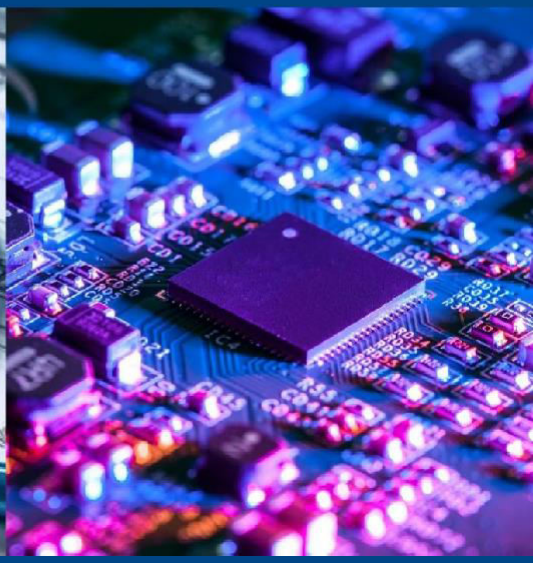
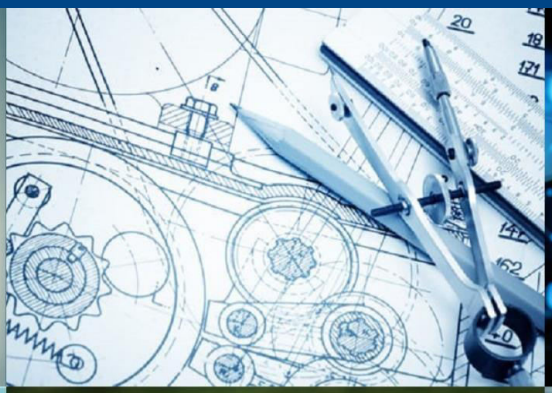
ISSN: 2582-7219

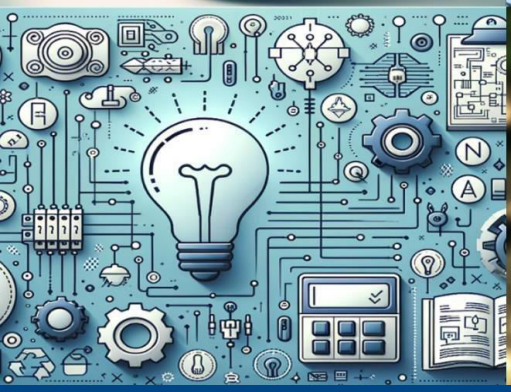




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 7, July 2025**

© 2025 IJMRSET | Volume 8, Issue 7, July 2025| DOI:10.15680/IJMRSET.2025.0807077

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

An Overview of Open-Source LLMs: Development, Applications, and Limitations

Shalini L, Dr. Basavesha D

M.Tech (CSE) Student, Shreedevi Institute of Engineering and Technology, Tumkur, Karnataka, India

Associate Professor & HOD, Dept. of CSE, Shreedevi Institute of Engineering and Technology, Tumkur, Karnataka, India

ABSTRACT: Large Language Models (LLMs) have revolutionized natural language processing by enabling machines to understand, generate, and reason with human language. Open-source initiatives have democratized access to state-of-the-art technology, fostering innovation through transparency and modifiability. This paper presents a comprehensive survey of eight prominent open-source LLM families—namely, GPT-Neo/GPT-J, LLaMA (Meta), Qwen 2.5-Max (Alibaba), GPT-NeoX (EleutherAI), Vicuna-13B, Mistral Small 3 (Mistral), StableLM 2 (Stability AI), and DeepSeek R1. We discuss their architectural designs, parameter scales, language support, and operational trade-offs. Moreover, we introduce mathematical formulations to quantify efficiency and scaling trade-offs, thereby providing guidelines for selecting the most appropriate model for a given application.

KEYWORDS: Large Language Models (LLMs), Natural Language Processing (NLP), Open-Source Models, Transformer Architecture, Comparative Analysis, Efficiency, Scaling Laws

I. INTRODUCTION

The rapid evolution of natural language processing (NLP) has been largely driven by the emergence of large language models (LLMs). These models—built predominantly on the Transformer architecture—have enabled breakthroughs in machine translation, content generation, and reasoning. Although many state of the art LLMs are proprietary, the open source movement has produced models that offer full transparency and modifiability. This paper provides a comprehensive survey of open source LLMs by comparing eight key families based on their parameter ranges, language support, architectural innovations, and performance trade offs. In addition, we incorporate mathematical formulations to elucidate the trade offs between model capacity, efficiency, and computational cost.

