 1.0, 1.2, 2.0	Analizzare l'effetto sulla creatività delle risposte
Max New tokens	Max tokens: 32, 64, 128, 256, 512	Analizzare l'effetto sulla creatività e coerenza delle risposte
Tipo di prompt	basic, chain-of-thought, few-shot, structured, constrained, role, self-verify	Valutare l'impatto del prompt engineering sulla robustezza e precisione delle risposte

05

Esperimento: Modello di Embedding

Fase: Retrieval

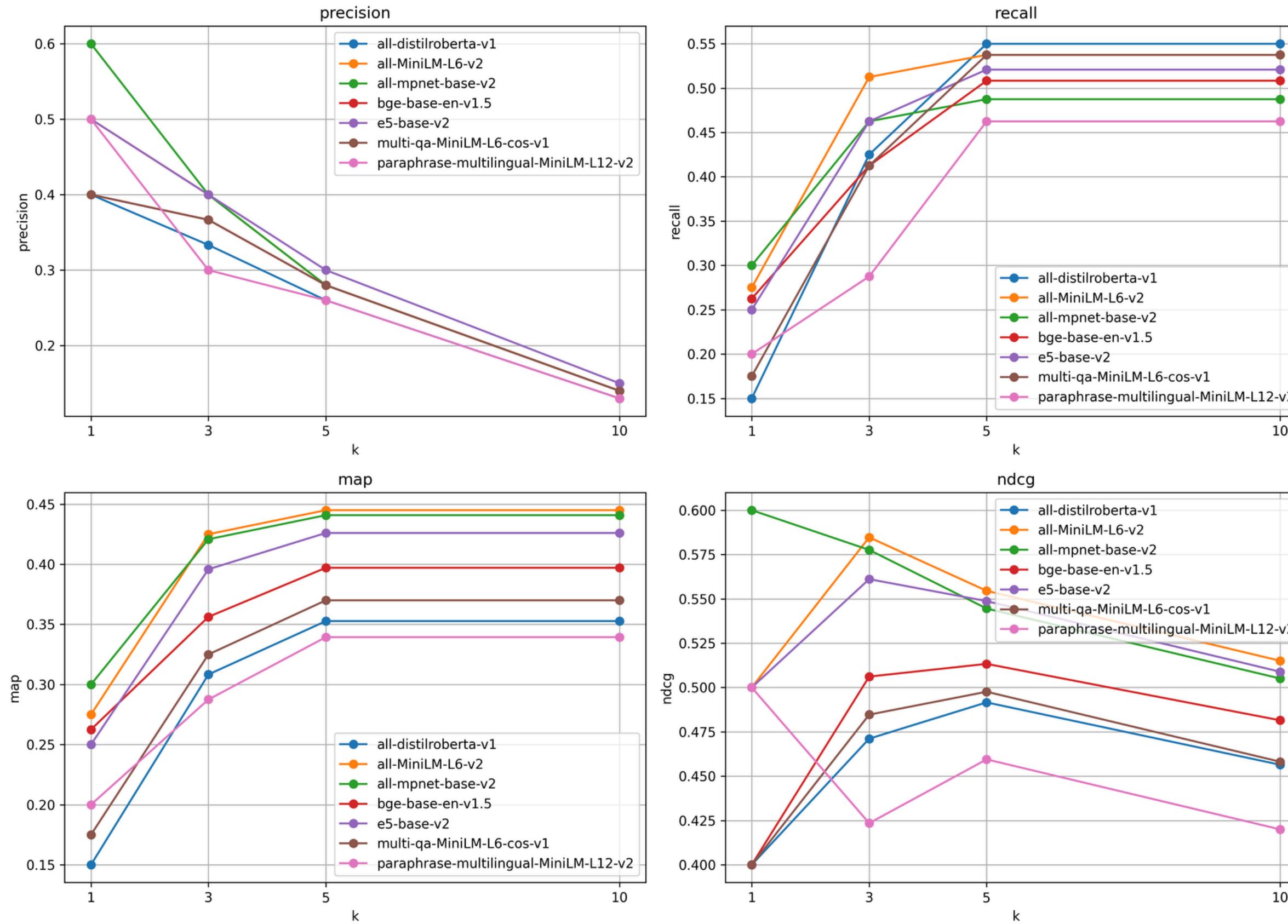
Modelli testati

- all-MiniLM-L6-v2: ottimizzato per semantic search
- bge-base-en-v1.5: migliore generalizzazione
- e5-base-v2: compiti di retrieval
- all-distilroberta-v1: versione leggera di RoBERTa
- all-mpnet-base-v2: dimensioni maggiori
- multi-qa-MiniLM-L6-cos-v1: ottimizzato per Q&A
- paraphrase-multilingual-MiniLM-L12-v2: modello multilingue

Obiettivo: valutare l'impatto del modello di embedding sulla fase di retrieval

05

Analisi: Modello di Embedding



Metriche di Retrieve al variare del Modello di Embedding

Analisi: Modello di Embedding

The image shows a code editor with three tabs open, each displaying a JSON document. The tabs are labeled "retrieved_docs.json all-MiniLM-L6-v2", "retrieved_docs.json paraphrase-multilingual-MiniLM-L12-v2", and "retrieved_docs.json all-mpnet-base-v2". Each tab contains a JSON object with multiple entries, each representing a retrieved document. The JSON structure includes fields like "index", "relevant_ids", and "retrieved_ids". The code editor highlights specific values in red, such as "doc_13" and "doc_18", which are common across all three models.

```
43 {  
48     "retrieved_ids": [  
49         "doc_40"  
50     ]  
51 },  
52 {  
53     "index": 18,  
54     "relevant_ids": [  
55         "doc_0",  
56         "doc_1",  
57         "doc_3",  
58         "doc_18"  
59     ],  
60     "retrieved_ids": [  
61         "doc_2",  
62         "doc_0",  
63         "doc_1",  
64         "doc_28",  
65         "doc_23"  
66     ],  
67 },  
68 {  
69     "index": 22,  
70     "relevant_ids": [  
71         "doc_0"  
72     ],  
73     "retrieved_ids": [  
74         "doc_0",  
75         "doc_6",  
76         "doc_7"  
77     ]  
78 },  
79 }  
80  
43 {  
48     "retrieved_ids": [  
49         "doc_13"  
50     ]  
51 },  
52 {  
53     "index": 18,  
54     "relevant_ids": [  
55         "doc_0",  
56         "doc_1",  
57         "doc_3",  
58         "doc_18"  
59     ],  
60     "retrieved_ids": [  
61         "doc_0",  
62         "doc_2",  
63         "doc_27",  
64         "doc_30",  
65         "doc_23"  
66     ],  
67 },  
68 {  
69     "index": 22,  
70     "relevant_ids": [  
71         "doc_0"  
72     ],  
73     "retrieved_ids": [  
74         "doc_0",  
75         "doc_0"  
76     ]  
77 },  
78     "retrieved_ids": [  
79         "doc_0",  
80         "doc_6",  
81     ]  
82 }  
83  
43 {  
48     "retrieved_ids": [  
49     ]  
50 },  
51 {  
52     "index": 18,  
53     "relevant_ids": [  
54         "doc_0",  
55         "doc_1",  
56         "doc_3",  
57         "doc_18"  
58     ],  
59     "retrieved_ids": [  
60         "doc_1",  
61         "doc_2",  
62         "doc_0",  
63         "doc_23",  
64         "doc_25"  
65     ],  
66 },  
67 {  
68     "index": 22,  
69     "relevant_ids": [  
70         "doc_0"  
71     ],  
72     "retrieved_ids": [  
73         "doc_0",  
74         "doc_0",  
75         "doc_0",  
76         "doc_1",  
77         "doc_6"  
78     ]  
79 }  
80
```

Documenti Recuperati con all-MiniLM-L6-v2 (sinistra), paraphrase-multilingual-MiniLM-L12 (centro) e all-mpnet-base-v2 (destra)

Esperimento: Chunk Overlap

Fase: Retrieval (+ Generation)

Valori testati:

- 1/10
- 1/5
- 1/2
- >1/2

Obiettivo: analizzare l'effetto della sovrapposizione nei chunk sulla qualità del contesto

05

Analisi: Chunk Overlap

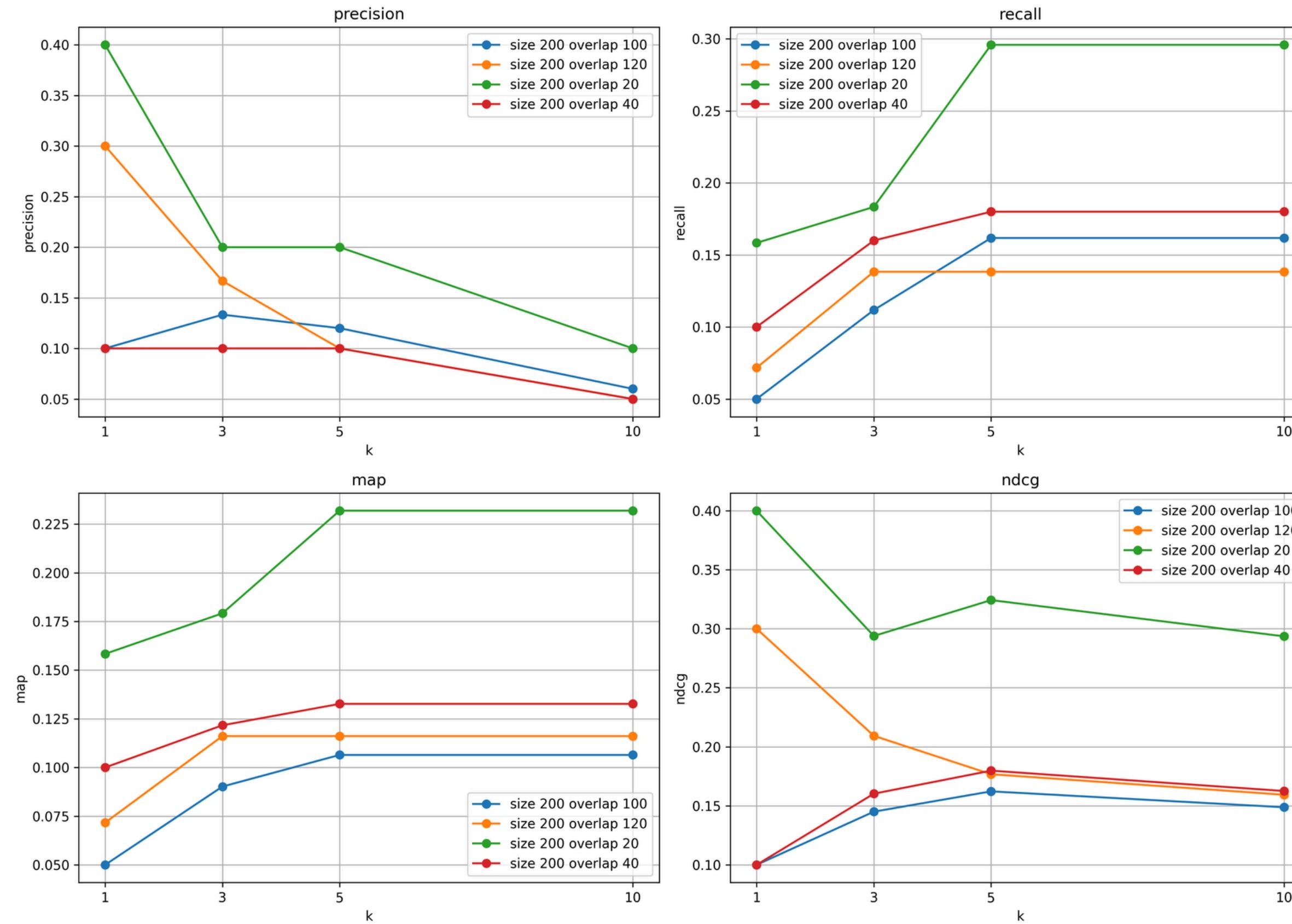
The image displays four separate code snippets, each representing a JSON document structure. The snippets are arranged horizontally and show the results of a retrieval process with different parameters:

- Leftmost Snippet:** titled "retrieved_docs.json size_200_overlap_20_2025_07_11_093 ...". It shows a list of documents indexed from 11 to 16. Each document entry includes "relevant_ids" (e.g., ["doc_3", "doc_4", "doc_5"] for index 11) and "retrieved_ids" (e.g., ["doc_3", "doc_0", "doc_39", "doc_4", "doc_5"] for index 11).
- Second Snippet:** titled "retrieved_docs.json size_200_overlap_40_2025_07_11_09 ...". It shows documents indexed from 11 to 16. The "relevant_ids" and "retrieved_ids" lists are identical to the first snippet.
- Third Snippet:** titled "retrieved_docs.json size_200_overlap_100_2025_07_11_09 ...". It shows documents indexed from 11 to 16. The "relevant_ids" and "retrieved_ids" lists are identical to the first snippet.
- Rightmost Snippet:** titled "retrieved_docs.json size_200_overlap_120_2025_07_11_09 ...". It shows documents indexed from 11 to 16. The "relevant_ids" and "retrieved_ids" lists are identical to the first snippet.

