Retrieval Augmented Generation

Duarte Santos – 124376

Table of Contents

What is Retrieval Augmented Generation?

RAG is a framework, proposed by Patrick Lewis in 2020, that combines retrieval-based systems with generative AI models to improve the relevance and accuracy of generated content.

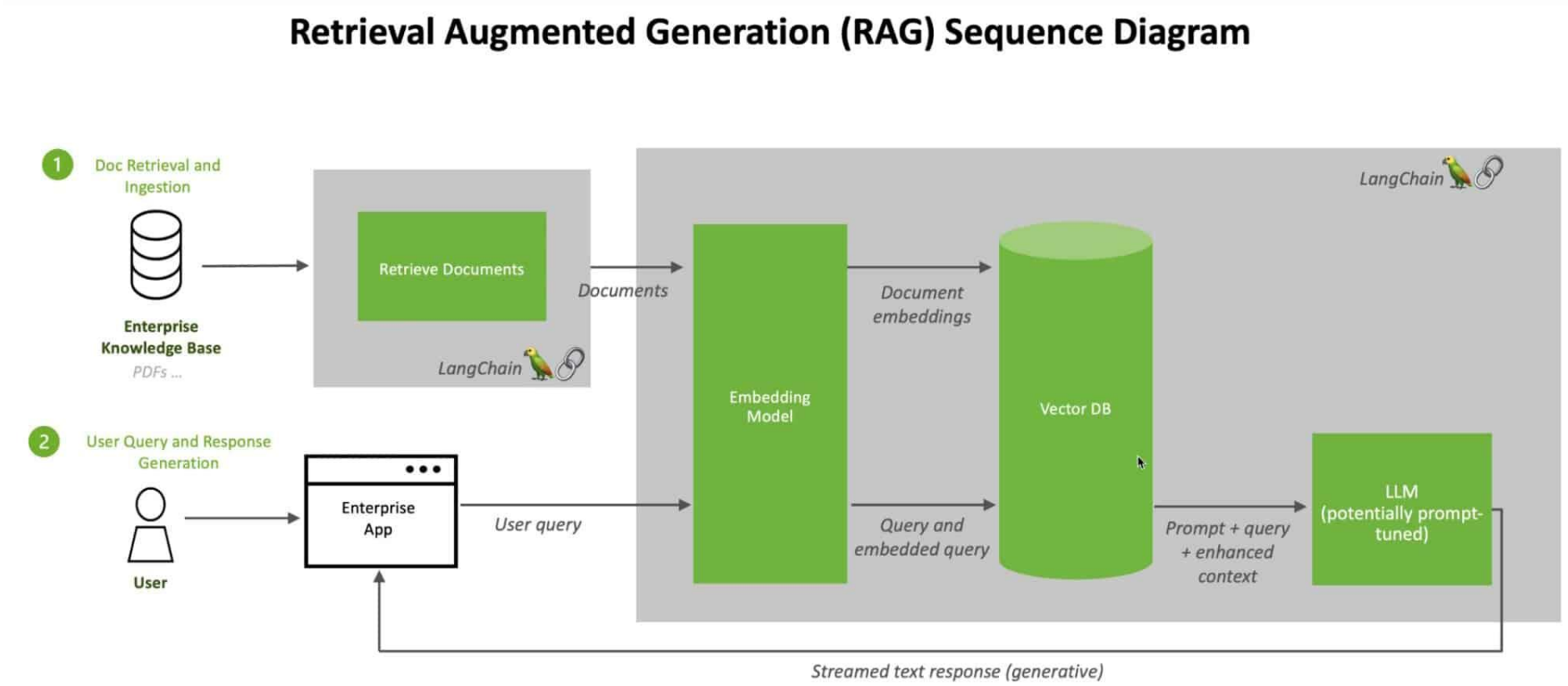
This approach leverages both explicit knowledge from retrieved documents and the implicit knowledge encoded in language models, resulting in more accurate and context-sensitive responses.

RAG Components

Retrieval: The system queries a knowledge base or external data source (e.g., databases, documents, or APIs) to fetch relevant context or information based on the input prompt

Generation: A generative model, such as GPT, uses the retrieved information to produce a coherent and contextually accurate response or output

How RAG works?



https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/

RAG Example

Who were the candidates for the 2024 USA election and who won in November?

I might not have the latest updates. For the most complete and up-to-date information about the U.S. Election, please visit news sources like The Associated Press and Reuters or check with your state or local election authority.

What is the last year this GPT model was trained on?

