STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 12, Issue 1, 2025
January–April
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Review JoAIRA

# Generative Artificial Intelligence with Emphasis on Large Language Models: Review and Current Trends

Joshua Michael[1,*], Fabian Barreto[2], Sujata Deshmukh[3]

***Abstract***

*Generative Artificial Intelligence deals with AI systems that generate new content, such as text, and images. It accomplishes this by using data patterns of texts and images that already exist. Generative AI began an era of major advancement in AI, producing more refined and human-like results. Large Language Models, LLMs, is a part of Generative AI with applications in Natural Language Processing such as text generation, translation, summarization, sentiment detection and question answering. This quickly advancing technology is influencing the future of creative and analytical tasks across multiple industries. OpenAI gave us ChatGPT that was trained on LLM. Tech Giants like Google and Meta then entered the race to develop and improve the existing models, with products like Gemma and Llama 3 respectively. LLMs also present ethical issues, including biases inherent in their training data and the risk of being used to create misleading information. This study reviews both proprietary and open-source LLMs in the literature and explains the cost consideration, current trends and future scope.*

**Keywords:** ChatGPT, deep learning, generative artificial intelligence, large language models, natural language processing, retrieval augmented generation

## INTRODUCTION

There have been significant advances in the field of Natural Language Processing (a branch of Machine Learning) primarily because of three main reasons: technology advances in Deep Learning, availability of huge amounts of text data over the internet and the availability of Compute power made possible because of special Graphical Processing Units (GPUs). There was a lot of investment from Silicon Valley to develop these Graphic Processors. Big Language Models (LLMs) are quickly changing question answering, sentiment analysis, translation, summarization, and text generation in a variety of industries. Figure 1 shows the training and adaptation process in LLMs for different tasks. LLMs are "large" because of the enormous size of the training set (billions of words or tokens (3/4th of a word) and the large number of parameters (weights or branches) of the neural network. Both are crucial for best performance of a LLM which is evaluated against benchmarks like General Language Understanding Evaluation (GLUE) [1] or Bilingual Evaluation Understudy (BLEU) [2]. The introduction of the Attention Mechanism significantly enhanced performance of LLMs for Machine Learning task.

***Author for Correspondence**
Joshua Michael
E-mail: joshua.michael@fragnel.edu.in

[1]Assistant Professor, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandra West, Mumbai, Maharashtra, India
[2]Assistant Professor, Department of Electronics and Telecommunication, Xavier Institute of Engineering, Mahim West, Mumbai, Maharashtra, India
[3]Professor, Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Fr. Agnel Ashram, Bandra West, Mumbai, Maharashtra, India

Received Date: August 05, 2024
Accepted Date: November 12, 2024
Published Date: December 30, 2024

**Citation:** Joshua Michael, Fabian Barreto, Sujata Deshmukh. Generative Artificial Intelligence with Emphasis on Large Language Models: Review and Current Trends. Journal of Artificial Intelligence Research & Advances. 2025; 12(1): 40–46p.

## TRANSFORMER NEURAL NETWORK

Vaswani *et al*. suggested the Transformer architecture, which does not use any recurrent or convolutional layers and instead depends only on

 

self-attention processes [3]. Because it can be parallelized, this method is quite effective for training on big datasets. The Transformer model captures long-range dependencies in text and has laid the groundwork for subsequent groundbreaking models. The Transformer's encoder-decoder architecture, which was adapted from Vaswani *et al*. [3], is seen in Figure 2.

**Self-Attention**

Transformer models are built using attention mechanisms at the core with an encoder and decoder. Typically, six encoders are stacked on top of each other. The number of encoders is a hyperparameter. All encoders share an identical structure, and do not share weights. A feed forward neural network and self-attention make up each encoder's two sublayers.

The Self-Attention sub-layer helps the encoder to look at relevant parts of the words as it encodes the central word in the input sequence. Self-Attention is a calculated probability score using Matrix Multiplication.

