O'REILLY

Hands-on Retrieval Augmented Generation

Jonathan Fernandes

Agenda

What is the latest in the world of RAG? (20 minutes)

Embeddings (40 minutes)

- Presentation: Limitations of LLMs; embeddings
- Hands-on exercise: Explore embeddings

Retrieval-augmented generation (45 minutes)

- Presentation: Using external sources of data; practical considerations; working with vector databases; introduction to RAG and LlamaIndex
- Hands-on exercise: Explore RAG

Using LlamaIndex (45 minutes)

- Presentation: Nodes; query; node parsers; retrievers and query engines
- Hands-on exercise: Work with LlamaIndex PDFs

Using OpenAI Assistant retrieval (30 minutes)

- Presentation: Retrieval using the Playground; retrieval using the OpenAI API; why RAG-based solutions are still relevant despite OpenAI Assistant
- Hands-on exercise: Use OpenAI Assistants for retrieval project

About me

Jonathan A. Fernandes · Verify now

AI Engineering & Large Language Models | Advisor AI & ML

United Kingdom · Contact info

AI & ML Advisory Services

University of Warwick - Warwick Business School

Live Course

Hands-on GPT-4-Turbo

With Jonathan Fernandes

3h 0m · June 20 • 5pm-8pm GMT+1

Live Course

How to Choose the Right LLM for your Application

With Jonathan Fernandes

3h 0m · May 16 • 5pm-8pm GMT+1

Quiz

In the context of LLMs, what does fine-tuning mean?

- 👍 Adjusting the model's hyperparameters
- 👎 Training the models on a specialized dataset

How can RAG contribute to data governance in LLMs?

👍 Regulatory control
👎 Using role-based access control

Which of these is not a requirement in RAG?

👍 Using a vector database
👎 Augmenting a user prompt

If you implement RAG correctly, you don't need finetuning

- True
- False

unsloth is a relatively new open source project.

The latest OpenAI model (gpt-4-turbo) gets information about the project incorrect. What is the best way to address this?

- Fine-tuning
- RAG

What is the main problem with using prompting instead of RAG for data retrieval in LLMs?

👍 More expensive to implement
👎 Constraints with token limits

What is an advantage of RAG over fine-tuning in Large Language Models?

👍 Data freshness
👎 Access to increased token limit

Your LLM isn't providing data output in the format that you want (e.g. json). What is the best option?

👍 Fine-tune
👎 RAG

What is RAG?

What is RAG?

Retrieval Augmented Generation

What is RAG?

Which planet in our solar system has the most moons?

What is RAG?

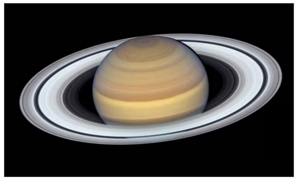
Saturn regains status as planet with most moons in solar system

Discovery of 62 new moons restores ringed planet's lead after it was briefly overtaken by Jupiter

Hannah Devlin *and* Nicola Davis

Fri 12 May 2023 18.07 BST

Share



The new moons will eventually be given names based on Gallic, Norse and Canadian Inuit gods, in keeping with convention for Saturn's moons. Photograph: NASA/JPL-Caltech/Space Science Institute/Reuters

Saturn has regained its crown as the planet with the most moons in the solar system, just months after being overtaken by its fellow gas giant Jupiter.

The leap-frog comes after the discovery of 62 new moons of Saturn, bringing its official total to 145. Jupiter, which added 12 moons to its tally in February, has 95 moons that have been formally designated by the International Astronomical Union (IAU).

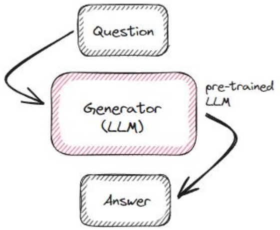
What does Llama-3 (8B) say?

What does Llama-3 (8B) say if we give it more up to date information?