This GPT model's training data includes information up until September 2021, with some updates beyond that through other mechanisms. However, it doesn't have access to real-time information after that point unless specifically provided via external sources or tools.

RAG Example

WIKIPEDIA
The Free Encyclopedia

Search Wikipedia | Search

2024 United States presidential election

82 languages

Article | Talk | Read | View source | View history | Tools

From Wikipedia, the free encyclopedia

For related races, see 2024 United States elections.

This article **may be too long to read and navigate comfortably**. Consider splitting content into sub-articles, condensing it, or adding subheadings. Please discuss this issue on the article's talk page. *(December 2024)*

Contents hide

(Top)
- Background
- Nominations
- Electoral map
- Campaign issues
- Third-party and independent candidates
- Timeline
- Opinion polling and forecasts
- Debates
- Results
- Analysis of results
- Voter demographics
- Aftermath
- Media analysis

See also
Notes
References
External links

Presidential elections were held in the United States on November 5, 2024.[a] The Republican Party's ticket—Donald Trump, who was the 45th president of the United States from 2017 to 2021, and JD Vance, the junior U.S. senator from Ohio—defeated the Democratic Party's ticket—Kamala Harris, the incumbent vice president, and Tim Walz, the governor of Minnesota.[6][7] Trump and Vance are scheduled to be inaugurated as the 47th president and the 50th vice president on January 20, 2025.[8]

The incumbent president, Joe Biden of the Democratic Party, initially ran for re-election as the party's presumptive nominee,[9] facing little opposition and easily defeating Representative Dean Phillips during the Democratic primaries;[10] however, what was broadly considered a poor debate performance in June 2024 intensified concerns about his age and health, and led to calls within his party for him to leave the race.[11] After initially declining to do so, Biden withdrew on July 21, becoming the first eligible incumbent president to withdraw since Lyndon B. Johnson in 1968.[12] Biden endorsed Harris,[13] who was voted the party's nominee by the delegates on August 5, 2024. Harris selected Walz as her running mate.[14][15]

Trump, who lost in 2020 to Biden, ran for re-election again.[16] He was nominated during the 2024 Republican National Convention along with his running mate, Vance, after winning the Republican primaries by

2024 United States presidential election

← 2020 | **November 5, 2024**[a] | 2028 →

538 members of the Electoral College
270 electoral votes needed to win

Opinion polls

Turnout	63.9% (▼2.7 pp)[1]	
Nominee	**Donald Trump**	Kamala Harris
Party	Republican	Democratic
Home state	Florida	California
Running mate	**JD Vance**	Tim Walz
Electoral vote	**312**	226
States carried	**31 + ME-02**	19 + DC + NE-02
Popular vote	**77,303,573**[2]	75,019,257[2]

https://en.wikipedia.org/wiki/2024_United_States_presidential_election

him the first Republican to do so since George W. Bush in 2004 but with what was the third smallest margin of victory since 1888.[42][43] Relative to the 2020 election, he improved his vote share among working class voters, particularly among young men, those without college degrees, and Hispanic voters.[44] Trump became the first president since Grover Cleveland in 1892 to be elected to non-consecutive terms.[45]

Question: Who were the candidates for the 2024 USA election and who won in November?

