Exploring AI with Transformers: The Engine Behind Generative AI and LLMs

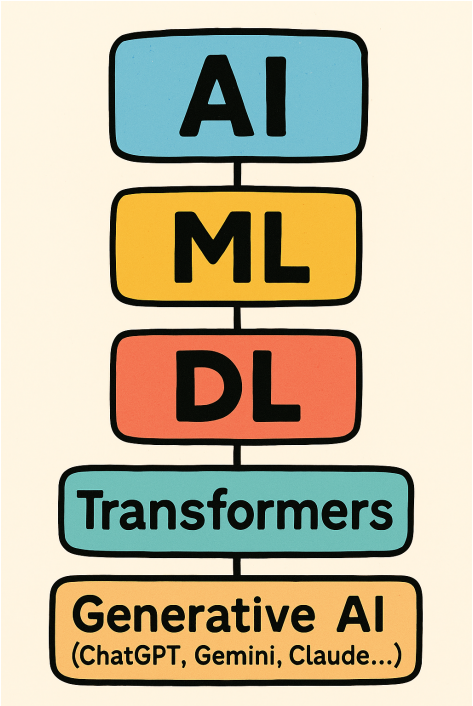
Premanand S

Assistant Professor
School of Electronics Engineering
Vellore Institute of Technology
Chennai Campus