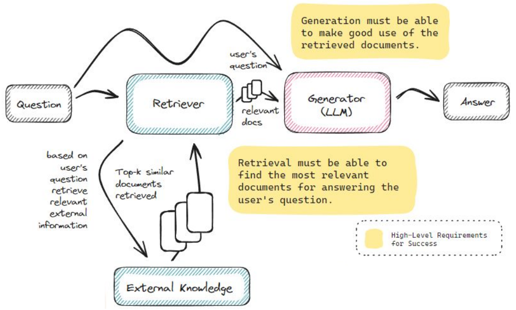
What is RAG?

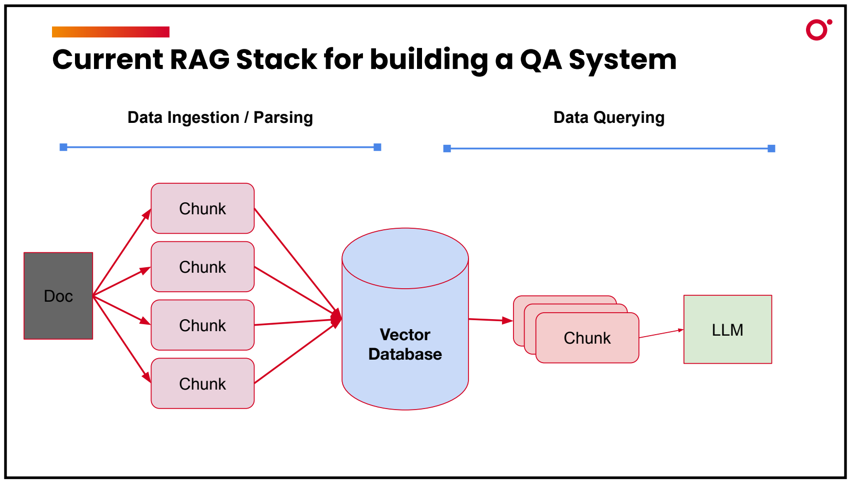
- Training date period
- Checking with a relevant source

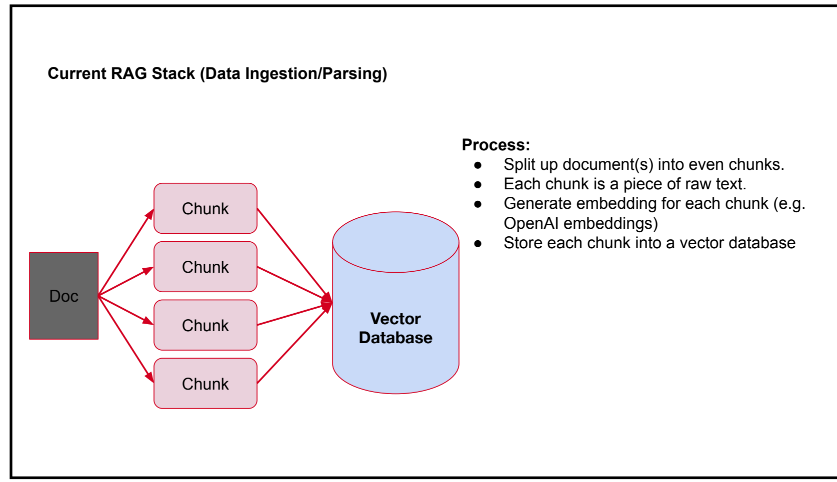
What is RAG?



What is RAG?

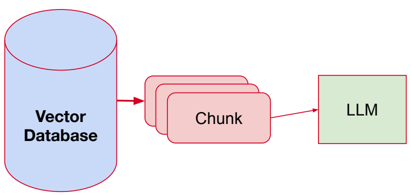



Current RAG Stack for building a QA System
Data Ingestion / Parsing
Data Querying
Doc
Chunk
Chunk
Chunk
Chunk
Vector Database
Chunk
LLM


Current RAG Stack (Data Ingestion/Parsing)
Doc
Chunk
Chunk
Chunk
Chunk
Vector Database
Process:
● Split up document(s) into even chunks.
● Each chunk is a piece of raw text.
● Generate embedding for each chunk (e.g. OpenAI embeddings)
● Store each chunk into a vector database

