# *The Evolution of Large Language Model: Models, Applications and Challenges*


Sindhu B
*Department of Computer Science and Engineering*
*Global Academy of Technology*
Bangalore, Karnataka, India
sindhubasavaraj13@gmail.com

Prathamesh R P
*Department of Computer Science and Engineering*
*Global Academy of Technology*
Bangalore, Karnataka, India
prathameshpatil1513@gmail.com

Sameera M B
*Department of Computer Science and Engineering*
*Global Academy of Technology*
Bangalore, Karnataka, India
mbsameera2001@gmail.com

KumaraSwamy S
*Department of Computer Science and Engineering*
*Global Academy of Technology*
Bangalore, Karnataka, India
skswamy@gat.ac.in



2024 International Conference on Current Trends in Advanced Computing (ICCTAC) | 979-8-3503-9580-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCTAC61556.2024.10581180



***Abstract*—**

**Large Language Models (LLMs) have attracted a lot of attention due to their success in natural language processing tasks. This paper provides a thorough overview by examining the architecture, applications, problems, assessment techniques, and future directions of LLM. With the constantly growing body of literature, a succinct yet comprehensive overview of recent developments is essential. Following the development of NLP, it highlights the move from rule-based systems to sophisticated transformer structures like as BERT and GPT. Important LLMs for text creation, translation, and summarization are mentioned, including T5, BART, and BioGPT. LLM performance is evaluated using metrics including as accuracy, perplexity, BLEU score, and ROUGE score. Research is still being done because of issues with bias, overfitting, and real-time processing. Future directions include managing longer contexts, lowering bias, and increasing efficiency through methods like federated learning. Continuous learning and multimodal LLMs are promising fields, as well as interpretive AI. In conclusion, LLMs have transformed natural language processing (NLP) and brought up both technical and ethical issues about the future of AI.**



**Keywords—Large Language Models (LLMs), Natural Language Processing (NLP), Context length improvements, Fine-tuning, Multi-modal LLMs.**


## I. INTRODUCTION

Language is a vital tool for human expression and communication, developing from infancy and evolving throughout life [1, 2]. However, machines lack the innate ability to understand and communicate in human language without advanced artificial intelligence (AI) [3]. Achieving human-like language skills in machines has been a longstanding scientific goal [4].

Advances in deep learning, vast computer resources, and abundant training data have led to the emergence of large language models (LLMs) [5]. LLMs, utilizing neural networks with billions of parameters, are trained on massive unlabeled text datasets through self-supervised learning [5]. They represent a significant advancement in natural language processing (NLP) and AI [6], excelling in various language tasks like text synthesis, translation, summarization, question-answering, and sentiment analysis [7].

LLMs trace their roots to earlier language model and neural network developments [8]. Statistical approaches like n-gram models were early attempts, but they lacked in expressing long-term interdependence and context [8]. The introduction of Recurrent Neural Networks (RNNs) was crucial for sequential data modeling, but they faced limitations due to vanishing gradients and long-term dependencies [9]. The transformer architecture, with its self-attention mechanism, addressed these limitations and served as the foundation for advanced LLMs like Google's BERT and OpenAI's GPT series [10, 11].

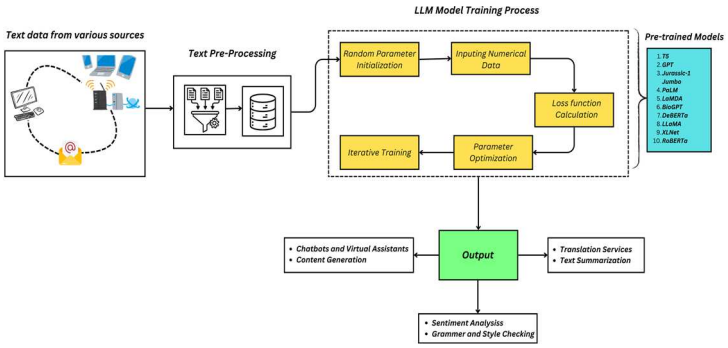


Figure 1 Pipeline of the LLM training phase.

The basic architecture of large language models (LLMs), depicted in Figure 1., involves receiving text data from various sources, preprocessing it, and undergoing training stages including parameter initialization, numerical input, loss calculation, optimization, and iterative training [12, 13]. LLMs have demonstrated potential in numerous natural language processing (NLP) tasks, including specialized domains like medicine and politics [14]. The development of sophisticated models like GPT, LLaMa, and Bard, as well as explorations into capabilities such as Alpaca and GPTHuggingface, have made LLMs a pivotal domain [15, 16]. However, comprehensive reviews of recent LLM developments and their limitations are lacking, with a scarcity of peer-reviewed articles focusing on technical complexities, taxonomy, architectures, API applications, domain-specific uses, and societal impacts [17, 18]. This paper aims to fill this gap by exploring and evaluating LLMs across domains, classifications, architectures of pre-trained

models, resources, and real-world applications [19]. It addresses open challenges including safety, ethics, privacy, economics, and environmental concerns, offering guidelines for future research and development in LLM utilization [20]. The contributions of this paper include an exhaustive overview of LLMs, comparative analysis of pre-trained model architectures, insights into their impact on society, and applications across various domains such as biomedical, healthcare, education, social media, business, and agriculture [21].

