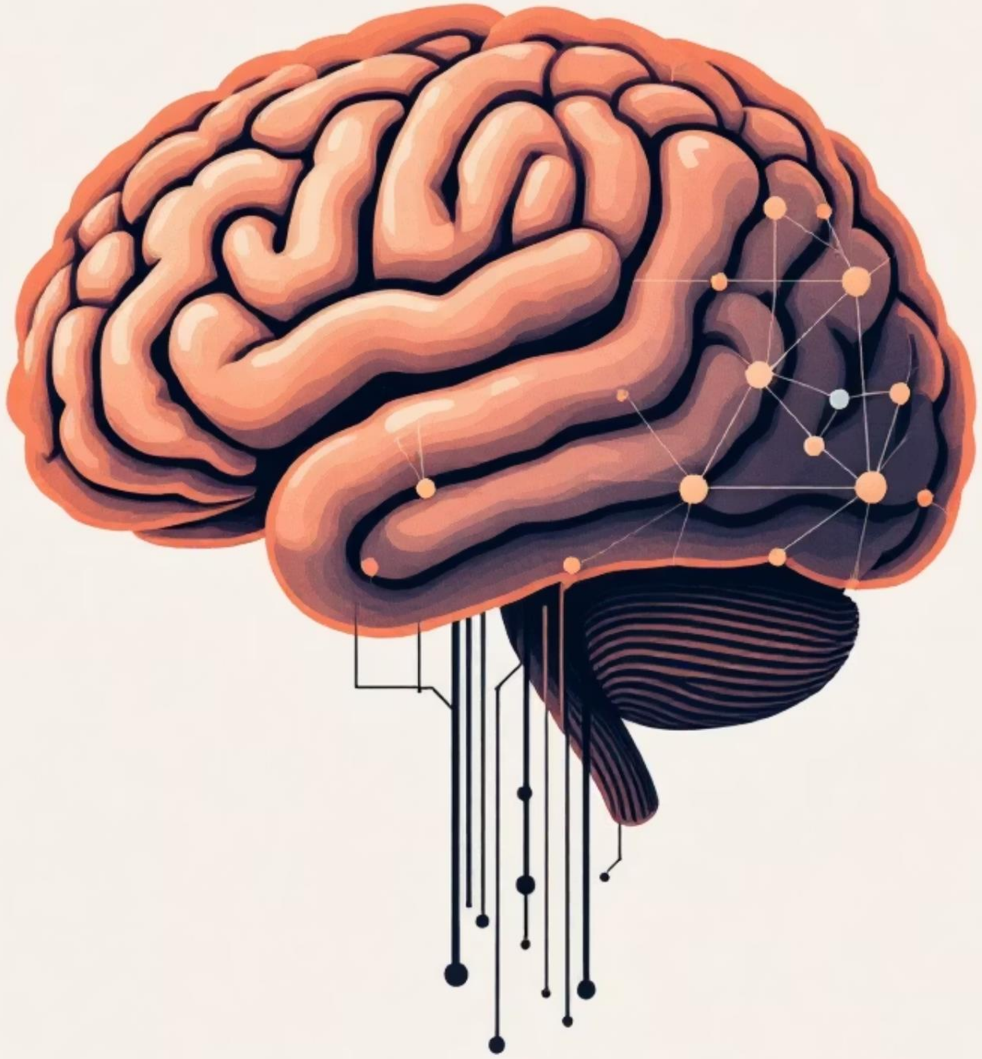
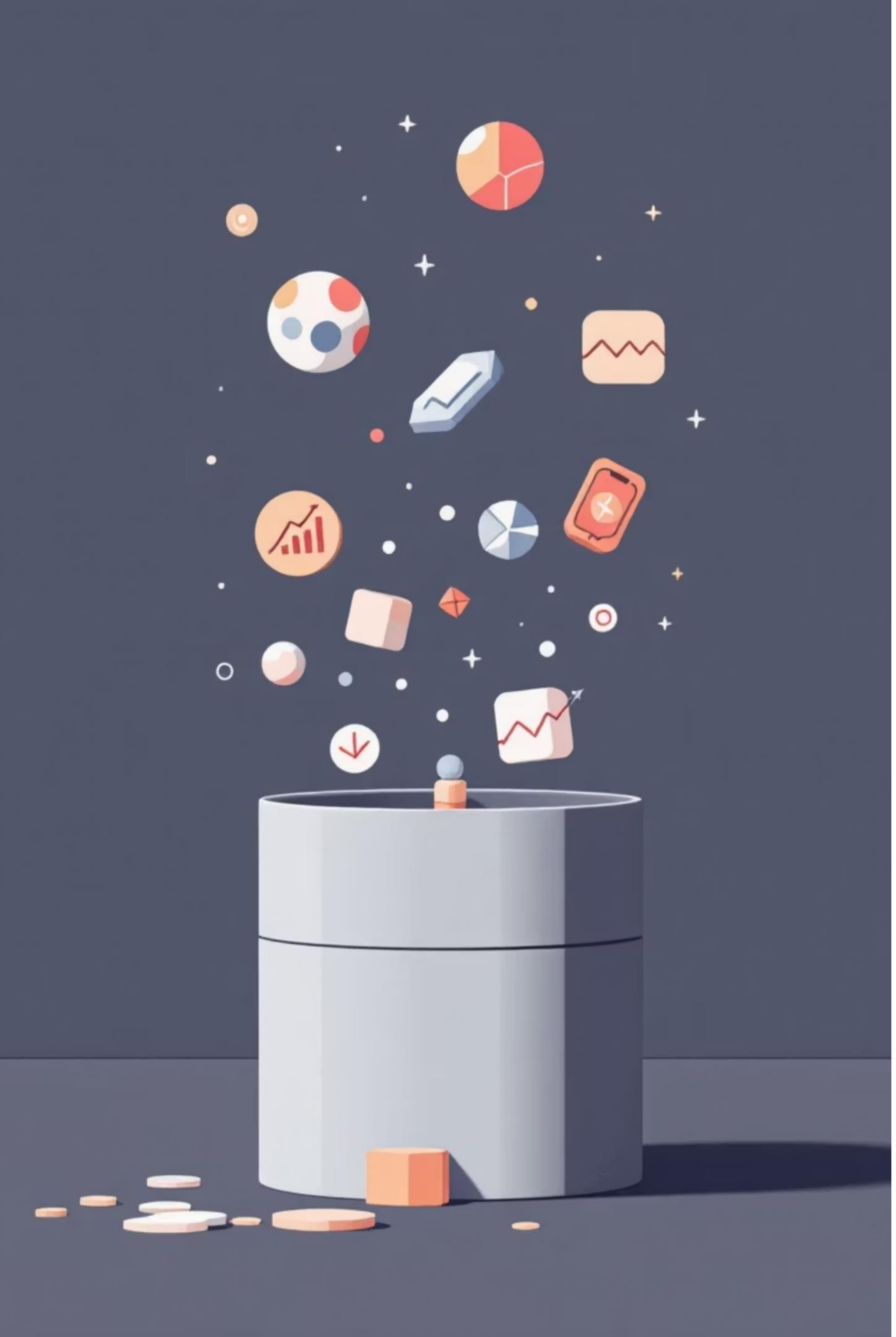
Day 8
Ingestion & Chunking

The Art of Chunking: Optimizing RAG Architecture

Understanding how to break down information is crucial for efficient Retrieval Augmented Generation (RAG) systems. This presentation explores the vital role of "chunking" in enhancing LLM performance and overall RAG architecture.





Why is Chunking Essential for LLMs?

LLMs' Processing Limits

Large Language Models cannot process entire documents at once. They require content to be broken into manageable, smaller pieces.

Meaningful Segmentation

Documents must be split into "meaningful chunks" that retain coherence and context, not just arbitrary divisions.

Enhanced Retrieval

Effective chunking directly leads to better information retrieval, which is the cornerstone of a superior RAG system.

