

# Retrieval-Augmented Generator (RAG)

Instructor

Apoorv Vishnoi

Head - Training Vertical at Analytics Vidhya



# LLM Limitations



- 1 Trained on vast data but **has a cutoff**.
- 2 Lacks knowledge of **recent events** and **niche topics**.

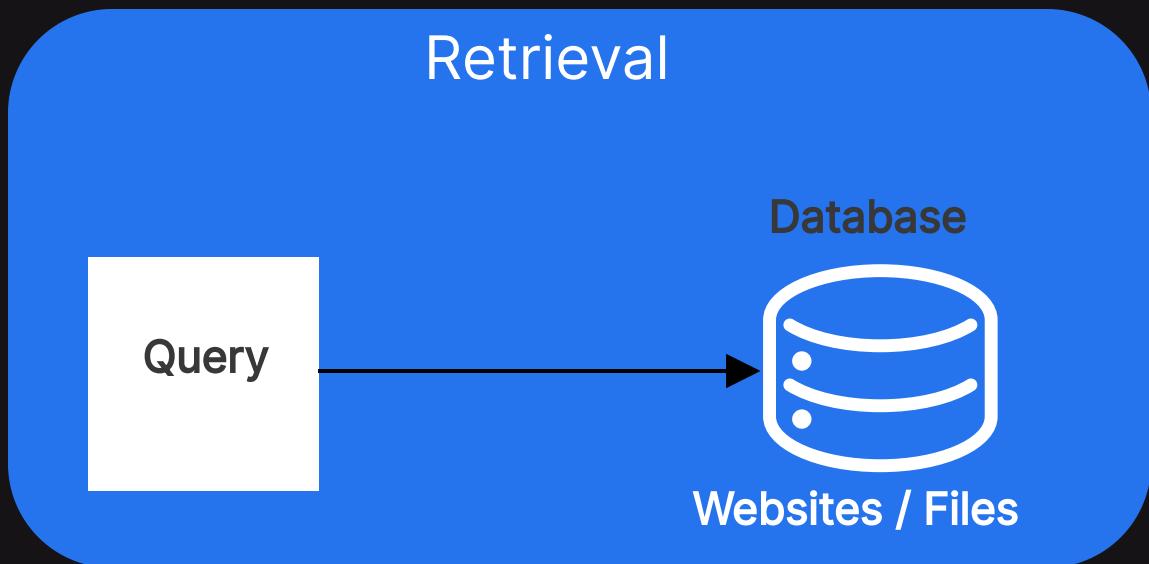
RAG

# RAG

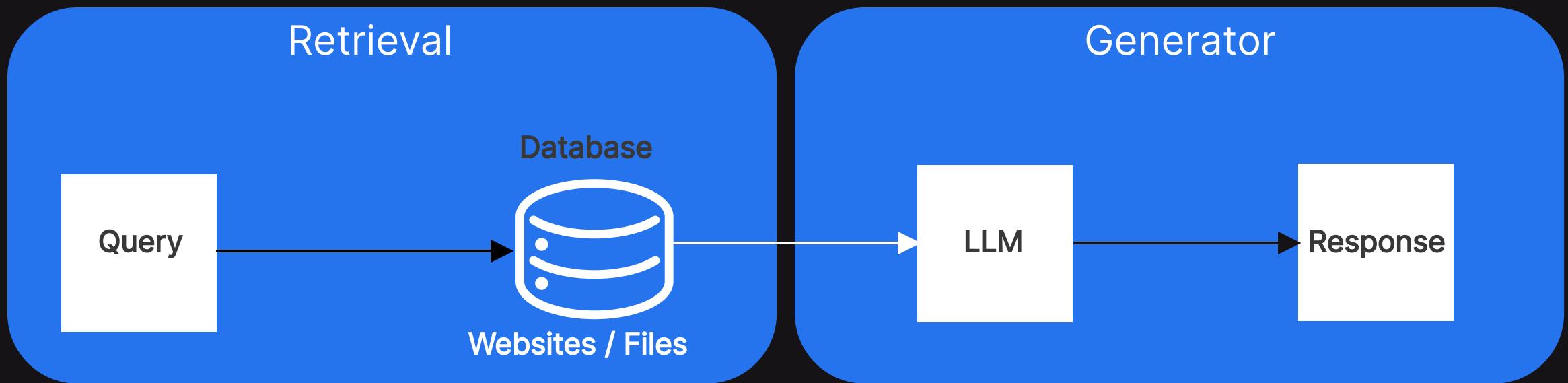


RAG enhances LLMs by retrieving information from databases, improving accuracy on current events and domain-specific queries.