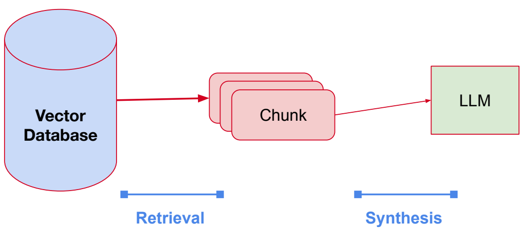
Current RAG Stack (Querying)

Process:
- Find top-k most similar chunks from vector database collection
- Plug into LLM response synthesis module



Current RAG Stack (Querying)

Process:
- Find top-k most similar chunks from vector database collection
- Plug into LLM **response synthesis module**



Using Llama Index

Digital piano

https://jonfernandes.github.io/files/digital-piano.pdf

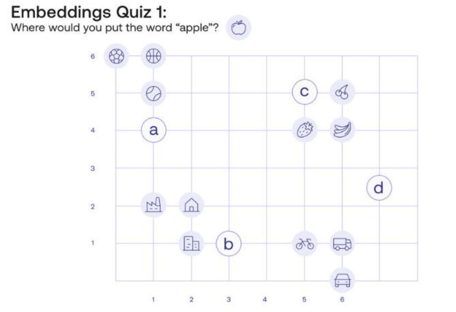
How do you play a “demo” using this digital piano? (3 min)

Hands-on Llama-index:

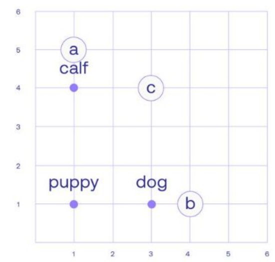
https://colab.research.google.com/drive/1eWbwmHYCzxw7dTnk0uYbWdD18pInOdzX?usp=sharing

Embeddings

Banana
Basketball
Bicycle
Building
Car
Castle
Cherry
House
Soccer
Strawberry
Tennis
Truck


Embeddings Quiz 1:
Where would you put the word “apple”?
c
a
d
b
1 2 3 4 5 6
1 2 3 4 5 6

What is c?


a
calf
c
puppy
dog
b
1 2 3 4 5 6
1 2 3 4 5 6

Word embeddings

Many more columns

Word	Numbers	
Apple	5	5
Soccer	0	6
House	2	2
Car	6	0

Word	Numbers			
A	-0.82	-0.32	...	-0.23
Aardvark	0.419	1.28	...	-0.06
...			...	
Zygote	-0.74	-1.02	...	1.35

4096

Sentence embeddings with Cohere (demo)

https://docs.google.com/spreadsheets/d/17AVE0M1mLgOVR1ptDUzP218rVrXbTTzwaQkxDpQlPIQ/edit?usp=sharing

Similarity between text

- Dot Product
- Cosine Similarity

The more similar two words or sentences are, the larger their Dot Product

Cohere's embeddings have 4096 dimensions

What do each of the dimensions mean?

	Dimension 0 (How citric?)	Dimension 1 (How large?)
Lemons are rich in vitamin C	8	2
Limes are tangy and acidic	9	1
Michael Jordan played for the Chicago Bulls	0	10

What do each of the dimensions mean?

	Dimension 0 (How citric?)	Dimension 1 (How large?)
Lemons are rich in vitamin C	8	2
Limes are tangy and acidic	9	1
Michael Jordan played for the Chicago Bulls	0	10

Dot-product between Lemons and Jordan sentence : 8 x 0 + 2 x 10 = 20

What do each of the dimensions mean?

	Dimension 0 (How citric?)	Dimension 1 (How large?)
Lemons are rich in vitamin C	8	2
Limes are tangy and acidic	9	1
Michael Jordan played for the Chicago Bulls	0	10

Dot-product between Limes and Jordan sentence : 9 x 0 + 1 x 10 = 10

What do each of the dimensions mean?

