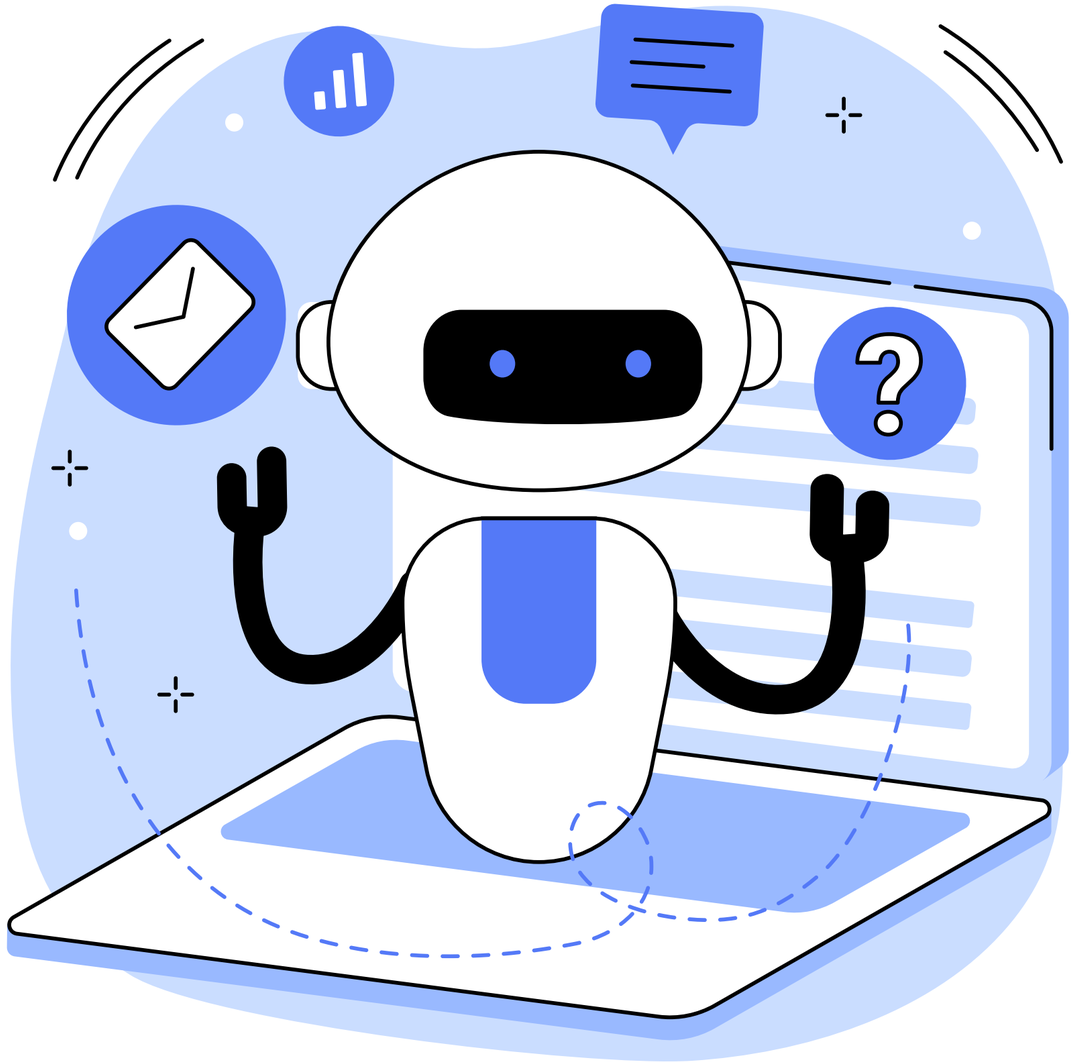
ARTIFICIAL INTELLIGENCE

Master Class

CONTENTS

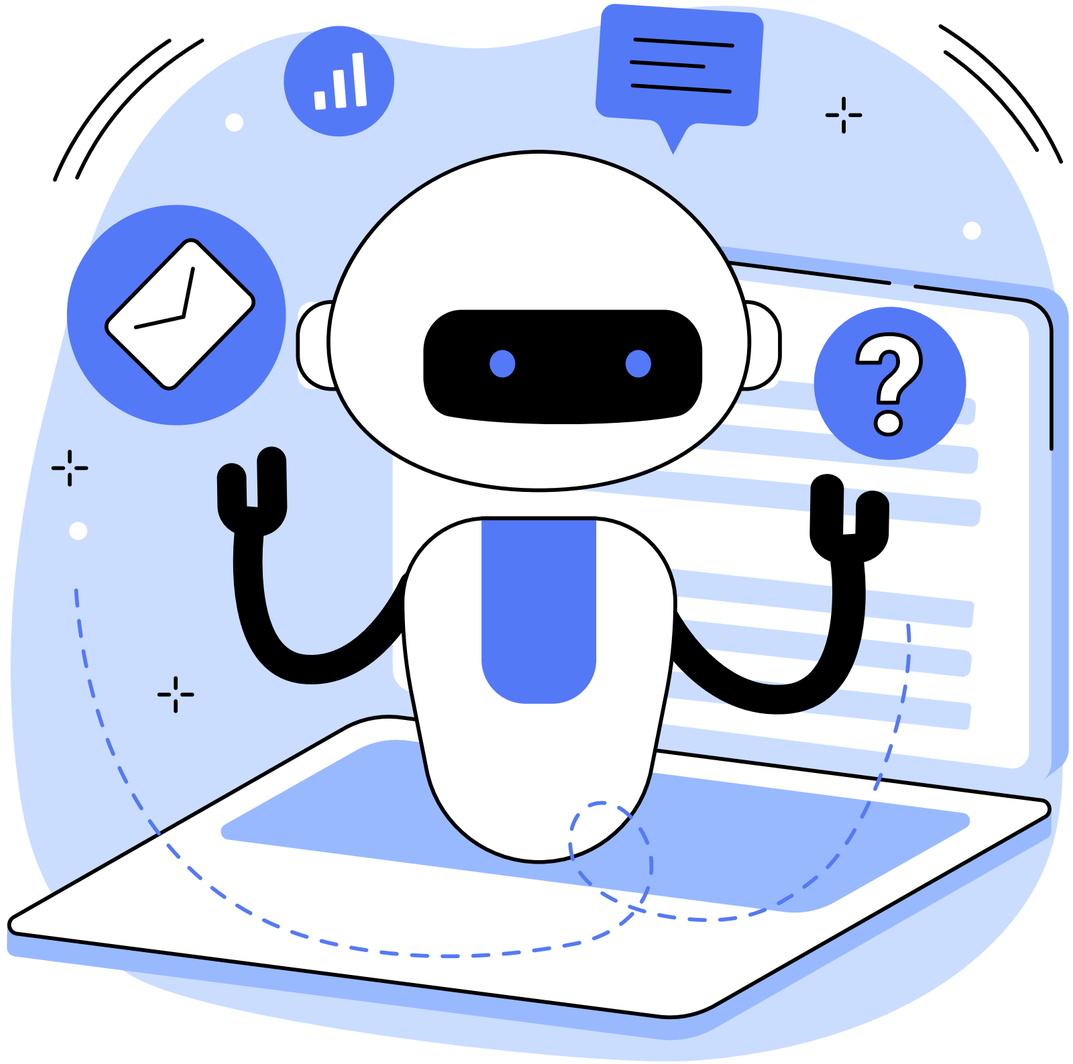
WHAT IS LLAMAINDEX?

LlamaIndex is an open-source library that connects