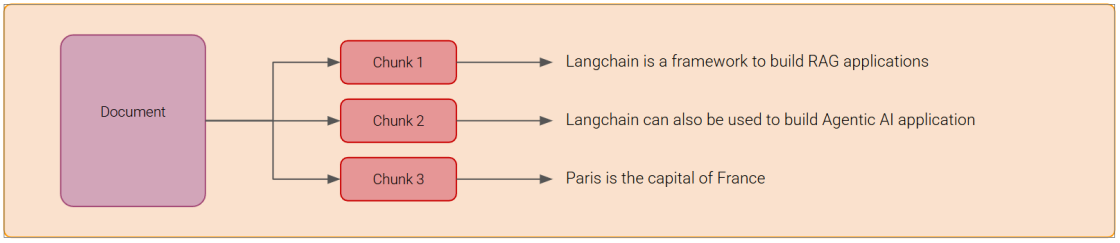
Advanced Chunking and Preprocessing

Semantic Chunking

Semantic chunking is the process of splitting a document into meaningful chunks based on semantic similarity - not just by number of tokens or lines.
This is important in RAG systems because

- Better chunks -> better retrieval -> better grounding -> better answers
- Chunks should be self-contained, contextually rich, and logically separated



Recursive Character Text Splitter

How does it work

1. **Document Segmentation**
2. **Perform Sentence Embeddings**
   Each sentence is converted into a vector representation
3. **Perform a Semantic Similarity Check**
   Cosine similarity between adjacent embedding
4. **Merging of Sentences**
   Merge adjacent sentences if they are semantically similar based on some threshold
5. **Form Chunks**

Example

1. LangChain is a framework for building LLM-powered applications.
2. It integrates with tools like OpenAI and Pinecone.
3. The Eiffel Tower is located in Paris.
4. France is a popular tourist destination.

After performing semantic similarity

1. LangChain is a framework for building LLM-powered applications. It integrates with tools like OpenAI and Pinecone.
2. The Eiffel Tower is located in Paris.
3. France is a popular tourist destination.