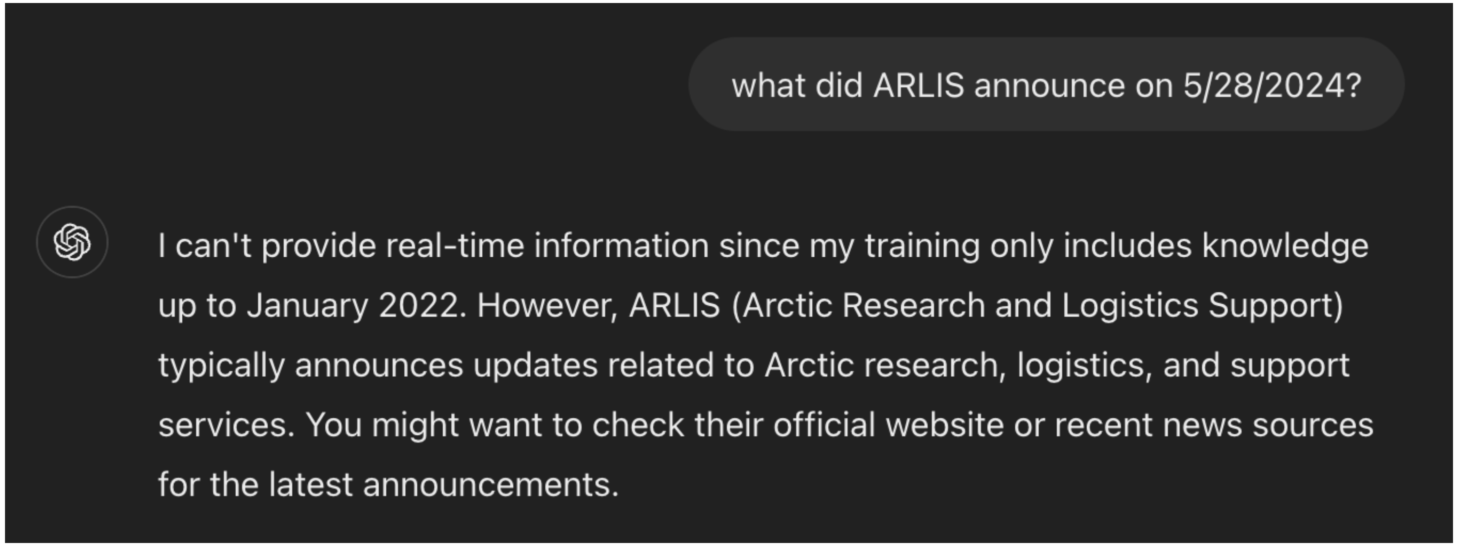
Introduction to Retrieval Augmented Generation (RAG)

Alan B McMillan, PhD

University of Wisconsin

6/4/2024

What is Retrieval Augmented Generation?



what did ARLIS announce on 5/28/2024?

I can't provide real-time information since my training only includes knowledge up to January 2022. However, ARLIS (Arctic Research and Logistics Support) typically announces updates related to Arctic research, logistics, and support services. You might want to check their official website or recent news sources for the latest announcements.

