


Evolution from RNNs to Transformers: Unleashing the Power of Attention!

- **RNNs: The Old Guard**

- Previous generations relied on Recurrent Neural Networks (RNNs) for generative tasks.



- While potent in their time, RNNs faced limitations due to compute and memory constraints. 

- Example: Next-word prediction task with RNNs showed poor performance with limited context. 😞❌


- Scaling RNNs to consider more preceding words required significant resource allocation.





- Despite scaling, RNNs struggled to grasp the full context, leading to flawed predictions.



- **Enter the Transformer: Revolutionizing Generative AI**

- In 2017, with the release of the "Attention is All You Need" paper, the Transformer architecture emerged. 

- Transformers marked a paradigm shift, offering scalable efficiency and parallel processing capabilities. 

- They could handle larger training datasets and efficiently utilize multi-core GPUs. 

- Crucially, Transformers introduced the concept of attention, enabling models to focus on word meanings. 👁️✨

- As the title suggests, attention became the cornerstone of this transformative approach.



In Summary:

The transition from RNNs to Transformers signifies a pivotal moment in generative AI. With attention at its core, Transformers revolutionized language understanding, offering scalability and efficiency unmatched by their predecessors. Let's explore further in the next segment!