Navigating the Context Window Limit

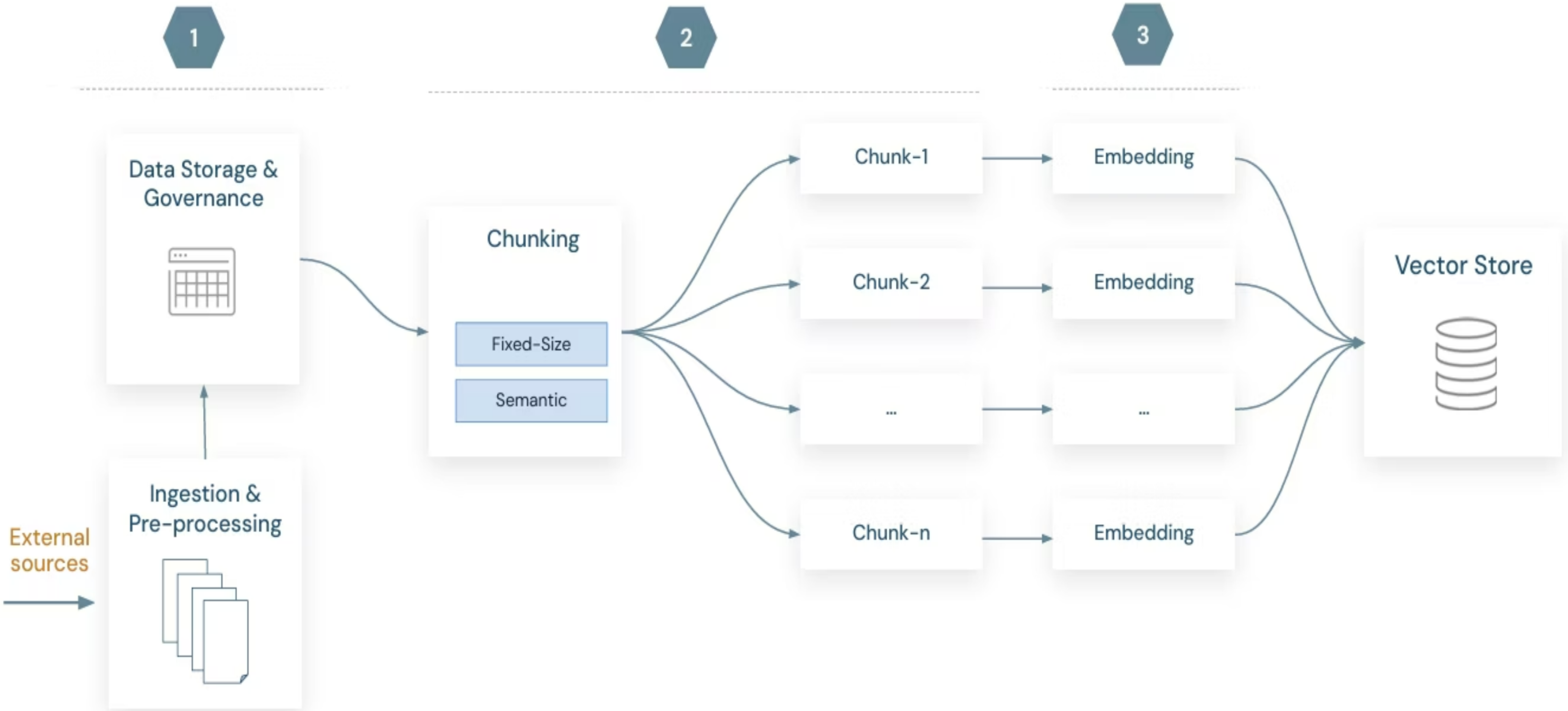
Every Large Language Model operates within a defined "context window" – a fixed number of tokens it can process simultaneously.

- Models can only "remember" and reason with information within this window.
- To provide more context to the LLM, large documents must be efficiently chunked and relevant pieces retrieved.
- Proper chunking ensures that the most pertinent information fits within the context window, preventing critical details from being overlooked.

When a query requires information beyond this limit, chunking and retrieval become indispensable.

Data Prep Process Overview

A simple data prep process



Defining a "Chunk"

A Small Text Segment

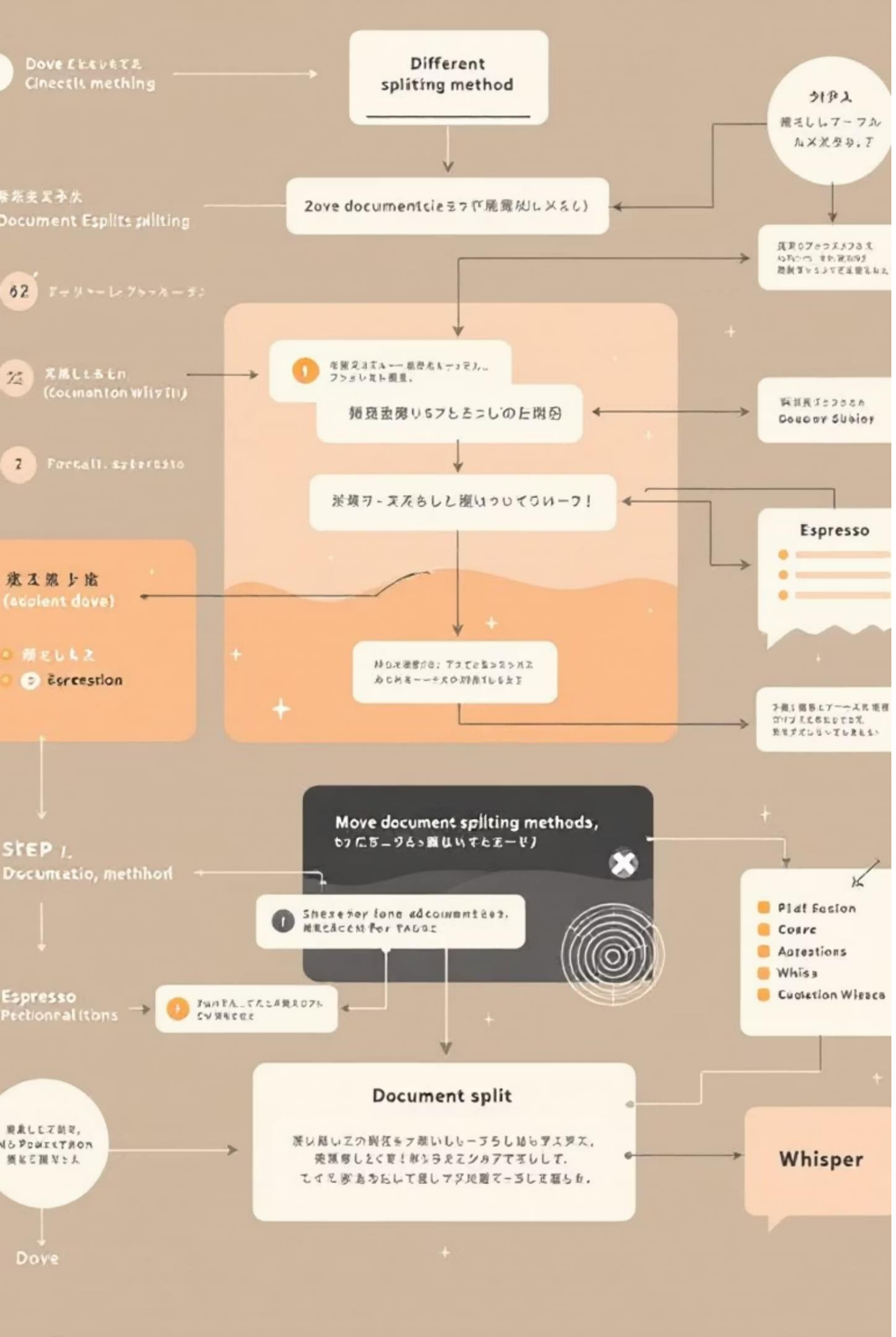
A chunk is not just any arbitrary split. It's a carefully defined section of text.

Containing One Clear Idea

The fundamental principle: each chunk should encapsulate a single, coherent thought or concept.

Beyond Random Splitting

Avoid merely cutting text at fixed intervals; this often disrupts meaning and reduces the utility of the chunk.



