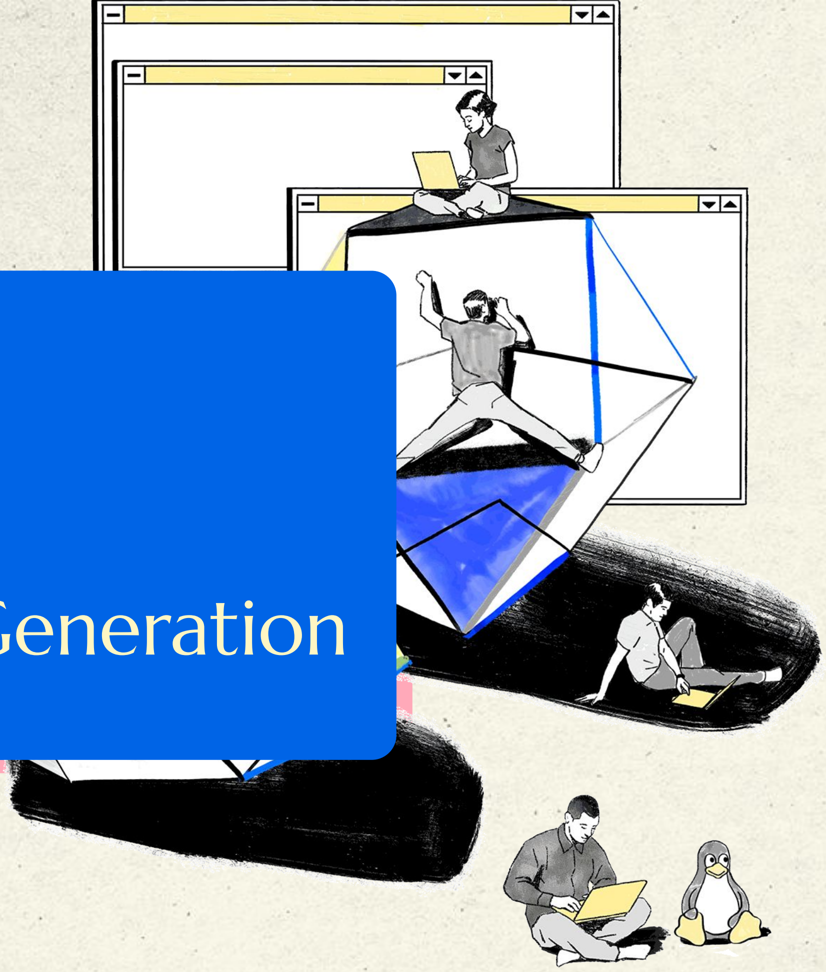
Unit AI-2

Retrieval-Augmented Generation



Codesmith

Become Irreplaceable.

Roadmap

- Overview
- How RAG works
- RAG optimizations
- RAG evaluations

Overview

What is RAG?

RAG combines information retrieval with generative AI models to enhance the accuracy and relevance of AI-generated content.

It provides an effective and efficient way of programmatically scaffolding queries with additional context.

Why focus on RAG?

LLMs are not designed to retrieve data - RAG is key to unlocking their potential.

The RAG layer is especially relevant for AI engineering.