language models to your data for search and retrieval. It streamlines document loading, chunking, embedding, indexing, and query pipelines.



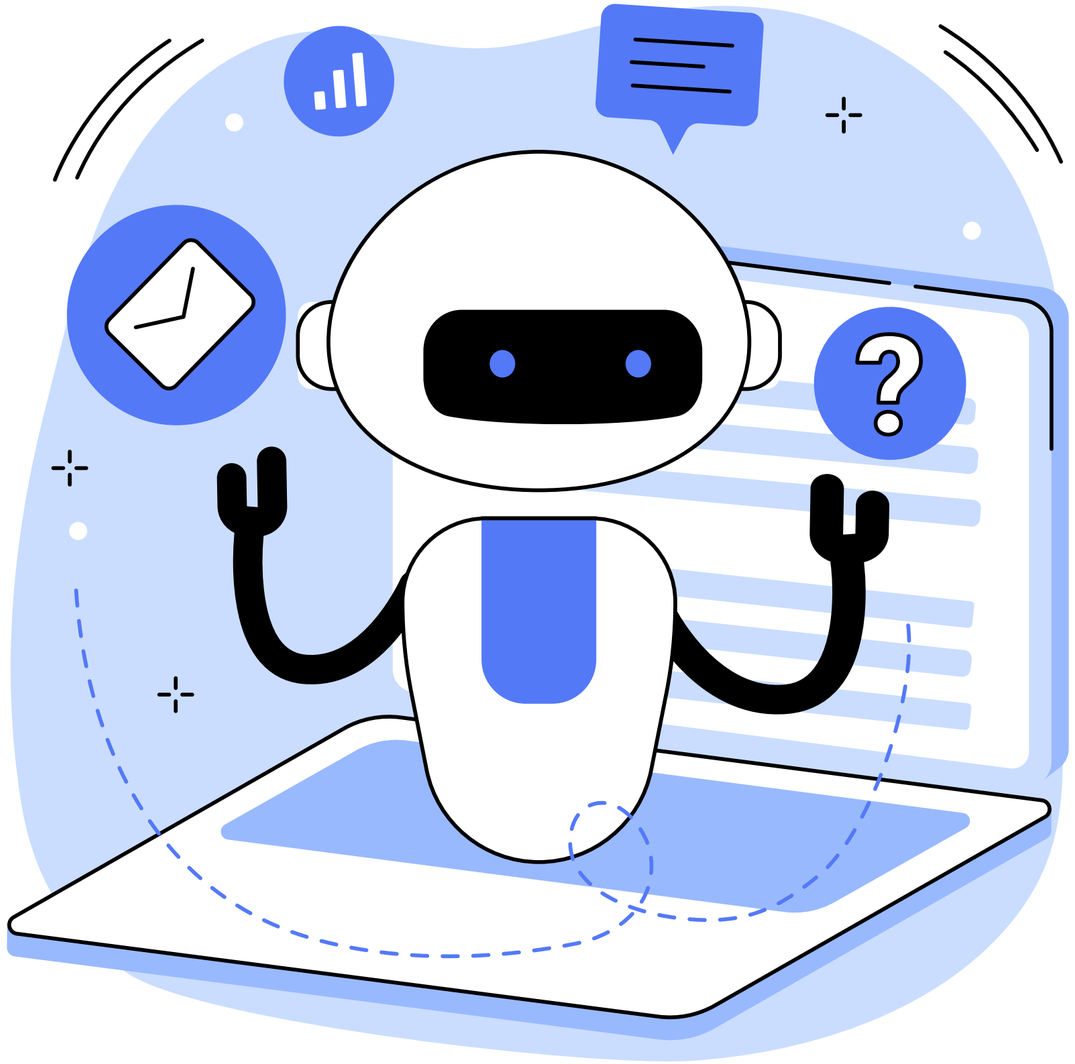
WHAT IS A PARSER?

