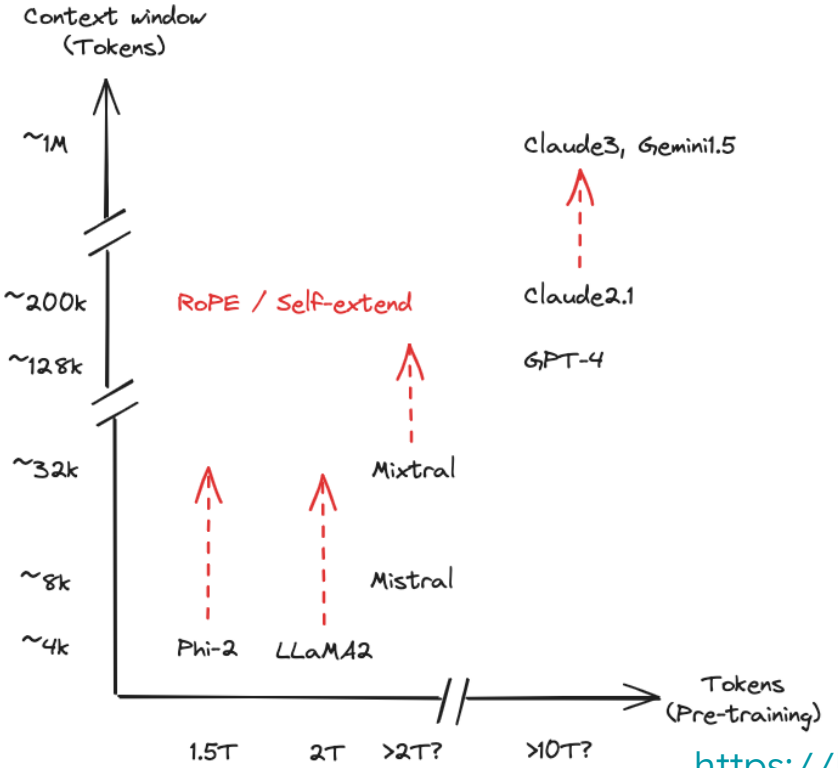
LangChain

# RAG from scratch: Overview

Lance Martin
Software Engineer, LangChain
@RLanceMartin

LangChain

> 95% of the world's data is "private", but we can "feed it" to LLMs

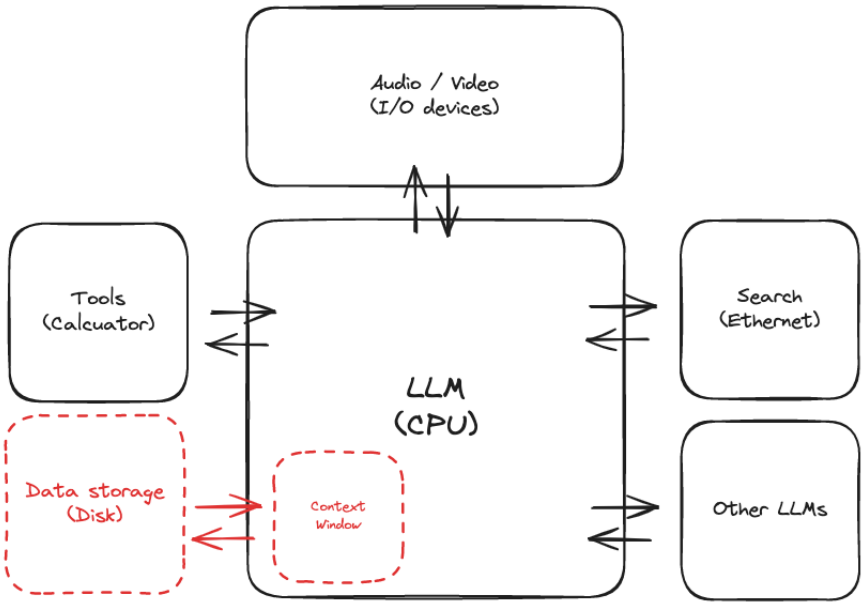



https://huggingface.co/blog/mixtral

https://x.com/RihardJarc/status/1778082161595208124

LangChain

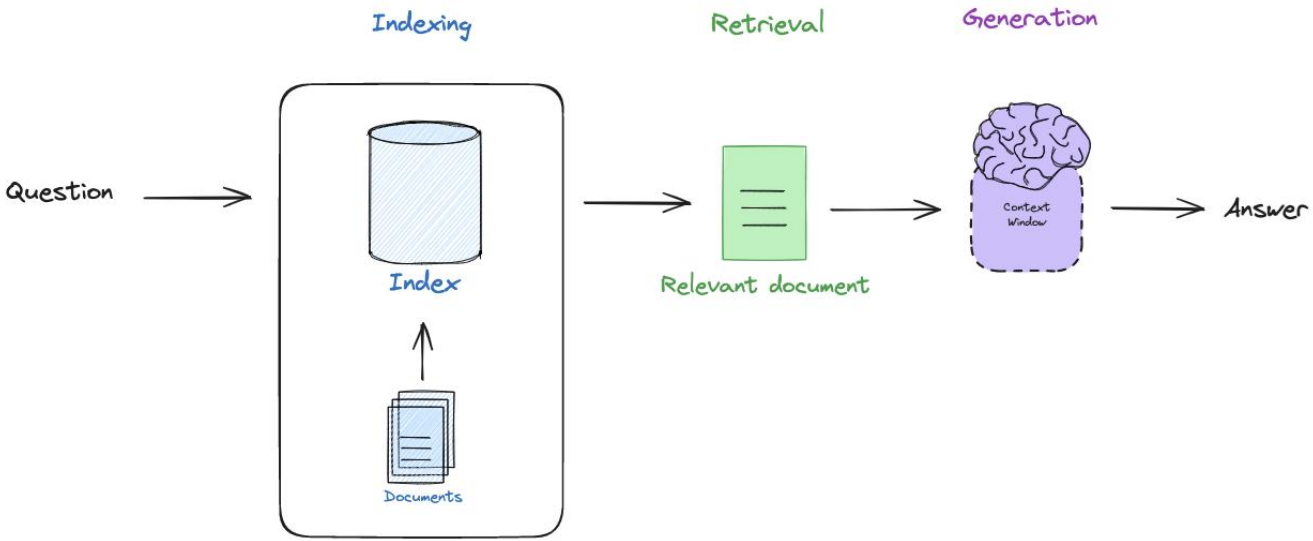
# Connecting LLMs to external data is a central need




