AI Literacy

Introduction to Transformers, Advanced AI Models, and Natural Language Processing (NLP)

Quick Recap

- Artificial Intelligence has evolved significantly, from early rule-based systems in the 1950s-1980s to the rise of machine learning in the 1990s and breakthroughs in deep learning with models like GPT.
- Machine learning (ML) and deep learning (DL) power AI advancements, with ML focusing on pattern recognition and decision-making, while DL uses neural networks for complex tasks like image processing, speech recognition, and automation across industries.
- AI's impact is seen in real-world applications such as Google Ads optimizing marketing campaigns, self-driving cars improving transportation, and AI chatbots enhancing customer interactions.

Engage and Think

Imagine you are a product manager at a growing tech company, exploring AI to improve customer engagement. You've heard about transformers and NLP models like BERT, GPT, and T5 but are unsure how to use them effectively.

Competitors already leverage AI-powered chatbots, personalized recommendations, and automated content to enhance customer experience. With transformers and NLP, businesses can analyze text, classify sentiment, and generate human-like responses effortlessly.

How will you use these AI advancements to stay competitive?

Learning Objectives

By the end of this lesson, you will be able to:

- Explain the fundamental concepts of transformers and NLP, including their architectures, components, and how they enable state-of-the-art AI applications
- Analyze how transformers process text using tokenization, embeddings, and positional encoding and evaluate their role in revolutionizing AI-driven natural language understanding
- Apply key natural language processing (NLP) techniques, such as text classification, sentiment analysis, and chatbots, to real-world use cases in customer service and business automation
- Evaluate the impact of advanced AI models (transformers & NLP) on industries by assessing their applications in chatbots, content generation, and automated decision-making

Attention Mechanism and Transformers

Attention Mechanism

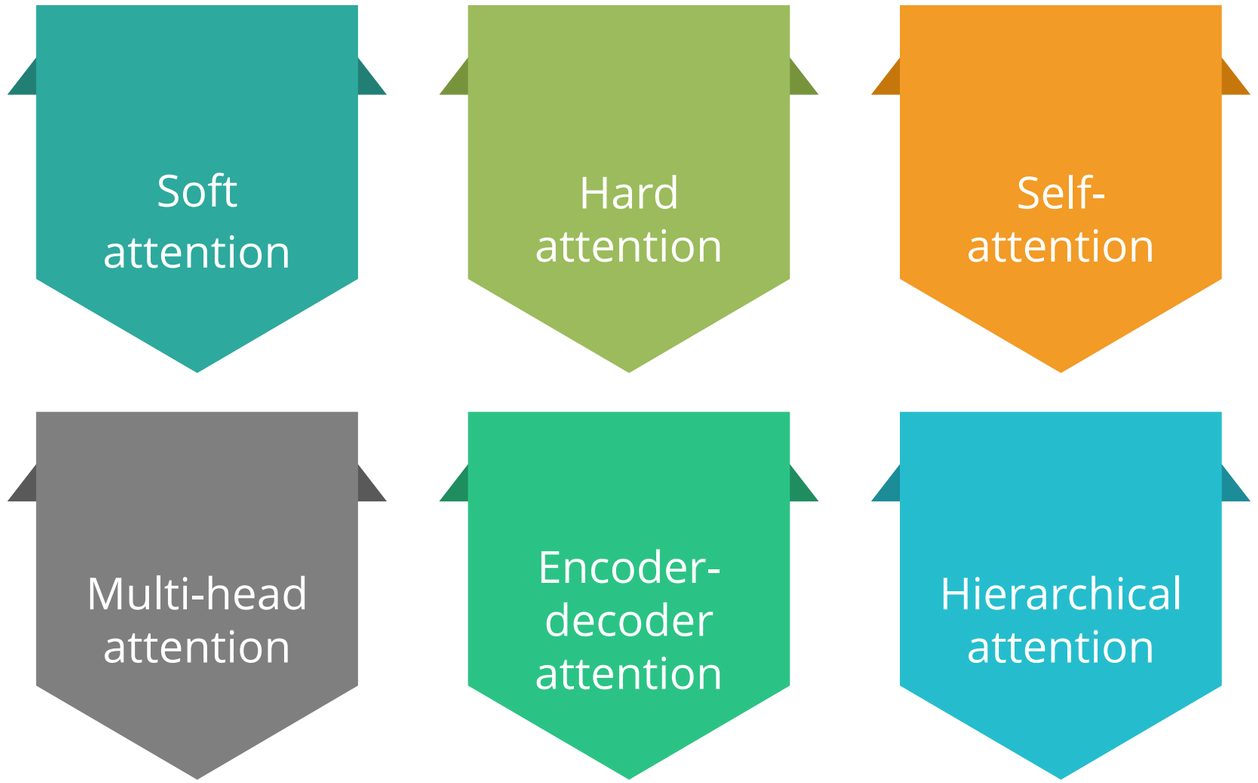
It is a core component of transformers, enabling them to process and understand sequences efficiently without relying on recurrent or convolutional structures.

Unlike traditional models, transformers do not process data sequentially; instead, they analyze all words at once using self-attention.

The attention mechanism helps transformers pay attention to the most relevant words in a sentence, even if they are far apart.

Attention Mechanism: Types

Different attention mechanisms serve various purposes in NLP, computer vision, and deep learning applications.



Introduction to Transformer Models

Transformer models are a type of deep learning architecture that leverages self-attention mechanisms to enable simultaneous processing of all sequence elements.

RNNs process data sequentially, requiring each step in a sequence to be handled one after another, unlike transformers, which process elements simultaneously.

Understanding Self-Attention

Self-attention, a key component in natural language processing (NLP), enables the network to focus on specific words or phrases in a sentence to improve context understanding.

While reading a novel, you simultaneously focus on the current page and recall earlier events, characters, and clues.

Example:

You are reading a mystery novel.

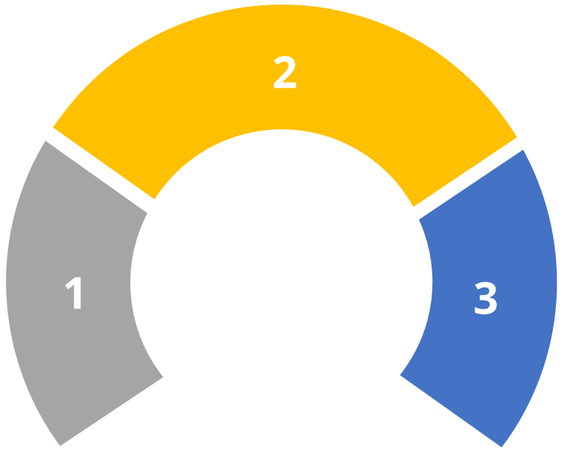
This context helps you understand the story better and predict future events.

Mechanics Behind Self-Attention

The self-attention layer calculates three vectors from each encoder's input vector:

Key vector (K): This vector scores the attentiveness of each word.

Query vector (Q): This vector scores each word regarding the extent of attention it needs.



Value vector (V): This vector represents the actual word content, which generate the final output.

Mechanics Behind Self-Attention

During training, vectors are iteratively trained and updated.

The following equation defines the attention score for each input word.

$$\text{Attention}\ (\text{Q,K,V}) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\text{V}$$

Understanding Mechanics Behind Self-Attention: Example

A person at a party tries to decide who to listen to in a crowded room. Imagine each person at the party has a story to tell, but not all stories are equally important to that person. So, everyone needs to decide how much attention each speaker should get.

Understanding Mechanics Behind Self-Attention: Example

Understand how self-attention works in this analogy:

Listening: Each person (data point or word in a sentence) listens to the stories (inputs) of others in the room (sequence).

