
Retrieval-Augmented Generation (RAG)

Going Beyond the LLM

Thaís de Araújo de Medeiros
thais.araujo.707@ufrn.edu.br

WHO AM I

THAÍS MEDEIROS

Member of the **Conect2ai**
Research Group | Projects:
CNPq and **Rota2030**



**Technician in
Textiles**
IFRN



**Master's Degree and Doctoral
Candidate**
Graduate Program in Electrical and
Computer Engineering
PPgEEC | UFRN



Computer Engineer
UFRN



**Bachelor of Science
and Technology**
UFRN

AGENDA



Introduction to RAG

Architecture & Key
Components

System Building

Hands-On

Which Brazilian actress won the Golden Globe in 2025?



I don't know which Brazilian actress won the Golden Globe in 2025 because my training data goes only up to **June 2024**. Any events that happened after that date, including the 2025 ceremony, are outside of my knowledge. Without searching the web, I cannot provide a confirmed answer.



Which Brazilian actress won the Golden Globe in 2025?

Context:

Brazil woke up to a historic achievement: Fernanda Torres became the first Brazilian actress to win the Golden Globe in 2025. She was awarded for her performance in *I'm Still Here (Ainda Estou Aqui)*, Globoplay's first original film, in the category of Best Actress in a Motion Picture – Drama.

At the ceremony in Hollywood, Walter Salles celebrated the recognition and highlighted the film's impact on Brazilian audiences. Based on the book by Marcelo Rubens Paiva, the feature film portrays the struggle of Eunice Paiva to find her husband, Rubens Paiva, who disappeared during the military dictatorship.



Fernanda Torres won the Golden Globe in 2025, becoming the first Brazilian actress to receive the award. She earned it for her performance in *I'm Still Here (Ainda Estou Aqui)*, Globoplay's first original film, in the category of Best Actress in a Motion Picture – Drama.



Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis^{†‡}, Ethan Perez^{*},

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

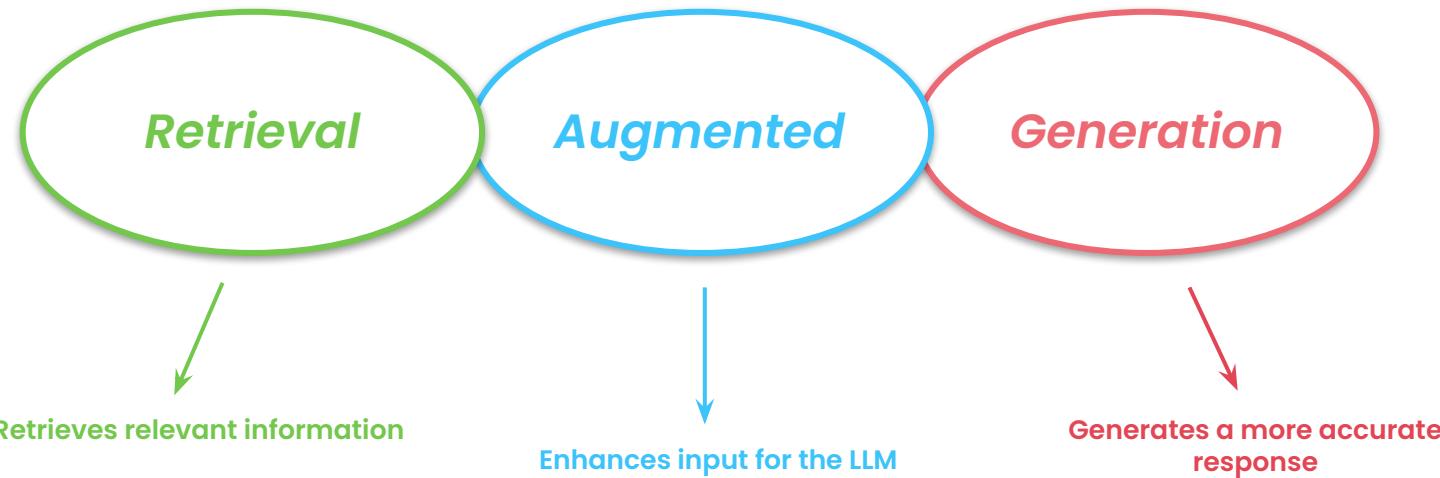
Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel^{†‡}, Sebastian Riedel^{†‡}, Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; ^{*}New York University;
plewis@fb.com