	Dimension 0 (How citric?)	Dimension 1 (How large?)
Lemons are rich in vitamin C	8	2
Limes are tangy and acidic	9	1
Michael Jordan played for the Chicago Bulls	0	10

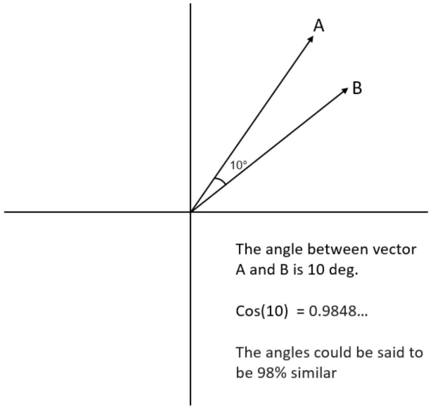
Dot-product between Limes and Lemons sentence : 8 x 9 + 2 x 1 = 74

Can we have a similarity score between 0 and 1?

Cosine Similarity:

- **2 sentences that are very dissimilar have a score close to 0.**
- **2 sentences that are similar have a score close to 1.**

Cosine Similarity



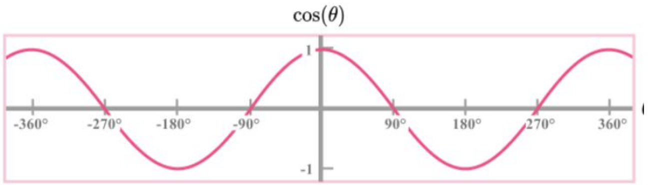
The angle between vector A and B is 10 deg.

Cos(10) = 0.9848...

The angles could be said to be 98% similar

Cosine Similarity:

- 2 sentences that are very dissimilar have a score close to 0.
- 2 sentences that are similar have a score close to 1.



Colab notebook (7 minutes):
https://colab.research.google.com/drive/1YVy0zrz42z2WexDYUFHMu9XMRIuJgKB5

Challenges with basic RAG

Source: llamaindex

Challenges with basic RAG

- **Quality-Related (Hallucination, Accuracy)**
- **Non-Quality-Related (Latency, Cost, Syncing)**

Challenges with basic RAG (Response Quality)

- Poor Retrieval
  - **Low Precision:** Not all chunks in retrieved set are relevant
    - Hallucination + Lost in the Middle Problems
  - **Low Recall:** Now all relevant chunks are retrieved.
    - Lacks enough context for LLM to synthesize an answer
  - **Outdated information:** The data is redundant or out of date.

Poor retrieval:

https://cohere-search-demos.vercel.app/?ref=cohere-ai.ghost.io&_gl=1*1b4m9f*_gcl_au*MTYyODQxMzMzMS4xNzA3MzYzNzcy

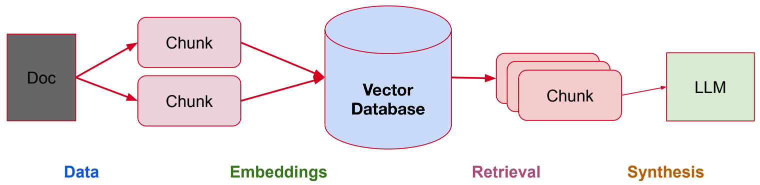
Challenges with Naive RAG (Response Quality)

- Poor Retrieval
  - **Low Precision:** Not all chunks in retrieved set are relevant
    - Hallucination + Lost in the Middle Problems
  - **Low Recall:** Now all relevant chunks are retrieved.
    - Lacks enough context for LLM to synthesize an answer
  - **Outdated information:** The data is redundant or out of date.

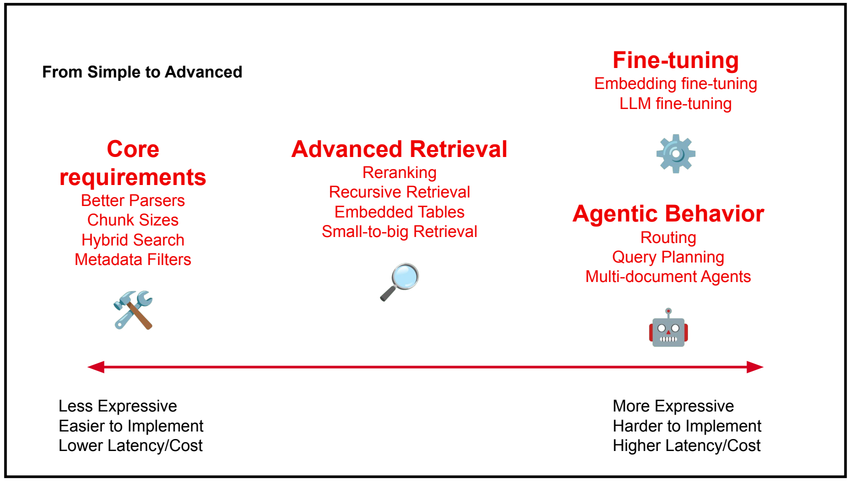
- Poor Response Generation
  - **Hallucination:** Model makes up an answer that isn't in the context.
  - **Toxicity/Bias:** Model makes up an answer that's harmful/offensive.

