Chapter 4
How Do LLMs Work?
A Deep Dive Into Transformer Models

Satvik Vats
https://orcid.org/0000-0002-9422-4915
Madan Mohan Malaviya University of Technology (MMMUT), Gorakhpur, India

Vikrant Sharma
Graphic Era Hill University, Dehradun, India

Priya Singh
Dataculture Technologies Private Limited, India

Samriti Thakur
Dataculture Technologies Private Limited, India

Daksh Rawat
https://orcid.org/0009-0007-9046-6988
Graphic Era Hill University, Dehradun, India

ABSTRACT

This chapter provides an in-depth overview of the Large Language Models (LLMs) by putting its lineage in the Transformer architecture- a framework that unites parallelism with long-range dependencies modeling via self-attention. It then introduces the most fundamental interfaces: multi-head attention, positional encoding and feed-forward networks, in a systematic fashion, followed by explanation of pre-training paradigms masked language modeling and causal language modeling. It also focuses on hyper-parameters of better tunes, especially reinforcement learning with human feedback (RLHF). Text generation Decoding routines, such as greedy decoding, beam or nucleus sampling, are discussed, and the limitations of the method

DOI: 10.4018/979-8-3373-3785-2.ch004

are highlighted in terms of such aspects as hallucinations, bias, interpretability, and environmental impact of LLMs. The chapter ends by questioning the ethical aspects of developing the LMM and by mapping future research agendas leading to the ethical discovery of these technologies.

INTRODUCTION

Large Language Models (LLMs) based on transformers have disrupted the areas of artificial intelligence and natural language processing. With the aid of the watershed Attention Is All You Need work (Vaswani et al., 2017), these models can perform text generation, text reading, and question answering, tasks previously limited to humans in terms of writing ability. Unlike recurrent neural networks, including variants of long-short term memory and transformers do not involve sequential processing of data; they process language at different points in parallel via self-attention. This improvement of architectures has led to training models of unprecedented scale and complexity with systems having billions or even trillions of parameters (Kaplan et al., 2020).

Modern LLMs such as the GPT-4, PaLM, LLaMA all have much more to offer than simple text generation. These models easily perform very complex tasks such as complex question answering and code-generation as well as on creative writing and basic ability to reason (Wei et al., 2022). They are able to excel because of a two-part training system: first, the system learns broad language skills from a lot of text data and then it is adjusted for specific jobs or preferences (Brown et al., 2020). At the same time, achieving this can be very difficult because it means using a lot of computing resources, the risk of models spreading inaccurate or harmful messages and the appearance of "hallucinations" where information produced by the model is convincing but wrong (Zhang et al., 2025).

LLMs have a huge influence on society now and the effect is still growing. AI is changing sectors such as education and healthcare, and it is also raising issues related to proper and improper uses of AI (Bender et al., 2021). At this stage where technology is changing fast, we should pay attention not only to what the models can accomplish but also to how they do it. This chapter explores, in detail, every aspect of LLMs, beginning from their main structure through the methods used to train them, their general output and what difficulties will face them as development moves forward.