EDGE FOR SCHOLARS

OpenAI in PHP

How AI works in theory & how to use it in PHP

Retrieval-Augmented Generation (RAG)

Weekly Topics

1. How Generative AI & LLMs Work
2. Prompt Engineering
3. **Retrieval-Augmented Generation (RAG)**
4. Fine-Tuning a Model
5. Agents
6. Using VUMC's OpenAI API in PHP

Key Concepts

- What is RAG?
  - Two Pipelines
- Retrieval Pipeline
  - Vectors & Embeddings
- Augmentation Pipeline
  - Cosine Similarity
- RAG Use Cases
- Advantages

What is RAG?

- Best way to supply an LLM with custom facts to add to a system

- Retrieval: Retrieves a fact from a document
- Augmented: Changes a prompt with new fact
- Generation: Generates a new response

RAG Example

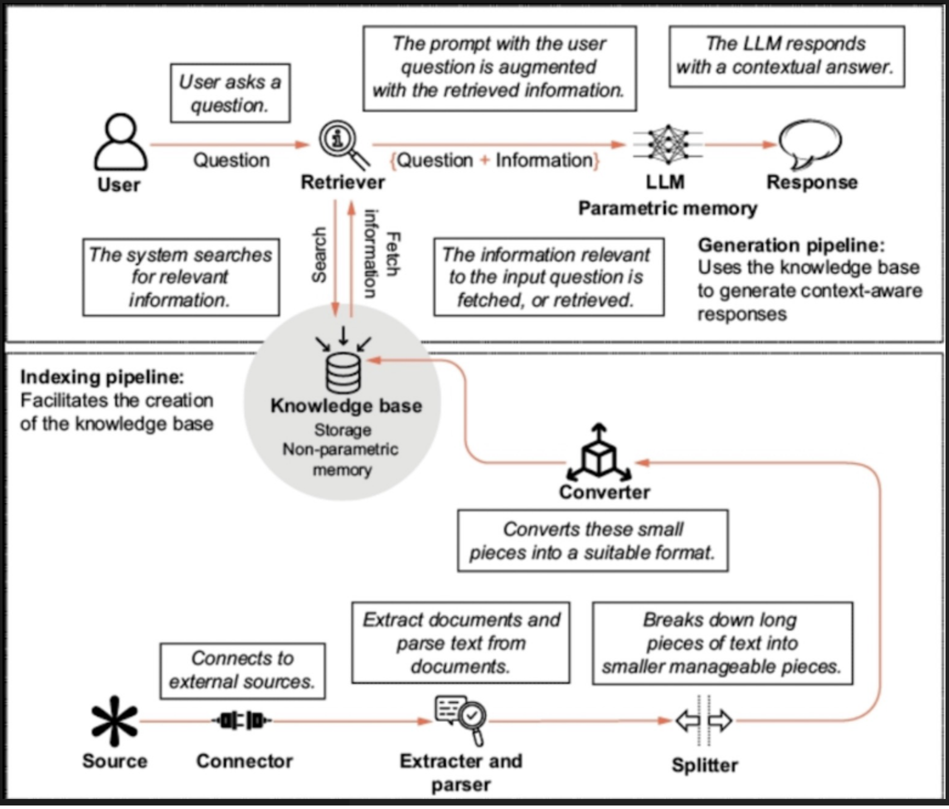
- Who won the Vanderbilt-Tennessee football game in 2025?
- Problem: LLM was compiled before the game was played

- Can use Wikipedia as a RAG data source to look up
- Scan documents to find answer
- Report answer

RAG Pipelines

- Generation / Augmentation Pipeline
- Indexing / Retrieval Pipeline

- Combine to give contextual answer



User asks a question.

The prompt with the user question is augmented with the retrieved information.

The LLM responds with a contextual answer.

The system searches for relevant information.

The information relevant to the input question is fetched, or retrieved.

Generation pipeline: Uses the knowledge base to generate context-aware responses

Indexing pipeline: Facilitates the creation of the knowledge base

Converts these small pieces into a suitable format.

Connects to external sources.

Extract documents and parse text from documents.