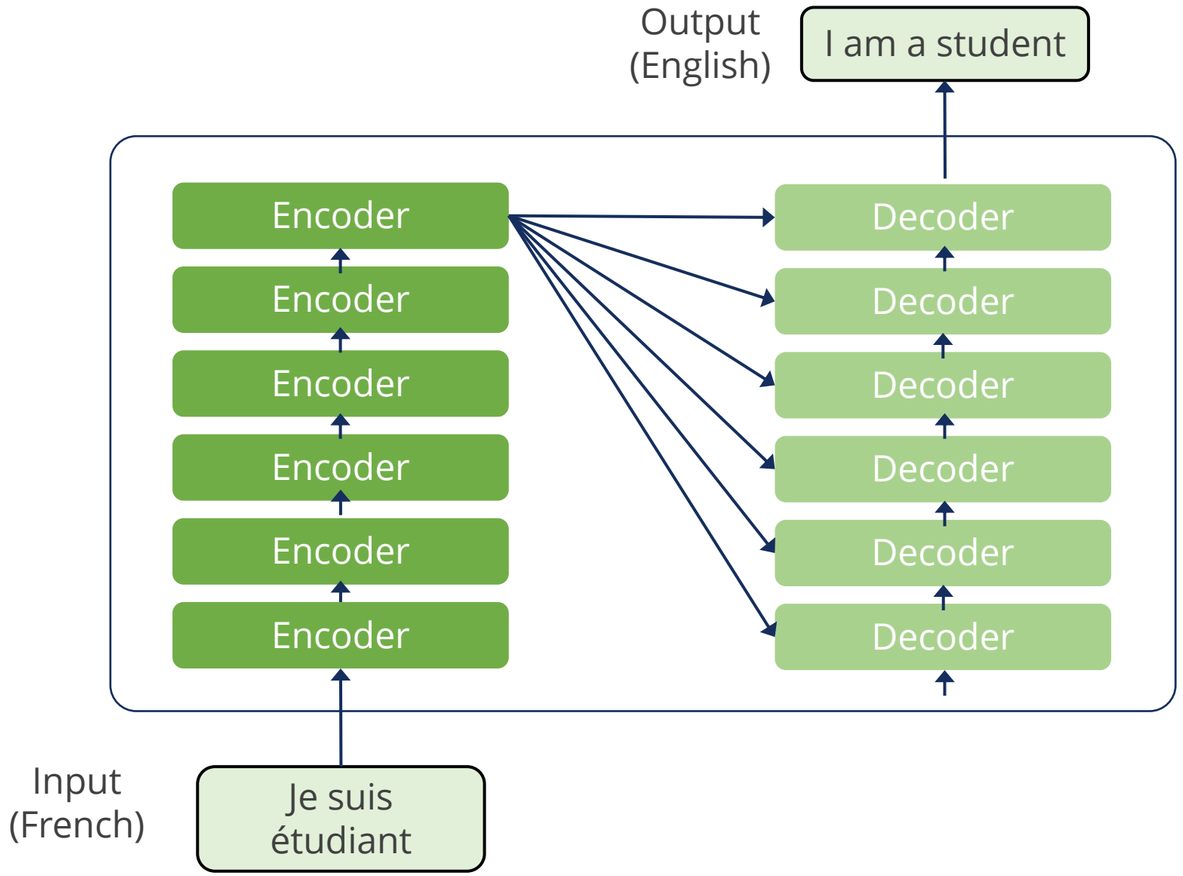
Scoring: Everyone assigns a score to each storyteller based on how relevant each story is to their current interests or the context of the conversation (query-key matching).

Focusing: The focus is more on the stories with higher scores (these get more attention).

Combining: Everyone creates a summary of what's important from all these stories, weighted by how much attention was paid to each (weighted sum of values).

Transformer Model Architecture

At a fundamental level, a transformer architecture comprises two primary components: an encoder and a decoder.



Transformer Model Architecture

Working of encoder and decoder

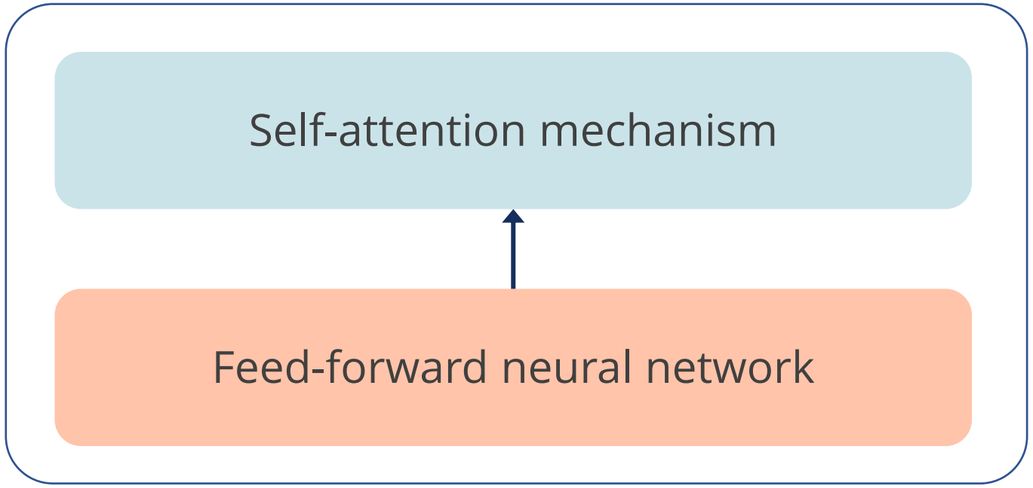
- The encoder processes input sequences and captures contextual information, employing self-attention to integrate information across the entire sequence.
- The decoder generates an output sequence and predicts the next word based on context.
- The architecture ensures accurate and coherent sequence processing and generation.

Transformer Working: Encoder

An example of language translation to understand the workings of a transformer:

Input: Je suis étudiant

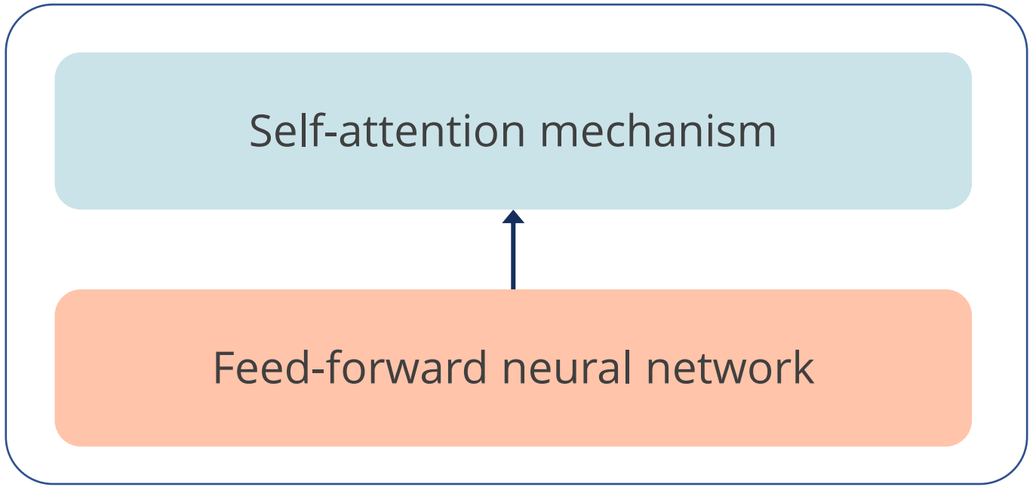
Each encoder has two layers:



- **Input processing**: The transformer model begins by taking an input, in this case, '**Je suis étudiant'** (I am a student in French). Each word is first converted into a vector through an embedding process.
- **Positional encoding**: After embedding, positional encoding is added to each word vector. This step is crucial as it injects information about the position of each word in the sequence into its vector representation, allowing the model to recognize word order, which is essential for understanding the syntax and semantics of the input sentence.

Transformer Working: Encoder

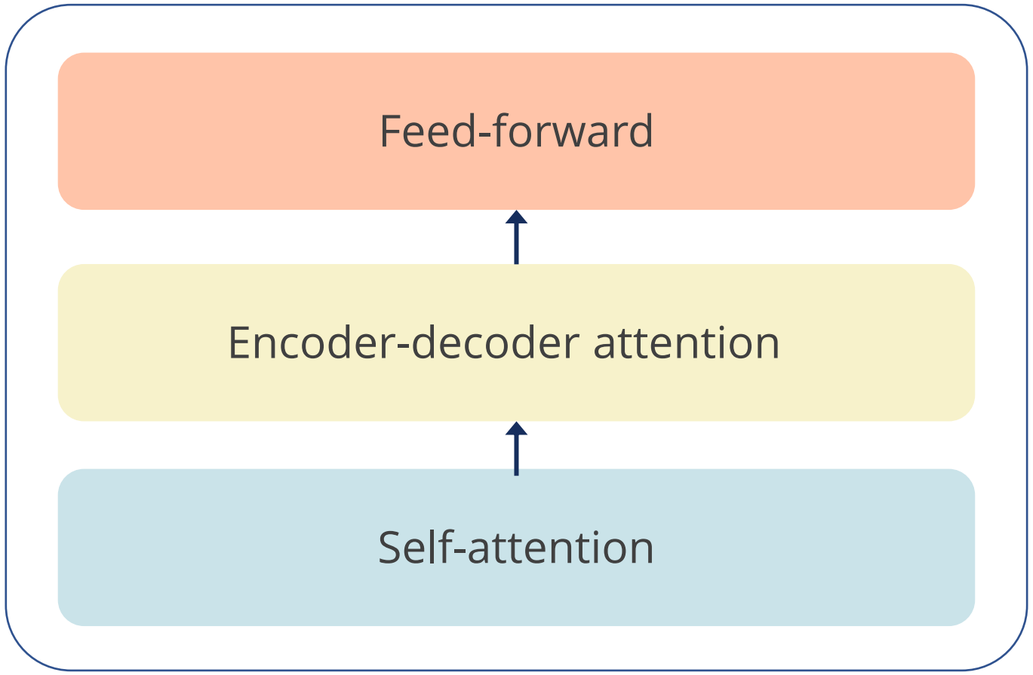
Each encoder has two layers:



- **Passing through encoders:** The word vectors, now enhanced with positional information, pass through multiple layers of encoders. Each encoder layer processes the vectors, refining and enriching the representations with context from the entire sentence. This process leverages self-attention and feed-forward neural networks within each layer.