## II. History of LLM through the years

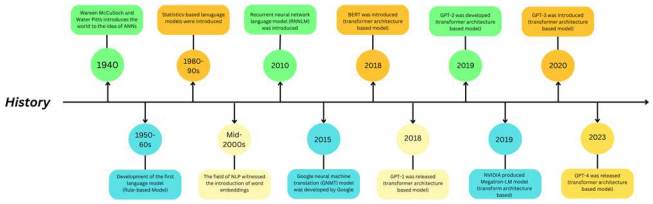


Figure 2 Development of LLMs through the years.

Language models have undergone a fascinating journey of evolution, paving the way for significant advancements in artificial intelligence (AI) and natural language processing (NLP). Warren McCulloch and Walter Pitts introduced the concept of artificial neural networks (ANNs) in the 1940s, laying the groundwork for future developments [22]. Subsequently, in the 1950s and 1960s, the first language models emerged, incorporating early neural networks and rule-based systems, which relied on established linguistic rules and features for language processing [23, 24].

The 1980s and 1990s witnessed the rise of statistics-based language models, which utilized probabilistic techniques to capture patterns and correlations within language data, surpassing earlier models in accuracy and data processing capabilities [25]. A significant breakthrough came in the mid-2000s with the introduction of word embeddings, which represented words in a vector space, capturing semantic relationships and contributing to advancements in NLP [26, 27].

In the mid-2010s, neural language models like the recurrent neural network language model (RNNLM) emerged, aiming to capture sequential dependencies in textual data and improve contextual understanding [28, 29]. Google's introduction of the Google Neural Machine Translation (GNMT) model in 2015 marked a milestone in machine translation, enhancing translation accuracy and meaningfulness [32, 33].

The landscape of language modeling underwent a transformative shift with the introduction of the transformer architecture in 2017 [35]. Models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) leveraged transformer architecture to overcome limitations of earlier models like RNNs and LSTM networks [36]. Google's BERT, introduced in 2018, addressed contextual understanding constraints by utilizing deep bidirectional representations across all layers [37, 38]. OpenAI's GPT series, starting with GPT-1 in 2018, showcased the transformative potential of transformers in advancing NLP tasks [39].

Subsequent iterations like GPT-2, developed in 2019, and GPT-3, introduced in 2020, further pushed the boundaries of language modeling with their vast parameter counts and superior text generation capabilities [40]. GPT-3, with its staggering 175 billion parameters, represented a significant leap in generating coherent and natural language text [18]. OpenAI continued this trajectory with the release of GPT-4 in 2023, further enhancing text generation capabilities and expanding into analyzing both textual and visual data [21].

These advancements in large language models offer promising opportunities for innovation and experimentation across various domains, including healthcare, education, and research, reshaping interactions with AI and NLP technologies [18]. The journey of language models reflects a continuous quest for enhancing our understanding and utilization of human language in AI systems

## III. Literature Survey

The literature review on Large Language Models (LLMs) includes a thorough examination of articles spanning various domains and applications.

[2] The study mentions the use of LangChain and LLM Model to develop a PDF chatbot. LangChain is a platform that simplifies the construction of chatbots and scalable AI/LLM applications. The LLM Model is a big language model that can generate text, translate languages, develop unique material, and deliver helpful replies. The document also references the use of Pinecone to store vectors of PDF files and React JS for the front-end building of a webpage to communicate with the chatbot.

[4] The suggested approach attempts to increase the accuracy and quality of job suggestions by utilizing large language models (LLMs) and generative adversarial networks (GANs). The research addresses the limits of solely using LLMs for resume completion and recommends the utilization of users' interactive actions to boost the accuracy of resume generation. It also proposes a GAN-based strategy to modify the representations of low-quality resumes for better recommendation outcomes.

[5] The literature review in this paper addresses the development and applications of Large Language Models (LLMs) in natural language processing, specifically in the context of retrieving information in conventional vehicle manuals. The system study investigates several LLM-based ways for increasing access to information within vehicle manuals, considering criteria such as response accuracy, cost effectiveness, and user experience.

