Transformers and Beyond: A Review of Key NLP Developments

Yash Rathore
Delhi Skill and Entrepreneurship University (DSEU)
ftd10622456@dseu.ac.in

***Abstract*—Natural Language Processing (NLP) has witnessed an unprecedented surge in capabilities and applications, driven by advancements in deep learning and transformer-based architectures. This review presents a comprehensive synthesis of twelve influential research contributions that span the breadth of modern NLP, encompassing foundational models and emerging frontiers. We begin by examining core architectures such as BERT, GPT, and T5, which have redefined pretraining, fine-tuning, and transfer learning paradigms. The survey further explores multilingual representation learning with XLM-R, domain-specific adaptations like BioBERT for biomedical text, and robust question answering systems exemplified by UnifiedQA. Addressing critical needs in trust and usability, we analyze explainability methods in NLP and ethical considerations around model biases and responsible deployment. Additionally, the paper reviews the rise of instruction-tuned and prompt-based learning in models like FLAN-T5, sentiment and emotion recognition from conversational data, empathetic response generation, and multimodal learning through vision-language models such as Flamingo. Collectively, these works highlight the evolution of NLP from task-specific systems to general-purpose language understanding frameworks, while underscoring challenges related to interpretability, multilingual fairness, and real-world reliability. This review aims to guide researchers and practitioners toward deeper insights and future directions in building inclusive, ethical, and high-performing NLP systems.**

***Index Terms*—Natural Language Processing, Transformers, BERT, GPT, T5, Multilingual NLP, Explainability, Sentiment Analysis, Prompt Engineering, Multimodal Learning, Biomedical Text Mining, NLP Ethics**

I. Introduction

Natural Language Processing (NLP) has become an indispensable branch of artificial intelligence, enabling machines to comprehend, interpret, and generate human language at scale. The field has rapidly evolved from rule-based systems and statistical models to deep neural networks powered by transformer architectures. Over the past few years, a series of groundbreaking models—such as BERT, GPT, and T5—have set new benchmarks in understanding and generating natural language. These models not only improved performance across a wide array of tasks but also introduced paradigm shifts in how NLP problems are framed and solved, particularly through transfer learning and pretraining on massive corpora.

This review synthesizes twelve highly impactful research papers that collectively span the most critical domains of modern NLP. The selected works represent innovations in language modeling, multilingual representation learning, domain-specific adaptations, explainability, ethical AI, and multimodal learning. The evolution of autoregressive models like GPT-3 [2] and unified frameworks such as T5 [3] has demonstrated the scalability and generalizability of transformer-based architectures. Simultaneously, multilingual models like XLM-R [4] address the global applicability of NLP, while domain-specific models such as BioBERT [8] cater to high-impact areas like healthcare.

In addition to architectural advancements, the field is increasingly addressing challenges such as transparency and fairness. Recent surveys and toolkits have proposed explainable NLP (XNLP) frameworks [5] to improve interpretability, while ethical critiques have drawn attention to the risks associated with data bias and large-scale deployment [12]. Emerging trends such as instruction tuning [9], emotion-aware conversational agents [6], and vision-language models [10] are pushing the boundaries of what language models can understand and do.

By consolidating insights from these twelve works, this review aims to provide a structured and critical understanding of how modern NLP has evolved and where it is heading. Each section of the paper focuses on one major topic, summarizing the core contributions of a representative research work and highlighting its impact on the field. The paper concludes by outlining open challenges and future research directions necessary to build inclusive, ethical, and adaptable NLP systems for the next generation.

II. BERT – Bidirectional Encoder Representations from Transformers

The introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. [1] marked a breakthrough in Natural Language Processing by enabling deep, bidirectional understanding of text. Prior models, including GPT and ELMo, were constrained by unidirectional or shallow context processing, whereas BERT leveraged a stack of transformer encoders to jointly condition on both left and right context in all layers. This structural innovation resulted in state-of-the-art performance across a wide array of NLP tasks.