Every row in the embedding matrix X corresponds to a word into the input sentence (dimensionality is 512). Matrix X is multiplied by the weight matrices ($W$^$Q$, $W$^K, and $W$^V, the values of these matrices are set during the training process) to calculate the Query (Q), Key (K), and Value (V) matrices (dimensionality is 64). The self-attention calculation in Matrix form can be expressed as shown in Eq. (1), ($d_k$ value is 64) [4]:

$$z = softmax\left(\frac{Q \times k^T}{\sqrt{d_k}}\right)V \tag{1}$$

**Multi-Head Attention**

Multi-layer attention enhances the model's capacity to concentrate on various positions, giving multiple subspaces. There are eight attention heads or layers in the transformer. Separate weight matrices ($W^Q$, $W^K$, and $W^V$) are maintained for each head/layer. We multiply X by the weight matrices to produce Q, K and V matrices. This is now performed eight times with different weight matrices, resulting in eight different Z matrices for each input embedding.

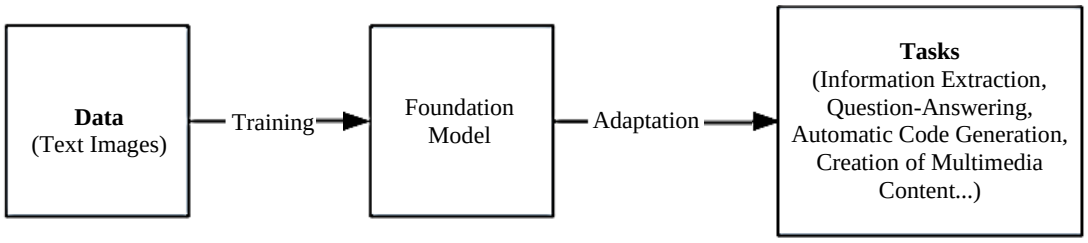


**Figure 1.** Training and adaptation process of large language models.

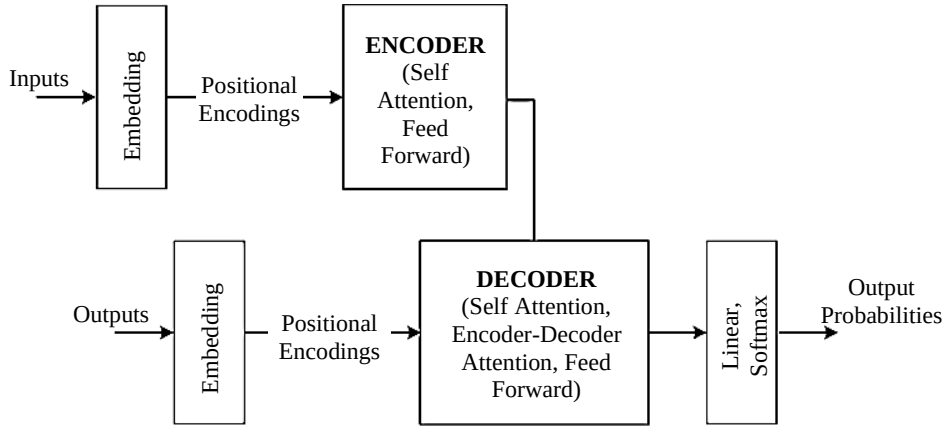


**Figure 2.** Transformer network architecture.

 

These Z matrices are condensed into a single Z matrix. In order to obtain the final Z matrix, which is supplied as input to the Feed Forward Neural Network (FFNN), we concatenate the eight Z matrices and then multiply them by an additional weight matrix $W$O, which is determined during the training phase. It is possible for multiple pathways to proceed through the FF layer simultaneously. The Encoder-Decoder attention layer is located between the Self-Attention and Feed-forward layers in the Decoder. This feature helps the decoder focus on relevant portions of the input phrase.

## LARGE LANGUAGE MODELS

Tech Giants have developed and released their LLMs based on the Transformer Architecture (with few variations). They can be classified as Encoder-only, Decoder-only and Encoder-Decoder Architectures.

### LLMs with Encoder-only Architecture

Bidirectional Encoder Representations from Transformers (BERT), a popular encoder-only architecture outperformed various existing state-of-the-art models on different evaluation metrics. BERT was developed by Google in 2018 [5]. BERT was trained on entire Wikipedia and BookCorpus.