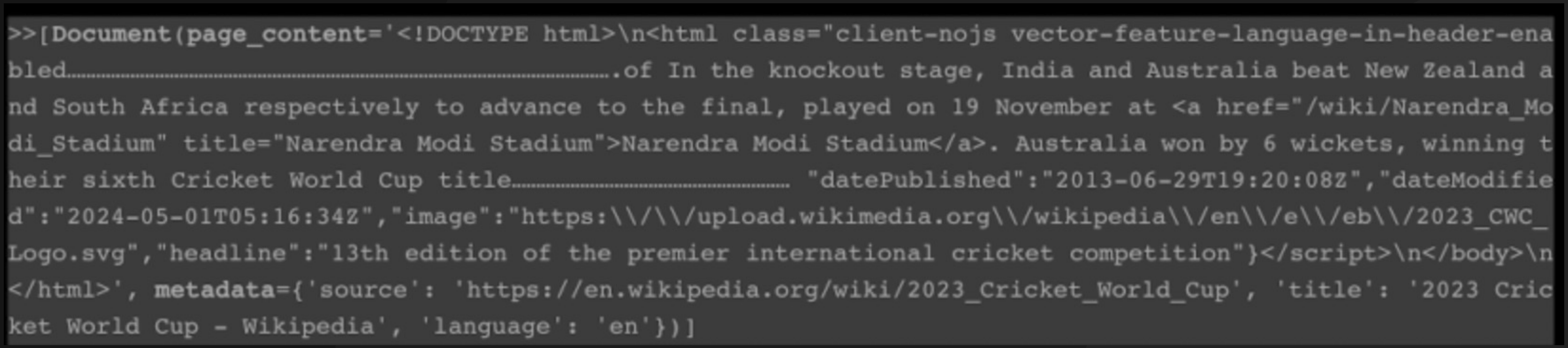
Breaks down long pieces of text into smaller manageable pieces.

Retrieval/Indexing Pipeline

- Get data as text
- Split text into chunks
- Turn chunks into vectors
- Store vectors in optimized data store with metadata

Getting the Text

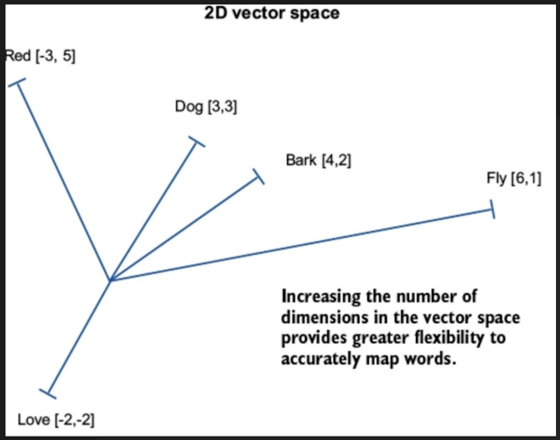
- Simplified text content + metadata
- Turn into overlapping chunks if too long



```
>>[Document(page_content='<!DOCTYPE html>\n<html class="client-nojs vector-feature-language-in-header-ena
bled.........................................................of In the knockout stage, India and Australia beat New Zealand a
nd South Africa respectively to advance to the final, played on 19 November at <a href="/wiki/Narendra_Mo
di_Stadium" title="Narendra Modi Stadium">Narendra Modi Stadium</a>. Australia won by 6 wickets, winning t
heir sixth Cricket World Cup title.................................. "datePublished":"2013-06-29T19:20:08Z","dateModifie
d":"2024-05-01T05:16:34Z","image":"https:\\/\\/upload.wikimedia.org\\/wikipedia\\/en\\/e\\/eb\\/2023_CWC_
Logo.svg","headline":"13th edition of the premier international cricket competition"}</script>\n</body>\n
</html>', metadata={'source': 'https://en.wikipedia.org/wiki/2023_Cricket_World_Cup', 'title': '2023 Cric
ket World Cup - Wikipedia', 'language': 'en'})]
```

Vectors & Embeddings

- Convert tokens into numbers & vectors/embeddings
- Vectors fill space (linear algebra)
- Vectors are n-dimensional (often 1500+ dimensions)
- Vectors are quickly manipulated by GPUs (and customized processors)


2D vector space

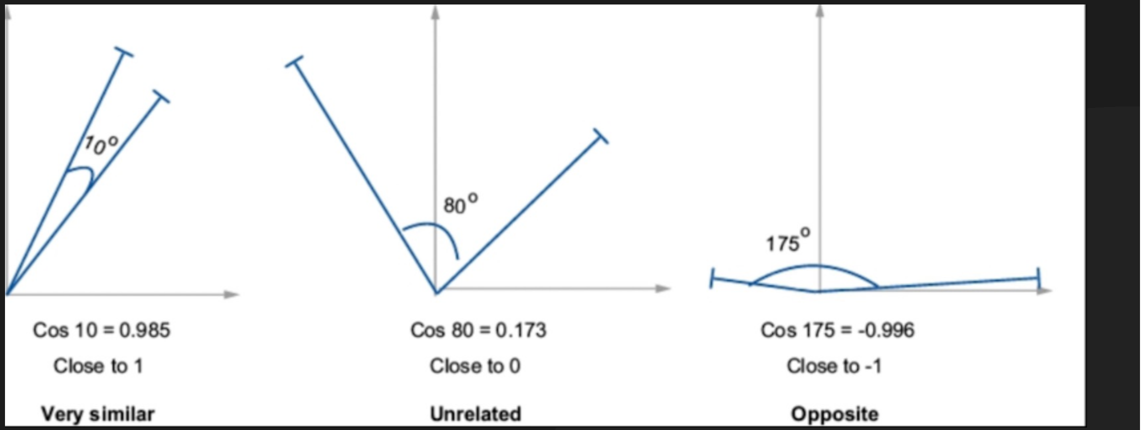
Increasing the number of dimensions in the vector space provides greater flexibility to accurately map words.