A parser splits large documents into smaller, manageable pieces (nodes or chunks). This step is essential for embedding, retrieval, and efficient LLM processing.



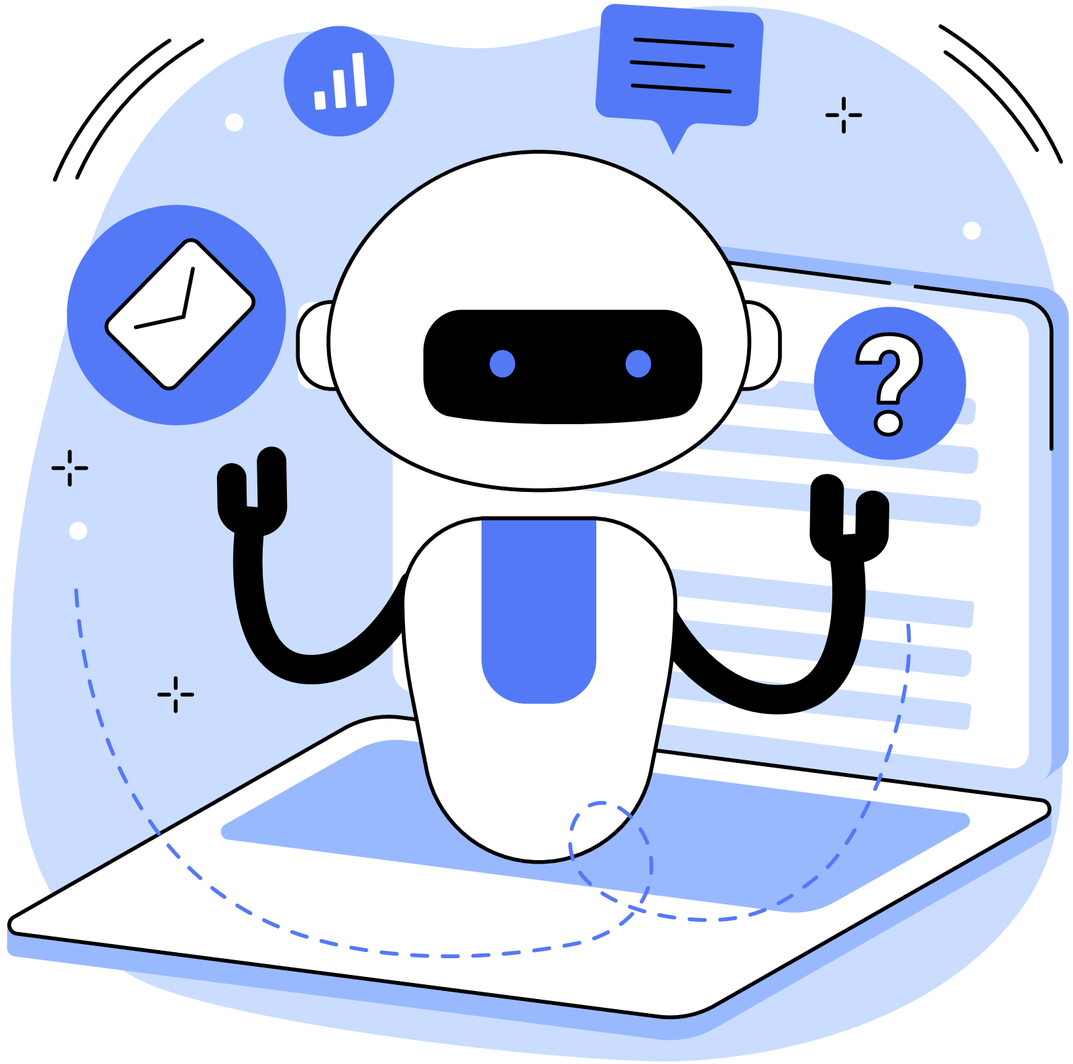
HIERARCHICALNODEPARSER

The HierarchicalNodeParser in LlamaIndex creates chunks at different levels (sections, paragraphs, sentences). It