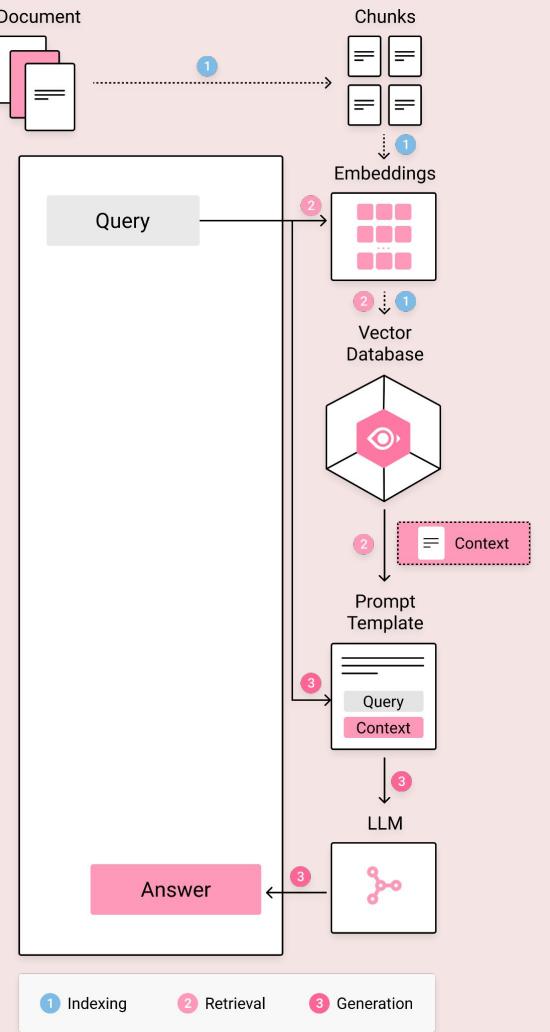
Abstract

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) — models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, and another which can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state of the art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