Augmentation/Generation Pipeline

- Prompt triggers RAG consultation
- Turn input string into vector
- Compare vector to data store’s vectors: Similarity
- Return top n results
- Combine results with prompt response

Cosine Similarity

- Vectors can be analyzed by various means to approximate similarity:
  - Cosine similarity
  - Distance apart
  - [Others, less common]



- Therefore, semantic similarity instead of exact similarity!

RAG Use Cases

- Supplies additional facts to a model
  - Fine-tuning provides tone and style

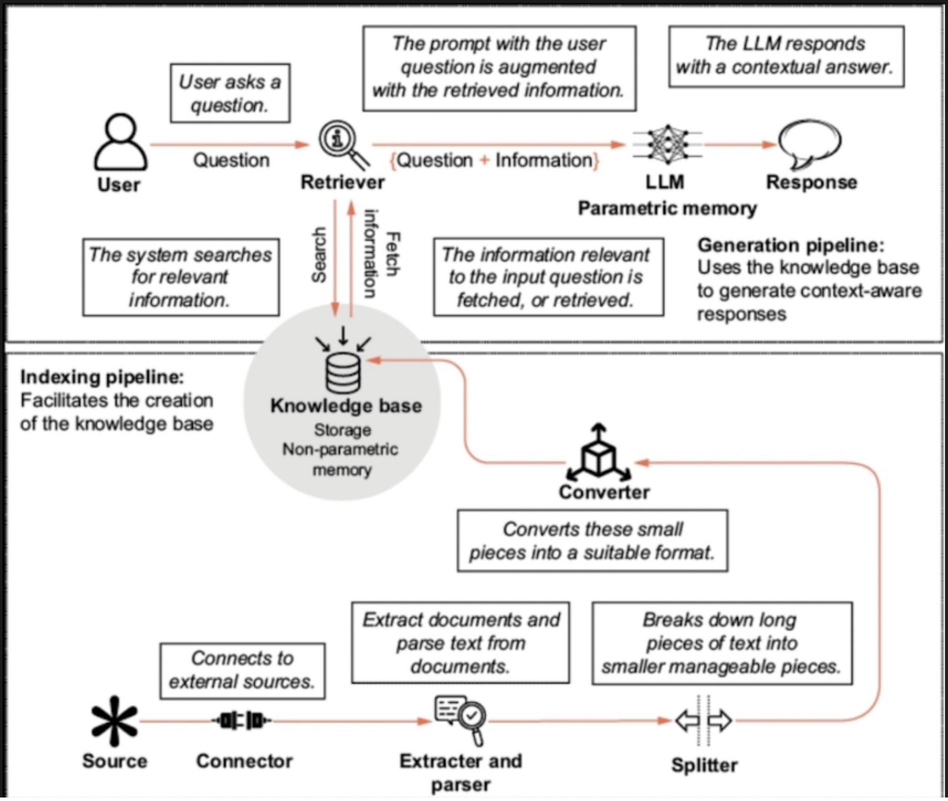
- Customize an organization’s philosophies & practices
- Provide up-to-date knowledge (e.g., medicine)
- Point to outside data sources (e.g., external website)

Advantages

- Directing to RAG can limit unruly hallucinations and confabulations by the foundational model
  - Reduce by 50%
- A quick way to contextualize a ChatBot
- A quick way to update information after last LLM build

Recap

- What is RAG?
  - Two Pipelines
- Retrieval Pipeline
  - Vectors & Embeddings
- Augmentation Pipeline
  - Cosine Similarity
- RAG Use Cases
- Advantages



User asks a question.

The prompt with the user question is augmented with the retrieved information.

The LLM responds with a contextual answer.

User — Question → Retriever — {Question + Information} → LLM — Response

Parametric memory

Search

Fetch information

The system searches for relevant information.

The information relevant to the input question is fetched, or retrieved.

Generation pipeline: Uses the knowledge base to generate context-aware responses

Indexing pipeline: Facilitates the creation of the knowledge base

Knowledge base
Storage
Non-parametric memory

Converter

Converts these small pieces into a suitable format.

