ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2026.0594

A COMPREHENSIVE OVERVIEW OF LARGE LANGUAGE MODELS AND GPT FAMILY

P.S. Vinayagam
Department of Computer Science, Pondicherry University Community College, India

Abstract

The field of natural language processing (NLP) underwent a sea change with the introduction of large language models (LLMs). The NLP systems have progressed from simple rule-based systems to sophisticated transformer-based generative pre-trained models. These models have been pre-trained on vast amounts of data and encompass tens to hundreds of billions of parameters. They demonstrate excellent emergent abilities, including complex reasoning, instruction following and in-context learning. This paper presents a comprehensive overview of the LLMs with a detailed discussion of one of the most widely used LLM families, namely, the GPT model family. The architecture of the LLMs is explored to understand the nuances of the modern approach to natural language processing. The GPT family is discussed in detail to understand its rapid evolution over a short period. The various application domains of LLMs and the challenges posed by LLMs are enumerated. This paper aims to serve as a foundational resource for researchers and practitioners navigating the rapidly evolving field of large language models.

Keywords:

Large Language Models (LLMs), Transformers, Natural Language Processing, Pre-trained Models, GPT Family

1. INTRODUCTION

The transition from rule-based systems and statistical methods to machine learning and neural networks has revolutionized the field of Natural Language Processing (NLP). The field of NLP has undergone a generational change, transitioning from task-specific models to a single general-purpose model capable of performing multiple tasks via prompting [1]-[3]. The spark started from the Shannon's application of information theory to human language which influenced n-gram language models to predict natural language text [4]. The n-gram models used statistical frequency of word sequences. Over the years the field of language models has seen multiple eras of major breakthroughs.

The early NLP systems of pre-1990s were rule-based, which made use of hand-crafted linguistic rules and symbolic representations. These models were non-scalable and highly labour-intensive, lacking learning or generalization capabilities. Examples include Eliza, Parry, A.L.I.C.E. and SHRDLU, which had limited conversational or task-specific competence. They were less adaptive beyond the use of predefined rules and domains [5]-[8].

The statistical language models (1990s-2000s) used data-driven language modelling techniques. These models were extensively used in speech recognition and machine translation applications. Though these models were effective for local pattern modelling, they suffered from limited capacity to capture long-range dependencies and did not possess explicit semantic understanding. n-gram language models, Hidden Markov Models (HMMs), Maximum Entropy Models and Conditional Random Fields (CRFs) were the dominant models of the era [5], [9], [10], [8].

The sequential neural language models (2000s – 2020s) made use of neural networks and distributed representations to capture semantic similarity. They were able to model variable-length sequences. These models faced challenges related to vanishing gradients [11] and poor parallelization. They faced difficulty in modelling long-term dependencies which limited their performance on long-context tasks. The dominant architectures were Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), Gated Recurrent Units (GRUs) and word embedding methods such as Word2Vec and GloVe [5], [8], [12], [13]. The Google Neural Machine Translation (GNMT) system was the first large-scale neural language model introduced by Google in the year 2015. This model was trained on massive amounts of bilingual text data and it exhibited excellent performance on machine translation tasks [14].

Early neural language models were designed for specific tasks. The training data and the learned embedding spaces were also task-dependent. The next major milestone was the introduction of transformer-based models (from late 2010s). The transformer-based neural network architectures with self-attention [15] had numerous advantages. They were able to learn the long-term dependencies in language and benefitted from parallel training on multiple Graphics Processing Units (GPUs). This made it possible to train larger models which resulted in the availability of pre-trained language models (PLMs) during the late 2010s. The PLMs were capable of learning universal language representations. For training these models, extensive datasets without task-specific human annotation were used [16], [5]. Transformer-based architectures have become the dominant paradigm in natural language processing since the late 2010s. These architectures utilize scalable self-attention mechanisms to capture rich contextual representations and model long-range dependencies effectively. Notable examples include models such as BERT and GPT series. However, transformer models are associated with substantial computational and energy costs [5], [8].

The term Large Language Models (LLMs) is used to refer to the transformer-based models which encompass tens or even hundreds of billions of parameters which have been trained on massive text datasets. Some examples of LLM models include PaLM, LLaMA, DeepSeek and GPT-4. The difference between the PLMs and LLMs stem from the scaling up of size of LLMs. This extreme size provides enhanced capabilities in language understanding and provides previously unavailable functionalities [3], [17].