BERT is pretrained using two self-supervised objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, random tokens in the input sequence are masked, and the model is trained to predict these tokens using the surrounding context. This strategy allows BERT to capture semantic and syntactic relationships more effectively than left-to-right models. The NSP objective trains BERT to determine whether a given sentence follows another in a coherent discourse, which is particularly useful for tasks involving sentence-level understanding such as natural language inference (NLI) and question answering.

The architecture of BERT is based on the transformer encoder introduced by Vaswani et al. [13], comprising multiple layers of multi-head self-attention, feed-forward networks, layer normalization, and residual connections. The original BERT-base model includes 12 transformer layers, 768 hidden units, and 12 attention heads, while BERT-large doubles these parameters, significantly improving performance but also increasing computational costs.

Fine-tuning is a key aspect of BERT's versatility. After pretraining, task-specific layers are added and the entire model is trained end-to-end on datasets such as GLUE, SQuAD, and CoLA. This has led to new state-of-the-art benchmarks in sentence classification, named entity recognition, and semantic similarity.

Following BERT's success, several optimized variants emerged. RoBERTa removed the NSP objective and trained longer with more data, achieving even better performance. DistilBERT compressed the model to reduce inference time, and ALBERT introduced parameter sharing to reduce redundancy. These innovations collectively contributed to the proliferation of transformer-based architectures in both academic and industrial NLP pipelines.

Despite its achievements, BERT's limitations include high memory consumption, slow inference, and the absence of generative capabilities due to its encoder-only design. Nonetheless, its role in shaping modern NLP architectures remains foundational.

III. GPT – Generative Pre-trained Transformers

The Generative Pre-trained Transformer (GPT) series, developed by OpenAI, introduced a paradigm shift in NLP by demonstrating that large-scale, autoregressive language models can generalize across tasks through simple prompting. The original GPT was released in 2018, followed by GPT-2 in 2019 and the landmark GPT-3 in 2020, which contains 175 billion parameters [2]. These models differ from BERT in architecture and training objective—they use only the transformer decoder stack and are trained with causal language modeling (CLM) to predict the next token in a sequence.

Unlike BERT's bidirectional attention, GPT employs unidirectional (left-to-right) attention, which maintains the natural flow of language and facilitates generation tasks such as story writing, summarization, and dialogue modeling. The model architecture consists of multiple layers of masked multi-head self-attention, feed-forward layers, residual connections, and layer normalization.

GPT-3's most distinctive capability is few-shot and zero-shot learning. By providing just a few task examples or descriptions in the prompt, GPT-3 can perform tasks it has never been explicitly trained on, such as translation, summarization, or question answering. This behavior is largely attributed to its massive parameter size and training on a broad dataset comprising Common Crawl, WebText, books, and Wikipedia.

