

Using Local LLMs and RAG for Exploring Research Papers

WORKSHOP ON EMPIRICAL RESEARCH IN THE AI ERA | BPLIM

Nelson Areal

nareal@eeg.uminho.pt

School of Economics and Management / University of Minho

December 16, 2024



Plan for the presentation

- What is RAG?
- Basic components
- Open source tools for each component
- Should I build one?
- Example
- Ready made solutions



What is RAG?

What is RAG?

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). Advances in Neural Information Processing Systems, 33, 9459-9474.



What is RAG?

What is the topic of this year's BPLIM workshop?



OI LiteLLM.gemini-2.0-flash-exp

Unfortunately, I don't have access to real-time information, including specific details about workshops like the BPLIM workshop.

To find the topic of this year's BPLIM workshop, I recommend the following:

- Check the official BPLIM website:** Look for a dedicated page for the workshop, often found under sections like "Events," "Conferences," or "Workshops." This page

OI LiteLLM.claude-3-5-sonnet

I don't have information about this year's BPLIM workshop or its topic. BPLIM (Banco de Portugal Microdata Research Laboratory) does host workshops, but without more specific context about which year you're referring to, I cannot tell you the topic of their current workshop.



OI OpenAI.gpt-4o

I'm sorry, but I don't have information on the specific topic of this year's BPLIM workshop. You may want to check the official BPLIM or related organizational websites or announcements for the most current details.



What is the topic of this year's BPLIM workshop?



What is RAG?

What is the topic of the 2024 BPLIM workshop?

< 2/2 >

OI LiteLLM.gemini-2.0-flash-exp

10:06 PM

The 2024 BPLIM workshop focused on the topic of "**Large Language Models for Multi-Modal Data**".

This means the workshop explored the use of Large Language Models (LLMs) in handling data that goes beyond just text, incorporating other modalities such as images, audio, video, and more.



OI LiteLLM.claude-3-5-sonnet

10:06 PM

I don't have information about the specific topic of the 2024 BPLIM workshop. Without being able to verify the details of this event, I cannot make any claims about its topic or content.



OI OpenAI.gpt-4o 10:06 PM

I'm sorry, but I don't have information on the specific topic of the 2024 BPLIM workshop. You might want to check the official website or contact the organizers for the most accurate and up-to-date information.



What is the topic of the 2024 BPLIM workshop?



What is RAG?

- What is the topic of the 2024 BPLIM workshop?
- Who is presenting?

Answer the questions using the information on <https://www.bportugal.pt/en/evento/workshop-empirical-research-ai-era-bplim>

Ol LiteLLM.gemini-2.0-flash-exp

Okay, I've reviewed the information on the provided link. Here are the answers to your questions:

- **What is the topic of the 2024 BPLIM workshop?**

The topic of the 2024 BPLIM workshop is "Empirical Research in the AI Era".

- **Who is presenting?**

The presenters at the workshop are:

- **Professor Pedro Pita Barros** (Nova School of Business and Economics)
- **Professor Paulo Pinho** (University of Porto)

Ol LiteLLM.claude-3-5-sonnet

According to the website:

The topic of the 2024 BPLIM workshop is "Empirical Research in the AI Era". The workshop will be held on June 7, 2024.

The presenters include:

- Matt Taddy (Amazon) as the keynote speaker
- Gary Koop (University of Strathclyde)
- Manasa Patnam (IMF)
- Additional speaker slots will be filled through a call for papers (submission deadline: March 31, 2024)

The workshop will focus on how artificial intelligence and machine learning are transforming empirical

Ol OpenAI.gpt-4o

I'm unable to browse the internet in real-time or access external websites. Please visit the provided URL to find detailed information on the 2024 BPLIM workshop, including the topic and the list of presenters.



With URL as context...



What is RAG?

- What is the topic of the 2024 BPLIM workshop?
- Who is presenting?

Answer the questions using the information below:

Banco de Portugal, through its Microdata Research Laboratory (BPLIM), is organizing the Workshop on Empirical Research in the AI Era, in collaboration with the Academia of Data Science.