BERT was trained on two different tasks to learn general language representations. It was trained in two variations: one model contains BERTBASE which has 12 layers in the Encoder with 110 million (M) parameters while BERTLARGE has 24 layers in the Encoder with about 340 million (M) parameters.

#### *Masked Language Modelling (MLM)*

For this objective, approximately 15% of the words within the input text are chosen at random and substituted with a designated (MASK) token, a unique vocabulary token. It helps to understand the contextual relationships between words to make precise predictions for missing words. Using the context that the neighboring words give, the model's objective is to predict the original words. BERT can comprehend the connections and dependencies between words in a phrase because of this bidirectional training.

#### *Next Sentence Prediction (NSP)*

In a given text, BERT is also trained to anticipate if a sentence comes after another. During the training process, sentence pairs are sampled, and the model learns to determine whether the second sentence directly follows the first one. This helps the model understand contextual relationships between sentences.

BERT's capacity for transfer learning and bi-directional attention mechanisms transformed the field of natural language processing. Search engines like Google use BERT to enhance the understanding of search queries, resulting in more relevant search results by considering the context and nuances of user input. BERT is also employed for sentiment analysis in social media, product reviews, or customer feedback to interpret the sentiment expressed in the text.

### LLMs with Decoder-only Architecture

Decoder-only architectures are designed primarily for generating text, unlike encoder-decoder models, which are used for tasks like translation or summarization.

#### *Generative Pretrained Transformers*

GPT stands for Generative (model can generate text), pre-trained (initially trained on some data and can be adapted to other data) Transformer (the specific neural network architecture). OpenAI has been a major driving force behind the development of Transformers Decoder-alone architecture. It first unveiled the Generative Pretrained Transformer (GPT) in 2017, and then GPT-2 in 2018, GPT-3.5 in 2022, and GPT-4 in 2023. GPT-4o (o for omni) was released by OpenAI on May 13, 2024. It can make outputs in any mix of text, audio, and graphics and accepts a variety of input formats, including text, audio, images, and video [6].

 

GPT can produce extremely fluid text because it has been trained on a vast volume of text. By projecting the next word based on the previous ones, the GPT is pretrained. It is used to autocomplete an input given an input prompt. GPT-3, in particular, is known for its large scale, with 175 billion (B) parameters, and it has demonstrated impressive performance in various language tasks because of its strong few-shot learning capabilities [7].

***Large Language Model Meta AI (LLaMa)***

In February 2023, Meta launched LLaMa (Large Language Model Meta AI), a state-of-the-art foundational LLM aimed at supporting researchers in advancing their work within AI [8]. Because foundation models are trained on large amounts of unlabeled data, they are excellent candidates for fine-tuning in a variety of applications. LLaMa models are transformer architecture based and have parameters ranging from 7 billion to 65 billion. During the training of a 65 billion-parameter model, Meta's code achieved approximately 380 tokens per second per GPU on a setup comprising 2048 A100 GPUs with 80 GB of RAM. This resulted in the processing of 1.4 trillion tokens for training over the LLaMa dataset.

**LLMs with Encoder-Decoder Architectures**

The T5 Text-to-text Transfer Transformer [9] combines Natural Language Understanding and Natural Language Generation Tasks by converting them to text-to-text tasks. BART is an additional encoder-decoder transformer architecture. Bidirectional and Auto-Regressive Transformer is referred to as BART. There are over 140 million characteristics in BART. Creating clear, semantically meaningful text from distorted textual data is the main task of BART. It can also be used for different NLP subtasks.

## CURRENT TRENDS

The current trends include enhancing logical reasoning and mitigating biases. There is also a felt need for Domain-specific, fine-tuned LLMs.

**Scale and Size**

Though industry experts are competing to create larger language models, a few businesses are also looking at smaller language models (SLMs) that suit them better as they prioritize efficiency and security. The newer LLM model parameter sizes increase from billions to trillions of parameters. The model performs better as its parameters are increased and when it is trained on a larger volume of material. Scaling up enhances their ability to perform numerous NLP tasks across diverse fields of knowledge, showcasing extensive general capabilities. GPT-4 Model (13T tokens) [10] outperforms GPT-3 (499B tokens) to answer Exam Questions.