This paper presents a comprehensive overview of the advancements in the field of natural language processing with special emphasis on LLMs and GPT family. The rest of the paper is organized as follows: Section 2 presents the architecture of the LLMs. Section 3 details the GPT model family. The various application domains of LLMs are explored in Section 4. The

challenges posed by the LLMs are discussed in Section 5 followed by Section 6 which concludes the paper.

2. LARGE LANGUAGE MODELS (LLMS)

The large language models (LLMs) are Transformer-based language models that contain tens to hundreds of billions of parameters and have been trained using massive text data. The LLMs take advantage of scaling laws and emergent abilities such as in-context learning, instruction following and multi-step reasoning. The key techniques which contribute to the LLMs becoming general and capable learners are scaling, training, ability eliciting, alignment tuning and tools manipulation [3], [17]. LLMs are built upon the Transformer architecture which has revolutionized the field of natural language processing. The Transformer architecture is based upon a dual-structured encoder-decoder design [18]. The encoder transforms a sequence of input symbols into numerical representations or vectors. This is fed to the decoder to generate the output text [18]–[20]. Figure 1 shows the overall Transformer architecture [15].

The encoder has two important components, namely, self-attention mechanism, and feed-forward neural layers. These two layers are used by the model to extract meaningful and rich features from the input data. To make sure that information flows smoothly across multiple layers, residual connections and layer normalization are used. The decoder uses the extracted representations to generate the output text which is done step by step in an auto-regressive manner. That is, each generated word is used as input to predict the next word. To maintain causality and to avoid the model from seeing future words during generation, masking is applied in the decoder. This allows Transformers to achieve high accuracy along with parallel processing. This improves efficiency and scalability in the case of large and complex datasets [18].

The earlier models such as RNNs and LSTMs relied on sequential computation. Transformers do not rely on sequential computation. They are fully based on the self-attention mechanism, which has greatly improved their speed and scalability. The encoder has access to the entire input text, such as a sentence in machine translation, and processes all words at the same time to understand how they relate to each other. First, the encoder applies self-attention to determine the importance of each word with respect to others in the sequence. Then, a position-wise feed-forward neural network is applied to each word to extract higher-level features and capture deeper meaning. The residual connections and layer normalization avoid large changes in intermediate values [18].

The model stacks multiple encoder layers to produce context-aware representations for each word. Each token contains information about the whole input sentence. These representations are later used by the decoder to generate accurate and meaningful outputs. The decoder has the same structure as that of the encoder along with an additional cross-attention sublayer. The decoder generates the output sentence based on the input text and previously generated words. During generation, the decoder can only attend to earlier tokens using causal masking. This ensures that future words are not visible. This enforces auto-regressive generation, where each word is predicted using only past information.

The decoder uses cross-attention to refer back to the encoder's output. As in the case of encoder, in the decoder also residual connections and layer normalization are used for stabilizing the training. By way of stacking multiple decoder layers, the model generates fluent, grammatically correct and contextually meaningful text step by step [18]. LLMs are categorized based on their neural architectures as, encoder-only models or autoencoding models, encoder-decoder models or sequence-to-sequence (seq-to-seq) models, decoder-only models or autoregressive models, multimodal models and mixture of experts (MoE) models [21].

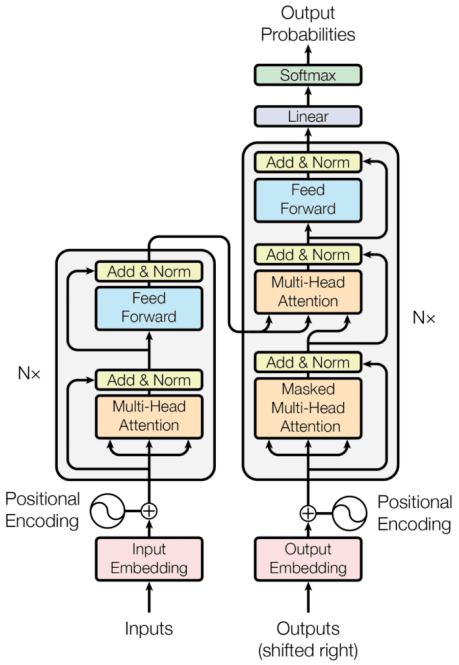

Fig.1. The Transformer Model Architecture [15]