Connects to external sources.

Extract documents and parse text from documents.

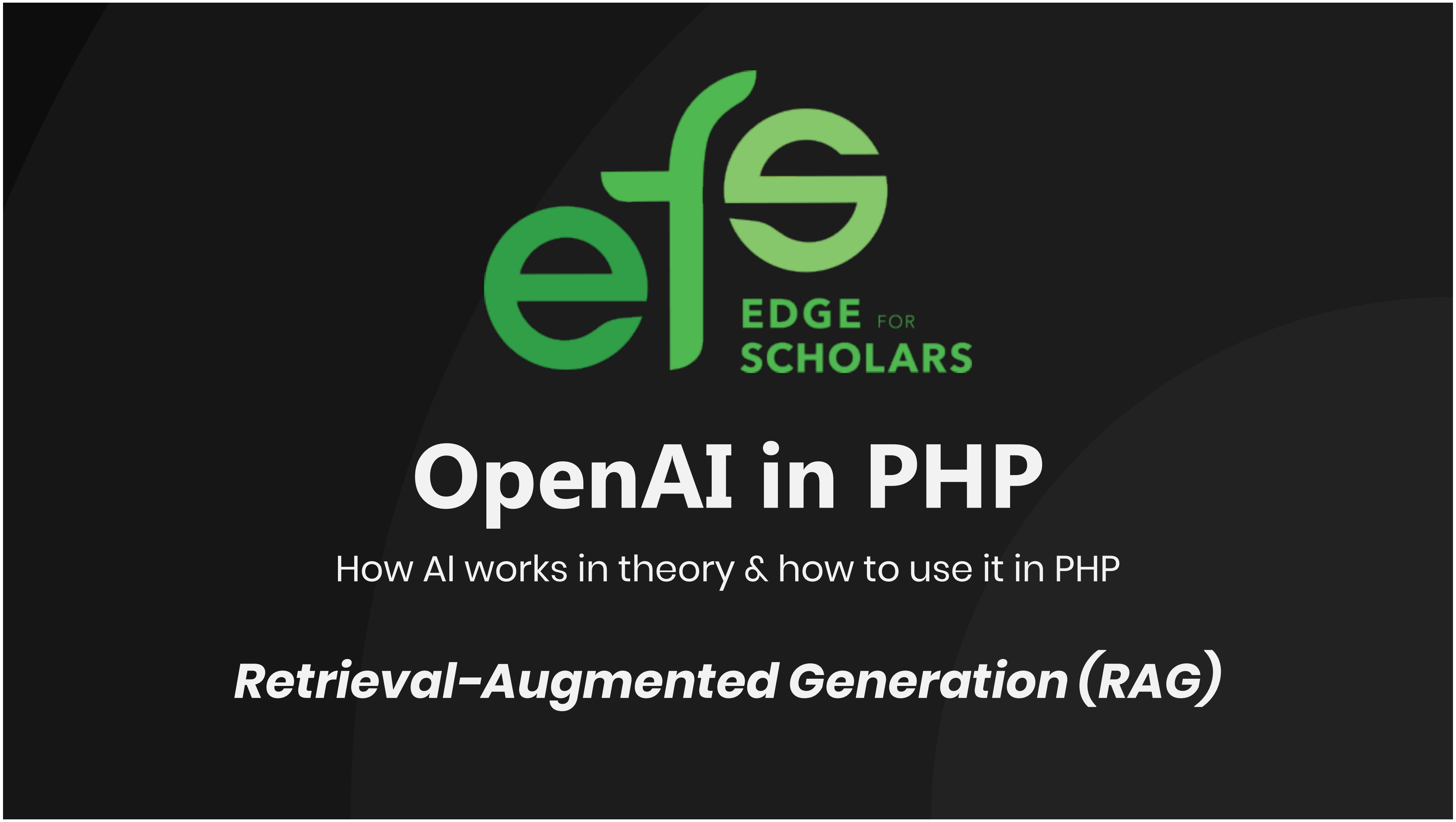
Breaks down long pieces of text into smaller manageable pieces.

Source — Connector — Extracter and parser — Splitter

Weekly Topics

1. ~~How Generative AI & LLMs Work~~
2. ~~Prompt Engineering~~
3. ~~Retrieval-Augmented Generation (RAG)~~
4. Fine-Tuning a Model
5. Agents
6. Using VUMC’s OpenAI API in PHP

Next Week: Fine-Tuning & Agents



EDGE FOR SCHOLARS

OpenAI in PHP

How AI works in theory & how to use it in PHP

Retrieval-Augmented Generation (RAG)