What do we do?

- **Data**: Can we store additional information beyond raw text chunks?
- **Embeddings**: Can we optimize our embedding representations?
- **Retrieval**: Can we do better than top-k embedding lookup?
- **Synthesis**: Can we use LLMs for more than generation?

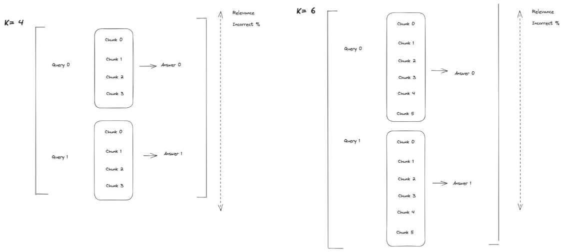



Improving RAG Systems


From Simple to Advanced
Core requirements
Better Parsers
Chunk Sizes
Hybrid Search
Metadata Filters
Advanced Retrieval
Reranking
Recursive Retrieval
Embedded Tables
Small-to-big Retrieval
Fine-tuning
Embedding fine-tuning
LLM fine-tuning
Agentic Behavior
Routing
Query Planning
Multi-document Agents
Less Expressive
Easier to Implement
Lower Latency/Cost
More Expressive
Harder to Implement
Higher Latency/Cost

Core requirements: Chunk Sizes

Tuning your chunk size can have outsized impacts on performance



Core requirements: Prompt Engineering

RAG uses core Question-Answering (QA) prompt templates
Ways you can customize:

- Adding few-shot examples
- Modifying template text

Accessing Prompts

Here we get the prompts from the query engine. Note that *all* prompts are returned, including ones used in sub-modules in the query engine. This allows you to centralize a view of these prompts!

```
prompts_dict = query_engine.get_prompts()
```

```
display_prompt_dict(prompts_dict)
```

Prompt Key: response_synthesizer:summary_template
Text:

```
Context information from multiple sources is below.
---------------------
{context_str}
---------------------
Given the information from multiple sources and not prior knowledge, answer the query.
Query: {query_str}
Answer:
```

Core requirements:
Customizing LLMs

Task performance on easy-to-hard tasks (RAG, agents) varies significantly among LLMs

https://docs.llamaindex.ai/en/stable/module_guides/models/llms/#llm-compatibility-tracking

Core requirements: Customizing Embeddings

Retriever - Many models from OpenAI, CohereAI, and open-source sentence transformers.
Rerankers - Many available from CohereAI and sentence transformers.

Retrieval quality = Your embedding model + reranker

Core requirements: Customizing Embeddings

Hit rate calculates the fraction of queries where the correct answer is found within the top-k retrieved documents. How often does the system gets it right within the top few guesses.

MRR evaluates the system's accuracy by looking at the rank of the highest-placed relevant document. It is the average of the reciprocals of these ranks across all the queries.
If the first relevant document is the top result, the reciprocal rank is 1; if it's second, the reciprocal rank is ½.