premanand.s@vit.ac.in

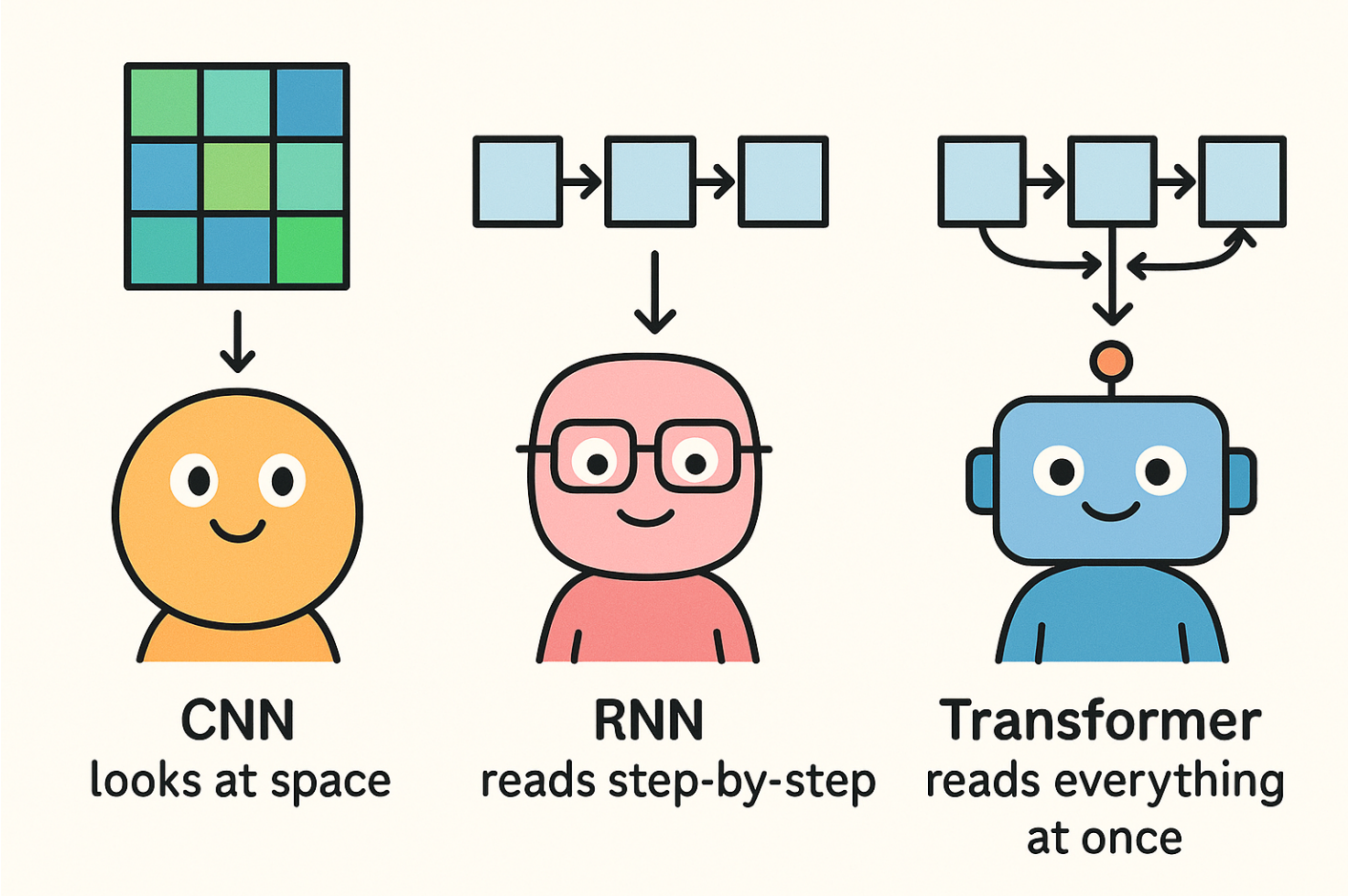
December 16, 2025

AI vs ML vs DL vs Transformers vs Generative AI



Takeaway: *Transformers are the bridge between Deep Learning and modern Generative AI.*

CNN Vs RNN Vs Transformer



Takeaway: *CNNs see space, RNNs follow time, Transformers see everything at once.*

The "Problem" with RNNs & LSTMs

Before Transformers, almost all NLP models used RNN-based architectures.

Their limitations:

- **Sequential bottleneck** — process inputs word-by-word → slow, cannot parallelize.
- **Long-term memory loss** — even LSTMs struggle with long dependencies (vanishing gradients).
- **Context understanding drops** as distance between words increases.
- **Poor scalability** for very large datasets.

Takeaway: *RNNs remember the recent past, but forget the important past.*

The Breakthrough: *Attention is All You Need* (2017)

Transformers (Vaswani et al., 2017): Removed recurrence entirely and introduced **pure attention-based sequence modeling**.

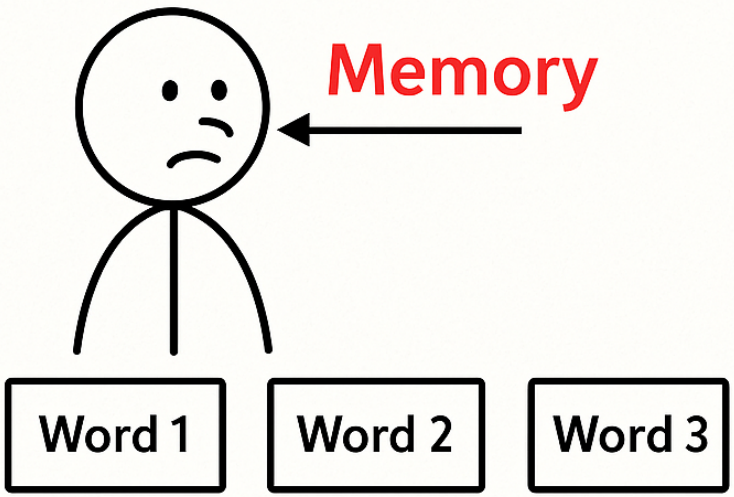
What changed?

- No step-by-step processing $\rightarrow$ **full parallelism**.
- Model can **attend to any token instantly**.
- **Captures long-range relationships** easily.
- **Trains massively faster** on GPUs/TPUs.
- Enabled scaling to **hundreds of billions of parameters**.

