UNIVERSITA' POLITECNICA DELLE MARCHE

DII – Dipartimento di Ingegneria dell'Informazione

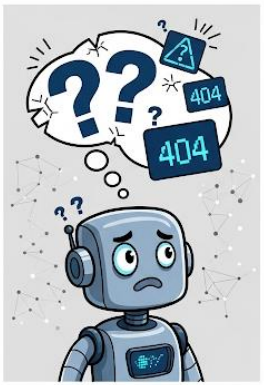
Question Answering through Retrieval-Augmented Large Language Models: an Experimental Evaluation

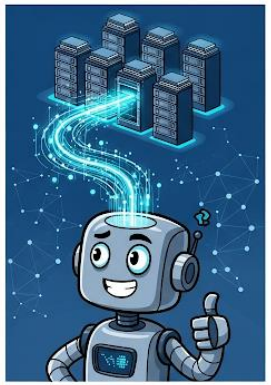
Claudia Diamantini, Alex Mircoli, Alessia Pagnotta,
Domenico Potena, Maria Cristina Recchioni, Emanuele Storti

{c.diamantini, a.mircoli, a.pagnotta, d.potena, c.recchioni, e.storti}@univpm.it

Scenario

- Large Language Models (LLMs) show impressive performance
- Limitations:
  - hallucinations
  - outdated/incomplete knowledge
- Retrieval-Augmented LLMs (RA-LLMs) integrate external knowledge
- Aim: improve accuracy and reliability through Retrieval-Augmented Generation (RAG)

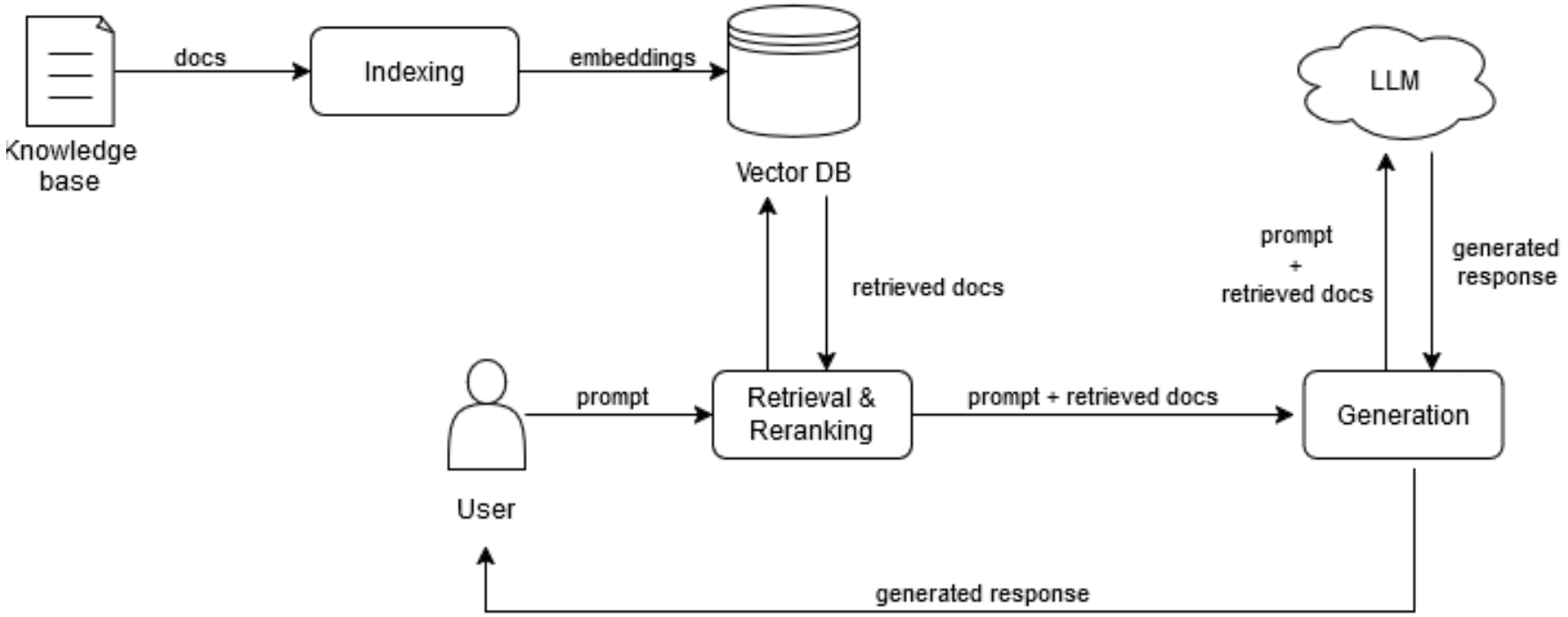




Main contributions

- Definition of a RAG pipeline for Question Answering
- Experimental evaluation of pipeline performance
  - Analysis of different techniques at each phase
  - Evaluation of several performance metrics
- Insights into optimization of RA-LLM architectures

Framework


docs
Indexing
embeddings
Knowledge base
Vector DB
LLM
retrieved docs
prompt + retrieved docs
generated response
prompt
Retrieval & Reranking
prompt + retrieved docs
Generation
User
generated response

Indexing

- **Process** and **store** data belonging to external knowledge bases
- Data are **optimized** for **efficient retrieval** during the following phase
- Main steps:
  - *Chunking*: texts are split into smaller units (chunks)
    - Recursive Character Chunker (RCC)
    - Semantic Chunker (SC)
  - *Embedding generation:* chunks are transformed into embedding vector (OpenAI Embedding API)
  - *Storage*: vector embeddings are stored and indexed into a vector databas