The Workshop will take place on December 16-17, 2024, at Porto Business School.

This is an on-site event with limited seating, but it will also be streamed live, allowing unlimited virtual participation.

Register here, until December 13.

Programme

16 December 2024

14

:00

Welcome

14

:15

Leveraging Generative AI in Research and Policy Analysis

Jesse Lastunen, UNU-WIDER

15

:00

Uses of Generative AI for Economics Research

Kevin Bryan, Lab University of Toronto

15



With text as context...



What is RAG?

Ol LiteLLM.gemini-2.0-flash-exp

10:17 PM

- **Topic of the 2024 BPLIM workshop:** The topic of the workshop is **Empirical Research in the AI Era.**
- **Presenters:** The presenters are:
 - Jesse Lastunen, UNU-WIDER
 - Kevin Bryan, Lab University of Toronto
 - Sérgio Correia, Board of Governors of the Federal Reserve System
 - Nelson Areal, Universidade do Minho
 - Michael Dowling, DCU Business School, Dublin City University
 - Ricardo Mora, Universidad Carlos III Madrid
 - Peter Gruber, USI Università della Svizzera Italiana
 - Ulrich Matter, Bern University of Applied Sciences, and University of St. Gallen
 - Lars Vilhuber (Cornell)

Ol LiteLLM.claude-3-5-sonnet

10:17 PM

The topic of the 2024 BPLIM workshop is "Empirical Research in the AI Era."

The presenters are:

- Jesse Lastunen (UNU-WIDER)
- Kevin Bryan (Lab University of Toronto)
- Sérgio Correia (Board of Governors of the Federal Reserve System)
- Nelson Areal (Universidade do Minho)
- Michael Dowling (DCU Business School, Dublin City University)
- Ricardo Mora (Universidad Carlos III Madrid)
- Peter Gruber (USI Università della Svizzera Italiana)
- Ulrich Matter (Bern University of Applied Sciences, and University of St. Gallen)
- Lars Vilhuber (Cornell)

Ol OpenAI.gpt-4o 10:17 PM

The topic of the 2024 BPLIM workshop is "Empirical Research in the AI Era."

The presenters at the workshop are:

- Jesse Lastunen, UNU-WIDER
- Kevin Bryan, Lab University of Toronto
- Sérgio Correia, Board of Governors of the Federal Reserve System
- Nelson Areal, Universidade do Minho
- Michael Dowling, DCU Business School, Dublin City University
- Ricardo Mora, Universidad Carlos III Madrid
- Peter Gruber, USI Università della Svizzera Italiana
- Ulrich Matter, Bern University of Applied Sciences and University of St. Gallen
- Lars Vilhuber, Cornell

With text as context...



What is RAG?

Ol llama3.2:latest 10:23 PM

Based on the provided information, here are the answers to your questions:

1. The topic of the 2024 BPLIM workshop is "Empirical Research in the AI Era".
2. The speakers presenting at the workshop are:
 - Jesse Lastunen (Leveraging Generative AI in Research and Policy Analysis)
 - Kevin Bryan (Uses of Generative AI for Economics Research)
 - Sérgio Correia (Unlocking Economic Data with LLM)
 - Nelson Areal (Using Local LLMs and RAG for Exploring Research Papers)
 - Michael Dowling (Financial Agents for Research and Practice)
 - Ricardo Mora (Integrating AI in Stata Programming: Perspectives on Enhancement and Constraints Across Skill Levels)
 - Peter Gruber (How Large Language Models Support Statistical Analysis)
 - Ulrich Matter (Using Large Language Models for Text-As-Data Studies in the Social Sciences)
 - Lars Vilhuber (Reproducibility and AI: Potential, Challenges, and Other Preliminary Thoughts)

With text as context...



What is RAG?

What is the topic of the 2024 BPLIM workshop?

Sources

- Registration for BPLIM... • 1 bplim
- Workshops | Banco d... • 2 bplim
- Registration for BPLIM... • 3 msites-dee-bplim...

