

# Vector Search and state-of-the-art retrieval for Generative AI apps

Pramesh Gautam

# Topics



- High-level overview on LLMs
- Retrieval Augmented Generation
- Vector Search
- Demo

# Large Language Models (LLMs) Usage

Question/Answering



Summarization

Information Extraction

# Large Language Models (LLMs) Issues

Data Cutoff

Hallucination

Bias



You

why was Sam Altman fired from OpenAI?



ChatGPT

As of my last update in January 2022, there was no information suggesting that Sam Altman, one of the co-founders of OpenAI, was fired from the organization. Sam Altman played a significant role in the early stages of OpenAI's development and served as the CEO before transitioning to the role of Chairman of the Board. However, organizational changes or events may have occurred after my last update. For the most accurate and current information, I recommend consulting recent news sources or official statements from OpenAI.



You

What do you know about the statue of Abraham Lincoln in Kathmandu?



ChatGPT

The statue of Abraham Lincoln in Kathmandu, Nepal, is a unique monument that symbolizes the friendship between Nepal and the United States. Erected in the capital city in 2017, this statue is one of only a few such monuments of Lincoln outside the United States. It signifies Nepal's appreciation for Lincoln's role in abolishing slavery and promoting human rights. The installation of the statue reflects Nepal's growing diplomatic relations with the United States and serves as a reminder of the shared values of freedom and democracy between the two nations.



# Large Language Models (LLMs) Issues

Data Cutoff

Hallucination

Bias



You

why was Sam Altman fired from OpenAI?



ChatGPT

As of my last update in January 2022, there was no information suggesting that Sam Altman, one of the co-founders of OpenAI, was fired from the organization. Sam Altman played a significant role in the early stages of OpenAI's development and served as the CEO before transitioning to the role of Chairman of the Board. However, organizational changes or events may have occurred after my last update. For the most accurate and current information, I recommend consulting recent news sources or official statements from OpenAI.



You

What do you know about the statue of Abraham Lincoln in Kathmandu?




ChatGPT

The statue of Abraham Lincoln in Kathmandu, Nepal, is a unique monument that symbolizes the friendship between Nepal and the United States. Erected in the capital city in 2017, this statue is one of only a few such monuments of Lincoln outside the United States. It signifies Nepal's appreciation for Lincoln's role in abolishing slavery and promoting human rights. The installation of the statue reflects Nepal's growing diplomatic relations with the United States and serves as a reminder of the shared values of freedom and democracy between the two nations.