ChatGPT 3.5

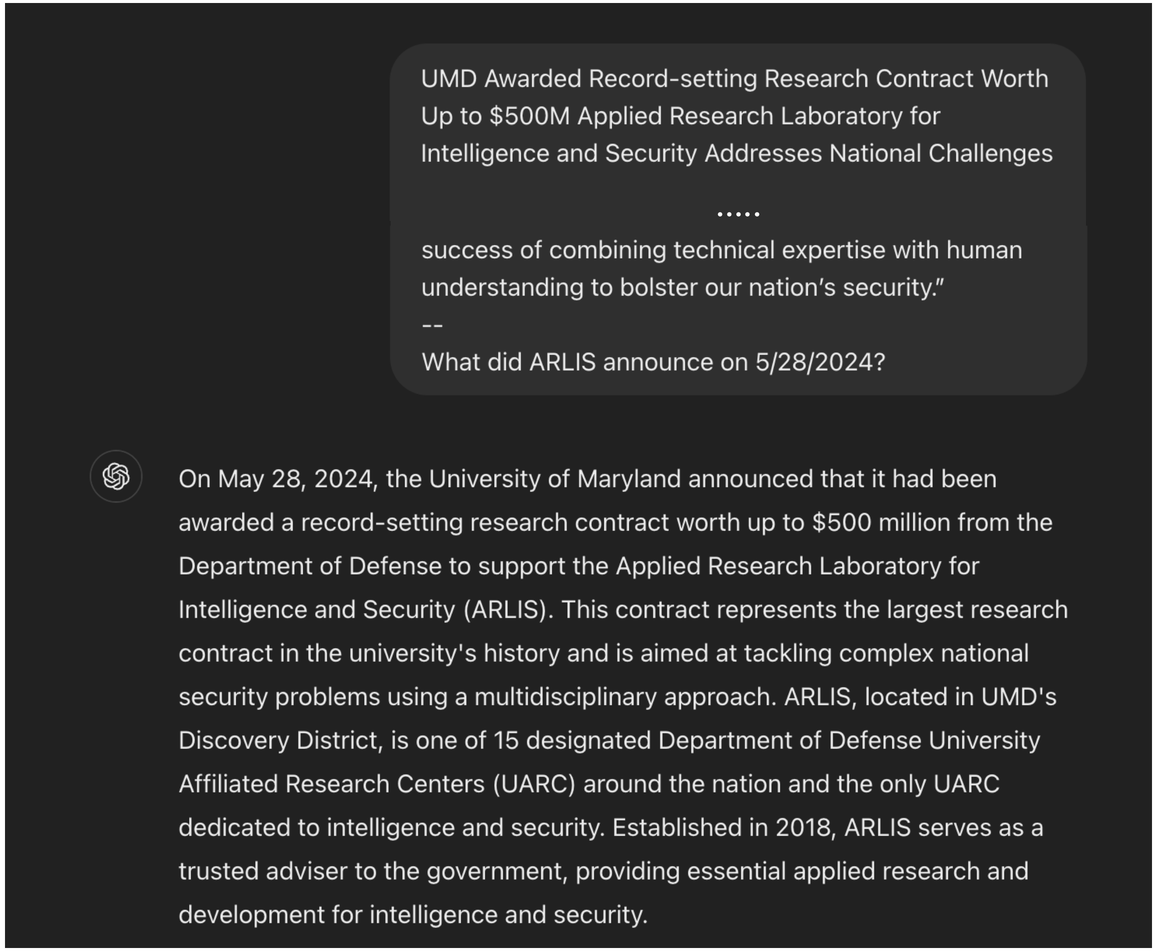
- Enhances the performance of a large language model by providing additional information and context.

Why is RAG Important?

- Enhances AI capabilities
  - Improves accuracy and relevance of generated responses
- LLMs have limited “context length”
  - We can only put in so much data into a prompt
    - This limit is rapidly increasing in 2024
- Real-world applications
  - Applications in various domains
    - Customer service
    - Document Q&A
    - Real-time data integration
    - Search engine
    - Personalized recommendation

Let’s try our example again

- Instead, first copy the text of a relevant news article: [UMD Awarded Record-setting Research Contract Worth Up to $500M | https://today.umd.edu/umd-awarded-record-setting-research-contract-worth-up-to-500m]
- Then ask our question.



UMD Awarded Record-setting Research Contract Worth Up to $500M Applied Research Laboratory for Intelligence and Security Addresses National Challenges

.....

success of combining technical expertise with human understanding to bolster our nation’s security.”

--

What did ARLIS announce on 5/28/2024?