It is crucial to highlight that a well-trained model with a smaller parameter size, when utilizing reinforcement techniques, can surpass the performance of a not so properly trained larger parameter model. Specifically, Llama-3 (70 billion) is a strong competitor for GPT-4 (1.75 trillion).

**Fine-tuning and Specialization**

Refinement and specialization of Large Language Models (LLMs) enhance their performance and applicability across diverse tasks and domains.

***Full-parameter Fine tuning***

Full parameter fine-tuning updates all the parameters (weights) of the model. These weight matrices are very large (in Billions). This type of fine-tuning requires significant computational resources [11] involving large GPUs or GPU Clusters.

***Parameter efficient Fine-tuning (PEFT)***

In 2021, a Microsoft research team introduced the Low-Rank Adaptation (LoRA) approach. Two distinct, smaller matrices that are multiplied together to create a matrix of the same size as the model's

 

weight matrix are used by LoRA to track the modifications that need to be made to the matrix weights. LoRA ensures that all the parameters of the model are trained, only here, we use two smaller matrices of low rank (Matrix Decomposition). As per Hu *et al.* [12], the LoRA technique results in a memory use reduction of three times. QLORA [13] (Quantized LoRA) is a quantized version of LoRA. 16-bit floats are stored as 4-bit floats reducing the memory footprint by a factor of 4, while retaining the original precision of the model.

### *Reinforcement Learning with Human Feedback*

Training LLMs requires the use of Reinforcement Learning from Human Feedback (RLHF) [14]. It ensures alignment with human values, making models more helpful, honest, and harmless. Human feedback in the form of judgments or ratings guides the learning process of the LLM algorithm, enhancing its performance and reliability.

### *Multi-modal AI and Open-source contributions*

There is an increasing trend in language models to be integrated with other modalities, such as visuals and audio. The goal of multimodal models was to comprehend and produce information from various data kinds. Many large language models like LLaMa are open-sourced, allowing researchers and developers to access and experiment with these models. This collaboration contributes to rapid advancements in the field.

## COST CONSIDERATIONS AND FUTURE SCOPE

### Why do LLMs cost so much?

LLMs cost a lot due to high computational demands, extensive training data, specialized hardware, ongoing research, and development efforts, as well as the energy costs associated with running large models. We briefly discuss a few factors affecting the cost of LLMs:

### *Use Case and Model Size*

The Business (Enterprise) needs to decide which platform partner or vendor to work with and the Generative AI use case. More computational power and money are required for larger, more parameterized models. Vendors vary their prices according to the size of the model. Enterprises can choose from among Flan (11B), Granite (13B) or Llama2 (70 B) models. The enterprise needs to determine which model best suits their use case.

### *Pre-training and Inferencing*

Pre-training is the process of building the LLM from scratch. It requires a tremendous amount of compute time and effort, though it gives the Enterprise control over the data to train an LLM. The cost factors for GPT3 were over a thousand GPUs, over a 30-day period costing over 4.6 million dollars. This is expensive and only a few players have emerged in the market.

Inferencing refers to the process that the model uses to understand the prompt question and the process that it uses to generate a response. Tokens are discrete units of information used in inferencing. Depending on the tokenizer being used, a token's size can change.

### *Tuning*

Tuning the model reduces the cost of scale, whereas Parameter Effective Fine Tuning reduces the cost at scale by deploying a smaller model (100 to 1000 labelled data sources).

### *Hosting and Deployment*

The LLMs can be deployed on a cloud platform or through a Software as a Service (SaaS) framework or it could be hosted within the premises of the Enterprise. The client has to pay a subscription (a known amount) to the vendor while inferencing from the LLM.

 

## Future Trends

LLMs currently have only a System 1 thinking capability where answers (responses) are fast and instinctive and cached. Another innovation is the Green AI that prioritizes sustainability in LLM development, aiming for energy efficiency and reduced environmental impact.

### *System 2 Thinking*

LLMs exhibit System 1 thinking: fast, instinctive responses akin to mental math or speed chess. System 2 thinking is more rational, slower, performs complex decision making and is a lot more conscious. Like when we have to multiply two big numbers in our mind or while playing Chess in a tournament setting and have time to keep track of the computations (tree search in chess) [15]. LLMs are allowed to take time and deliver accurate responses.