Transformer Working: Decoder

A decoder has three layers:

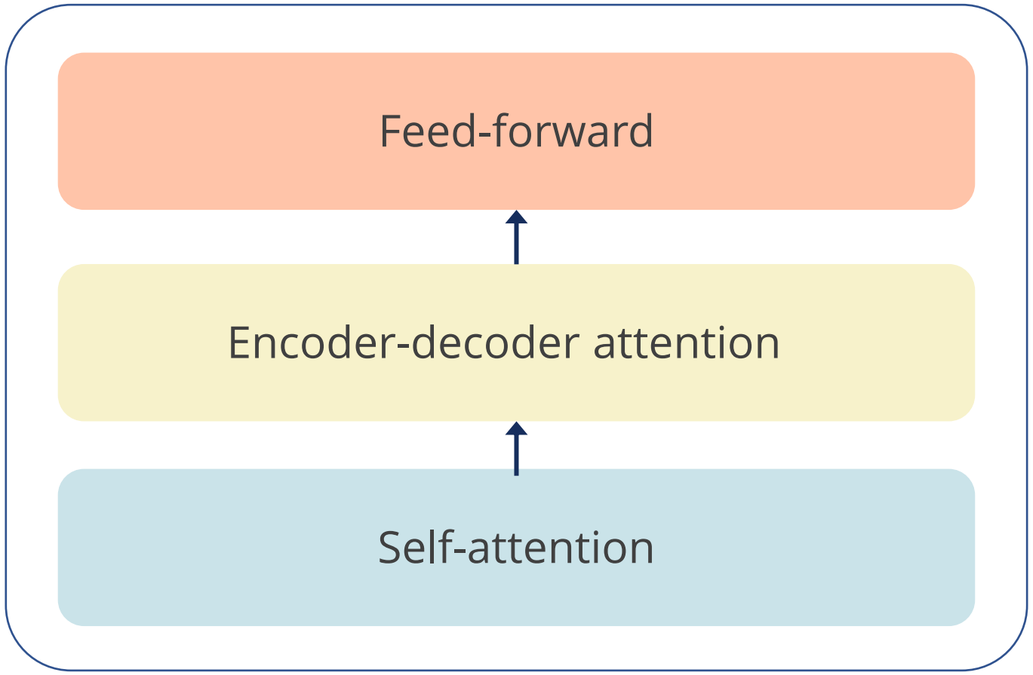


- **Receiving encoder outputs**: The decoder begins its process by receiving the entire sequence of outputs from the encoder. These outputs contain encoded information about every word in the input sequence, providing a comprehensive context that the decoder will use to generate the translation.

- **Output sequence initialization:** The decoder generates the output sequence by receiving a special start token. This token serves as the initial input for the decoder layers.

Transformer Working: Decoder

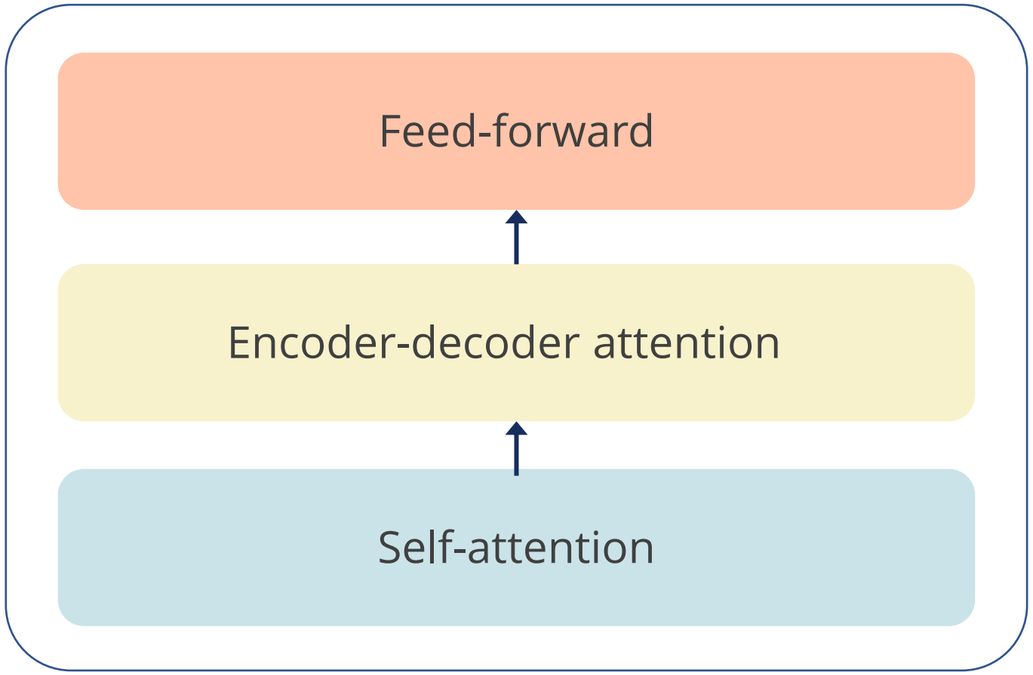
A decoder has three layers:



- **Self-attention mechanism:** Each decoder layer first applies a self-attention mechanism with a restriction. This self-attention only allows each position in the decoder to attend to earlier positions in the output sequence. This ensures that the predictions for each word are dependent only on the known previous words, preserving the auto-regressive property necessary for coherent generation. This mechanism helps the decoder understand the context within the output sequence itself, enhancing the flow and grammatical structure of the translation.

Transformer Working: Decoder

A decoder has three layers:

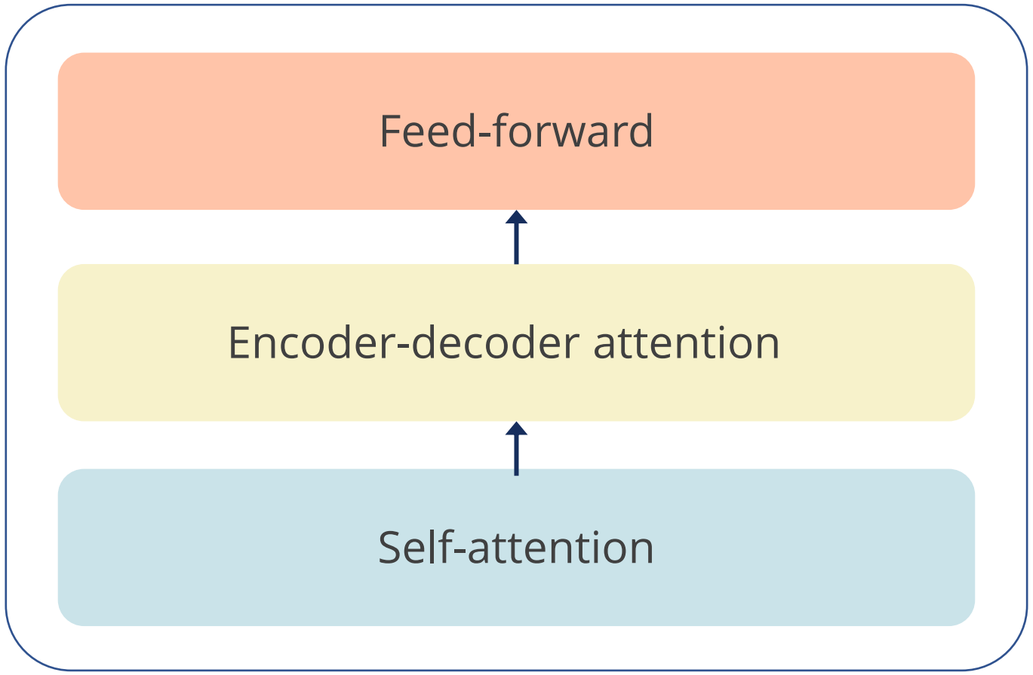


- **Encoder-decoder attention:** After processing through its self-attention layer, each decoder layer uses an encoder-decoder attention mechanism. This crucial layer allows the decoder to focus on relevant parts of the input sequence for each word being generated in the output. By attending to the encoder's outputs, the decoder effectively utilizes the input sequence's contextual information, ensuring that the generated translation is semantically aligned with the input.