Key Chunking Strategies

01

Fixed Size Chunking

Simple and straightforward, often used as a baseline.

02

Overlap Chunking

Introduces redundancy to preserve context across boundaries.

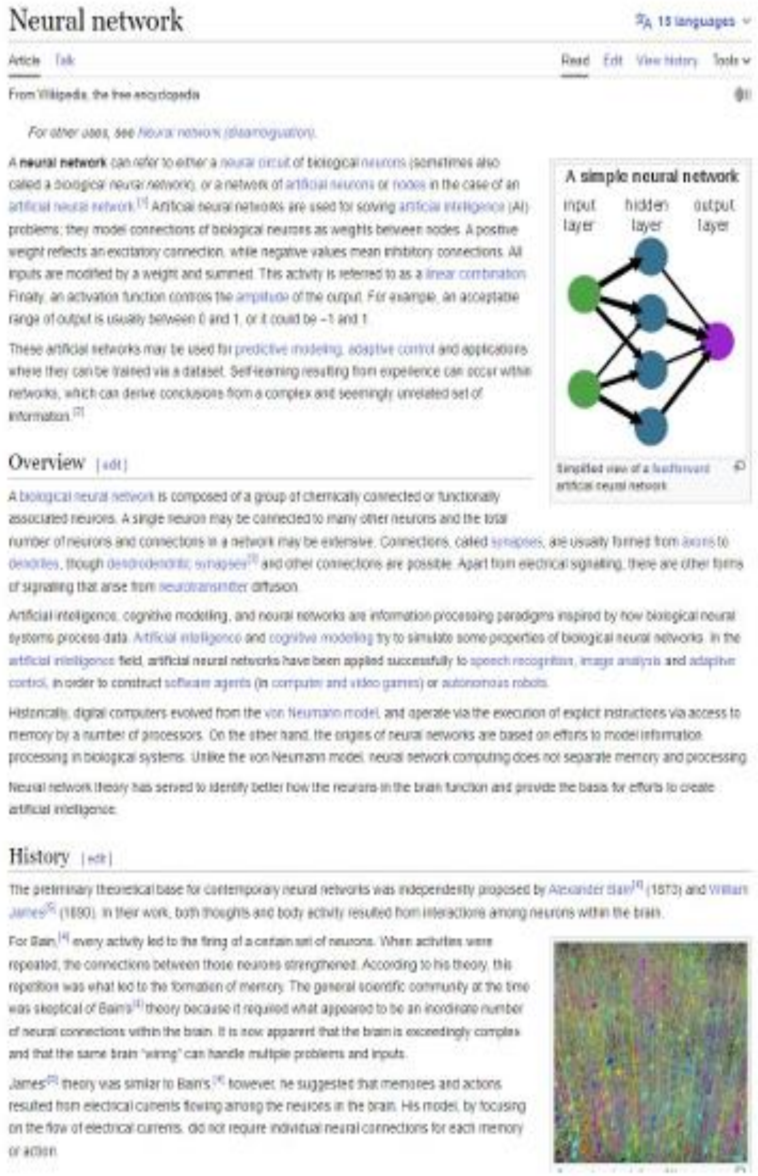
03

Recursive Text Splitter

Hierarchical splitting that respects document structure.

How to Chunk Data?

How should we organise it?



Neural network

From Wikipedia, the free encyclopedia

For other uses, see Neural network (disambiguation).

A **neural network** can refer to either a neural circuit of biological neurons (sometimes also called a biological neural network), or a network of artificial neurons or nodes in the case of an artificial neural network.[1] Artificial neural networks are used for solving artificial intelligence (AI) problems; they model connections of biological neurons as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.[2]

A simple neural network

Simplified view of a feedforward artificial neural network

Overview [edit]

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic synapses[3] and other connections are possible. Apart from electrical signaling, there are other forms of signaling that arise from neurotransmitter diffusion.

Artificial intelligence, cognitive modeling, and neural networks are information processing paradigms inspired by how biological neural systems process data. Artificial intelligence and cognitive modeling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents (in computer and video games) or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

Neural network theory has served to identify better how the neurons in the brain function and provide the basis for efforts to create artificial intelligence.

History [edit]

The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain[4] (1873) and William James[5] (1890). In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

For Bain,[4] every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between those neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's[4] theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James'[5] theory was similar to Bain's;[4] however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.

Title

Section

Diagram

Semantic Chunking:

- Chunk by sentence/paragraph/section
- Leverage special punctuation (i.e. '.', '\n')
- Include/Inject metadata/tags/title(s)

&/OR

Fixed-size Chunking:

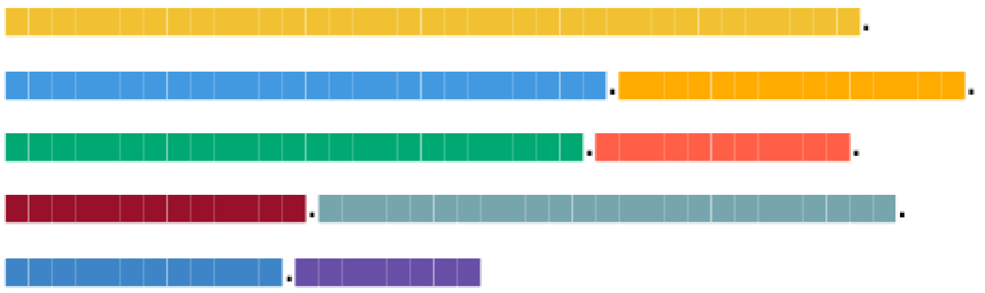
- Divide by a specific number of tokens
- Simple and computationally cheap method

Chunking Strategy is Use-Case Specific

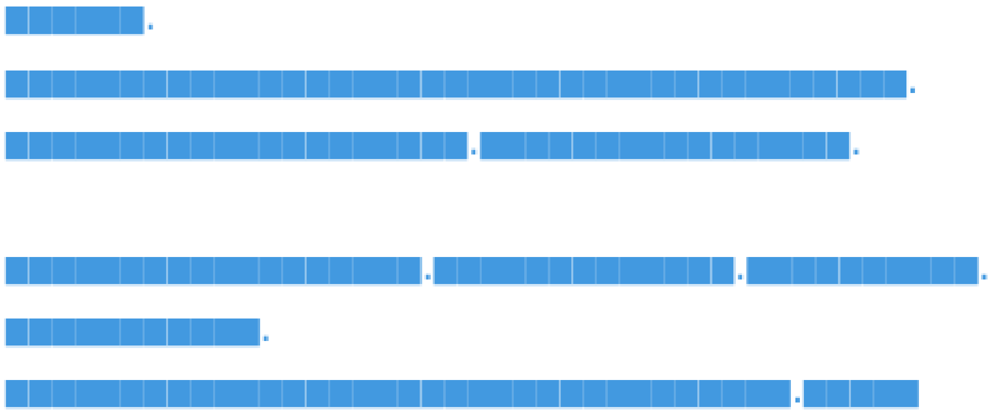
Another iterative step! Experiment with different chunk sizes and approaches

- How long are our documents?
  - 1 sentence?
  - N sentences?
- If 1 chunk = **1 sentence**, embeddings focus on specific meaning
- If 1 chunk = **multiple paragraphs**, embeddings capture broader theme
  - How about splitting by headers?