### *Learning by Self improvement*

This approach too has two major steps [16]. LLMs can become more intelligent than humans and beat humans in Game playing. First, LLM learn by imitating best human players. At present, human labelers are writing answers for LLMs. So, LLMs cannot go above human response accuracy. Second, Learn by Self-improvement (reward = win the game). LLMs can surpass human intelligence. In the domain of Open Language modeling, there is no easy way to evaluate reward function. However, in narrow domains, scope for reward function can be possible! LLMs can, in other words, self-improve in contexts where there is a reward function.

### *1-bit LLM*

A 1-bit LLM version [17] that meets the full-precision of the FP-16 Transformer LLM is one in which each weight of the LLM is ternary {–1, 0, 1}. In terms of latency, memory, throughput, and energy usage, this paradigm is more economical.

### *RAG and GraphRAG*

Large Language Models (LLMs) sometimes hallucinate and generate speculative and potentially inaccurate responses based on outdated or limited training data. Retrieval-Augmented Generation (RAG) addresses this by sourcing up-to-date facts from external knowledge bases, improving response accuracy [18]. LangChain and LLamaIndex are key frameworks for RAG, using sentence embeddings stored in vector databases to enhance context retrieval. Integration with knowledge graphs, as seen in GraphRAG [19], further refines LLM responses, exemplified by LinkedIn's efficiency gains in ticket resolution.

## CONCLUSION

This study offers a comprehensive analysis of recent studies on LLMs, highlighting their profound impact on artificial intelligence, particularly their varied applications in natural language processing. Future research prospects include the development of Retrieval-Augmented Generation (RAG) systems, which integrate retrieval and generative approaches to optimize performance. Additionally, the study underscores the significance of fine-tuning LLMs for specific tasks, enhancing their efficacy across diverse domains. The study intends to contribute to the conversation on utilizing LLMs for creative AI solutions by addressing these themes.

## Declaration of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

## REFERENCES

1. Wang A, *et al*. GLUE: A MultiTask Benchmark and Analysis Platform for Natural Language Understanding. arXiv: 1804.07461. 2018.
2. Papineni K, *et al*. Bleu: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002; 311–318.

 

3. Vaswani A, *et al.* Attention is all you need. In Adv Neural Inf Process Syst. 2017; 6000–60.
4. Alammar J. (2022). The Illustrated Stable Diffusion: Visualizing Machine Learning One Concept at a Time. [Online] Available from: https://jalammar.github.io/illustrated-stable-diffusion/.
5. Devlin J, *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. preprint arXiv: 1810.04805. 2018.
6. Hurst A, Lerer A, Goucher AP, Perelman A, Ramesh A, Clark A, Ostrow AJ, Welihinda A, Hayes A, Radford A, Mądry A. Gpt-4o system card. arXiv preprint arXiv:2410.21276. 2024 Oct 25.
7. Brown T, *et al*. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020; 33: 1877–1901.
8. Hugo Touvron, *et al*. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. 2023a.
9. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020; 21(140): 1–67.
10. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R. Gpt-4 technical report. arXiv preprint arXiv:2303.08774. 2023 Mar 15.
11. Hennings M. (2023 Nov 24). LoRA Fine-tuning & Hyperparameters Explained (in Plain English). [Online]. Retrieved on 5th March 2024 from https://www.entrypointai.com/blog/lora-fine-tuning/
12. Hu EJ, *et al.* LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685. 2021.
13. Dettmers T, *et al*. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv preprint arXiv:2305.14314. 2023.
14. Chaudhari S, *et al*. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. arXiv preprint arXiv:2404.08555. 2024.
15. Yao S, *et al*. Tree of Thoughts: Deliberate Problem Solving with Large Language Models, arXiv:2305.10601. 2023.
16. Silver D, *et al*. Mastering the game of Go with deep neural networks and tree search. Nature. 2016; 529(7587): 484–489.
17. Shuming M, *et al*. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. arXiv:2402.17764. 2024.
18. Gao Y, *et al*. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. 2023.
19. Xu Z, *et al*. Retrieval-augmented generation with knowledge graphs for customer service question answering. arXiv preprint arXiv:2404.17723. 2024.

 