[12] This paper offers an extensive analysis of Automatic Text Summarization (ATS) with an emphasis on practical applications utilizing a "Process-Oriented Schema" perspective. Additionally, it explores the latest developments in ATS using Large Language Models (LLMs), with the goal of filling in the gaps in the literature from the previous two years.

[15] This paper offers a new agent structure for resume screening based on LLM, which increases the effectiveness of hiring procedures. It presents agents that use LLMs to make decisions, greatly increasing classification accuracy and processing speed. Analyses on actual resume data show an 11-fold speedup over manual approaches, as well as improved results in the summary and grading stages.

Analysis demonstrates the effectiveness of LLM agents during the decision-making process and indicates how they can change resumes.

Figure 3., represents the number of papers we have used as reference in this research over years.

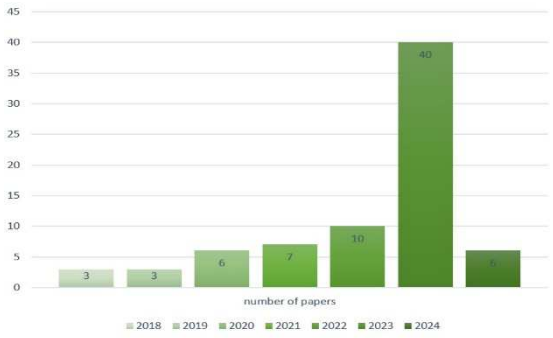


Figure 3 Number of Papers over years

## IV. LLM ARCHITECTURE OVERVIEW

The architecture of a Large Language Model (LLM) is influenced by several aspects, such as the purpose of the particular model design, the computational resources at hand, and the types of language processing tasks that the LLM is expected to do. Several layers, including feed forward layers, embedding layers, and attention layers, make to the general architecture of LLM. A text which is embedded inside is collaborated together to generate predictions.

- *Transformer-Based LLM Model Architectures*

Transformer-based models, which have transformed activities related to natural language processing, generally adhere to a general architecture as in Figure 4., including of the following elements:

**Input Embeddings:** Every token in the input text is embedded into a continuous vector representation after it has been tokenized into smaller units like words or subwords. The input's syntactic and semantic content is captured in this embedding step.

**Positional Encoding:** Since transformers do not inherently encode the order of the tokens, positional encoding is applied to the input embeddings to provide information about the positions of the tokens. This makes it possible for the model to handle the tokens while accounting for their sequential order.

**Encoder:** The encoder, which protects the context and meaning of text data, analyzes the input text using a neural network technique to generate several hidden states. The transformer architecture consists of several encoder layers. [42]The two main constituents of every encoder layer are the feed-forward neural network and the self-attention mechanism.

**Self-Attention Mechanism:** By calculating attention scores, self-attention allows the model to assess the relative relevance of various tokens in the input sequence. It enables the model to take into account, in a context-aware manner, the dependencies and relationships among various tokens.

**Feed-Forward Neural Network:** Each token is subjected to a feed-forward neural network separately following the self-attention stage. With its completely connected layers and non-linear activation functions, this network enables the model to represent intricate relationships between tokens.

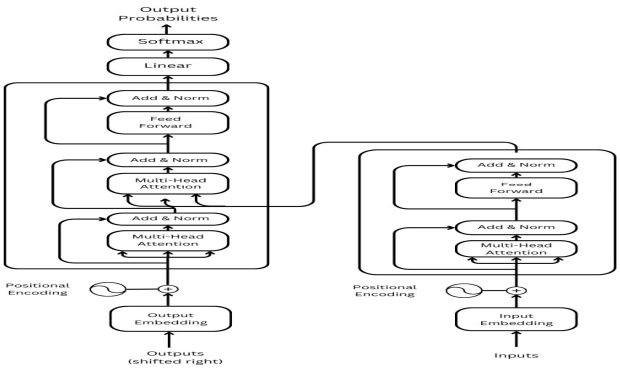


Figure 4 Transformer Model Architecture

**Decoder Layers:** Some transformer-based models come with both an encoder and a decoder component. By focusing on the previously created tokens, the model can produce sequential outputs through autoregressive generation, which is made possible by the decoder layers.

**Multi-Head Attention:** Transformers frequently use multi-head attention, which is the simultaneous application of many learned attention weights to self-attention. As a result, the model may simultaneously handle different portions of the input sequence and capture different kinds of interactions.

**Layer normalizing:** In the transformer architecture, layer normalizing is applied following each sub-component or layer. It enhances the learning process' stability and the model's capacity to generalize across various inputs.

Output Layers: Depending on the particular task at hand, the transformer model's output layers may change. For instance, in language modeling, the probability distribution over the next token is typically generated using a linear projection with SoftMax activation.