Benefits of RAG

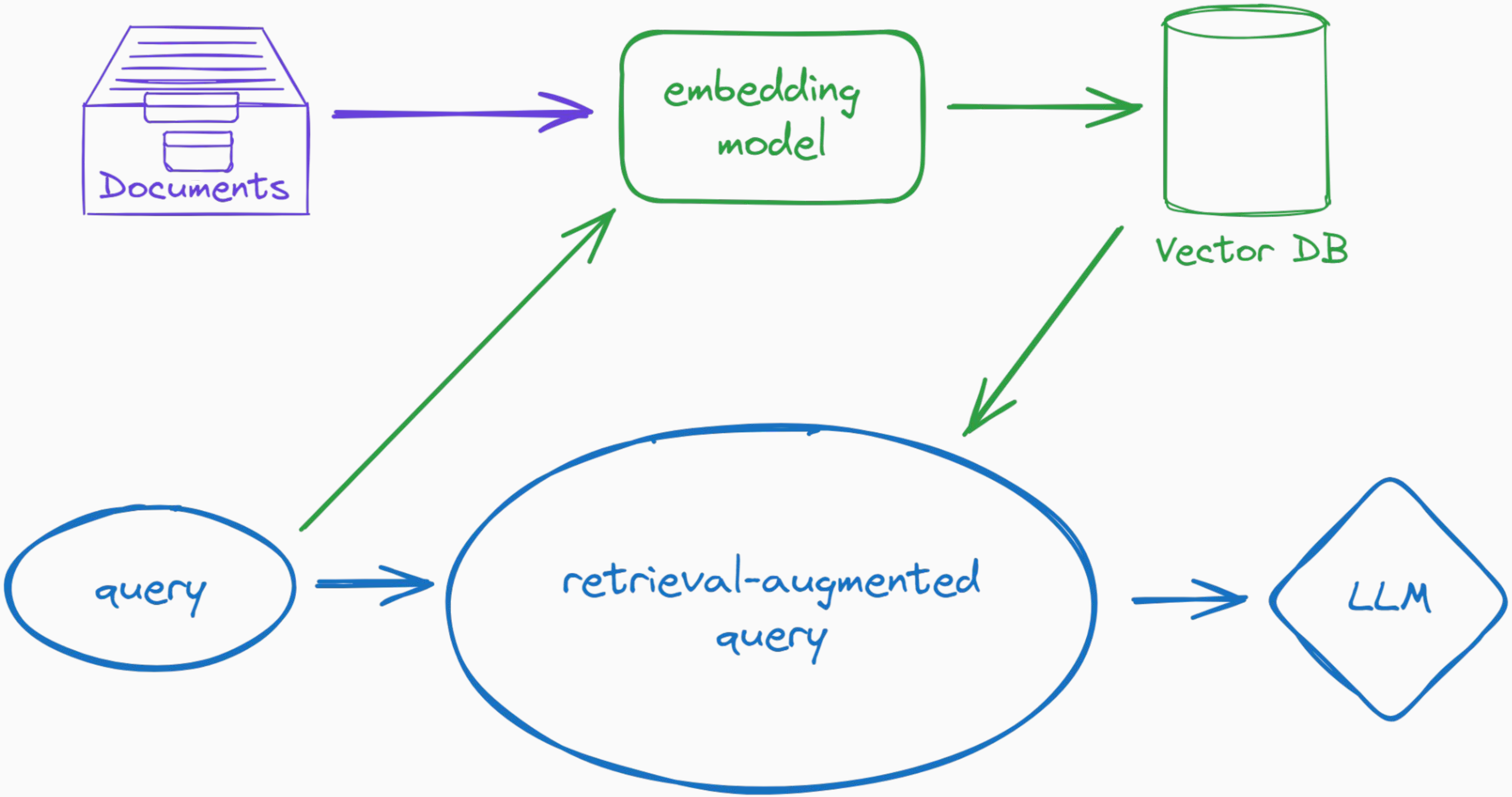
- Groundedness
    - Relevance
    - Accuracy
    - Currentness
    - Domain / proprietary specificity
- Interpretability
- Efficiency

Prompting vs RAG vs fine-tuning

Goal	Prompting	RAG	Fine-Tuning
Grounding	😐	✅	😐
Consistency	✅	😐	✅
Confidence	✅	😐	✅
Interpretability	❌	✅	❌
Alignment	😐	😐	✅
Robustness	😐	❌	✅
Latency	❌	😐	✅
Cost	❌	😐	😐