In all four cases, the "relevant_ids" and "retrieved_ids" lists are identical, indicating no overlap between chunks. The "index" field is present in every document entry, and the "retrieved_ids" field is present in every document entry except the last one (index 16).

Documenti Recuperati con Chunk Size a 200 e Overlap 20, 40, 100, 120

Analisi: Chunk Overlap



Metriche di Retrieve con Chunk Size a 200 e Overlap 20 40 100 e 120

Esperimento: Chunk Size

Fase: Retrieval (+ Generation)

Valori testati:

- 100
- 200
- 500
- 1000
- 2500

Obiettivo: analizzare l'effetto della segmentazione del testo sulla qualità del contesto

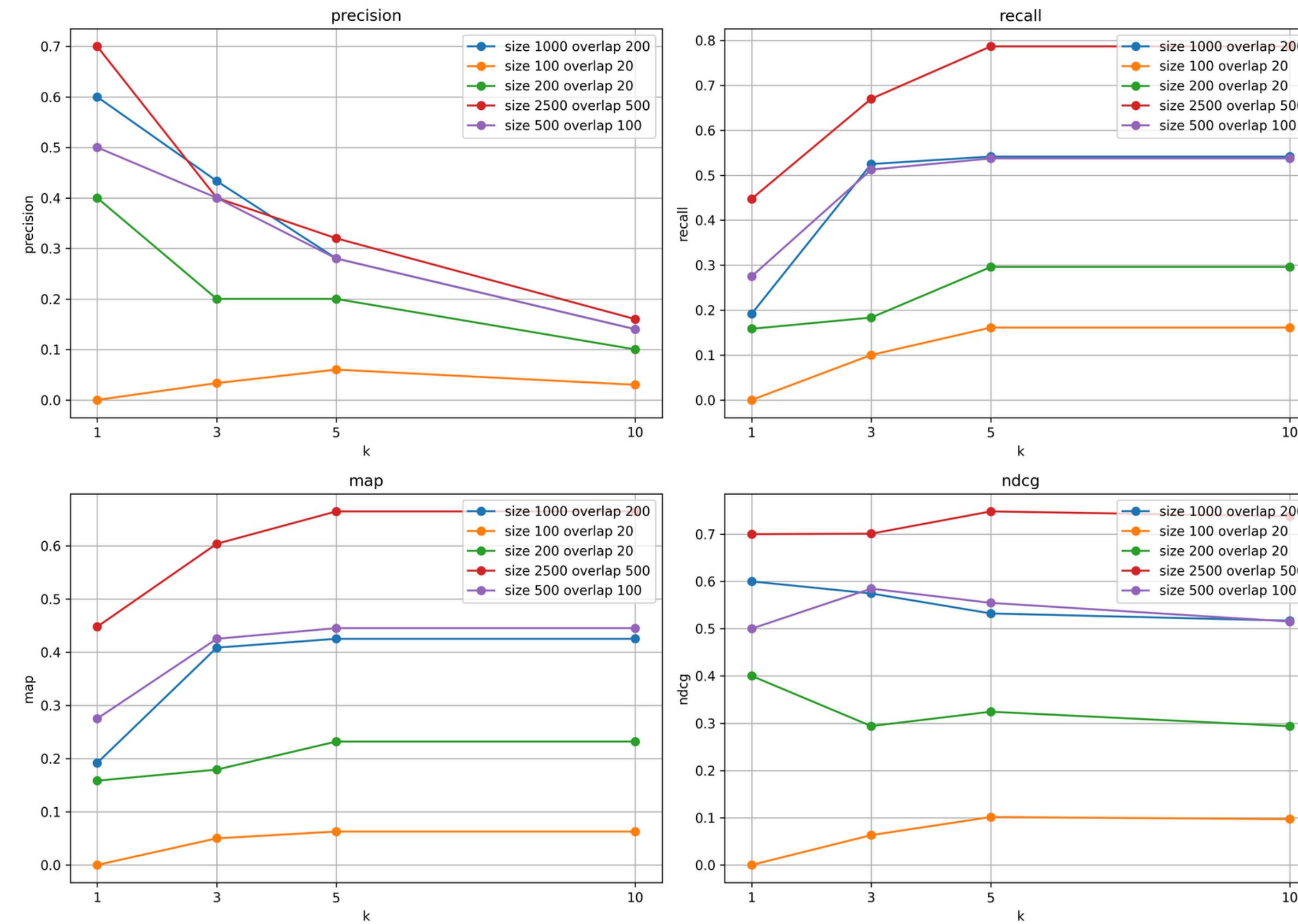
05

Analisi: Chunk Size

{ retrieved_docs.json size_100_overlap_20_2025_07_11_092... }	{ retrieved_docs.json size_200_overlap_20_2025_07_11_092... }	{ retrieved_docs.json size_500_overlap_100_2025_07_11_092... }	{ retrieved_docs.json size_1000_overlap_200_2025_07_11_092... }	{ retrieved_docs.json size_2500_overlap_500_2025_07_11_092... }
1 [2 { 3 "index": 11, 4 "relevant_ids": [5 "doc_9", 6 "doc_13" 7], 8 "retrieved_ids": [9 "doc_0", 10 "doc_11", 11 "doc_7", 12 "doc_89", 13 "doc_10" 14] 15 }, 16 { 17 "index": 13, 18 "relevant_ids": [19 "doc_7" 20], 21 "retrieved_ids": [22 "doc_50", 23 "doc_10", 24 "doc_22", 25 "doc_30", 26 "doc_28" 27] 28 }, 29 { 30 "index": 16, 31 "relevant_ids": [32 "doc_8", 33 "doc_246" 34], 35 "retrieved_ids": [36 "doc_0", 37 "doc_119", 38 "doc_40".	1 [2 { 3 "index": 11, 4 "relevant_ids": [5 "doc_3", 6 "doc_4", 7 "doc_5" 8], 9 "retrieved_ids": [10 "doc_3", 11 "doc_0", 12 "doc_39", 13 "doc_4", 14 "doc_5" 15], 16 { 17 "index": 13, 18 "relevant_ids": [19 "doc_1" 20], 21 "retrieved_ids": [22 "doc_22", 23 "doc_13", 24 "doc_4", 25 "doc_4", 26 "doc_10", 27 "doc_12" 28], 29 { 30 "index": 16, 31 "relevant_ids": [32 "doc_3", 33 "doc_109" 34], 35 "retrieved_ids": [36 "doc_0", 37 "doc_0", 38 "doc_18".	1 [2 { 3 "index": 11, 4 "relevant_ids": [5 "doc_1", 6 "doc_2" 7], 8 "retrieved_ids": [9 "doc_2", 10 "doc_0", 11 "doc_1", 12 "doc_17", 13 "doc_12" 14], 15], 16 { 17 "index": 13, 18 "relevant_ids": [19 "doc_1" 20], 21 "retrieved_ids": [22 "doc_1", 23 "doc_2", 24 "doc_4", 25 "doc_6", 26 "doc_5" 27], 28], 29 { 30 "index": 16, 31 "relevant_ids": [32 "doc_1", 33 "doc_49" 34], 35 "retrieved_ids": [36 "doc_0", 37 "doc_35", 38 "doc_8".	1 [2 { 3 "index": 11, 4 "relevant_ids": [5 "doc_0", 6 "doc_1" 7], 8 "retrieved_ids": [9 "doc_1", 10 "doc_0", 11 "doc_8", 12 "doc_6", 13 "doc_5" 14], 15], 16 { 17 "index": 13, 18 "relevant_ids": [19 "doc_0" 20], 21 "retrieved_ids": [22 "doc_5", 23 "doc_2", 24 "doc_0", 25 "doc_1", 26 "doc_4" 27], 28], 29 { 30 "index": 16, 31 "relevant_ids": [32 "doc_0", 33 "doc_24" 34], 35 "retrieved_ids": [36 "doc_0", 37 "doc_11", 38 "doc_17".	1 [2 { 3 "index": 11, 4 "relevant_ids": [5 "doc_0" 6], 7 "retrieved_ids": [8 "doc_0", 9 "doc_2", 10 "doc_3", 11 "doc_1" 12], 13 { 14 "index": 13, 15 "relevant_ids": [16 "doc_0" 17], 18 "retrieved_ids": [19 "doc_0", 20 "doc_2", 21 "doc_0", 22 "doc_1", 23 "doc_1" 24], 25 { 26 "index": 16, 27 "relevant_ids": [28 "doc_0", 29 "doc_9" 30], 31 "retrieved_ids": [32 "doc_0", 33 "doc_9", 34 "doc_7", 35 "doc_8", 36 "doc_2" 37].

Documenti Recuperati con i migliori Overlap per ciascun Size

Analisi: Chunk Size



Metriche di Retrieve con i migliori Overlap per ciascun Size

Analisi: Chunk Size

The image shows a code editor with two tabs open, comparing JSON responses generated with different chunk sizes.

Left Tab (Chunk Size 100 Overlap 20):

