Course Overview

Chunking is a critical design decision when building retrieval-based AI systems and Retrieval-Augmented Generation (RAG) pipelines. The way documents are divided into chunks directly affects how information is indexed, retrieved, and ultimately understood by large language models. Poor chunking can lead to irrelevant retrieval, missing context, and unreliable responses—even when the rest of the system is well designed.

This course explores chunking not just as a preprocessing step, but as an important component of system architecture. You will learn how different chunking strategies work, how they influence retrieval quality, and how to select the right approach based on data type, model constraints, and application goals.

By the end of the course, you will understand how to design, implement, and evaluate chunking strategies for real-world AI applications that rely on document retrieval.

Course Objectives

- Understand why chunking is a foundational component of retrieval-based AI systems
- Identify factors that influence chunking decisions across different datasets and tasks
- Apply multiple chunking strategies suited for different types of content
- Recognize and troubleshoot common issues caused by poor chunking design
- Compare different chunking approaches and understand their trade-offs
- Design hybrid chunking pipelines for complex document structures
- Evaluate the effectiveness of chunking strategies using practical metrics

---

What You Will Learn

- How chunking affects retrieval accuracy in AI and RAG systems
- Heuristics for selecting chunk size and overlap
- Fixed-size, sentence-based, and sliding window chunking methods

- Semantic chunking using meaning-based segmentation
- Structure-aware chunking for formats like HTML, Markdown, PDFs, tables, and code
- Differences between semantic and structure-based approaches
- Techniques for combining multiple chunking strategies
- Methods for evaluating chunk quality across different use cases

---

Prerequisites

- Basic Python programming
- Familiarity with Natural Language Processing concepts
- Basic understanding of embeddings and vector databases
- Introductory knowledge of LLM-based applications

Who This Course Is For

- AI engineers building retrieval or RAG-based systems
- ML engineers working with document search pipelines
- Backend developers integrating LLMs into applications
- Data scientists working with large document collections
- Developers interested in improving AI-driven search and knowledge retrieval

---

Course Syllabus

1. Introduction: Chunking as a Design Decision

2. Why Chunking Exists
3. Factors Influencing Chunking Choices (data type, task, retrieval method, model constraints)
4. Chunk Size and Overlap Heuristics
5. Common Failure Modes of Poor Chunking
6. Overview of Chunking Strategies
7. Fixed-Size Chunking
8. Sentence-Based Chunking
9. Sliding Window Chunking
10. Semantic Chunking
11. Structure-Aware Chunking in Practice (HTML, Markdown, PDFs, tables, code)
12. Comparing Semantic vs Structure-Aware Chunking
13. Hybrid Chunking Approaches
14. Evaluation of Chunking Strategies
15. Metrics and Techniques for Evaluating Chunk Quality Across Use Cases

---

Skills You Will Gain

- Designing effective chunking strategies for AI pipelines
- Implementing multiple chunking approaches for different data formats
- Optimizing chunk size and overlap for retrieval performance
- Building structure-aware preprocessing pipelines
- Diagnosing retrieval issues caused by poor chunking
- Evaluating chunking strategies using systematic metrics

- Designing scalable preprocessing pipelines for RAG systems