How RAG works

RAG workflow



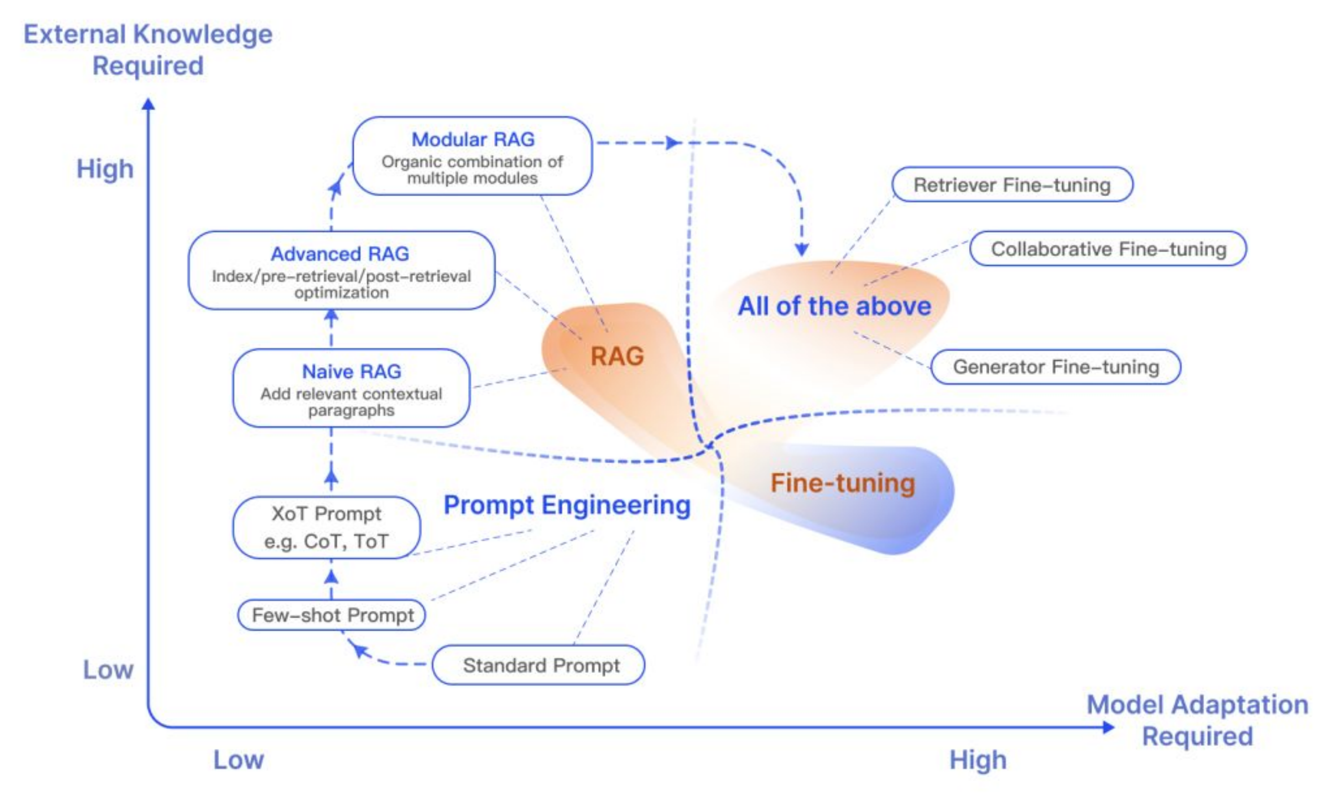
Naive RAG challenges

- Precision
- Recall
- Faithfulness
- Answer relevance
- Accuracy

Data ingestion

- Discovery
- Acquisition
- Validation
- **Transformation**
- **Loading**

RAG vs fine-tuning


External Knowledge Required
High
Low
Modular RAG
Organic combination of multiple modules
Advanced RAG
Index/pre-retrieval/post-retrieval optimization
Naive RAG
Add relevant contextual paragraphs
XoT Prompt
e.g. CoT, ToT
Few-shot Prompt
Standard Prompt
Prompt Engineering
RAG
All of the above
Retriever Fine-tuning
Collaborative Fine-tuning
Generator Fine-tuning
Fine-tuning
Model Adaptation Required
Low
High

[*Retrieval-Augmented Generation for Large Language Models: A Survey*](#)

Retrieval optimization

Chunk optimization

Balance retrieval specificity against generation context.