# Large Language Models (LLMs): Work with Private Data




Pass data as prompt

Context length  
Cost  
Processing time

Fine-tuning

Retrieval Augmented Generation

# Large Language Models (LLMs): Work with Private Data



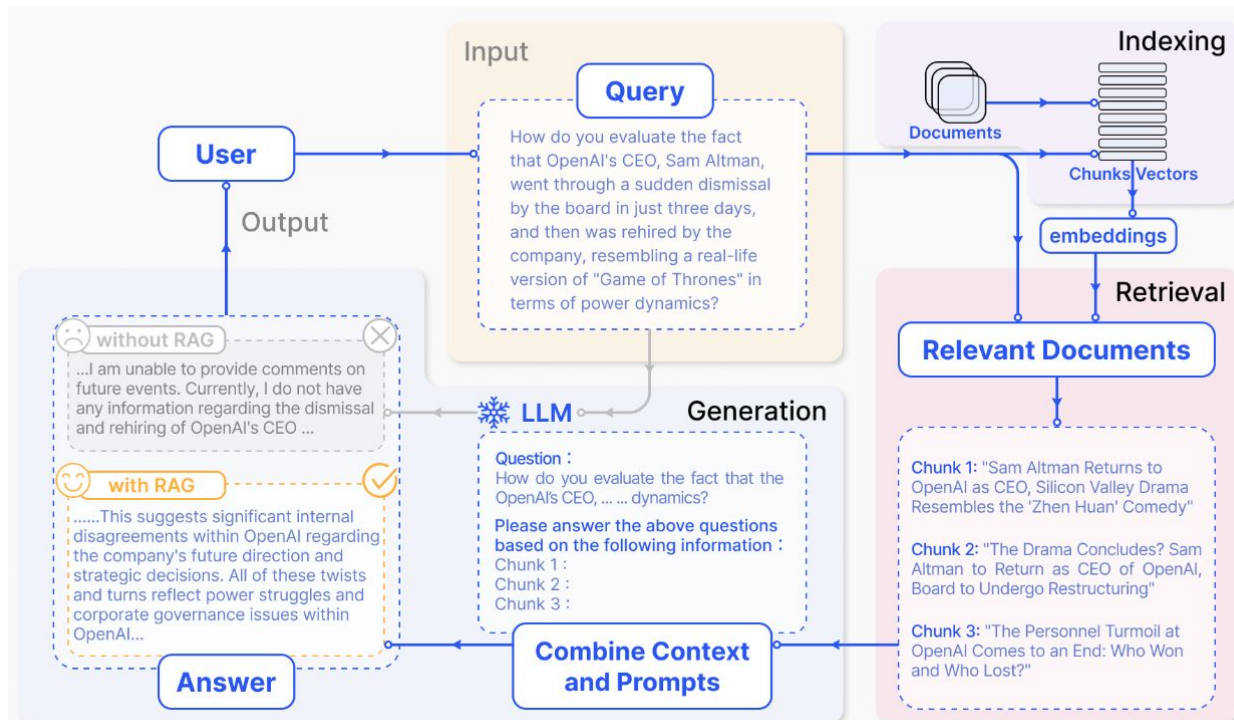
Pass data as prompt

Fine-tuning

Compute  
Technical Complexity  
Keep on iterating

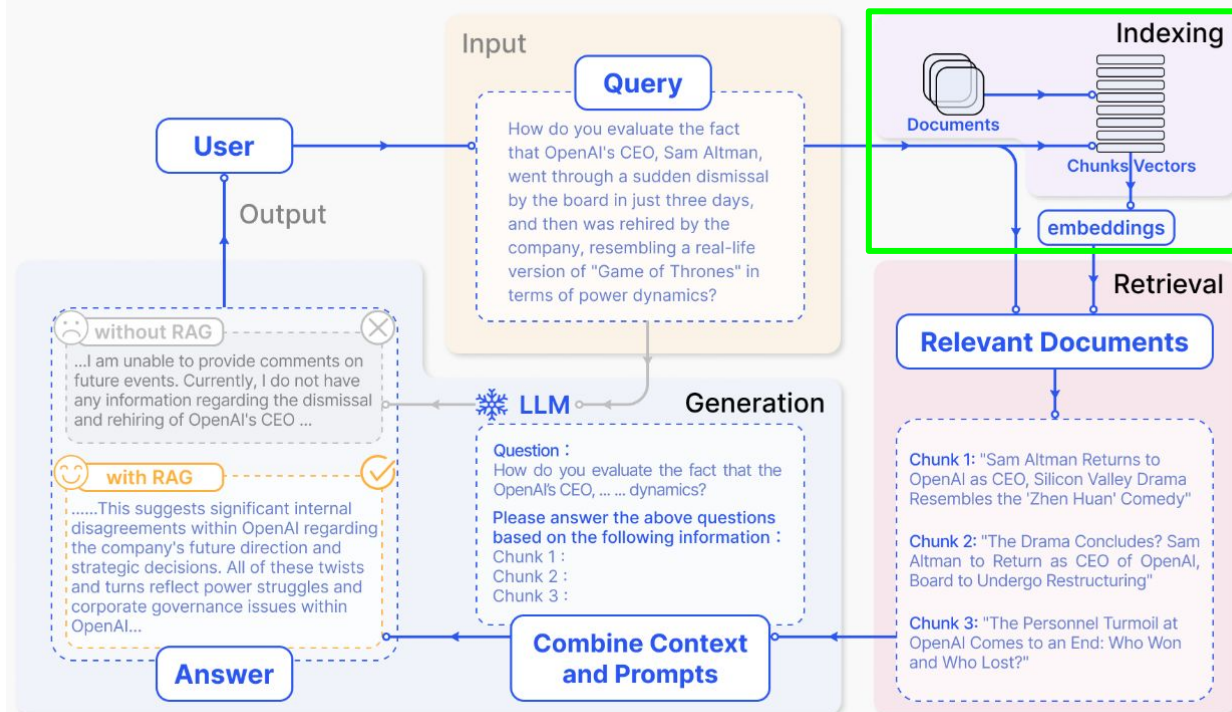
Retrieval Augmented Generation