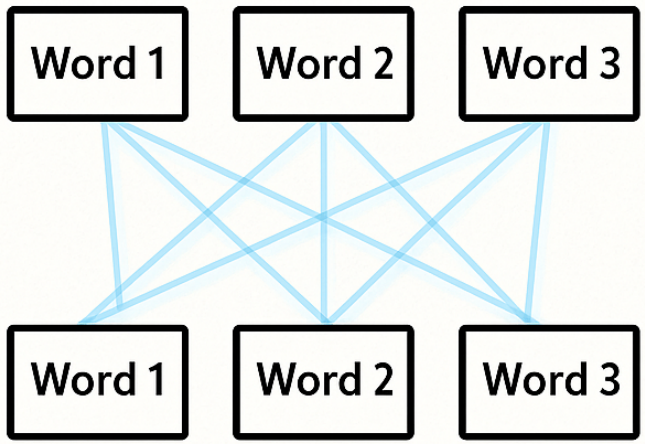
Takeaway: *The Transformer wasn't just a new architecture — it was a revolution in how AI learns patterns.*

Paper link: `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`

Recurrent

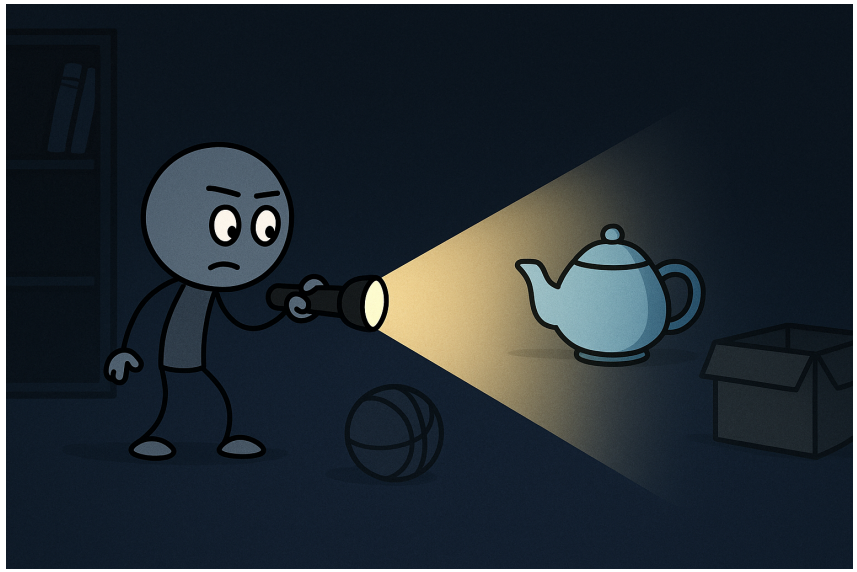


Attention



Self-Attention = A Spotlight in a Dark Room

- For every word, the model decides: *"Which other words should I focus on?"*
- Strong light $\rightarrow$ **high attention**
- Weak light $\rightarrow$ **low attention**
- The model shines the spotlight **everywhere at once** *(parallelism)*



Example sentence: *"He ate the food because he was hungry."*

Why RNNs fail here: They cannot maintain such long-distance relationships reliably.

Real-World Examples Powered by Transformers

ChatGPT

- Uses a **decoder-only Transformer (GPT)** architecture.
- Generates human-like text, code, explanations.

Google Gemini

- A multimodal Transformer $\rightarrow$ processes text, images, audio, video.

GitHub Copilot

- Trained on billions of code lines $\rightarrow$ predicts functions, auto-completes code.

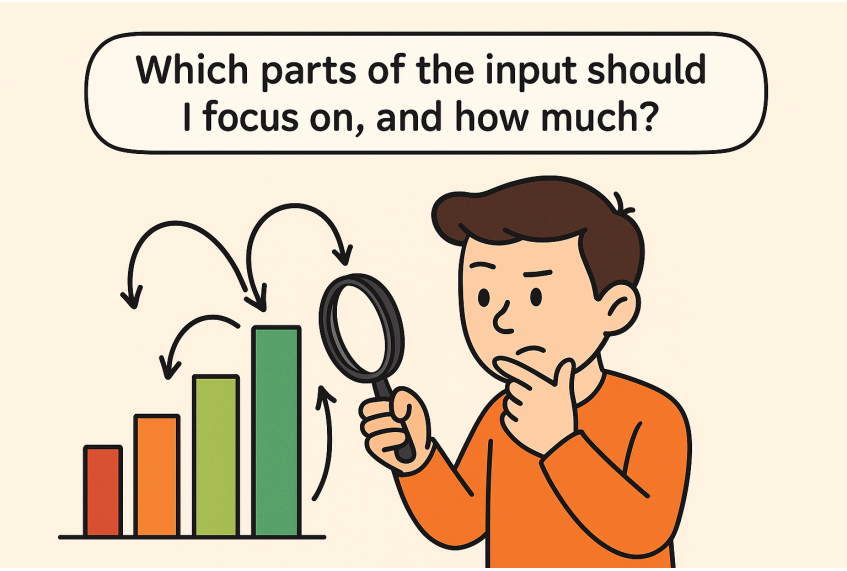
Vision Transformer (ViT)

- Treats image patches like "tokens".
- Captures global structure better than CNNs in many tasks.