https://x.com/karpathy/status/1707437820045062561?s=20

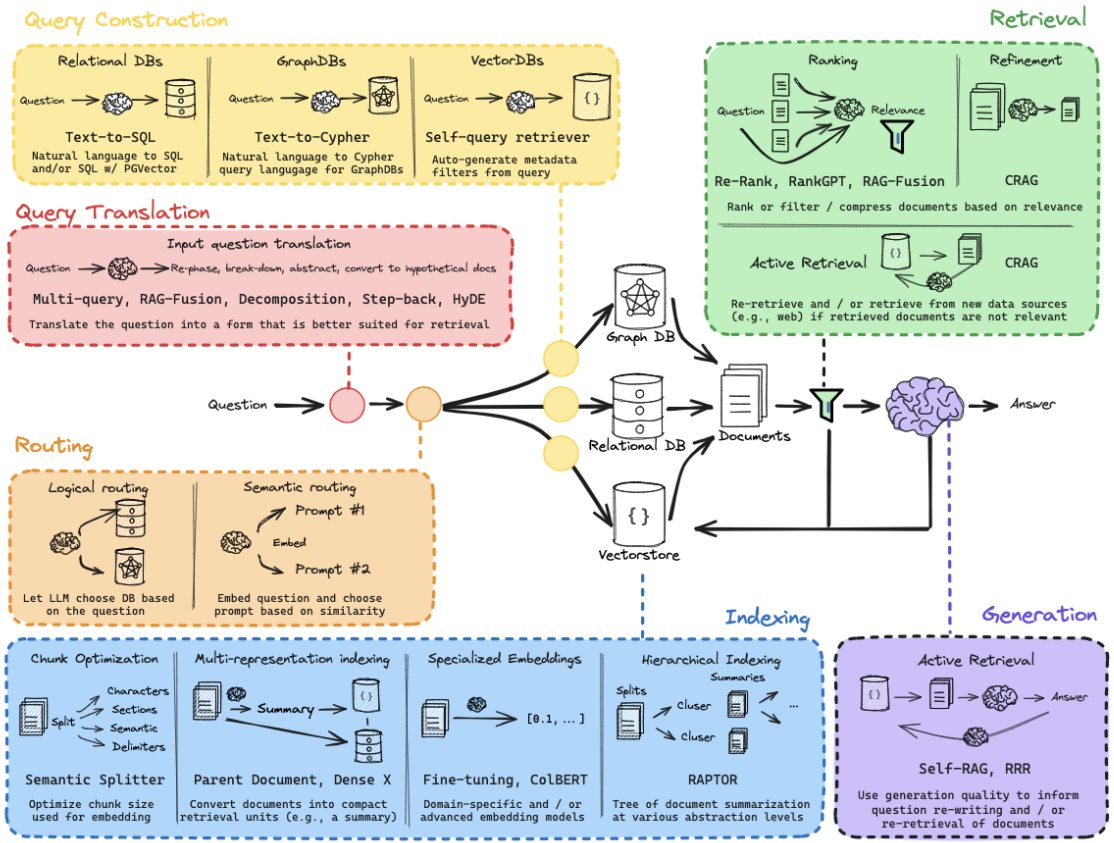
LangChain

# **R**etrieval **A**ugmented **G**eneration



LangChain

# We'll start from scratch, and build up to state of the art in RAG



LangChain

# Outline

**Basics**

- Indexing
- Retrieval
- Generation

**Advanced**

- Query transformations
- Routing
- Query construction
- Indexing
- Retrieval
- Generation