# Retrieval Augmented Generation (RAG)





# Retrieval Augmented Generation (RAG)



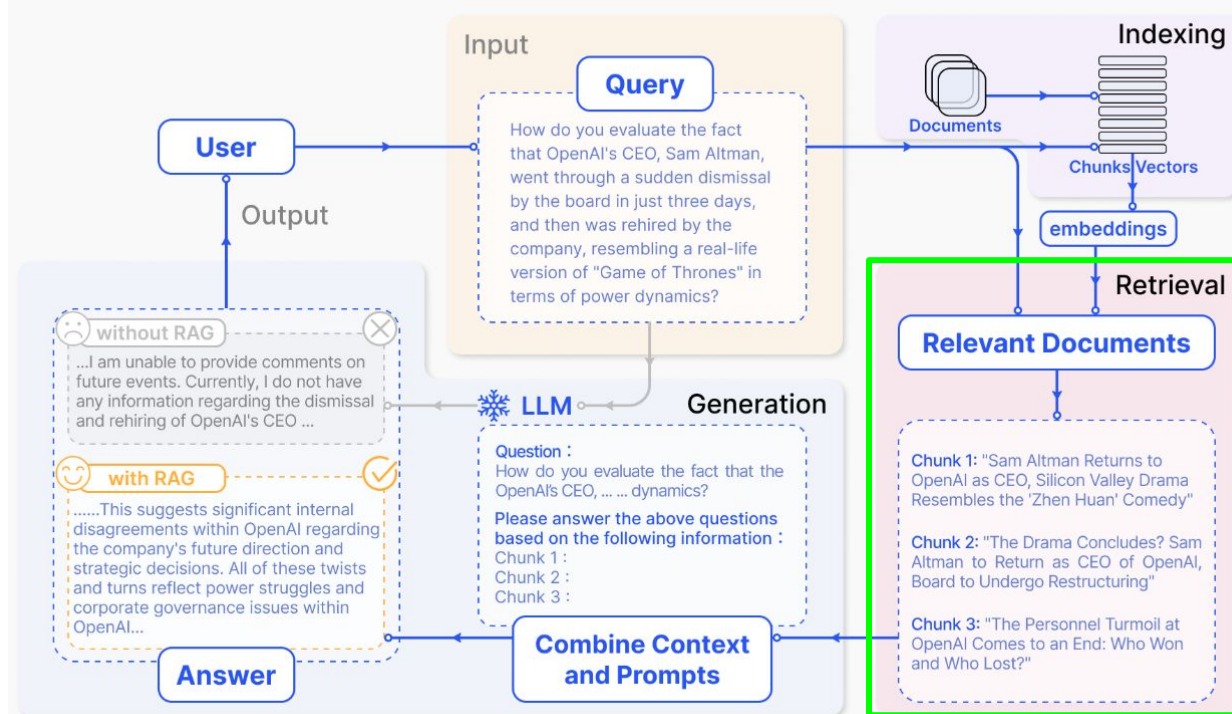
## Step 1: Indexing

# Retrieval Augmented Generation (RAG): Indexing

- Store documents to make them easier to retrieve
- Examples
  - Vector search, Graphs, Elasticsearch etc

Doc 1
Doc 2
Doc 3
Doc 4
Doc 5
Doc 6
...
Doc n

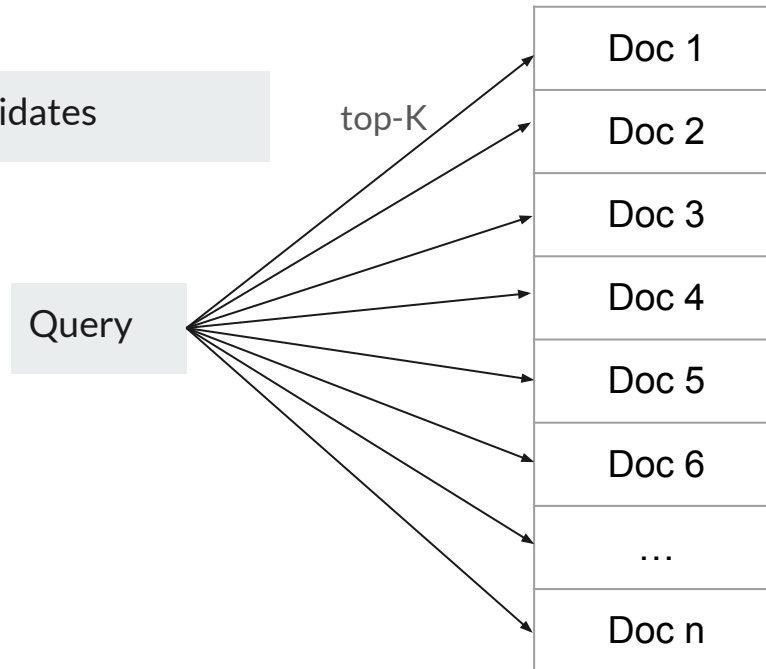
# Retrieval Augmented Generation (RAG)



