Advancing Natural Language Processing: A Journey Through the Evolution of Language Models

Akanksha Bisht[1*],Aditi Gupta[1],Aarsh Mall[1] and Nidhi Rana[2]

[1]Shivalik College of Engineering, Dehradun, Uttarakhand, India (248197)
[2]College of Engineering Roorkee, Uttarakhand, India (247667)

akankshabisht0107@gmail.com

Abstract. Natural language processing (NLP) has come a long way from a basic rule-based system to the sophisticated AI models we use today. The system used to adhere to rigid rules, which worked well for simple tasks but had trouble with the complexity of real-world language. The introduction of machine learning, including neural networks like MLPs, CNNs, RNNs, brought some huge improvements, but the real ground breaker came in 2017 with the transformer model. Transformers, with their ability to process text parallel and focus on key parts of the sentences, which completely transformed NLP. Today, models like BERT and GPT are at the soul of applications like chatbots, translations and summarization. This paper looks at the journey of language models, the breakthroughs in the field, the challenges that remain, and the interesting future possibilities, including combining text with other forms of data.

Keywords: Natural Language Processing (NLP), Transformers, RNNs, CNNs, MLPs, LSTMs

1 Introduction

Natural language processing (NLP) is one of among the most interesting and significant domains of computer science and artificial intelligence (AI). Simply put, it is about training computers to understand and use human language, which is a huge task given how complex, ambiguous, and context-dependent language can be.Consider the last time you asked Alexa to play your favorite song or used Google Translate to understand a foreign language. These everyday conveniences are driven by NLP[1]. What's more astonishing is how much this discipline has changed over time. But what is more remarkable is how much this field has evolved over the years.

In the early days, NLP were like rigid rule followers. Natural language processing (NLP) has emerged as one of the most exciting. They relied on hand-coded rules that linguists and programmers carefully developed. These systems were capable of performing fundamental tasks such as sentence parsing, but struggled when confronted with real-world linguistic variety. For example, rule-based chatbots

P. Ranjan et al. (eds.), *Proceedings of the International Conference on Recent Advances in Artificial Intelligence for Sustainable Development (RAISD 2025)*, Advances in Intelligent Systems Research 196,
https://doi.org/10.2991/978-94-6463-787-8_32

may understand "How are you?" But are you completely thrown off by "how it's going?" [2].

With the advent of statistical methodologies in the 1990s, things began to shift. Researchers started to use mathematical models to examine patterns in huge text databases rather than depending just on rules. Because it enabled the system to learn from examples instead of being manually designed for every scenario, this was a significant advancement [3]. These statistical models did have limitations, though. They were able to tally the frequency of word combinations, but they were unable to decipher their meaning.

Machine learning and then deep learning, which brought techniques like Multilayer Perceptrons (MLPs), were the true breakthroughs. Using neural networks for NLPs began with MLPs, which are basic neural networks made up of several layers of neurons. However they have significant limitations in capturing sequential data like sentences or paragraphs. Around the same time researchers explored Conventional neural networks (CNNs) originally designed for image processing to analyze text by focusing on patterns like n-grams or word embedding in more localized manner [4]. While CNNs improved performance on certain tasks like sentence classifications, they too struggled with understanding long term dependency in text. The gap was eventually connected with Recurrent Neural Networks (RNNs), which were especially created to process data consecutively. RNNs could remember information across sequence, making them a natural fit for language tasks. However they have their own limitations, like the gradient that disappears problems, which caused it to difficult for them to record long term dependencies, this challenge was addressed with the development of Long Short Term memory networks(LSTMs), A more advanced type of RNN that introduced mechanism to store and selectively forget information[5].

Then, in 2017 everything changed again with the introduction of the transformer model. This architecture introduced in the paper “Attention is all you need” by Vaswani et al., shifted the pattern in NLP. Transformers made it possible to process entire sequences of text in parallel, which was a game-changer for efficiency. The "attention mechanism," which enabled models to focus on the most pertinent portions of phrases while digesting them, was the true magic, though [6]. Models like BERT (Bidirectional Encoder Representations from Transformers) and the most delinquent sensation, GPT (Generative pre-trained transformer), were developed as a result of this discovery and have since taken over the field [7], [8].

Nowadays, the foundation of NLP is transformer-based models. They enable automated text summarization, machine translation, chatbots, and virtual support. Consider DeepSeek, Gemini, and GPT-4. Throughout this process, BERT has been frequently used to comprehend and extract meaning from text [8]. The fast growth of natural language processing and creative AI has sparked debates about AI ethics, discrimination, and sustainability [9].