preserves document structure and allows searches at multiple granularities.



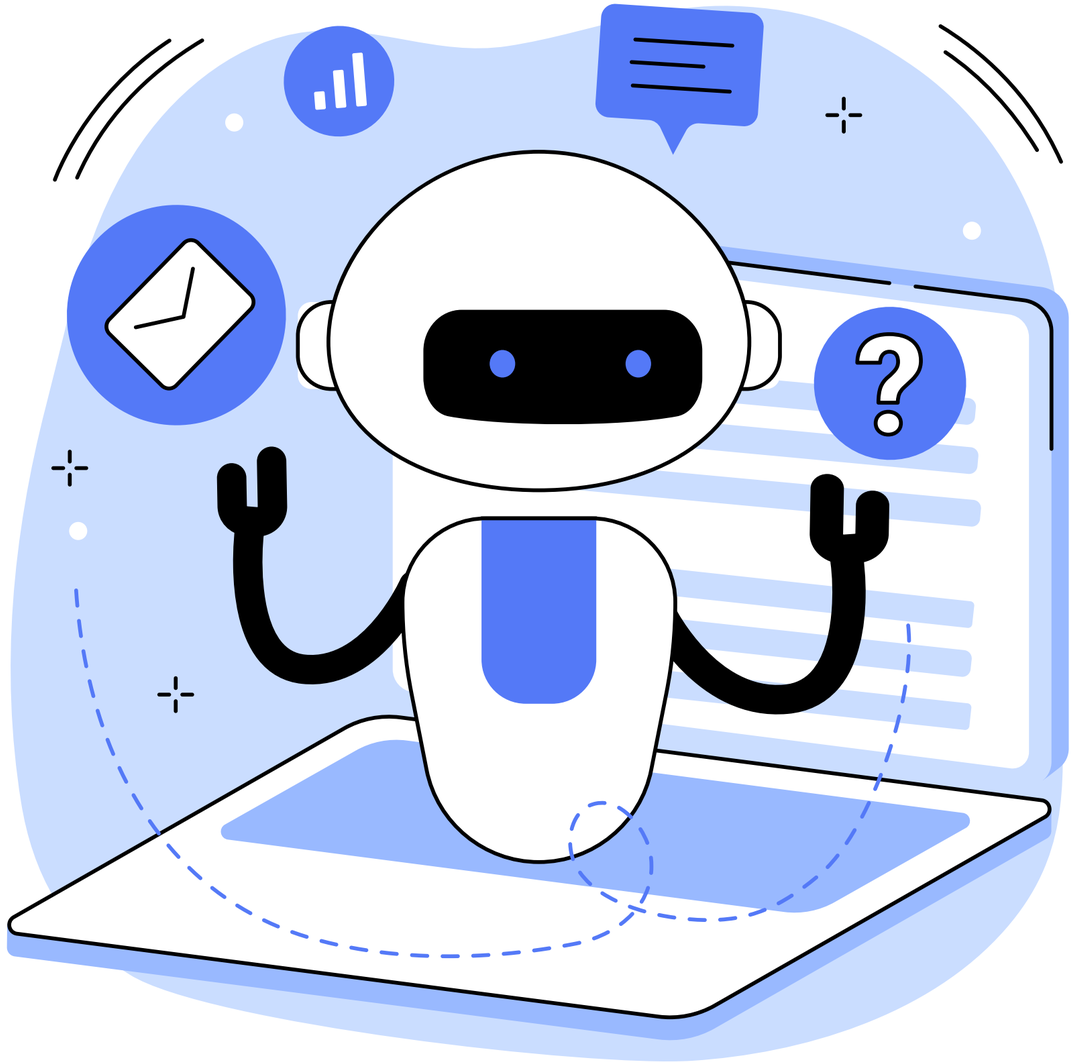
CHUNKING DOCUMENTS

Documents are split into manageable chunks for embedding. Each chunk contains a portion of text, making it easier for models to process and retrieve relevant information.