Transformer Architecture — The Big Picture

Transformer answers one question:

- Transformers do **not read sequentially**.
- They process the **entire sequence at once**.
- Understanding comes from **attention**, not memory.



Key idea:
Transformers replace recurrence with attention.

Types of Transformers — Big Picture

- Transformers are **not a single model**.
- They are a **family of architectures**.
- All share the same core idea: **Attention**.

Different Transformer designs solve different problems.

Types of Transformers — Core Models

Type	Architecture	Key Strength
Encoder-only (BERT)	Encoder blocks (Bidirectional)	Language understanding, classification
Decoder-only (GPT)	Decoder blocks (Masked attention)	Text generation, LLMs
Encoder–Decoder (T5, BART)	Encoder + Decoder (Cross-attention)	Translation, summarization
Vision Transformer (ViT)	Image patches as tokens	Global image context

Takeaway: *Same attention idea, different architectures.*

Types of Transformers — Advanced Variants

Type	Focus	Key Strength
Multimodal (CLIP, Gemini)	Text + Image + Audio	Cross-modal reasoning
Efficient / Long-Sequence (Longformer)	Long inputs	Reduced attention cost
Domain-Specific (BioBERT)	Specialized data	Domain expertise
Hybrid Models (CNN + Transformer)	Local + Global features	Best of CNN + Attention

Takeaway: *Transformers adapt to data, scale, and domain.*

Attention — The Heart of the Transformer

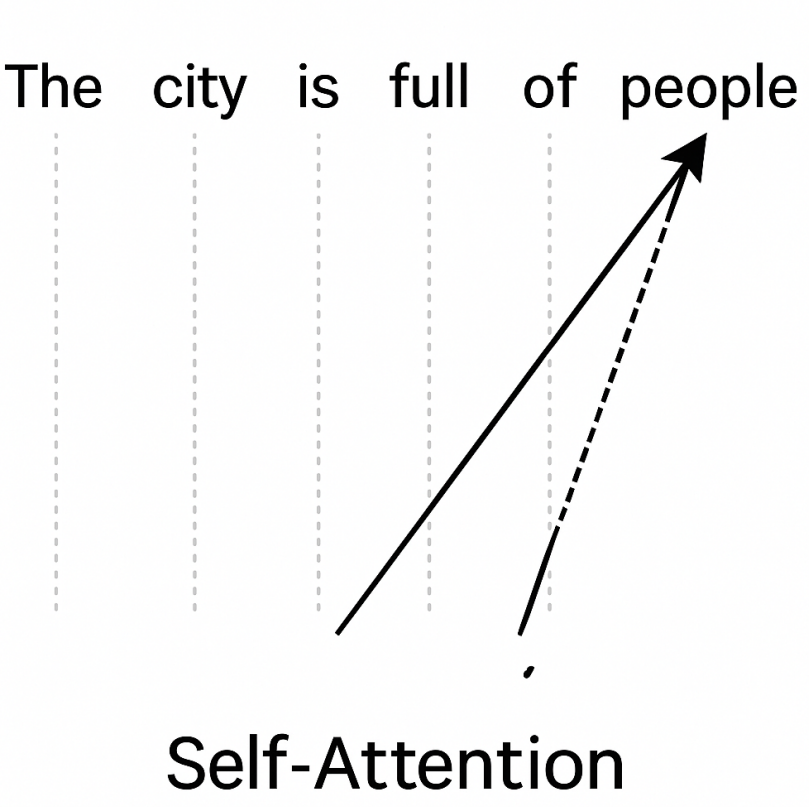
What is Attention?

- Every word looks at **all other words** in the sentence.
- The model decides:
  - Who is important?
  - Who is related?
  - Who can be ignored?