Code walk-through

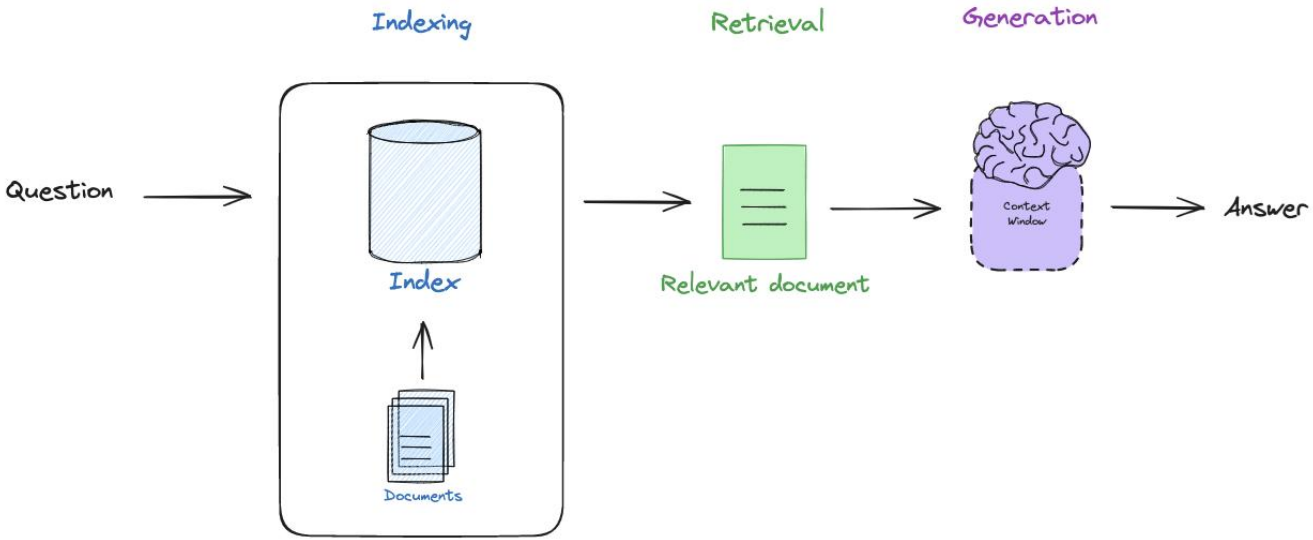
LangChain

# RAG from scratch: Indexing

Lance Martin
Software Engineer, LangChain
@RLanceMartin

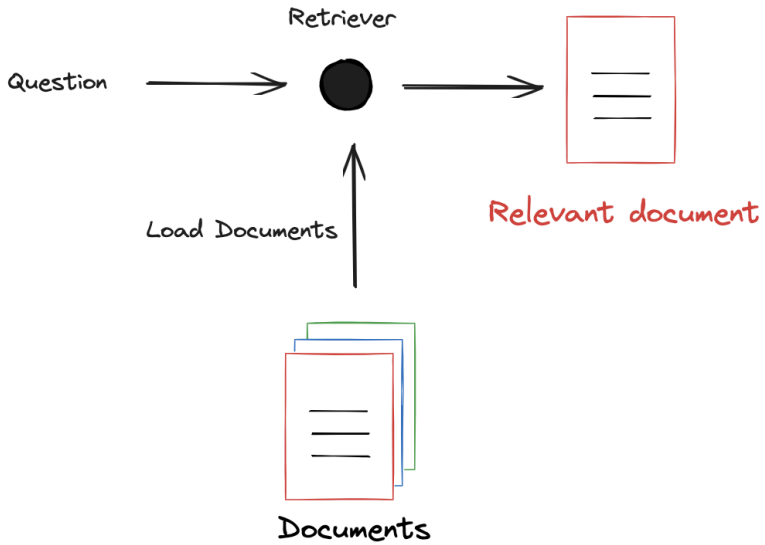
LangChain

# RAG motivation



LangChain

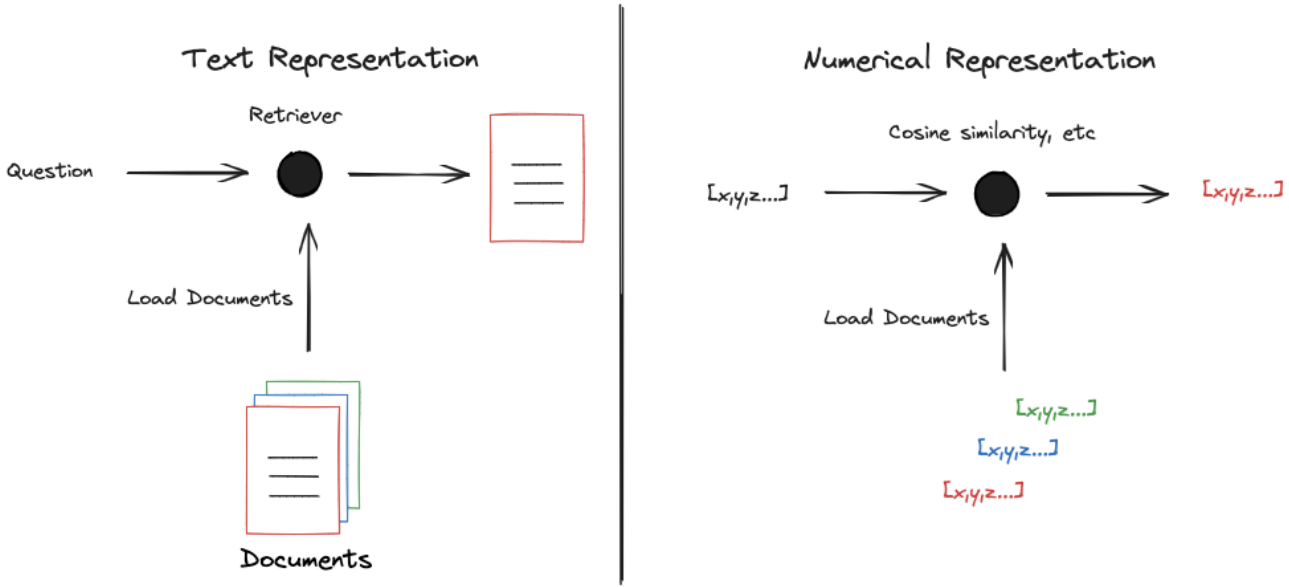
# Document loading




https://integrations.langchain.com/

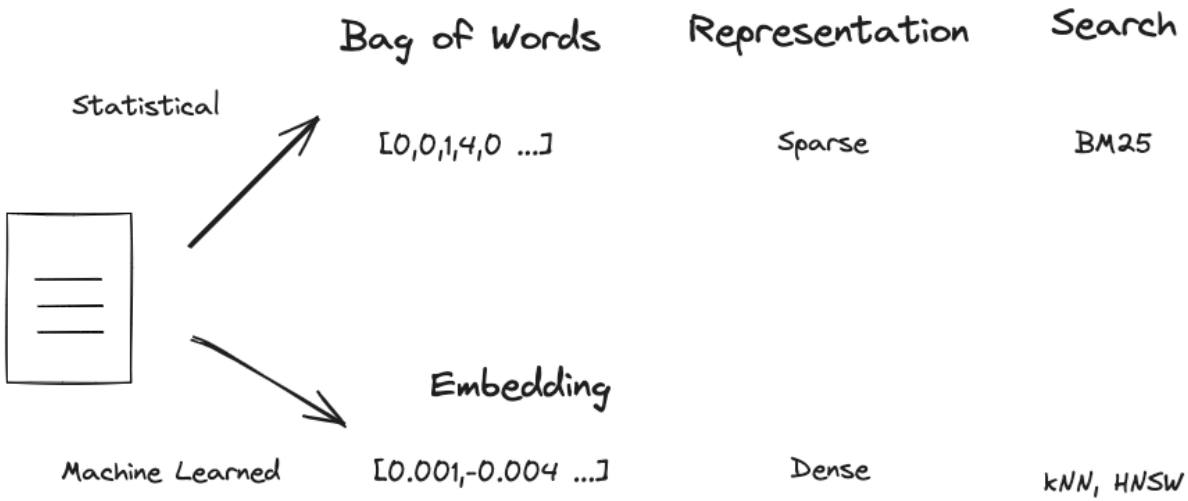
