

Artificial Intelligence
2024/2025



RAG System for University Information Assistance

Pallini Vincenzo - 907303

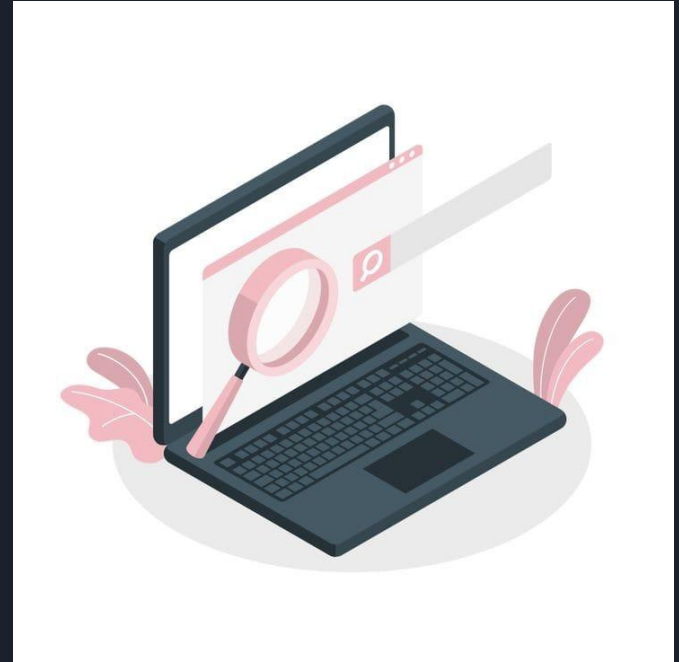
Challenges in Accessing University Information

- Fragmented information
- Accessibility barriers
- Lack of interactivity



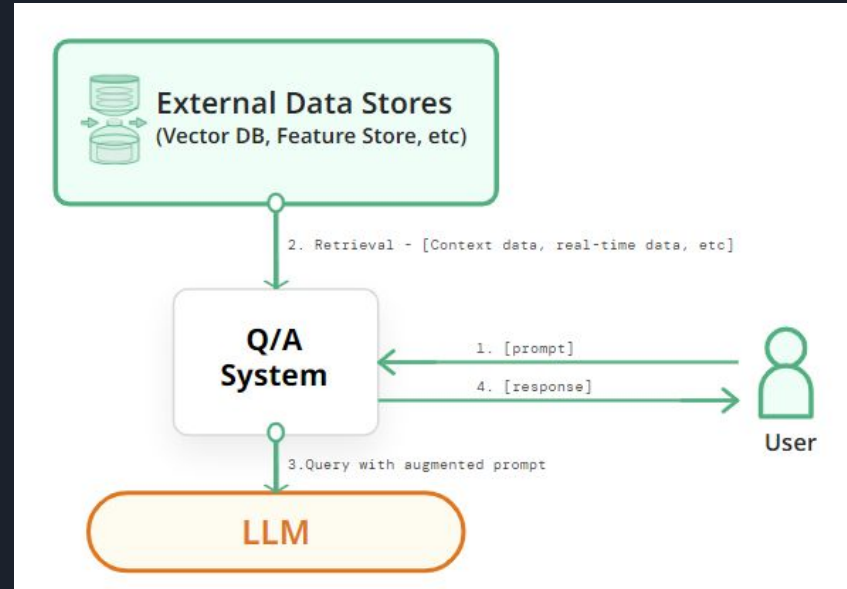
Benefits of the RAG System for University Assistance

- Quick access to documents
- Generation of contextualised answers
- Simplification of information