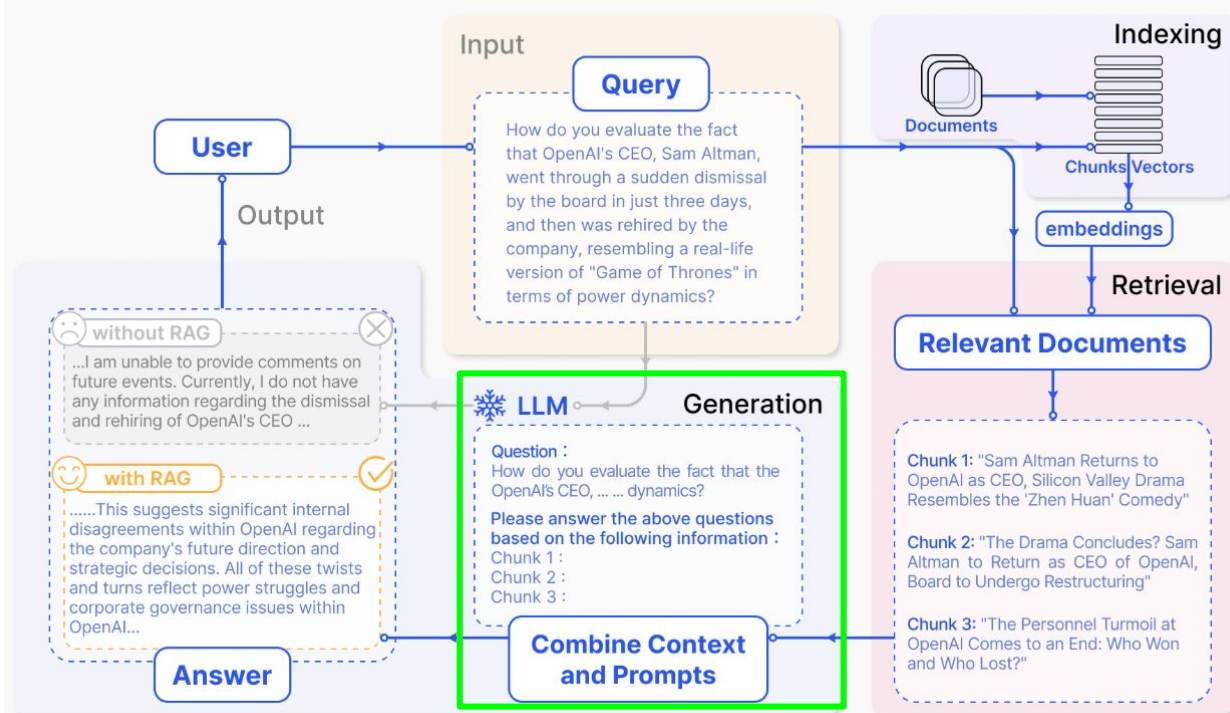
## Step 2: Retrieval

# Retrieval Augmented Generation (RAG): Retrieval

- Query the index and get top-K candidates



# Retrieval Augmented Generation (RAG)




Step 3: Generation

# Retrieval Augmented Generation (RAG): Generation

- Use the retrieved documents to answer the question

Answer the given question: {question} using  
the context: {context}.

# Retrieval Augmented Generation (RAG): Applications



Recommender Systems

Q/A Engine

Search Engine

Dialogue Systems

# Retrieval Augmented Generation (RAG): Paradigms



Naive RAG


- Naive RAG process: Indexing, retrieval, generation.
- Limitations: Low precision, recall, outdated information.
- Augmentation challenges: Redundancy, ranking, reconciling style.

Advanced RAG

Modular RAG



# Retrieval Augmented Generation (RAG): Paradigms



Naive RAG

Advanced RAG

- Advanced RAG improves retrieval quality.
- Pre-retrieval optimizes data indexing: chunk size, metadata etc.
- Retrieval stage enhances embedding model: fine-tuning.
- Post-retrieval focuses on noise handling: re-ranking, prompt compression

Modular RAG

# Retrieval Augmented Generation (RAG): Paradigms

Naive RAG

Advanced RAG

Modular RAG

- Naive and Advanced RAG are subsets of Modular RAG.
- Modules can be rearranged based on task needs.
- Optimization techniques for RAG pipelines include:
  - Hybrid Search Exploration
  - Recursive Retrieval and Query Engine
  - StepBack-prompt
  - Sub-Queries
  - Hypothetical Document Embeddings

# Retrieval Augmented Generation (RAG): Evaluation

- Evaluate retrieval and generation
- Retrieval
  - NDCG and Hit Rate from recommender systems
- Generation
  - Test faithfulness, relevance, and so on
- Manual and automatic tests
- RAG triad by **TruLens**
  - Context Relevance
    - Is the retrieved context relevant to the query?
  - Groundedness
    - Is the response supported by the context?
  - Answer Relevance
    - Is the answer relevant to the query?

# Retrieval Augmented Generation (RAG): Evaluation



- Three qualities and four abilities
  - Context Relevance
  - Faithfulness
  - Answer Relevance
  - Noise Robustness
  - Negative Rejection
  - Information Integration
  - Counterfactual Robustness
- Tools
  - TruLens, RAGAS
  - TruLens RAG triad

# Why vector search?

To capture semantic similarity

## Queries

Query 1: "Find me a cozy hotel near the beach with ocean views."

Query 2: "Discover a quaint seaside inn with a picturesque view of the sea."

Query 3: "Locate a charming beachside lodge offering scenic ocean vistas."

## Documents

**Title:** "Beachfront Resort: Enjoy Stunning Ocean Views"

**Description:** "Our beachfront resort offers cozy rooms with breathtaking views of the ocean. Relax and unwind in our serene coastal setting."

**Title:** "Seaside Inn: Experience Tranquility by the Sea"

**Description:** "Nestled along the coast, our seaside inn provides a quaint retreat with panoramic views of the sea. Escape the hustle and bustle and immerse yourself in the beauty of nature."

**Title:** "Charming Beachside Lodge: Your Gateway to Coastal Bliss"

**Description:** "Indulge in the rustic charm of our beachside lodge, where every room offers stunning vistas of the ocean. Experience the magic of seaside living and create lasting memories."