2.1 ENCODER-ONLY MODELS OR AUTOENCODING MODELS

Encoder-only large language models use only the encoder part of the Transformer to understand sentences and capture relationships between words. These models are usually trained by masking some words in a sentence and asking the model to predict the missing words. This training approach is self-supervised and can be applied to very large text datasets. Examples of models that are encoder-only are BERT, ALBERT, RoBERTa, and ELECTRA. To solve specific downstream tasks, these models require an additional prediction layer. Encoder-only models work best for tasks that need a full understanding of the sentence, such as text classification and named entity recognition [8].

2.2 ENCODER–DECODER MODELS OR SEQUENCE-TO-SEQUENCE (SEQ-TO-SEQ) MODELS

Encoder–decoder models make use of both the encoder and decoder modules. The encoder converts the input sentence into hidden representations. The decoder takes care of generating the target output text. These models support flexible training strategies. The T5 (Text-to-Text Transfer Transformer) model is

trained by masking and predicting spans of text. The UL2 model combines multiple training objectives with different masking strategies. Examples of models under encoder-decoder category include T0 and ST-MoE. These models are well suited for tasks that require text generation based on context. They also support summarization, translation, and question answering. BART (Bidirectional and Auto-Regressive Transformers) is also based on the encoder-decoder model [8].

2.3 DECODER-ONLY MODELS OR AUTOREGRESSIVE MODELS

Decoder-only large language models make use of only the decoder component to generate output text. They are trained to predict the next-word. The model learns to predict the next token based on previous ones. Large decoder-only LLMs are capable of performing many tasks using only a few examples or simple instructions. They do not need extra prediction layers or fine-tuning. Examples of this model are GPT-2, GPT-3, GPT-4, PaLM, LLaMA and DeepSeek-V3 [8], [18]. Open-source models such as Alpaca and Vicuna are decoder-only derivatives.

2.4 MULTIMODAL MODELS

Multimodal models process and integrate multiple data modalities such as text, images, audio or video. Examples of this model are DeepSeek-VL and GPT-4o. These models fuse heterogeneous inputs thereby enabling complex cross-modal tasks such as visual question answering, image captioning, and cross-modal retrieval. The multimodal models are capable of understanding context and semantics across different data types [8].

2.5 MIXTURE OF EXPERTS (MOE) MODELS

In the conventional models all parameters are activated for every input. Whereas in MoE models numerous smaller expert sub-networks are managed by a router network. For each input token, the router dynamically selects a subset of experts to process it. This allows the model to maintain a very large parameter count but at the same time the inference costs are kept low. Examples of this model are Mistral's Mixtral 8×7B and Google's Gemini models. These model have reduced computational overhead compared to the dense models [8].

3. GPT MODEL FAMILY

Many LLM families have emerged recently and have attained wide-spread usage. One of the most notable and widely used LLM family, namely, GPT family lineage is discussed in this section.

3.1 GPT-1

After the introduction of the Transformer model by Google in 2017 [15], the OpenAI team adapted their language modelling to the new neural network architecture. GPT-1 was the first instance of the Generative Pre-trained Transformers (GPTs) paradigm released in the year 2018 by OpenAI [3]. This included unsupervised pre-training followed by supervised fine-tuning for NLP tasks. GPT-1 is an open-source model and is based on a decoder-only Transformer architecture with masked self-attention. This model has been trained using a next-token prediction technique on a large corpus of BooksCorpus text [22]. GPT-1 contains approximately 117 million parameters. This has exhibited good performance on tasks such as textual entailment, question answering, and sentiment classification. GPT-1 exhibited that a single pre-trained language model can handle multiple tasks. This is possible because of task-specific fine-tuning. This served as the foundational training paradigm for the future GPT models [3], [23].

3.2 GPT-2

GPT-2 exhibited scaling in both model size and training data. The model has 1.5 billion parameters and has been trained on a large, diverse WebText. GPT-2, released in 2019, has zero-shot generalization across a wide range of NLP tasks without explicit fine-tuning. GPT-2 is also based on the decoder-only transformer design. It utilizes deeper networks, wider hidden layers, and improved tokenization via byte pair encoding. GPT-2 exhibited fluent long-form text generation and task-agnostic capabilities such as translation and summarization from prompting [3], [23].