- **Feed-forward neural networks:** Like the encoder, each decoder layer includes a position-wise feed-forward neural network. This network processes the output of the attention mechanisms and produces the intermediate representations that generate the next word in the output sequence.

Transformer Working: Decoder

A decoder has three layers:



- **Output generation:** The final layer in the decoder transforms the intermediate representations into logits, which pass through a softmax layer to form a probability distribution over the possible output words. The word with the highest probability is selected as the next word in the output sequence.
- **Repeat the process:** This process repeats for each word in the output sequence until the decoder generates an end-of-sequence token, signaling the completion of the output generation.

Quick Check

Which component of the transformer model processes the input sequence and captures contextual information before passing it to the decoder?

A. Encoder

B. Decoder

C. Attention layer

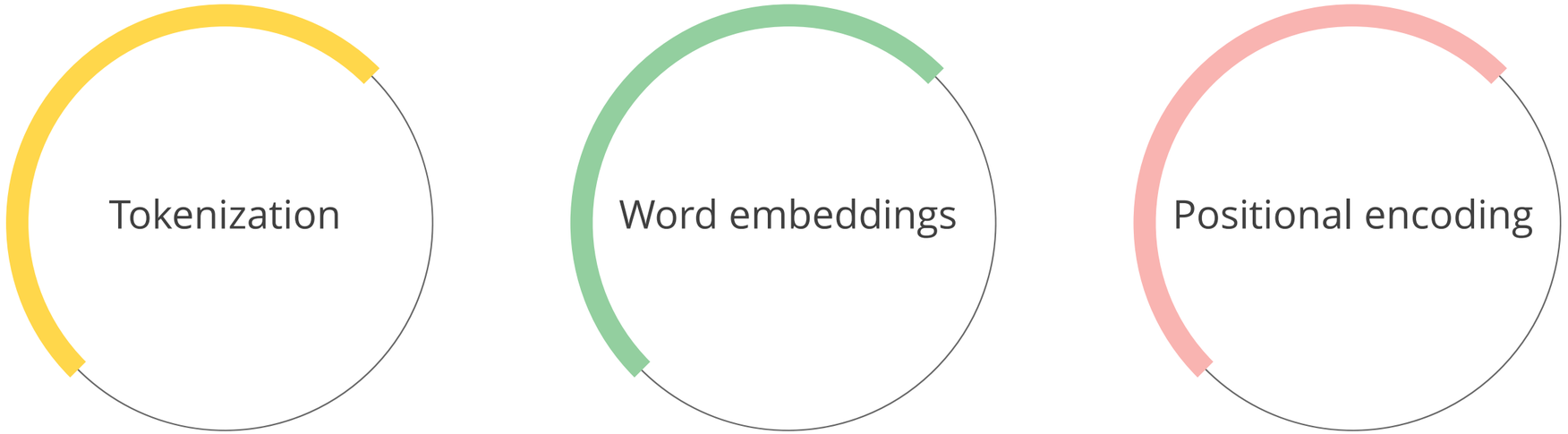
D. Feedforward network

Text Processing in Transformers

Text Processing in Transformers

Transformers process text by converting words into numerical representations that the model can understand.

The three key steps are:



This helps models like BERT and GPT analyze and generate human-like text.

Text Processing in Transformers

Step 1: Tokenization

Breaking text into smaller units (tokens) that can be processed by the transformer

Types of tokenization are:

Word-based tokenization: Splits text into words (for example, "AI is great" → ["AI", "is", "great"]).

Subword tokenization: Splits words into smaller parts (for example, "transformers" → ["trans", "former", "s"]).

Character tokenization: Breaks text into individual characters (for example, "AI" → ["A", "I"]).

Text Processing in Transformers

Step 2: Word embeddings

Converting tokens into numerical vectors (embeddings)

Instead of treating words as just numbers, embeddings capture meaning and relationships between words.

Words with similar meanings have similar embeddings in vector space.

Text Processing in Transformers

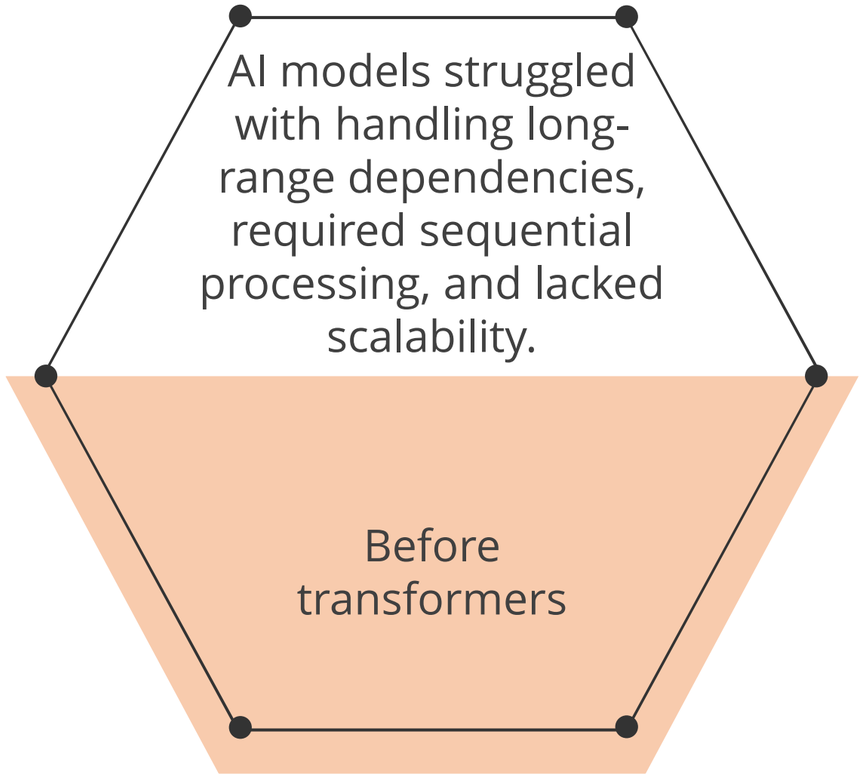
Step 3: Positional encoding

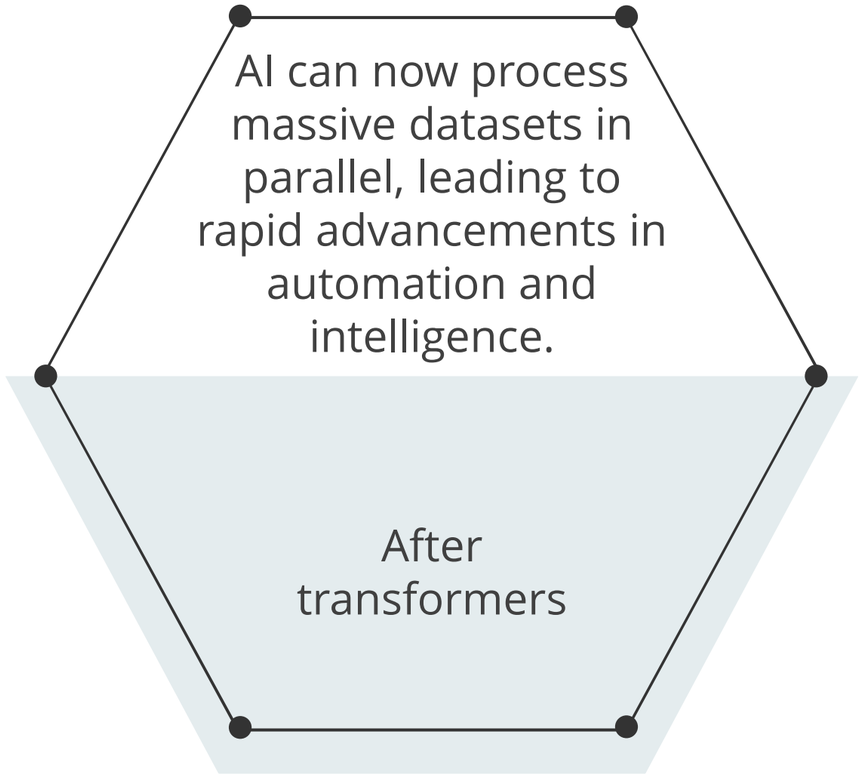
Adding position information to maintain order

- Transformers analyze all words simultaneously rather than processing text sequentially.
- This means the model needs a way to retain word order information.
- Positional encoding assigns unique values to each word’s position in a sentence.

