Improving Retrieval for Q-A on Financial Documents

Anote

Presented by Spurthi Setty

Anote

Introduction

What is Retrieval Augmented Generation (RAG)?

01

The Problem

The Current Limitations of the RAG pipeline

02

Methods to Improve

Chunking Strategies, Metadata Annotations & Filtering, Query Expansion

03

Evaluation

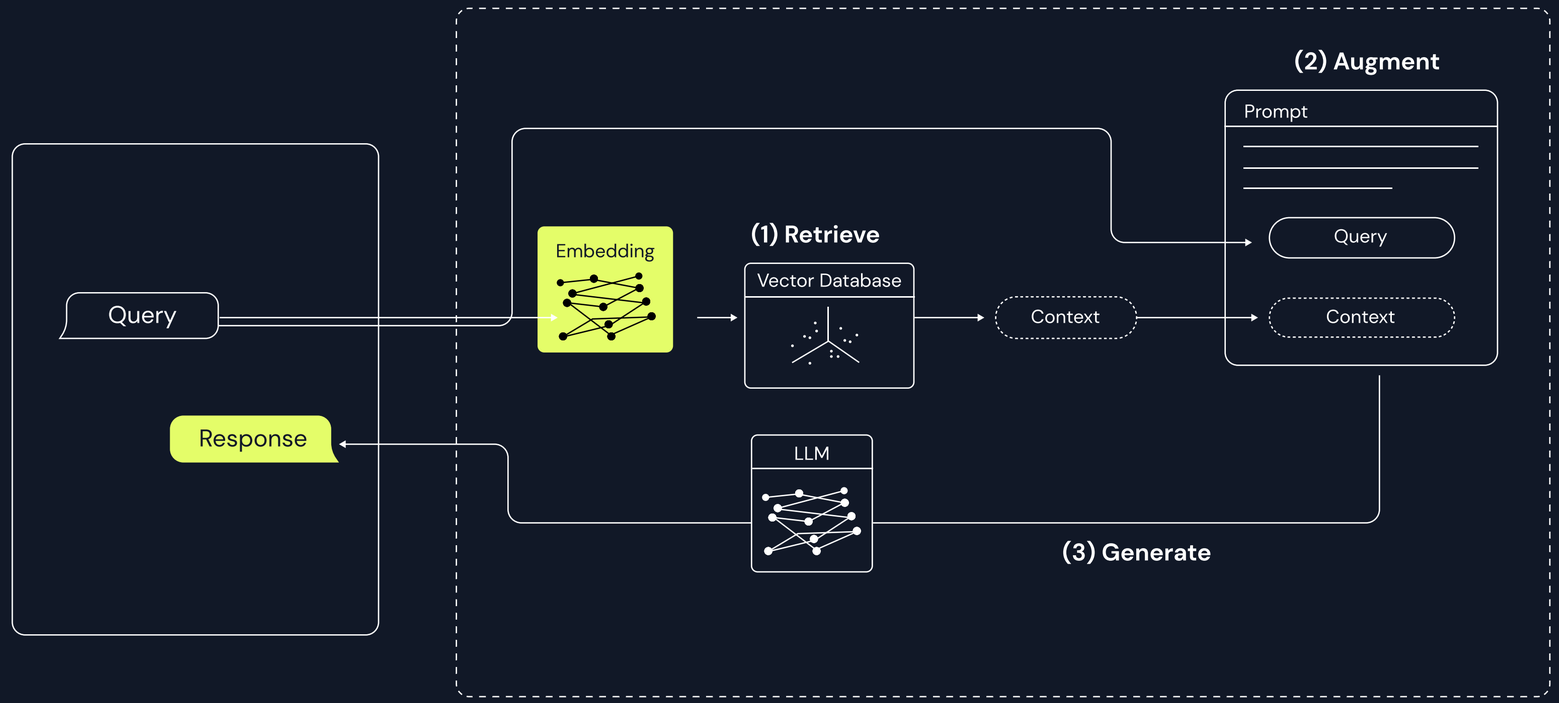
Methods to measure the accuracy and quality of results

04

Anote

Retrieval Augmented Generation

- In-context knowledge injection
- Adds additional knowledge sources directly to the input query
- Effective in preventing hallucinations



Anote

Limitations of Current RAG Pipe

Similarity
≠
Relevance

Ineffective Chunking
↓
Disregards document structure

Information is sparse
+
located in multiple different locations

Lack of Domain Specific Knowledge

Anote

Chunking Strategies

Before Base RAG

Uniform Chunking

- All chunks are of the same size
- Disregards the document structure and the nature of the queries being asked
- Leads to chunks being incomplete or containing too much irrelevant information

New Improved

Recursive Chunking

- Dynamic strategy to divide the text
- Adapt more fluidly to the content
- Uses punctuation and other NLP indicators

Element Based Chunking

- Define sections based on headings and subheadings
- Be able to identify and separate tables as different chunks
- Can use predefined rules based on nature of financial documents for RAG

Anote

Metadata Annotations and Filtering

By date, topic, company etc.

