

# Introduction to RAG

This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# Limitations of Current Generative Models

- **Lack of Up-to-Date Information:** If you ask a generative model about recent events, it may provide outdated answers or refuse to respond.
- **Hallucinations:** Generative models sometimes produce plausible-sounding but incorrect or unverifiable information.

# Challenges in Retrieval-Based Systems

- **Lack of Synthesis:** They return multiple relevant sources but do not summarize or interpret the data cohesively.
- **User Burden:** Users must manually review and extract insights from retrieved content.

# What is RAG?

---

- **Information Retrieval Systems**
- **Generative Models**

# RAG's Strengths: Why it Stands out

---

- **Contextual Accuracy**
- **Factual Grounding**
- **Scalability**
- **Customizability**

# Real-World Applications of RAG

---

- **Conversational Agents (Customer Support)**
- **Healthcare Support Systems**
- **Education and Learning Platforms**

# Transformative Potential of RAG

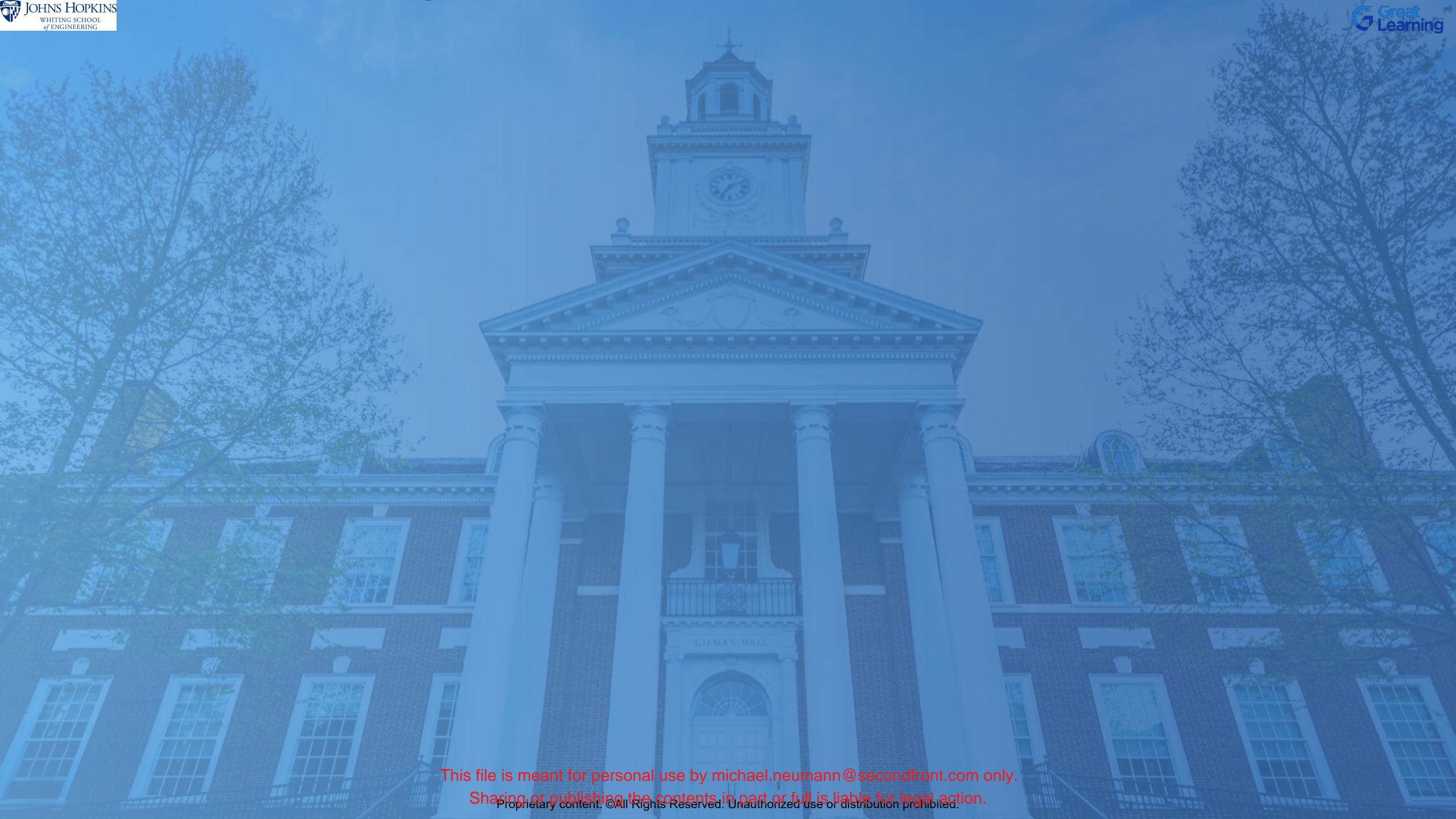
---

- **Bridging the Knowledge Gap**
- **Enhancing Decision-Making**
- **Personalization**

This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.



This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# RAG vs. Traditional Search

This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# Traditional Search

popular polish shoe brands

Business Insider  
[https://www.businessinsider.com › Reviews › Style](https://www.businessinsider.com/reviews/style/the-best-shoe-polishes-for-shining-leather-shoes)

The 6 best shoe polishes for shining leather shoes

Jan 29, 2024 — Our top picks for the best shoe polishes in 2024 · Best conditioner: Saphir Renovateur · Best wax: Saphir Medaille d'Or Pate de Luxe · Best color: ...

Ranking Shoe Polishes (19 BEST & WORST Brands) ft ...  
YouTube · Gentleman's Gazette 1 month ago

What is the Best Shoe Polish? | Shoe Polish Review  
YouTube · Kirby Allison

- Traditional search lacks context.
- Google has algorithms to correct.

# Generative Search

list some popular polish shoe brands

- Generative search has context to improve the quality of search.
- Avoids "keyword" based SEO optimization.
- More accurate and relevant.



Here are some popular Polish shoe brands:

1. **Wojas** - Known for high-quality leather shoes and a wide range of styles for both men and women.
2. **Gino Rossi** - Offers elegant and fashionable shoes, often with a focus on sophisticated designs.
3. **Badura** - Renowned for its stylish and comfortable footwear, including both casual and formal options.
4. **CCC** - A well-known brand offering a variety of shoe styles, including many popular international brands.
5. **Ryłko** - Specializes in comfortable and stylish shoes made from high-quality materials.
6. **Kazar** - Known for its luxurious and trendy shoe collections, often featuring unique designs.
7. **Venezia** - Offers a wide range of fashionable shoes, blending classic and contemporary styles.
8. **Lasocki** - Provides affordable yet stylish shoes, often focusing on comfort and durability.
9. **Bartek** - Specializes in children's footwear, known for its quality and ergonomic designs.
10. **Conhpol** - Offers high-quality men's shoes, combining traditional craftsmanship with modern styles.



# Generative Search

**Generative AI** can provide sourcing for information.  
Search for "Which planet in the solar system has the most moons?"



Space.com

<https://www.space.com> > The Universe > Solar System

Jupiter now has the most moons in the solar system ...

Feb 3, 2023 — Jupiter now has the most moons in the solar system, beating Saturn thanks to 12 newfound satellites.



KiwiCo

<https://www.kiwico.com> > blog > stem > which-planet-h...

Which Planet Has the Most Moons?

Jul 4, 2023 — Within our solar system, Saturn stands tall as the planet with the most moons, boasting an incredible count of 145.



The Planets Today

<https://www.theplanetstoday.com> > which\_planet\_has\_th...

Which Planet is has the most moons?

Which planet has the most moons? Uranus as 27 moons, Neptune has 14 moons, Mars has 2 moons, Earth has 1 moon, Venus and Mercury have no moons.



As of now, Jupiter holds the title for the planet with the most moons in the solar system. Jupiter has 95 confirmed moons. This number is constantly updated as new moons are discovered and confirmed by astronomers.

## References

- NASA. (2023). Jupiter Overview. Retrieved from <https://solarsystem.nasa.gov/planets/jupiter/overview/>



As of now, Saturn has 83 confirmed moons. This number is constantly updated as new moons are discovered and confirmed by astronomers.