This paper examines the lengthy history of natural language processing, focusing on how language models have evolved over time. We will begin by reviewing

rule-based systems and statistical approaches before delving into the birth of neural networks such as CNNs, RNNs, MLPs, and LSTMs, as well as the transformational potential of transformers. Here, we have studied and mentioned the challenges these models face like controlling bias in training data and finding a balance between performance and the computational cost. NLP has grown tremendously in the past decade, from enhancing model implementation to creating systems that can understand a variety of data, including text, audio, and pictures [10].

2 Historical Overview of NLP and Early Language Models

The lofty goal of teaching computers to comprehend and produce human language was the origin of natural language processing, or NLP. Linguistics and mathematical logic served as the foundation for early systems, but as computing power increased, natural language processing (NLP) moved from rigid rule-based techniques to more adaptable statistical models and, ultimately, machine learning. Let's examine how these advancements lay the groundwork for the cutting-edge technology of today.

2.1 Rule Based System: the foundation of NLP

In the early days of NLP, the most common approach was rule-based systems. These systems were developed based on the handcraft rules– essentially, a set of instructions programmed by linguistics and computer scientists. The idea was to encode the knowledge of human language in the form of syntactic rules that a machine could follow. Early instances include Joseph Weizenbaum's chatbot ELIZA, developed in the 1960s, and used a set of pattern matching rules to stimulate conversation [2].

While ELIZA could mimic a physiotherapist, its capabilities were limited to simple, script based interactions. Rule-based often flattered when faced with the complexities of the natural language, as they required a tiresome set of rules to cover all possible linguistic scenarios. Though they have drawbacks, rule-based systems established the groundwork for more sophisticated language models. They made clear the importance of understanding systematic structure and marked the first attempt at machine understanding language.

2.2 Statistical Models: A New Era in Language Modeling

The 1990s saw a dramatic departure from rule-based systems and a move toward statistical models. To find trends in the data, researchers began utilizing statistical methods and big text corpora. The method was based on n-gram models, one of the earliest statistical methods for language modeling.

An n-gram is a string of "n" words in which the model forecasts the subsequent word by using the frequency of word combinations found in the training set. Fig. 1 depicts the transition from rule-based systems to statistical models, showcasing the evolution of AI-driven decision-making and pattern recognition.

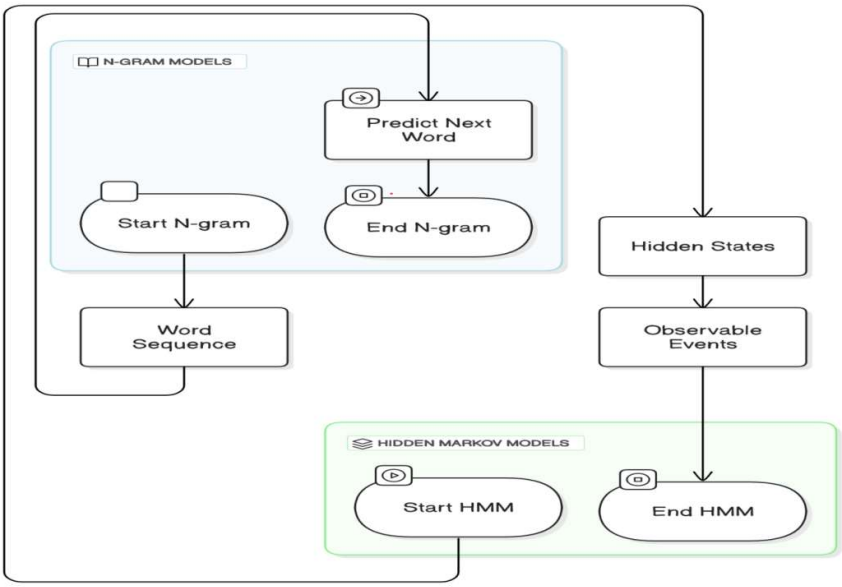

Fig. 1. 1990s: The Shift to Statistical Models

During this time, the Hidden Markov Model (HMM) also became well-known. They worked well for tasks like part-of-speech tagging and speech recognition. An effective abstraction for NLP tasks was provided by HMMs, which described the sequence of observable events (words) and their corresponding hidden states (such as syntactic categories) [11].

While these models could capture certain patterns in language, their reliance on fixed sequences (such as bigrams ortrigrams) meant they lacked an understanding of long-term dependencies in language, which became an obvious limitation as the task grew more complex. The introduction of machine entropy models and conditional random fields further improved the ability of models to capture complex relationships in the text. However, statistical methods still have crucial drawbacks: they rely heavily on the availability of vast amounts of label data to learn patterns. This often led to challenges in tasks like syntactic parsing, where rules and context were not easily captured by statistical method alone.