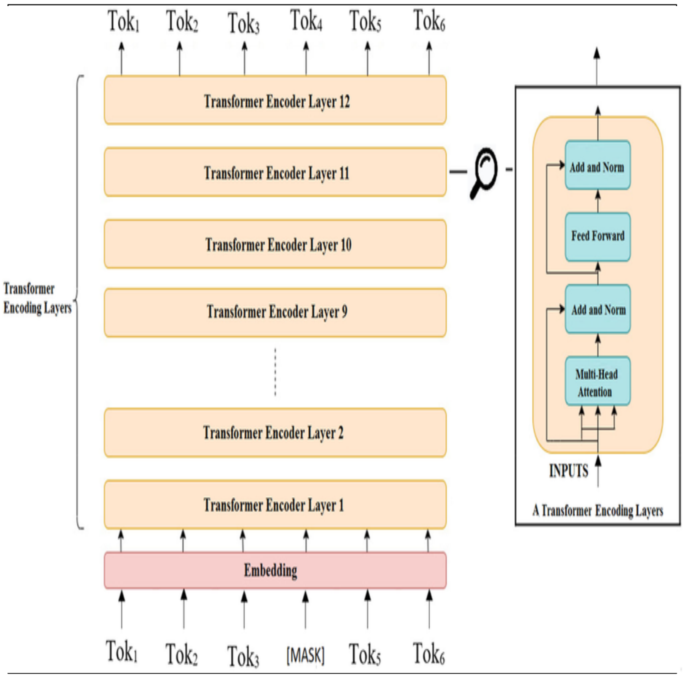

Tok1
Tok2
Tok3
Tok4
Tok5
Tok6
Transformer Encoder Layer 12
Transformer Encoder Layer 11
Transformer Encoder Layer 10
Transformer Encoder Layer 9
Transformer Encoder Layer 2
Transformer Encoder Layer 1
Transformer Encoding Layers
Embedding
Tok1
Tok2
Tok3
[MASK]
Tok5
Tok6
Add and Norm
Feed Forward
Add and Norm
Multi-Head Attention
INPUTS
A Transformer Encoding Layers

Fig. 1. BERT architecture illustrating stacked bidirectional transformer encoders with masked language modeling and next sentence prediction [1].

The training objective of GPT follows the standard autoregressive formula:

$$P(x) = \prod_{t=1}^{T} P(x_t | x_1, x_2, ..., x_{t-1}) \tag{1}$$

This makes GPT particularly powerful for text generation tasks where coherence and fluency over long passages are critical. However, GPT models also face criticism for hallucination, bias reproduction, and lack of transparency in their decision-making.

With the introduction of GPT-4, OpenAI extended the GPT family to multimodal inputs, enabling processing of both text and image data. GPT-4 also shows improved factuality, safety, and alignment, although details about its architecture and training remain partially undisclosed.

IV. T5 – Text-to-Text Transfer Transformer

The Text-to-Text Transfer Transformer (T5), introduced by Raffel et al. [3], redefined the formulation of NLP problems by casting all tasks—including classification, translation, summarization, and question answering—into a text-to-text framework. Unlike BERT, which uses only the encoder, or GPT, which relies on the decoder, T5 adopts a full encoder-decoder transformer architecture, similar to the original Transformer proposed by Vaswani et al. [13].

The key innovation of T5 lies in its unified task format. For every task, both the input and output are treated as text strings. For instance:

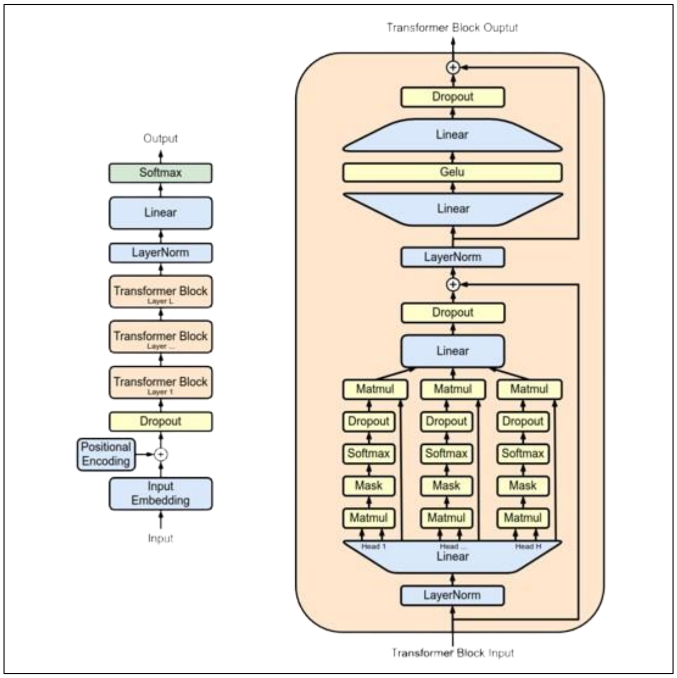


Fig. 2. GPT architecture: Transformer decoder blocks using masked self-attention for left-to-right language modeling [2].