3.3 GPT-3

GPT-3 released in 2020 is yet another milestone in large language model research. It has 175 billion parameters [1]. It is an autoregressive model and it has been trained using licensed data, human-created data and publicly available web data. GPT-3 utilized few-shot or zero-shot and in-context learning capabilities. GPT-3 performed new tasks by conditioning on natural language instructions and a small number of examples provided directly in the prompt, without gradient-based fine-tuning. The performance of GPT-3 can be attributed to the benefits accrued from extreme scale. The performance of the model has been excellent in tasks such as translation, question answering, cloze tasks, arithmetic reasoning and domain adaptation. GPT-3 also catalyzed research into prompt engineering and task formulation as alternatives to traditional supervised learning [1], [23], [24].

3.4 CODEX

CODEX was introduced in the year 2021. It is a specialized descendant of GPT-3 model. It has been fine-tuned on large-scale code corpora which has been sourced primarily from public GitHub repositories [25]. This model is capable of translating natural language descriptions into executable code in multiple programming languages. Basically CODEX is a decoder-only transformer. But it has been adapted to handle programming language syntax, indentation, and long-range dependencies which are common in the case of source code. The model is capable of performing tasks such as code completion, bug fixing, and function generation. CODEX is used in GitHub Copilot. GitHub Copilot has exhibited good performance in providing software engineering assistance. CODEX is an example of performance improvement owing to domain-specific fine-tuning on large pre-trained language models [23].

3.5 INSTRUCTGPT

InstructGPT was released in the year 2022. GPT-3 had the issue of producing outputs that were not very helpful. At times the output were not in line with the user intent. To address this issue, InstructGPT was built on top of GPT-3. It uses Reinforcement

Learning from Human Feedback (RLHF) which combines supervised fine-tuning with preference-based reward modeling. Human annotators are used to provide demonstrations and rank model outputs. This was then used to optimize the model via reinforcement learning. InstructGPT and GPT-3 share the same architecture. But both differ in training methodology that has been used. InstructGPT had fewer parameters than GPT-3 models but this model was preferred by human evaluators. The use of RLHF influenced the design of conversational AI systems [23], [26].

3.6 WEBGPT

WebGPT is also built on top of GPT-3. It enables interaction with a text-based web browser. The model is trained to answer open-ended questions by issuing search queries, navigating web pages, and extracting relevant information. In this case also training combines imitation learning from human demonstrations with reward modeling and reinforcement learning. WebGPT integrates search and navigation commands in the language mode. The pre-trained models are limited by static knowledge. This model improves factual accuracy through information retrieval [23], [27].

3.7 GPT-3.5

GPT-3.5 released in 2022 is a family of intermediate models built upon GPT-3. GPT-3.5 takes advantage of additional training, refinement, and alignment techniques to improve performance. GPT-3.5 models demonstrate improved reasoning, reduced hallucination, and better instruction following. These models are the initial backbone for ChatGPT released in late 2022. GPT-3.5 also utilizes RLHF and this models benefits from large-scale conversational data which enables multi-turn dialogue. GPT-3.5 bridges the gap between GPT-3 and GPT-4 [23], [24].

3.8 CHATGPT

ChatGPT was released in November 2022. It is an application-level system built initially on GPT-3.5. Later models such as GPT-4 and GPT-5 are integrated as and when new models are introduced. The main feature for the popularity of ChatGPT can be attributed to its interactive dialogue ability. ChatGPT integrates conversational fine-tuning, safety filters and system-level orchestration for multi-turn interaction. Additionally features such as clarification requests and context retention are also made available. ChatGPT is based on the autoregressive Transformer architecture. It has been optimized for conversational coherence and user engagement. The release of ChatGPT brought LLMs into daily use by people from all walks of life. It is capable of summarization, tutoring, coding assistance and information retrieval [23], [24], [28].

3.9 GPT-4

GPT-4 brought multimodal capability within the GPT family. It was released in March 2023. It exhibited improvements in reasoning and robustness. OpenAI stopped release of architecture details starting from GPT-4. However, GPT-4 is pre-trained using next-token prediction and subsequently fine-tuned using RLHF. GPT-4 is trained on a larger and diverse dataset than the earlier models. It supports both text and image inputs. This model has exhibited good performance on academic, legal and professional benchmarks. Because of its multimodal design, this model is able to reason over both visual and textual information. This enables applications such as document understanding and visual question answering [23], [24], [29].