## References

- NASA. (2023). Saturn Overview. Retrieved from <https://solarsystem.nasa.gov/planets/saturn/overview/>

Why did you say 83 confirmed moons when the NASA website you referenced says there are 146 moons?



Memory updated

You are correct. Saturn has 146 moons, with 83 of them confirmed and 63 still awaiting confirmation. Here is the corrected information:

# Generative Search

- After correction, subsequent searches may be updated.

which planet in the solar system has the most moons? Include references in APA format.



As of now, Saturn holds the title for the planet with the most moons in the solar system. Saturn has a total of 146 moons, with 83 confirmed and 63 awaiting confirmation.

## References

- NASA. (2023). Saturn Overview. Retrieved from <https://solarsystem.nasa.gov/planets/saturn/overview/>



This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# Bridging Retrieval and Generation in AI-Powered NLP

This file is meant for personal use by [michael.neumann@secondfront.com](mailto:michael.neumann@secondfront.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# The Era before RAG

---

- **Retrieval-Only Models:**
  - BM25, TF-IDF, and embedding-based retrieval.
  
- **Generative-Only Models:**
  - GPT-3, GPT-3.5, GPT-4, and other LLMs.

# What makes RAG different?

---

- **Live and up-to-date knowledge integration.**
- **More context-aware and accurate responses.**
- **A reduction in hallucinated outputs.**

# What makes RAG different?

---

- **Live and up-to-date knowledge integration.**
- **More context-aware and accurate responses.**
- **A reduction in hallucinated outputs.**
- **RAG uses retrieved documents dynamically to inform its generation, resulting in more grounded and fact-based responses.**

# BERT (Retrieval-Oriented)

---

- **Bidirectional Contextual Understanding:**

"The bank is on the river"

# BERT (Retrieval-Oriented)

- **Bidirectional Contextual Understanding:**

"The bank is on the river"

- **Strengths:**

- Document ranking, NER, and question answering.
- Retrieving relevant information.

- **Weaknesses:**

- Cannot generate fluent responses.
- Lacks creativity and coherence.

# GPT (Generation-Oriented)

- **Unidirectional Language Modeling:**
  - Process text left-to-right.

# GPT (Generation-Oriented)

- **Unidirectional Language Modeling:**

- - Process text left-to-right.

- **Strengths:**

- - Highly fluent.
  - Excels in content creation.

- **Weaknesses:**

- - Lacks real-time retrieval.
  - Prone to hallucinations.

# RAG (A Unified Approach)

---

- **Uses retrieval (BERT-like)**
- **Uses generation (GPT-like)**

# RAG (A Unified Approach)

- **Uses retrieval (BERT-like)**
- **Uses generation (GPT-like)**
- **Strengths:**
  - Retrieves live and up-to-date data.
  - Generates coherent responses.
  - Handles complex and context-dependent questions.

# Advantages of RAG

- **RAG bridges the gap by retrieving AND summarizing.**

**Query:** "What happens in a breach of contract?"

- **Retrieval-Only Output:**

- "Here are five articles on breach of contract." (*User has to read all of them.*)

- **RAG Output:**

- "A breach of contract occurs when one party fails to fulfill their obligations.

Common causes include non-performance or delayed performance, as outlined in these legal sources."

*(A synthesized and meaningful answer.)*

# Advantages of RAG

- **RAG grounds responses in real-world retrieved data.**

**Query:** "Does intent matter in a breach of contract?"

- **GPT Output:**

- "A breach of contract only occurs when the contract is intentionally broken." (*Incorrect — intent is not always required.*)

- **RAG Output:**

- "A breach of contract can occur when either party fails to perform their duties, regardless of intent, as specified under contract law and confirmed by legal standards."  
*(Accurate, well-sourced, and fact-based.)*

# The Future with RAG

---

- **Bridging the Knowledge Gap**
- **Improved Decision-Making**
- **Personalization**



This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# Retrieval-Augmented Generation

This file is meant for personal use by [michael.neumann@secondfront.com](mailto:michael.neumann@secondfront.com) only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# Agenda

---

1. Refresher on Components of RAG
2. Simple RAG Code
3. RAGAs

This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# Why RAG?



