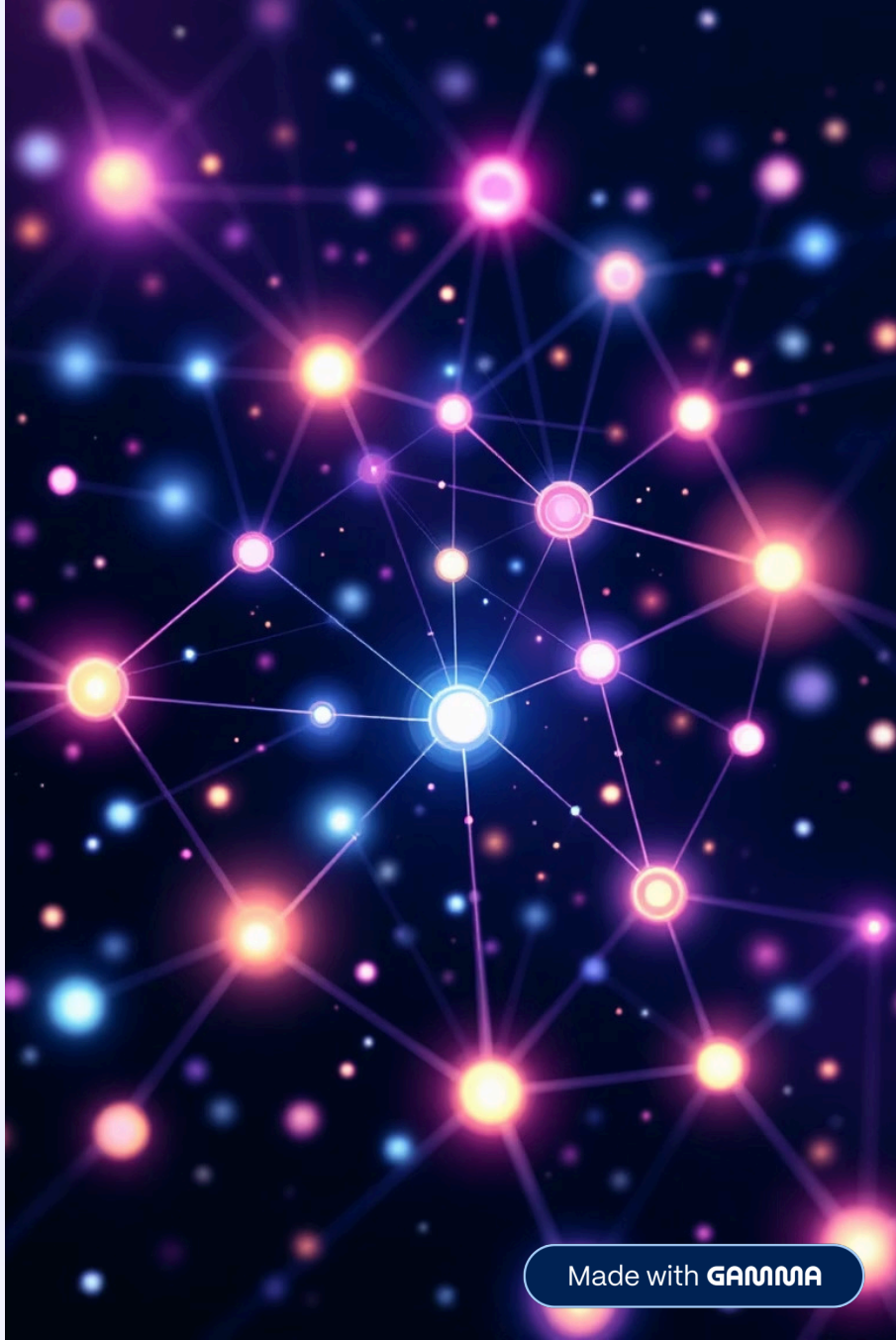
Transformers: The Brains Behind Modern AI

Understanding Self-Attention and Context in Language Models

Based on "Attention is All You Need" — Vaswani et al. (2017)



What Is a Transformer?

A **Transformer** is a revolutionary neural network architecture designed to understand language by reading **entire sentences at once** instead of processing them word-by-word like traditional models.

💡 **Key Idea:** Each word **pays attention** to every other word in a sentence — capturing context, meaning, and complex relationships across the entire input.

Powers Modern AI Tools

- ChatGPT and the GPT series
- Google Search with BERT
- Grammarly's language suggestions
- Google Translate's accuracy

Why Transformers Were Needed

Older models like RNNs and LSTMs faced significant limitations that held back progress in natural language processing.

Problem	Why It Mattered
Sequential processing	Cannot process in parallel → extremely slow training times
Poor long-range memory	Tendency to forget earlier words in lengthy texts
Complex training	Harder to optimise and prone to vanishing gradients
Limited context	Struggle to connect distant words meaningfully

🚀 Transformers solved all of these challenges elegantly.