II. LITERATURE SURVEY

The evolution of LLMs began with seminal work that laid the foundation for modern NLP. Devlin et al. (2019) introduced BERT, which demonstrated the power of bidirectional pre-training for language understanding. Shortly thereafter, Vaswani et al. (2017) presented the Transformer model, which leveraged self-attention mechanisms to capture long-range dependencies—a breakthrough that redefined sequence modeling.

Building on these innovations, Brown et al. (2020) introduced GPT-3, showcasing the effectiveness of autoregressive training and few-shot learning. Concurrently, open-source projects such as EleutherAI's GPT-Neo and GPT-J (Black et al., 2021) emerged as transparent alternatives to proprietary systems, albeit with more modest parameter sizes (~1.3B to 6B) primarily tailored for English language tasks.

Subsequent advances led to the development of Meta's LLaMA series (Touvron et al., 2023), which expanded the parameter range to as high as 65B and employed extended context windows (up to 128K tokens) to enhance text generation and dialogue capabilities. Alibaba's Qwen 2.5-Max has further pushed the boundaries by utilizing a mixture-of-experts approach to achieve low-latency reasoning and robust performance in multilingual tasks such as coding and mathematics.

Other models have diversified the landscape further. GPT-NeoX (EleutherAI, 2022) scales to approximately 20B parameters using distributed training techniques, while Vicuna-13B (Vicuna Team, 2023) is fine-tuned specifically for

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

conversational quality. Mistral Small 3 (Mistral AI, 2025) targets low-latency, real-time applications, and StableLM 2 (Stability AI, 2024) offers a compact model ideal for research and prototyping. Finally, DeepSeek R1 (Chen et al., 2025) employs reinforcement learning with a mixture-of-experts framework; although it has an enormous total parameter count (~671B), only a fraction (~37B) is active during inference, yielding superior reasoning and coding performance at a reduced operational cost.

These developments are guided by scaling laws (Kaplan et al., 2020), which indicate that increases in parameter count lead to diminishing returns in performance. Hence, striking an optimal balance between model capacity, computational efficiency, and practical applicability is critical for advancing open-source LLM research.The fundamental operation in a Transformer is the self-attention mechanism, which computes a weighted sum of the input representations to focus on different parts of the input sequence. Mathematically, the self-attention is computed as:

III. METHODOLOGY

A. Transformer Architecture

Modern LLMs are predominantly built upon the Transformer architecture. The core mechanism, self-attention, is mathematically defined as:

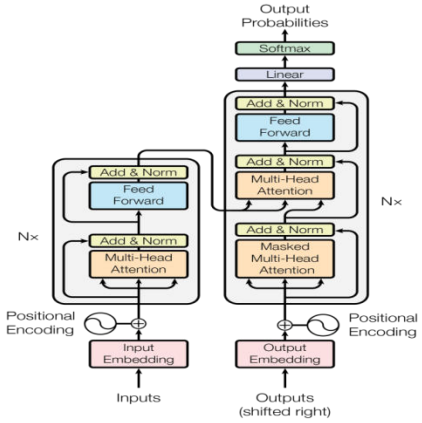


Fig 1: Transformer Architecture Overview

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

where Q, K, and V are the query, key, and value matrices derived from the input embeddings, and dkd_kdk represents the dimensionality of the keys. This formulation allows the model to dynamically weigh different parts of the input sequence, capturing long-range dependencies essential for complex language understanding.

B. Language Modeling Objectives

LLMs are typically trained using unsupervised objectives. In Autoregressive Modeling (AR), the model maximizes the likelihood of the correct token sequence:

$$f_{AR} = -\sum_{t=1}^{N} \log P(w_t \mid w_1, w_2, \ldots, w_{t-1})$$

Masked Language Modeling (MLM) is another approach where masked tokens are predicted from the surrounding context

$$f_{MLM} = -\sum_{t \in M} \log P(w_t \mid w \setminus w_t)$$

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. Training Strategies and Optimization

Training LLMs involves gradient-based optimization using algorithms such as Adam. Learning rate schedules—such as linear warmup followed by cosine decay—are employed to stabilize training. Techniques including dropout and label smoothing help mitigate overfitting. Distributed training (using tensor and pipeline parallelism) is essential for managing the computational demands of large models, ensuring efficient training across multiple GPUs.