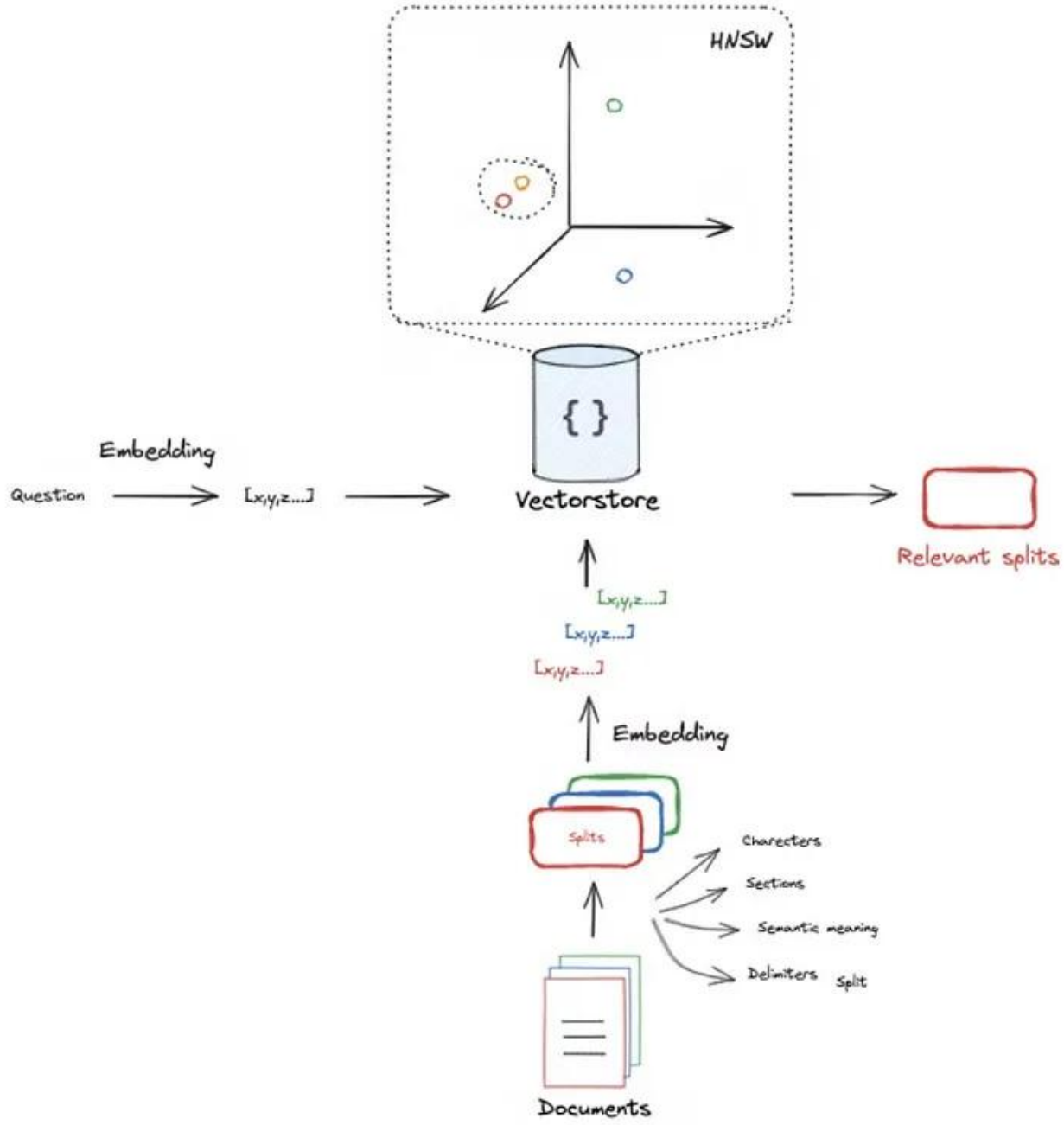
The candidates for the 2024 United States presidential election were:

1. **Republican Party**:
   - **Donald Trump** (former president from 2017 to 2021)
   - **J.D. Vance** (junior U.S. senator from Ohio)
2. **Democratic Party**:
   - **Kamala Harris** (incumbent vice president)
   - **Tim Walz** (governor of Minnesota)

Winner: Donald Trump and his running mate J.D. Vance won the election. They received 312 Electoral College votes to Kamala Harris' 226 and also won the national popular vote with 49.9%. Trump is scheduled to be inaugurated as the 47th president on January 20, 2025, making him the first president since Grover Cleveland to serve non-consecutive terms.

Embeddings improvements

Split the documents into smaller chunks to enable faster searching and more specific retrieval.



https://medium.com/@dinabavli/rag-basics-basic-implementation-of-retrieval-augmented-generation-rag-e80e0791159d

Advantages and Limitations of RAG

- Improved Accuracy
- Reduced LLMs Hallucinations
- Modularity
- Dynamic Knowledge Update
- Domain Adaptability

- Complex Architecture
- Latency
- Potential for Irrelevant Context
- Scalability of Retrieval

RAG Use Cases

CUSTOMER SUPPORT CHATBOTS

SCIENTIFIC RESEARCH

LEGAL ASSISTANCE

ENTERPRISE DOCUMENT SEARCH

Future of RAG

Integration with Real-Time and Dynamic Knowledge

Multilingual retrieving and synthesizing evidence from diverse multilingual sources

Multimodal data (images, vídeo, sound etc...) support

Smaller Models with External Knowledge

Conclusions

Retrieval Augmented Generation offers flexibility, adaptability, and efficiency, making it ideal for a wide range of use cases.

As research and technology progress, RAG systems will become more sophisticated, real-time, and domain-specific, opening new frontiers for AI applications

Questions?