- *Scaling laws*

Model scaling involves expanding the size of a LLM by increasing its parameter count, with the goal of boosting its learning and generalisation capabilities[45]. The findings from the research highlight a significant advantage of model scaling in enhancing the fine-tuning performance of LLMs

## V. PRE-TRAINED LLM MODELS

Pretrained language models are vital for natural language processing since they can capture broad language understanding and generating skills from various text sources. They provide significant advantages by reducing processing resources and data needed for fine-tuning specific activities. Table 1., shows some of the most prominent pre-trained LLM models.

Each model is meticulously categorized based on its unique features, datasets utilized for pre-training, fine-tuning capabilities, and key applications. Ranging from Google AI's adaptable text-to-text model, T5, to OpenAI's revolutionary GPT series, and incorporating specialty LLMs like BioGPT for biomedical text analysis[43,44]. Additionally, it highlights models such as BART, which amalgamate qualities from diverse architectures for tasks like text summarization. Furthermore, it contains developing models like BLOOMZ, emphasizing their potential for real-world applications and developments in the field.

**Table 1 Overview of Pretrained Language Models**

| Model Name | Description | Key Features | Training Data | Fine-Tuning Data | Fine-Tuning Tasks | Applications |
|---|---|---|---|---|---|---|
| T5 (Text-to-Text Transfer Transformer) | Versatile text-to-text model by Google AI | Unified framework, extensive pre-training | Internet text data | Task-specific datasets | Text classification, translation, summarization | Language translation, summarization |
| GPT (Generative Pre-trained Transformer) | OpenAI's versatile LLM for various NLP tasks | Extensive pre-training, deep language understanding | Internet text data | Custom datasets | Text generation, translation, Q&A, and more | Chatbots, content generation, NLP domains |
| Jurassic-1 Jumbo | Open-source LLM by AI21 Labs, excels in factual language understanding | Strong factual language capabilities, question answering | Web text and code | Task-specific datasets | Question answering, factual search | Research, information retrieval |
| PaLM (Pathway Language Model) | Google AI's powerful LLM with understanding of code and instructions | Multimodal and multilingual capabilities | Massive dataset of text, code, and other modalities | Task-specific datasets | Diverse NLP tasks, code understanding | Research, generative tasks |
| LaMDA (Language Model for Dialogue Applications) | Google AI's conversational LLM focused on natural dialogue | Dialogue fluency and coherence | Dialogue data | Dialogue datasets | Dialogue modeling, chatbot development | Conversational AI assistants, chatbots |
| BioGPT | Specialized biomedical LLM with state-of-the-art results | Biomedical literature pretraining, excels in biomedical tasks | Biomedical literature | Biomedical datasets | Biomedical text analysis | Biomedical text analysis, research |
| BART (**B**idirectional and **A**utoregressive Transformer) | Combines strengths of BERT and GPT-2 for text summarization | Strong summarization capabilities | Web text | Task-specific datasets | Text summarization, translation, question answering | Summarization of news articles, scientific papers |
| DeBERTa | Improves upon RoBERTa with masked language modeling and entity masking | Improved performance on various NLP tasks | Book Corpus, English Wikipedia, entity mentions | Task-specific datasets | Various NLP tasks | Further improvement on NLP benchmarks |
| LLaMA (Language Models for Large Applications) | Meta AI's open-source LLM with focus on efficiency | Designed for large-scale deployment | Web text | Task-specific datasets | Diverse NLP tasks | Research, real-world applications |
| GPT-NeoX-20B | Open-source LLM by Eleuther AI, successor to GPT-Neo | Impressive generation capabilities, large model size | Web text | Task-specific datasets | Text generation, summarization, code generation | Creative writing, code completion |
| Megatron-Turing NLG | Highly-parameterized LLM by NVIDIA and Google AI | Impressive text generation abilities | Web text | Task-specific datasets | Text generation, summarization, code generation | Creative writing, code completion |
| BLOOMZ (Big-Learning Open-source Open Multimodal Information Network) | Open-source LLM by AI for Good Foundation | Focuses on factual language understanding and multilingual capabilities | Text and code in multiple languages | Task-specific datasets | Diverse NLP tasks | Research, social good applications |
| XLNet | Combines autoregressive pre-training with bidirectional context learning | Bidirectional context learning, versatile approach | Internet text data | Task-specific datasets | Diverse NLP tasks | Research, applications |
| RoBERTa | BERT-based model with refined hyperparameters | Significance of design decisions, publicly available, top-tier NLP results | Book Corpus, Wikipedia | Task-specific datasets | Various NLP tasks | Benchmark improvements, research |
| Speech XLNet | Unsupervised acoustic model withrobust regu larization. | Robust regularizer, improved recognition accuracy | Speech datasets | TIMIT, WSJ datasets | Speech recognition | Speech recognition systems |

## VI. APPLICATIONS OF LARGE LANGUAGE MODELS(LLMS)

The application of Large Language Models (LLMs) spans across diverse sectors, showcasing their adaptability and transformative potential. These models, adept at comprehending and generating human-like text, are increasingly prevalent in various industries and research communities [46]. From medicine to finance, education robotics, LLMs are revolutionizing traditional processes and enabling novel solutions.