3.10 GPT-4 TURBO

GPT-4 Turbo is a more powerful, efficient and cost-effective version of the GPT-4 model. It was announced in November 2023. This model is capable of processing longer input sequences. This feature enables support for applications such as long-document analysis and codebase reasoning. This model reduces latency and cost while providing excellent performance. Though architecture details are not publicly available, but optimization and training refinements are decipherable [30].

3.11 GPT-4O AND GPT-4O MINI

GPT-4o (“omni”) introduced native multimodality and was announced in May 2024. This enabled processing of text, images and audio within a single model architecture. This model integrates modalities end-to-end which reduces latency and also improves cross-modal reasoning. Real-time voice interaction and visual understanding is supported by this model. The GPT-4o mini is a smaller and cost-efficient variant that has multimodal capability. But it supports high-throughput even in resource-constrained deployments [31].

3.12 REASONING-ORIENTED GPT MODELS (O-SERIES)

The o-series models released in September 2024 focus on multi-step reasoning and deliberative computation. They have been designed to allocate more computation to complex tasks. This greatly improves performance on mathematics, logic and scientific reasoning benchmarks. In the training the intermediate reasoning steps have been emphasized. Though these models are slower than other standard GPT models they provide improved correctness [32].

3.13 GPT-4.5

GPT-4.5 announced in 2024 is a refinement of GPT-4 model. Improvements are apparent in instruction following, conversational stability and factual accuracy. The main focus of this model is on usability and robustness across general tasks. Various evaluations suggest reduced hallucination and improved consistency over the earlier models [33].

3.14 GPT-5 AND GPT-5-SERIES MODELS

The GPT-5 series released in 2025 is based on system-level architectures that dynamically route tasks to specialized internal components. This series models support fast conversational responses, deep reasoning and professional workflows. GPT-5.1 and GPT-5.2 refine the basic architecture and show improvements in efficiency, reasoning depth, and long-context knowledge work [34].

The Table.1 presents the technical specifications of the entire GPT family. A context window is the maximum amount of data (text, code, images, or audio) a model can consider at one time, acting as its “working memory”.

Table. 1. GPT family comparison

Model	Release Year	Architecture	Parameters	Context Window	Training Data	Features
GPT-1	2018	Transformer Decoder	117M	512 tokens	BookCorpus (~5GB)	Unsupervised pre-training + task fine-tuning, basic text generation
GPT-2	2019	Transformer Decoder	1.5B	1024 tokens	WebText (~40GB)	Zero-shot task transfer, improved coherence
GPT-3	2020	Transformer Decoder	175B	2048 tokens (original)	CommonCrawl, WebText, books, Wikipedia (~570GB)	Few-shot learning, scale effects
GPT-3.5	2022	Transformer Decoder	Undisclosed	4096 tokens	Code, dialogue, web text	Instruction tuning, RLHF, Cost optimization
GPT-4	2023	Mixture of Experts (speculated)	Undisclosed	8K tokens	Multimodal (text+images)	Multimodal, improved reasoning, longer context
GPT-4 Turbo	2023	Optimized MoE	Undisclosed	128K tokens	Updated to April 2023	Knowledge cutoff update, cheaper API, lower latency, cost reduction
GPT-4o	2024	Multimodal architecture	Undisclosed	128K tokens	Audio, vision, text	Native multimodal, real-time audio, emotion detection
GPT-4.5	2024	Enhanced MoE	Undisclosed	128K+ tokens	Expanded multimodal	Improved reasoning, reliability, reduced hallucinations
GPT-5	2025	Next-gen architecture	Undisclosed	400K+ tokens	Comprehensive multimodal	AGI-level capabilities, complex problem-solving, true reasoning

Sources: [23], [28]–[34]

Measured in tokens, this capacity determines how much information the model retains within a single interaction before forgetting earlier inputs. Long-context models can process millions of tokens, equivalent to thousands of pages.

4. APPLICATIONS OF LARGE LANGUAGE MODELS

LLMs have shown remarkable efficiency in handling various tasks such as machine translation, text generation, text classification, text summarization, sentiment analysis, spam filtering and question answering. LLMs owing to these capabilities, find applications in various domains such as healthcare, education, finance and legal services [35].