“The Hello World of Large Language Models” — Pedro Rodriguez

<https://www.solulab.com/rag-app-development-and-its-applications-in-ai/>

This file is meant for personal use by michael.neumann@secondfront.com only.

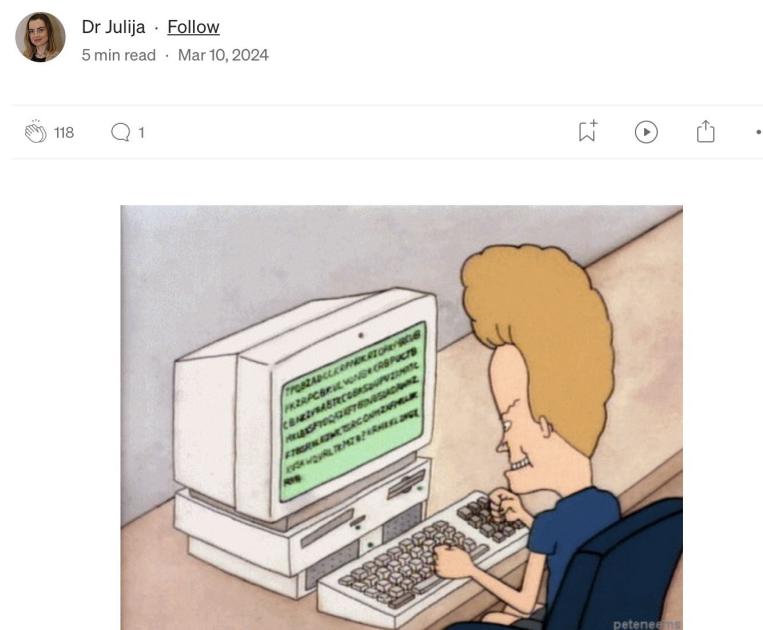
Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# Code a Simple RAG from Scratch

## Use OpenAI (External)

**How I built a Basic RAG for PDF QA in a few lines of python code**



<https://medium.com/@drjulija/how-i-built-a-basic-rag-for-pdf-qa-in-a-few-lines-of-python-code-9849c32e59f0>

<https://medium.com/@danushidk507/rag-with-llama-using-ollama-a-deep-dive-into-retrieval-augmented-generation-c58b9a1cfcd3>

## Completely Local

**Code a simple RAG from scratch**

Community Article Published October 29, 2024



ngxson  
Xuan-Son Nguyen



**RAG with LLaMA Using Ollama: A Deep Dive into Retrieval-Augmented Generation**

DhanushKumar · Follow  
5 min read · Nov 30, 2024

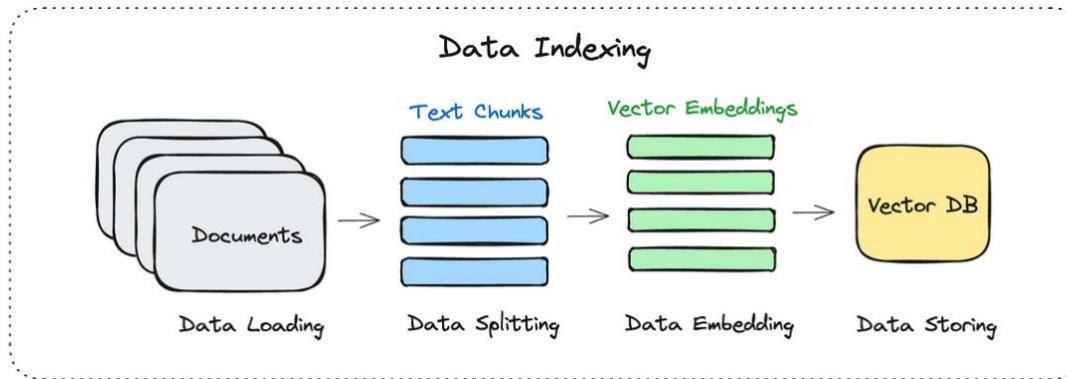
This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