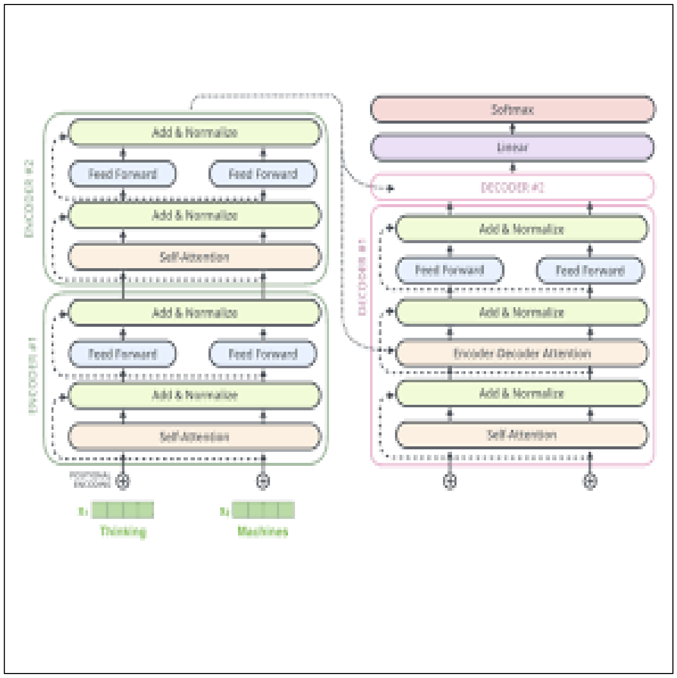


Fig. 3. T5 architecture: Encoder-decoder model trained with a span corruption objective using the text-to-text paradigm [3].

Translation: ”translate English to French: I love apples.” → ”J’aime les pommes.”

Summarization: ”summarize: The article discusses...” → ”The article highlights...”

T5 was pretrained on a large cleaned corpus known as C4 (Colossal Clean Crawled Corpus) using a denoising objective called Span Corruption, where random spans of text are replaced by a sentinel token, and the model is trained to reconstruct the missing content. This differs from BERT’s token masking and allows T5 to learn both language understanding and generation simultaneously.

T5’s architecture supports multiple model sizes, from T5-small (60M) to T5-11B (11 billion parameters), enabling its deployment across tasks of varying complexity. The model is fine-tuned for downstream tasks by simply prepending task-specific prefixes (e.g., ”summarize:”, ”translate:”) to inputs, eliminating the need for architectural modifications.

T5 achieved state-of-the-art results on benchmarks such as GLUE, SuperGLUE, SQuAD, and CNN/DailyMail summarization. Its flexibility, scalability, and simplicity have made it a backbone for many subsequent instruction-tuned models, including FLAN and UnifiedQA.

Limitations of T5 include its high computational demands during training and inference, particularly for the larger variants. Nonetheless, its design philosophy has influenced many modern NLP models by demonstrating the power of a unified sequence-to-sequence formulation.

V. Multilingual NLP – XLM-RoBERTa

In a globally connected world, the ability of NLP models to understand and generate text across multiple languages is crucial for inclusivity and scalability. The XLM-RoBERTa (XLM-R) model, introduced by Conneau et al. [4], is one of the most successful efforts in this direction. It builds on the masked language modeling objective of BERT and RoBERTa, while scaling training to 100 languages using massive multilingual corpora.

XLM-R is pretrained on a dataset derived from Common-Crawl, called CC100, which includes cleaned and filtered content from 100 different languages. Unlike its predecessor XLM (which used translation language modeling), XLM-R adopts the masked language modeling (MLM) approach without requiring parallel corpora. This makes it more scalable and adaptable to low-resource languages.

The architecture of XLM-R is identical to RoBERTa, consisting of a deep stack of transformer encoder layers with GELU activation, byte-pair encoding (BPE) tokenization via SentencePiece, and training without NSP. However, XLM-R is trained on 2.5 TB of multilingual text, which is significantly more than multilingual BERT (mBERT), allowing it to learn richer cross-lingual representations.