# Components of RAG

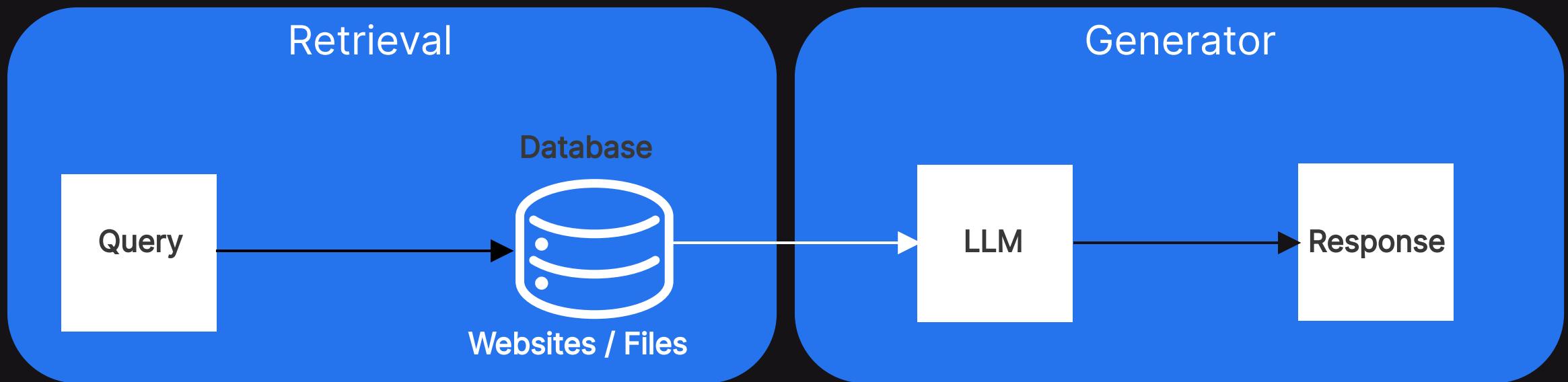


# Components of RAG



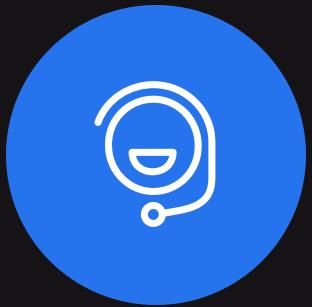
# Components of RAG

- AI with External Knowledge Integration



# Hands-On

# Applications



## Customer Support

Enhance chatbots with internal knowledge.

# Applications



## Customer Support

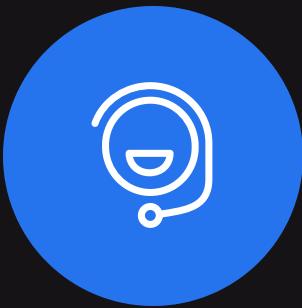
Enhance chatbots with internal knowledge



## Healthcare

Retrieves patient data for informed responses.

# Applications



## Customer Support

Enhance chatbots with internal knowledge for accurate responses.



## Healthcare

Retrieve patient data for informed responses.



## Content Creation

Gather insights and checks facts using online sources.

# Applications



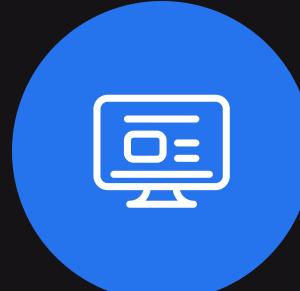
## Customer Support

Enhance chatbots with internal knowledge.



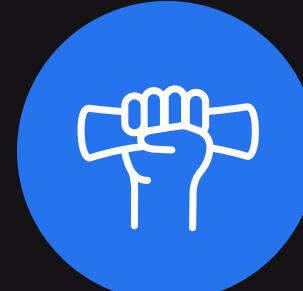
## Healthcare

Retrieve patient data for informed responses.



## Content Creation

Gather insights and checks facts using online sources.



## Education

Generate personalized study guides by retrieving specific learning materials.

# Challenges



1

## Complexity

Seamless integration is required between retrieval and generation components.



2

## Latency

Delays in real-time applications due to retrieval step.

# RAG: Summary

- RAG enhances generative AI with real-time, relevant data.
- Key considerations: data quality & system complexity management.

# Thank You