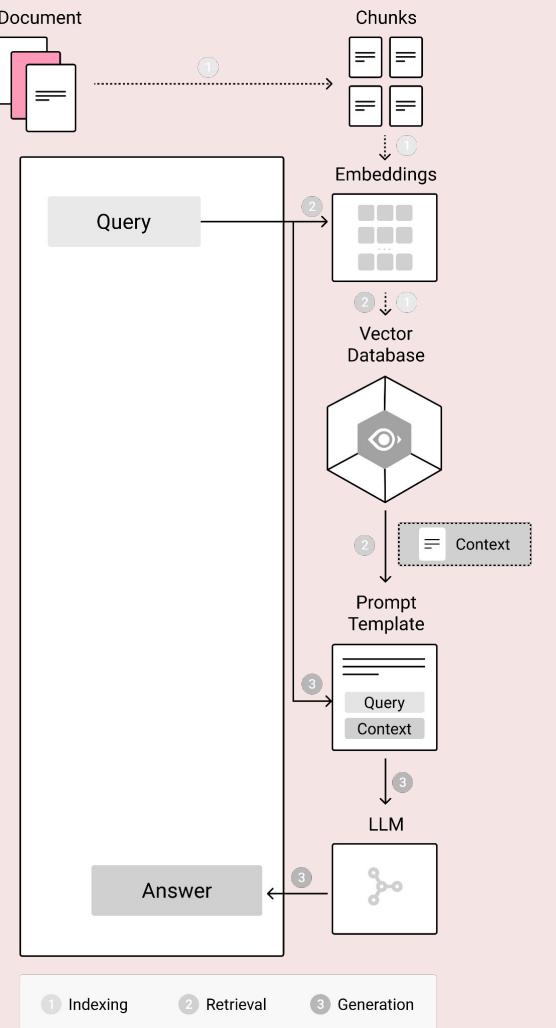
CONCEPT



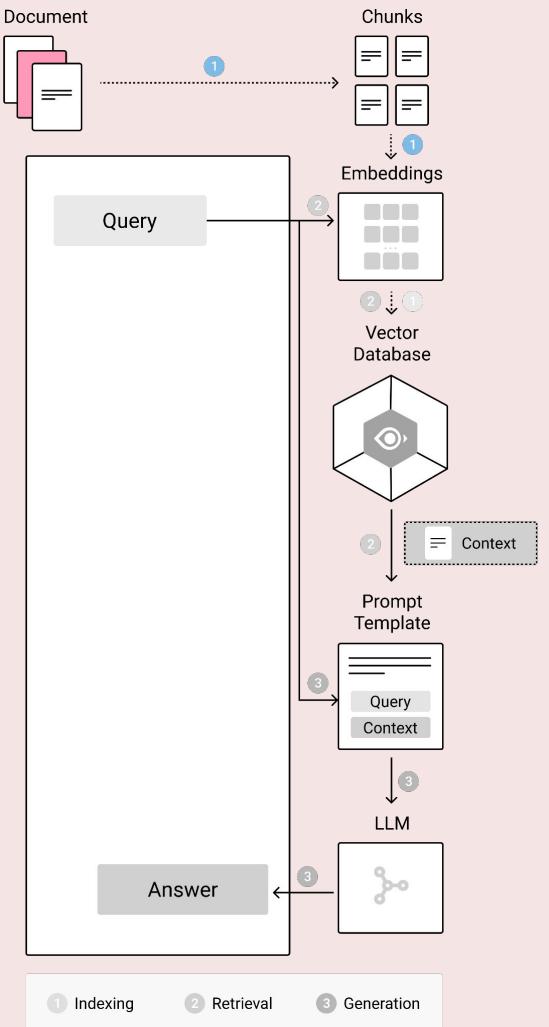
A TYPICAL RAG SETUP



TYPES OF DOCUMENTS



CHUNKING



The air filter is a component responsible for preventing impurities from entering the engine.

chunk 1

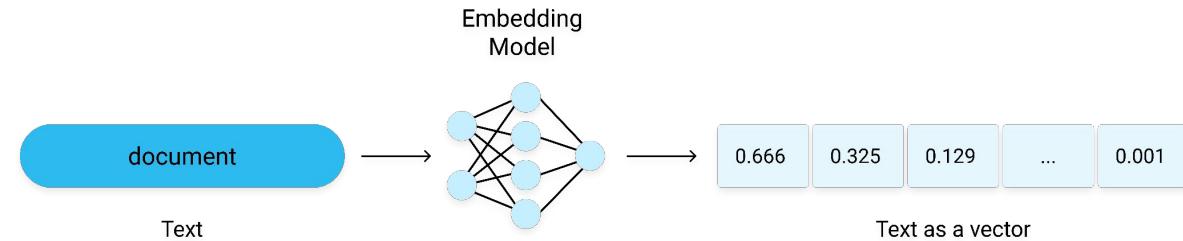
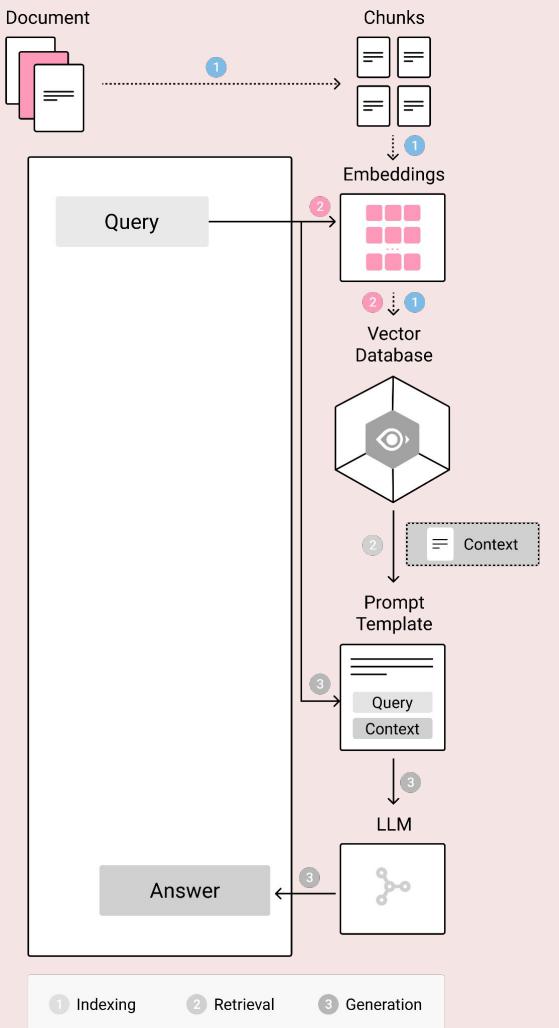
overlap