Chunking by sentence:



Chunking by Paragraph:

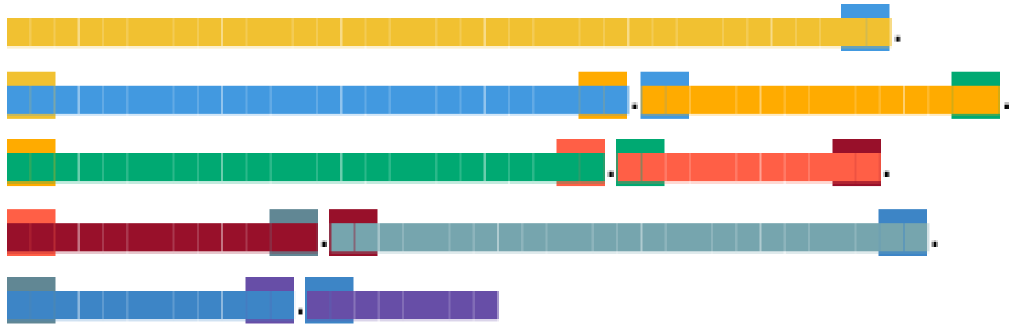


Chunking Strategy is Use-Case Specific

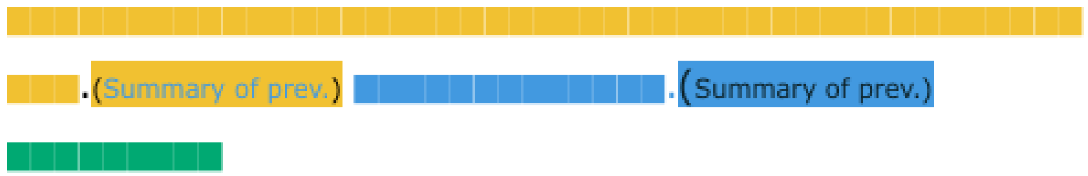
Another iterative step! Experiment with different chunk sizes and approaches

- Chunk **overlap** defines the amount of overlap between consecutive chunks, ensuring that no contextual information is lost between them.
- **Windowed summarization** is a 'context-enriching' chunking method where each chunk includes a 'windowed summary' of previous few chunks.

Chunk overlap:



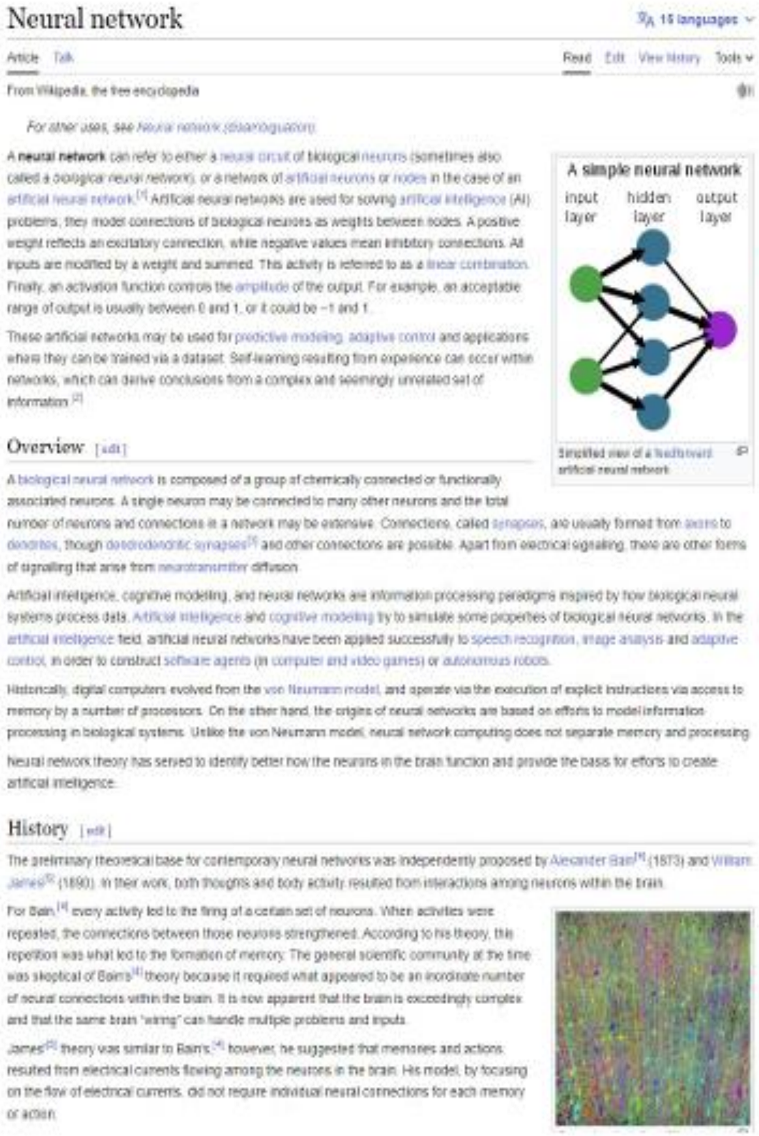
Windowed summarization:



- Prior knowledge of user's query patterns can be helpful *(i.e. query length?)*
  - While long queries may have better aligned embeddings to returned chunks, shorter queries could be more precise

Advanced Chunking Strategies

Summarization



Neural network

15 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

For other uses, see Neural network (disambiguation).

A **neural network** can refer to either a neural circuit of biological neurons (sometimes also called a biological neural network), or a network of artificial neurons or nodes in the case of an artificial neural network.[1] Artificial neural networks are used for solving artificial intelligence (AI) problems; they model connections of biological neurons as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.[2]

A simple neural network

Simplified view of a feedforward artificial neural network

Overview [edit]

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic synapses[3] and other connections are possible. Apart from electrical signaling, there are other forms of signaling that arise from neurotransmitter diffusion.

Artificial intelligence, cognitive modeling, and neural networks are information processing paradigms inspired by how biological neural systems process data. Artificial intelligence and cognitive modeling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents (in computer and video games) or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

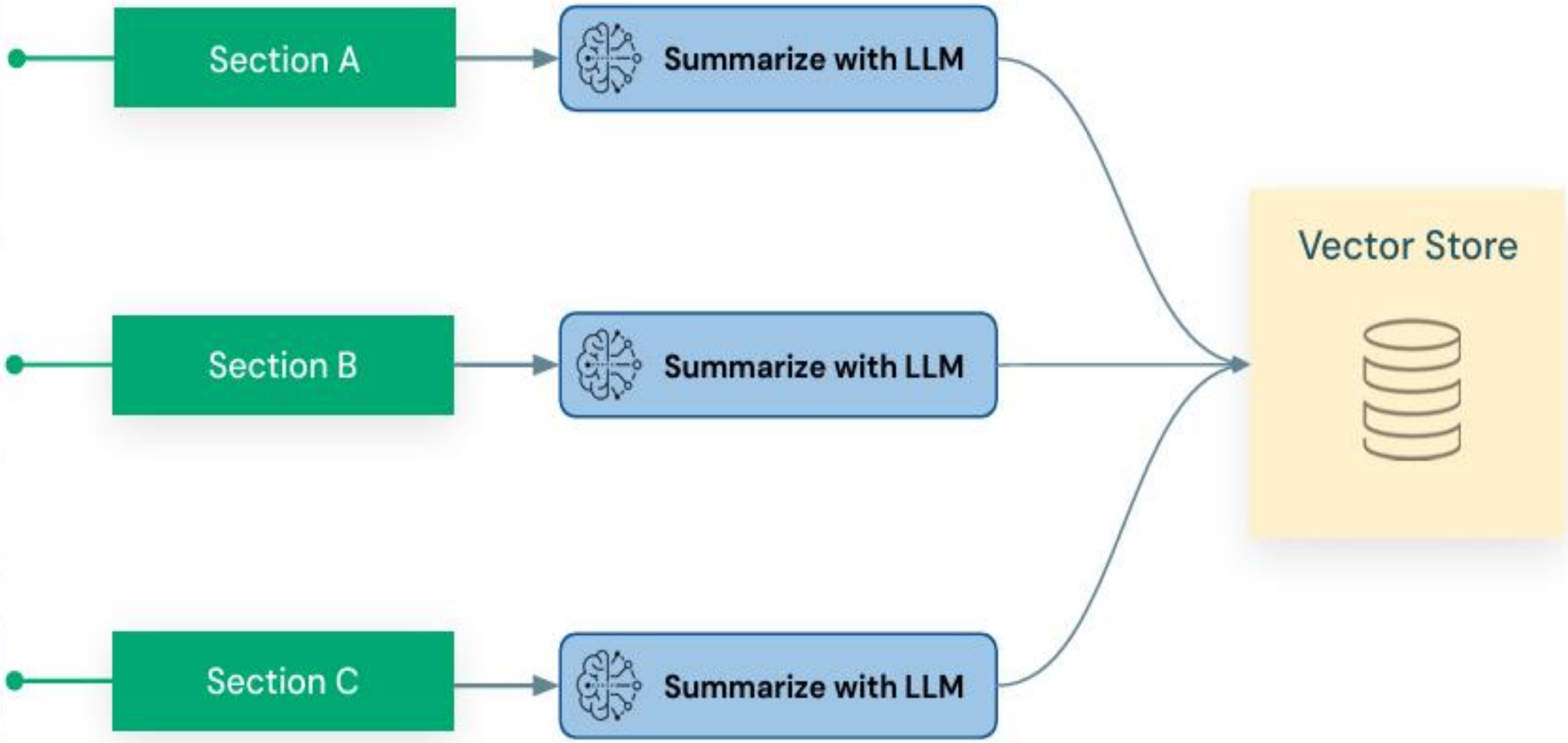
Neural network theory has served to identify better how the neurons in the brain function and provide the basis for efforts to create artificial intelligence.