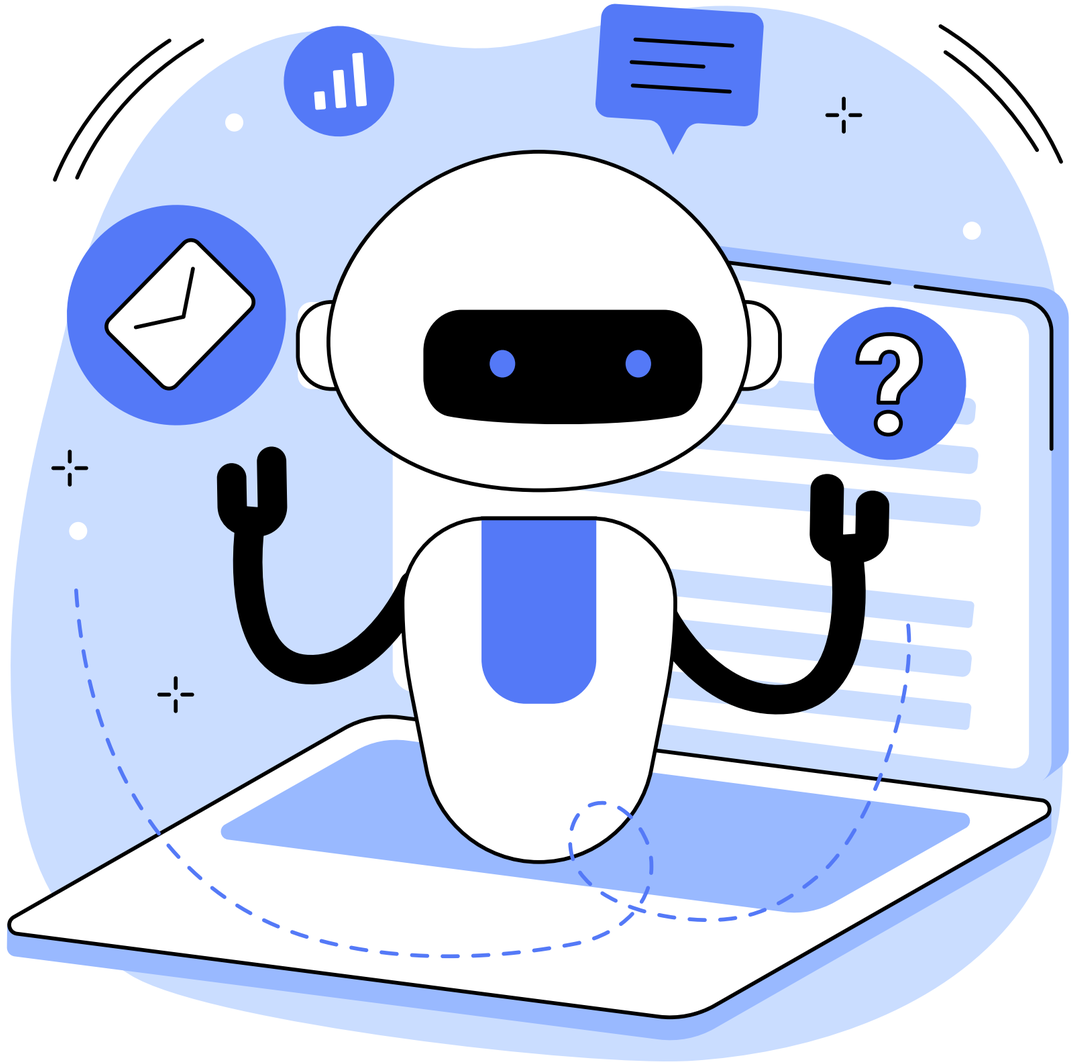
WHY OVERLAP CHUNKS?

Overlapping chunks include shared tokens or sentences between adjacent pieces, ensuring that important context isn't lost at the edges during retrieval or generation.