4.1 HEALTHCARE

ChatGPT is assisting as an interactive tool for learning and problem-solving in the field of medical education. The performance of ChatGPT in the United States Medical Licensing Exam (USMLE) exceeded the passing threshold or was atleast comparable without requiring any specialized training [5], [36]. Various tools have been developed to assist medical practitioners. XrayGPT is one such tool which automatically analyses X-ray images and answers the questions of the patients [5]. Yet another useful tool in the field of medicine is segment anything (SAM) model by meta. It can be used to analyse a variety of medical images [5], [37]. For assisting in drug discovery tasks DrugGPT tool is used [5], [38].

4.2 EDUCATION

The impact of LLMs in the domain of education cannot be overstated. By way of enabling personalized learning experiences, providing constructive feedback for assessments and supplementing tutors by way of providing virtual tutoring, LLMs have brought in a revolutionary change in the field of education. Khanmigo by Khan Academy provides various specialized tools for both the tutors and the learners spanning varied subjects [8], [39]. To assist the tutors in the grading and feedback tasks, tools such as Gradescope AI are being widely used. These tools have the capability to even grade short-answers and provide constructive feedback [8], [40]. The availability of 24/7 virtual tutoring has made anytime and anywhere learning a reality. Duolingo's GPT-4-powered AI tutor assists learners in improving language skills with special emphasis on learning grammar in an interactive way [8], [41].

4.3 FINANCE

In the domain of finance, LLMs are being used for investment analysis and strategy, compliance and fraud detection, financial reporting and also as virtual assistants. To identify market trends, LLMs perform sentiment analysis on data available through various channels such as financial news and social media. Another area where LLMs are being used is for supporting algorithmic trading and portfolio management [8], [42].

4.4 LEGAL

Increasingly law firms have started using AI as an integral tool in their daily workflow. Chatlaw is an opensource legal language model which has surpassed GPT-4 in the Lawbench and Unified Qualification Exam for Legal Professionals by good margin. This

model has received highest scores in providing feedback on legal cases as evaluated on four dimensions, namely, completeness, correctness, guidance and authority, by legal experts [43]. Models are also available to support legal research, drafting and decision making.

5. CHALLENGES OF LARGE LANGUAGE MODELS

Even though LLMs are being extensively used in all the domains, they do present various challenges which needs to be addressed to maximize benefits while safeguarding human values. The major challenges such as bias and prejudice, hallucination and explainability have been discussed below. Other major challenges include high computational and energy costs, privacy and security issues, ethical issues, environmental impact and sustainability issues.

5.1 BIAS AND PREJUDICE

LLMs can inadvertently allow unfairness and prejudices. The vast datasets on which the LLMs have been trained may contains various biases, such as racial, gender or socioeconomic status. These biases may be inadvertently reproduced in the responses of the model. Any historical biases in the datasets may generate misleading information. The onus lies on the researchers to preprocess training data and identify inherent biases. The model outputs need to be regularly validated to avoid biases [18]. As chatbots are interactive, they tend to get influenced by the user input. If the prompts contain biased questions, it is likely that the language model generates biased responses. If the models are trained for a specific metric, the model may tend to develop algorithmic bias. If chatbots are presented with biased context, the model may generate biases responses owing to contextual bias [5].

5.2 HALLUCINATION

Owing to the model's training process, dataset and architectural design, the language models may hallucinate and provide factually incorrect responses in an attempt to fill in gaps in knowledge or context [5]. Hallucinations in LLMs may either be intrinsic or extrinsic. The intrinsic hallucinations distort the factual correctness and conflict with the source input. Whereas extrinsic hallucinations introduce speculative information not available in the source input [23]. The incorrect responses because of hallucination may prove to be a problem in sensitive applications. Methods, such as penalizing hallucinations or using factually correct data during training, have been proposed to mitigate the issue of hallucination [5].

5.3 EXPLAINABILITY

Understanding the decision-making process of LLMs utilizing billions of parameters is a herculean task. This lack of transparency results in the inability to understand how outputs are generated for a particular input. The vast amount of data from multiple sources may result in implicit biases and associations that are not interpretable. The very nature of the deep neural networks proves to be a challenge for explainability. This lack of transparency raises issues pertaining to accountability, trust and ethical considerations [5].

