5
Chunking
Techniques

🤜🤛

YOU

need to know!

Weaviate

Photosynthesis is one of nature's most vital processes.

chunk 1   chunk 2   chunk 3
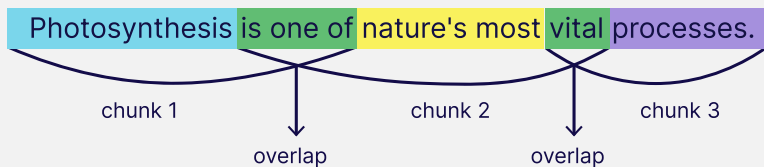
overlap   overlap

Fixed-Size Chunking

This technique **splits** the text into chunks of a **fixed size**, without considering natural breaks or the structure of the content. It's **simple** and **cost-effective**, but lacks **contextual awareness**. To improve this, **overlapping chunks** can be used, allowing adjacent chunks to share some content.

Quantum entanglement is a key concept in quantum physics. It occurs when particles become linked, so the state of one instantly affects the state of another, no matter the distance between them

This connection challenges our understanding of space and time. When you measure one entangled particle, the other's state changes instantly.

Recursive Chunking

Text is initially split using a **primary separator,** like paragraphs. If the resulting chunks are too large, **secondary separators**, like sentences, are applied recursively until the desired chunk size is achieved. This technique respects the **document's structure** and is **flexible** for various use cases.

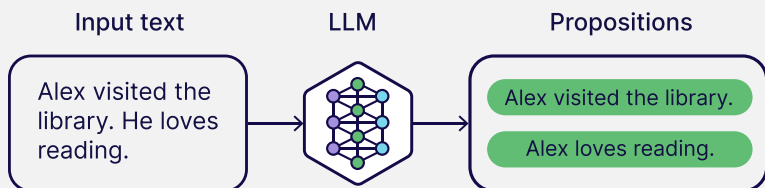# Heading This is a heading. ---## Subheading This is a subheading. We can continue with more content here.

Document–Based Chunking

This technique creates chunks based on the **natural divisions** within the document, such as headings or sections. It's very effective for structured data like **HTML**, **Markdown**, or **code files** but it's less useful when the data lacks clear structural elements.

The water cycle is a continuous process by which water moves through the Earth and atmosphere. It involves processes such as evaporation, condensation, precipitation, and collection. Evaporation occurs when the sun heats up water in rivers, lakes, or oceans, turning it into vapor or steam. This vapor rises into the air and cools down, forming clouds. Eventually, the clouds become heavy and water falls back to the earth as precipitation, which can be rain, snow, sleet, or hail. This water then collects in bodies of water, continuing the cycle.

Semantic Chunking

In this technique, the text is divided into **meaningful units**, such as sentences or paragraphs, which are then **vectorized**. These units are then combined into chunks based on the **cosine distance** between their embeddings, with a new chunk formed whenever a significant **context shift** is detected.

Input text	LLM	Propositions

Alex visited the library. He loves reading.

Alex visited the library.

Alex loves reading.

LLM–Based Chunking

This advanced technique uses a **Language Model (LLM)** to generate chunks. The LLM processes the text and generates **semantically isolated** sentences or **propositions** that can stand alone. While this method is **highly accurate**, it is also the most **computationally demanding**.