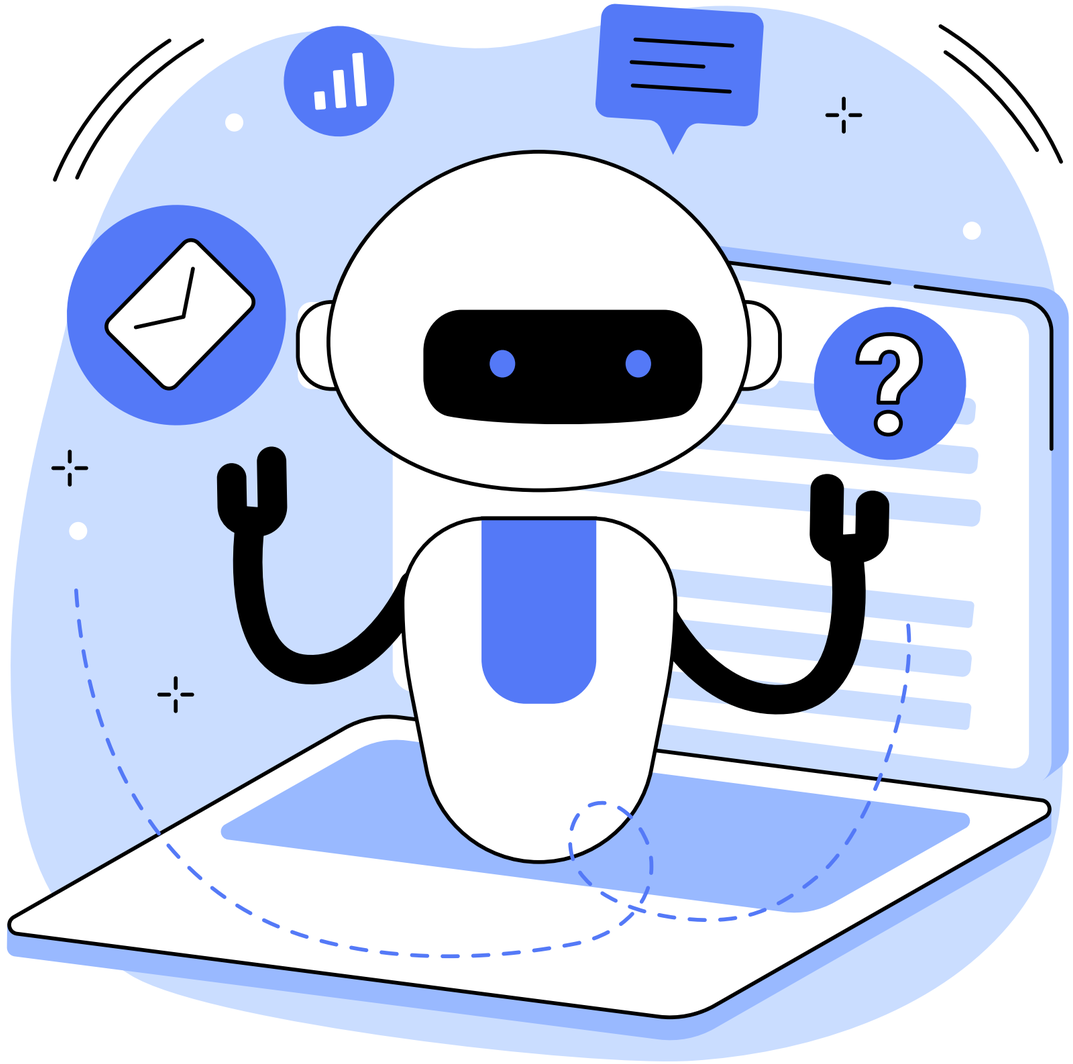
RETRIEVAL-AUGMENTED GENERATION (RAG)

RAG pipelines retrieve relevant document chunks using vector search and pass them as context to language models, producing grounded, fact-rich answers.