**Medicine:** LLMs revolutionize healthcare by offering evidence-based treatment recommendations, powering chatbots for patient interaction, aiding medical research with data analysis, and enhancing medical education through personalized training materials[41]. They also contribute to public health initiatives by analyzing media data for disease outbreak detection and sentiment monitoring.

**Education:** LLMs personalize learning experiences by providing customized study materials and practice questions for students. They streamline tasks for teachers by aiding in lesson planning, grading assignments, and generating diverse educational content. LLMs also improve accessibility by supporting students with disabilities and facilitating language learning.

**Science:** LLMs expedite scientific research by analyzing and summarizing scientific literature, aiding in hypothesis generation, and assisting in scientific writing. They enhance interdisciplinary collaboration by translating complex concepts for non-specialists, facilitating effective communication among researchers[47,50].

**Mathematics:** LLMs provide step-by-step explanations for solving mathematical problems, assist in proof verification, and bridge the gap between theoretical mathematics and applied contexts. They enhance accessibility by translating complex concepts for non-specialists, facilitating interdisciplinary collaboration and knowledge dissemination.

**Finance:** LLMs enable applications such as robo-advising, algorithmic trading, and low-code solutions in financial services. [48]They improve decision-making processes, facilitate transparency, and contribute to accessibility by providing open-source resources for novel applications and customization.

**Robotics:** LLMs enhance human-robot interaction, assist in task planning, motion planning, navigation, and object manipulation, and facilitate continuous learning in robots. They contribute to the advancement of robotics research and the development of intelligent and adaptive robotic systems.

## VII. PERFORMANCE EVALUATION METHODS

Evaluating the performance of large language models (LLMs) is crucial for understanding their capabilities and limitations across various tasks and applications. As LLMs continue to gain prominence in both research and industry settings, it becomes imperative to employ robust evaluation methods to ensure their effectiveness and reliability.

The development of sophisticated Large Language Models (LLMs) like the GPT series and Retrieval-Augmented production (RAG) models is a major step toward the goal of data-driven decision-making and human-like text production in the field of artificial intelligence. Understanding and implementing advanced assessment metrics is essential for professionals involved in the development or use of these technologies[49,52], especially data scientists and quantitative researchers, as it helps with both model validation and performance optimization in complex settings.

- **Accuracy, Precision, and Recall:**

Accuracy in the context of LLMs and RAG models is often more complex than a straightforward ratio due to the nuances of language and the model's purpose. [51] It's essential in tasks like classification and question answering where responses can be distinctly validated against a ground truth. Precision and Recall are critical in environments where the consequence of false positives and false negatives varies. For example, in a legal or financial context, the precision of a model might be prioritized to avoid costly errors.

**Formulas:**

$$Accuracy = \frac{(True\ Positives + True\ Negatives)}{Total\ Number\ of\ Samples} \quad (1)$$

$$Precision = \frac{True\ Positives}{(True\ Positives + False\ Positives)} \quad (2)$$

$$Recall = \frac{True\ Positives}{(True\ Positives + False\ Positives} \quad (3)$$

- **Perplexity and Cross-Entropy**

Perplexity provides a measure of a model's certainty in its output, with lower values indicating better performance. It is directly derived from Cross-Entropy [54,56], which measures the dissimilarity between the predicted probability distribution and the actual distribution in the test data. These metrics are crucial for models that generate or complete text, as they quantify how well the model predicts a sample.

**Formulas:**

$$\text{Cross} - \text{Entropy} = -\sum x\, p(x) \log q(x) \quad (4)$$

$x$ : Represents a possible sample in the data.
$p(x)$: Denotes the actual probability of sample x occurring in the data distribution.
$q(x)$: Represents the probability of sample x predicted by the model

$$Perplexity = 2 - Cross - Entropy \quad (5)$$

- **BLEU**

BLEU score is short for Bilingual Evaluation Understudy has been developed for evaluating the performance of text translation models[53].

**Formula:**

$$\boldsymbol{BLEU = BP * (\sum \eta = 1\, N(w\_n * \log(p\_n)))} \quad (6)$$

$\boldsymbol{BP}$ (Brevity Penalty) **:** This term penalizes short translated sentences to avoid favoring shorter outputs over longer, more informative ones. It is calculated using another formula that considers the ratio between the reference and candidate text lengths.