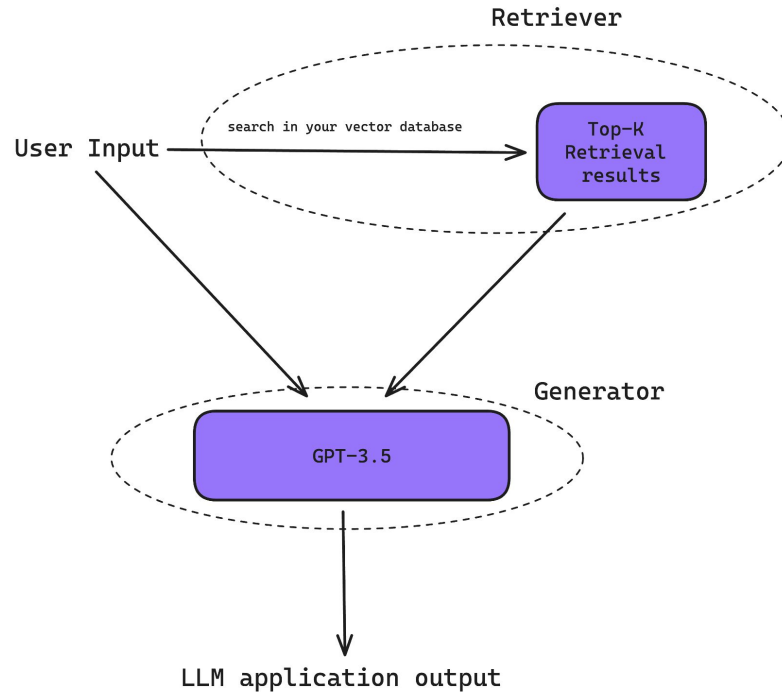


RAG System: Architecture Overview

- External Data Management
- Retrieval process
- Processing with LLM

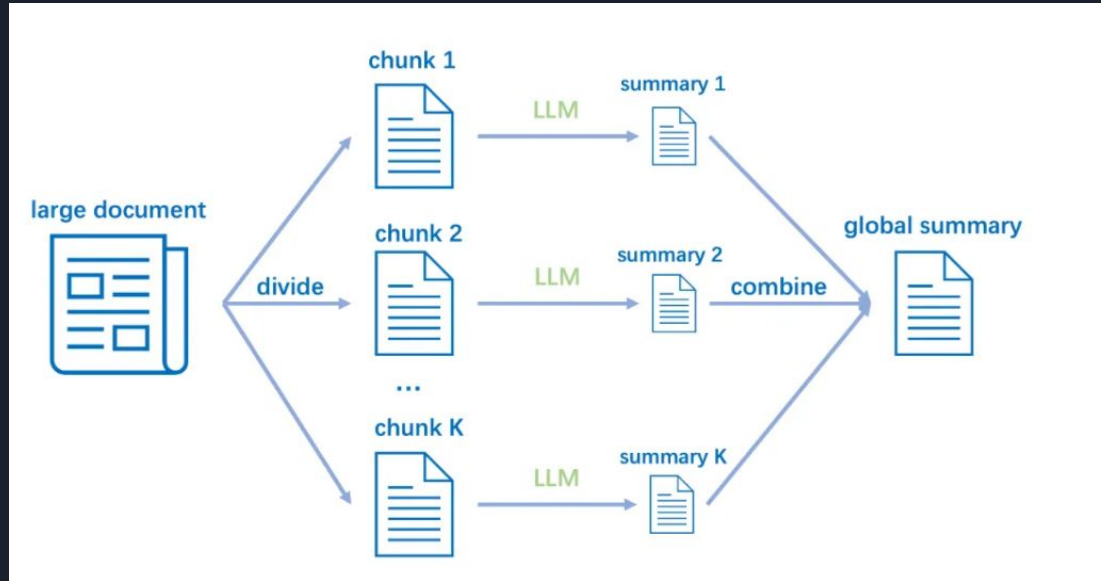


Components of the RAG system

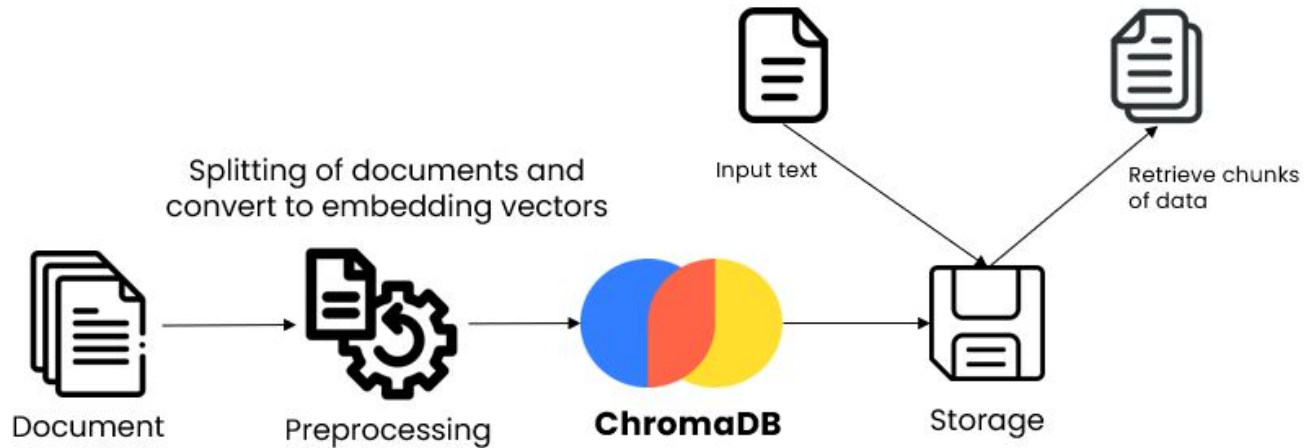


Methodology - Overview

- Database population
- Extraction and chunking
- Query and generation with LLM