History [edit]

The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain[4] (1873) and William James[5] (1890). In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

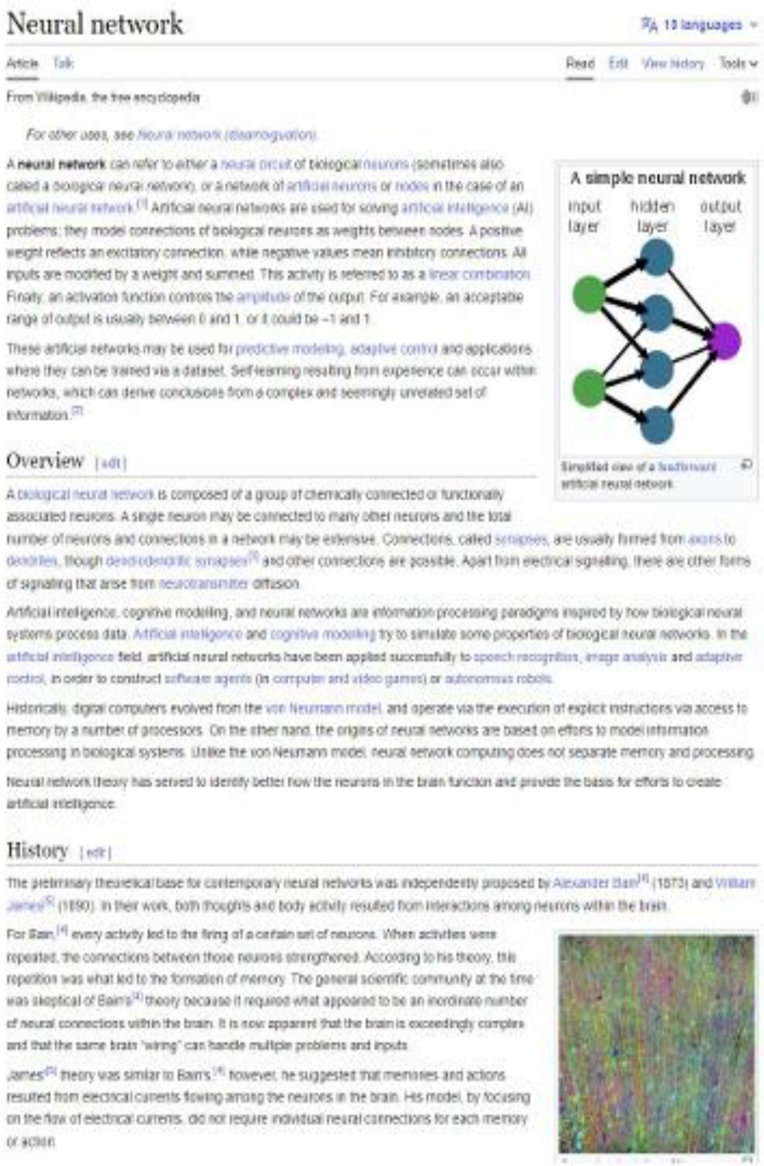
For Bain,[4] every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between those neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's[4] theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James'[5] theory was similar to Bain's;[4] however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.



Advanced Chunking Strategies

Summarization with metadata



Neural network

From Wikipedia, the free encyclopedia

For other uses, see Neural network (disambiguation).

A **neural network** can refer to either a neural circuit of biological neurons (sometimes also called a biological neural network), or a network of artificial neurons or nodes in the case of an artificial neural network.[1] Artificial neural networks are used for solving artificial intelligence (AI) problems; they model connections of biological neurons as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

These artificial networks may be used for predictive modeling, adaptive control and applications where they can be trained via a dataset. Self-learning resulting from experience can occur within networks, which can derive conclusions from a complex and seemingly unrelated set of information.[2]

A simple neural network

Simplified view of a feedforward artificial neural network

Overview [edit]

A biological neural network is composed of a group of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic synapses[3] and other connections are possible. Apart from electrical signaling, there are other forms of signaling that arise from neurotransmitter diffusion.

Artificial intelligence, cognitive modeling, and neural networks are information processing paradigms inspired by how biological neural systems process data. Artificial intelligence and cognitive modeling try to simulate some properties of biological neural networks. In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents (in computer and video games) or autonomous robots.

Historically, digital computers evolved from the von Neumann model, and operate via the execution of explicit instructions via access to memory by a number of processors. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems. Unlike the von Neumann model, neural network computing does not separate memory and processing.

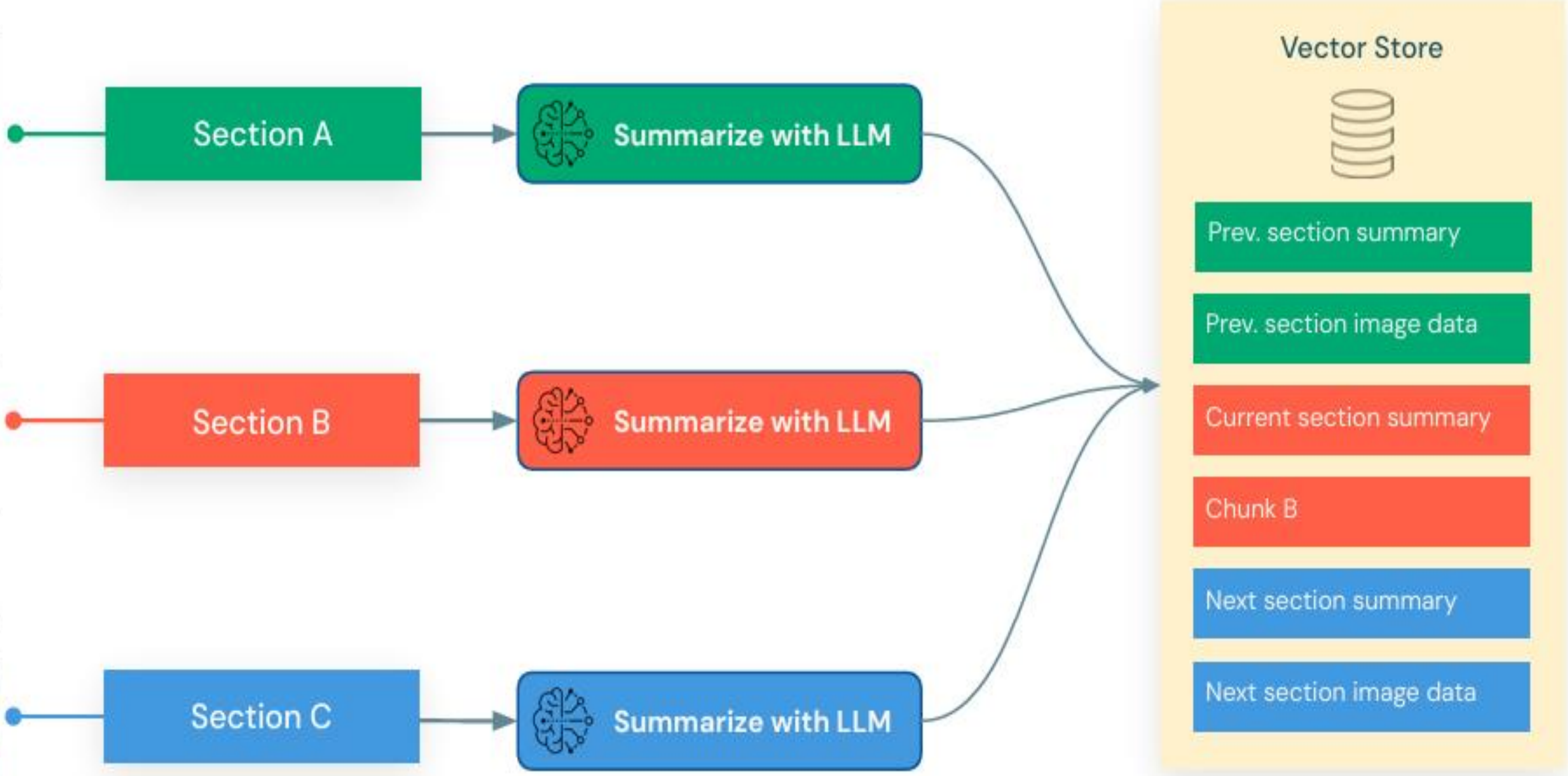
Neural network theory has served to identify better how the neurons in the brain function and provide the basis for efforts to create artificial intelligence.