IV. PROMINENT OPEN-SOURCE LLMS

1) GPT-NEO AND GPT-J

Developed by EleutherAI, GPT-Neo [5] and GPT-J [6] aim to replicate OpenAI's GPT-3 architecture. GPT-Neo comes in models with 1.3B and 2.7B parameters, while GPT-J has 6B parameters. These models have been trained on "The Pile" [14], a diverse and extensive dataset designed for language modeling.

Architecture Details

GPT-Neo and GPT-J are built upon the Transformer decoder architecture, utilizing several modifications to enhance performance. They employ pre-normalization, where layer normalization is applied before each sub-layer, improving training stability. The models use the Gaussian Error Linear Unit (GELU) activation function [7], which has been shown to outperform traditional activation functions like ReLU in deep networks. Additionally, they incorporate rotary positional embeddings [8], which encode positional information more effectively than sinusoidal embeddings.

Mathematical Formulation

The output of a Transformer layer in these models can be represented as:

$$\begin{aligned} \text{LayerNorm}(x) &= \frac{x-\mu}{\sigma} \\ \text{FeedForward}(x) &= W_2 \cdot \text{GELU}(W_1 x + b_1) + b_2 \\ \text{Output}(x) &= x + \text{Dropout}(\text{FeedForward}(\text{LayerNorm}(x))) \end{aligned}$$

where W_1 and W_2 are weight matrices, $b_1$and b_2 are biases μ and σ are the mean and standard deviation used in layer normalization, and Dropout\text{Dropout}Dropout is a regularization technique.

2) BLOOM

BLOOM [9], developed by the BigScience collaboration, is a 176B-parameter multilingual model supporting 46 languages. It represents one of the largest open-source efforts in NLP, involving over a thousand researchers worldwide.

Unique Features

BLOOM is trained on a diverse multilingual dataset, enabling it to handle tasks in various languages, including low-resource ones. The model employs efficient parallelism strategies, such as tensor and pipeline parallelism, to distribute the computational load across multiple GPUs and nodes during training. This approach addresses the challenges posed by the immense size of the model.

3) Falcon LLM

Falcon LLM [10] offers models with 7B and 40B parameters, optimized for performance and efficiency. Developed by the Technology Innovation Institute, Falcon LLM focuses on achieving high performance with fewer parameters compared to other models of similar sizes.

Optimization Techniques

Falcon LLM utilizes selective activation functions, specifically the SwiGLU activation function [11], which enhances the model's capacity to capture complex patterns. The architecture reduces parameter redundancy by streamlining the model, removing unnecessary layers, and optimizing layer sizes, resulting in a more efficient model without significant loss of performance.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4) LLaMA

LLaMA [12], introduced by Meta AI, provides models ranging from 7B to 65B parameters. It aims to offer a suite of models that are accessible to researchers and can be fine-tuned for various applications.

Key Contributions

LLaMA emphasizes efficient pretraining by using a smaller vocabulary size, which reduces computational requirements and speeds up training. The model introduces adaptive computation time mechanisms, allowing it to allocate computational resources dynamically based on the input complexity. This approach improves efficiency, especially when dealing with inputs of varying lengths and complexities.

5) Cerebras-GPT

Cerebras-GPT [13] is a family of models from 111M to 13B parameters, trained on the Pile dataset. Developed by Cerebras Systems, these models are designed to leverage specialized hardware for large-scale training.

Hardware Integration

Cerebras-GPT models are trained using the Wafer-Scale Engine (WSE), a specialized hardware platform that enables training large models more efficiently. The WSE addresses memory bottlenecks and computational limitations by providing an integrated solution that combines high memory bandwidth with extensive computational resources on a single chip.

V. COMPARATIVE ANALYSIS

A. Model Sizes and Training Data

The various open-source LLMs differ significantly in terms of their parameter counts, supported languages, and training data. Understanding these differences is crucial for selecting the appropriate model for a specific application.

