A Concise Report on Transformer Architecture and the Road to AGI

Yash Mishra

December 18, 2025

Abstract

This report summarizes the Transformer architecture—the foundational technology behind modern Large Language Models (LLMs)—and reflects on its role in the pursuit of Artificial General Intelligence (AGI).

Report

The Transformer, introduced in *Attention is All You Need*, replaces recurrence with self-attention, enabling parallel sequence processing and efficient capture of long-range dependencies. This architectural shift powers LLMs such as GPT, Gemini, and Claude across tasks in generation, reasoning, and multimodal understanding.

GPT (Generative Pre-trained Transformer) combines large-scale unsupervised pretraining with task-specific fine-tuning to predict the next token with high fidelity. Open ecosystems like Hugging Face lower barriers to experimentation and deployment for language and vision-language models. Intuitive resources (e.g., 3Blue1Brown) further demystify attention, embeddings, and optimization, while LaTeX remains essential for clearly formalizing the mathematical groundwork of these systems.

Beyond narrow task performance, the Transformer's scalability and versatility make it a cornerstone in the broader quest for **AGI**. Achieving AGI will likely require richer long-term memory, better grounding in the real world, and more autonomous learning and planning. Current LLMs hint at these capabilities, but sustained advances in architecture, training, and alignment are needed to approach human-level generality.

In summary, the Transformer is not merely a model; it is the catalyst accelerating progress from specialized AI toward the vision of AGI.