How Transformers Revolutionized AI?

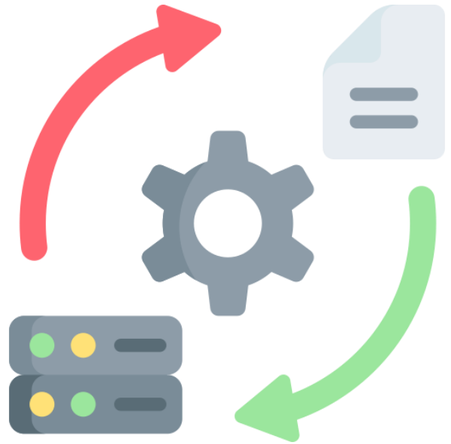
Transformers have reshaped the field of AI, enabling models to understand, generate, and process data more efficiently than ever before.





Transformer Models: Advantage

Transformer models can excel at capturing long-range dependencies through self-attention mechanisms, making them ideal for tasks such as machine translation and document summarization.



Prior models, such as RNNs and LSTMs, had challenges dealing with long-range sentence dependencies, which were resolved by transformer models.

Quick Check

A company is developing an AI-powered chatbot capable of understanding long and complex customer queries instantly. However, their current model struggles with sequential processing, leading to slow responses and difficulty understanding long-range dependencies in conversations.
Which AI model architecture should they adopt to overcome these challenges?

A. Recurrent neural networks (RNNs)

B. Convolutional neural networks (CNNs)

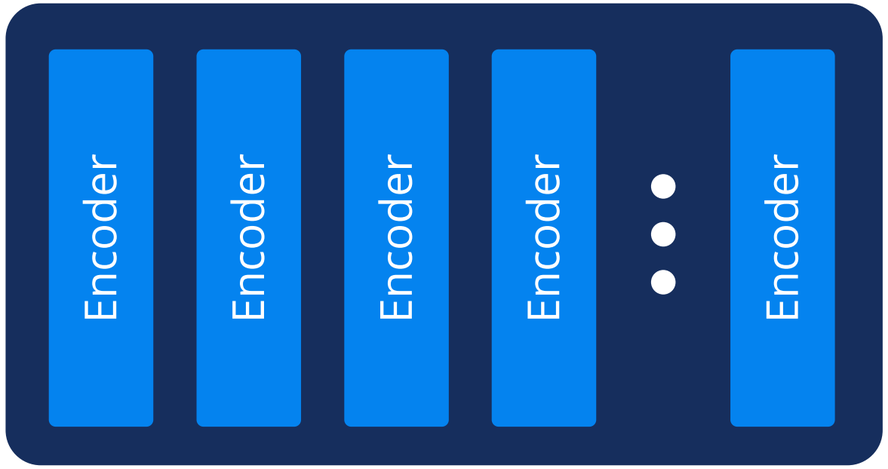
C. Transformers

D. Decision trees

Introduction to BERT Model

BERT Model

BERT stands for Bidirectional Encoder Representations from transformers.



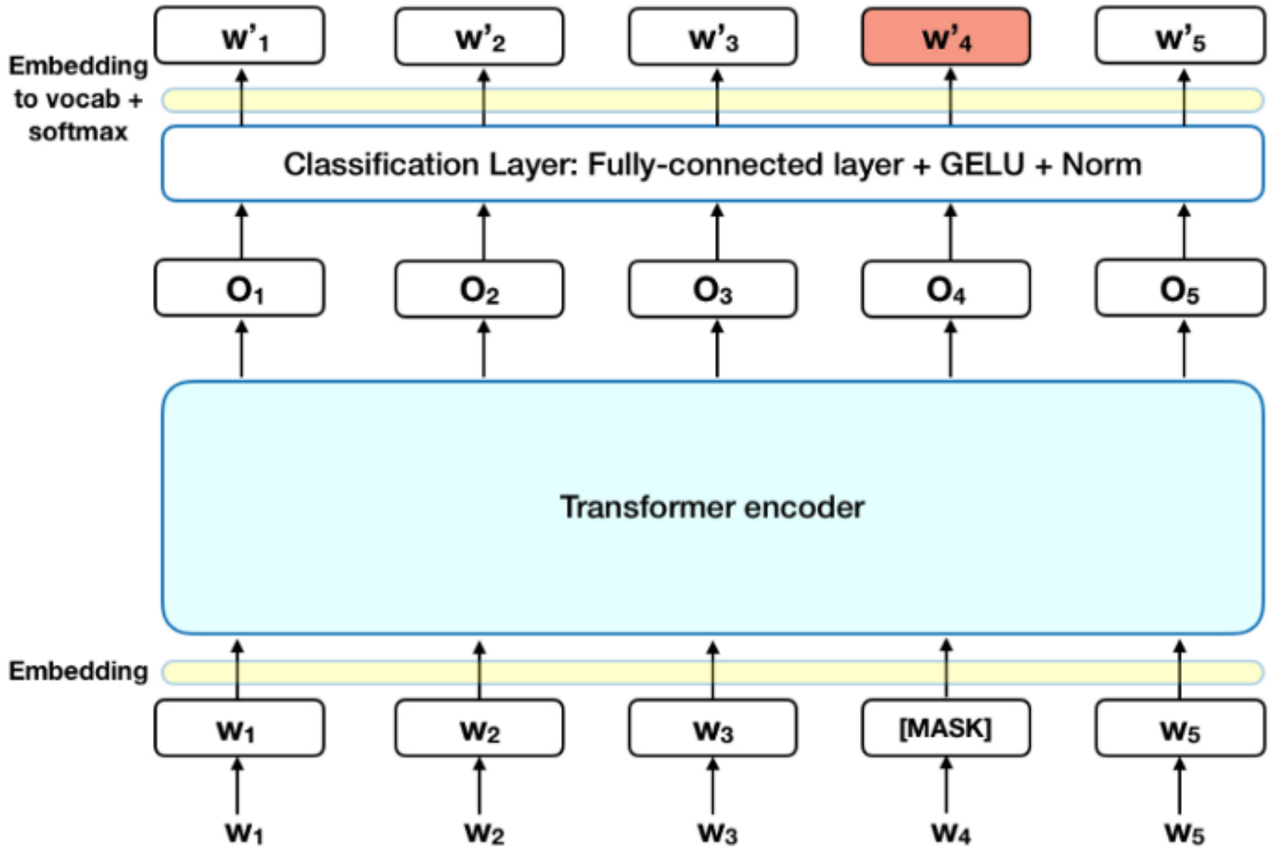
It is a transformer model without decoder modules and only has a trained encoder stack.

The transformer encoder tends to read the entire sequence of words at once.

It learns the context of the given input rather than learning it in sequence. It is called contextual learning.

BERT and Masked Language Modeling

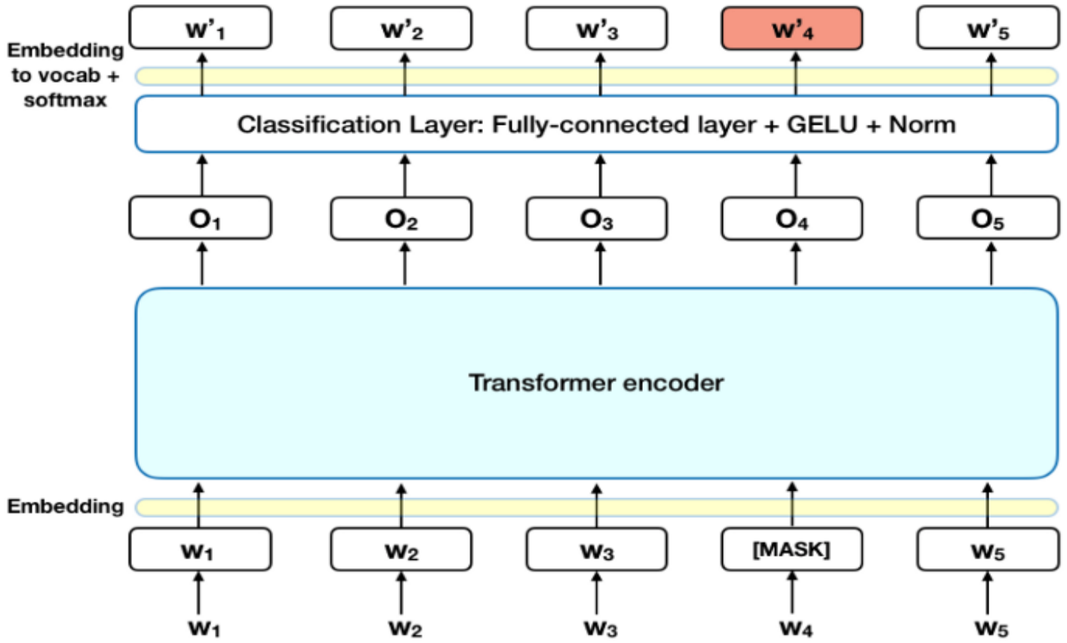
A Masked Language Model (MLM) is a type of language model used in the pre-training of models like BERT (Bidirectional Encoder Representations from transformers). The working of an MLM is as shown below:



Source: Lappin, Shalom. *Deep learning and linguistic representation*. Chapman and Hall/CRC, 2021.