$\sum$ (Sigma): This symbol represents a summation operation.

$\boldsymbol{\eta}$ :iterates over n-grams (sequences of n words) from 1 to N.

$\boldsymbol{N}$**:** This denotes the maximum n-gram size considered in the calculation. Typically, N ranges from 1 to 4.

$\boldsymbol{w\_n}$**:** This represents the weight assigned to the n-gram precision score. By default, all weights are set to be equal $(\boldsymbol{w\_n} = {}^{1}/_{\boldsymbol{N}})$ (7)

$\boldsymbol{p\_n}$**:** This refers to the modified n-gram precision, which is calculated as the ratio between the number of n-grams in the candidate text that match perfectly with an n-gram in the reference text, clipped at the maximum number of n-grams that could possibly occur in the reference text.

The BLEU score combines two key factors:

**N-gram precision:** This captures how well the model-generated text matches the reference text in terms of n-gram overlap[60].
**Brevity penalty:** This discourages shorter outputs that might be easier for the model to generate but lack content.The exponential term in the formula amplifies the impact of lower n-gram precision scores[73,75]. This means that a significant decrease in precision for any n-gram level will have a stronger negative influence on the overall BLEU score.

**Limitations of BLEU Score:**
**Sensitivity to Sentence Order:** BLEU score doesn't consider word order within the n-grams, so rearranging words in a grammatically correct way can significantly affect the score[57].
**Bias Towards Short Sentences:** The brevity penalty might not always accurately capture how informative a shorter sentence can be[55].

- **ROUGE**

The objective of Rouge score is the evaluation of summarization models. Its intuition is pretty simple[74]. The rouge score does not care about the word order and only focuses on matching words between two texts. There are various ROUGE metrics, such as ROUGE-N, ROUGE-L:
**ROUGE-N (N-gram Overlap):**
ROUGE-N measures the overlap of n-grams (contiguous sequences of n words) between the candidate (generated) text and the reference (human-generated) text[72].
It computes precision, recall, and F1-score for unigrams (ROUGE-1), bigrams (ROUGE-2), trigrams (ROUGE-3)

The precision, recall, and F1-score for ROUGE-N are calculated as follows:

$$\text{Precision} = \frac{\text{Number of overlapping n-grams in candidate and reference}}{\text{number of n-grams in candidate}} \quad (8)$$

$$\text{Recall} = \frac{\text{Number of overlapping n-grams in candidate and reference}}{\text{Number of n-grams in reference}} \quad (9)$$

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

**ROUGE-L (Longest Common Subsequence):**
ROUGE-L measures the longest common subsequence (LCS) between the candidate and reference texts. The LCS is the longest sequence of words that appear in the same order in both texts[70].

Precision, recall, and F1-score for ROUGE-L are calculated based on the length of the LCS:

$$Precision = \frac{Length\ of\ LCS}{Number\ of\ words\ in\ candidate} \quad (11)$$

$$Recall = \frac{Length\ of\ LCS}{Number\ of\ words\ in\ reference} \quad (12)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (13)$$

## VIII. CHALLENGES AND FUTURE DIRECTIONS

LLMs have rapidly developed from non-existent to prevalent in machine learning within a few years. Their capacity to generate human-like text has received attention and applications in various sectors. However, this rapid rise in prominence has also revealed many challenges that must be addressed to unlock the full potential of LLMs[[59,71].
**Bias and Fairness:** LLMs trained on biased data can perpetuate those biases. To ensure fair and inclusive outcomes, researchers are exploring techniques to debias training data and develop fairer evaluation metrics.
**Overfitting and Memorization vs. Generalization:** LLMs can become overly reliant on specific patterns in their training data, hindering their ability to handle new situations. Finding the right balance between memorization and generalization is crucial.
**Hallucinations:** LLMs can generate outputs that seem plausible but are factually incorrect. Techniques to improve the factual grounding of LLM responses are being investigated[61].
**Prompt Engineering:** The way humans prompt LLMs significantly influences the output. Developing effective prompt engineering methods is essential for guiding LLMs towards desired responses.
**Limited Knowledge and Outdated Information:** Keeping LLM knowledge up-to-date is challenging. Techniques like retrieval-augmented generation are being explored to address this issue[68].
**Real-Time Processing:** The large size and complexity of LLMs make real-time processing difficult, especially for mobile applications[62,63]. Hardware acceleration and model optimization techniques are being developed to address this.
**Long-Term Dependencies:** LLMs can struggle with maintaining context and handling long-term dependencies in complex tasks. Research into improving their ability to follow long sequences and conversations is ongoing[67].