Core requirements: Customizing Embeddings

Embedding	WithoutReranker		bge-reranker-base		bge-reranker-large		Cohere-Reranker	
	Hit Rate	MRR	Hit Rate	MRR	Hit Rate	MRR	Hit Rate	MRR
OpenAI	0.876404	0.718165	0.91573	0.832584	**0.910112**	**0.855805**	**0.926966**	**0.86573**
bge-large	0.752809	0.597191	0.859551	0.805243	0.865169	0.816011	0.876404	0.822753
llm-embedder	0.814607	0.587266	0.870787	0.80309	0.876404	0.824625	0.882022	0.830243
Cohere-v2	0.780899	0.570506	0.876404	0.798127	0.876404	0.825281	0.876404	0.815543
Cohere-v3	0.825843	0.624532	0.882022	0.806086	0.882022	0.834644	0.88764	0.836049
Voyage	0.831461	0.68736	0.926966	0.837172	**0.91573**	**0.858614**	**0.91573**	**0.851217**
JinaAI-Small	0.831461	0.614045	0.91573	0.843071	**0.926966**	**0.857303**	**0.926966**	**0.868633**
JinaAI-Base	0.848315	0.68221	0.938202	0.846348	**0.938202**	**0.868539**	**0.932584**	**0.873689**
Google-PaLM	0.865169	0.719476	0.910112	0.833708	**0.910112**	**0.85309**	**0.910112**	**0.855712**

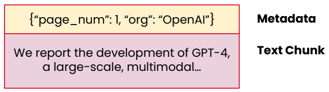
Hit rate - How often does the system gets it right within the top few guesses.

MRR evaluates the system's accuracy by looking at the rank of the highest-placed relevant document.

Core requirements: Metadata Filtering

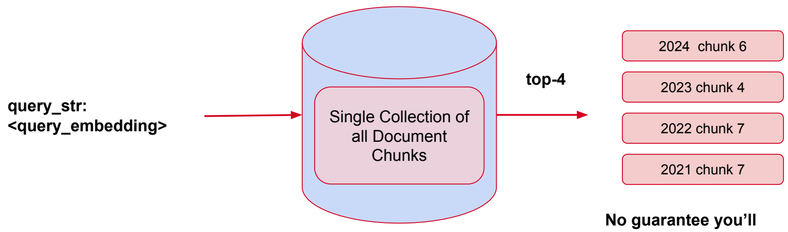
- **Metadata:** context you can inject into each text chunk
- Examples
  - Page number
  - Document title
  - Summary of adjacent chunks
  - Questions that chunk can answer
- **Benefits**
  - **Can Help Retrieval**
  - **Can Augment Response Quality**
  - **Integrates with Vector DB Metadata Filters**

Example of Metadata



Core requirements: Metadata Filtering

Question: "What were the biggest economic risks in 2024?"
Raw Semantic Search is **low precision.**



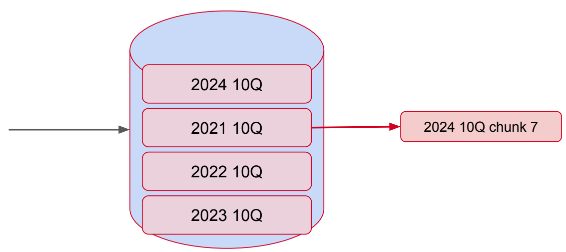
No guarantee you'll return the relevant document chunks!

Core requirements: Metadata Filtering

Question: “What were the biggest economic risks in 2024?
If we can *infer* the metadata filters (year=2024), we remove irrelevant candidates, **increasing precision**!

query_str:
<query_embedding>

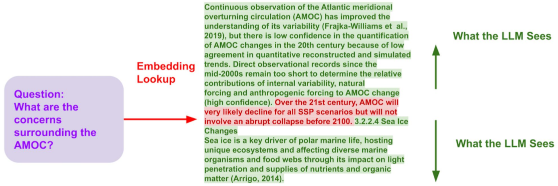
Metadata tags:
{“year”: 2021}



Advanced Retrieval: Small-to-Big

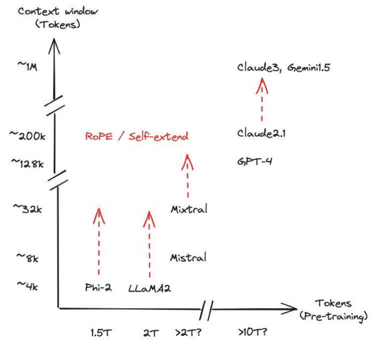
Intuition: Embedding a big text chunk feels suboptimal.
Solution: Embed text at the sentence-level - then **expand** that window during LLM synthesis

Embed Sentence → Link to Expanded Window



Context lengths of 1 million tokens. Do we need RAG?