Performance-wise, XLM-R significantly outperforms mBERT and even strong monolingual baselines on cross-lingual benchmarks like XNLI, MLQA, BUCC, and XTREME. In zero-shot transfer settings—where the model is fine-tuned on one language (e.g., English) and tested on others—it delivers strong generalization capabilities across typologically diverse languages.

XLM-R has also proven effective in real-world applications, such as multilingual search, machine translation pretraining, and cross-lingual sentiment analysis. Its ability to bridge language gaps without parallel supervision makes it a preferred model for inclusive and globally deployable NLP systems.

However, like all multilingual models, XLM-R faces the challenge of language imbalance. High-resource languages dominate training data, sometimes leading to performance degradation on low-resource languages. Additionally, token fragmentation issues in agglutinative or morphologically rich languages can limit performance.

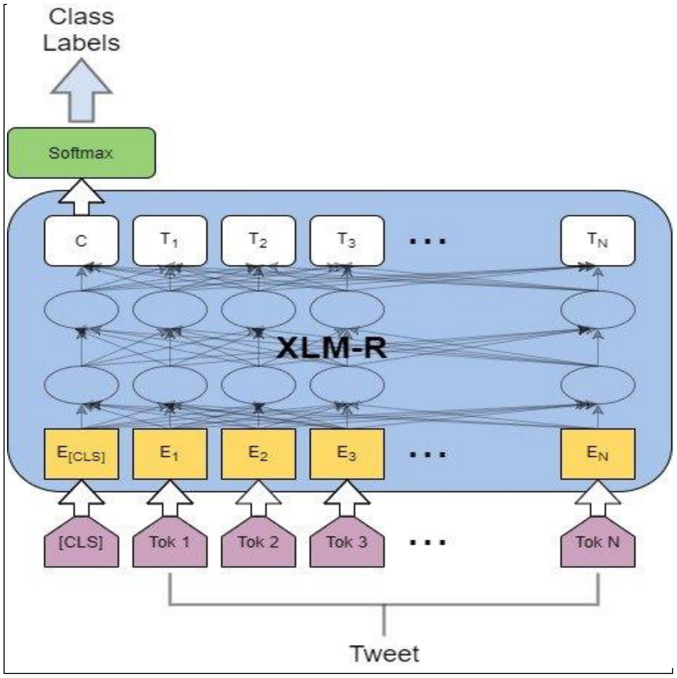


Fig. 4. XLM-R multilingual transformer trained with MLM objective across 100 languages using CC100 corpus [4].

VI. Explainability in NLP – XNLP Techniques and Challenges

As neural networks become increasingly integral to high-stakes NLP applications—such as healthcare, legal systems, and finance—the demand for explainability and transparency in their predictions has become critical. Traditional transformer-based models like BERT, GPT, and T5 are often referred to as black-box models due to their opacity in decision-making, despite their remarkable performance. To bridge this interpretability gap, the field of Explainable NLP (XNLP) has emerged, aiming to provide human-understandable justifications for model predictions.

A comprehensive survey by Mohammadi et al. [5] classifies XNLP methods into post-hoc and intrinsic approaches. Post-hoc techniques, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), generate explanations after model predictions are made. These are model-agnostic and widely used for classification tasks. LIME perturbs input tokens to assess their impact on the output, while SHAP calculates each token's marginal contribution using game-theoretic principles.

Intrinsic methods, on the other hand, modify model architecture or training to produce explanations natively. Examples include attention-based explanations, rational generation modules, and interpretable embedding spaces. For instance, models that highlight input spans during sentiment analysis or QA offer insights into which parts of the text influenced the decision. However, attention weights themselves are not always reliable indicators of importance, as highlighted in recent studies.

XNLP techniques are evaluated using metrics such as fidelity, plausibility, and stability. Fidelity measures how well the explanation aligns with the model's behavior, while plausibility assesses human-perceived correctness. High-fidelity explanations are essential in domains like medicine or criminal justice, where a wrong justification may have serious consequences.