# Vector Search: Process



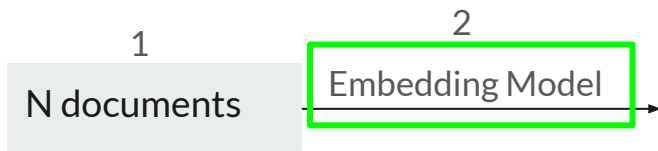
1

N documents

Chunk Document into multiple chunks

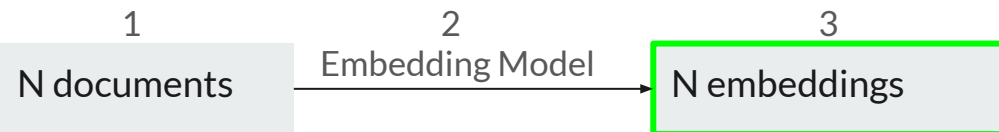
- Fixed length
- Sentence level

# Vector Search: Process



- Convert text into embedding
- Open source/self-hosting vs APIs
- Same embedding model for indexing and search
- Factors to decide: context length, dimension, quality

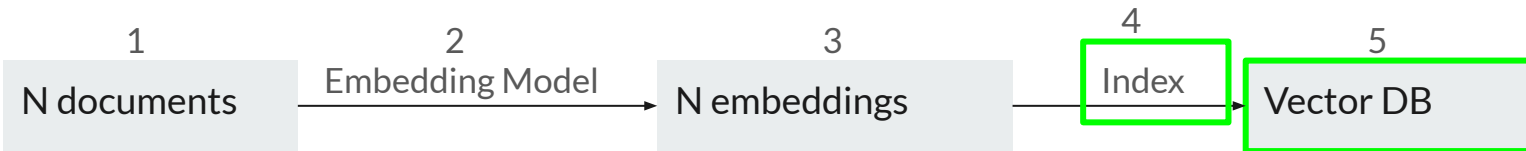
# Vector Search: Process



Get N embeddings from N documents

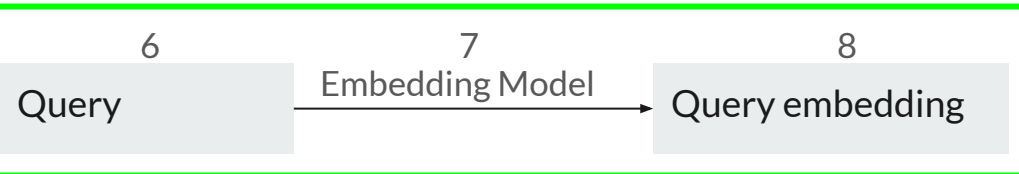
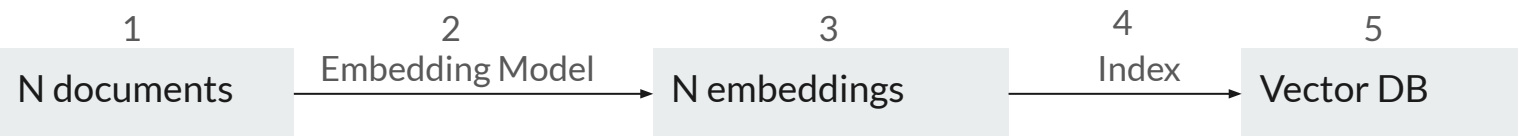


# Vector Search: Process



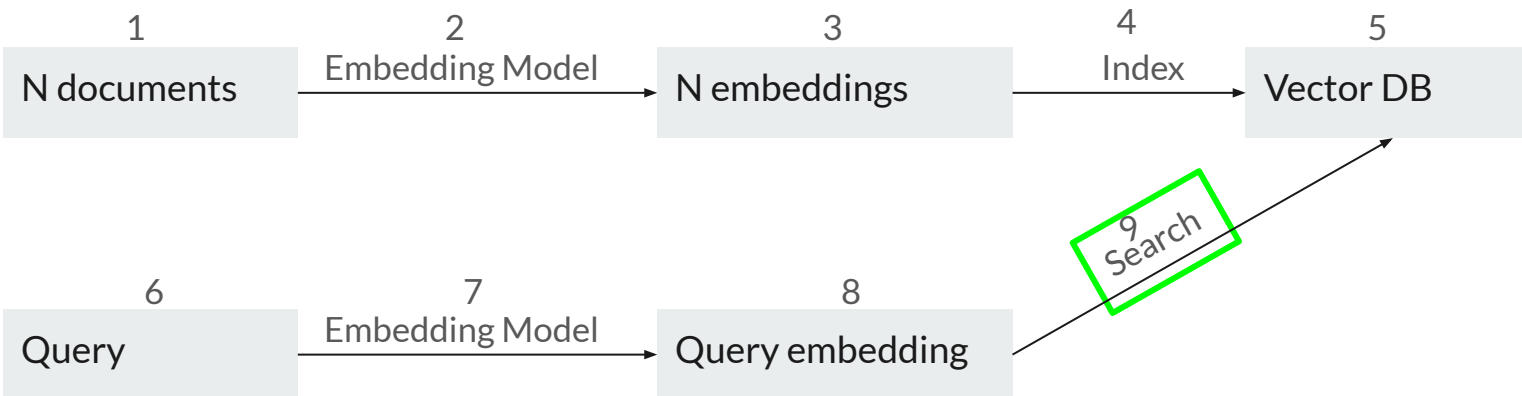
Pinecone, Chroma  
Open source/self-host vs managed solution