On May 28, 2024, the University of Maryland announced that it had been awarded a record-setting research contract worth up to $500 million from the Department of Defense to support the Applied Research Laboratory for Intelligence and Security (ARLIS). This contract represents the largest research contract in the university's history and is aimed at tackling complex national security problems using a multidisciplinary approach. ARLIS, located in UMD's Discovery District, is one of 15 designated Department of Defense University Affiliated Research Centers (UARC) around the nation and the only UARC dedicated to intelligence and security. Established in 2018, ARLIS serves as a trusted adviser to the government, providing essential applied research and development for intelligence and security.

ChatGPT 3.5

How does RAG work?

1. Input Query
   - User provides an initial query or prompt.
2. Retrieve Relevant Information
   - Search external databases or documents for relevant data.
   - Utilize vector databases for efficient similarity searches.
3. Combine Retrieved Information
   - Select the most relevant chunks of information.
4. Integrate this information into an enhanced prompt for the LLM
   - The LLM processes the enhanced prompt.
   - Generates a response that incorporates the additional context.

Let’s Revisit Context Window Quickly

- LLMs have limited input and output capability
  - GPT 3.5: 16k tokens in, 4k tokens out
  - GPT 4o: 128k tokens in, 4k tokens out
- Some models have much larger input context
  - Claude 3: up to 1M tokens in, 4k tokens out
  - Gemini 1.5 Pro: up to 2M tokens in, 8k tokens out
- We can put in increasingly huge amounts of data, but not get huge amounts out.
  - For example, we cannot translate an entire book or document to another language in one step
- So for the moment (June 2024), RAG remains very important

RAG works at scale

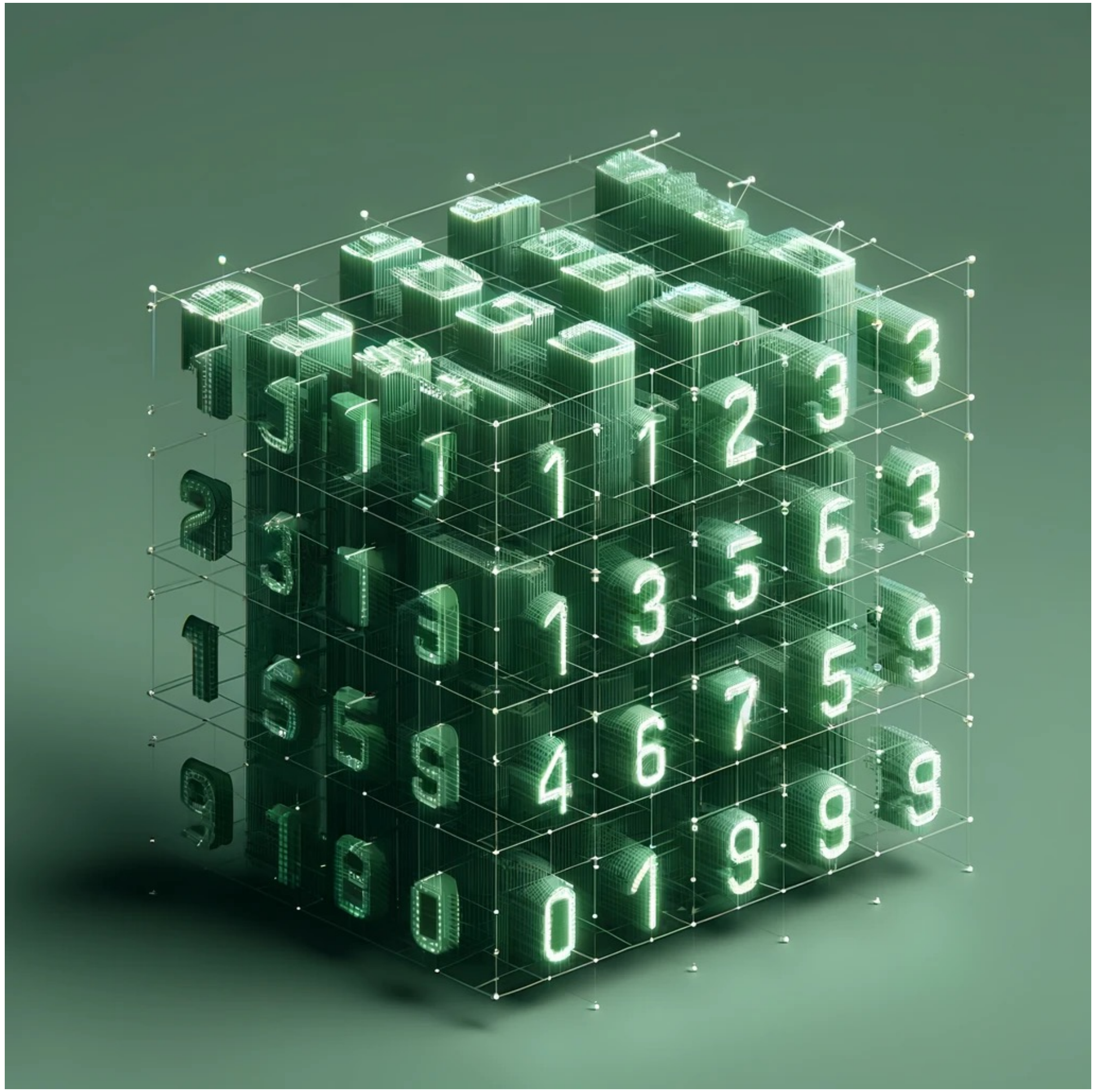
- Easy if we have a small document or set of information
- Imagine a huge dataset of documents
  - We need to spend time finding the relevant sections
  - This is where we need a RAG system
- RAG Ingredients
  - LLM(s)
  - Tokenizer/Embeddings
  - (Vector) Database