2.3 Early machine learning approaches: bridging the gap

The next leap came with machine learning (ML) approaches that came in the late 1990s and early 2000s. Support Vector Machines (SVMs), decision trees, and logis-

tics regression became popular for tasks such as text classification, sentiment analysis, as well as named entity recognition [12].

So these models are able to recognize patterns in data, and they represent a substantial improvement over hand coded rules. They were nevertheless still limited by their inability to comprehend linguistic sequential dependencies. The traditional machine learning algorithms that relied on fixed length features vectors faced difficulties since language is fundamentally sequential, with words building upon one and another. Researchers started experimenting with artificial neural networks (ANNs) at the same time. Multilayer perceptron's (MLPs), one of the earliest neural network models, began to make progress in identifying patterns in linguistic data. However, their inability to efficiently analyze interconnections between distant words in a phrase made their limits clear when dealing with sequential information, such as sentences or paragraphs [13].

2.4 The Development of Neural Networks: Handling Extended Backbone

The First machine learning model, the creation of the language model was RNN, which is the first doorstep of neural networks. This model is especially created to handle consequences in the reference of NLP in the reference of jobs, where context is crucial. RNNs have the ability to "remember" words from a sequence for the prediction in the future. At last it's found that it is able to handle the consequences of the previous model: The Sequential nature of language.

RNNs were far from flawless, though. It was challenging for RNNs to capture long-range dependencies in text because of a major issue called the vanishing gradient problem. As an example, RNNs would often "forget" the words that came before them at the end of a long sentence. To address [5] suggested LSTMs, or long short term memory networks. Longer sequences and connection recognition were far more far more easily handled by LSTMs because of their gating mechanism, which allowed them to selectively store and forget information. Later in the 2000s LSTM became the method of choice for tasks like machine translation and speech recognition.

LSTMs and RNNs continued to suffer difficulties in spite of these developments, particularly with regard to parallelization and effectively handling big datasets. This prepared the way for the transformer model, which was the next significant innovation in NLP.

2.5 Transformers: the revolution in media.

The way NLPs tasks were tackled underwent a significant change in 2017 when Vaswani et al. introduced the transformer model. By processing full text sequences in parallel, transformers significantly increased efficiency compared to RNNs and LSTMs, which processed data sequentially. By enabling models to concentrate on the most pertinent segments of the input sequence during prediction, the attention mechanism was the primary breakthrough in transformers. Transformers might "attend"

to every word in a sentence at once rather than processing them one at a time, producing complex contextual embedding of words.

Transformers outperformed RNNs or LSTMs due to their capacity to process all of a sequence at once, especially for large scale tasks like question answering, text summarization, and language translation. Transformers paved the way for the creation of models such as GPT(Generative Pre Trained transformer) and BERT (Bidirectional Encoder Representations from Transformers), which have raised the bar for NLP tasks[14],[7].

By enhancing efficiency and making it possible to train much bigger models utilizing enormous datasets, Transformers created new possibilities for AI applications. These models serve as the basis for NLP applications in industries including healthcare and finance, enabling systems to do tasks that were previously unimaginable.

3The Emergence of Neural Language Models

Neural Language Models Development An important turning point in Natural Language Processing (NLP) was the creation of neural language models. These models opened the door for robots to comprehend and produce human language more effectively by introducing a novel method of representing and processing language that transcended the constraints of previous techniques.

During this time, word embedding is one of the earliest innovations. It was possible to map words into a continuous space that captured semantic links thanks to these dense vector representations. In contrast to the inflexible one-hot encodings of the past word embeddings reflected the contextual meaning of words. For instance, more unrelated phrases like “book” and “king” would be placed farther apart in this area, whereas words like “queen” and “king” would be placed closer together. Glove, which was first presented by [15], and Word2vec, which was created by [16], Glove used global statistics of word co-occurrence to train embeddings words that caught brother relationships in text, while Word2Vec invented Effective architectures like skip-gram and CBOW to predict word associations.

Neural networks gained prominence as a result of these developments in word representation. Feed-forward neural networks and other early attempts to include neural techniques were comparatively simple. Although they were comparatively simple. Although they were able to do basic NLP tasks like text categorization, they had trouble handling data sequences like sentences or paragraphs, which are essential for language comprehension.As a result of these developments in word representation, neural networks gained prominence. Initially, attempts to incorporate neural techniques, including feed-forward neural networks, were somewhat simple. Text classification and other basic NLP tasks were within their grasp, but they had trouble with data sequences like sentences and paragraphs, which are essential to language comprehension.