Context windows are getting larger



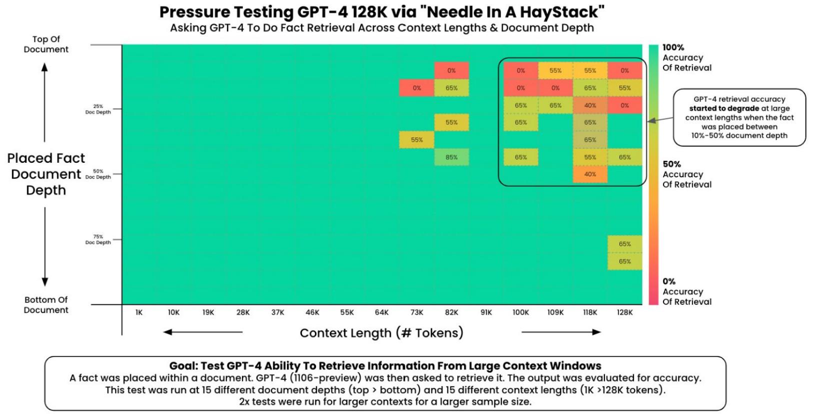
Do we need RAG anymore?

Shaun.AGI
@agishaun

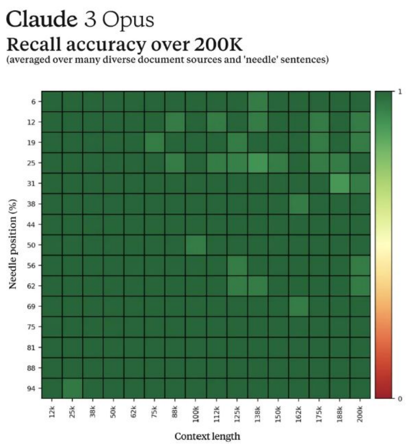
⚠️ RAG might be dead, after reading 58 pages of Genimi 1.5 Pro tech report. Here's my thoughts as AI founder,

1. Simple RAG system like similarity search with vector db will be dead. But more customized RAG will still live. The goal of RAG is mostly on retrieval relevant information. After reading the report, I am convinced LLM can do retrieval really really well.
2. RAG itself may not be dead totally, but 90% of people won't need it anymore. Most dataset can fit in 1M tokens. Just like OpenAI's assistant API, once Gemini API can handle large files, the only thing matters is the cost. However based on the report, 1.5 Pro's training cost and inference cost is much much lower than Gemini 1.0📉

https://twitter.com/agishaun/status/1758561862764122191



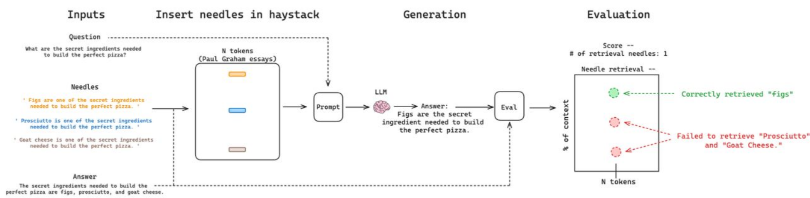
https://github.com/gkamradt/LLMTest_NeedleInAHaystack