6. CONCLUSION

The advent of LLMs has transformed the entire landscape of natural language processing. Emergent abilities not available in the past have made inways in every walk of life. This paper provided a structured overview of the changes that the NLP field has undergone over the years starting from the rule-based systems to the transformer-based architecture models with special emphasis on the architecture of the LLM models. Of the many LLM model families, the GPT family has imbibed excellent qualities in its products over a short period of time. The entire lineage of the GPT family has been explored. Though LLMs have found applications in all domains, they pose serious challenges which needs to be addressed to provide safe solutions to maximize benefits while safeguarding human values. Nevertheless, as we progress towards achieving artificial general intelligence (AGI) these LLMs prove to be the stepping stones of success.

REFERENCES

[1] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, C. Wu, C. Winter and D. Amodei, "Language Models are Few-Shot Learners", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1-25, 2020.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, "Language Models are Unsupervised Multitask Learners", *OpenAI*, Vol. 2, pp. 1-24, 2019.

[3] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu and J.R. Wen, "A Survey of Large Language Models", *Proceedings of International Conference on Artificial Intelligence*, Vol. 13, pp. 1-144, 2023.

[4] E. Shannon, "Prediction and Entropy of Printed English", *Bell System Technical Journal*, Vol. 30, No. 1, pp. 50-64, 1951.

[5] M.U. Hadi, Q. Al Tashi and R. Qureshi, "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations and Future Prospects", *TechRxiv*, Vol. 8, pp. 1-57, 2025.

[6] E.D. Liddy, "Natural Language Processing", *Encyclopedia of Library and Information Science*, Vol. 2, pp. 1-16, 2001.

[7] J. Weizenbaum, "ELIZA-A Computer Program for the Study of Natural Language Communication between Man and Machine", *Communications of the ACM*, Vol. 9, No. 1, pp. 36-45, 1966.

[8] S.M. Mohammadabadi, B.C. Kara, C. Eyupoglu, C. Uzay, M.S. Tosun and O. Karakuş, "A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges and Societal Implications", *Electronics*, Vol. 14, No. 18, pp. 1-8, 2025.

[9] X. Liu and W.B. Croft, "Statistical Language Modeling", *Annual Review of Information Science and Technology*, Vol. 39, pp. 1-80, 2004.