One major barrier to the adoption of Recurrent Neural Networks (RNNs) was the difficulty of handling sequential data. RNNs are perfect for language problems like speech recognition and translation because they have the capacity to process and remember information across sequences. Despite their potential, RNNs had drawbacks, most notably the vanishing gradient issue. Because of this problem, these models had trouble learning dependencies across lengthy sequences [17]. The below diagram, Fig 2 depicts the evolution of neural language models.

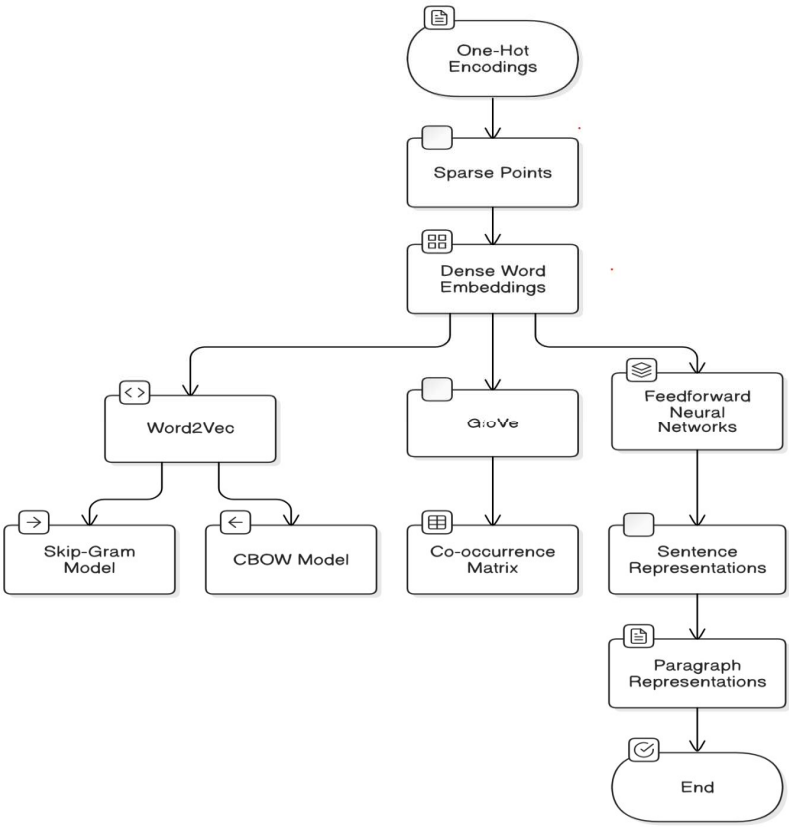

Fig. 2. Evolution of Neural Language Models in NLP

To overcome the problem, researchers created a more advanced version of RNNs called Long Short Term Memory Networks (LSTMs). These networks incorporated techniques to selectively forget and preserve information in order to better represent long-term dependency [5].LSTMs were widely used in applications that needed sequential comprehension, such text generation and automatic speech recognition, and represented a significant breakthrough in NLP. Another important neural design at this time was the multilayer perceptron (MLP). Because of their layered topologies of

linked neurons, MLPs were used for classification tasks including spam detection and sentiment analysis, according to [13]. The inability of early neural approaches to handle complex language difficulties was further demonstrated by the fact that they struggled with tasks that had sequential context since they processed inputs as fixed vectors.

Not with standing these developments, the existing designs had drawbacks: Furthermore, although word embeddings such as Word2vec and GloVe offered useful representation, their applicability to increasingly difficult language problems was limited due to their inability to adapt dynamically to contextual information. Due to their requirement for sequential information processing, RNNs and LSTMs were computationally costly, particularly when dealing with huge datasets.

But this time was crucial in moving NLP away from strict, rule-based frameworks and weak statistical models and toward more flexible, context-sensitive approaches. Using neural networks, researchers created the foundation for the next generation of revolutionary models, paving the way for the creation of transformers and contextual embedding that would fundamentally change the field.

4 The Revolution of Transformers

An important turning point in the development of Natural Language processing (NLP) was the release of the transformer architecture. The Transformer revolutionized the way Natural Language processing (NLP) systems interpret & process language [6]. first presented it in their seminal paper "Attention is All You Need." By addressing the drawbacks and inefficiencies of earlier models such as Recurrent Neural Networks (RNNs) and Long Short Term Memory networks (LSTMs), the Transformer transformed language modeling.