What are Tokens?

- We need to prepare text for further processing by machine learning models
  - Break down text into smaller units called tokens.
  - Tokens can be words, subwords, or characters.
- Types of Tokenization
  - Word Tokenization: splits text into individual words.
    - Example: "National security" → ["National", "security"]
  - Subword Tokenization: splits text into meaningful subunits.
    - Example: "National" → ["Nation", "al"]
  - Character Tokenization: splits text into individual characters.
    - Example: "NS" → ["N", "a", "t", "i", "o", "n", "a", "l"]

What are Embeddings?

- Embeddings are numerical representations of text tokens
  - Input Text -> Tokens -> Embeddings -> LLM
  - Can capture semantic meaning and relationships between words
  - The format that machine learning models can process
  - Facilitates tasks like text classification, sentiment analysis, and more



Creating Embeddings

- Represent tokens as vectors in a continuous vector space.
  - Example: "National" → [0.25, -0.13, 0.40, ...] (simplified representation)
- Example Techniques:
  - Bag of Words: Simple representation counting word occurrences
  - Word2Vec: Uses neural networks to learn word associations
  - BERT (Bidirectional Encoder Representations from Transformers): Contextual embeddings capturing meaning based on surrounding words
  - OpenAI, HuggingFace, and open-source embedding models
- Embeddings can encode contextual meaning!
- Are language and task specific (e.g., code, mathematics)

Tokens, Embeddings, Now What?

- Importance of “Chunking” Documents
  - We either cannot or do not need to put the entire document into our augmented prompt
  - Need manageable and relevant data chunks
  - Done before tokenization/embedding
- Challenges
  - Maintaining context and coherence.
  - Balancing chunk size for performance and computational limitations