ChromaDB: Document Processing Pipeline



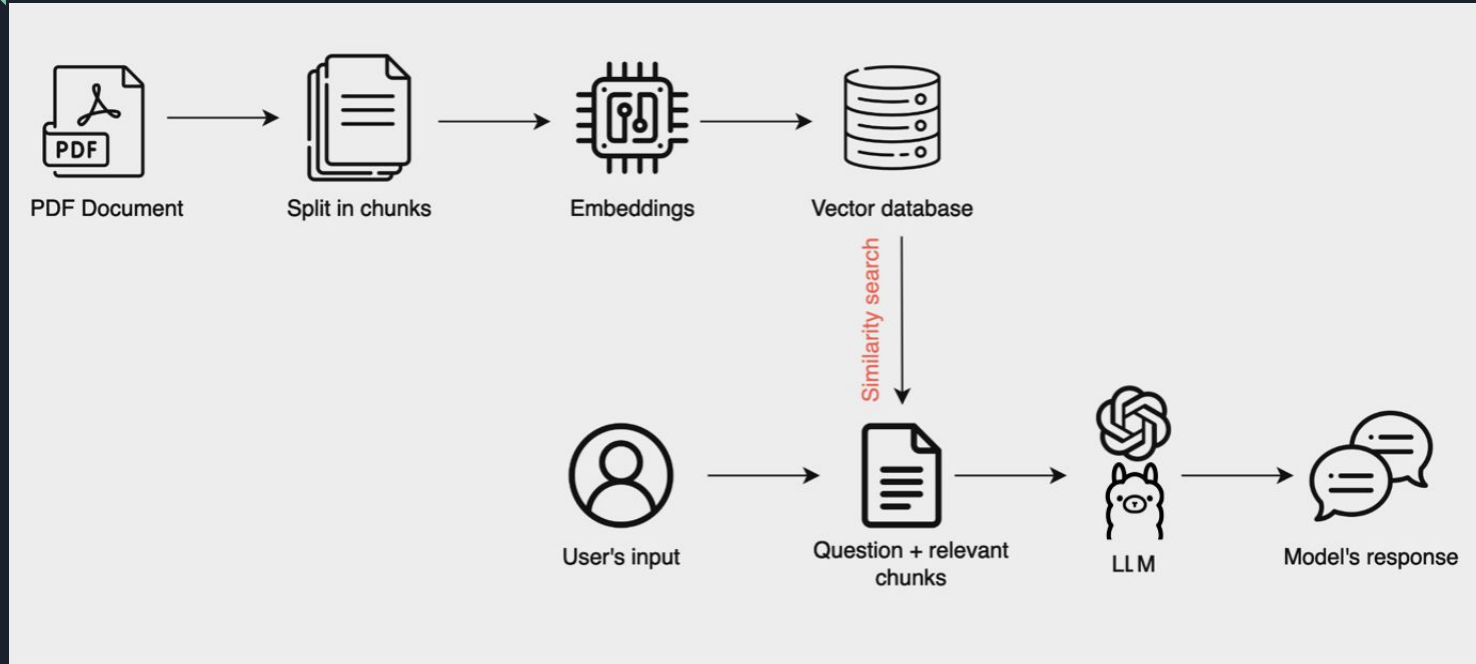
LangChain Framework– as orchestrator

Is a framework for developing applications powered by language models(LLMs). It enables applications that:

- Are context-aware: connect a language model to sources of context
- Reason: rely on a language model to reason (about how to answer based on provided context).



