Understanding Text Chunking

Text chunking is the process of dividing a large body of text into smaller, manageable pieces or "chunks". This is especially useful when working with large documents such as PDFs, where processing the entire document at once may be inefficient or exceed token limits in language models.

Why Chunk Text?

1. To handle large documents in manageable parts.
2. To improve performance in language model queries.
3. To apply operations like summarization, search, or translation on individual chunks.

How to Chunk Text?

1. Extract text from the source (e.g., PDF).
2. Decide chunk size (e.g., 300-500 words or characters).
3. Optionally allow overlap between chunks to preserve context.
4. Store or process each chunk as needed.

Example:

Original Text (1000 words) -> Chunk 1 (0-300), Chunk 2 (250-550), Chunk 3 (500-800), ...

Overlap ensures continuity in content, which is important for tasks like summarization or Q&A.

Applications:

- Document summarization
- Semantic search

- Conversational agents