Chunking Strategies

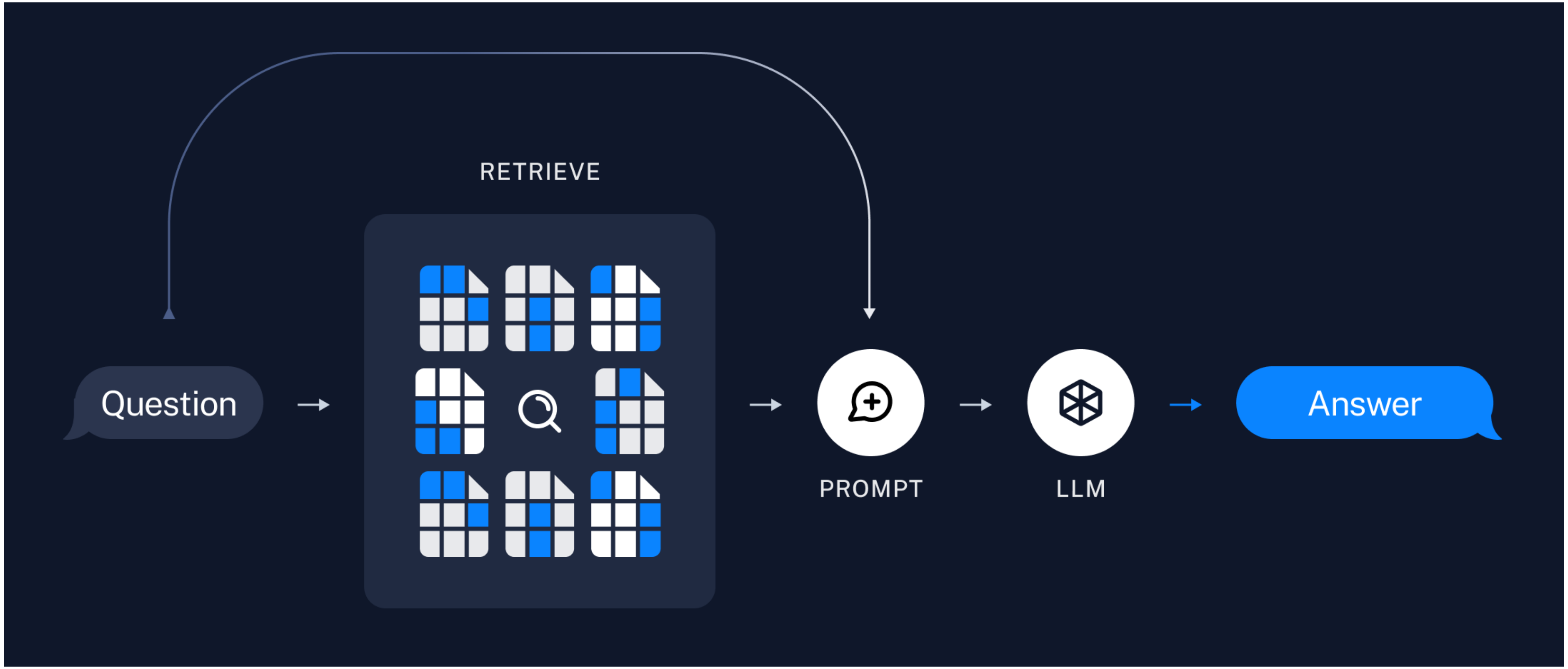
- Fixed
  - Can be overlapping
- Document
  - Page, section, etc.
  - Works well only with well-formatted
- Semantic
  - Split document into sections based on similarity of content
  - Can be computationally expensive