The core element of the Transformer is itself-attention mechanism, which allows the model to evaluate the relative significance of many words in phrases. For example, a Transformer can determine that, in spite of the intervening clause, “Boy” is related to “is happy” when examining the sentence “The boy who found the treasure is happy”. The Transformer processes complete sequences at once, in contrast to RNNs and LSTMs that process sequences word by word. Without the sequential processing bottleneck because of this parallelism, which also increases performance [6].

4.1 Comparing Benefits for Parallel Processing between RNN & LSTM

For lengthy sentences or documents, RNN and LSTM consumes more time for training and effectively cost increases for the Sequential Text Processing , on the other hand as we compared to Transformers , it uses hardware acceleration to analyses each and every word in a sequence manner and it less time in training as well[18].

Managing Extended Duration dependencies Transformers employ self-attention to take into account the relationships between every word in a sequence at

once, but RNNs find it difficult to capture long term dependencies because of the problem of disappearing gradients. This makes them especially useful for tasks where comprehension of the larger context is essential, such as translation and summarization [19].

More straight forwardly scalable buildings with transformers, a stack of encoder and decoder modules takes place of recurrent layers. Compared to earlier designs, the model can handle significantly larger datasets and more complicated jobs because of its improved scalability due to its simple architecture [20].

The Development of Contemporary NLP Models Some of the most significant NLP models of our time was made possible by the Transformer. For example:

1. Developed by [7] leverages the Transformer's bidirectional self-attention to exact context from words that come from and after. For tasks like entity recognition, sentiment analysis, and question answering, this makes it incredibly successful.
2. GPT (Generative Pre Trained Transformer) Models that are based on the Transformer architecture, such as GPT-3 and GPT-4, are excellent at producing writing that seems human. They are capable of summarizing articles, writing essays and even handling deep discussions [21].

4.2 The Transformers Beyond NLP

Beyond text processing the Transformer has a significant impact. Its architecture has been effectively modified for use in different fields:

1. Vision Transformers (ViTs): These models provide state-of-the art outcome in image recognition tasks by applying the Transformer's self-attention mechanism [22].
2. Multimodal Learning to accomplish tasks like producing through image descriptions or producing artwork in response to text prompts models such as CLIP and DALL-E integrate text and image data [23].

4.3 Issues and Moral Determinations

Despite Transformers' phenomenal success, there remain challenges. They require a lot of processing power, which raises concerns about energy consumption and the impact on the environment. Furthermore, there are ethical concerns around inclusion and fairness because these algorithms usually pick up biases from the massive datasets they are trained on [9]. Efforts are underway to make transformers less biased and more effective so that they may be utilized appropriately in real-world scenarios [24].

5The Pre Trained Model Era

A revolutionary era in Natural Language Processing (NLP) has begun with the introduction of Pre Trained models, which have radically altered the methods used to approach and resolve language tasks. By using enormous volumes of unlabeled data for pre-training and requiring very little labeled data for fine - tuning, models such as BERT, GPT, RoBERTa, and their successor have significantly advanced the field. In addition to raising NLP standards, this paradigm shift driven by transfer learning has made sophisticated language understanding more accessible for a wider range of applications.

The idea of fine tuning and pre training fundamentally, pre training is teaching a model general language patterns like grammar, syntax, and semantic linkages by using a large corpus of text. The resultant model is a flexible basis that may be optimized for particular applications using relatively small labeled datasets. Models can attain remarkable performance using significantly less data and computational resources because of this two-step procedure, which consists of pre-training and task specific fine tuning [25].

By learning bidirectional context, for example pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers) may comprehend a word depending on the words that surround it, whether those words are preceding or after it. Earlier unidirectional methods such as GPT 2, process text from left to right; this is in contrast. In tasks including named entity recognition, sentiment analysis, and question answering, researchers have achieved state-of-the-art results by optimizing BERT [26].

5.1 Pre trained key models

1. Transformer Based Bidirectional Encoder Representations, or BERT, BERT was first presented by [26] and its capacity to understand the reciprocal relationships in text made it significant advancement in NLP. It employs next sentence prediction (NSP) to capture inter sentence links and a masked language model (MLM) objective, in which some words in a phase are masked and the model learns to predict them. These training goals enable BERT to perform exceptionally well on tasks that call for a thorough comprehension of context.
2. GPT or Generative Pre Trained Transformer The generative methodology of OpenAI's GPT models focusses on anticipating subsequent word in a series. Although GPT-2 showed promise in producing coherent writing, GPT 3 and GPT 4 transformed conversational AI by producing human-like responses, summarizing texts, and even doing zero-shot learning- in which they tackled tasks for which they had not received explicit training [8].
3. RoBERTa: A Strurdily Enhanced BERT pre training Method, In contrast to BERT, RoBERTa employs more robust training techniques, including longer training durations, larger batches, and a variety of data sources, and does away with the NSP aim. A model that continuously beats BERT on a benchmark is the outcome of these improvements [27].