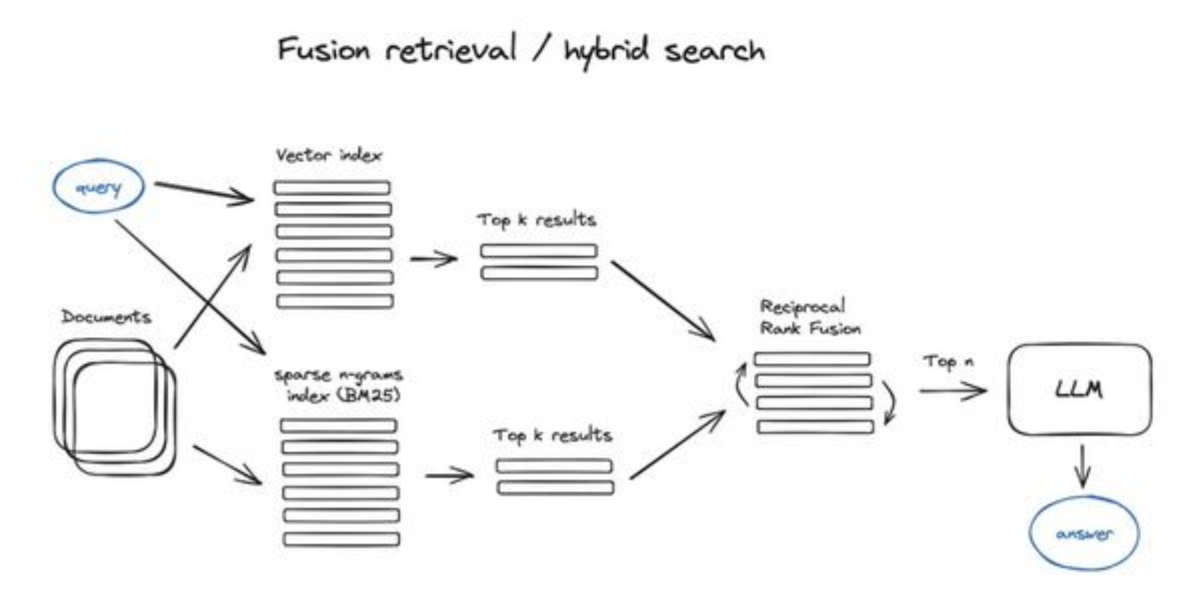
- Fixed-size
- Intent-based & Recursive
- Strategy-based

Hybrid retrieval

Capitalize on traditional information retrieval techniques.

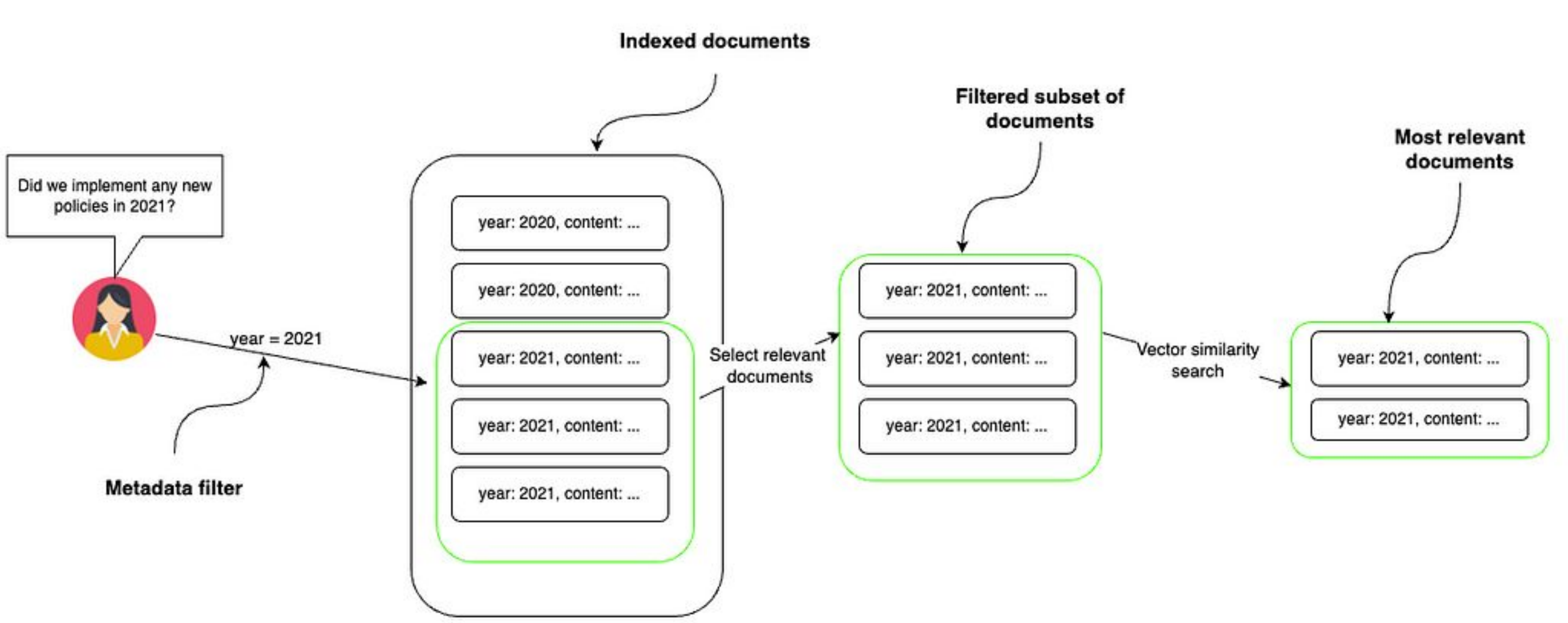
- Ensemble (rank aggregation / weighted averaging)
- Cascade (filtering layers)