https://github.com/FullStackRetrieval-com/RetrievalTutorials/

Storing our data for RAG

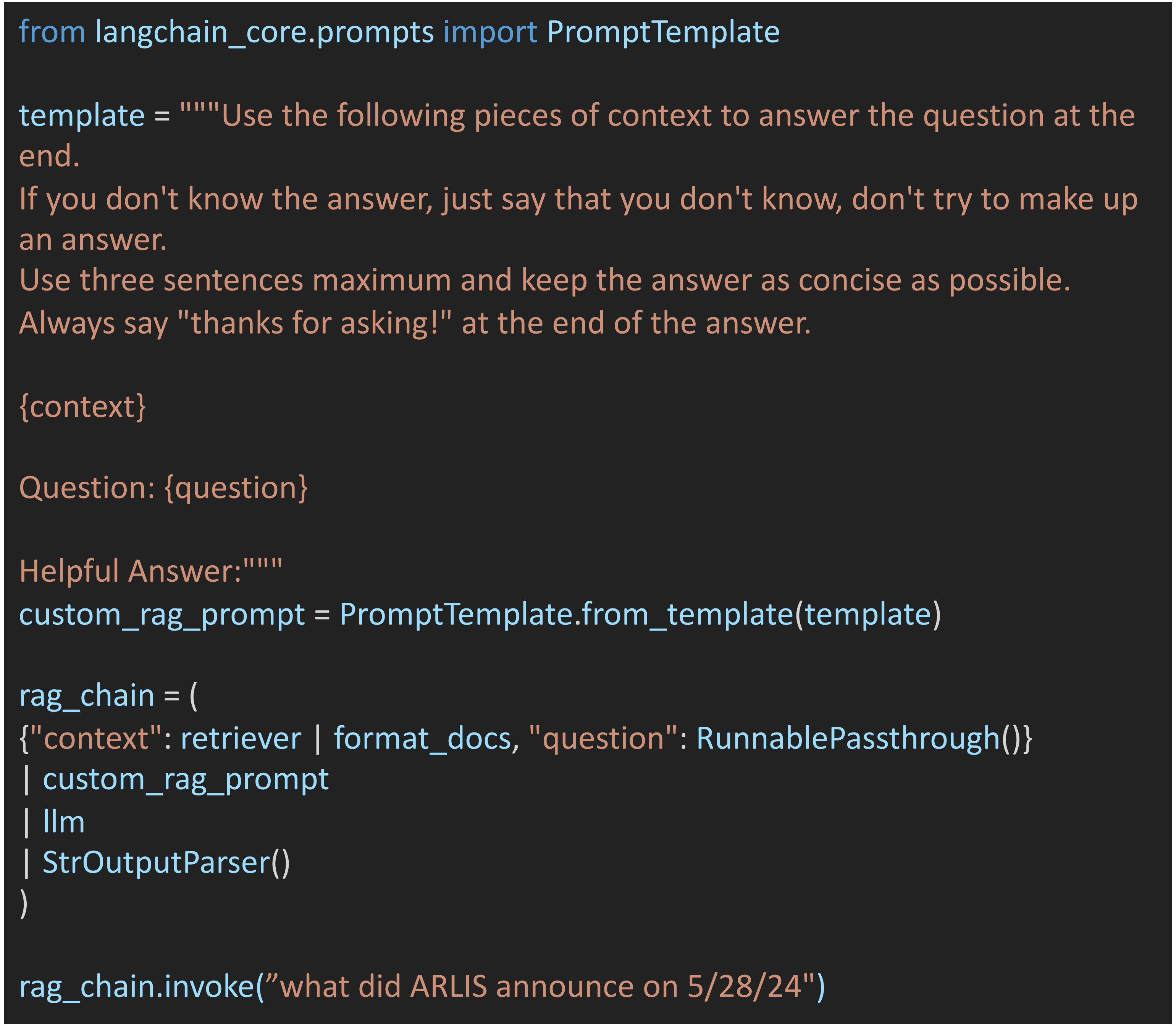
- Vector Databases
  - Highly efficient for similarity searching on our embedded text
  - Many existing implementations
    - Chroma, Pinecone, FAISS, Lance, etc.
- Vector search
  - With a good embedding, entries with mathematically small differences in the embedding vector are semantically related
  - E.g., cosine similarity: $\cos(\theta) = \dfrac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$
  - Take top N similar chunks and use them to augment our LLM prompt

Putting it all together



langchain.com

What does a RAG prompt actually look like?