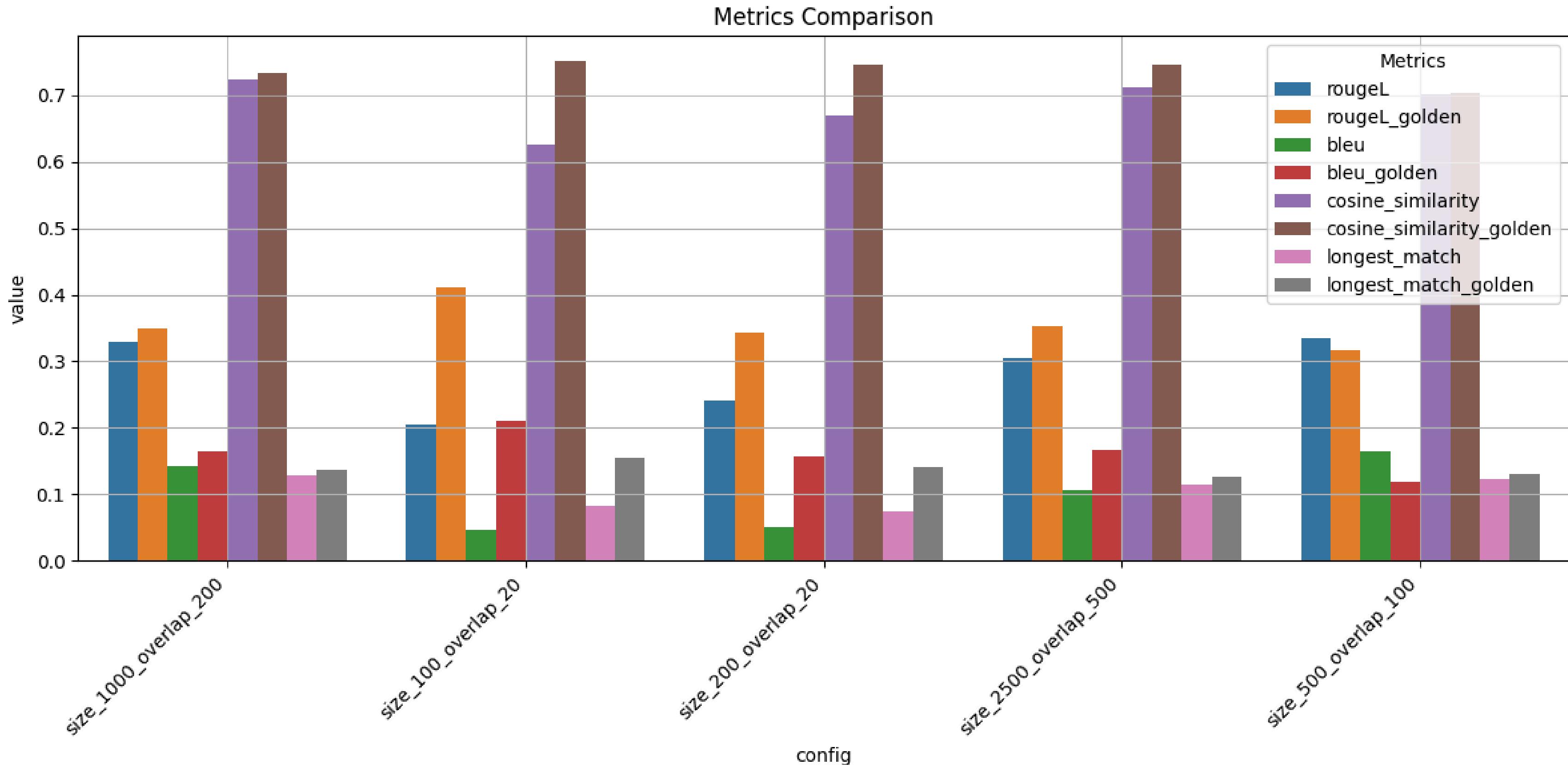
```
1 [ "index": 11, "prediction": "The context contains the information that \"Now That's What I Call Music (original UK album)\" is an album, and it is part of the Virgin / EMI Series chronology which includes \"Now That's What I Call Music (1983)\". However, the exact release date of \"Now That's What I Call Music\" is not explicitly mentioned in the provided context. Therefore, based solely on the given context, I don't know the exact release date.", "prediction_golden": "Now That's What I Call Music (also simply titled Now or Now 1) was released in the United Kingdom on 28 November 1983.", "semantic_evaluation": "The statement is correct. The context explicitly mentions that the album \"Now That's What I Call Music (original UK album)\" was released in 1983 as part of the Virgin / EMI Series, which is confirmed by the mention of the \"1983\" title and the series' chronology.", "semantic_evaluation_golden": "Yes, the phrase is correct. The album was indeed released in the United Kingdom on 28 November 1983." }, { "index": 13, "prediction": "The 2018 Grey Cup is being played in Edmonton.", "prediction_golden": "The 2018 Grey Cup is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.", "semantic_evaluation": "Yes, the phrase is correct. The 2018 Grey Cup is being played in Edmonton, as stated in the original information.", "semantic_evaluation_golden": "The phrase is correct. The 2018 Grey Cup was indeed scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on November 25, 2018. The original statement accurately reflects this event." }, { "index": 16, "prediction": "The band Bastille got their name based uniquely on the context that \"The name of the band derives from Bastille Day , which is celebrated on 14 July , the date of\".", "prediction_golden": "The band Bastille got their name based on Bastille Day, which is celebrated on 14 July, the date of lead vocalist Dan Smith's birthday.", "semantic_evaluation": "Yes, the phrase is correct. The band's name, Bastille, is derived from Bastille Day (14 July), which is celebrated on the lead singer's birthday. The statement accurately reflects the unique context (Bastille Day) that inspired the band's name.", "semantic_evaluation_golden": "Yes, the phrase is correct. The band Bastille was named after Bastille Day (14 July), which coincides with the lead vocalist Dan Smith's birthday." }, { "index": 17, "prediction": "The Arkansas River is located in the United States, flowing through the states of Colorado, Kansas, Oklahoma, and Arkansas. It originates in the Rocky Mountains near Leadville, Colorado, and flows southeastward, eventually emptying into the Mississippi River near Napoleon, Arkansas.", "prediction_golden": "The Arkansas River is located in Colorado, Kansas, Oklahoma, and Arkansas, and its watershed also drains parts of Texas, New Mexico, and Missouri.", "semantic_evaluation": "The phrase is correct. It accurately describes the Arkansas River's location - states it flows through (Colorado, Kansas, Oklahoma, Arkansas) and its watershed (Texas, New Mexico, Missouri)."
```

Right Tab (Chunk Size 2500 Overlap 500):

```
1 [ "index": 11, "prediction": "Now That's What I Call Music (original UK album) was released on 28 November 1983.", "prediction_golden": "Now That's What I Call Music (also simply titled Now or Now 1) was released in the United Kingdom on 28 November 1983.", "semantic_evaluation": "Yes, the statement is correct. The original UK album \"Now That's What I Call Music\" (released as \"Now\" or \"Now 1\") was indeed released on 28 November 1983.", "semantic_evaluation_golden": "The statement is correct. The album \"Now That's What I Call Music\" (also titled \"Now\" or \"Now 1\") was indeed released in the United Kingdom on 28 November 1983, as explicitly stated in the provided information." }, { "index": 13, "prediction": "The 2018 Grey Cup is being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.", "prediction_golden": "The 2018 Grey Cup is scheduled to be played at Commonwealth Stadium in Edmonton, Alberta.", "semantic_evaluation": "Yes, the statement is correct. The 2018 Grey Cup was played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on November 25, 2018.", "semantic_evaluation_golden": "Yes, the phrase is correct. The 2018 Grey Cup was indeed scheduled to be played at The Brick Field (a nickname for Commonwealth Stadium in Edmonton, Alberta) on November 25, 2018." }, { "index": 16, "prediction": "The band Bastille got their name from Bastille Day, which is celebrated on 14 July, the date of lead vocalist Dan Smith's birthday.", "prediction_golden": "The band Bastille got their name based on Bastille Day, which is celebrated on 14 July, the date of lead vocalist Dan Smith's birthday.", "semantic_evaluation": "Yes, the phrase is correct. The band Bastille got their name from Bastille Day (14 July), which is the birthday of lead vocalist Dan Smith.", "semantic_evaluation_golden": "Yes, the statement is correct. The band Bastille was named after Bastille Day (14 July), which is the lead vocalist Dan Smith's birthday." }, { "index": 17, "prediction": "The Arkansas River is located in the United States, flowing through the states of Colorado, Kansas, Oklahoma, and Arkansas. It originates in the Rocky Mountains near Leadville, Colorado, and flows southeastward, eventually emptying into the Mississippi River near Napoleon, Arkansas.", "prediction_golden": "The Arkansas River is located in Colorado, Kansas, Oklahoma, and Arkansas, and its watershed also drains parts of Texas, New Mexico, and Missouri.", "semantic_evaluation": "The phrase is correct. It accurately describes the Arkansas River's location - states it flows through (Colorado, Kansas, Oklahoma, Arkansas) and its watershed (Texas, New Mexico, Missouri)."
```

Confronto Risposte tra Chunk 100 Overlap 20 e Chunk 2500 Overlap 500

Analisi: Chunk Size



Metriche Generation con i migliori Overlap per ciascun Size

Esperimento: Temperature

Fase: Generation

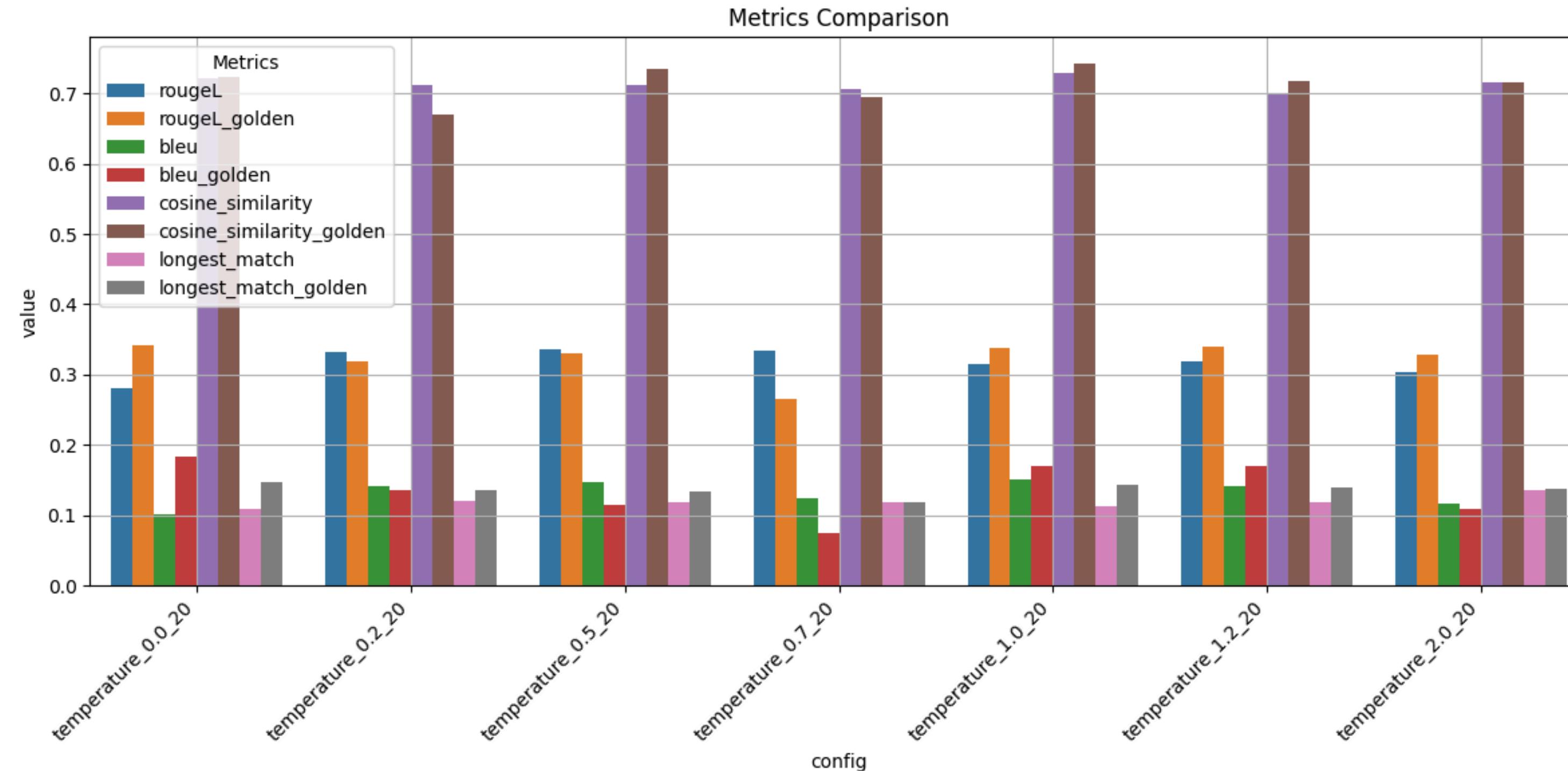
Valori testati:

- 0.0
- 0.2
- 0.5
- 0.7
- 1.0
- 1.2
- 2.0

Obiettivo: analizzare l'effetto della temperatura sulla creatività e coerenza delle risposte

