

# Chapter 2

## A Comprehensive Review on Large Language Models

### Exploring Applications, Challenges, Limitations, and Future Prospects

Asmita Yadav

*Jaypee Institute of Information Technology, India*

#### **ABSTRACT**

*In the realm of computer science and language, large language models (LLMs) stand out as remarkable tools of artificial intelligence (AI). Proficient in deciphering intricate language nuances, LLMs offer sensible responses and find applications in natural language understanding, language translation, and question answering. This chapter delves into the history, creation, training, and multifaceted applications of LLMs. It explores the basics of generative AI, focusing on generative pre-trained transformers (GPT). Examining the evolution of LLMs and their diverse applications in medicine, education, finance, and engineering, the chapter addresses real-world challenges, including ethical concerns, biases, comprehensibility, and computational requirements. It serves as an informative guide for researchers, practitioners, and enthusiasts, elucidating the potential, challenges, and future of LLMs in AI.*

#### **1. INTRODUCTION**

Large Language Models (LLMs) signify a noteworthy leap forward in the realms of natural language processing and artificial intelligence research (Hochreiter & Schmidhuber, 1997). These models have substantially elevated machines' capacity to comprehend and generate language resembling human expression (Li et al., 2023). Employing deep learning methodologies and extensive datasets, LLMs have showcased their adeptness across diverse language-oriented tasks such as text creation, translation, summarization, question answering, and sentiment analysis. The origins of LLMs can be traced back to early language model and neural network development. Initial attempts centered around statistical tech-

DOI: 10.4018/979-8-3693-3502-4.ch002

niques and n-gram models (Moor et al., 2023), yet these approaches struggled with capturing extensive contextual dependencies in language.

The turning point for LLMs occurred with the inception of the Transformer architecture in the seminal work “Attention is All You Need” by Vaswani et al. in 2017 (Arisoy et al., 2012). Leveraging the self-attention mechanism, the Transformer model facilitated parallelization and efficient handling of long-range dependencies. This laid the groundwork for influential models like OpenAI’s GPT series and Google’s BERT, both achieving groundbreaking results across various language tasks (Mikolov et al., 2010).

Subsequently, LLMs have progressed through multiple developmental stages, with models evolving in size and intricacy. The GPT series, starting with GPT-1 and extending to GPT-2 and GPT-3, has incrementally expanded in parameter count, enabling more sophisticated language comprehension and generation capabilities. Similarly, BERT-inspired models have seen advancements in pre-training strategies, exemplified by models like ALBERT (A Lite BERT) and RoBERTa (Pachouly et al., 2022; Vaswani et al., 2017), enhancing performance and efficiency.

Furthermore, LLM advancements have extended into specialized domains, with models tailored for specific tasks like medical language processing, scientific research, and code generation. Addressing ethical concerns, interpretability, and mitigating biases in LLMs has been a focus, ensuring responsible and equitable utilization. The progression of Large Language Models has revolutionized natural language processing and AI research, resulting in notable achievements across diverse language tasks.

In summary, the evolution of language modeling research has undergone four key development stages: statistical language models, neural language models, pre-trained language models, and large language models (Devlin et al., 2018). This research primarily concentrates on LLMs, aiming to shed light on the data sources for pre-training LLaMA, as outlined in Table I and Figure 1.

OpenAI developed a contemporary language model known as ChatGPT, utilizing the GPT-3.5 architecture and training it with a substantial volume of text data sourced from the internet, including books, articles, wikis, and websites. II. ChatGPT excels in generating responses that closely mimic human language, facilitating engaging conversations with users. In the realm of computer vision (CV), researchers are actively involved in creating vision-language models inspired by ChatGPT’s capabilities. These models are specifically crafted to enhance multimodal dialogues, where both visual and textual information play crucial roles (Lan et al., 2019). Various type of large language models are shown in Figure 1.

Furthermore, progress in this field has given rise to GPT-4 (Liu et al., 2019), which extends the capabilities of language models by seamlessly integrating visual information into the input. This inte-

*Table 1. Pre-training data: Mixtures of data used for pretraining LLaMA (Chen et al., 2021)*

<b>Dataset</b>	<b>Sampling pops(%)</b>	<b>Epochs</b>	<b>Disk size(GB)</b>
Wikipedia	4.7	2.47	84 GB
AeXiv	2.6	1.1	93 GB
Stack Exchange	1.9	1.08	77 GB
CommonCrawl	69	1.15	3.6 TB
Github	4.7	.61	331GB
C4	17	1.14	791 GB
Books	5.1	2.41	89 GB

21 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:  
[www.igi-global.com/chapter/a-comprehensive-review-on-large-language-models/346321](http://www.igi-global.com/chapter/a-comprehensive-review-on-large-language-models/346321)

## Related Content

---

### IT-ASSIST: Towards Usable Applications for Elderly People

Claas Ahlrichs, Daniel Kohlsdorf, Michael Lawoand Gerrit Kalkbrenner (2011). *International Journal of Ambient Computing and Intelligence* (pp. 33-41).

[www.irma-international.org/article/assist-towards-usable-applications-elderly/52039](http://www.irma-international.org/article/assist-towards-usable-applications-elderly/52039)

### Behavioral Implicit Communication (BIC): Communicating with Smart Environments

Cristiano Castelfranchi, Giovanni Pezzuloand Luca Tummolini (2010). *International Journal of Ambient Computing and Intelligence* (pp. 1-12).

[www.irma-international.org/article/behavioral-implicit-communication-bic/40346](http://www.irma-international.org/article/behavioral-implicit-communication-bic/40346)

### CoAP-Based Lightweight Interoperability Semantic Sensor and Actuator Ontology for IoT Ecosystem

Sukhavasi Suman, Thinagaran Perumal, Norwati Mustapha, Razali Yaakob, Mohd Anuaruddin Bin Ahmadonand Shingo Yamaguchi (2021). *International Journal of Ambient Computing and Intelligence* (pp. 92-110).

[www.irma-international.org/article/coap-based-lightweight-interoperability-semantic-sensor-and-actuator-ontology-for-iot-ecosystem/275760](http://www.irma-international.org/article/coap-based-lightweight-interoperability-semantic-sensor-and-actuator-ontology-for-iot-ecosystem/275760)

### Big Data Analytics-Based Agro Advisory System for Crop Recommendation Using Spark Platform

Madhuri J.and Indiramma M. (2023). *Handbook of Research on AI and Machine Learning Applications in Customer Support and Analytics* (pp. 227-247).

[www.irma-international.org/chapter/big-data-analytics-based-agro-advisory-system-for-crop-recommendation-using-spark-platform/323123](http://www.irma-international.org/chapter/big-data-analytics-based-agro-advisory-system-for-crop-recommendation-using-spark-platform/323123)

### A Genetic Algorithm-Based Multivariate Grey Model in Housing Demand Forecast in Turkey

Miraç Eren, Ali Kemal Çelikand brahim Huseyni (2016). *Intelligent Techniques for Data Analysis in Diverse Settings* (pp. 42-65).

[www.irma-international.org/chapter/a-genetic-algorithm-based-multivariate-grey-model-in-housing-demand-forecast-in-turkey/150287](http://www.irma-international.org/chapter/a-genetic-algorithm-based-multivariate-grey-model-in-housing-demand-forecast-in-turkey/150287)