View 12 more

Answer

Workshop Topic: Empirical Research in the AI Era

The Banco de Portugal Microdata Research Laboratory (BPLIM) is organizing a workshop on "Empirical Research in the AI Era" in collaboration with the Academia of Data Science. This event aims to explore the intersection of artificial intelligence (AI) and empirical research, focusing on its applications and implications for policy analysis and decision-making.

According to source 3, the workshop will take place on December 16-17, 2024, at Porto Business School in Porto, Portugal. The program includes presentations from experts in AI and data science, discussing topics such as leveraging generative AI in research and policy analysis (Source: 11).

Using Perplexica

Perplexica: Perplexica is an AI-powered search engine. It is an Open source alternative to Perplexity AI



Advantages of using RAG

- Use updated information
- Enhances accuracy
- Reduces alucinations

as the model responds to questions **grounded in current and relevant information**, and even **smaller models** can perform well these tasks.



Advantages of using RAG

By using **local models** you can do it **locally**,
privately, with **no costs**, with **greater
reproducibility**.



Basic components

RAG basic components

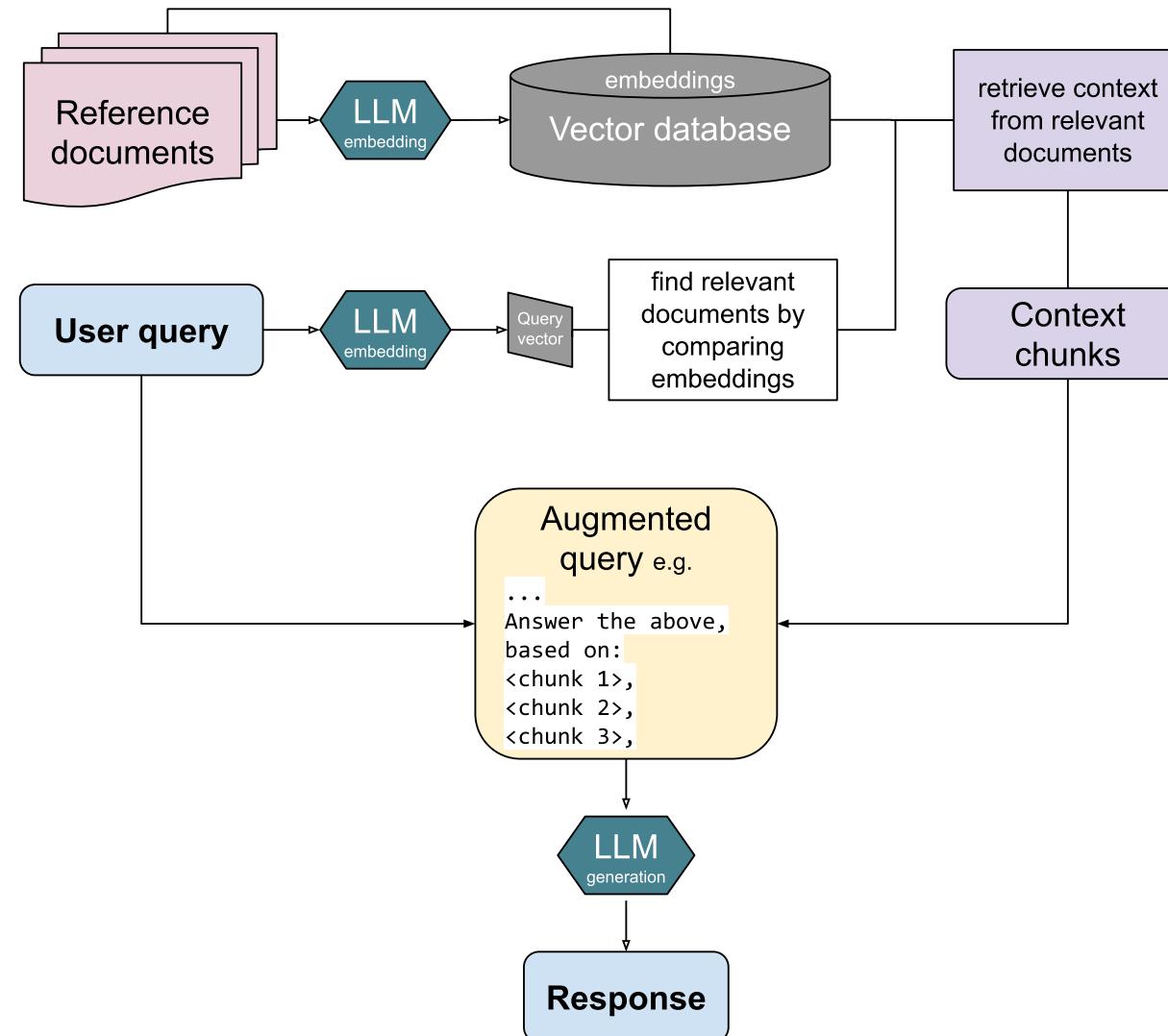
Embeddings are numerical representations (vectors) of words or tokens in a continuous vector space, where similar words are mapped to nearby points.