Annotate chunks of tables based on what they are

Enables chatting with multiple docs

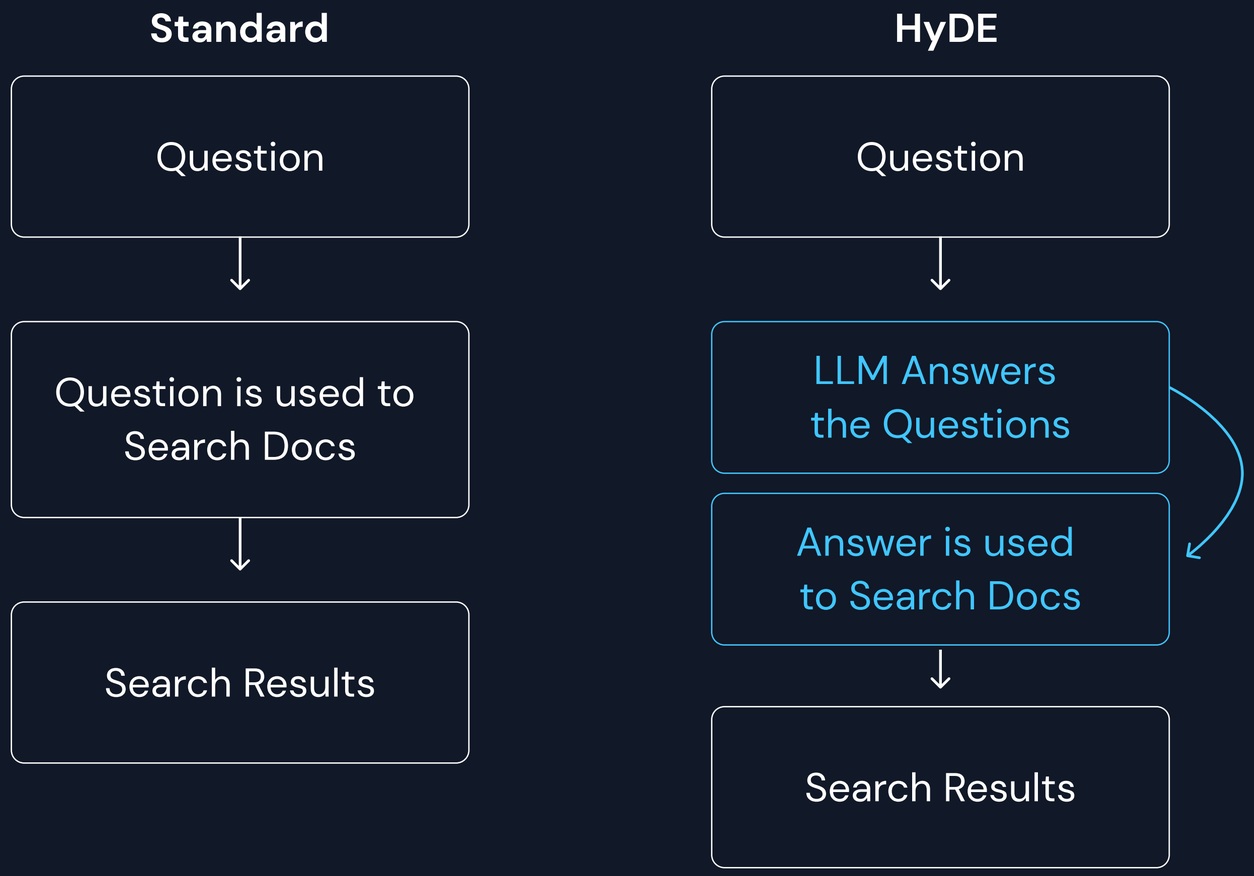
Annotate chunks with summary or additional context