Keyword + embedding retrieval



[Wei, Huang, and Wang: Retrieval-Augmented Generation for LLM Applications](#)

Metadata filtering



Neo4j - Graph-based Metadata Filtering to Improve Vector Search in RAG Applications

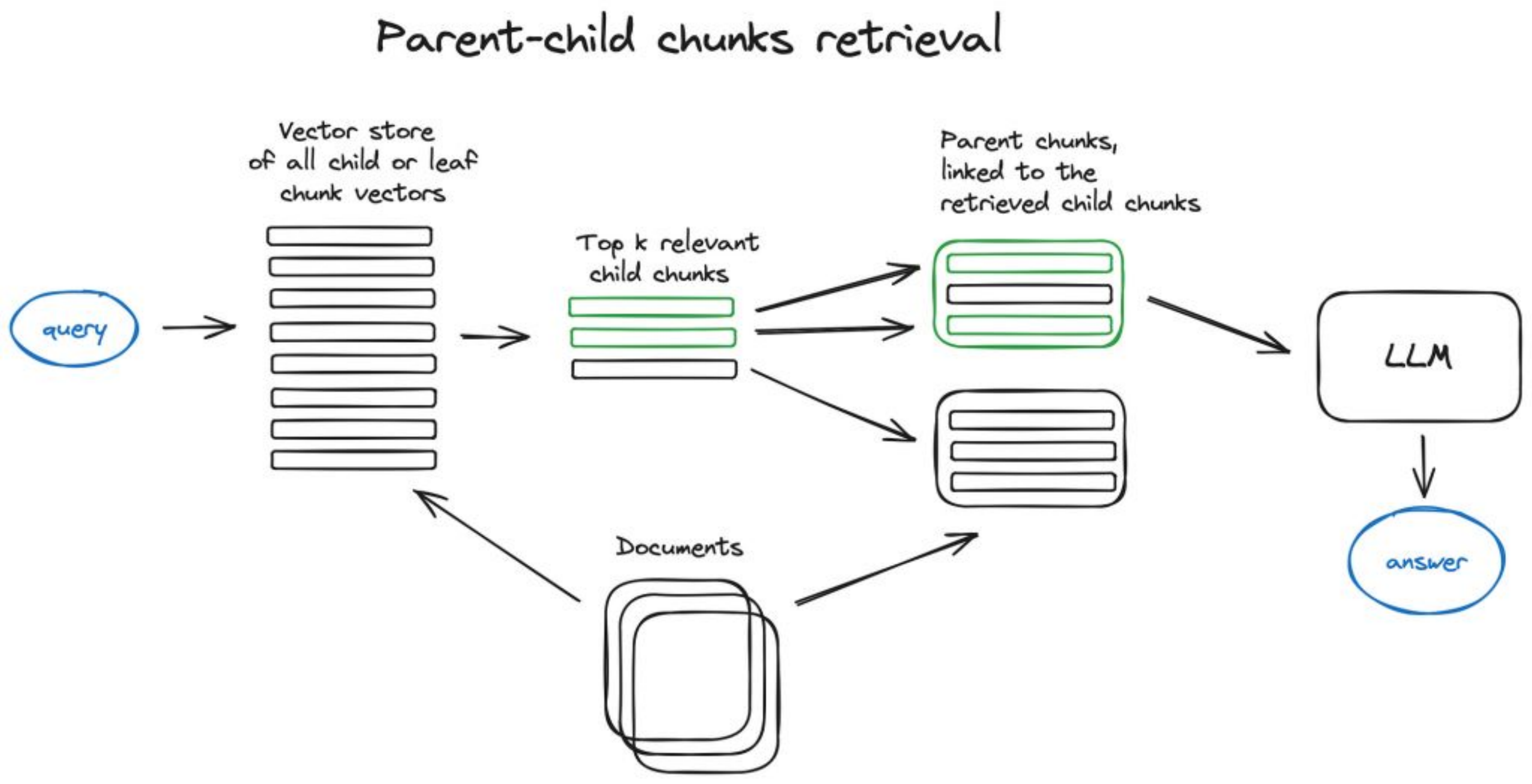
Small-to-big retrieval

Optimize for retrieval *and* generation by expanding context after initial search.

- Sentence window retrieval
- Parent document crawler

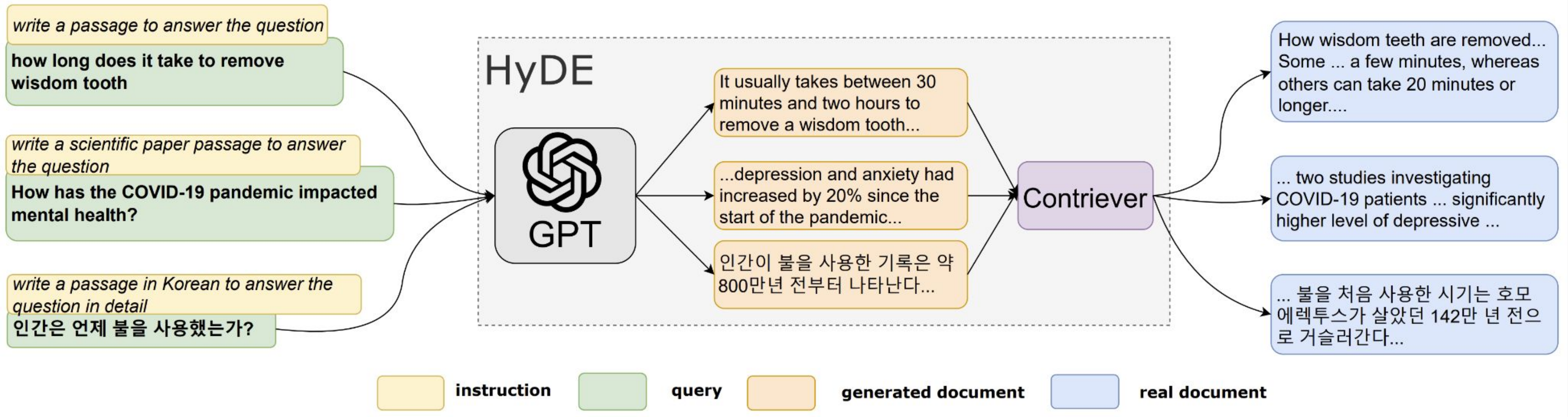
Similarly, retrieval could proceed recursively to build context.

Parent document crawler


Parent-child chunks retrieval
Vector store of all child or leaf chunk vectors
Top k relevant child chunks
Parent chunks, linked to the retrieved child chunks
query
LLM
answer
Documents

ChatGen.ai - The Ultimate Guide on Retrieval Strategies - RAG

Hypothetical Document Embeddings (HyDE)



write a passage to answer the question

how long does it take to remove wisdom tooth

write a scientific paper passage to answer the question

How has the COVID-19 pandemic impacted mental health?

write a passage in Korean to answer the question in detail

인간은 언제 불을 사용했는가?

It usually takes between 30 minutes and two hours to remove a wisdom tooth...

...depression and anxiety had increased by 20% since the start of the pandemic...

인간이 불을 사용한 기록은 약 800만년 전부터 나타난다...

How wisdom teeth are removed... Some ... a few minutes, whereas others can take 20 minutes or longer....