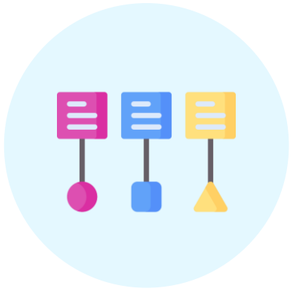
BERT and Masked Language Modeling

In Masked Language Modeling (MLM), 15% of the words are replaced with a MASK token before feeding the sequence of words into BERT.



The model attempts to predict the original values of the masked words by leveraging the contextual information available.

Use Cases for BERT



Text classification: It identifies text characteristics like fraud detection.

Text generation: It generates text, specifically chatbot responses.

Use Cases for BERT

Search engine optimization: It improves search relevance for user queries.

Question-answering (Q&A) system: It helps in accurate Q&A responses.

Quick Check

A search engine wants to improve its ability to understand the full context of a user's query, especially when the meaning of a word depends on the words before and after it. The system needs a model to process entire sentences at once instead of reading word-by-word in sequence.
Which AI model would be the best choice for this task?

A. GPT-3

B. RNN

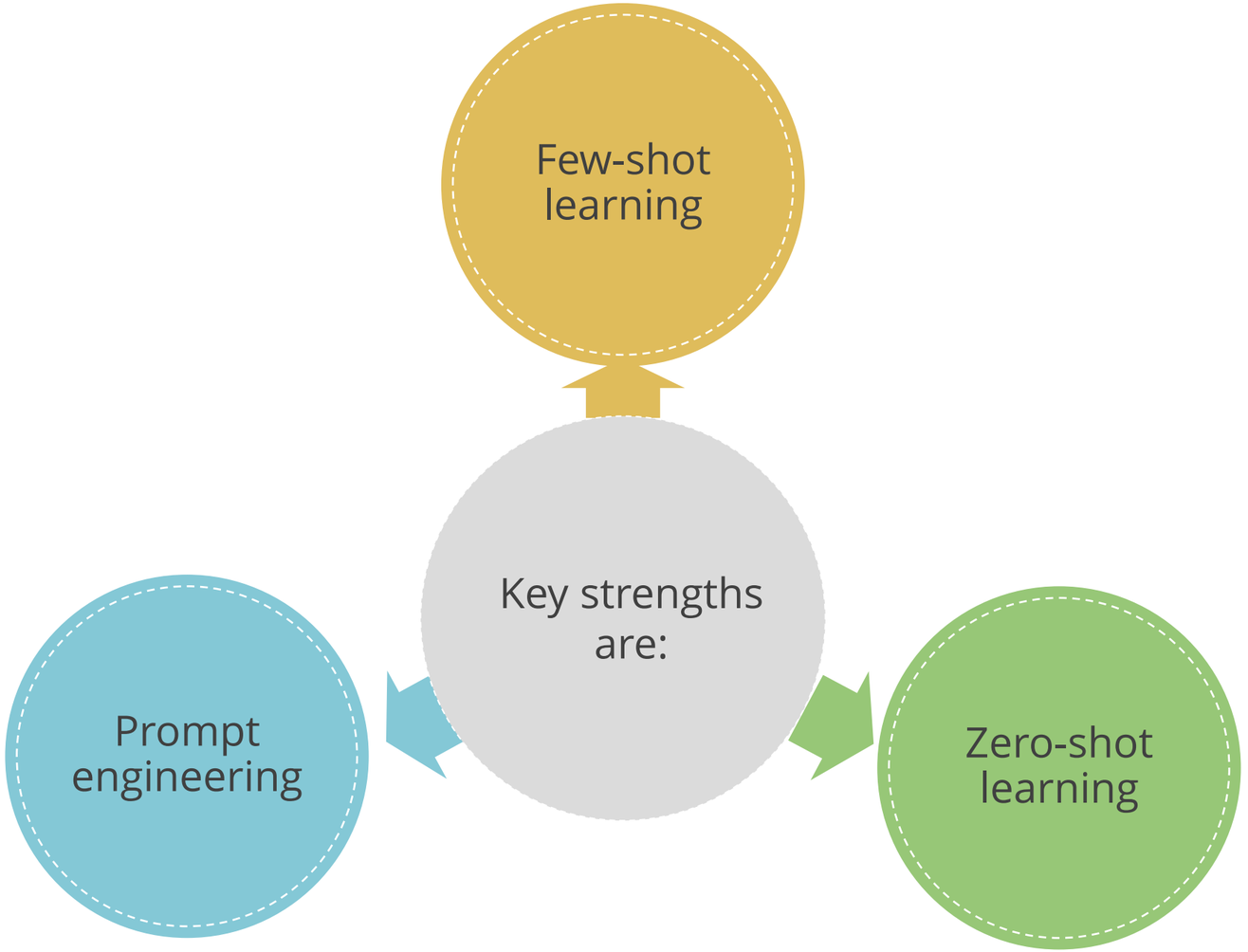
C. BERT

D. CNN

Key Features of Generative Pre-Trained Transformer (GPT) Models and Their Applications

Introduction to GPT Models

Generative pre-trained transformer (GPT) models, developed by OpenAI, use transformer architecture to generate human-like text by understanding the context from user input.



Few-Shot Learning

In this, GPT models learn new tasks from a few examples in the prompt, adapting based on context without retraining.

Example

User prompt:
Translate the following sentences from English to French:
Hello, how are you? → Bonjour, comment ça va?

Application areas

Machine translation, sentiment analysis, text summarization

Zero-Shot Learning

In this, GPT models perform tasks without examples by using pre-trained knowledge to generate responses.

Example

User prompt:
Classify the following sentence as Positive, Negative, or Neutral:
I really enjoy using AI for my projects.
GPT output: Positive

Application areas

Text classification, question answering, named entity recognition

Prompt Engineering

Prompt engineering is designing and optimizing prompts to guide AI models like GPT in generating accurate responses by improving query structure.

Example

User prompt:

Prompt 1: Tell me about Python: Generic response
Prompt 2: Explain Python as if I'm a 5-year-old: Simplified response
Prompt 3: List three advantages of using Python for AI development: Concise and relevant response

Application areas

Chatbots and virtual assistants, content generation, AI-driven tutoring

Quick Check

A company wants to develop an AI-powered customer support chatbot. They want the chatbot to handle a wide range of queries, including questions it has never seen before, without requiring additional training. Which feature of GPT models would allow the chatbot to respond effectively to new questions without prior examples?

A. Few-shot learning

B. Zero-shot learning

C. Fine-tuning

D. Manual rule-based programming

Introduction to Natural Language Processing (NLP)

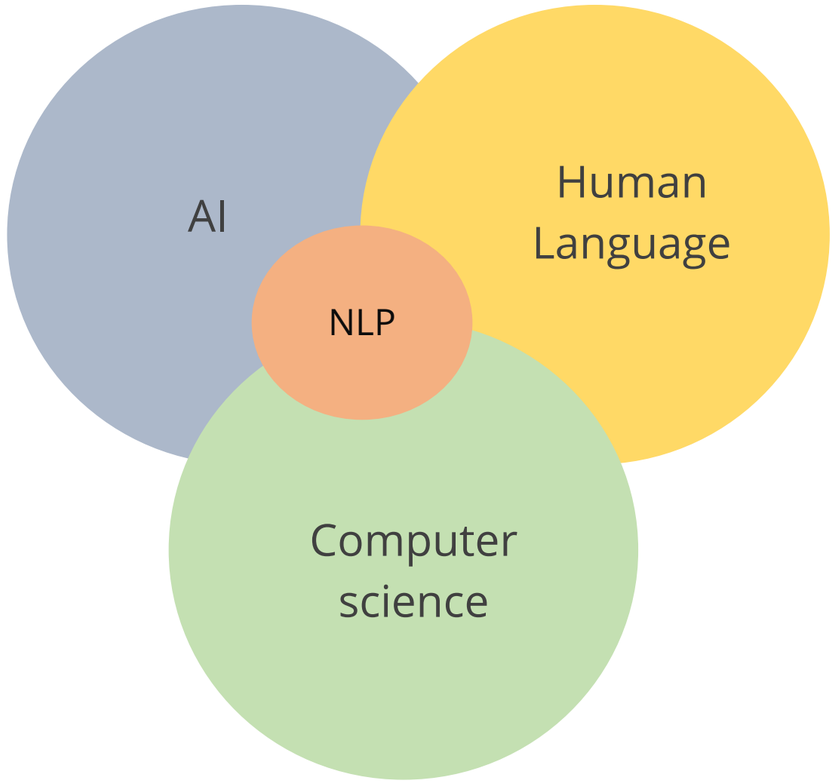
What is NLP?