Think of them as transforming words into lists of numbers that capture semantic relationships.

[What are embeddings? | Vicki Boykis](#)



RAG basic components

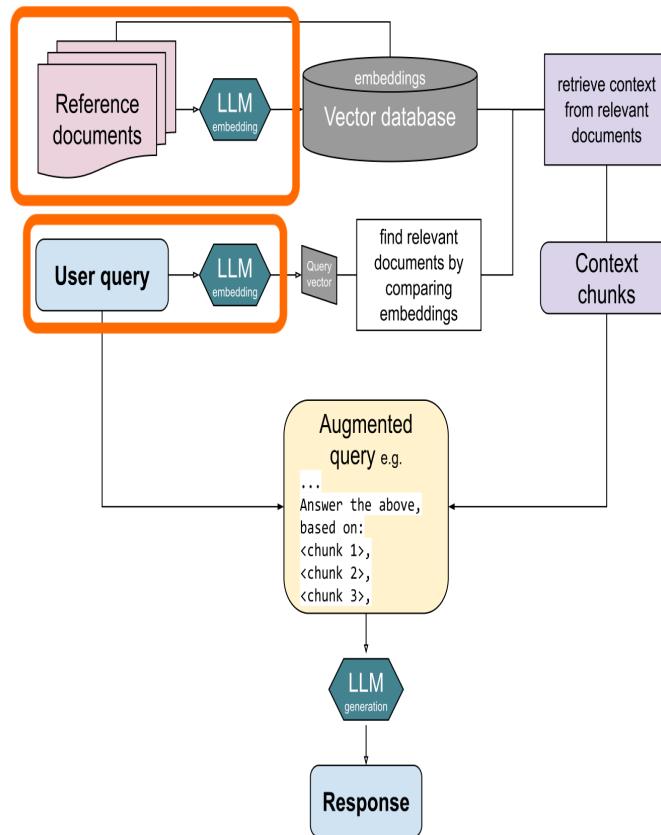


By Turtlecrown - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=150390279>

Open source tools for each component

Open source tools for each component

Convert the PDF to text



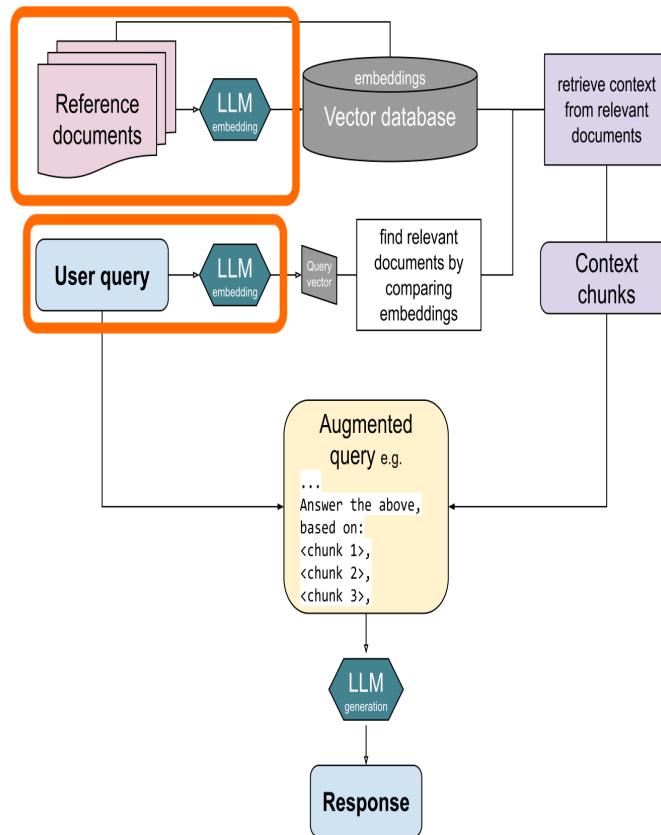
- MuPDF
- Doclign | IBM

By Turtlecrown - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=150390279>