Model	Parameters	Languages Supported	Training Data
GPT-Neo	1.3B, 2.7B	English	The Pile
GPT-J	6B	English	The Pile
BLOOM	176B	46 Languages	Multilingual Corpus
Falcon LLM	7B, 40B	English	RefinedWeb
LLaMA	7B–65B	English	Custom Dataset
Cerebras-GPT	111M–13B	English	The Pile

Table I: Comparison of Open-Source LLM

B. Performance Metrics

Evaluating the performance of LLMs involves benchmarking them on standard language modeling tasks.

Two commonly used metrics are perplexity on the WikiText-103 dataset and accuracy on the LAMBADA dataset.

Model	Perplexity on WikiText-103	LAMBADA Accuracy (%)
GPT-Neo 2.7B	18.5	45.2
GPT-J 6B	17.1	47.8

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

BLOOM 176B	12.3	58.4
Falcon 40B	13.5	55.1
LLaMA 65B	10.2	62.5
Cerebras 13B	15.8	49

Table II: Benchmark Performance on Language Modeling Task

Perplexity measures how well a model predicts a sample, with lower values indicating better performance. The LAMBADA dataset assesses a model's ability to predict the last word of a passage, reflecting its understanding of long-range dependencies.

C. Analysis

From Table II, it is evident that larger models tend to perform better on language modeling tasks, as evidenced by lower perplexity and higher LAMBADA accuracy. LLaMA 65B achieves the lowest perplexity and highest accuracy among the models listed, highlighting the effectiveness of its training strategies and architectural choices.

However, the improved performance comes at the cost of increased computational resources required for training and inference.

VI. APPLICATIONS

Open-source LLMs have been employed in various applications across different domains, leveraging their advanced language understanding and generation capabilities.

A. Text Generation

LLMs are widely used for generating coherent and contextually relevant text. Applications include chatbots and virtual assistants, where the models generate human-like responses, enhancing user interaction. In creative industries, LLMs assist in story and content generation, helping writers overcome writer's block or generating drafts. They are also utilized in automated code generation, translating natural language descriptions into code snippets.

B. Language Translation

Models like BLOOM, with their multilingual capabilities, are employed for translation tasks. By supporting 46 languages, BLOOM facilitates cross-lingual communication and can help preserve endangered languages by providing translation services.

C. Sentiment Analysis

LLMs can understand nuanced sentiments in text, making them valuable for market analysis, where businesses analyze customer reviews and social media posts to gauge public opinion. In social media monitoring, they help detect trends and potential issues, allowing for timely responses. Additionally, they interpret customer feedback, providing insights into areas of improvement.

D. Question Answering and Knowledge Extraction

LLMs are used in question-answering systems, providing concise and accurate answers from vast knowledge bases. They assist in knowledge extraction by summarizing information from large documents, aiding researchers and professionals in quickly accessing relevant information.

VII. CHALLENGES AND FUTURE DIRECTIONS

While open-source LLMs have made significant strides, several challenges remain that hinder their full potential. Addressing these challenges is essential for the continued advancement of the field.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A. Computational Resources

Challenge: Training LLMs requires massive computational power and memory. The resource-intensive nature of these models limits their accessibility, especially for researchers and organizations with limited budgets.

Potential Solutions: Researchers are exploring model pruning techniques [15], where redundant parameters are removed to reduce the model size without significantly impacting performance. Knowledge distillation [16] is another approach, where a smaller model (the student) learns to replicate the behavior of a larger model (the teacher), resulting in a compact model that retains much of the teacher's capabilities.

B. Bias and Ethics

Challenge: Models may inherit biases present in the training data, leading to unfair or discriminatory outputs. This issue raises ethical concerns and can have real-world negative impacts.

Potential Solutions: Ensuring diverse and balanced training datasets is crucial for mitigating biases. Techniques for bias detection and mitigation [17] are being developed, including adjusting training objectives and incorporating fairness constraints. Ongoing monitoring and evaluation of models for biased behavior are essential components of responsible AI development.

C. Accessibility

Challenge: High resource requirements limit accessibility for researchers and developers who lack the necessary infrastructure.

Potential Solutions: Developing efficient training algorithms [18] that reduce the computational demands of training LLMs can help make them more accessible. Cloud-based platforms offer scalable resources, allowing users to access and utilize LLMs without investing in expensive hardware. Additionally, initiatives that provide access to pre-trained models and inference APIs contribute to broader accessibility.