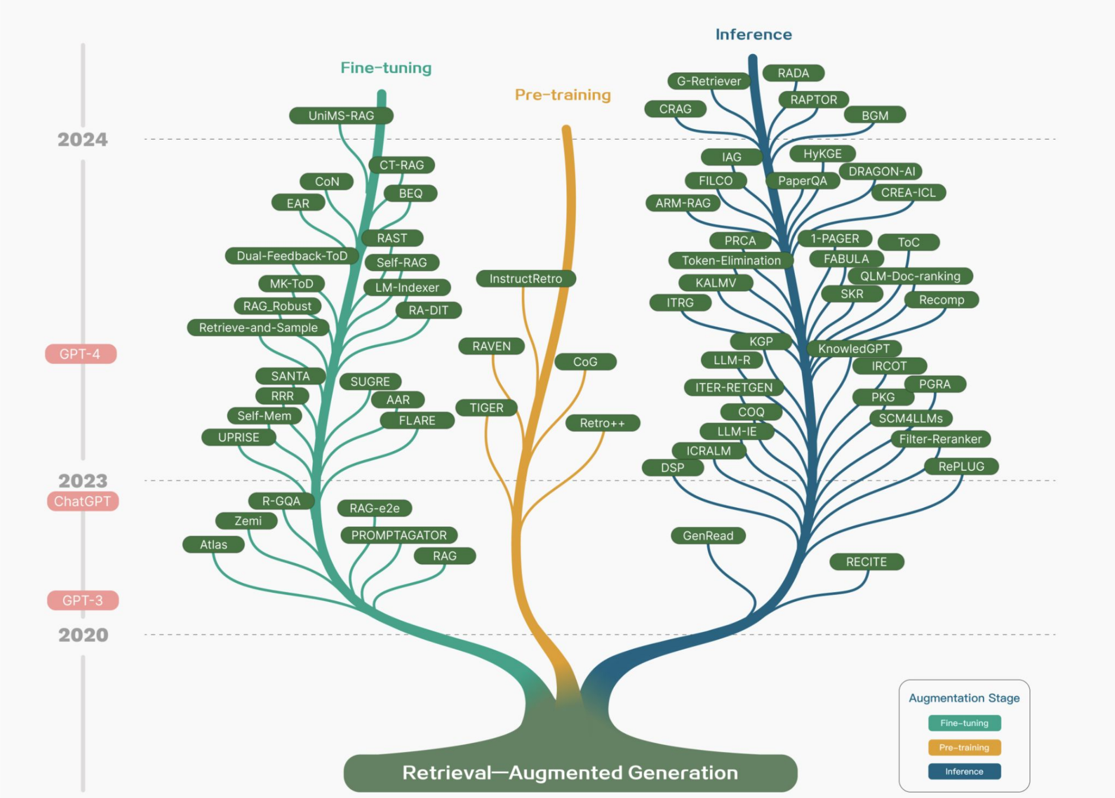
... two studies investigating COVID-19 patients ... significantly higher level of depressive ...

... 불을 처음 사용한 시기는 호모 에렉투스가 살았던 142만 년 전으로 거슬러간다...