Retrieval & Re-ranking

- **Retrieve** the most relevant set of texts from the knowledge base stored in the vector DB
- User **prompts** are transformed into **embeddings**
  - They are compared with those in the vector DB
- Comparison strategies
  - Semantic search (Euclidean distance)
  - Cosine similarity
  - Hybrid search (Okapi BM25 ranking function + semantic search)

- **Re-ranker**: reorder and filter candidates

Generation

- Use LLMs to generate natural language responses
- Input: user prompt + retrieved documents
- Output quality depends on LLM performance
- Evaluated models:
  - GPT-4o-mini
  - Mistral-7B
  - Mixtral-8x7B
  - Llama-70B

Dataset

- Knowledge base: 74 documents (Italian cuisine, nutrition)
- Prompts: 320 questions across 3 complexity levels
  - Simple questions (100): questions about basic facts explicitly present in the documents.
  - intermediate questions (102): require contextual understanding and the integration of multiple pieces of information
  - multi-document (112): require in-depth analysis, synthesis skills, and critical thinking

Dataset - Examples

Simple questions
- What is Arrabbiata sauce and how is it used in Italian cooking? - What are some common ingredients in Lucanian cuisine? - Tell me about the history of agnolotti and its origin
Intermediate questions
- What are the specific differences in filling and preparation between agnolotti di magro and agnolotti di grasso, and how do these distinctions influence the overall flavor profile of the dishes? - How did the introduction of new ingredients from the Americas in the 18th century impact the development and diversity of Italian cuisine? - What are the key regional variations in the preparation of timballo, particularly regarding the types of ingredients and methods used in different Italian regions?
Multi-document questions
- How do the ingredient choices and preparation techniques in Beef Braciole and Pasta Alla Norma reflect the culinary traditions and regional influences of southern Italy, particularly in the use of local produce and meats? - How does the preparation method of ossobuco, with its detailed cooking process involving vegetables and wine, compare to the traditional cooking techniques used in the preparation of cotechino with lentils, which also involves specific vegetable accompaniments and cooking methods? - How do the traditional cooking methods and ingredient combinations in Roman-style roasted lamb and tagliatelle al ragù reflect the regional culinary practices of Lazio and Bologna, particularly in terms of the significance of local ingredients and historical influences on these dishes?

Evaluation

- Approach: LLM-as-a-judge
  - OpenAI GPT-4o-mini (temperature = 0)

- Metrics:
  - *Contextual Precision*: how well the system prioritizes the most useful documents for a specific query
  - *Contextual Recall*: how completely the system retrieves all necessary information for a comprehensive answer to the query
  - *Contextual Relevancy*: how relevant the retrieved context is with the question that is specified in the prompt
  - *Answer Relevancy*: the proportion of the generated output that is relevant to the given input
  - *Faithfulness*: how consistent the actual output is with the content of the knowledge base

Results - Indexing

- Experimental results

System	Chunking strategy	Embedding
Naïve RAG	Recursive chunker	text-embedding-3-small
Semantic Chunker	Semantic chunker	text-embedding-3-small
OpenAI Large Embedding RAG	Recursive chunker	text-embedding-3-large

Metrics	Naïve RAG	Semantic Chunker RAG	OpenAI Large Embedding RAG
ContextualPrecisionScore	0.826	0.878 (+6.28%)	**0.884 (+7.02%)**
ContextualRecallScore	0.931	**0.960 (+3.09%)**	0.958 (+2.86%)
ContextualRelevancyScore	0.266	0.256 (-3.68%)	**0.276 (+3.69%)**
AnswerRelevancyScore	0.826	0.865 (+4.65%)	**0.921 (+11.51%)**
FaithfullnessScore	**0.954**	0.946 (-0.84%)	0.942 (-1.25%)

- Large embeddings improved:
  - Contextual Precision
  - Answer Relevancy
- • Faithfulness stable across variants

Results – Retrieval & Re-ranking

- Experimental results

Metrics	Naïve RAG	Cosine Similarity	Reranker	Hybrid Search
ContextualPrecision	0.826	0.814 (-1.54%)	**0.882 (+6.78%)**	0.836 (+1.21%)
ContextualRecall	0.931	0.941 (+0.99%)	**0.965 (+3.59%)**	0.956 (+2.67%)
ContextualRelevancy	0.266	0.273 (+2.77%)	0.271 (+2.04%)	**0.292 (+9.69%)**
AnswerRelevancy	0.826	0.839 (+1.52%)	**0.860 (+4.11%)**	0.860 (+4.07%)
Faithfullness	**0.954**	0.952 (-0.19%)	0.947 (-0.72%)	0.941 (-1.36%)

- Cosine similarity similar to Naïve RAG (Euclidean distance)
- Best performance with **Reranker** and **Hybrid Search**
- Hybrid search improved Contextual Relevancy (+9.7%)
  - small drop in Faithfulness

Results - Generation

- Experimental results

Metrics	GPT-4o-mini	Mistral 7B RAG	Mistral 7x8B RAG	Llama 70B RAG
AnswerRelevancyScore	0.826	0.697 (-15.59%)	0.510 (-38.33%)	0.751 (-9.13%)
FaithfullnessScore	0.954	0.895 (-6.13%)	0.901 (-5.51%)	0.901 (-5.55%)

- GPT-4o-mini outperforms open-source models
- Llama-70B best among open-source models
- Open-source models struggle with complex queries

Conclusion & future work

- Evaluation of the main phases of a RAG system
  - Indexing: semantic chunking and large embeddings improve retrieval
  - Retrieval: re-ranking and hybrid search enhance relevancy
  - Generation: Closed-source LLMs best for generation
    - Size matters

- Future work:
  - Larger datasets
  - Optimization of chunk size/overlap
  - Additional LLMs (Claude, DeepSeek)