# Vector Search: Process



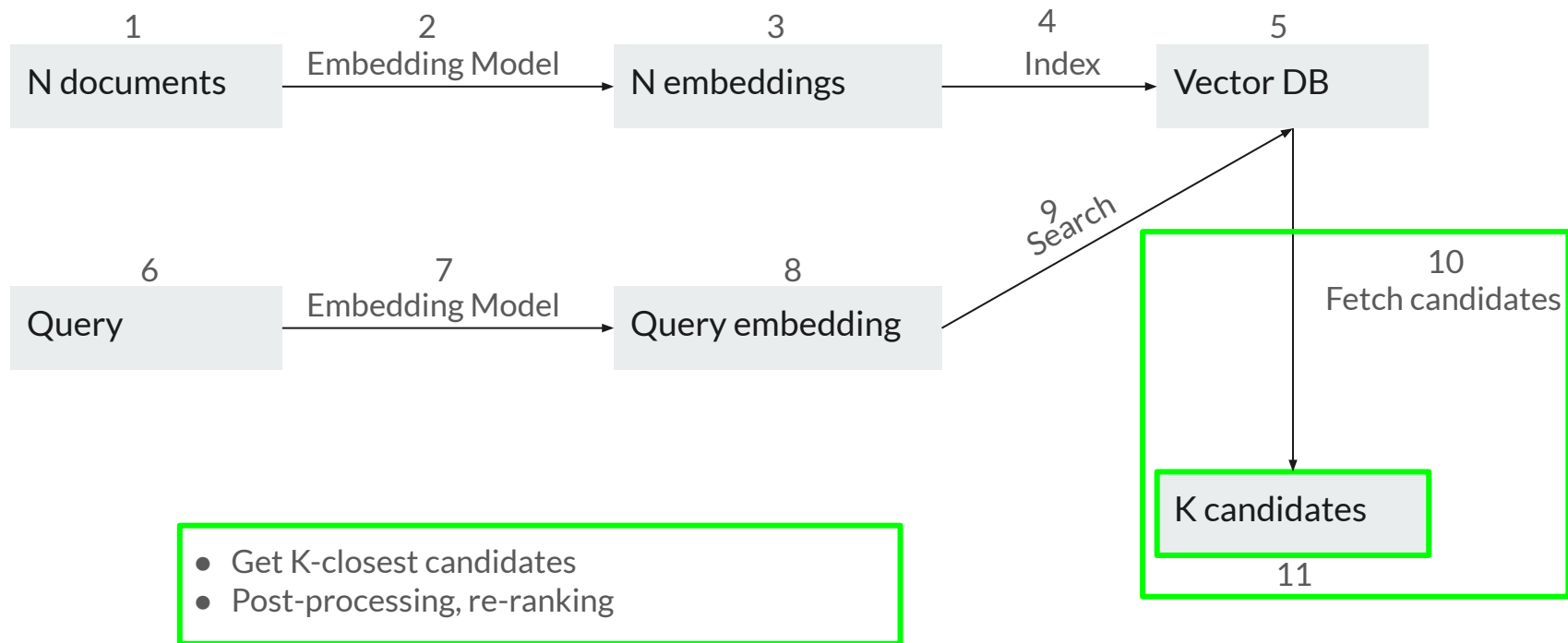
- Same embedding model
- Generate query embedding