LangChain

# Numerical representation for search



LangChain

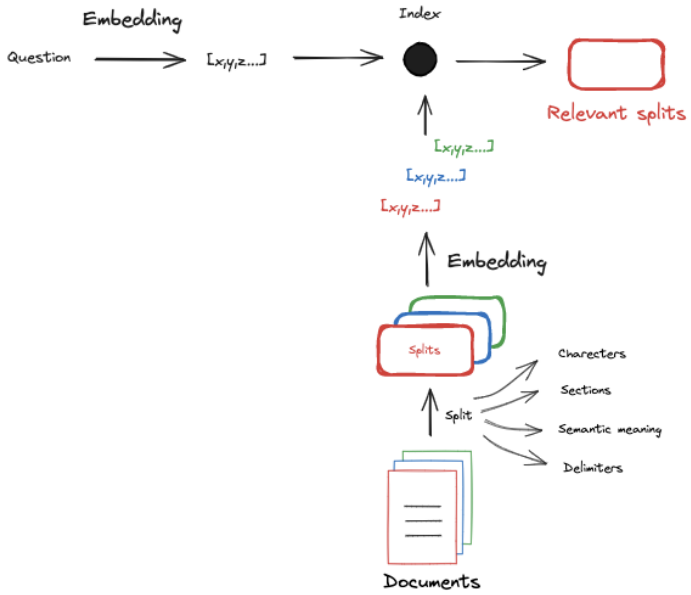
# Statistical and machine learned representations




https://simonwillison.net/2023/Oct/23/embeddings/
https://www.pinecone.io/learn/series/nlp/dense-vector-embeddings-nlp/
https://cameronrwolfe.substack.com/p/the-basics-of-ai-powered-vector-search?utm_source=profile&utm_medium=reader2