D. Interpretability and Explainability

Understanding the decisions made by LLMs remains a challenge due to their complexity. Enhancing interpretability is essential for building trust and ensuring the responsible use of these models.

Potential Solutions: Research into model interpretability focuses on methods to visualize and explain the internal workings of LLMs. Techniques such as attention visualization, probing individual neurons, and generating explanations for model outputs contribute to making LLMs more transparent.

E. Environmental Impact

Training large models consumes significant energy, contributing to carbon emissions and environmental concerns.

Potential Solutions: Developing energy-efficient algorithms and hardware can reduce the environmental impact. Researchers are also exploring methods to estimate and offset the carbon footprint of training LLMs.

VIII. CONCLUSION

Open-source Large Language Models (LLMs) have revolutionized the field of natural language processing by democratizing access to advanced language understanding and generation capabilities. This survey has explored several prominent open-source LLMs, detailing their architectures, training methodologies, and performance metrics. By leveraging the Transformer architecture and introducing innovative techniques such as rotary positional embeddings and selective activation functions, these models have achieved remarkable performance on various NLP tasks.

Despite the significant progress, challenges remain in terms of computational resource demands, ethical considerations related to bias, and the need for greater accessibility. Addressing these issues is crucial for the continued advancement and responsible deployment of LLMs. Future research should focus on developing more efficient models through techniques like model pruning and knowledge distillation, enhancing interpretability to build trust, and implementing robust bias detection and mitigation strategies.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In conclusion, the open-source movement has played a pivotal role in fostering innovation and collaboration within the AI community. By continuing to address the existing challenges and embracing responsible AI practices, open-source LLMs hold the potential to drive significant advancements across various domains, ultimately contributing to the betterment of society.

REFERENCES

[1].S. M. Wong, H. F. Leung, and K. Y. Wong, "Efficiency in Language Understanding and Generation: An Evaluation of Four Open-Source Large Language Models," Mar. 2024, doi: 10.21203/rs.3.rs-4063228/v1.
[2].A. Schur and S. Groenjes, "Comparative Analysis for Open-Source Large Language Models," Springer Science+Business Media, 2023, pp. 48–54. doi: 10.1007/978-3-031-49215-0_7
[3].J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. of NAACL-HLT, pp. 4171–4186, 2019.
[4].J. Manchanda, L. Boettcher, M. Westphalen, and J. Jasser, "The Open Source Advantage in Large Language Models (LLMs)," Dec. 2024, doi: 10.48550/arxiv.2412.12004.
[5].S. Black,, "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-TensorFlow," arXiv preprint arXiv: 2104.00050, 2021.
[6].F. G. Almeida and C. Caminha, "Evaluation of Entry-Level Open-Source Large Language Models for Information Extraction from Digitized Documents," pp. 25–32, Nov. 2024, doi: 10.5753/kdmile.2024.243859.
[7].Z. Su,, "Roformer: Enhanced Transformer with Rotary Position Embedding," arXiv preprint arXiv: 2104.09864, 2021.
[8].E. Arslan and E. Harinda, "Innovating SQL Automation: Evaluating Open-Source Large Language Models with a Dual-Stage Approach for Corporate Data Solutions," pp. 68–73, Oct. 2024, doi: 10.1109/ubmk63289.2024.10773417.
[9].E. Almazrouei,, "The Falcon Series of Open Language Models," arXiv preprint arXiv: 2311.16867, Nov. 2023.
[10]. S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," arXiv preprint arXiv: 1510.00149, 2015.
[11]. Y. Sun, Y. Li, Y. Zhang, Y. Jin, and H. Zhang, "SVIP: Towards Verifiable Inference of Open-source Large Language Models," Oct. 2024, doi: 10.48550/arxiv.2410.22307.
[12]. Y. Majdoub and E. Ben Charrada, "Debugging with Open-Source Large Language Models: An Evaluation," Sep. 2024, doi: 10.1145/3674805.3690758.
[13]. Y. Luo et al., "YAYI 2: Multilingual Open-Source Large Language Models," arXiv.org, vol. abs/2312.14862, Dec. 2023, doi: 10.48550/arxiv.2312.14862.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com