05

Analisi: Temperature



Metriche Generation al variare della Temperatura

Analisi: Temperature

```
1 [ ... ] responses.json temperature_0.0_2025_07_11_155513
2 { ...
3   "index": 11,
4   "prediction": "The context states that the first album in the UK Now! series, titled \"Now That's What I Call Music (original UK album)\" was released in the United Kingdom on 28 November 1983.",
5   "prediction_golden": "Now That's What I Call Music (also simply titled Now or Now 1) was released in the United Kingdom on 28 November 1983.",
6   "semantic_evaluation": "Yes, the statement is correct. The context confirms that the first album in the UK Now! series, titled \"Now That's What I Call Music (also simply titled Now or Now 1)\", was released in the UK on 28 November 1983.",
7   "semantic_evaluation_golden": "Yes, the statement is correct. The album \"Now That's What I Call Music\" (also titled \"Now\" or \"Now 1\") was released in the United Kingdom on 28 November 1983."
8 },
9 {
10   "index": 13,
11   "prediction": "The 2018 Grey Cup is being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.",
12   "prediction_golden": "The 2018 Grey Cup is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.",
13   "semantic_evaluation": "Yes, the statement is correct. The 2018 Grey Cup was scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on November 25, 2018.",
14   "semantic_evaluation_golden": "Yes, the phrase is correct. The 2018 Grey Cup was scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on November 25, 2018."
15 },
16 {
17   "index": 16,
18   "prediction": "The band Bastille got their name based on the French word for \"stronghold\" or \"fortress,\" which is \"Bastille.\" The name was chosen by lead singer Dan Smith, who was inspired by the French Revolution and the storming of the Bastille in Paris. The name is stylized as BASTILLE, with the delta symbol representing the French letter \"é\" in the name.",
19   "prediction_golden": "The band Bastille got their name based on Bastille Day, which is celebrated on 14 July, the date of lead vocalist Dan Smith's birthday.",
20   "semantic_evaluation": "The statement is **incorrect** based on the given information. The original text only states that the band's name derives from Bastille Day (14 July, Dan Smith's birthday), without mentioning the French word for \"stronghold\" or the storming of the Bastille. The stylization detail with the delta symbol is also not mentioned in the original passage.",
21   "semantic_evaluation_golden": "Yes, the phrase is correct. The band's name is derived from Bastille Day, which aligns with lead vocalist Dan Smith's birthday on July 14."
22 },
23 { ... ] responses.json temperature_1.0_2025_07_11_163459
24 { ...
25   "index": 11,
26   "prediction": "The context states that \"Now That's What I Call Music ( 1983 ) Now That's What I Call Music II ( 1984 ) Now That's What I Call Music ( also simply titled Now or Now 1 ) is the first album from the popular Now ! series that was released in the United Kingdom on 28 November 1983.\nBased on this information, the album \"Now That's What I Call Music\" was released in the United Kingdom on 28 November 1983.",
27   "prediction_golden": "The album \"Now That's What I Call Music\" (also simply titled \"Now\" or \"Now 1\") was released in the United Kingdom on 28 November 1983.",
28   "semantic_evaluation": "The phrase is **incorrect**. The original context only mentions the 1983 release of \"Now That's What I Call Music (also simply titled Now or Now 1)\" as the first album in the series. The user's statement incorrectly includes the 1984 album (\"Now That's What I Call Music II\") and repeats the 1983 release, which is not supported by the provided information.",
29   "semantic_evaluation_golden": "Yes, the statement is correct. The album \"Now That's What I Call Music\" (also titled \"Now\" or \"Now 1\") was released in the United Kingdom on 28 November 1983."
30 },
31 {
32   "index": 13,
33   "prediction": "The 2018 Grey Cup is being played at Commonwealth Stadium in Edmonton, Alberta, as reported by several news outlets on March 24, 2017. The game is scheduled to be held on Sunday, November 25, 2018.",
34   "prediction_golden": "The 2018 Grey Cup is being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.",
35   "semantic_evaluation": "The second sentence is correct. The original statement confirms the game is scheduled for November 25, 2018. The second sentence states the 2018 Grey Cup is being played at Commonwealth Stadium in Edmonton, Alberta, as reported in 2017, aligning with the first statement's date. The tense (\"being played\") matches the scheduled date (\"scheduled to be played\"), and the reporting date in 2017 is feasible for a future event.",
36   "semantic_evaluation_golden": "The statement is correct. The 2018 Grey Cup is indeed being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on Sunday, November 25, 2018."
37 },
38 {
39   "index": 16,
40   "prediction": "The band Bastille got their name based on the French word \"Bastille,\" which refers to the historic fortress in Paris, France. The name is stylized as BASTILLE, which is an anagram of \"Bastille.\" This unique choice of name reflects the band's eclectic musical style and their global appeal, drawing inspiration from various cultural influences.",
41   "prediction_golden": "The band Bastille got their name based on Bastille Day, which is celebrated on 14 July, the date of lead vocalist Dan Smith's birthday."
42 }
```

Confronto Risposte tra Temperature 0 (sinistra) e Temperature 1 (destra)

Esperimento: Max New Token

Fase: Generation

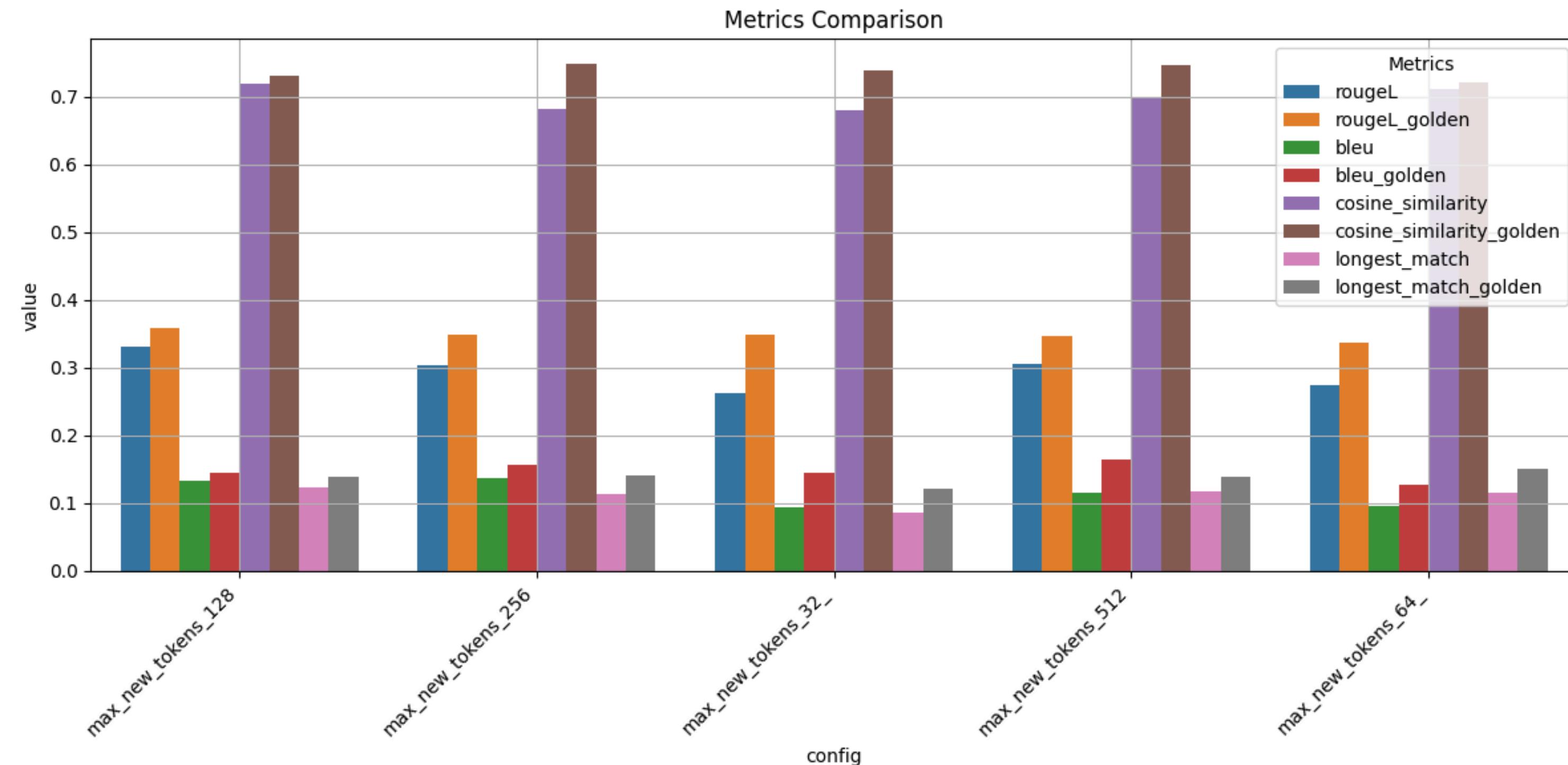
Valori testati:

- 32
- 64
- 128
- 256
- 512

Obiettivo: analizzare la completezza, coerenza e qualità delle risposte

05

Analisi: Max New Token



Metriche Generation al variare di Max New Tokens

Analisi: Max New Token

The image shows a code editor with two tabs open, comparing JSON responses generated by different models based on the same input. The left tab is for a model with 32 maximum new tokens, and the right tab is for a model with 256 maximum new tokens. Both responses include line numbers and JSON structures.

Left Tab (32 Max New Tokens):

```
43 },
44 {
45     "index": 23,
46     "prediction": "Star Trek: The Next Generation was announced on October 10, 1986.",
47     "prediction_golden": "Star Trek: The Next Generation first aired on 26 September 1987.
",
48     "semantic_evaluation": "Yes, the statement is correct. Star Trek: The Next Generation
was officially announced on October 10, 1986. The first season premiered on September
28, 1987, as listed in the data provided.",
49     "semantic_evaluation_golden": "The statement is incorrect. The first episode of Star
Trek: The Next Generation aired on **September 28, 1987**, not September 26."
50 },
51 {
52     "index": 32,
53     "prediction": "Based on the provided context, Jackie Ward, who is better known as
Robin Ward, sang the song \"Wonderful Summer\". The context mentions that she was
hired to make a demo recording of \"Wonderful Summer\" in 1963, which was later
released by Dot Records as a 45rpm single. The song sold over one million copies and
reached #14 on the Billboard Hot 100 chart in December 1963. Therefore, Jackie Ward
(Robin Ward) sang \"Wonderful Summer\".",
54     "prediction_golden": "The context mentions Jackie Ward, also known as Robin Ward, as
an American singer who had a 1963 million-selling single called \"Wonderful Summer.\" It
does not mention any other songs or artists associated with \"The Most Wonderful
Summer of My Life.\" Therefore, based on the provided context, Jackie Ward did not
sing \"The Most Wonderful Summer of My Life.\",
55     "semantic_evaluation": "Yes, the statement is correct. The context explicitly states
that Robin Ward (Jackie Ward) sang \"Wonderful Summer,\" which became a
million-selling hit in 1963, and references the demo recording by Dot Records. The
provided details align with the assertion that she sang the song.",
56     "semantic_evaluation_golden": "The statement is correct. The context does not mention
\"The Most Wonderful Summer of My Life\" in relation to Jackie Ward's work. The only
song explicitly tied to her is \"Wonderful Summer\" from 1963. There is no indication
of her involvement with the other song."
57 },
58 {
59     "index": 37,
60     "prediction": "Based on the context provided, On Golden Pond won two Oscars: Best
Actor for Henry Fonda and Best Actress for Katharine Hepburn.",
61     "prediction_golden": "On Golden Pond won three Oscars based on the context provided.
The context mentions that the film won three awards: Best Actor (Henry Fonda), Best
Actress (Katharine Hepburn), and Best Adapted Screenplay (Thompson).",
62     "semantic_evaluation": "The statement is incorrect. On Golden Pond won **three**
Oscars: Best Actor (Henry Fonda), Best Actress (Katharine Hepburn), and Best Adapted
Screenplay. The user's version only lists two, omitting the third.",
63     "semantic_evaluation_golden": "The statement is correct. The context explicitly lists
the three Academy Award wins for *On Golden Pond* as Best Actor (Henry Fonda), Best
Actress (Katharine Hepburn), and Best Adapted Screenplay (Thompson)."
```

Right Tab (256 Max New Tokens):

```
37 {
38 },
39 {
40     "index": 23,
41     "prediction": "Star Trek: The Next Generation was first announced on October 10, 1986,
42     and its cast was revealed in May 1987.",
43     "prediction_golden": "Star Trek: The Next Generation first aired on September 28, 1987.
",
44     "semantic_evaluation": "The statement is incorrect. Star Trek: The Next Generation was
45     announced in October 1986, but the cast was revealed in May 1987, not in October 1986.
46     The first season started in September 1987, aligning with the announcement and cast
47     reveal timing. The user's phrasing conflates the original announcement with the cast
48     reveal date.",
49     "semantic_evaluation_golden": "Yes, the statement is correct. Star Trek: The Next
50     Generation first aired on September 28, 1987."
51 },
52 {
53     "index": 32,
54     "prediction": "The song \"Wonderful Summer\" was sung by Jackie Ward, also known as
55     Robin Ward, based on the context provided.",
56     "prediction_golden": "The context mentions that Jackie Ward, better known as Robin
57     Ward, is an American singer who is known for her 1963 million-selling song \"Wonderful
58     Summer.\" However, it does not mention any song called \"Most Wonderful Summer of My
59     Life.\" Based on the given context, it is not clear which singer sang \"Most Wonderful
60     Summer of My Life.\" The context does not provide any information about this song or
61     its singer.",
62     "semantic_evaluation": "Yes, the statement is correct. The context clearly identifies
63     Jackie Ward as Robin Ward's real name and attributes \"Wonderful Summer\" to her. The
64     mention of The Partridge Family is a separate, possibly conflicting detail but does
65     not invalidate the first part.",
66     "semantic_evaluation_golden": "The context does not mention the song \"Most Wonderful
67     Summer of My Life\" or attribute it to Jackie Ward (Robin Ward). Based on the given
68     information, it is not possible to confirm the accuracy of the claim. \n\n**Answer:**\nNo, the context does not provide evidence that Jackie Ward sang \"Most Wonderful
69     Summer of My Life.\""
70 },
71 {
72     "index": 37,
73     "prediction": "On Golden Pond won 3 Oscars based on the context provided. Henry Fonda
74     won his only competitive Oscar for Best Actor, and Katharine Hepburn won her fourth
75     Best Actress award, extending her record for the most Oscars won by a thespian.
76     Additionally, the film was nominated for five major Academy Awards (Best Picture,
77     Director, Actor, Actress, and Screenplay).",
78     "prediction_golden": "On Golden Pond won three Oscars based on the context provided.
79     The context states that the film won three awards: Best Actor (Henry Fonda), Best
80     Actress (Katharine Hepburn), and Best Adapted Screenplay (Thompson).",
81     "semantic_evaluation": "The statement is correct. The context explicitly lists
82     the three Academy Award wins for *On Golden Pond* as Best Actor (Henry Fonda), Best
83     Actress (Katharine Hepburn), and Best Adapted Screenplay (Thompson)."
```