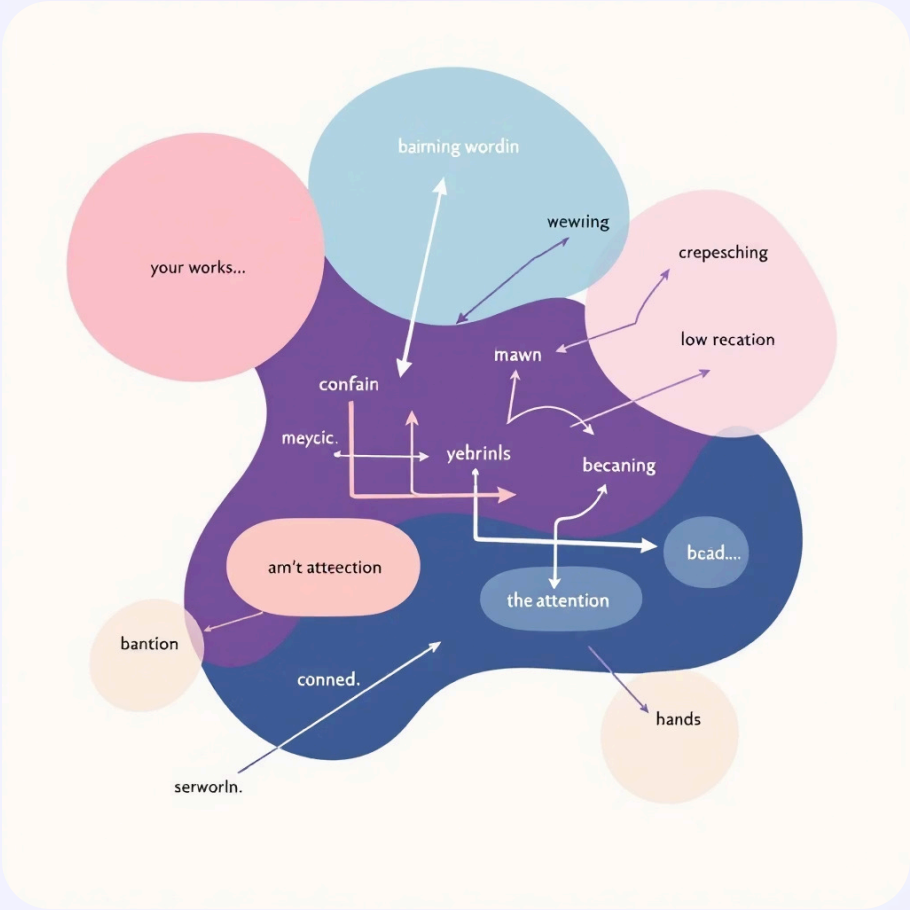
The Core Idea: Self-Attention

The **Self-Attention Mechanism** allows each word to examine every other word in the sentence to determine what's contextually important.

> **Example:** In "The **bank** of the river was calm," the word "bank" pays attention to "river" (not "money") to derive the correct meaning.

⚙️ **Every word has three representations:**

- **Query (Q)** → what it's looking for
- **Key (K)** → how relevant it is to others
- **Value (V)** → the actual information it carries



📊 The output is a **weighted sum of values** based on calculated attention scores between all word pairs.

Transformer Architecture Overview

🧩 The Transformer consists of two major components working in harmony:

Encoder

Function: Reads and understands input text

Example Models: BERT, RoBERTa

Builds context-aware representations of the input sequence

Decoder

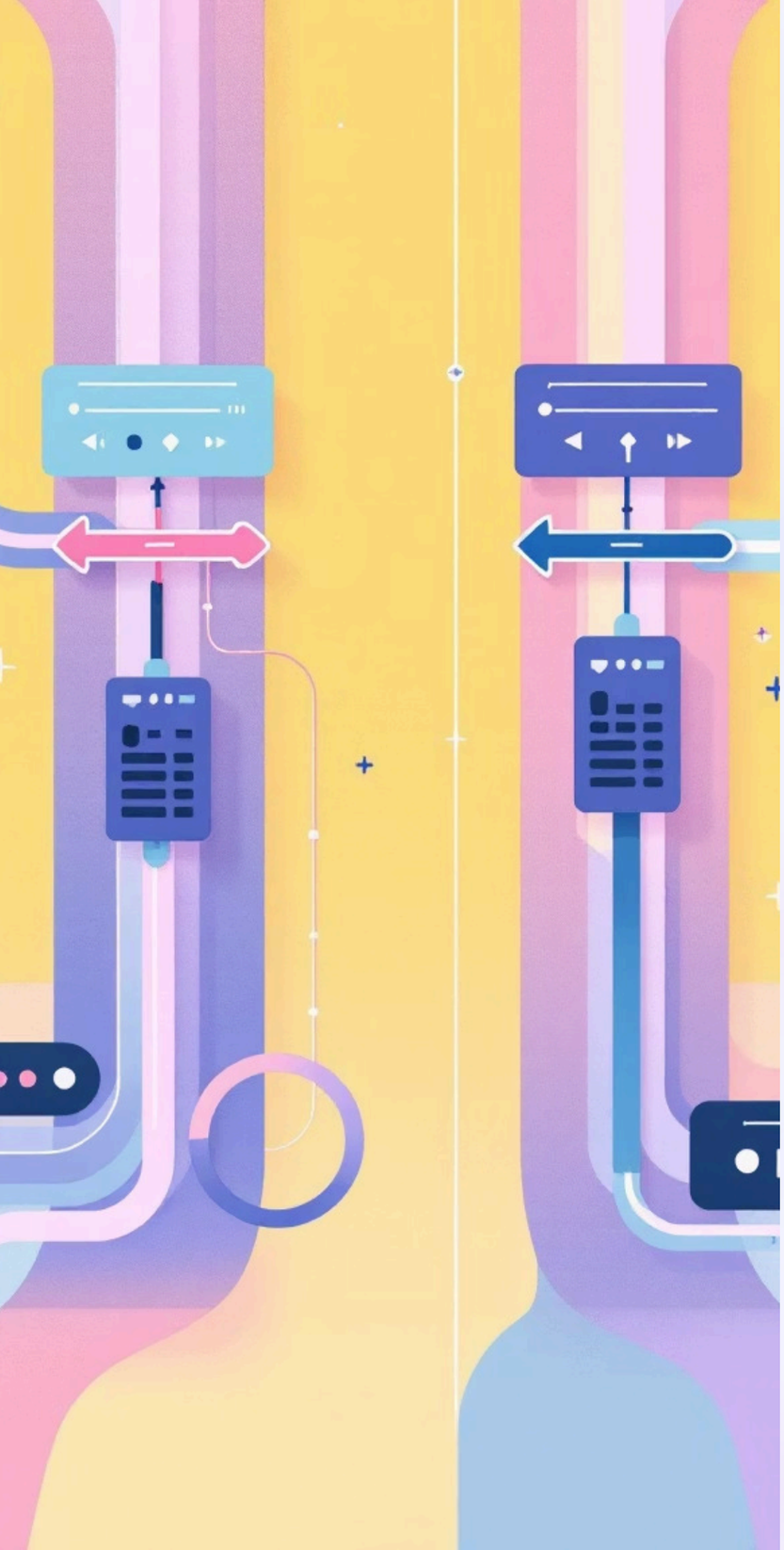
Function: Generates text based on understanding

Example Models: GPT, T5

Predicts output sequences auto-regressively

Each component is built using:

1. Multi-Head Self-Attention
2. Feed-Forward Neural Network
3. Layer Normalisation & Residual Connections



Encoder vs Decoder: The Dynamic Duo

Encoder

Input Processing:

- Takes input sentence or text
- Uses self-attention to analyse relationships
- Builds context-aware representations
- All tokens processed simultaneously

Decoder

Output Generation:

- Takes encoder output plus previous tokens
- Predicts the next word in sequence
- Uses masked self-attention
- Generates text auto-regressively

📘 **Working together:** Used in translation, summarisation, question answering, chatbots, and many more applications.

Parallelisation: The Big Advantage

Unlike RNNs, Transformers don't process words sequentially. They handle **all tokens simultaneously**, which unlocks massive advantages:

Faster training on GPUs

Parallel processing leverages modern hardware efficiently

Better scalability to long texts

No bottleneck from sequential dependencies

Easier optimisation

Allows deeper architectures without gradient issues

⚡ This parallelisation is what made large models like GPT and BERT possible.

Advantages of Transformers

Transformers brought revolutionary improvements across multiple dimensions:

Parallel Processing

Dramatically faster training and inference compared to sequential models

Long-Range Attention

Captures distant word relationships effectively across entire sequences

Scalability

Can be scaled up to billions of parameters whilst maintaining efficiency

State-of-the-Art

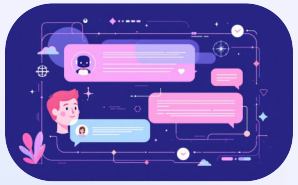
Consistently outperforms previous models across NLP benchmarks

Transfer Learning

Pretrain once on large datasets, then fine-tune for many specific tasks

Impact & Applications

Transformers are the foundation of **all modern NLP** and have even revolutionised **vision models (ViTs)** today.

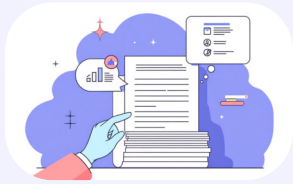


Text Generation

ChatGPT, Bard, Claude

Understanding & Classification

BERT, RoBERTa



Summarisation

T5, BART

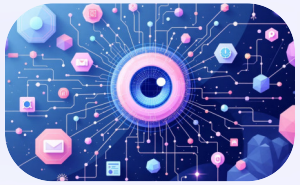


Image Processing

Vision Transformers (ViT)

💡 **"Attention truly changed everything."** — From language to vision, transformers have reshaped the entire AI landscape.

Summary

Holistic Processing

Transformers process entire sentences at once, not word-by-word

Self-Attention Magic

Use self-attention mechanism to understand complex word relationships

Speed & Context

Solve the speed and context problems that plagued RNNs and LSTMs

Modern AI Foundation

Power nearly every modern AI application we use today

🌟 Introduced in 2017 with "Attention is All You Need," transformers are now the foundation of today's AI revolution — from language models to computer vision and beyond.