# Vector Search: Process

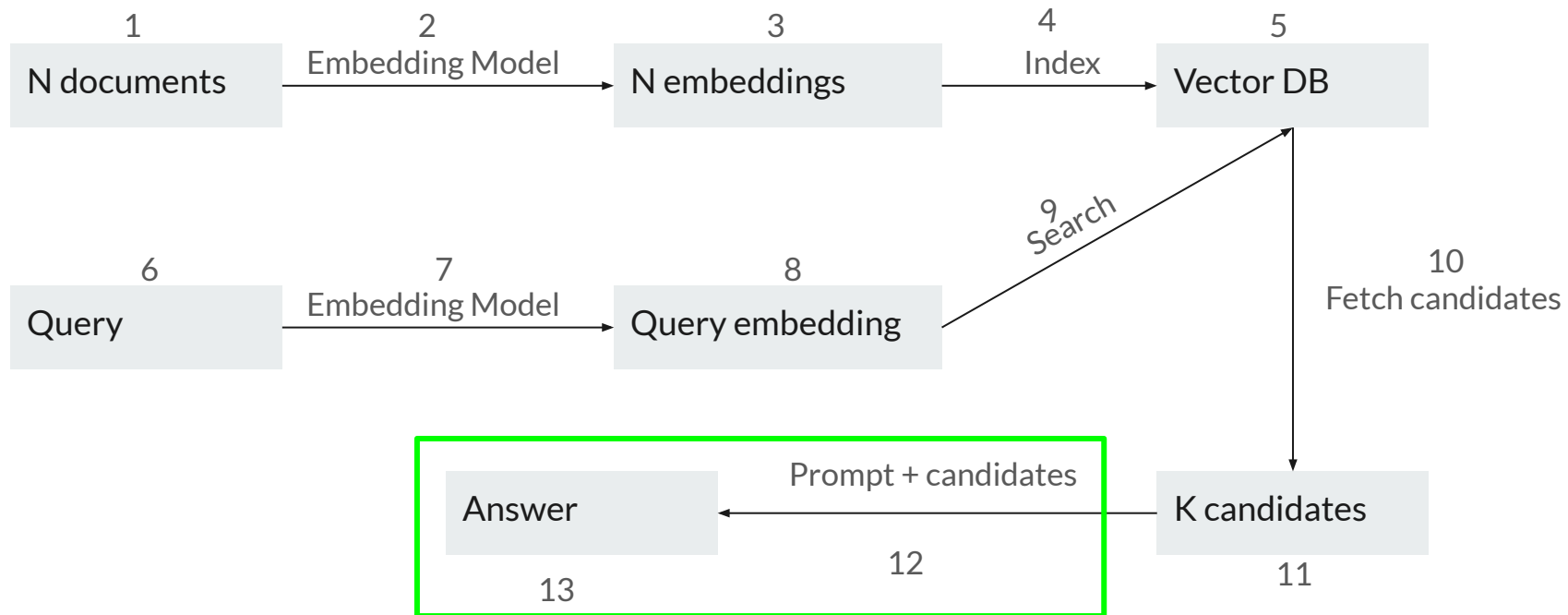


- Compare query vector with indexed vectors
- Cosine similarity, euclidean distance
- Filters, pre-processor