Open source tools for each component

Embeddings



- **mxbai-embed-large-v1**: 1024 dimensions; 512 context size; 334M parameters;
- **nomic-embed-text-v1**: 768 dimensions; 8192 context size; 137M parameters;

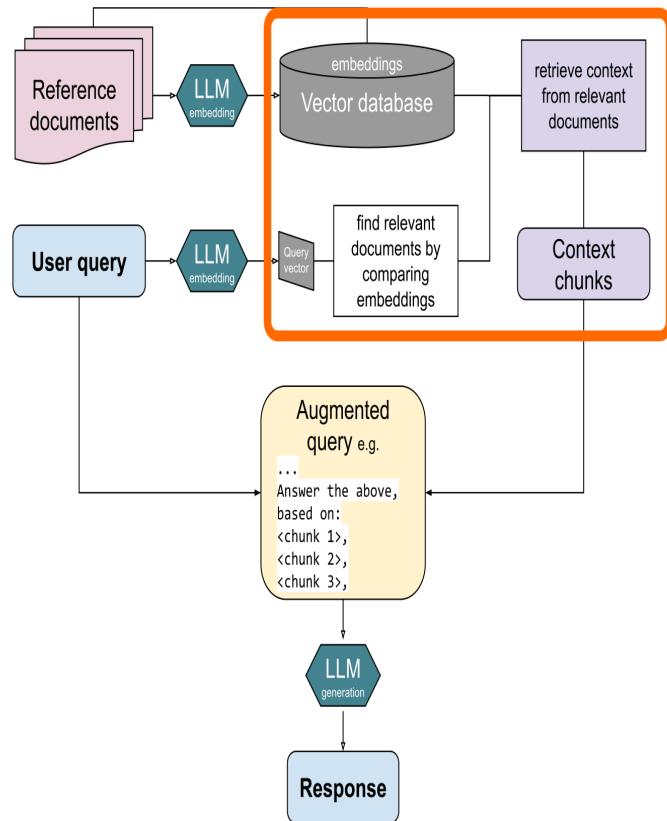
Binary and Scalar Embedding Quantization
for Significantly Faster & Cheaper Retrieval

By Turtlecrown - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=150390279>



Open source tools for each component

Database



- PostgreSQL

- Elastic

Vector databases:

- milvus
- Weaviate
- Qdrant

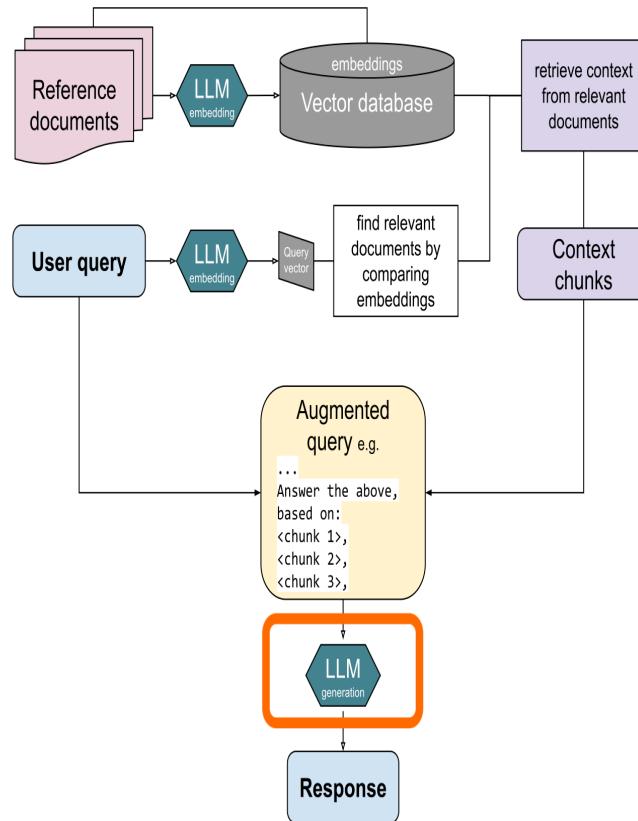
No server databases:

- SQLite
- DuckDB
- LanceDB

By Turtlecrown - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=150390279>

Open source tools for each component

LLM



LLM inference:

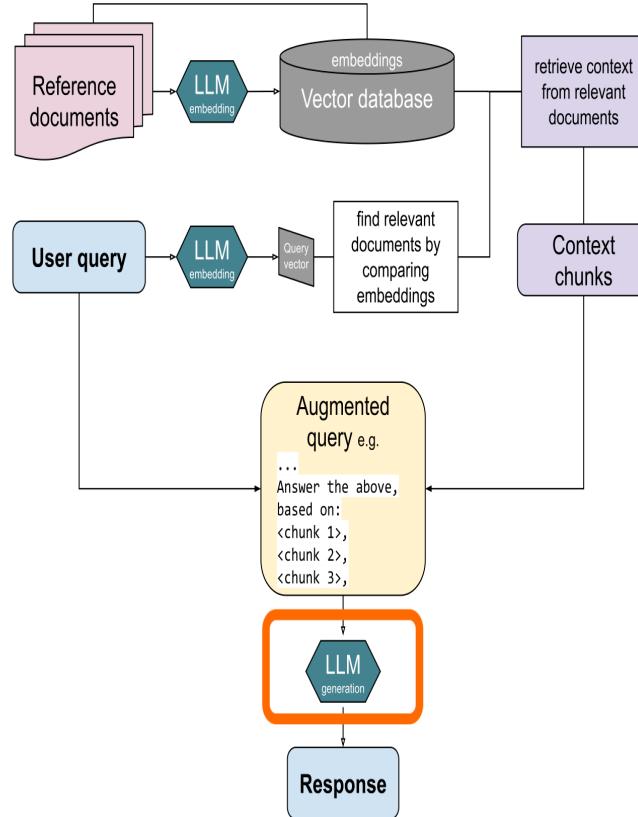
- Ollama
- Ilama.cpp: LLM inference in C/C++
- Ilamafile: Distribute and run LLMs with a single file

By Turtlecrown - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=150390279>



Open source tools for each component

LLM



LLM models:

- OLMo 2 | Ai2
- Meta's Llama models: 3.1(8B, 405B), 3.2(1B, 3B) and 3.3 (70B)
- Microsoft's Phi-3 (3.8B, 14B);
- Mistral 7B;
- Google's Gemma 2 (2B, 9B, 27B);
- Alibaba's Qwen2.5 (0.5B, 1.5B, 3B, 7B, 14B 32B, 72B)

By Turtlecrown - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=150390279>





Should I build
one?



Should I build one?



- Create a custom made tool for your **own needs**
- Don't want to be dependent on third party services
- Opportunity to learn more about LLMs and AI tools



- Don't want to waste time
- Don't want to learn a new things
- I don't know how to program

Example

Ready made solutions

Ready made solutions

- GPT4All
- AnythingLLM
- weaviate/Verba
- n8n
- RAGFlow
- Haystack

...and **many** other commercial applications