4. XLNet by using permutation based training target, XLNet, a BERT successor, overcomes the drawbacks of BERT's bidirectional pre training. This eliminates the need of masking tokens and enables it to capture both forward and backward dependencies [28].

5.2 The Effects of Transfer Learning

The development of NLPs has evolved to rely heavily on transfer learning. Researchers and developers can drastically cut down on resources required for task specific development by using pretrained models. This strategy has produced groundbreaking outcomes on benchmarks such as SuperGLUE leaderboard and General Language Understanding Evaluation (GLUE). The widespread use of pretrained models has led to unparalleled accuracy in tasks including sentiment analysis, machine translations, and natural language interface [19].

6Current Trends and Innovations

Recent developments are expanding the realms of what is feasible in the ever changing field of natural language processing (NLP). Improvements in multi model systems, instruction tuned models, and efficient transformers have broadened the use of language models while resolving some of the ethical and computational issues that arose with the previous architectures. In addition to improving performance, these developments are expanding the range of industries in which NLP can be used.

Effective Transformers: Enhancing Efficiency High computational cost is one of the biggest problems with transformer based models, such GPT and BERT. This has led to research on efficient Transformer topologies that aim to maintain or improve performance while reducing computing complexity.

Models such as Reformer and Linformer, for instance, provide techniques to reduce memory requirements and computational overhead. While Reformer effectively approximates attention processes using locality-sensitive hashing, Informer compresses attention matrices using low rank approximations [29], [30].By focusing on the most relevant parts of the input utilizing sparse attention approaches, Sparse Transformers decreased the model's resource intensity without sacrificing accuracy, according to [31]. Thanks to effective transformer designs, which enable the implementation of complex models without requiring a significant amount of processing power, NLP technology is now accessible to smaller enterprises.

6.1 Human Alignment and Models Accommodated by instruction

Instruction-tuned models are a major development in NLP. Because these models are taught to follow user-given directions in simple English, they are incredibly flexible and easy to use. Some models, such as Open AI's GPT 4 and Google's PaLM 2, enhance performance on a range of tasks through instruction tweaking, eliminating the requirement for task-specific fine-tuning.

Additionally, the significance of adjusting to human inputs has grown. Strengthening Models like Instruct GPT employ Learning from Human Feedback (RLHF) to better align their outputs with user expectations. This approach involves training the model with an annotated dataset created by human reviewers to guarantee that the machine's replies are accurate and relevant to the scenario [32].

The capacity of instruction-tuned algorithms to provide logical and practical solutions has greatly improved with the addition of human preferences in the training loop. Combining text, image, and audio in multimodal models. Another ground-breaking step is the development of multimodal models, which can generate and evaluate data from several modalities such as text, pictures, and audio [23-30]. Examples of trends include OpenAI's DALL.E model, CLIP, and Google's Flamingo. Fig 3 illustrates advancements in AI models for Natural Language Processing, highlighting innovative approaches that enhance language understanding, generation, and contextual reasoning.