Intuition:

- Not all words are equally important.
- Meaning comes from **relationships**, not position.

Takeaway: *Attention lets the model understand context globally.*



Why Attention is So Powerful

Attention enables:

- Long-range dependency modeling.
- Parallel processing of sequences.
- Context-aware understanding.

Example:

"The patient was treated with antibiotics because he had an infection."

- Attention links

Takeaway: *Transformers understand meaning, not just order.*

Different Types of Attention — Big Picture

- Attention decides **what to focus on and how much**.
- Transformers use **multiple attention mechanisms**.

Different attention types solve different problems.

Different Types of Attention — One-Glance View

Attention Type	**What It Does**	**Uniqueness / Use**
Self-Attention	Tokens attend to other tokens in the same sequence	Captures global context; replaces recurrence
Masked Self-Attention	Tokens attend only to past tokens	Enables next-token prediction (LLMs)
Cross-Attention	One sequence attends to another sequence	Links input to output (translation, summarization)
Multi-Head Attention	Multiple attention heads run in parallel	Learns different relationships simultaneously
Global Attention	Every token attends to every token	High accuracy, high computation cost
Local / Sparse Attention	Tokens attend to selected nearby tokens	Efficient for long sequences
Soft Attention	Uses continuous attention weights	Differentiable; used in Transformers

Takeaway: *Attention is not one mechanism, but a family of focusing strategies.*

Positional Encoding — Order Still Matters

Problem:

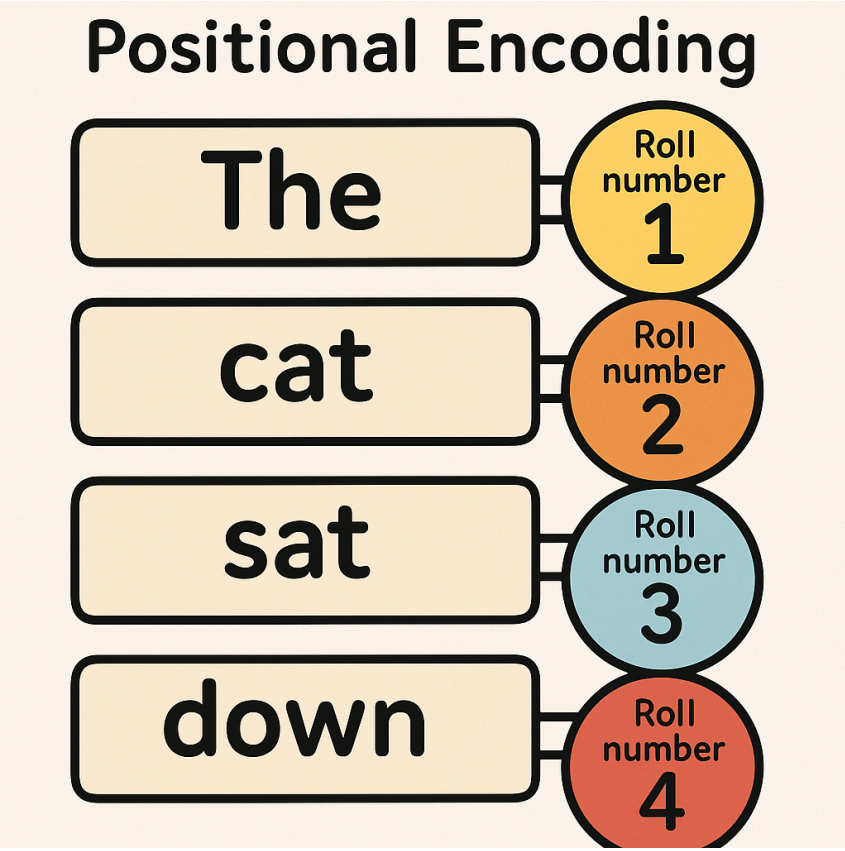
- Transformers process words in parallel.
- Parallel processing loses word order.

Solution: Positional Encoding

- Injects position information into word embeddings.
- Helps the model understand:
  - which word comes first,
  - which comes next,
  - overall sentence structure.

