# Introduction

Semantic chunking is crucial for retrieval-augmented generation (RAG) systems. The process involves splitting documents into meaningful units, allowing large language models to operate efficiently and retrieve relevant knowledge. The effectiveness of chunking directly impacts retrieval accuracy.

## Motivation

Chunking enables granular access to content. This supports improved search, summarization, and question answering. It also supports scalable document processing.

# Related Work

Prior work in chunking has explored both syntactic and semantic methods. Syntactic approaches use structure such as headings or paragraph breaks.

Semantic chunking attempts to preserve topic coherence, even across multiple paragraphs.

### Syntactic Chunking

Methods relying on formatting cues, such as Markdown or HTML headings, often split text by section. While easy to automate, this can fragment related ideas or isolate small sections.

### Semantic Chunking

Semantic chunking uses content analysis—often via embeddings or topic models—to find natural boundaries in meaning. This often results in larger, more meaningful text blocks.

# Implementation Guidelines

The following rules ensure coherent and useful chunks for downstream applications:

- Each chunk should ideally be around 100 words.
- Chunks must not be shorter than 10 lines or 100 words, except at the end or in cases of necessity.
- Headings are grouped with their content.
- Very short sections may be merged to preserve coherence.

# Example Cases

## Case 1: Long Section

This section is deliberately long. It demonstrates how a chunker should group several paragraphs under one heading if they address the same topic. The chunk should not be split mid-topic, even if it exceeds the ideal word count.

Additional supporting details are included here to ensure the chunk is long enough for the test. The boundary should be set at a clear topic change, not an arbitrary line or word limit. This paragraph continues the topic, making it part of the same chunk. If another subheading followed with just a sentence, the chunker should consider merging it with more content.

# Case 2: Two Short Headings

## Short A

A very brief point.

## Short B

Another short statement.

Both of these are too brief to stand alone, so they should be grouped with the next semantically related section.

# Case 3: Minimal Content at End