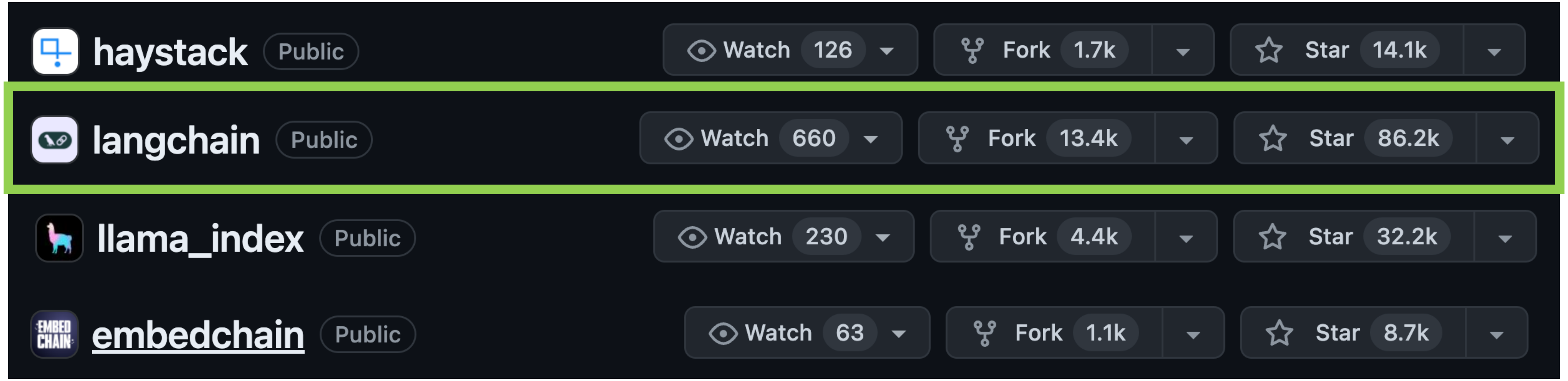
```
from langchain_core.prompts import PromptTemplate

template = """Use the following pieces of context to answer the question at the end.
If you don't know the answer, just say that you don't know, don't try to make up an answer.
Use three sentences maximum and keep the answer as concise as possible.
Always say "thanks for asking!" at the end of the answer.

{context}

Question: {question}

Helpful Answer:"""
custom_rag_prompt = PromptTemplate.from_template(template)

rag_chain = (
{"context": retriever | format_docs, "question": RunnablePassthrough()}
| custom_rag_prompt
| llm
| StrOutputParser()
)

rag_chain.invoke(”what did ARLIS announce on 5/28/24")
```

RAG Software Packages

- Many existing software libraries that helps automate all of these steps of chunking, embedding, vector databases, and prompting an LLM:
  - Haystack
  - OpenAI Assistants
  - LlamaIndex
  - Langchain
  - Embedchain

For today we will explore Langchain

- Currently the most popular framework of those I listed:



haystack	Public	Watch 126	Fork 1.7k	Star 14.1k
langchain	Public	Watch 660	Fork 13.4k	Star 86.2k
llama_index	Public	Watch 230	Fork 4.4k	Star 32.2k
embedchain	Public	Watch 63	Fork 1.1k	Star 8.7k

Why a Tool Like Langchain?

- Directly implements document ingestion
  - Website, text file, PDF, etc.
- Configurable backend LLM (OpenAI, others)
- Integrated with vector database
- Highly configurable and modular

Conclusion - What is the future of RAG?

- As models improve (know more and have increased context), it may be less important
- Concept of integrating information is key
  - Emergence of intelligent agents
    - Software entities that perform tasks autonomously, making decisions and acting on behalf of users.
  - Capabilities:
    - Gather and process information from diverse sources.
    - Analyze data to provide insights and recommendations.
    - Perform complex tasks that require contextual understanding.
    - Adapt to improve performance over time.
  - ChatGPT versus GPT4
    - One is a multi-agent chatbot the other is the LLM