chunk 2

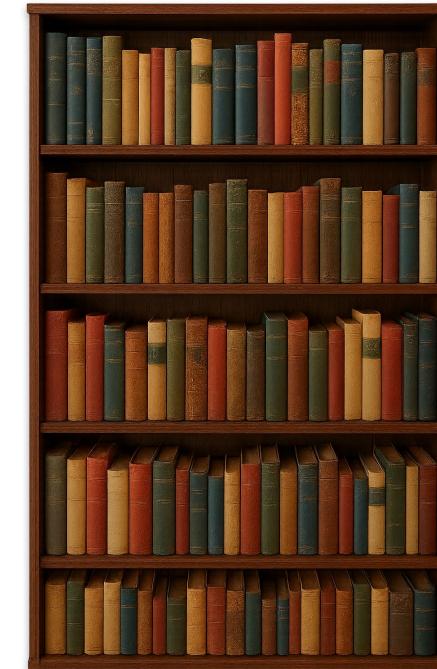
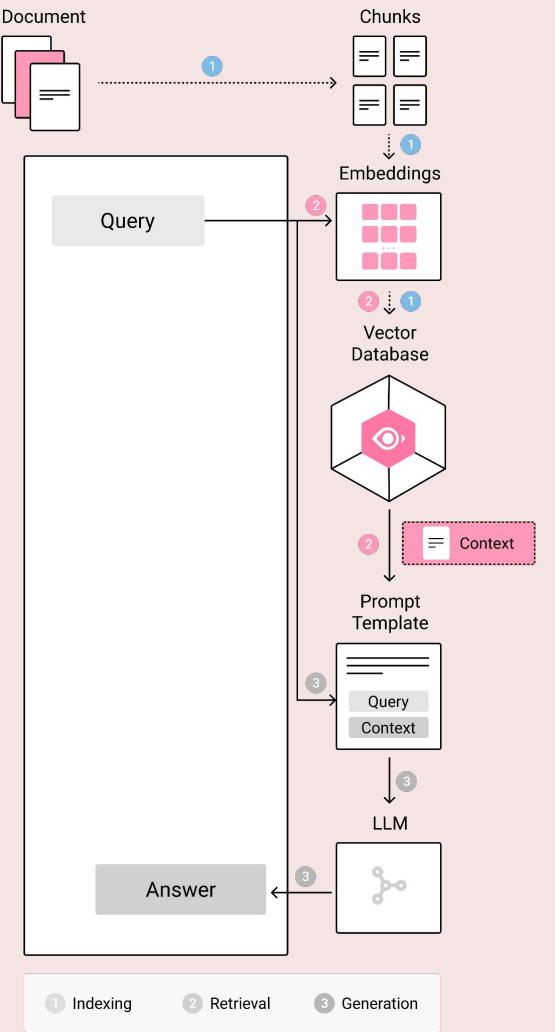
overlap