[10] F. Jelinek, "*Statistical Methods for Speech Recognition*", 1998.
[11] S. Hochreiter, "Recurrent Neural Net Learning and Vanishing Gradient", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 6, No. 2, pp. 107-116, 1998.
[12] P. Azunre, "*Transfer Learning for Natural Language Processing*", 2021.
[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, Vol. 9, No. 8, pp. 1735-1780, 1997.
[14] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao and K. Macherey, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation", *Proceedings of International Conference on Machine Learning*, Vol. 88, pp. 1-23, 2016.
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All you Need", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 1-11, 2017.
[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf and M. Funtowicz, "Transformers: State-of-the-Art Natural Language Processing", *Proceedings of International Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Vol. 9, pp. 38-45, 2020.
[17] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E.H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean and W. Fedus, "Emergent Abilities of Large Language Models", *Proceedings of International Conference on Computation and Language*, Vol. 6, pp. 1-30, 2022.
[18] P. Peykani, F. Ramezanlou, C. Tanasescu and S. Ghanidel, "Large Language Models: A Structured Taxonomy and Review of Challenges, Limitations, Solutions and Future Directions", *Applied Sciences*, Vol. 15, pp. 1-12, 2025.
[19] H. Naveed, A.U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N.P. Barnes and A. Mian, "A Comprehensive Overview of Large Language Models", *ACM Transactions on Intelligent Systems and Technology*, Vol. 16, No. 5, pp. 1-72, 2025.
[20] W. Xu, W. Hu, F. Wu and S. Sengamedu, "DeTiME: Diffusion-Enhanced Topic Modeling using Encoder-Decoder based LLMs", *Association for Computational Linguistics: EMNLP 2023*, Vol. 6, pp. 9040-9057, 2023.
[21] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, "Unifying Large Language Models and Knowledge Graphs: A Roadmap", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 36, No. 7, pp. 3580-3599, 2024.
[22] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, "Improving Language Understanding by Generative Pre-Training", *Open AI*, Vol. 9, pp. 1-12, 2024.
[23] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain and J. Gao, "Large language Models: A Survey", *Proceedings of International Conference on Computation and Language*, Vol. 8, pp. 1-44, 2024.
[24] K.S. Kalyan, "A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4", *Natural Language Processing Journal*, Vol. 6, pp. 1-8, 2023.
[25] M. Chen, J. Tworek, H. Jun, Q. Yuan, H.P. De Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray and W. Zaremba, "Evaluating Large Language Models Trained on Code", *Proceedings of International Conference on Machine Learning*, Vol. 12, pp. 1-35, 2021.
[26] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P.F. Christiano, J. Leike and R. Lowe, "Training Language Models to Follow Instructions with Human Feedback", *Proceedings of Conference on Neural Information Processing Systems*, Vol. 23, pp. 1-15, 2022.
[27] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess and J. Schulman, "WebGPT: Browser-Assisted Question-Answering with Human Feedback", *Open AI*, Vol. 78, pp. 1-30, 2021.
[28] Open AI, "Introducing ChatGPT", Available at: https://openai.com/blog/chatgpt, Accessed in 2022.
[29] OpenAI, "GPT-4 Technical Report", Available at: https://openai.com/index/gpt-4-research/, Accessed in 2023.
[30] OpenAI, "GPT-4 Turbo Model Documentation", Available at: https://platform.openai.com/docs/models/gpt-4-turbo, Accessed in 2023.
[31] OpenAI, "GPT-4o System Card", Available at: https://openai.com/index/gpt-4o-system-card/, Accessed in 2024.
[32] OpenAI, "OpenAI o3 and o4-Mini System Card", Available at: https://openai.com/index/o3-o4-mini-system-card/, Accessed in 2025.
[33] OpenAI, "Introducing GPT-4.5", Available at: https://openai.com/index/introducing-gpt-4-5/, Accessed in 2025.
[34] OpenAI, "Introducing GPT-5", Available at: https://openai.com/index/introducing-gpt-5/, Accessed in 2025.
[35] A. Mohammed and R. Kora, "A Comprehensive Overview and Analysis of Large Language Models: Trends and Challenges", *IEEE Access*, Vol. 13, pp. 95851-95875, 2025.
[36] A. Gilson, C.W. Safranek, T. Huang, V. Socrates, L. Chi, R.A. Taylor and D. Chartash, "How does ChatGPT Perform on the United States Medical Licensing Examination? The implications of Large Language Models for Medical Education and Knowledge Assessment", *JMIR Medical Education*, Vol. 9, No. 1, pp. 1-11, 2023.
[37] J. Ma, Y. He, F. Li, L. Han, C. You and B. Wang, "Segment Anything in Medical Images", *Nature Communications*, Vol. 15, No. 1, pp. 1-7, 2024.
[38] Y. Li, C. Gao, X. Song, X. Wang, Y. Xu and S. Han, "DrugGPT: A GPT-based Strategy for Designing Potential Ligands Targeting Specific Proteins", *bioRxiv*, Vol. 3, pp. 1-30, 2023.
[39] S. Shetye, "An Evaluation of Khanmigo, a Generative AI Tool, as a Computer-Assisted Language Learning App", *Studies in Applied Linguistics and TESOL*, Vol. 24, pp. 1-11, 2024.

[40] J. Hidalgo-Reyes, J. Alvarez, L. Guevara-Chavez and Z.G. Cruz-Netro, "Gradescope as a Tool to Improve Assessment and Feedback in Engineering", *Proceedings of International Conference on Institute for the Future of Education*, Vol. 30, pp. 1-7, 2025.

[41] J. Vega, M. Rodriguez, E. Check, H. Moran and L. Loo, "Duolingo Evolution: From Automation to Artificial Intelligence", *Proceedings of International Conference on Applications of Computational Intelligence*, Vol. 24, pp. 54-71, 2025.

[42] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg and G. Mann, "BloombergGPT: A Large Language Model for Finance", *Proceedings of International Conference on Machine Learning*, Vol. 9, pp. 1-76, 2023.

[43] J. Cui, Z. Li, Y. Yan, B. Chen and L. Yuan, "ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases", *Proceedings of International Conference on Computation and Language*, Vol. 67, pp. 1-15, 2023.