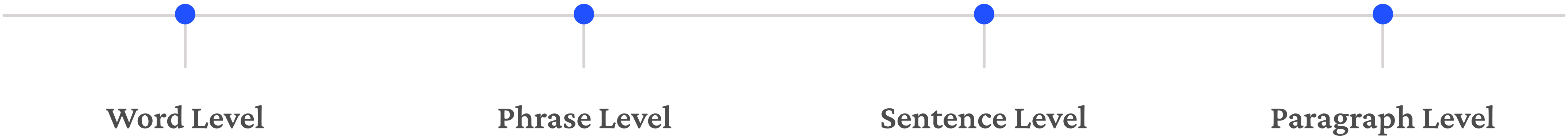
Applied NLP

Session 5

Lecturer: Narges Chinichian

Winter Semester 2025-2026

Our Journey Through Text Analysis



Word Level

Phrase Level

Sentence Level

Paragraph Level

Today, we ascend to the **whole text level**: analyzing complete documents and books with state-of-the-art NLP techniques.

Analyzing Complete Books

No 5-measure style today!

We focus on **one technique** (to rule them all! ;))

RAG (Retrieval Augmented Generation)

RAG combines search and generation for intelligent text analysis.

This approach enables sophisticated questions on entire books, providing accurate, contextually grounded answers with precise source references.

The ChatGPT Problem

You've probably all consulted ChatGPT at some stage of this course and faced:

Limited Knowledge Base

Cannot access your specific book; only knows what's in its training data or explicitly provided.

No Source Attribution

Cannot cite passages or page numbers; no access to original text.

Hallucination Risk

May invent details or produce incorrect statements without your actual text.

RAG solves these issues by grounding responses in your actual documents with explicit source references.

The Power of RAG Systems

Your Documents

Supply specific texts, books, or datasets

LLM Intelligence

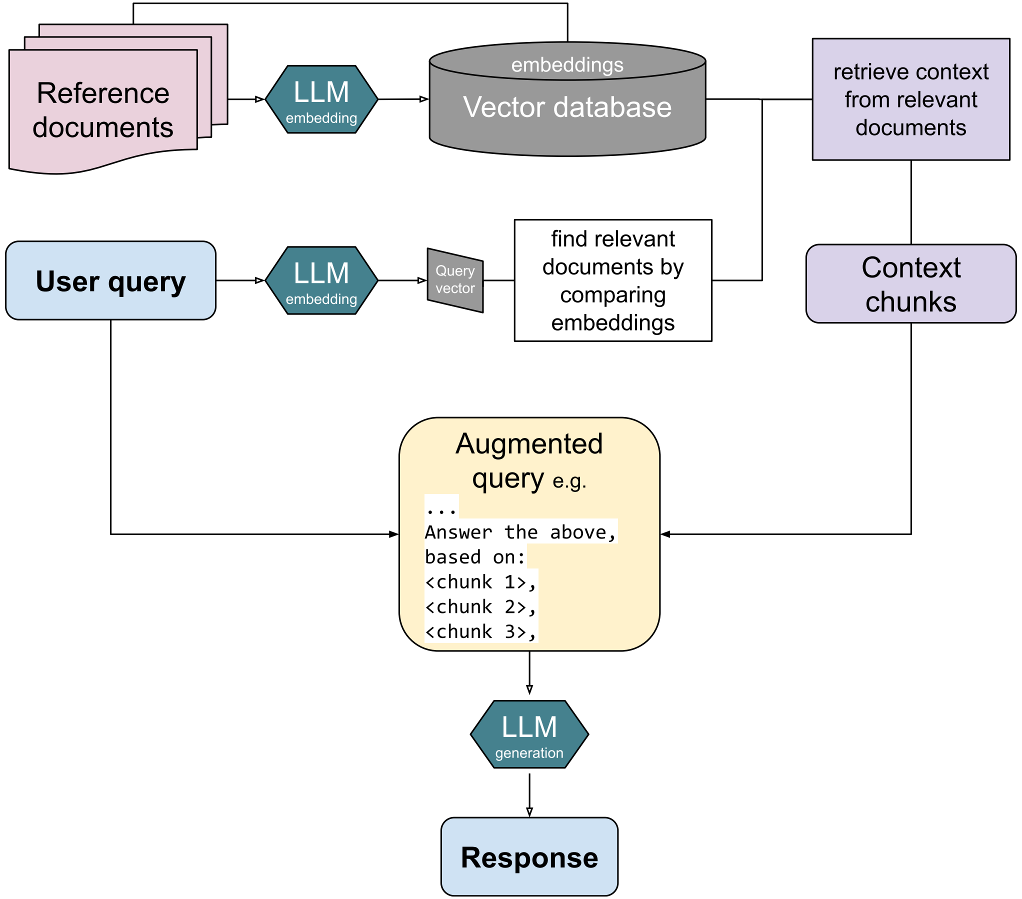
Leverage language understanding and generation

Grounded Answers

Get accurate, source-backed responses

RAG bridges the gap between the general knowledge of large language models and the specific information in your supplemental materials, creating a system that's both intelligent and factually grounded.

RAG System Architecture



```
...
Answer the above,
based on:
<chunk 1>,
<chunk 2>,
<chunk 3>,
```

Source: Wikipedia - Retrieval-augmented generation

Customizing Your RAG Pipeline

LLM Selection

Options: GPT-4, Claude, Llama 2, Mistral

Considerations: Cost, speed, context window size, and output quality

Embedding Models

Options: OpenAI embeddings, Sentence-BERT, Cohere

Considerations: Semantic accuracy, dimension size, and computational efficiency

Chunking Strategy

Options: Fixed-size (512, 1024 tokens), semantic boundaries, paragraph-based

Considerations: Overlap size, coherence preservation, retrieval precision

Vector Database

Options: Pinecone, Weaviate, ChromaDB, FAISS

Considerations: Scale, query speed, persistence needs, and integration ease

Why RAG for Your Project Now?

Verify Previous Analyses

Cross-check findings from word, phrase, and sentence-level analyses against the full text context

Explore Narrative Structures

Engage in dynamic dialogue with your book to uncover themes, character arcs, and plot developments

Extract Presentation Content

Identify compelling passages, quotes, and examples to strengthen your final presentation

This is your opportunity to **synthesize insights** across all analysis levels and build a coherent narrative about your book.

Real-World Applications of RAG

Enterprise Knowledge Management

Analyze company-specific documents, policies, contracts, and internal wikis to answer employee questions instantly

Medical Research

Query vast collections of clinical studies and patient records to support evidence-based treatment decisions

Legal Document Analysis

Search through case law, precedents, and regulatory documents with precise citation and context extraction

Customer Support

Provide accurate, up-to-date information from product manuals, FAQs, and troubleshooting guides

RAG systems deliver **precise, current information** grounded in authoritative sources — a critical advantage over general-purpose LLMs.

Today's Hands-On Session

1 Setup Environment

2 Ingest Your Book

3 Build RAG Pipeline

4 Query and Analyze