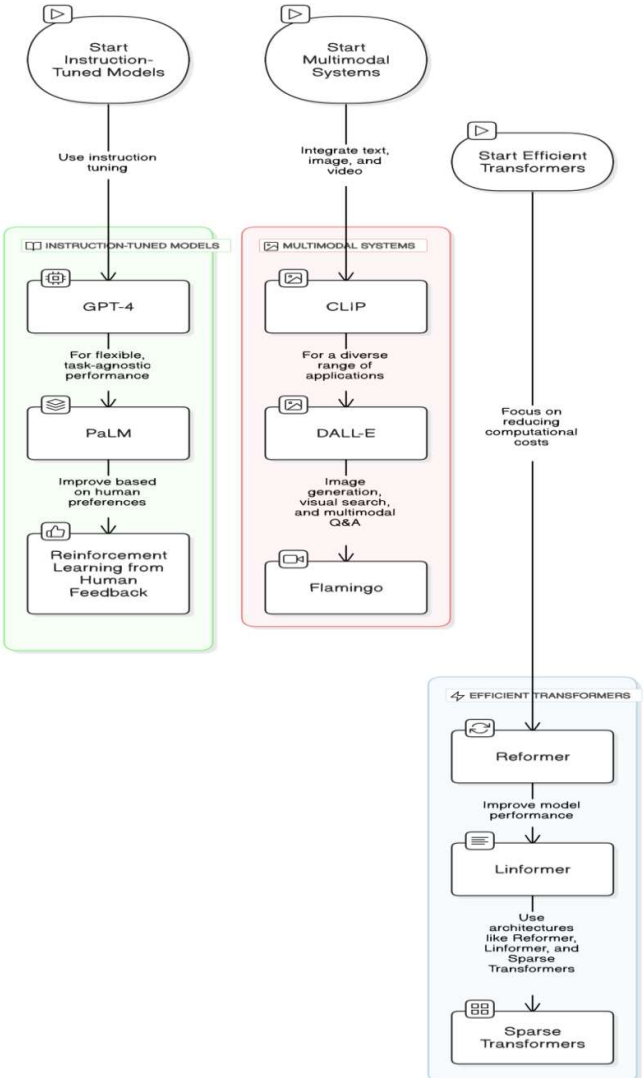

Fig. 3. Innovation in AI models for Natural Language Processing

While CLIP bridges text and images modalities for tasks like image capturing and visual search, DALL.E integrates text and image understanding to produce high quality images based on textual descriptions [33], [10]. In contrast, flamingo combines text, images and videos data, making it possible to use it in multi-modal's question answering, video summarization, and narrative [34].

These models are creating fascinating opportunities like education, where multimodal systems may produce interactive learning experiences, and health care where text to image models can produce visualizations of medical disorders.Prospect of a new trend in the future of NLPs can still be improved, Even though recent developments have pushed its limits [34-38]. For efficient transformers to support larger datasets and more complex activities, they must continue to develop while preserving their environmental sustainability. More advancements are needed in instruction-tuned models to address possible biases and enhance generalizability. Multimodal systems must deal with similar challenges in integrating several data sources without compromising contextual accuracy. If academics and practitioners focus on these topics, NLP can remain a dynamic and important subject that can satisfy both technology and societal goals.

7Challenges and Ethical Consideration

As the NLP system evolves and becomes integrated into daily life, it presents a number of challenges and ethical dilemmas. There has been a lot of discussion about concerns including model bias, data privacy risks, interpretability problems, and environmental costs. These problems need to be fixed to ensure that NLP technology remains useful and inclusive.

7.1 Linguistic Model Bias

Language model bias is now a major ethical concern. These models frequently pick up on and intensify biases in the training set. They might, for example, create products that portray racial, gender,or socioeconomic stereotypes. Studies have demonstrated that, depending on past trends in their datasets, models such as GPT and BERT might provide damaging or biased material [9], [35].

Curating more representative and varied datasets and creating methods for de-biasing models are two ways to lessen bias. For instance, methods that are presently being studied include adversarial training and fine-tuning on bias-reduced data [36].But attaining total neutrality is still a difficult task.

7.2 Privacy of Data

Large Volumes of data are frequently needed to train language models, which raises privacy issues. Particularly in customized apps like chatbots and virtual assistants, user data may unintentionally be compromised or exploited. High-profile events brought attention to the significance of protecting sensitive data, including breaches utilizing AI models[37].

Methods such as federated learning and differential privacy have surfaced as viable remedies. While federated learning enables models to learn from decentralized data without moving it to a central server, differential privacy guarantees that individual user data cannot be identified in aggregated datasets [38]. These methods are opening the door to NLP systems that are more secure.

7.3 Interpretability

The interpretability of massive language models is another noteworthy obstacle. It is more difficult to understand how models reach particular conclusions as they get more complex. Trust may be weakened by this lack of openness, particularly in high-stakes domains like the legal or medical systems.

Increasing interpretability can be achieved, for example, by developing explainable AI (XAI) techniques that shed insight on model selection. For example, saliency maps and attention visualizations are used to demonstrate which elements of the input data the model focused on while generating predictions [39]. These technologies are helping to shed light on the "black box" nature of NLP models.

7.4 Impact of the environment

The environmental impact of developing and deploying large language models is another pressing concern. The carbon impact of many transatlantic trips is equivalent to the power needed to train a single big model, like GPT-3 [24]. This raises questions about the viability of scaling NLP technology.

Researchers are investigating methods such as algorithm optimization, the use of energy-efficient hardware, and algorithm optimization to reduce the environmental impact of NLP [40]. These efforts are crucial to balancing technological advancements with environmental responsibility. Mitigation and Conscientious AI In order to address these concerns, the industry has begun to place a greater focus on responsible AI. This means establishing moral guidelines, assembling multidisciplinary teams of ethicists and sociologists, and promoting transparency in the development of models. Organizations like Open AI and the Partnership on AI have taken steps to create best practices and encourage the appropriate use of NLP technology [41].