Confronto risposte tra 32 (sinistra) e 256 (destra) Max New Tokens

Esperimento: Tipo di Prompt

Fase: Generation

Obiettivo: valutare l'impatto del prompt engineering sulla robustezza e precisione delle risposte

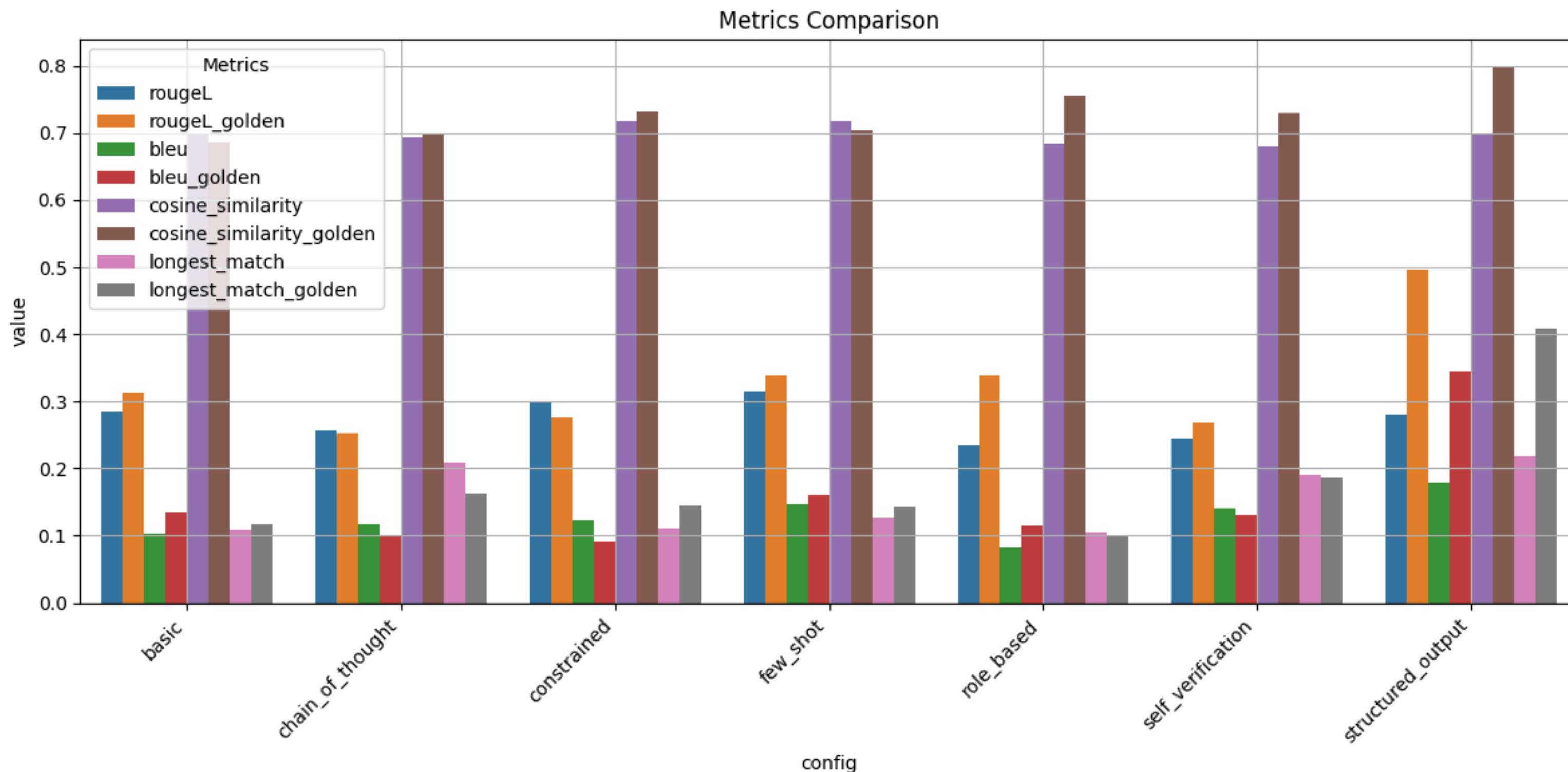
```
CHAIN_OF_THOUGHT_PROMPT = {  
    "system": """You are a precise question-answering assistant. Follow these steps:  
1. First, carefully read and understand the provided context  
2. Identify the key information relevant to the question  
3. Structure your reasoning step by step  
4. Provide a concise answer based ONLY on the context  
5. If information is missing, clearly state what you don't know  
  
Remember: Never make assumptions or add information not present in the context.""""  
    "human": """Context: {context}  
  
Question: {query}  
  
Let's solve this step by step:  
1. What specific information from the context is relevant?  
2. How does this information answer the question?  
3. What is the precise answer based solely on the context?  
  
Answer:"""  
}  
  
FEW_SHOT_PROMPT = {  
    "system": """You are a precise question-answering assistant.  
Answer questions based ONLY on the provided context.""""  
    "human": """Here are some examples of good responses:  
  
Context: The Eiffel Tower was completed in 1889. It stands 324 meters tall.  
Question: When was the Eiffel Tower built?  
Answer: The Eiffel Tower was completed in 1889.  
  
Context: The Eiffel Tower was completed in 1889. It stands 324 meters tall.  
Question: What is its color?  
Answer: I cannot answer this question as the context does not contain any information  
about the Eiffel Tower's color.  
  
Now please answer the following:  
Context: {context}  
Question: {query}"""  
}
```

```
STRUCTURED_OUTPUT_PROMPT = {  
    "system": """You are a precise question-answering assistant.  
Answer questions based ONLY on the provided context.""""  
    "human": """Context: {context}  
  
Question: {query}  
  
Please provide your answer in this format:  
RELEVANT CONTEXT: [Quote the specific parts of the context that are relevant]  
REASONING: [Explain your step-by-step thought process]  
CONFIDENCE: [High/Medium/Low - based on how directly the context answers the question]  
FINAL ANSWER: [Concise answer based only on the context]"""  
}  
  
CONSTRAINED_PROMPT = {  
    "system": """You are a precise question-answering assistant.  
Answer questions based ONLY on the provided context.""""  
    "human": """Important constraints:  
1. Maximum answer length: 50 words  
2. Use only information explicitly stated in the context  
3. If multiple interpretations are possible, list them  
4. For numerical answers, include units if provided  
5. Express uncertainty when context is ambiguous  
  
Context: {context}  
Question: {query}"""  
}  
  
ROLE_BASED_PROMPT = {  
    "system": """You are an expert research assistant with these key traits:  
1. Extreme precision in citing information  
2. Strong analytical skills for understanding context  
3. Honest about knowledge limitations  
4. Clear and concise communication  
5. Methodical in analyzing questions  
  
Answer questions based ONLY on the provided context.""""  
    "human": """Context: {context}  
Question: {query}"""  
}
```

```
SELF_VERIFICATION_PROMPT = {  
    "system": """You are a precise question-answering assistant.  
Answer questions based ONLY on the provided context.""""  
    "human": """Context: {context}  
Question: {query}  
  
After formulating your answer, please:  
1. Verify that every statement is supported by the context  
2. Check if any assumptions were made  
3. Confirm the answer directly addresses the question  
4. Ensure no external knowledge was used  
5. Validate confidence level assessment  
| Ctrl+L to chat, Ctrl+K to generate  
Your response:"""  
}  
  
BASIC_PROMPT = {  
    "system": """Just answer queries based on the provided context.""""  
    "human": """Answer the query: {query} based uniquely on the context: {context},  
don't make up anything, just say what the context contains.  
If the information is not in the context, you must say you don't know.  
You must answer only the specified question and nothing else.""""  
}
```

05

Analisi: Tipo di Prompt



Metriche Generation al variare del Prompt

Analisi: Tipo di Prompt

```
{ responses.json few_shot_2025_07_11_163206 }  
output > generation > prompt_type > few_shot_2025_07_11_163206 > { responses.json > ...  
9 {  
10   correct name for the stadium where the Grey Cup was held. The correct location is simply **Commonwealth Stadium in Edmonton, Alberta**."  
11 }  
12 {  
13   "index": 16,  
14   "prediction": "The band Bastille got their name from the French word \"Bastille,\" which refers to the former royal fortress in Paris that was a symbol of the French Revolution. The band members, including lead singer Dan Smith, were inspired by the historical significance of the Bastille and chose it as their band name.",  
15   "prediction_golden": "The band Bastille got their name from **Bastille Day**, a French national holiday celebrated on **14 July**, which is also the date of lead vocalist Dan Smith's birthday. The name was chosen to reflect this connection.",  
16   "semantic_evaluation": "The statement is correct. The band's name originates from the French word \"Bastille,\" which refers to the historical fortress in Paris that symbolized the French Revolution. The band members, including lead singer Dan Smith, were inspired by this historical significance and chose it as their band name. The original statement about Bastille Day (July 14th) is also accurate, as Smith's birthday aligns with the holiday's date. Both explanations are consistent with the band's origin story.",  
17   "semantic_evaluation_golden": "Yes, the statement is correct. The band Bastille was named after **Bastille Day**, a French national holiday celebrated on **14 July**, and this date also coincides with the birthday of lead vocalist Dan Smith. The connection between the name and the date is explicitly mentioned in the original text."  
18 }  
19 {  
20   "index": 17,  
21   "prediction": "The Arkansas River flows through the U.S. states of Colorado, Kansas, Oklahoma, and Arkansas, and its watershed also drains parts of Texas, New Mexico, and Missouri. So, on a map, you would typically see the Arkansas River located in the central and southern parts of the United
```

```
{ responses.json role_based_2025_07_11_172345 X
output > generation > prompt_type > role_based_2025_07_11_172345 > {} responses.json > ...
9   {
10     "text": "The band Bastille's name is derived from the French word \"b\u00e2ille,\" which means \"open mouth\" or \"wide-eyed.\" The band members chose this name because they wanted to capture the feeling of being wide-eyed and open to the world, as well as to suggest a sense of wonder and curiosity. The stylized name \"B\u00c3STILLE\" is an anagram of \"Bastille,\" created by the band to give their name a unique and memorable appearance.",
11     "index": 16,
12     "prediction": "The band Bastille got their name from the French word \"b\u00e2ille,\" which means \"open mouth\" or \"wide-eyed.\" The band members chose this name because they wanted to capture the feeling of being wide-eyed and open to the world, as well as to suggest a sense of wonder and curiosity. The stylized name \"B\u00c3STILLE\" is an anagram of \"Bastille,\" created by the band to give their name a unique and memorable appearance.",
13     "prediction_golden": "The band Bastille took their name from **Bastille Day**, a French national holiday celebrated on July 14th. This date corresponds to the birthdate of lead vocalist Dan Smith, who chose the name to reflect both the holiday and his personal connection to it.",
14     "semantic_evaluation": "The statement is **incorrect**. The band Bastille's name originates from **Bastille Day**, not the French word \"b\u00e2ille.\"
15     "semantic_evaluation_golden": "The phrase is not fully correct. The original statements only mention the name derives from Bastille Day because it's Smith's birthday. The added detail about the holiday's significance (\\"a French national holiday celebrated on July 14th\\") and the personal connection (\\"who chose the name to reflect both the holiday and his personal connection to it\\") are not explicitly stated in the given information."
16   },
17   {
18     "text": "The Arkansas River is located in the central United States, flowing through **Colorado, Kansas, Oklahoma, and Arkansas**. Its headwaters
19     "index": 17,
20     "prediction": "The Arkansas River is located in the central United States, flowing through **Colorado, Kansas, Oklahoma, and Arkansas**. Its headwaters
21     "prediction_golden": "The Arkansas River is located in the central United States, flowing through **Colorado, Kansas, Oklahoma, and Arkansas**. Its headwaters
22     "semantic_evaluation": "The statement is **incorrect**. The original statement only mentions the river flows through Colorado, Kansas, Oklahoma, and Arkansas. It does not mention its headwaters.
23     "semantic_evaluation_golden": "The phrase is not fully correct. The original statements only mention the river flows through Colorado, Kansas, Oklahoma, and Arkansas. It does not mention its headwaters."
24   },
25   {
26     "text": "The stadium was built in 1926 and has a capacity of approximately 65,000 spectators. It is located in the city of Edmonton, Alberta, Canada. The stadium is known for its distinctive red exterior and its unique retractable roof system. The stadium has hosted numerous major events, including the Commonwealth Games, the Pan American Games, and the World Cup. The stadium is also home to the Edmonton Eskimos football team and the Edmonton Oilers ice hockey team."}
```