Natural language processing (NLP) is a branch of AI.

It helps the machine deal with human languages.

It helps machine to understand, interpret, and manipulate human languages.

Most of the natural language processing techniques depend on machine learning to derive meaning from human languages.



NLP

01 NLP is the ability of a computer to analyze, understand, and generate human languages.

02 A language is a system, a set of rules or set of symbols.

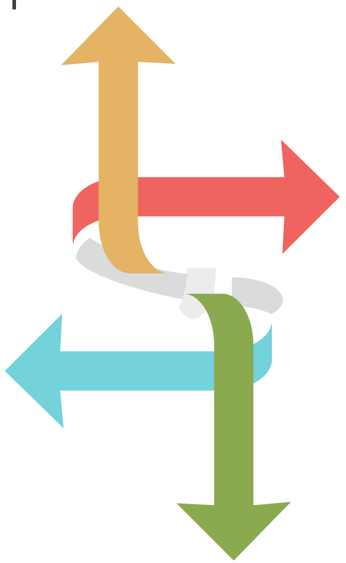
03 Symbols are combined and used for conveying information or broadcasting the information.

04 In NLP, rules of grammar are used for handling symbols.

NLP

NLP enables machines to understand and process human language, transforming unstructured data into meaningful insights and powering AI-driven applications.

It processes and analyzes unstructured text and speech data.



It automates tasks like customer support and document summarization.

It leverages neural networks to capture meaning and context in language.

It powers applications like chatbots, translation, and sentiment analysis.

Categories of NLP

Rule-based NLP

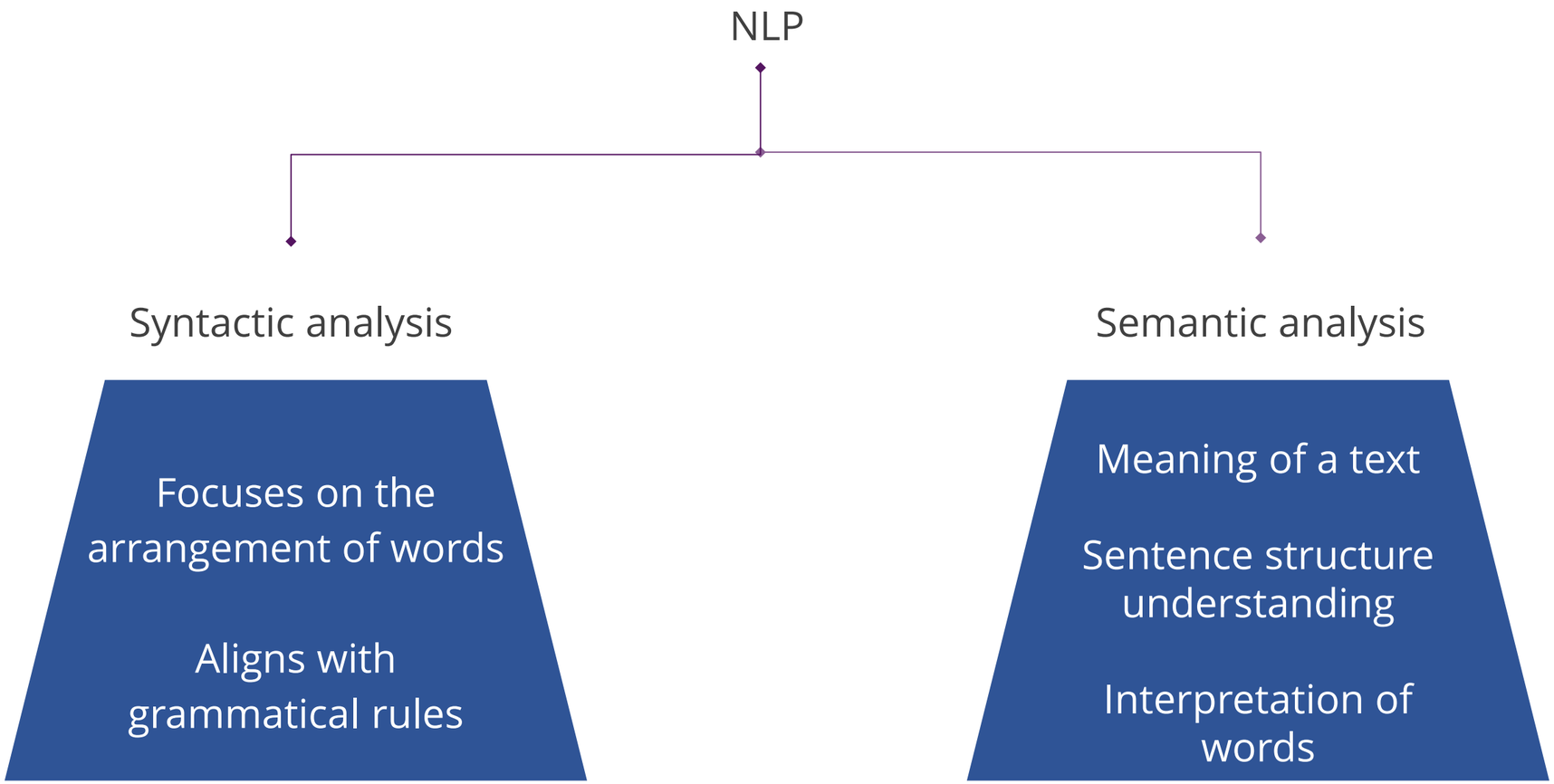
- Designed by creating a set of rules
- Developed by heuristic rules

Statistical revolution

Statistical NLP

- Relies heavily on machine learning
- Applies automatic learning procedure

Techniques Used in NLP



Components of NLP

1

Natural Language Understanding (NLU)

NLU

Natural Language Processing

NLG

Natural Language Generation (NLG)

2

Components of NLP

Natural Language Understanding (NLU)

Taking some sentences and finding out what they mean

Components of NLP

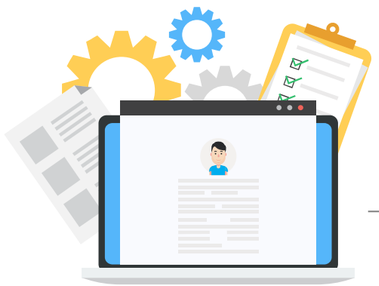
Natural Language Generation (NLG)

Key steps in NLG:

1. Converting a formal representation of information into a natural language expression.
2. Mapping the given input in the natural language with a useful representation
3. Producing output in the natural language from some internal representation
4. Applying different levels of analysis, including morphological, syntactic, semantic, and discourse analysis.

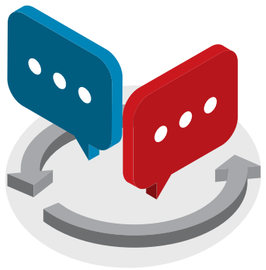
NLP: Uses

Use of NLP in conversational bot in each step:



Users request

Intent identification



Entity extraction

Content management, history processing, session management

Natural Language Understanding

Response generation

Natural Language Generation

Quick Check

A company is developing an AI chatbot for customer support. Initially, they used manually crafted rules to respond to customer queries, but the chatbot struggled with understanding new or unseen questions. To improve accuracy and adaptability, they decided to implement a system that learns from data automatically instead of relying on predefined rules. Which NLP approach should they switch to?

A. Rule-based NLP

B. Statistical NLP

C. Keyword matching

D. Manual scripting

Text Classification in NLP

Text Classification in NLP

It automatically classifies text into predefined topics, helping organize content for news platforms, search engines, and customer support systems.