LangChain

# Loading, splitting, and embedding




https://chunkviz.up.railway.app/
https://www.youtube.com/watch?v=8OJC21T2SL4

Code walk-through

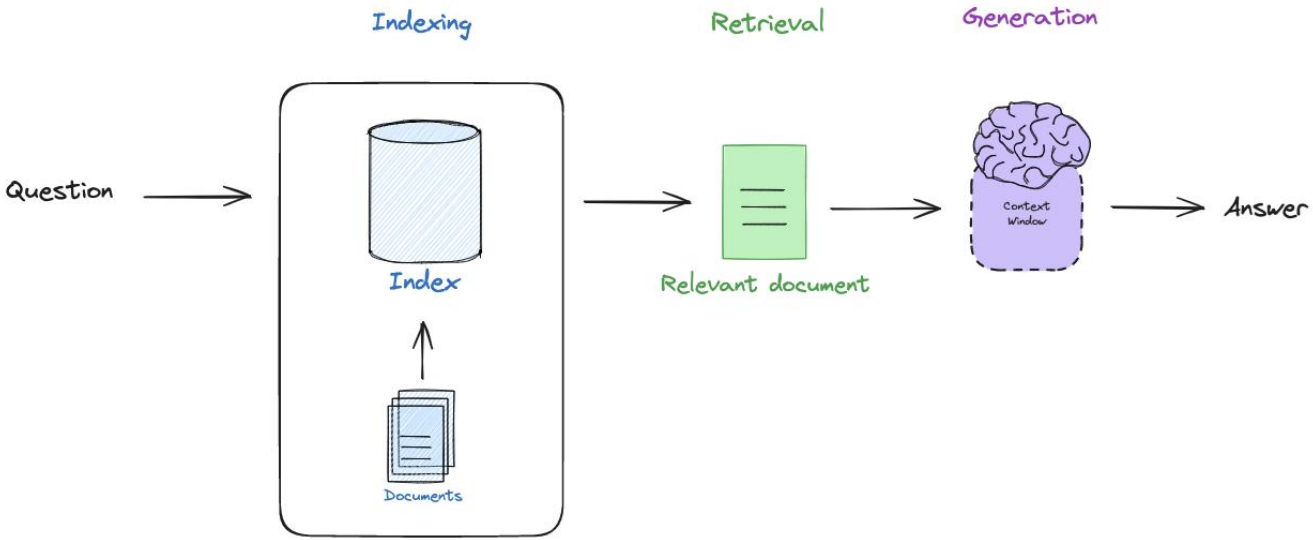
LangChain

# RAG from scratch: Retrieval

Lance Martin
Software Engineer, LangChain