instruction **query** **generated document** **real document**

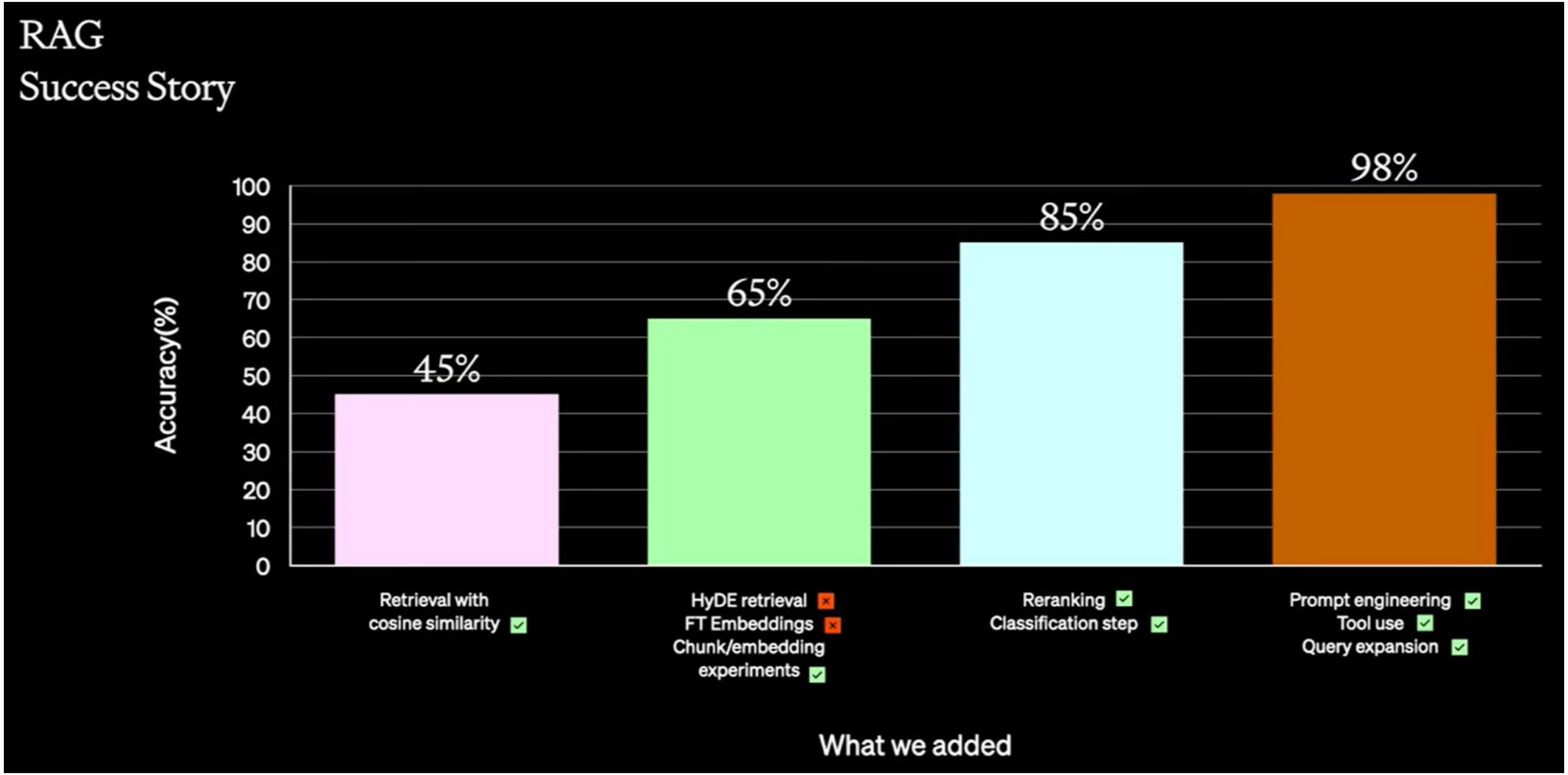
[Precise Zero-Shot Dense Retrieval without Relevance Labels]

And many more...



Retrieval-Augmented Generation for Large Language Models: A Survey

OpenAI “RAG Success Story”



A Survey of Techniques for Maximizing LLM Performance

RAG evaluation

Information retrieval metrics

- Recall
- Precision
- Mean reciprocal rank (MRR)
- Normalized discounted cumulative gain (NDCG)

Building an evaluation dataset

- Benchmarks
- Web data
- Synthetic data
- Proprietary data

RAG testing pyramid

As with traditional applications, it’s beneficial to test each unit (retrieval, augmentation, and generation), their integration, and the end-to-end result.

RAG evaluation dimensions

- Retrieval relevance
- Groundedness
- Adaptability
- Toxicity
- Efficiency

Retrieval relevance

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
- RAGAs (RAG Assessment)

Groundedness

- Entity linking accuracy
- Semantic similarity scores
- FactScore

Generation-specific dimensions

- Noise robustness
- Negative rejection
- Information integration
- Counterfactual robustness

Further considerations

RAG vs long context

SOTA models boast long context windows and have dramatically improved their performance - but RAG may still be preferable given:

- Expanded scope
- Increased precision
- Reduced cost

RAGOps

- Vector DBs
- Frameworks + libraries
- RAG 2.0