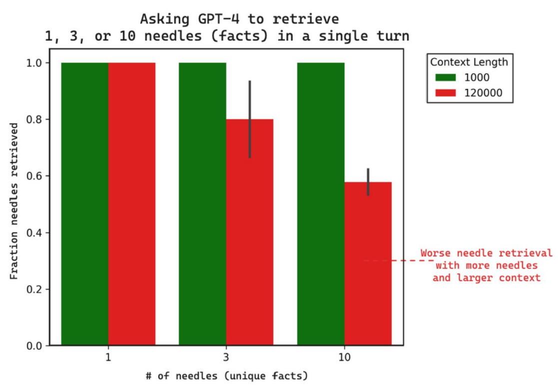
Long context and near-perfect recall

https://www.anthropic.com/news/claude-3-family

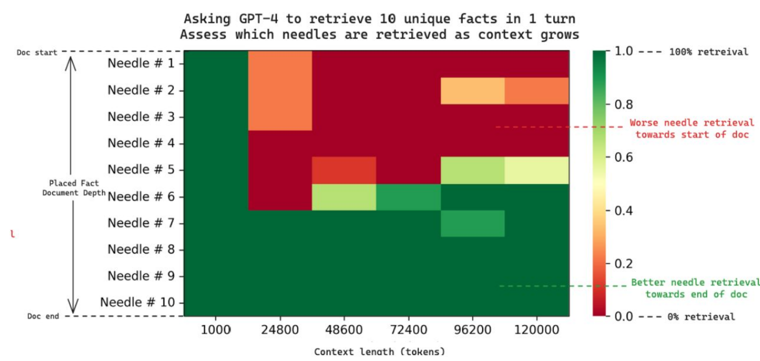
Multiple needles

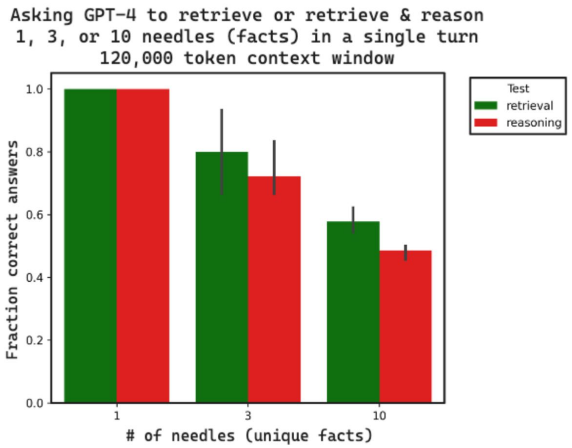


Multiple needles



Multiple needles

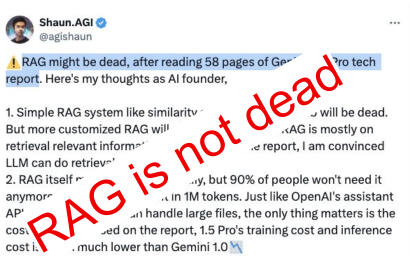




Langsmith - tracking and tracing RAG

https://docs.google.com/spreadsheets/d/1XnXBF_PF0HMmsiopvJPskNhWx7q-m0kE/edit?usp=sharing&ouid=113379868449114672319&rtpof=true&sd=true

Do we need RAG anymore?



Using Openai Assistant Retrieval

Digital piano

How do you play a “demo” using this digital piano using the OpenAI platform?

What has happened behind the scenes for the Assistant to get you your answer?

Limitations and other considerations

- Maximum file size is 512 MB. Must be < 2 million tokens
- Maximum number of files per assistant: 20
- Size for all files in organization < 100 GB
- Pricing: $0.20/GB per assistant per day.
- No support for fine-tuned models.
- Support for notifications without needing to poll

Live Course

[Hands-on GPT-4-Turbo]

With Jonathan Fernandes

3h 0m June 20 • 5pm-8pm GMT+1

Live Course

How to Choose the Right LLM for your Application

With Jonathan Fernandes

3h 0m May 16 • 5pm-8pm GMT+1

Agenda

What is the latest in the world of RAG? (20 minutes)

Embeddings (40 minutes)

- Presentation: Limitations of LLMs; embeddings
- Hands-on exercise: Explore embeddings

Retrieval-augmented generation (45 minutes)

- Presentation: Using external sources of data; practical considerations; working with vector databases; introduction to RAG and LlamaIndex
- Hands-on exercise: Explore RAG

Using LlamaIndex (45 minutes)

- Presentation: Nodes; query; node parsers; retrievers and query engines
- Hands-on exercise: Work with LlamaIndex PDFs

Using OpenAI Assistant retrieval (30 minutes)

- Presentation: Retrieval using the Playground; retrieval using the OpenAI API; why RAG-based solutions are still relevant despite OpenAI Assistant
- Hands-on exercise: Use OpenAI Assistants for retrieval project

O'REILLY®