@RLanceMartin

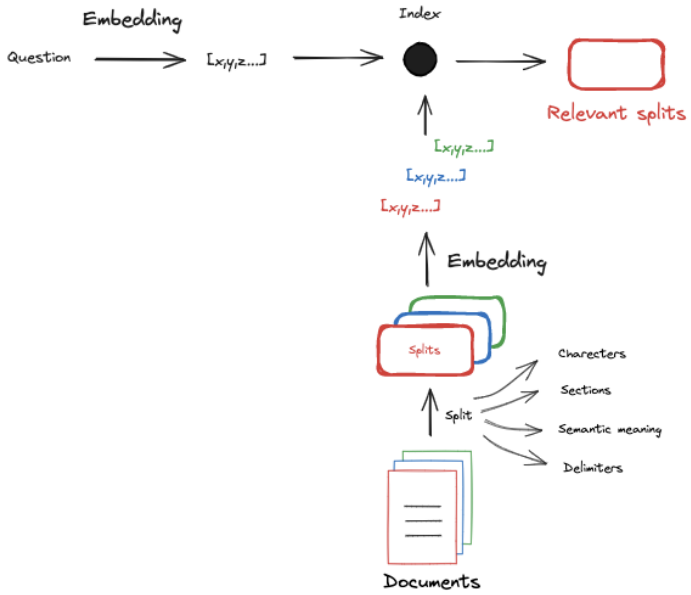
LangChain

# RAG motivation



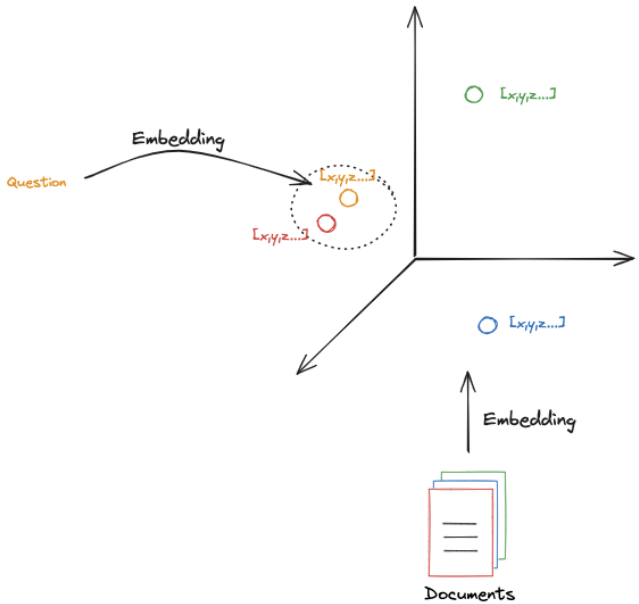
LangChain

# Index makes documents easy to retrieve



LangChain

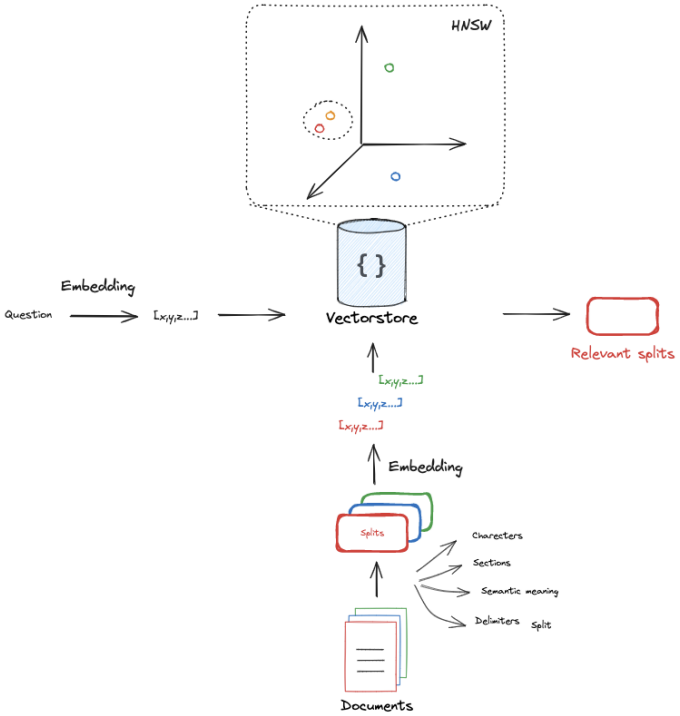
# Retrieval powered via similarity search