# Vector Search: Process

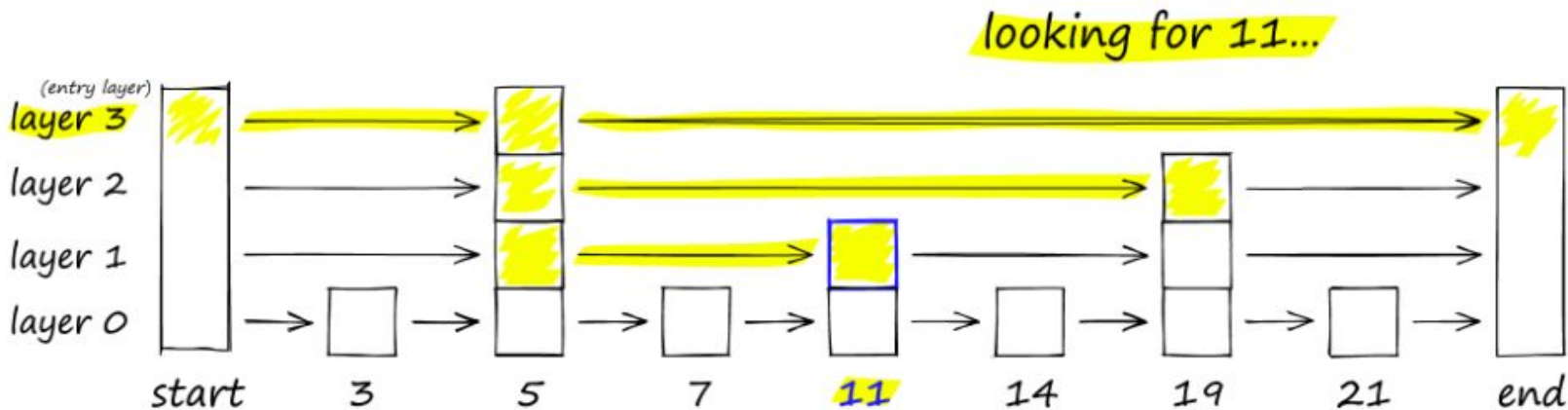


# Vector Search: Process



# Vector Search: Algorithm (HNSW)

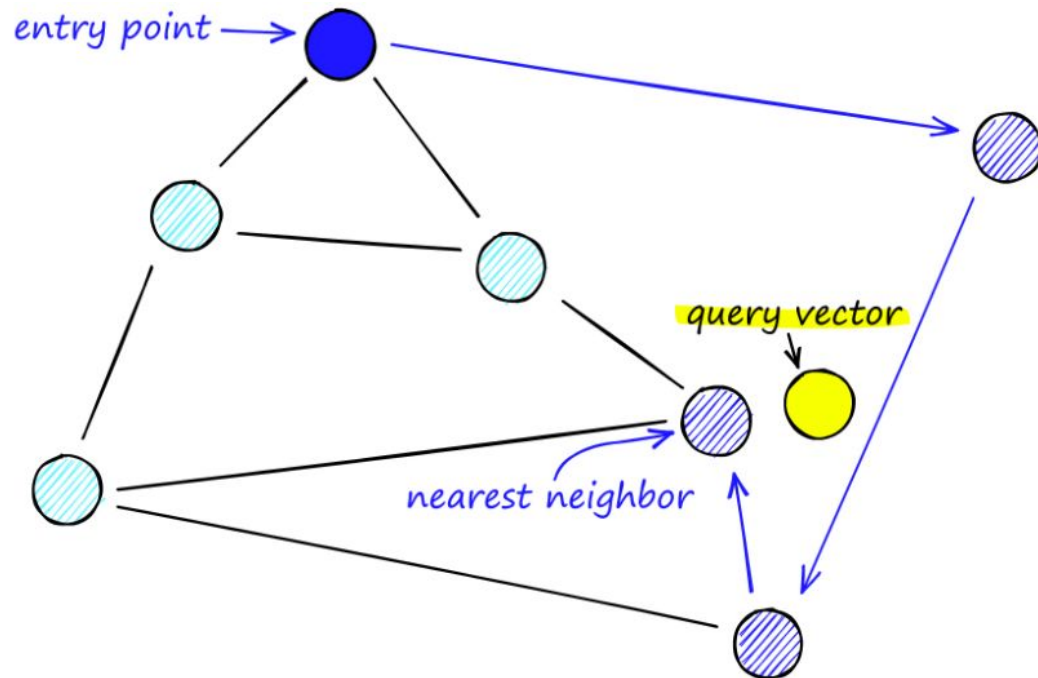
- Hierarchical Navigable Small Worlds
- Proximity graph
- Combination of probability skip list and Navigable Small World (SNW) graphs



A probability skip list structure, we start on the top layer. If our current key is greater than the key we are searching for (or we reach end), we drop to the next layer.

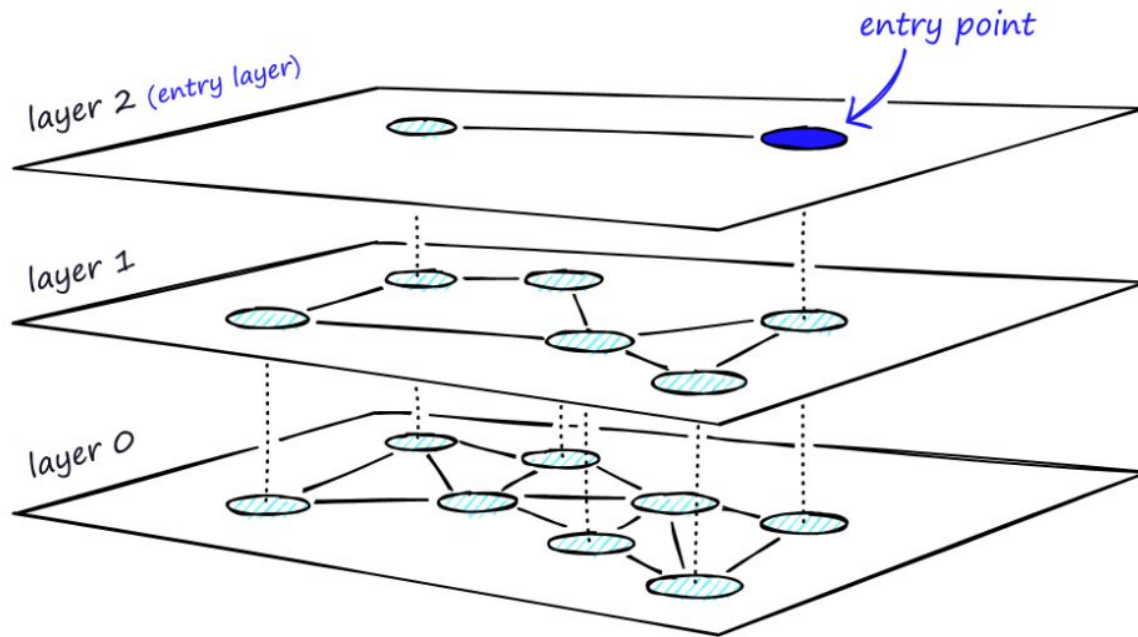
Probability Skip Lists

# Vector Search: Algorithm (HNSW)



The search process through a NSW graph. Starting at a pre-defined entry point, the algorithm greedily traverses to connected vertices that are nearer to the query vector.

# Vector Search: Algorithm (HNSW)



Layered graph of HNSW, the top layer is our entry point and contains only the longest links, as we move down the layers, the link lengths become shorter and more numerous.

Hierarchical Navigable Small Worlds (HNSW)



# Retrieval Augmented Generation: Demo



# References



1. <https://arxiv.org/abs/2312.10997>
2. <https://arxiv.org/abs/2005.11401>
3. <https://www.pinecone.io/learn/series/faiss/hnsw/>
4. <https://pub.towardsai.net/advanced-rag-techniques-an-illustrated-overview-04d193d8fec6>
5. <https://learn.deeplearning.ai/courses/building-evaluating-advanced-rag/lesson/4/sentence-window-retrieval>



**Thank You**

**Q/A**