chunk 3

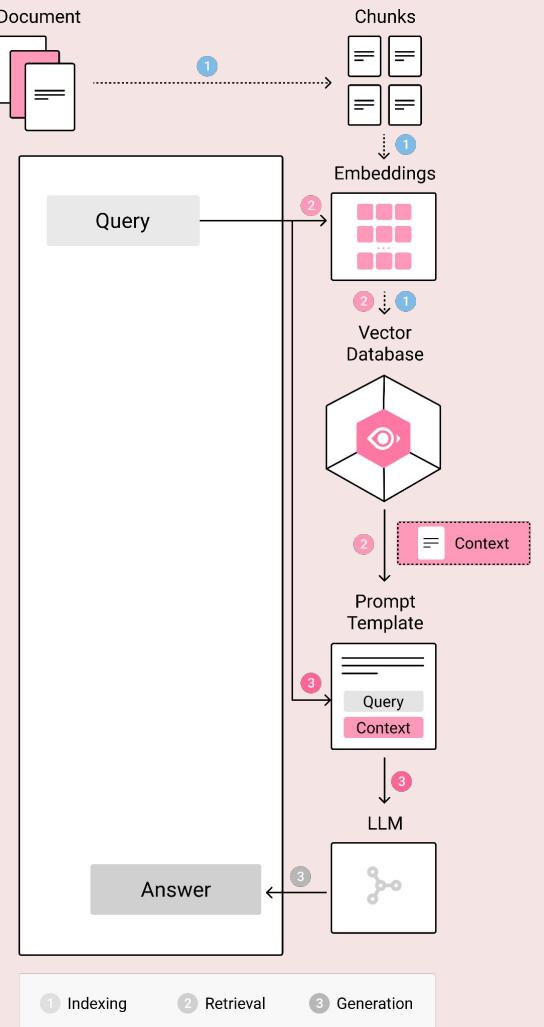
EMBEDDING



VECTOR DATABASE



PROMPT TEMPLATE



You're a friendly customer support. You know this:

<EXTERNAL KNOWLEDGE>



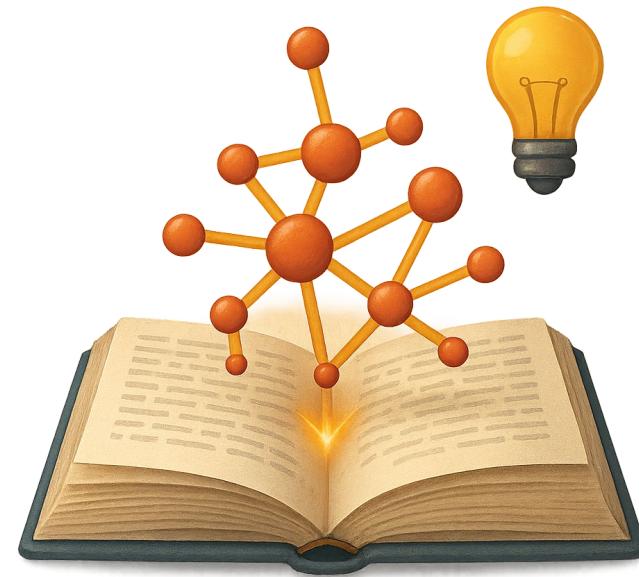
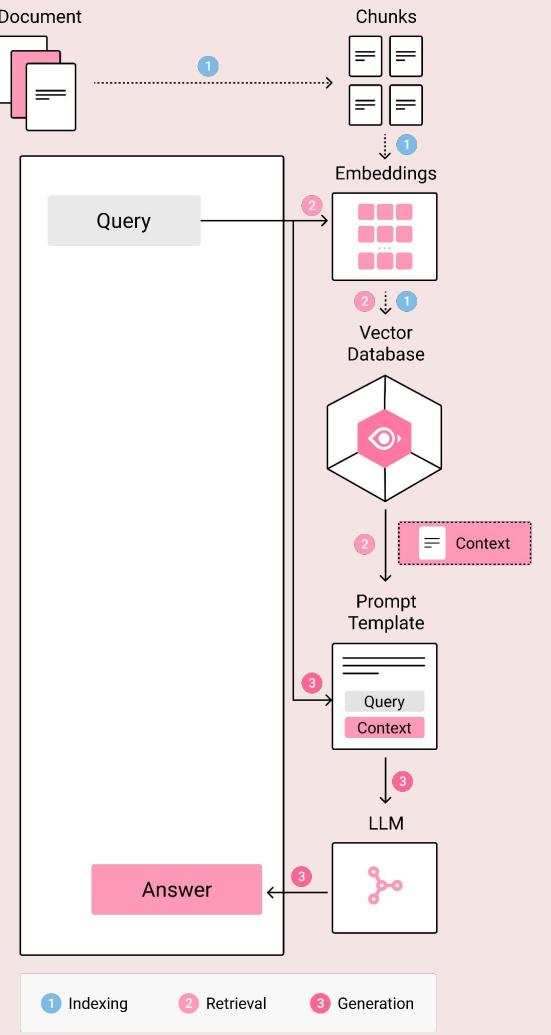
A customer asked this:

<USER QUERY>

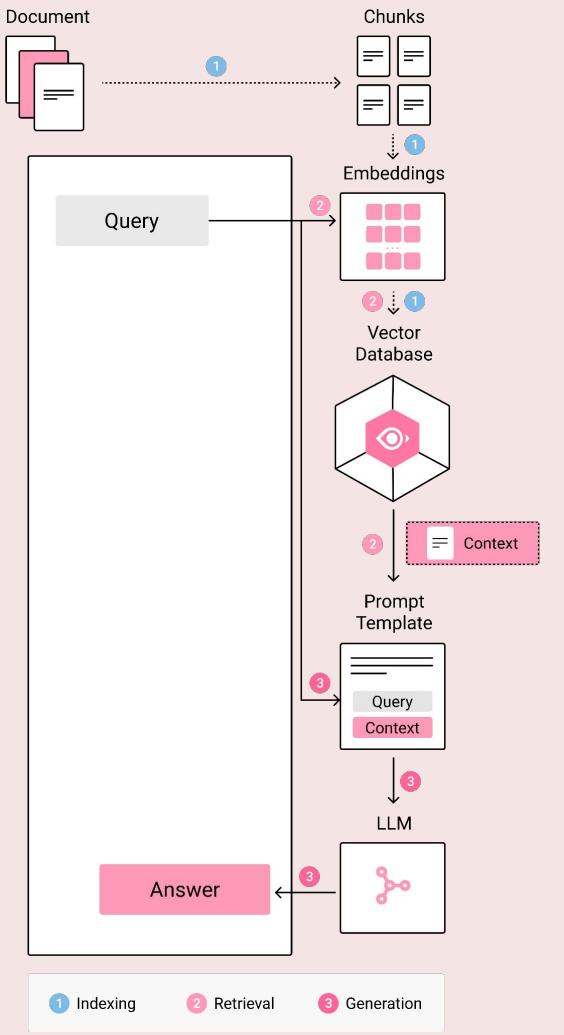


Please generate an answer for the customer.

LARGE LANGUAGE MODEL



LARGE LANGUAGE MODEL



🤗 Hugging Face

🦙 Ollama

🔗 OpenAI
⭐ Gemini

HOW TO BUILD?

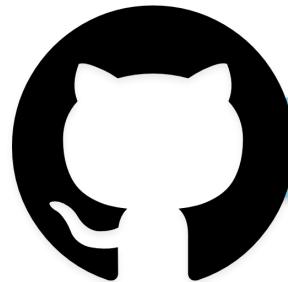


LlamaIndex



LangChain

REPOSITÓRIO



<https://github.com/conect2ai/RAG-Bootcamp>

HANDS-ON