Additionally, recent research aims to take preventative actions in order to avoid potential dangers. For instance, anticipatorily identifying and addressing potential abuse scenarios, including disseminating false information, is now an essential part of model construction.

8Conclusion

With the rise of natural language processing (NLP), rule-based systems were superseded by statistical techniques, neural networks, and the ground-breaking transformer architecture. By addressing earlier issues, each phase cleared the path for the more sophisticated models of today, including BERT and GPT. These pre-trained transformer-based models have revolutionized natural language processing (NLP), increasing the precision and effectiveness of tasks such as summarization translations and conversational AI.

Recent developments like instruction-turned and multimodal models hint at a future when AI systems can easily understand and integrate several data sources, including text and visuals. However, there are challenges associated with these improvements. Bias, data privacy, and the environmental effects of large-scale models remain significant issues. These problems must be resolved to create inclusive and moral AI systems.

As the field develops, the emphasis will be on striking a balance between ethical responsibilities and innovation. NLP can continue to influence how we engage with technology and one another in significant and disruptive ways by improving interpretability, streamlining model efficiency, and encouraging ethical AI activities.

Disclosure of Interests. The authors declare that they have no conflict of interest.

References

1. Jurafsky,D.,Martin,J.H.: Speech and Language Processing. 3rd ed.Prentice Hall (2021)
2. Weizenbaum, J.**:** ELIZA is a computer program for the study of natural language communication between man and machine.Communications of the ACM**9**(1), 36–45 (1966)
3. Manning, C.D., Schütze, H.:Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
4. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation**9**(8), 1735–1780 (1997)
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need.Advances in Neural Information Processing Systems**30**, 5998–6008 (2017)
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agrawal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Amodei, D., Sutskever, I.: Language models are few-shot learners. Advances in Neural Information Processing Systems33, 1877–1901 (2020)
9. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.:On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. ACM, New York (2021)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.**:** Learning transferable visual models from natural language supervision.arXiv preprint arXiv:2103.00020 (2021)
11. Jelinek, F.:Statistical Methods for Speech Recognition. MIT Press, Cambridge (1997)
12. Cortes, C., Vapnik, V.:Support vector networks.Machine Learning**20**(3), 273–297 (1995)
13. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by backpropagating errors. Nature**323**, 533–536 (1986)

14. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners.OpenAI Blog (2019).
15. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In:Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. ACL, Doha (2014)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
17. Bengio, Y., Simard, P., Fransconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks**5**(2), 157–166 (1994)
18. Henderson, J., Titov, I., Cloete, A.: Leveraging neural attention mechanisms for language understanding.Journal of Machine Learning Research**21**(92), 1–33 (2020)
19. Wang, J., Liu, Y., Zhou, M.: Transforming NLP: From RNNs to transformers. Communications of the ACM**62**(10), 45–52 (2019)
20. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research**21**, 5485–5555 (2020)
21. Zhang, T., Sun, X., Han, J.: Generative pre-trained transformers for NLP: A review. Artificial Intelligence Review**54**(3), 2435–2465 (2020)
22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16*16 words: Transformers for image recognition at scale. arXiv preprintarXiv:2010.11929 (2021)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In:Proceedings of the International Conference on Machine Learning (ICML) (2021)
24. Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243 (2019)
25. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of ACL 2018, pp. 328–339 (2018)
26. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of* NAACL-HILT 2019, pp. 4171–4186 (2019)
27. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pre-training approach. arXiv preprintarXiv:1907.11692 (2019)
28. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: Generalized autoregressive pre-training for language understanding. Advances in Neural Information Processing Systems**32**, 5753–5763 (2019)
29. Kitaev, N., Kaiser, L., Levskaya, A.: Reformer: The efficient transformer. In: Proceedings of ICLR 2020(2020)
30. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv preprintarXiv:2006.04768 (2020)
31. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprintarXiv:1904.10509 (2019)

32. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A.: Training language models to follow instructions with human feedback.arXiv preprintarXiv:2203.02155 (2022)
33. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. arXiv preprintarXiv:2102.12092 (2021)
34. Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelovic, R., Ramapuram, J., Gedik, A.: Flamingo: A visual language model for few-shot learning. In: Proceedings of NeurIPS 2022(2022)
35. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of NeurIPS 2016(2016)
36. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: Proceedings of NAACL 2018 (2018)
37. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Proceedings of IEEE S&P 2017 (2017)
38. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of AISTATS 2017 (2017)
39. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
40. Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O.: Green AI. Communications of the ACM**63**(12), 54–63 (2020)
41. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K.: Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. Minds and Machines**29**(4), 555–580 (2019)