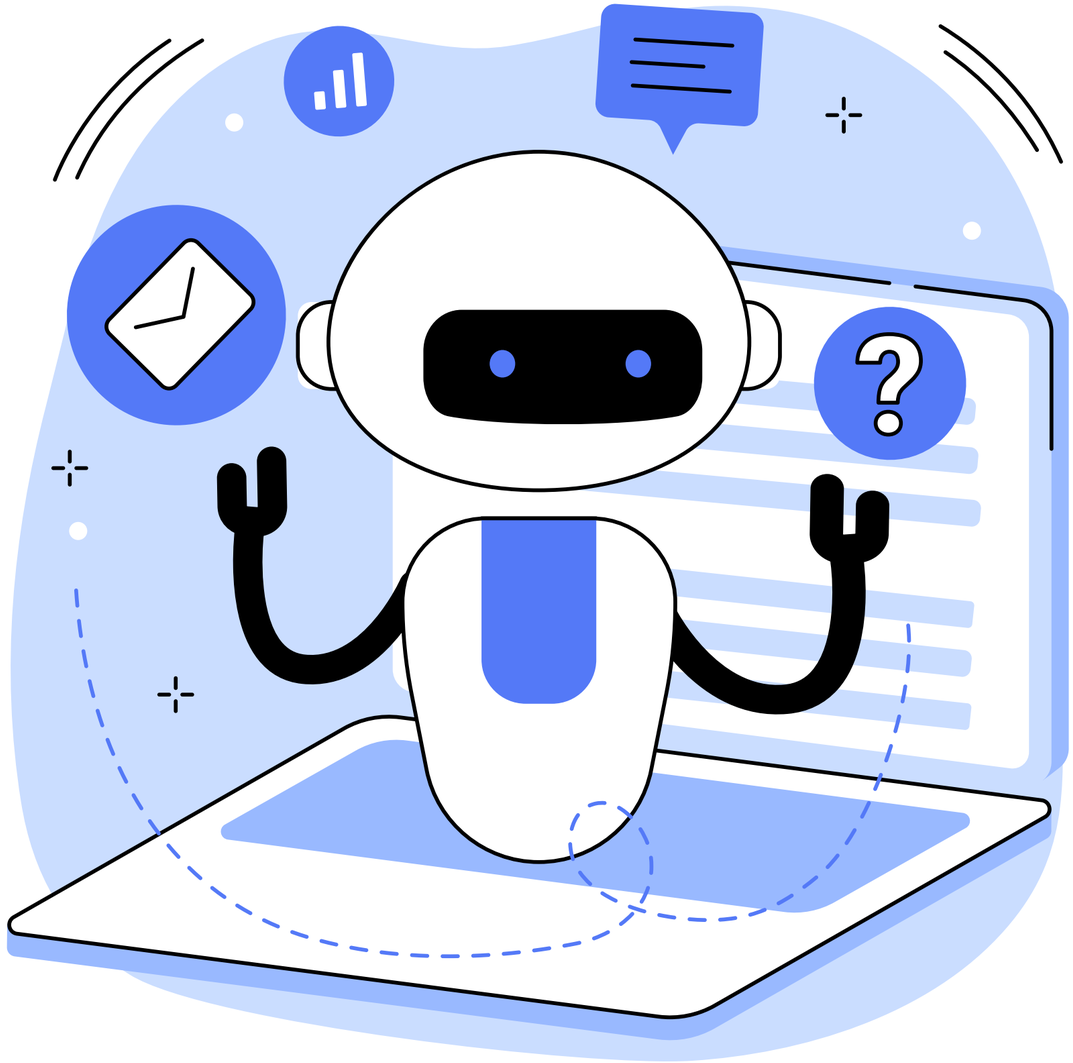
QUERIES YOU CAN ASK

You can ask fact-based, summary, and explanatory questions. The system finds the most relevant chunks and uses them for answers or generation.



THE ROLE OF CHUNK OVERLAP SIZE

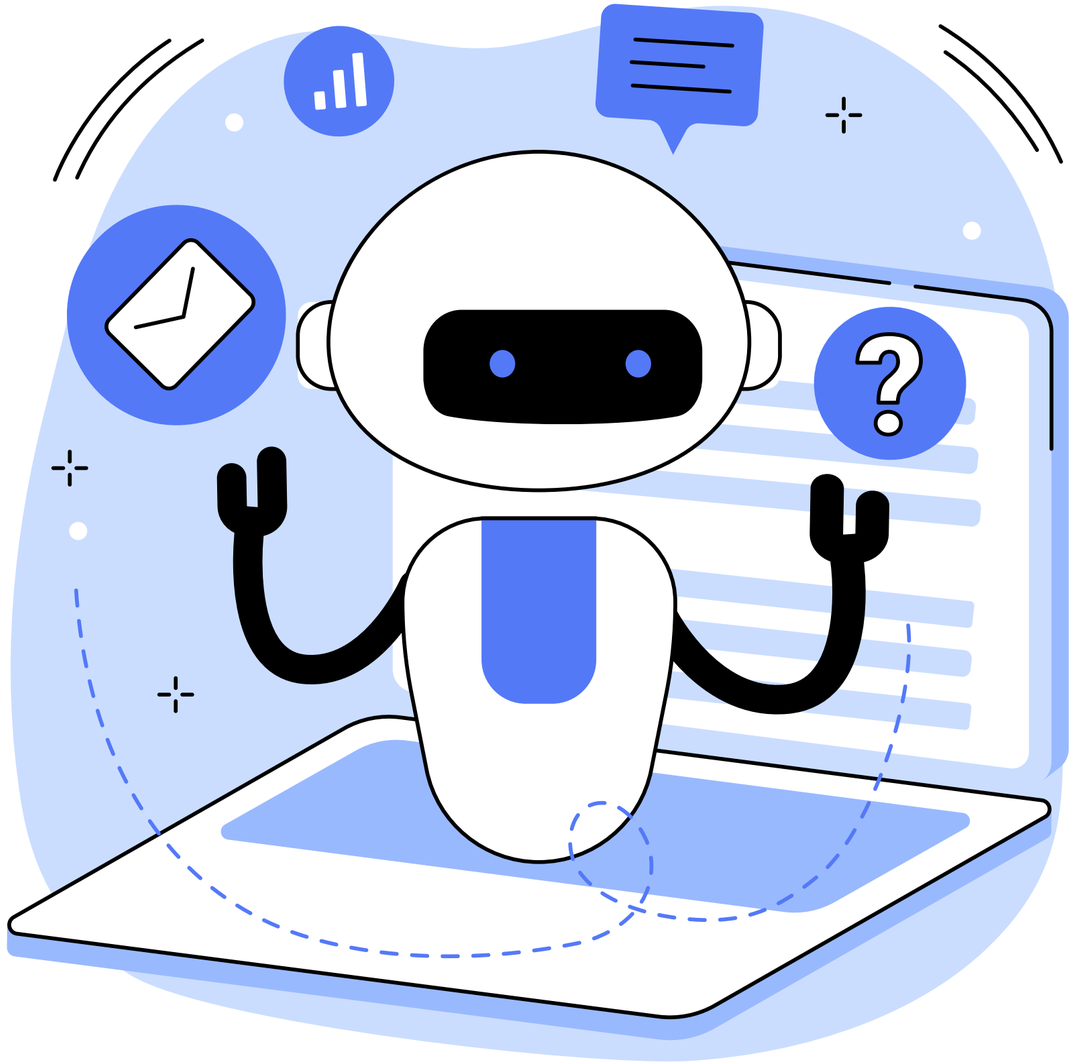
Typical chunk overlap is 10-40 tokens. More overlap means richer context, but also more data. Tune overlap for your use-case to maximize retrieval quality.