https://simonwillison.net/2023/Oct/23/embeddings/

LangChain

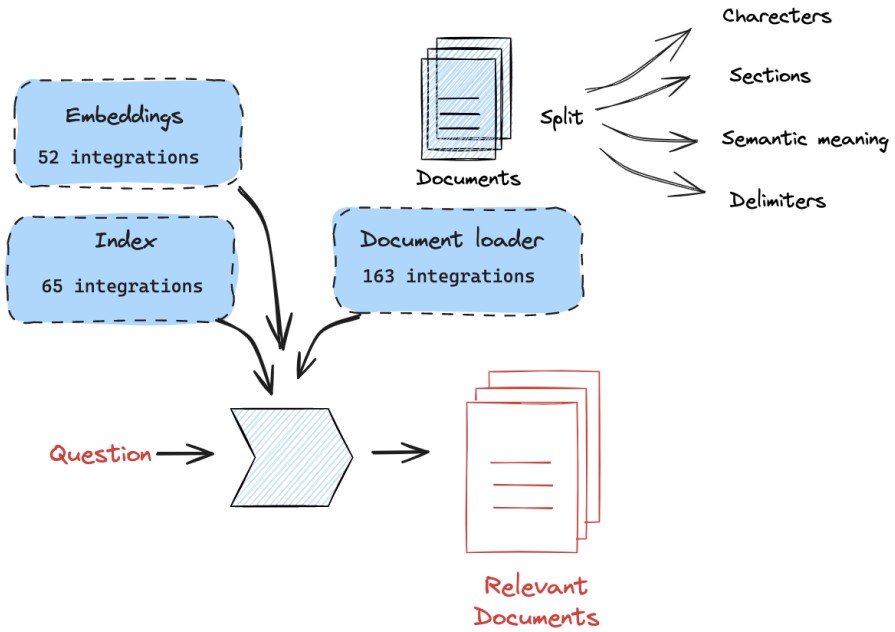
# Vectorstores implement this for you




https://www.pinecone.io/learn/series/faiss/hnsw/

LangChain

# LangChain has many integrations to support this




https://integrations.langchain.com/
https://vdbs.superlinked.com/

Code walk-through

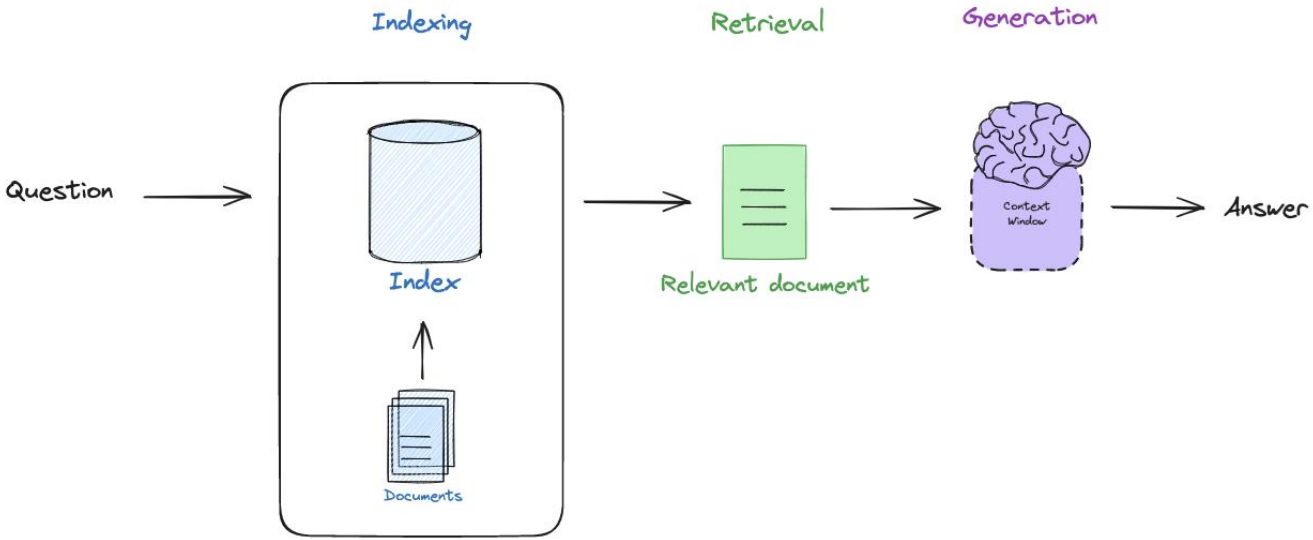
LangChain

# RAG from scratch: Generation

Lance Martin
Software Engineer, LangChain