The future of LLMs is being shaped by exciting research. This includes reducing bias in training data, making LLMs more efficient with techniques like federated learning, and enabling them to handle longer contexts in conversations[64,65]. Researchers are also working on continuous learning for LLMs to stay up-to-date and interpretable AI for users to trust AI decisions. Multimodal LLMs that process different data types and foster human-AI collaboration are another promising area. Finally, research on dynamic evaluation metrics, personalization, and ethical frameworks is crucial for ensuring LLMs are beneficial to society[69].

## IX. CONCLUSION

In conclusion, this review paper extensively covers the landscape of Large Language Models (LLMs) by digging into many elements of their architecture, applications, performance evaluation, ethical considerations, recent breakthroughs, problems, and future directions. Beginning with an overview of Natural Language Processing (NLP) and the evolution of LLMs, we clarified the significance and wide-ranging uses of this model. The study identified

important LLMs, including BERT, GPT, XLNet, and T5, explaining their key characteristics and contributions to the field. Evaluation criteria and comparison with existing NLP models underscored the better performance and scalability of LLMs. Ethical factors, including as bias mitigation and responsible use, stressed the significance of ethical AI governance in LLM development and deployment. Recent breakthroughs demonstrated the tremendous growth in LLM technology, while obstacles and future prospects outlined avenues for further study and innovation. In summary, LLMs have transformed NLP, enabling unparalleled capabilities in understanding and generating natural language, while presenting crucial ethical and technical questions for the future of AI.

## REFERENCES


[1] Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). Transformers:"the end of history" for natural language processing?. In Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21 (pp. 677-693). Springer International Publishing.

[2] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, *32*.

[3] Adiwardana, D., Luong, M. T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., ... & Le, Q. V. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

[4] y Arcas, B. A. (2022). Do large language models understand us?. *Daedalus*, *151*(2), 183-197.

[5] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[8] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. NAACL-HLT. *arXiv*.

[9] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1-67.

[11] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

[12] Zhang, Z., Gu, Y., Han, X., Chen, S., Xiao, C., Sun, Z., ... & Sun, M. (2021). Cpm-2: Large-scale cost-effective pre-trained language models. *AI Open*, *2*, 216-224.

[13] Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Al-Shaibani, M. S. (2023). Bloom: A 176b-parameter open-access multilingual language model.

[14] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... & Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

[15] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, *24*(240), 1-113.

[16] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1-53.

[17] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

[18] Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., ... & Khashabi, D. (2022). Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

[19] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2022). Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

[20] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

[21] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[22] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

[23] Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526-1541.

[24] Boiko, D. A., MacKnight, R., & Gomes, G. (2023). Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.

[25] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. (2022). Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

[26] Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... & Florence, P. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

[27] Parisi, A., Zhao, Y., & Fiedel, N. (2022). Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*.

[28] Zhang, B., & Soh, H. (2023, October). Large language models as zero-shot human models for human-robot interaction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 7961-7968). IEEE.

[29] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., ... & Huang, F. (2023). mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

[30] Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., ... & Dai, J. (2024). Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, *36*.

[31] Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., & Shan, Y. (2024). Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, *36*.

[32] Saravia, E. (2022). Prompt engineering guide. *GitHub. URL: https://github. com/dair-ai/Prompt-Engineering-Guide [accessed 2023-06-01]*.

[33] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., ... & Tang, J. (2022). Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

[34] Wang, Y., Le, H., Gotmare, A. D., Bui, N. D., Li, J., & Hoi, S. C. (2023). Codet5+: Open code large language models for code understanding and generation. *arXiv preprint arXiv:2305.07922*.

[35] Wang, S., Sun, Y., Xiang, Y., Wu, Z., Ding, S., Gong, W., ... & Wang, H. (2021). Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2112.12731*.

[36] Rasley, J., Rajbhandari, S., Ruwase, O., & He, Y. (2020, August). Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp.3505-3506).

[37] Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020, November). Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* (pp. 1-16). IEEE.

[38] He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., & Neubig, G. (2021). Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*.

[39] Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E. P., Bing, L., ... & Lee, R. K. W. (2023). Llm-adapters: An adapter family for parameter-

efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.
[40] Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.
[41] Wu, S., Fei, H., Qu, L., Ji, W., & Chua, T. S. (2023). Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519.
[42] Pesaru, A., Gill, T. S., & Tangella, A. R. (2023). AI assistant for document management Using Lang Chain and Pinecone. International Research Journal of Modernization in Engineering Technology and Science.
[43] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
[44] Du, Y., Luo, D., Yan, R., Liu, H., Song, Y., Zhu, H., & Zhang, J. (2023). Enhancing job recommendation through llm-based generative adversarial networks. arXiv preprint arXiv:2307.10747.
[45] Medeiros, T., Medeiros, M., Azevedo, M., Silva, M., Silva, I., & Costa, D. G. (2023). Analysis of language-model-powered chatbots for query resolution in pdf-based automotive manuals. Vehicles, 5(4), 1384-1399.
[46] Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2023). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. Authorea Preprints.
[47] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. Authorea Preprints.
[48] Antu, S. A., Chen, H., & Richards, C. K. (2023). Using LLM (Large Language Model) to Improve Efficiency in Literature Review for Undergraduate Research.
[49] Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. AI and Ethics, 1-31.
[50] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
[51] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.
[52] Jin, H., Zhang, Y., Meng, D., Wang, J., & Tan, J. (2024). A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. arXiv preprint arXiv:2403.02901.
[53] Vizcarra, J., Haruta, S., & Kurokawa, M. (2024, February). Representing the Interaction between Users and Products via LLM-assisted Knowledge Graph Construction. In 2024 IEEE 18th International Conference on Semantic Computing (ICSC) (pp. 231-232). IEEE.
[54] Zeng, F., Gan, W., Wang, Y., & Philip, S. Y. (2023, December). Distributed training of large language models. In 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS) (pp. 840-847). IEEE.
[55] Gan, C., Zhang, Q., & Mori, T. (2024). Application of llm agents in recruitment: A novel framework for resume screening. arXiv preprint arXiv:2401.08315.
[56] Wang, H., & Li, Y. F. (2023, October). Large Language Model Empowered by Domain-Specific Knowledge Base for Industrial Equipment Operation and Maintenance. In 2023 5th International Conference on System Reliability and Safety Engineering (SRSE) (pp. 474-479). IEEE.
[57] Arnautov, K. V., & Akimov, D. A. (2024, January). Application of Large Language Models for Optimization of Electric Power System States. In 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon) (pp. 314-317). IEEE.
[58] Yu, S., Fang, C., Ling, Y., Wu, C., & Chen, Z. (2023, October). LLM for test script generation and migration: Challenges, capabilities, and opportunities. In 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security (QRS) (pp. 206-217). IEEE.
[59] Dzeparoska, K., Lin, J., Tizghadam, A., & Leon-Garcia, A. (2023, October). LLM-based policy generation for intent-based management of applications. In 2023 19th International Conference on Network and Service Management (CNSM) (pp. 1-7). IEEE.
[60] Jeong, C. (2023). Generative AI service implementation using LLM application architecture: based on RAG model and LangChain framework. Journal of Intelligence and Information Systems, 29(4), 129-164.
[61] Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., ... & Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. ACM Transactions on Knowledge Discovery from Data.
[62] Ziems, N., Yu, W., Zhang, Z., & Jiang, M. (2023). Large language models are built-in autoregressive search engines. arXiv preprint arXiv:2305.09612.
[63] Spatharioti, S. E., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Comparing traditional and llm-based search for consumer choice: A randomized experiment. arXiv preprint arXiv:2307.03744.
[64] Yao, B., Jiang, M., Yang, D., & Hu, J. (2023). Empowering LLM-based machine translation with cultural awareness. arXiv preprint arXiv:2305.14328.
[65] Karpinska, M., & Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. arXiv preprint arXiv:2304.03245.
[66] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature medicine, 29(8), 1930-1940.
[67] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.
[68] Yang, Y., Tang, Y., & Tam, K. Y. (2023). Investlm: A large language model for investment using financial domain instruction tuning. arXiv preprint arXiv:2309.13064.
[69] Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X. Y. (2023, November). Enhancing financial sentiment analysis via retrieval augmented large language models. In Proceedings of the Fourth ACM International Conference on AI in Finance (pp. 349-356).
[70] Mbakwe, A. B., Lourentzou, I., Celi, L. A., Mechanic, O. J., & Dagan, A. (2023). ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS digital health, 2(2), e0000205.
[71] Pearce, H., Tan, B., Ahmad, B., Karri, R., & Dolan-Gavitt, B. (2023, May). Examining zero-shot vulnerability repair with large language models. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 2339-2356). IEEE.
[72] Xia, C. S., Paltenghi, M., Tian, J. L., Pradel, M., & Zhang, L. (2023). Universal fuzzing via large language models. arXiv preprint arXiv:2308.04748.
[73] Cai, Y., Mao, S., Wu, W., Wang, Z., Liang, Y., Ge, T., ... & Duan, N. (2023). Low-code llm: Visual programming over llms. arXiv preprint arXiv:2304.08103, 2.
[74] Bommasani, R., Liang, P., & Lee, T. (2023). Holistic evaluation of language models. Annals of the New York Academy of Sciences, 1525(1), 140-146.
[75] Zhang, C., Bai, M., Zheng, Y., Li, Y., Xie, X., Li, Y., ... & Liu, Y. (2023). Understanding large language model based fuzz driver generation. arXiv preprint arXiv:2307.12469.