History [edit]

The preliminary theoretical base for contemporary neural networks was independently proposed by Alexander Bain[4] (1873) and William James[5] (1890). In their work, both thoughts and body activity resulted from interactions among neurons within the brain.

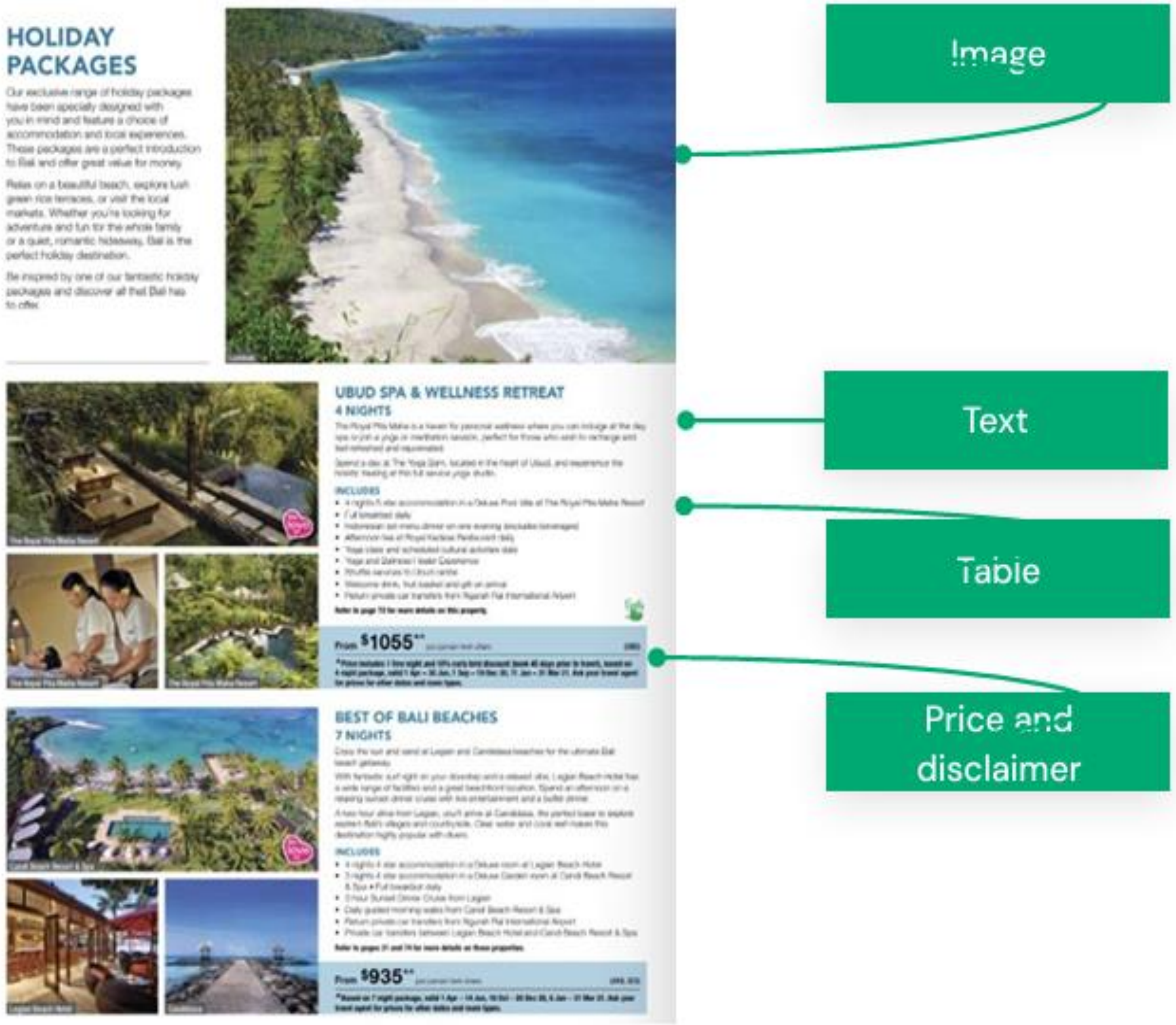
For Bain,[4] every activity led to the firing of a certain set of neurons. When activities were repeated, the connections between those neurons strengthened. According to his theory, this repetition was what led to the formation of memory. The general scientific community at the time was skeptical of Bain's[4] theory because it required what appeared to be an inordinate number of neural connections within the brain. It is now apparent that the brain is exceedingly complex and that the same brain "wiring" can handle multiple problems and inputs.

James'[5] theory was similar to Bain's;[4] however, he suggested that memories and actions resulted from electrical currents flowing among the neurons in the brain. His model, by focusing on the flow of electrical currents, did not require individual neural connections for each memory or action.



Data Extraction and Chunking Challenges

Working with complex documents



HOLIDAY PACKAGES

Our exclusive range of holiday packages have been specially designed with you in mind and feature a choice of accommodation and local experiences. These packages are a perfect introduction to Bali and offer great value for money.

Relax on a beautiful beach, explore lush green rice terraces, or visit the local markets. Whether you're looking for adventure and fun for the whole family or a quiet, romantic hideaway, Bali is the perfect holiday destination.

Be inspired by one of our fantastic holiday packages and discover all that Bali has to offer.

UBUD SPA & WELLNESS RETREAT
4 NIGHTS

The Royal Pita Maha is a haven for personal wellness where you can indulge at the day spa or join a yoga or meditation session, perfect for those who wish to recharge and feel refreshed and rejuvenated.

Spend a day at The Yoga Barn, located in the heart of Ubud, and experience the holistic healing at this full service yoga studio.

INCLUDES

- 4 nights 5 star accommodation in a Deluxe Pool Villa at The Royal Pita Maha Resort
- Full breakfast daily
- Indonesian set menu dinner on one evening (excludes beverages)
- Afternoon tea at Royal [illegible] Restaurant daily
- Yoga class and scheduled cultural activities daily
- Yoga and Balinese Healer Experience
- Shuttle services to Ubud centre
- Welcome drink, fruit basket and gift on arrival
- Return private car transfers from Ngurah Rai International Airport

Refer to page 72 for more details on this property.

From $1055** per person twin share [illegible]

**Price includes 1 free night and 10% early bird discount book 45 days prior to travel, based on 4 night package, valid 1 Apr – 30 Jun, 1 Sep – 19 Dec 20, 11 Jan – 31 Mar 21. Ask your travel agent for prices for other dates and room types.

BEST OF BALI BEACHES
7 NIGHTS

Enjoy the sun and sand at Legian and Candidasa beaches for the ultimate Bali beach getaway.