Analogy:

- Same words, different order $\rightarrow$ different meaning.



Why Transformers Matter Beyond ChatGPT

- Transformers are **not limited to text**.
- They are a **general-purpose sequence modeling framework**.

Any data that has:

- Order
- Context
- Long-range dependencies

can benefit from Transformers.

Key idea: *If your data is sequential, Transformers are relevant.*

Applications in NLP

- Text classification (spam, sentiment, fake news)
- Question answering systems
- Machine translation
- Summarization of documents
- Conversational AI and chatbots

Used Models:

- BERT, RoBERTa $\rightarrow$ understanding tasks
- GPT, LLaMA $\rightarrow$ generation tasks

Impact: *Transformers replaced classical NLP pipelines entirely.*

Applications in Vision & Multimodal AI

- Vision Transformer (ViT) for image classification
- Object detection and segmentation
- Medical imaging (X-ray, MRI, CT analysis)
- Image captioning
- Text-to-image generation (DALL·E, Stable Diffusion)

Key idea:

- Images $\rightarrow$ split into patches $\rightarrow$ treated like tokens

Takeaway: *Transformers unified vision and language.*

Applications in Time-Series & Signals

- ECG, EEG signal classification
- Speech recognition
- Financial forecasting
- Sensor data analysis (IoT)
- Anomaly detection

Why Transformers here?

- Capture long-range dependencies
- Better than RNN/LSTM for long signals

Takeaway: *Transformers are powerful for biomedical and industrial signals.*

Transformers in Generative AI Systems

- Large Language Models (ChatGPT, Gemini, Claude)
- Code generation and software assistants
- Research paper summarization
- Educational tutoring systems

System-level concept:

- Retrieval-Augmented Generation (RAG)
- Tool-using agents
- Multimodal assistants

Insight: *Modern AI systems are built around Transformers, not standalone models.*

Open Research Directions with Transformers

- Efficient Transformers (Linformer, Performer, Longformer)
- Hybrid models (CNN + Transformer)
- Lightweight Transformers for edge devices
- Explainability and interpretability
- Robustness and bias mitigation

Research trend: *Making Transformers smaller, faster, and more trustworthy.*

Research Ideas You Can Start Immediately

- Replace RNN/LSTM with Transformer in existing projects
- Apply Transformers to local datasets (medical, education, agriculture)
- Combine domain knowledge with attention mechanisms
- Study attention maps for interpretability
- Benchmark CNN vs Transformer vs Hybrid models

Message: *Transformers are a tool — innovation comes from application.*

Generative AI — Myths vs Reality

Common Myths

- LLMs understand language like humans
- Bigger models are always better
- Attention means intelligence

Reality

- LLMs predict probabilities, not meaning
- Smaller, efficient models can outperform
- Attention measures relevance, not reasoning

Takeaway: *Understanding limitations is as important as knowing capabilities.*

What Transformers Cannot Do

- Do not possess true reasoning or consciousness
- Do not understand the real world
- Can hallucinate confident but wrong answers
- Inherit bias from training data
- Are sensitive to prompt wording

Key Message: *Transformers are powerful pattern learners — not thinkers.*

Where Attention Can Fail

- Attention weights are not always explanations
- High attention does not imply causal importance
- Attention maps can be misleading

Research Insight:

- Interpretability of attention is still an open problem

Takeaway: *Attention helps models decide — not explain.*

Why Blind Trust Is Dangerous

- LLMs can sound confident even when wrong
- Hallucinations appear fluent and convincing
- No built-in truth verification

Important Reminder: *Fluency is not correctness.*

If You Remember Only 5 Things

- Attention replaced recurrence
- Transformers scale extremely well
- Meaning comes from relationships
- Same model supports generation and classification
- Understanding matters more than training

Transformers are tools — wisdom lies in their use.

Final Thought

Transformers changed AI,

but thoughtful humans shape its impact.

Thank You!

Stay Connected

Premanand S

Email: premanand.s@vit.ac.in
Phone: +91-7358679961

LinkedIn: linkedin.com/in/premsanand
Instagram: instagram.com/premsanand
WhatsApp Channel: anandsDataX

Google Scholar: Google Scholar Profile
GitHub: github.com/anandprems