Complete RAG System Architecture



Performance Analysis: Custom RAG System vs ChatGPT

Tu: Quanto è il totale di tasse da pagare per il corso area A sotto ISEE < 27.500 Meritevoli per quest'anno

Bot: Per calcolare il totale delle tasse da pagare per un corso nell'Area A con ISEE inferiore a 27.500 e lo studente rientra nella categoria dei meritevoli, seguiamo questi passaggi:

1. ****Determinazione del Contributo Unico:****
 - Per gli ISEE fino a 27.000,00 €, il contributo unico è di € 0,00.
 - Dal calcolo per ISEE compresi tra 27.000,01 e 28.000,00:
\\
\\text{Contributo} = (ISEE/1000)^2 \\times 0,5937 \\times 0.87 \\times R
\\]
- Dove \\(R = [(ISEE - 27.000) / 1.000] \\backslash).
2. ****Calcolo per ISEE < 27.500:****
 - Supponiamo un ISEE di 27.400,00 € (per esempio):
\\
R = (27.400 - 27.000) / 1.000 = 0,4
\\]
\\text{Contributo} = (27.4/1000)^2 \\times 0,5937 \\times 0.87 \\times 0,4
\\]
\\text{Contributo} = 0,00075076 \\times 0,5937 \\times 0.87 \\times 0,4
\\]
\\text{Contributo} = 0,0001545
\\]
3. ****Contributo Minimo:****
 - Il contributo minimo è di € 200,00 se il calcolo risulta inferiore.

Quindi, per ISEE < 27.500, lo studente dovrebbe pagare almeno € 200,00 come contributo unico poiché questo valore supera quello calcolato.

4. ****Prima Rata Accollo:****
 - L'importo della prima rata acconto è di € 156,00 (comprende l'imposta di bollo e la tassa regionale).
5. ****Totale delle Tasse per Quest'Anno:****
 - Totale = Contributo Unico + Prima Rata Accollo
 - Totale = € 200,00 + € 156,00 = € 356,00

Pertanto, **il totale delle tasse da pagare per l'anno in questione è di **€ 356,00**.**

Fonti: ['data\\Guida TASSE 24_25.pdf:2:1', 'data\\Guida TASSE 24_25.pdf:1:2', 'data\\Guida TASSE 24_25.pdf:1:1', 'data\\Guida TASSE 24_25.pdf:9:5', 'data\\Guida TASSE 24_25.pdf:1:0']

Performance Analysis: Custom RAG System vs ChatGPT

ChatGPT ▾

Condividi



Guida TASSE 24_25.pdf
PDF

Quanto è il totale di tasse da pagare per il corso area A sotto ISEE
< 27.500 Meritevoli per quest'anno

< 3/3 >

Per il corso dell'Area A, con ISEE inferiore a 27.500€, gli studenti meritevoli **non pagano alcuna tassa**.
L'importo totale da pagare è 0€ .



Scrivi un messaggio a ChatGPT



Cerca



Avvia il ragionamento



ChatGPT può commettere errori. Considera di verificare le informazioni importanti.





Evaluation: Automated QA Generation

Question Generation (QA):

We use a scheme to instruct the LLM model to generate 1 QA per text chunk.

Ground Truth creation:

An answer (Ground Truth) is generated for each question based solely on the extracted context.

Internal Dataset for Evaluation:

The generated QA pairs constitute the dataset for evaluating the performance of the RAG system.

RAGelo Evaluation System: Performance Metrics

Metrics:

- Quality
- Trustworthiness
- Originality

Additional Metric:

- Fact score



Accuracy Results: Gemma 2 vs Phi-4

Gemma 2: 2b

Accuracy: 55.21%



Phi 4: 14b

Accuracy: 85.86%





Future Developments

- Upgrade to a more powerful LLM for improved accuracy
- Access to complete university data for comprehensive answers
- Integration with existing university platforms
- Advanced RAG Architectures



Thank you!
for your attention



Bibliography

[1]: Retrieval-Augmented Generation for AI-Generated Content: A Survey. Authors: Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, Bin Cui. Link: <https://arxiv.org/abs/2402.19473>

[2]: Evaluating RAG-Fusion with RAGElo: an Automated Elo-based Framework. Authors: Zackary Rackauckas and Arthur Câmara and Jakub Zavrel. Link: <https://arxiv.org/abs/2406.14783>

[3]: LangChain: <https://python.langchain.com/docs/introduction/>

[4]: Chroma: <https://docs.trychroma.com/docs/overview/introduction>