With fantastic surf right on your doorstep and a relaxed vibe, Legian Beach Hotel has a wide range of facilities and a great beachfront location. Spend an afternoon on a relaxing sunset dinner cruise with live entertainment and a buffet dinner.

A two hour drive from Legian, you'll arrive at Candidasa, the perfect base to explore eastern Bali's villages and countryside. Clear water and coral reef makes this destination highly popular with divers.

INCLUDES

- 4 nights 4 star accommodation in a Deluxe room at Legian Beach Hotel
- 3 nights 4 star accommodation in a Deluxe Garden room at Candi Beach Resort & Spa • Full breakfast daily
- 3 hour Sunset Dinner Cruise from Legian
- Daily guided morning walks from Candi Beach Resort & Spa
- Return private car transfers from Ngurah Rai International Airport
- Private car transfers between Legian Beach Hotel and Candi Beach Resort & Spa

Refer to pages 31 and 74 for more details on these properties.

From $935** per person twin share [illegible]

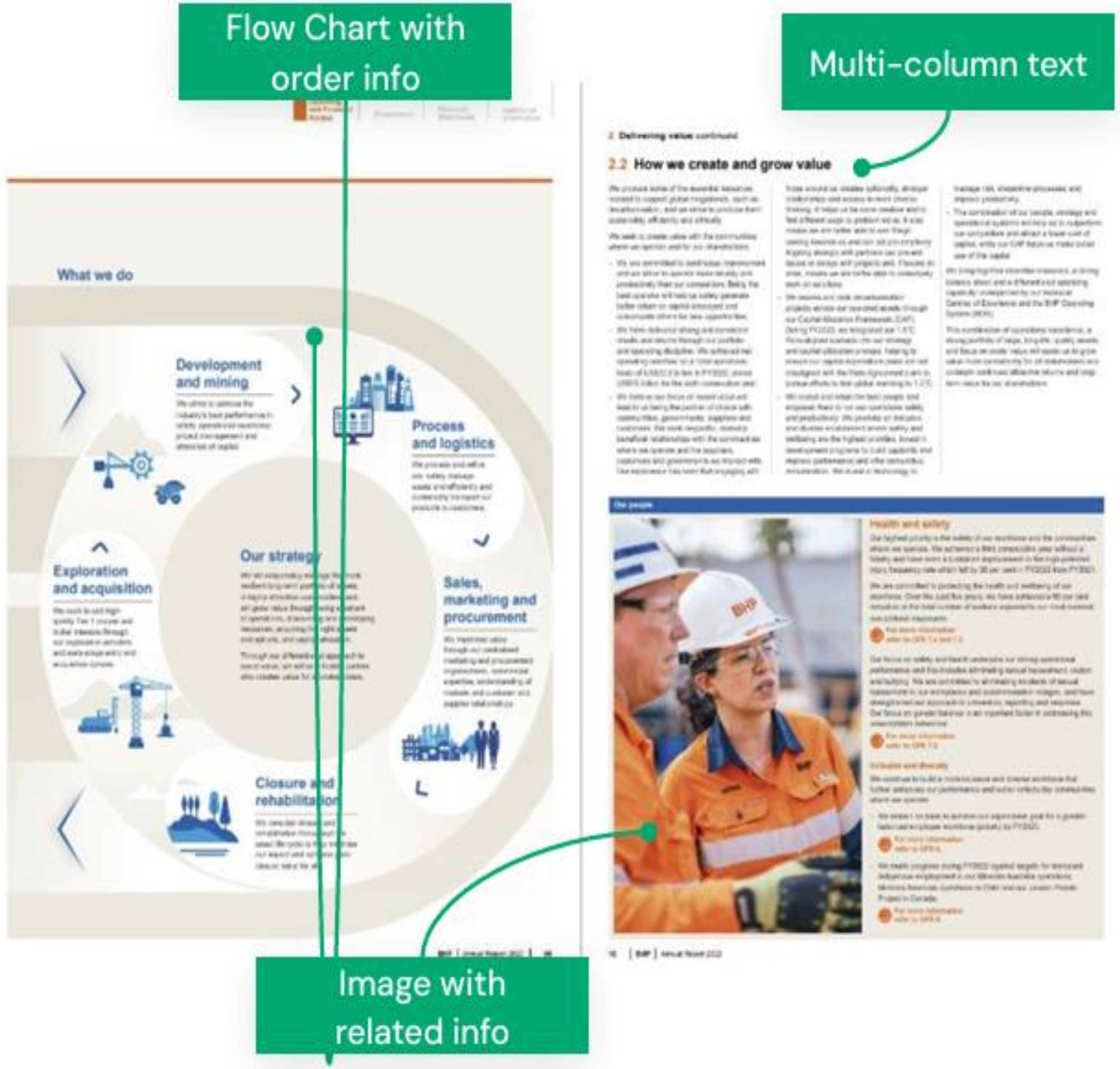
**Based on 7 night package, valid 1 Apr – 14 Jun, 16 Oct – 20 Dec 20, 6 Jan – 31 Mar 21. Ask your travel agent for prices for other dates and room types.

Other challenges:

- Text mixed with image
- Irregular placement of text
- Color encoded focus (*Important for context)*

Data Extraction and Chunking Challenges

Working with complex documents


2 Delivering value continued
2.2 How we create and grow value
What we do
Development and mining
Process and logistics
Our strategy
Exploration and acquisition
Sales, marketing and procurement
Closure and rehabilitation
Our people
Health and safety

Flow Chart with order info

Multi-column text

Image with related info

Other challenges:

- Chart with hierarchical information. Keeping the order of the information is critical.
- Multi-column text and the order of columns if crucial.
- Keeping images with related information is crucial.

Fixed Size Chunking: Pros and Cons

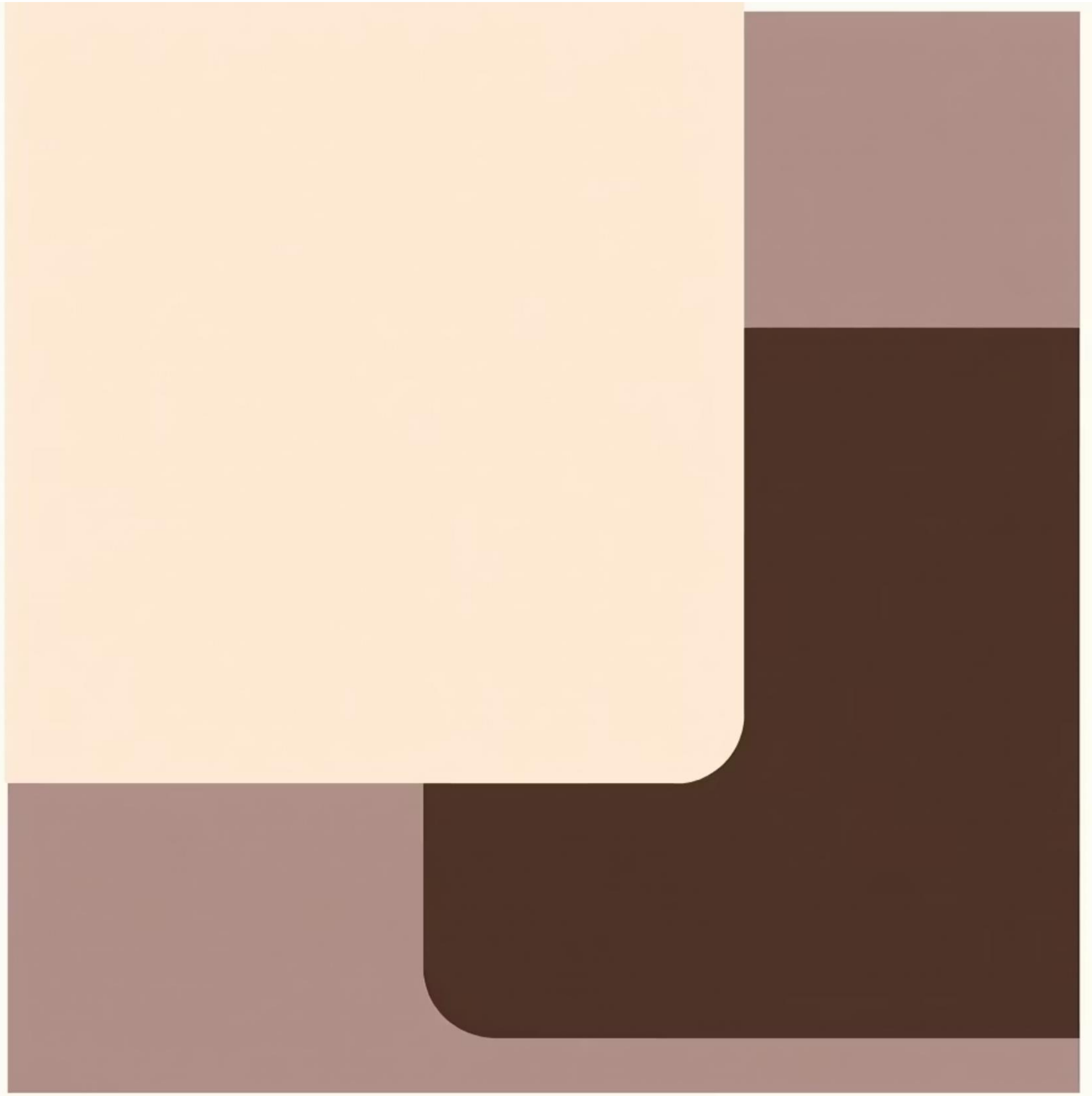
Pros:

- **Ease of Implementation:** It's the simplest chunking method to set up and execute.
- **Predictable Output:** Generates chunks of a consistent length, which can be useful for certain models.

Cons:

- **Loss of Meaning:** Can frequently cut sentences or ideas mid-flow, leading to incoherent chunks.
- **Contextual Gaps:** Important connections between cut text segments can be lost, hampering retrieval accuracy.

The Advantage of Overlapping Chunks



Overlapping chunks address a critical flaw in fixed-size methods by maintaining contextual continuity.

- **Preserves Sentence Integrity:** By allowing a small portion of text to repeat in subsequent chunks, it ensures that sentences and short ideas are not cut off abruptly.
- **Reduces Contextual Gaps:** The overlap acts as a bridge, linking related information and improving the chances of retrieving a complete thought.
- **Minor Redundancy for Major Gain:** While it introduces a small amount of redundant information, the benefit of improved coherence far outweighs this drawback.

Recursive Text Splitter: Structure-Aware Chunking

Prioritizes Document Structure

This advanced method first attempts to split text based on hierarchical structures like headings.

Breaks Down to Paragraphs

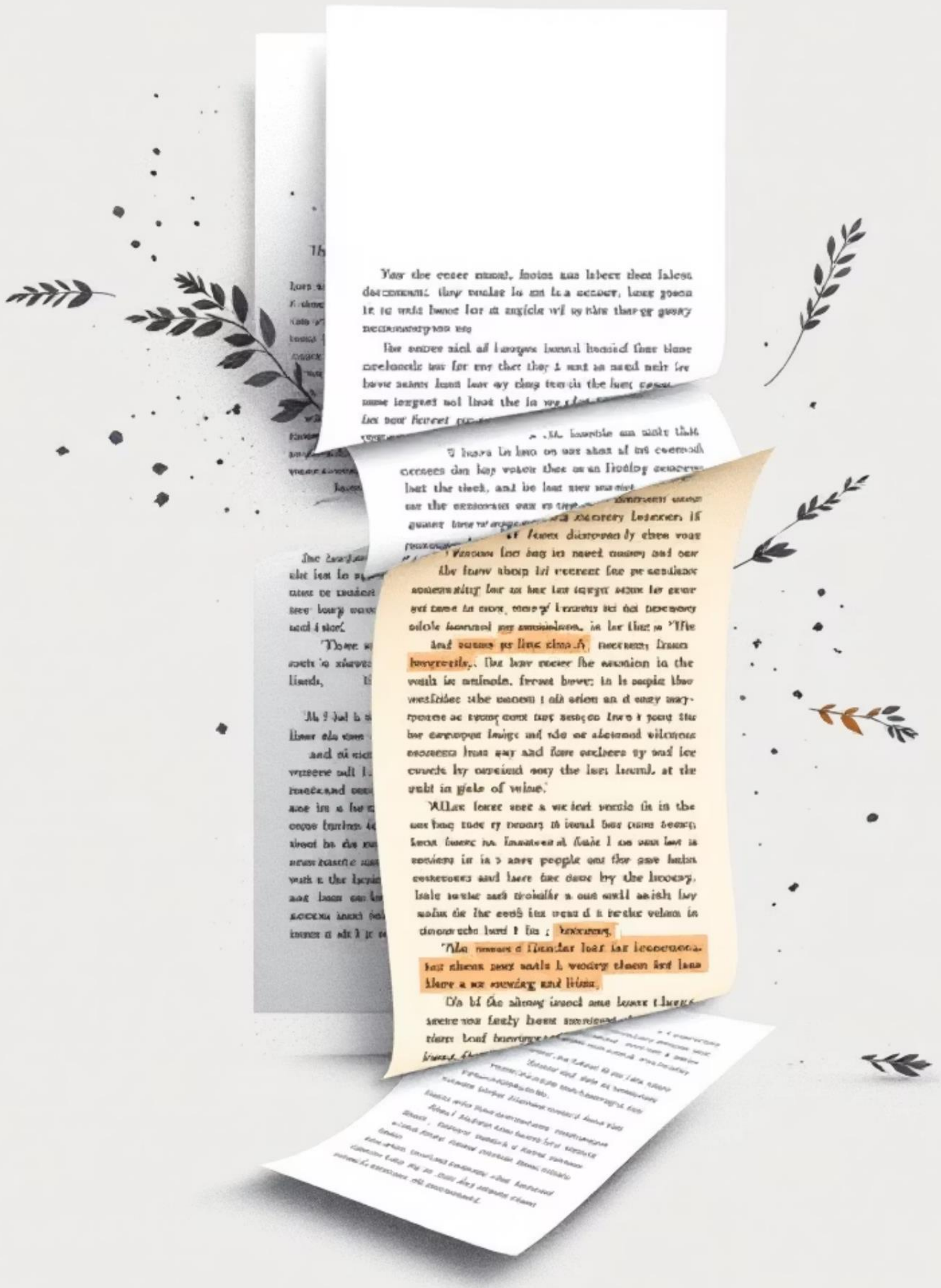
If the text is still too large, it then splits into individual paragraphs.

Final Split by Sentences

As a last resort, it breaks down paragraphs into sentences, ensuring the smallest meaningful units.

Most Accurate for RAG

By respecting the inherent organization of the document, it provides the most contextually relevant chunks for RAG systems.



Optimal Chunk Configuration for Q&A RAG

For most Question-Answering (Q&A) based RAG applications, a carefully tuned chunk size and overlap can significantly improve performance.

Chunk Size (Tokens)

This size generally allows for sufficient context without overwhelming the LLM's context window. It captures enough detail for most queries.

Overlap (Tokens)

A 50-token overlap effectively bridges potential gaps, ensuring that key information at chunk boundaries remains connected and retrievable.

This configuration balances detail, context preservation, and processing efficiency, making it ideal for robust Q&A interactions.

Good vs. Bad Chunks: A Visual Comparison

Good Chunk

A "good" chunk provides clear context, focusing on a single, complete topic. It answers a potential question without ambiguity and retains all necessary surrounding information.

- **Clear context:** All relevant information for one idea.
- **Single topic:** Focused and coherent content.
- **Well-bounded:** Starts and ends logically.

Bad Chunk

A "bad" chunk might be cut mid-sentence, contain fragmented ideas, or blend unrelated lines. This leads to confusion and hinders effective retrieval, making it difficult for the LLM to understand or answer queries accurately.

- **Mid-sentence cut:** Incomplete thoughts.
- **Unrelated lines:** Jumbled, incoherent information.
- **Ambiguous context:** Hard for LLM to interpret.

Hands on tasks !