Confronto risposte tra prompt base (sinistra), few shot (centro) e role based (destra)

Analisi: Tipo di Prompt

```
responses.json constrained_2025_07_11_164532
output > generation > prompt_type > constrained_2025_07_11_164532 > responses.json > {} 1
1 [ "index": 11, "prediction": "The first album, \"Now That's What I Call Music\" (original UK album), was released in the United Kingdom on 28 November 1983. It was initially available on vinyl and audio cassette.", "prediction_golden": "The album \"Now That's What I Call Music\" was released on **28 November 1983**.", "semantic_evaluation": "Yes, the statement is correct. The original UK album \"Now That's What I Call Music\" (Now 1) was indeed released in the UK on 28 November 1983, initially available on vinyl and audio cassette.", "semantic_evaluation_golden": "Yes, the statement is correct. The album \"Now That's What I Call Music\" was indeed released on **28 November 1983**." }, { "index": 13, "prediction": "The 2018 Grey Cup is being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta, on Sunday, November 25, 2018.", "prediction_golden": "The 2018 Grey Cup is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on Sunday, November 25, 2018.\n**Answer:** The Brick Field, Commonwealth Stadium, Edmonton, Alberta.", "semantic_evaluation": "Yes, the phrase is correct. The 2018 Grey Cup, scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta, on Sunday, November 25, 2018, is accurately described. The original statement confirms the date, location, and event, and adding \"The 2018 Grey Cup\" enhances clarity.", "semantic_evaluation_golden": "The statement is correct. The 2018 Grey Cup was scheduled to be played at The Brick Field in Edmonton, Alberta on Sunday, November 25, 2018. This is because Commonwealth Stadium was undergoing renovations during that time, and The Brick Field served as a temporary venue for the event.\n**Answer:** The Brick Field at Commonwealth Stadium in Edmonton, Alberta." }, { "index": 16, "prediction": "Bastille got their name from the French word for \"royal palace,\" which symbolizes their ambition and aspiration. The band was formed in London, England, in 2010.", "prediction_golden": "Bastille's name derives from Bastille Day (14 July), the French national holiday, and the lead vocalist Dan Smith's birthday. The band members are Dan Smith (vocals), Kyle Simmons (keyboard), Will Farquarson (bass/guitar), and Chris Wood (drums).", "semantic_evaluation": "The statement is **not correct** based on the provided information. The band's name originates from **Bastille Day**, not the French word for \"royal palace.\" The original text explicitly states that the name derives from the holiday celebrated on 14 July, which is Dan Smith's birthday. There is no mention of the French word \"palais royal\" or any connection to a royal palace in the given context. The name is stylized as BASTILLE, a play on the word \"Bastille,\" referencing the French Revolution's storming of the Bastille Fortress." }, responses.json structured_output_2025_07_11_164007
output > generation > prompt_type > structured_output_2025_07_11_164007 > responses.json > ...
1 [ "index": 11, "prediction": "RELEVANT CONTEXT:\n\"Now That 's What I Call Music ( 1983 ) Now That 's What I Call Music II ( 1984 ) Now That 's What I Call Music ( also simply titled Now or Now 1 ) is the first album from the popular Now ! series that was released in the United Kingdom on 28 November 1983.\n\nREASONING:\nThe question asks when \"Now That's What I Call Music\" was released. The context clearly states that the first album in the series was released in November 1983.\n\nCONFIDENCE: High\nFINAL ANSWER: November 1983", "prediction_golden": "RELEVANT CONTEXT: Now That 's What I Call Music ( also simply titled Now or Now 1 ) is the first album from the popular Now ! series that was released in the United Kingdom on 28 November 1983 .\n\nREASONING: The context explicitly states the release date of the album in the first sentence.\n\nCONFIDENCE: High\nFINAL ANSWER: 28 November 1983", "semantic_evaluation": "The statement is correct. The first album in the Now series, \"Now That's What I Call Music,\" was released on **28 November 1983**.\n\n**Final Answer:** November 1983\n**Confidence:** High", "semantic_evaluation_golden": "The statement \"28 November 1983\" is correct, as explicitly mentioned in the context. The album's initial release date is confirmed to be on 28 November 1983 in the United Kingdom.\n\n**Final Answer:** 28 November 1983\n**Confidence:** High" }, { "index": 13, "prediction": "RELEVANT CONTEXT: \"The game will be televised in Canada nationally on TSN. The game is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on Sunday, November 25, 2018.\n\nREASONING: The context explicitly states the location of the game, which is Commonwealth Stadium in Edmonton, Alberta.\n\nCONFIDENCE: High\nFINAL ANSWER: The 2018 Grey Cup is being played at Commonwealth Stadium in Edmonton, Alberta.", "prediction_golden": "RELEVANT CONTEXT: The game is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on Sunday, November 25, 2018.\n\nREASONING: The context explicitly states the venue as \"The Brick Field at Commonwealth Stadium in Edmonton, Alberta.\"
\n\nCONFIDENCE: High\nFINAL ANSWER: The 2018 Grey Cup is being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.", "semantic_evaluation": "The statement is correct. The 2018 Grey Cup was indeed played at Commonwealth Stadium in Edmonton, Alberta on Sunday, November 25, 2018. The context explicitly confirms this location and date, and the game was nationally televised on TSN in Canada.\n\n**Final Answer:** The 2018 Grey Cup is being played at Commonwealth Stadium in Edmonton, Alberta.", "semantic_evaluation_golden": "The context explicitly states the game is scheduled for \"The Brick Field at Commonwealth Stadium in Edmonton, Alberta\" on November 25, 2018. The phrase \"The 2018 Grey Cup is being played...\" is correct because the Grey Cup is the championship game of the Canadian Football League (CFL), and the given venue and date align with the 2018 Grey Cup location. The Brick Field at Commonwealth stadium is" }
```

Risposte del miglior modello sintattico: constrained (sinistra) e semantico: structured (destra)

Analisi: Tipo di Prompt

```
responses.json self_verification_2025_07_11_173245
output > generation > prompt_type > self_verification_2025_07_11_173245 > responses.json > {} 1 > prediction
2 {
    "index": 13,
    "prediction": "Based on the provided context, the 2018 Grey Cup game is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta, Canada. This information is confirmed by the statement: \"The game is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on Sunday, November 25, 2018.\" The context also mentions that Calgary and Edmonton were bidding to host the game, and Edmonton was officially named the host city on June 5, 2017.",
    "prediction_golden": "1. Verification: The context states that \"The game is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on Sunday, November 25, 2018.\"\\n2. Assumptions: None were made beyond what the context provided.\\n3. Answer: The 2018 Grey Cup is being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.\\n4. External knowledge: None was used.\\n5. Confidence level: High, as the information is directly provided in the context.",
    "semantic_evaluation": "Yes, the statement is correct. The context confirms the 2018 Grey Cup game is scheduled for November 25, 2018, at The Brick Field at Commonwealth Stadium in Edmonton, Alberta. The location and year are accurately stated, even though the specific date is also provided in the context.",
    "semantic_evaluation_golden": "The context explicitly states the game is scheduled for Sunday, November 25, 2018, at The Brick Field at Commonwealth Stadium in Edmonton, Alberta. The answer correctly identifies this as the 2018 Grey Cup, which aligns with the provided details. \\n\\n**Answer:** Yes, the statement is correct. \\n**Confidence Level:** High."
}
{
    "index": 16,
    "prediction": "The band Bastille got their name from the French word \"Bastille,\" which refers to the Parisian fortress that was stormed during the French Revolution. The band members chose this name because they wanted to evoke a sense of revolution and change, similar to the events that took place during the French Revolution. The name was also chosen to emphasize the band's unique sound and style, which blends elements of indie pop, synth-pop, and indie rock.",
    "prediction_golden": "The band Bastille got their name from **Bastille Day**, a French national holiday celebrated on **14 July**, which is also the date of lead vocalist Dan Smith's birthday. This information is explicitly stated in the context. \\n\\n**Verification:** Supported by the context's last sentence. \\n**Assumptions:** None made; the origin is directly attributed to Bastille Day and Dan Smith's birthday. \\n**Answer Directness:** The question is fully addressed by stating the origin as Bastille Day and its connection to Smith's birthday. \\n**External Knowledge:** No external information used. \\n**Confidence Level:** High (100%)
}

responses.json chain_of_thought_2025_07_11_162546
output > generation > prompt_type > chain_of_thought_2025_07_11_162546 > responses.json > {} 2 > prediction
2 {
    "index": 13,
    "prediction": "The 2018 Grey Cup is being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.",
    "prediction_golden": "1. The specific information from the context relevant to the question is: \"The game is scheduled to be played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta on Sunday, November 25, 2018.\"\\n2. This information answers the question by providing the exact location where the 2018 Grey Cup is being played: The Brick Field at Commonwealth Stadium in Edmonton, Alberta.\\n3. The precise answer based solely on the context is: The 2018 Grey Cup is being played at The Brick Field at Commonwealth Stadium in Edmonton, Alberta.",
    "semantic_evaluation": "The phrase is not entirely correct because it uses the present tense (\"is being played\") instead of the scheduled future tense (\"was scheduled to be played\"). However, the location and event details are accurate.",
    "semantic_evaluation_golden": "The provided sentence correctly states the location of the 2018 Grey Cup as **The Brick Field at Commonwealth Stadium in Edmonton, Alberta**, on **Sunday, November 25, 2018**. This information directly answers the question by specifying the exact venue and date. The answer is accurate and concise."
}
{
    "index": 16,
    "prediction": "The band Bastille got their name from the French word \"Bastille,\" which refers to the historical fortress in Paris that was stormed in 1789, symbolizing the start of the French Revolution. The name was chosen by the band members as a nod to this iconic event, reflecting their inspiration and the symbolic meaning behind it.",
    "prediction_golden": "1. The relevant information from the context is: \"The name of the band derives from Bastille Day, which is celebrated on 14 July, the date of Smith's birthday.\"\\n2. This information answers the question by explaining the origin of the band's name, specifically stating that it is derived from Bastille Day, the date of the lead vocalist Dan Smith's birthday.\\n3. The precise answer based solely on the context is: The band Bastille got their name from Bastille Day, which is celebrated on 14 July, the date of lead vocalist Dan Smith's birthday.",
    "semantic_evaluation": "The statement is incorrect. The original text only mentions that the band's name derives from Bastille Day (14 July, Dan Smith's birthday), without referencing the historical Bastille fortress or the French Revolution.",
    "semantic_evaluation_golden": "All three statements are correct. The band's name"
}
```

Risposte dei modelli peggiori: self verification (sinistra) e chain of thought (destra)

Esperimento: Top-k

Fase: Generation (+ Retrieve)