Anote

Query Expansion

HyDE – Retrieve chunks based on more than just the original query

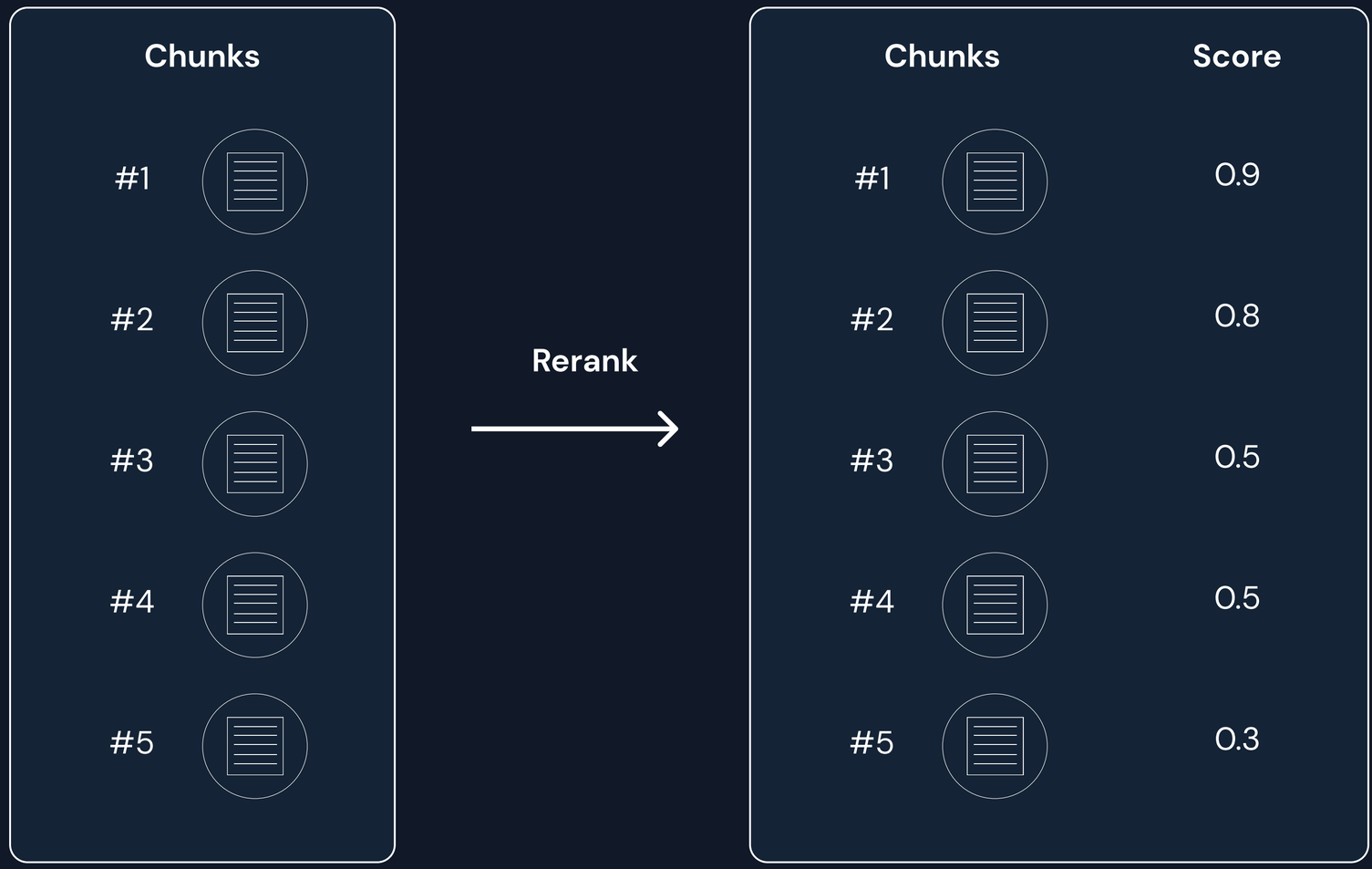
Hypothetical Document Embeddings



Anote

Reranking Chunks

Separate algorithm to reranks chunks by relevance rather than just similarity



Anote

Evaluation

Retrieval Accuracy

- Page level and paragraph level accuracy
- RAGAS Score for Retrieval Quality

Answering Accuracy

- LLM evaluation (ex: GPT Score)
- Cosine Similarity/Rouge Score