@RLanceMartin

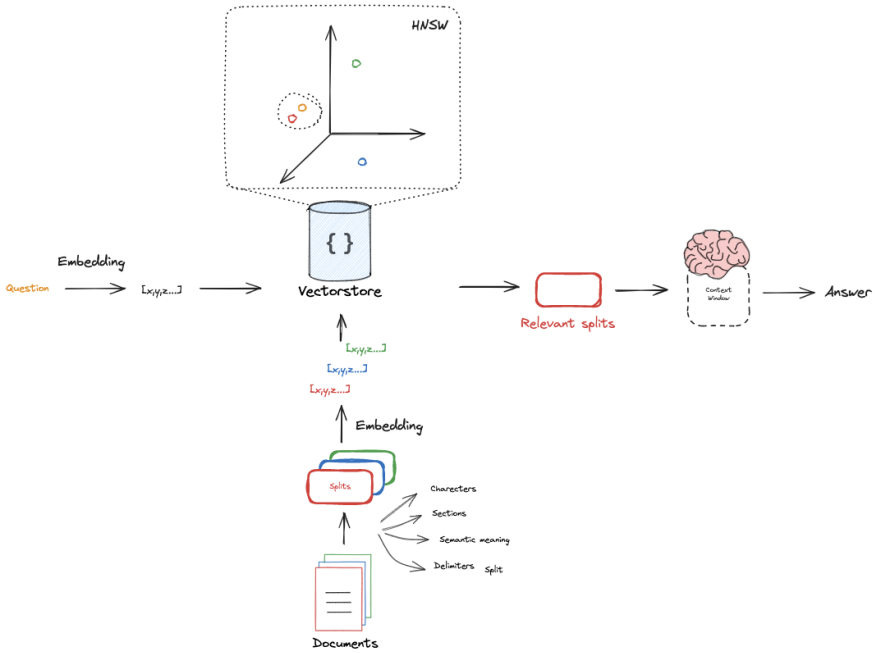
LangChain

# RAG motivation



LangChain

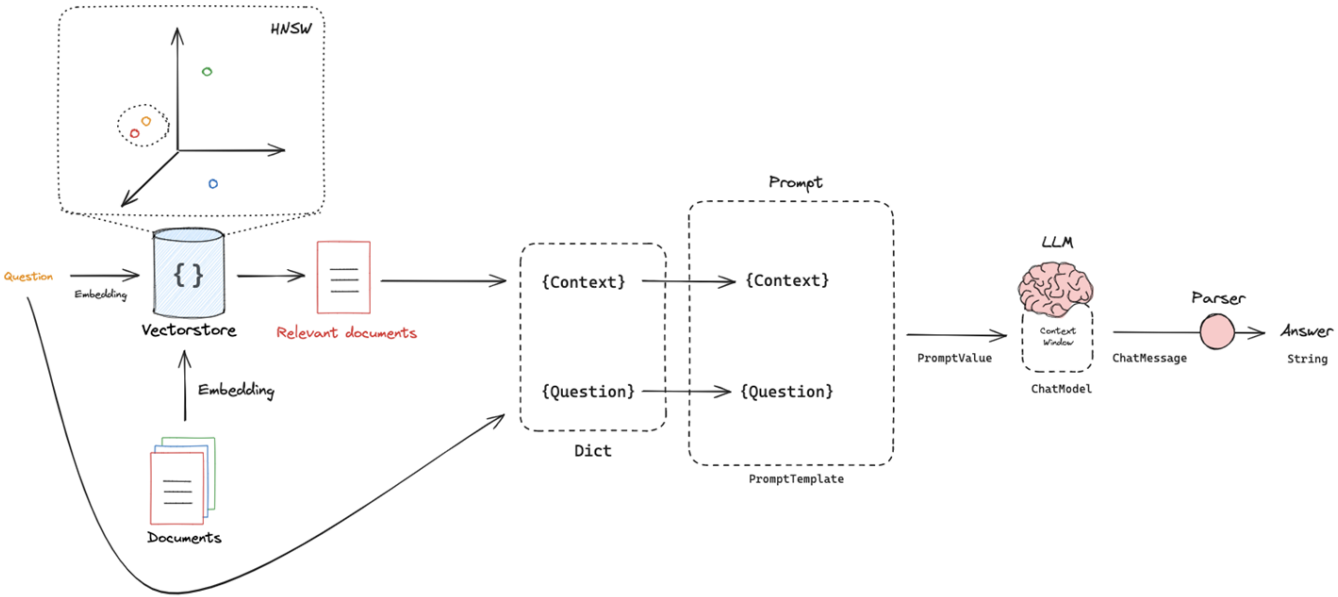
# Adding docs to context window




https://python.langchain.com/docs/use_cases/question_answering/

LangChain

# Connecting retrieval with LLMs via prompt




https://python.langchain.com/docs/expression_language/get_started

https://smith.langchain.com/hub/rlm/rag-prompt?organizationId=1fa8b1f4-fcb9-4072-9aa9-983e35ad61b8

Code walk-through