Common applications include:

- **Sentiment analysis:** Understanding opinions from text
- **Spam detection:** Filtering out unwanted messages
- **Topic categorization:** Organizing text into relevant topics

Sentiment Analysis

It identifies whether text expresses positive, negative, or neutral sentiments and is used in social media monitoring, customer feedback analysis, and brand reputation management.

Common applications include:

- Analyzing customer reviews on Amazon, Yelp, and Google
- Monitoring public opinion on Twitter and social media
- Evaluating user sentiment in market research and surveys

Spam Detection

It uses NLP and machine learning to differentiate between spam and legitimate messages, filtering emails, SMS, and online comments to enhance security and user experience.

Common applications include:

- Email filtering
- SMS fraud detection
- Fake news and comment moderation

Topic Categorization

It automatically classifies text into predefined topics, organizing content for news platforms, search engines, and customer support systems.

Common applications include:

- News categorization
- Customer support ticket classification
- Academic paper classification

Quick Check

A news website wants to automatically sort articles into categories like politics, sports, technology, and entertainment so that users can easily find relevant content.
Which NLP technique would best help the website achieve this?

A. Sentiment analysis

B. Spam detection

C. Topic categorization

D. Named entity recognition

NLP: Applications

NLP: Applications

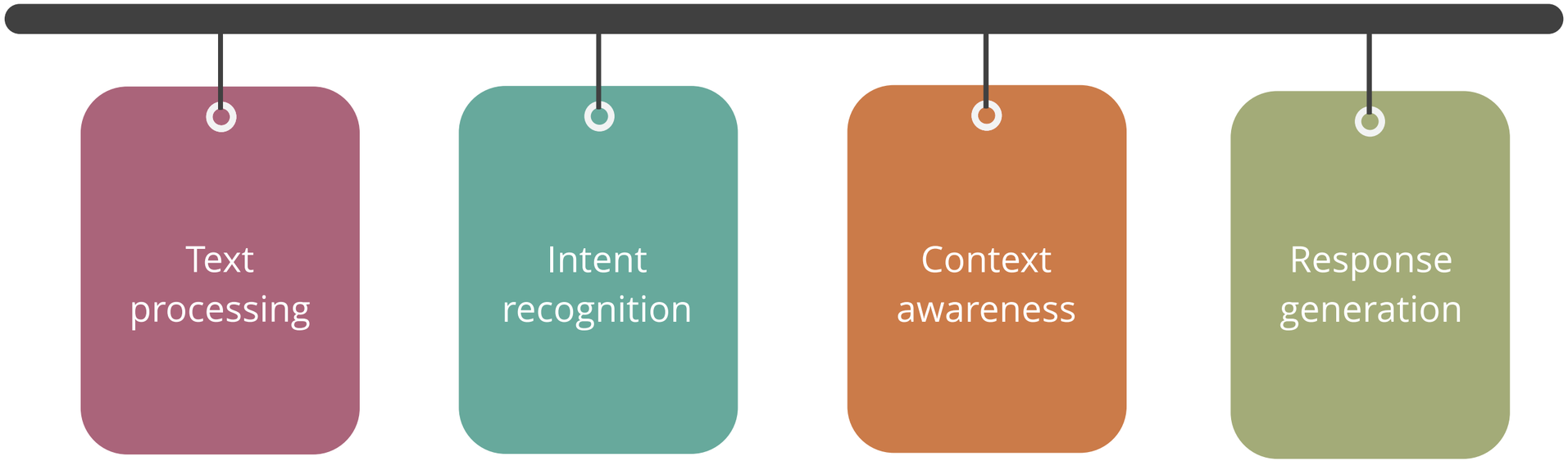
Chatbots and conversational AI are real-world applications of NLP, enabling machines to understand, process, and respond to human language.

They simulate human-like interactions in customer support, virtual assistants, and business automation.

They use NLP techniques like intent recognition, sentiment analysis, and text generation to provide meaningful conversations.

How NLP Powers Chatbots and Conversational AI?

NLP enables chatbots and conversational AI to understand, process, and generate human-like responses by leveraging key components such as:



Example: Customer support chatbots analyze text input and determine if a query is about order tracking, refunds, or technical support.

Quick Check

Which of the following is a real-world application of Natural Language Processing (NLP)?

A. Image classification in autonomous vehicles

B. Chatbots and virtual assistants in customer support

C. Predicting stock market trends using numerical analysis

D. Performing mathematical calculations

NLP in Customer Service

Case Study: Back of America

BANK OF AMERICA

Background:

Bank of America, one of the largest financial institutions in the U.S., wanted to improve customer engagement and support by integrating AI into their service channels. Traditional customer support was becoming costly, time-consuming, and unable to scale efficiently to meet growing customer demands.

To enhance customer experience, Bank of America introduced **Erica,** an AI-powered virtual assistant that uses natural language processing (NLP) and machine learning to assist customers with banking needs.

Challenges faced

- High call volume and wait times
- Inconsistent responses
- Scalability issues
- Demand for 24/7 support

Source: https://newsroom.bankofamerica.com/content/newsroom/press-releases/2023/07/bofa-s-erica-surpasses-1-5-billion-client-interactions--totaling.html?utm_source=chatgpt.com

Case Study: Back of America

Solution: Implementing Erica, an AI-powered NLP chatbot

Conversational AI with NLP: Processes text and voice queries to assist users

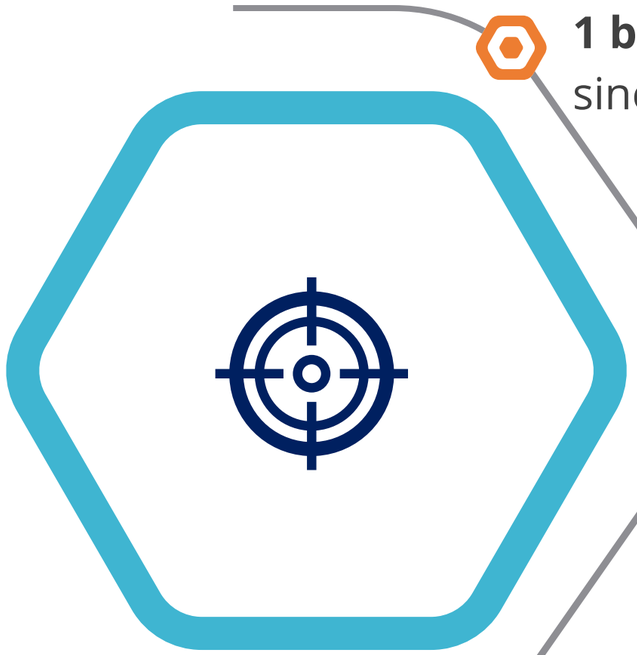
Financial guidance: Provides spending insights, budgeting advice, and credit score updates

Automated banking services: Enables users to check balances, review transactions, schedule payments, and lock/unlock debit cards

Smart search and query handling: Uses machine learning to improve response accuracy over time

Case Study: Back of America

Key results and business impact:



1 billion + interactions: Erica has processed over 1 billion customer requests since launch.

90% query resolution rate: Erica has handled most inquiries without human intervention.

Enhanced customer satisfaction: Erica provides 24/7 instant, accurate, and personalized AI assistance.

Applications of NLP in Business

Applications of NLP in Business

NLP drives innovation across industries by automating language-based tasks and improving customer engagement.

Customer support

Enhances customer experience through chatbots and virtual assistants

Speech processing

Enables transcription, voice recognition, and virtual assistants

Text analysis

Extracts insights using sentiment analysis, summarization, and spam filtering

Language translation

Breaks language barriers with tools like real-time translators

Quick Check

A global e-commerce company receives thousands of customer reviews daily. They want to analyze feedback to understand customer sentiment, identify common complaints, and filter out spam or irrelevant reviews. Which NLP application would best help them achieve this?

A. Speech processing

B. Customer support

C. Text analysis

D. Language translation

Key Takeaways

- Transformers revolutionized AI by enabling parallel processing, allowing models like BERT and GPT to analyze text efficiently without relying on sequential data processing. This has led to major advancements in natural language understanding and generation.
- Transformers process text using tokenization, embeddings, and positional encoding, helping AI models understand word relationships and context, which improves translation, summarization, and conversational AI.
- Natural language processing (NLP) powers AI applications, such as sentiment analysis, spam detection, and text classification, enabling businesses to extract insights from large volumes of text.
- Chatbots and AI-driven conversational agents enhance customer service, automating responses and improving user engagement. Companies leverage NLP models to provide personalized, context-aware interactions.

Q&A