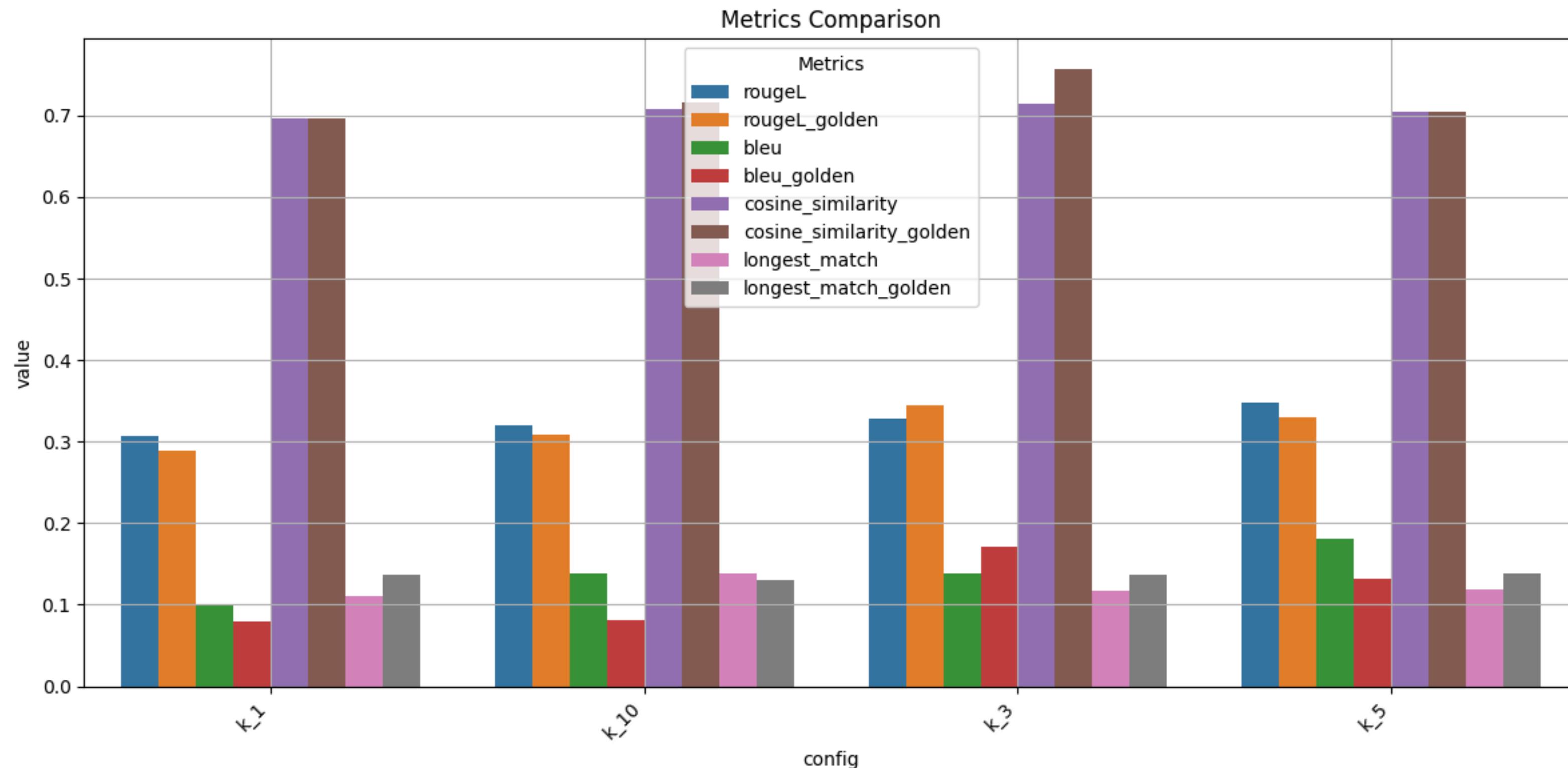
Valori testati:

- 1
- 3
- 5
- 10

Obiettivo studiare la quantità di contesto ottimale da fornire al generatore

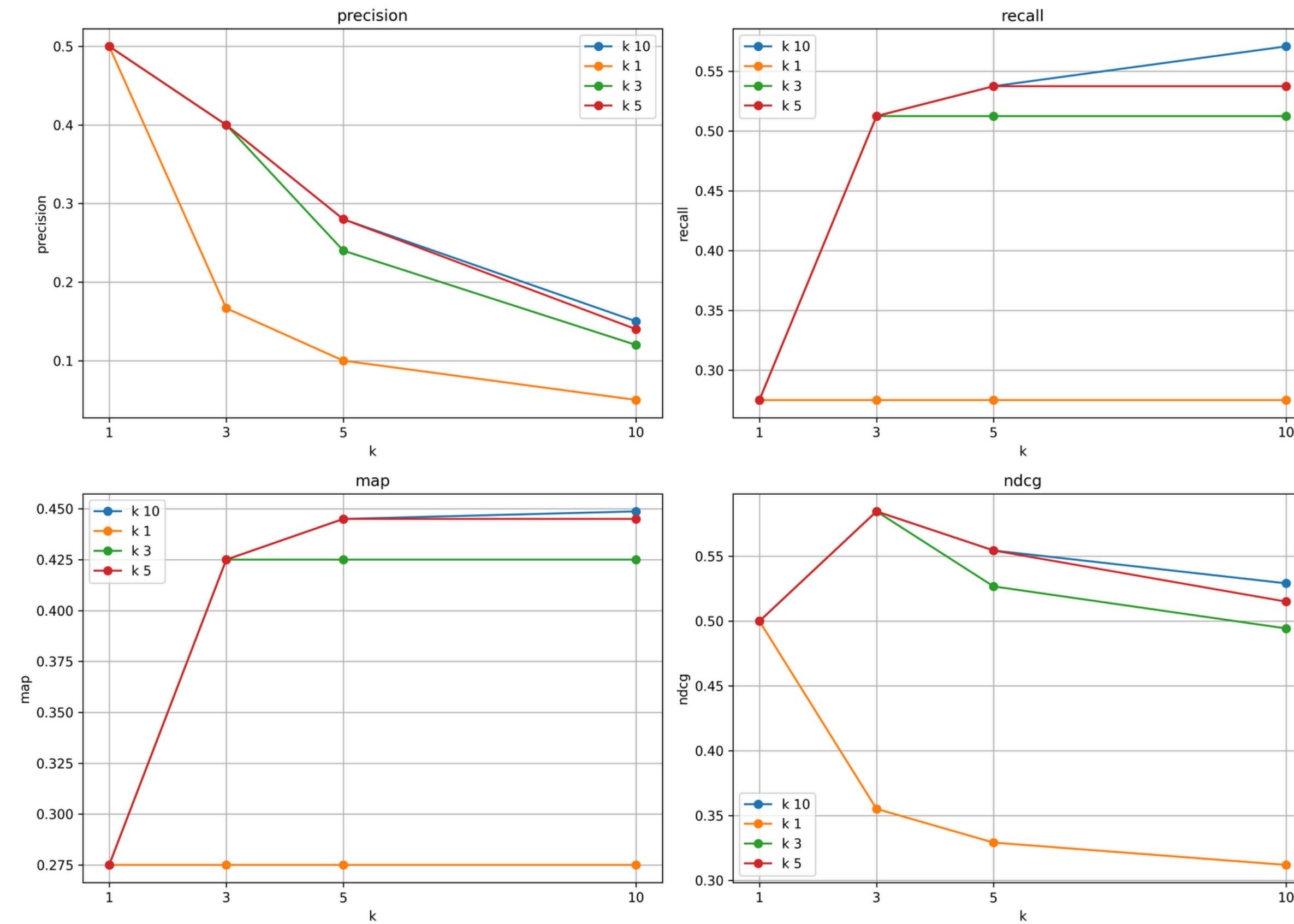
05

Analisi: Top-k



Metriche Generation al variare del parametro k

Analisi: Top-k



Metriche Retrieve al variare del parametro k

Analisi: Top-k

The image shows three separate code editor windows side-by-side, each displaying a JSON document. The leftmost window is for k=1, the middle for k=3, and the rightmost for k=5. Each window has a title bar with the file name 'retrieved_docs.json' and a timestamp. The code is identical across all three windows, showing a list of documents indexed by their ID. Each document entry includes 'relevant_ids' (the query terms) and 'retrieved_ids' (the documents found). The retrieved documents are ordered by relevance, with higher IDs appearing first.

```
1 [ ... 55 ]
```

```
1 [ ... 55 ]
```

```
1 [ ... 55 ]
```

Documenti Recuperati con k=1 (sinistra), k=3 (centro), k=5 (destra)

Esperimenti Pipeline BEIR

Embedding models:

- all-MiniLM-L6-v2: ottimizzato per semantic search
- bge-base-en-v1.5: migliore generalizzazione
- e5-base-v2: compiti di retrieval
- all-distilroberta-v1: versione leggera di RoBERTa
- all-mpnet-base-v2: dimensioni maggiori
- multi-qa-MiniLM-L6-cos-v1: ottimizzato per Q&A
- paraphrase-multilingual-MiniLM-L12-v2: modello multilingue

Score function:

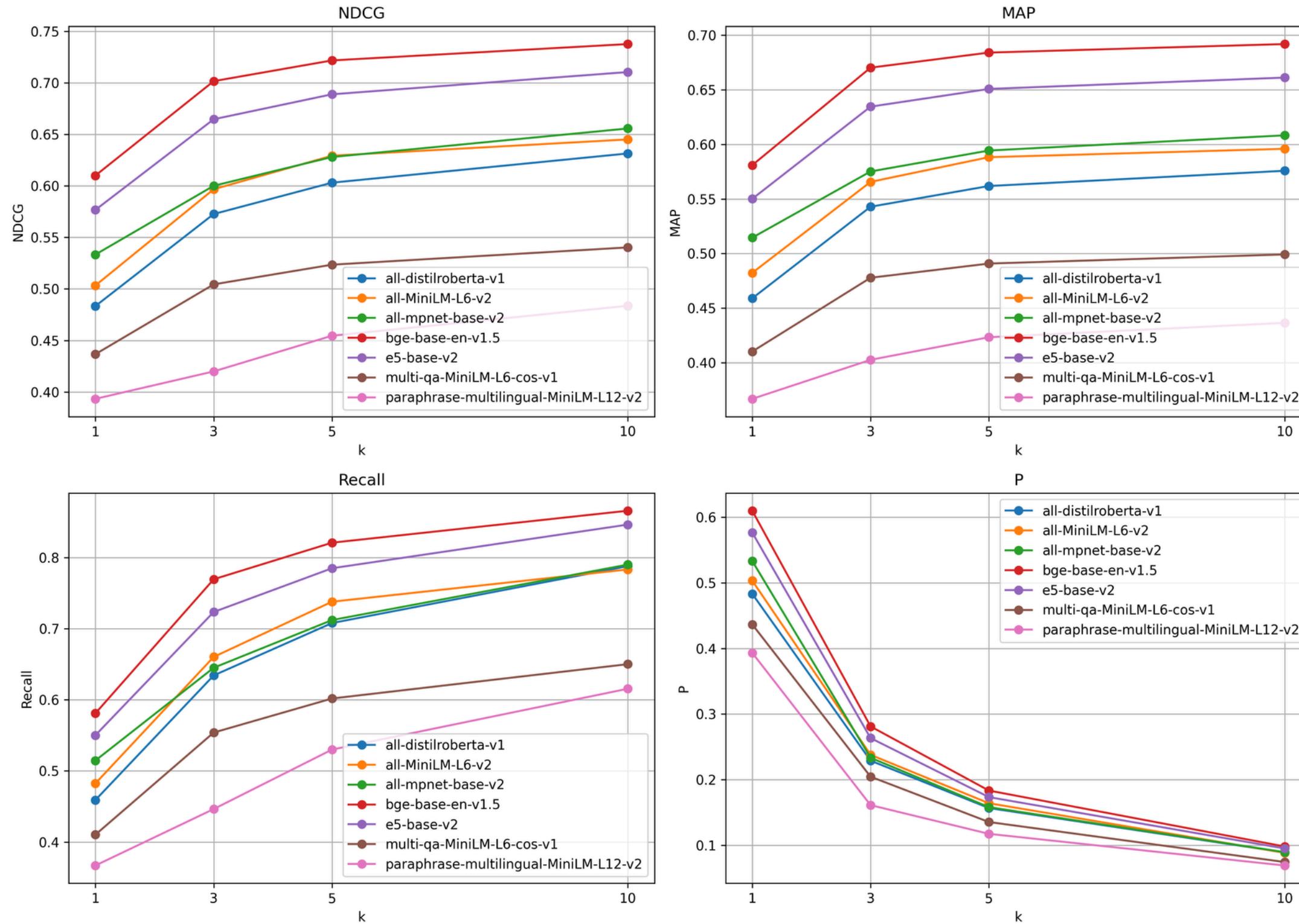
- cosine similarity
- dot product

Dataset BEIR:

- SciFact
- SCIDOCs
- FiQA-2018

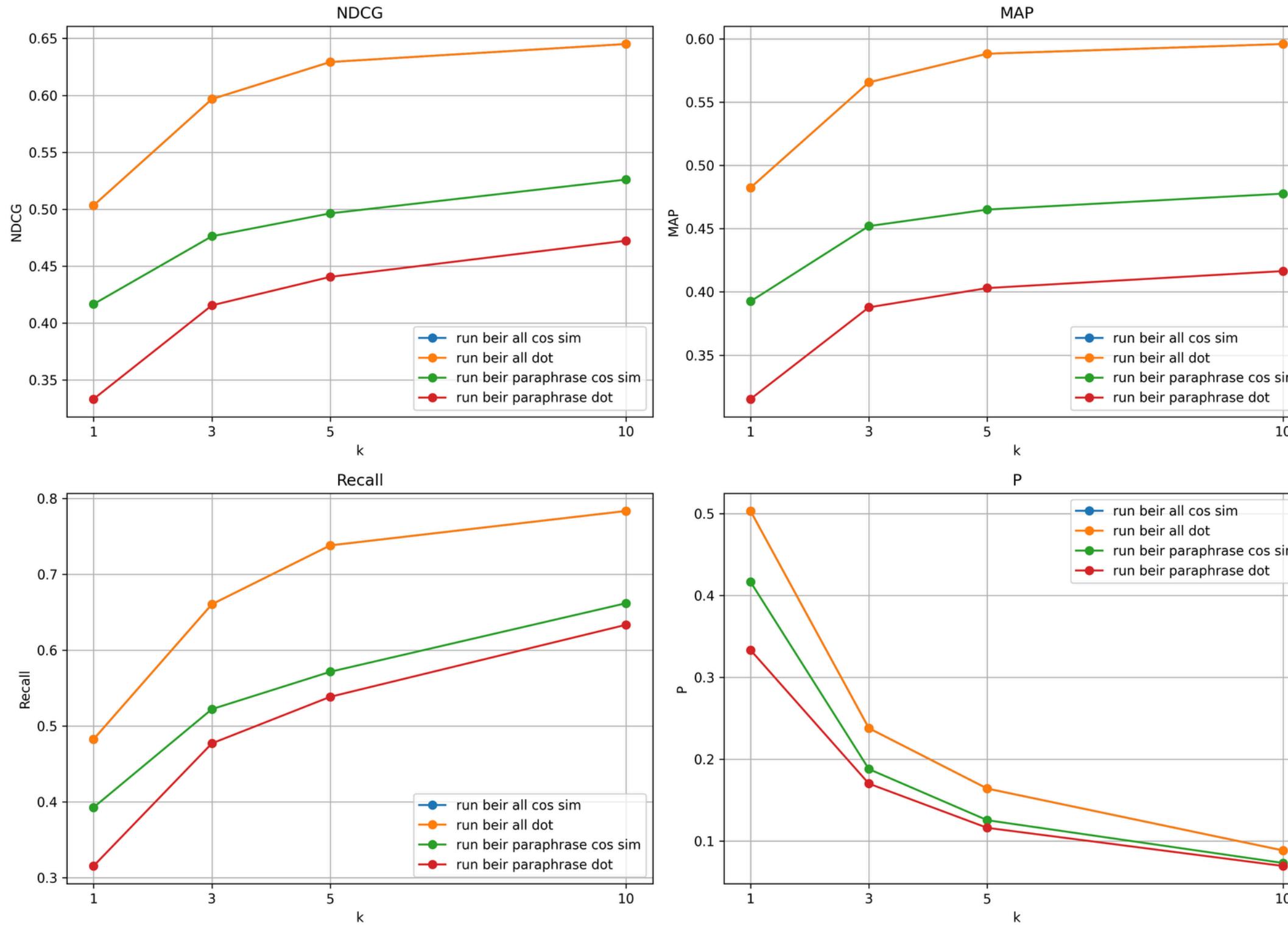
05

Embedding Model



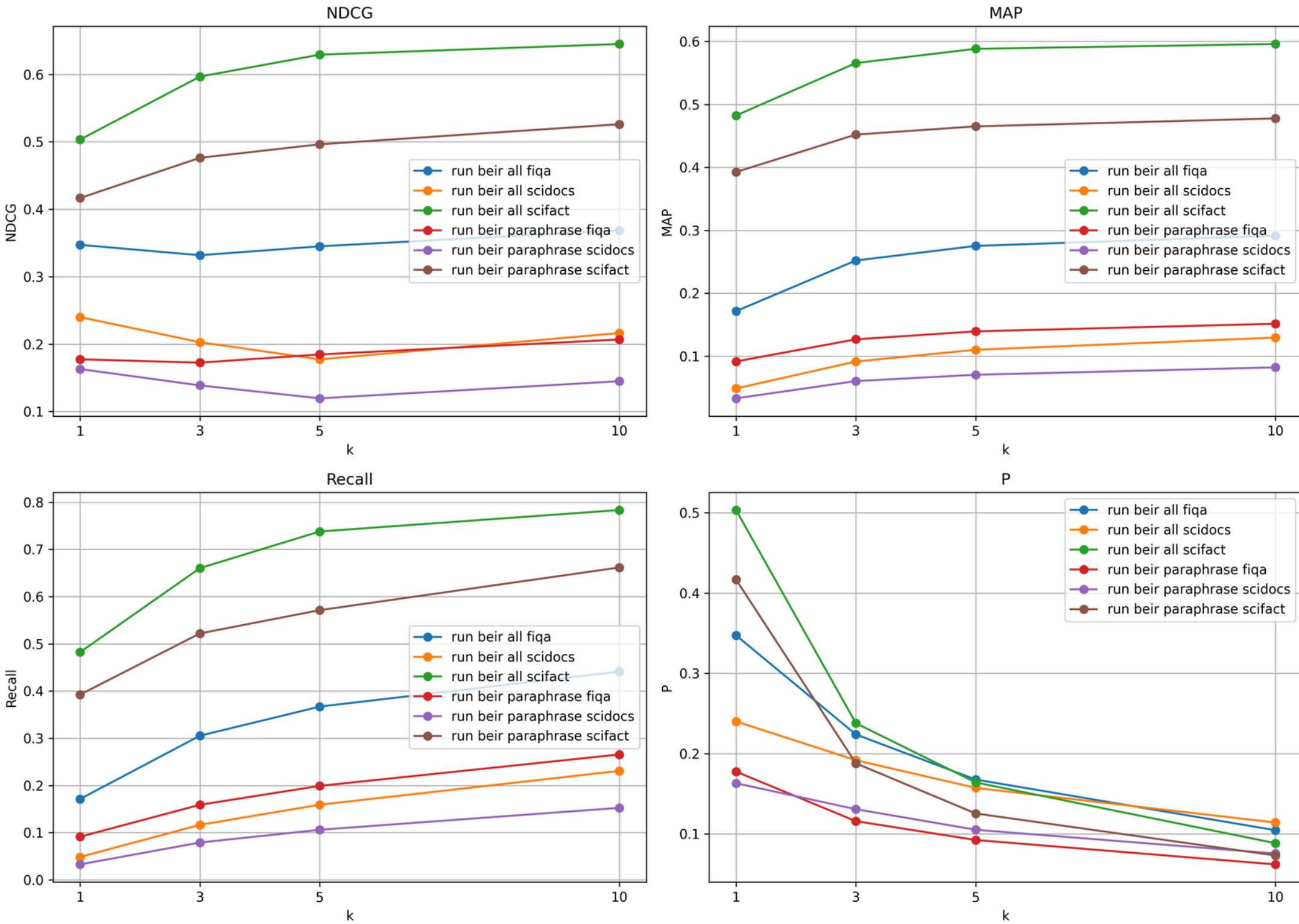
Metriche Retrieval al variare del Modello di Embedding

Score Function



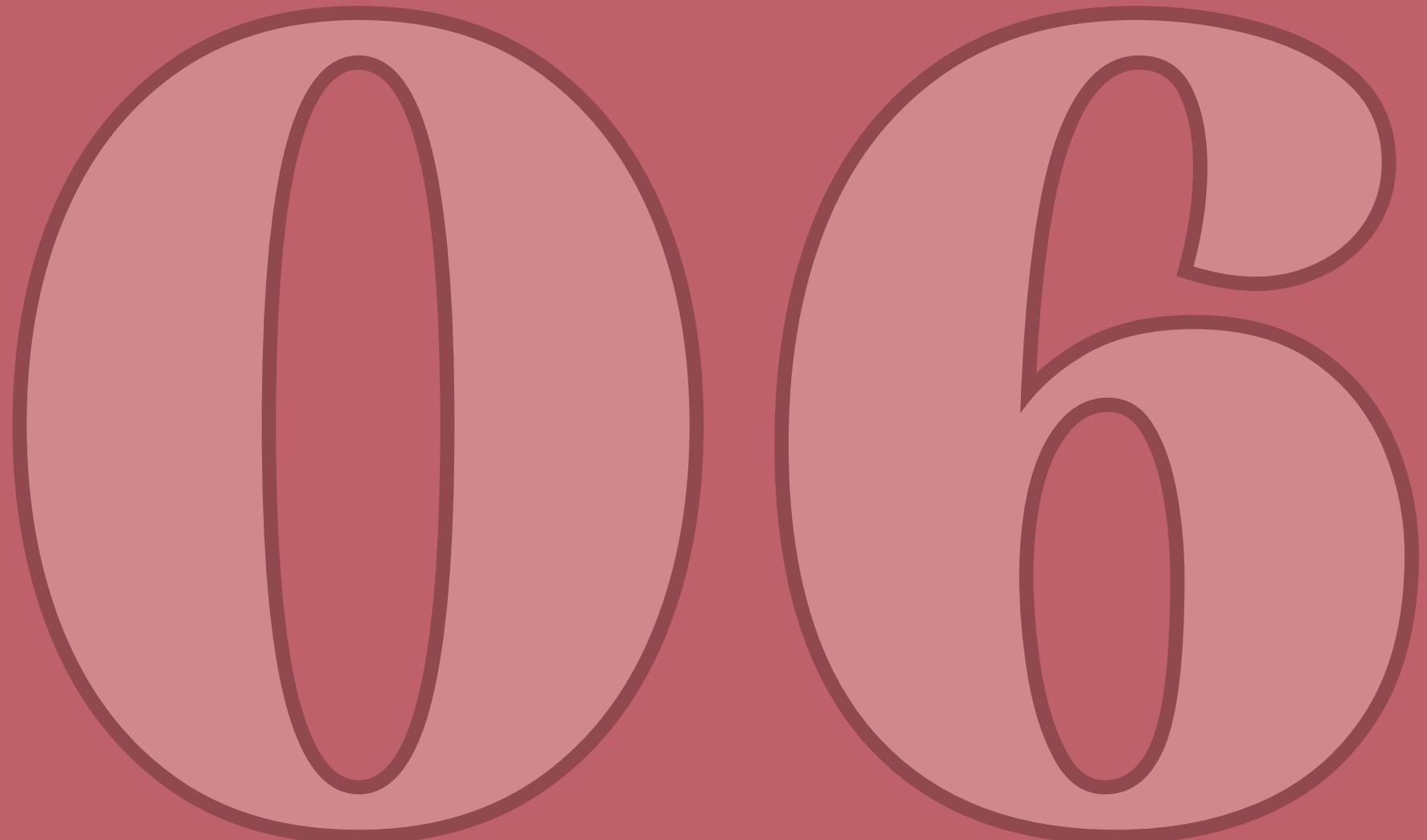
Metriche Retrieval al variare della Score Function

Dataset



Metriche Retrieval al variare del Dataset

Conclusioni e Sviluppi Futuri



Sviluppo e Metodologia

Due pipeline complementari: custom (flessibilità) + BEIR (standardizzazione)

Analisi univariata: variazione un parametro alla volta

Modularità: codice organizzato in funzioni dedicate

Configurazione JSON: tracciabilità completa degli esperimenti

Plotter dedicati: analisi visiva automatica dei risultati

Risultati Pipeline Custom

Embedding: all-miniLM-L6-v2

Chunking: dimensioni del chunk elevati, dimensioni di overlap bassi.

Top-k: 3 documenti sufficienti per generazione ottimale

Temperature: relativamente alta offre risultati migliori.

Max New Token: non viene rispettato da LLM.

Prompt engineering: structured e constraint migliori

Risultati Pipeline BEIR

Embedding: bge-base-en-v1.5 migliore

Score Function: cosine similarity meglio di dot product

Dataset: prestazioni sensibili a tale parametro.

06

Limiti Attuali e Direzioni Future

Limiti: dataset ridotto (10 esempi), modello LLM limitato (SmolLM3-3B)

Mancanza integrazione: FAISS vs exact search

Monitoraggio risorse: assenza metriche CPU/RAM dettagliate

Sviluppi hardware: supporto GPU per FAISS, profiling dettagliato

Configurazioni estese: dataset completi, chunking granulare

Modelli avanzati: API premium per LLM più potenti

Domande?

Grazie mille dell'attenzione!

Ruben-2828/rag-system-evaluation



A lightweight tool for evaluating RAG (Retrieval-Augmented Generation) systems using automated metrics and APIs.

2 Contributors 0 Issues 1 Star 0 Forks

[GitHub](#)

Ruben-2828/rag-system-evaluation: A lightweight tool for evaluating RAG (Retrieval-Augmented Generation)...

A lightweight tool for evaluating RAG (Retrieval-Augmented Generation) systems using automated metrics and APIs. - Ruben-2828/rag-system-evaluation

[GitHub](#)

06