**A Comprehensive Study on Retrieval Augmented Generation Methods for More Robust LLMs**

**1. Introduzione**

**2. Fondamenti di Retrieval-Augmented Generation**

* Come funziona
* Problemi principali da risolvere
* Database vettoriali
* Mitigazione allucinazioni

**3. Metodi di Retrieval e Ottimizzazione**

* Tecniche di chunking: Late Chunking vs Naive Chunking
* Modelli di embedding diversi
* Contextual embedding
* BM25 e TF-IDF per il reranking
* Sistema che combina similarità con BM-25 o TF-IDF per reranking
* Dynamic Similarity Thresholding

**4. Prompt Engineering**

* Nozioni sul prompt engineering
* Stili di prompt
* Valutazione dei prompt rispetto a un golden dataset con DeepEval, usando metriche che sfruttano LLM as a Judge
  + Metriche di Valutazione
    - GEval, DAG, Hallucination, Answer Relevancy e Faithfulness

**5. LLM diversi**

* Valutazione performance in base al modello, con context window diversi

**6. Restituzione chunk rilevanti**

* Confronto chunk con risposta per capire quali sono quelli più rilevanti per la risposta

**7. Esperimenti e risultati**

**8. Conclusioni e lavori futuri**