CONVERTING RESPONSE WITH LLM

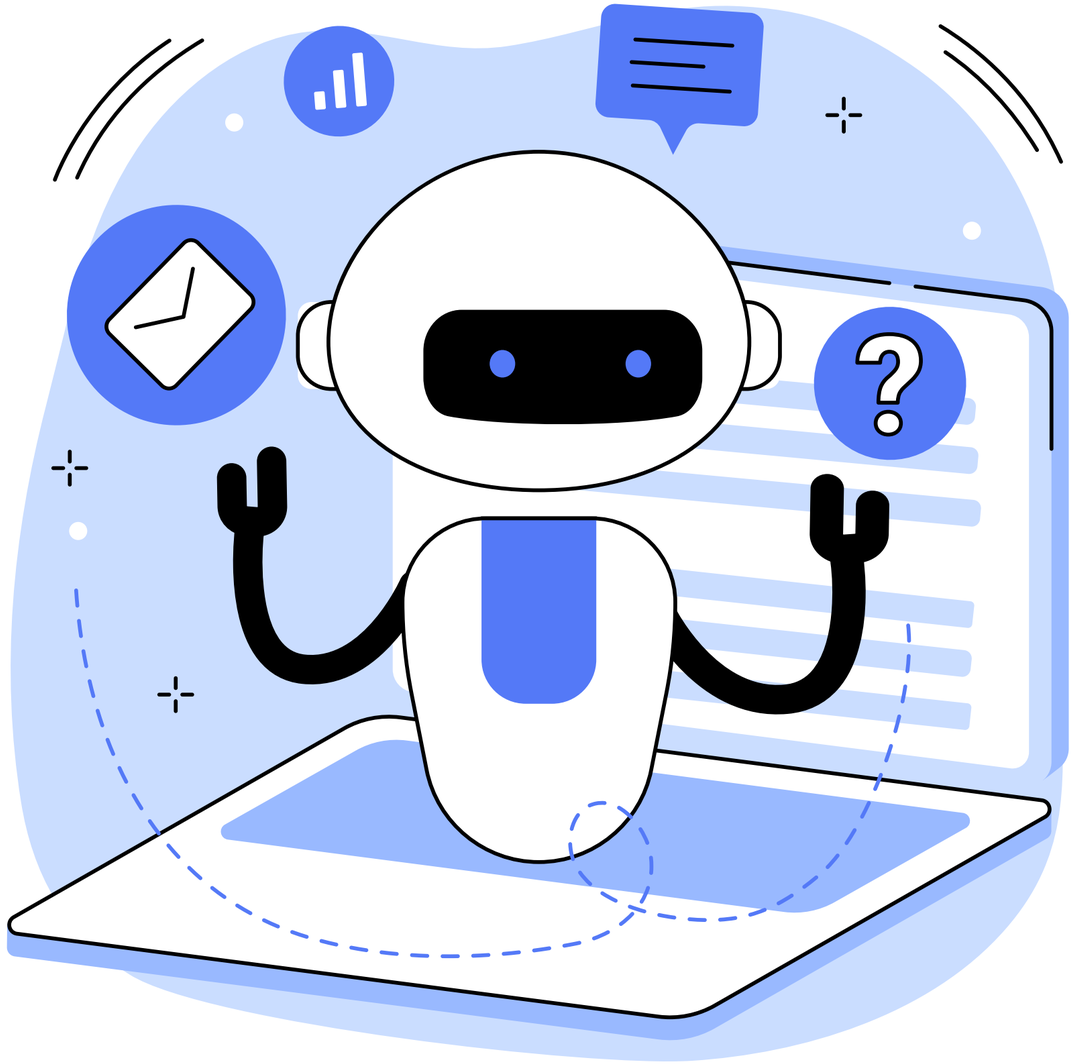
After retrieving the most relevant chunks, we use a Large Language Model (LLM) to convert raw results into fluent, user-friendly answers.

This final step bridges the gap between raw document data and high-quality answers.



PUTTING IT ALL TOGETHER

Combine chunking, embedding, overlap, and vector retrieval to build a scalable, context-aware document search and Q&A system.



THANK YOU FOR LISTENING!