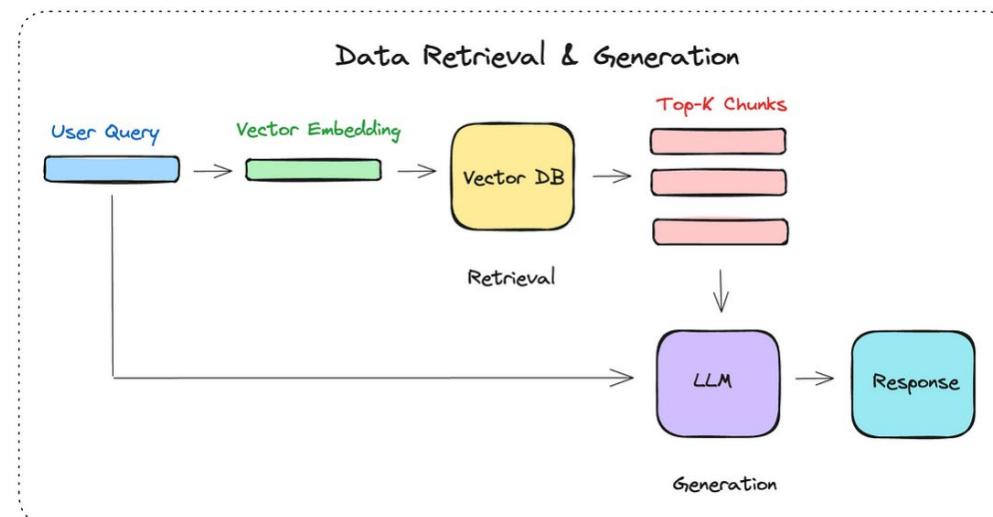
Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# What are we doing today?

Basic RAG Pipeline



Data Retrieval & Generation



This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.

# RAGAs

- RAGAs (**R**etrieval-**A**ugmented **G**eneration **A**sessment) is a framework ([GitHub](#), [Docs](#)) that provides you with the necessary ingredients to help you evaluate your RAG pipeline on a component level.
- **Retriever Component:** It retrieves additional context from an external database for the LLM to answer the query.
- **Generator Component:** It generates an answer based on a prompt augmented with the retrieved information.

<https://medium.com/data-science/evaluating-rag-applications-with-ragas-81d67b0ee31a>

## Evaluating RAG Applications with RAGAs

A framework with metrics and LLM-generated data to evaluate the performance of your Retrieval-Augmented Generation pipeline

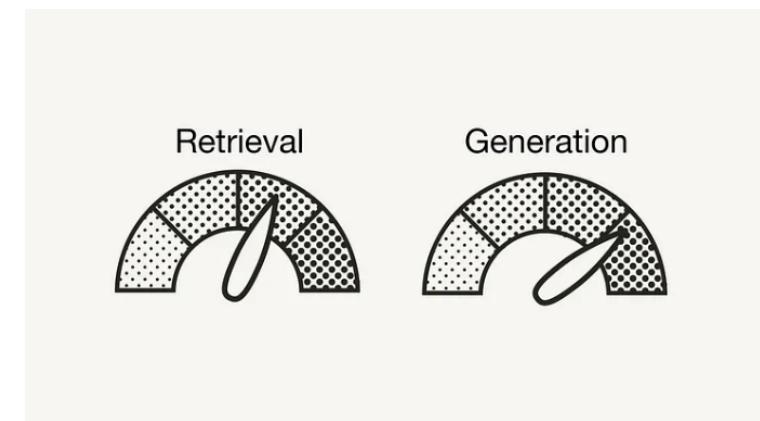


Leonie Monigatti · [Follow](#)

Published in TDS Archive · 8 min read · Dec 13, 2023

1.6K

Q 9





This file is meant for personal use by michael.neumann@secondfront.com only.

Sharing or publishing the contents in part or full is liable for legal action.

Proprietary Content. © All Rights Reserved. Unauthorized use or distribution prohibited.