The survey also emphasizes domain-specific challenges. In healthcare, explanations must align with clinical logic, while in finance, explanations must comply with transparency regulations. In dialogue systems, emotional reasoning must be interpretable to build trust with users. Furthermore, the paper identifies a gap in evaluation benchmarks for XNLP—most existing datasets are not annotated with ground-truth explanations, making standard evaluation difficult.

Recent advances include rationale extraction models, counterfactual reasoning, and explanation generation using large language models. Some LLMs, such as GPT-4, are now capable of generating free-text justifications alongside their predictions, although these are often heuristic and may not reflect true reasoning paths.

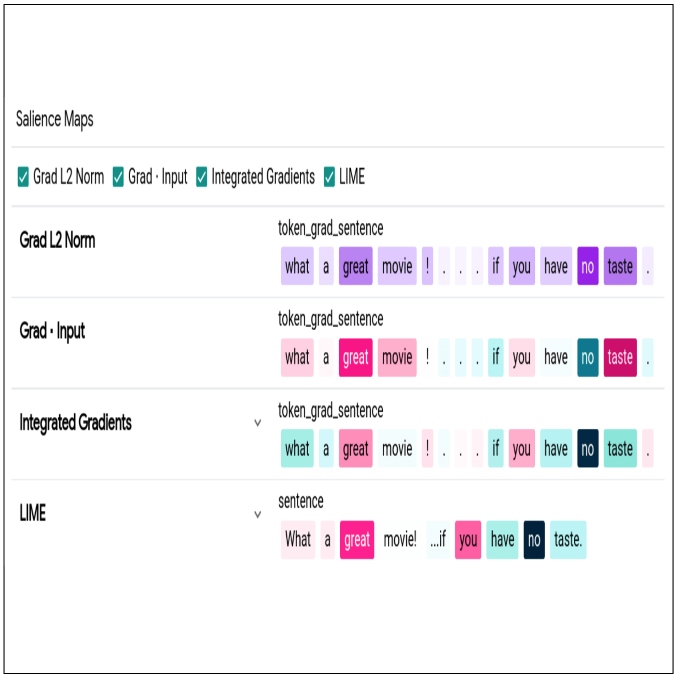
While XNLP is a growing field, it still grapples with several unresolved issues: scalability to large datasets, explanation consistency across similar inputs, and potential trade-offs between model performance and interpretability.

VII. Emotion and Sentiment Analysis in NLP

Emotion and sentiment analysis has become a cornerstone of NLP applications in domains such as social media monitoring, customer feedback, mental health analysis, and political opinion mining. These tasks aim to identify underlying emotional or sentiment-related cues within text, going beyond simple polarity classification to recognize fine-grained emotions like anger, joy, sadness, or surprise.

The EmotionX benchmark proposed by Huang et al. [6] represents one of the early large-scale efforts to tackle multilingual emotion detection in dialogues. It introduces datasets in both English and Chinese and provides an evaluation framework for models to classify utterances into emotions such as ”joy,” ”neutral,” ”sadness,” and ”anger.” The dataset comprises real human conversational data, which adds complexity due to colloquialisms, context dependency, and code-switching.

In their baseline experiments, the authors tested several models including LSTM, CNN, and attention-based RNNs. Attention mechanisms significantly improved performance by focusing on emotionally relevant parts of the sentence. However,



Salience Maps

☑ Grad L2 Norm ☑ Grad · Input ☑ Integrated Gradients ☑ LIME

Grad L2 Norm
token_grad_sentence
what a great movie ! . . . if you have no taste .

Grad · Input
token_grad_sentence
what a great movie ! . . . if you have no taste .

Integrated Gradients ˅
token_grad_sentence
what a great movie ! . . . if you have no taste .

LIME ˅
sentence
What a great movie! ...if you have no taste.

Fig. 5. Overview of XNLP techniques: post-hoc explanation (LIME, SHAP), attention-based heatmaps, and intrinsic rationale extraction [5].

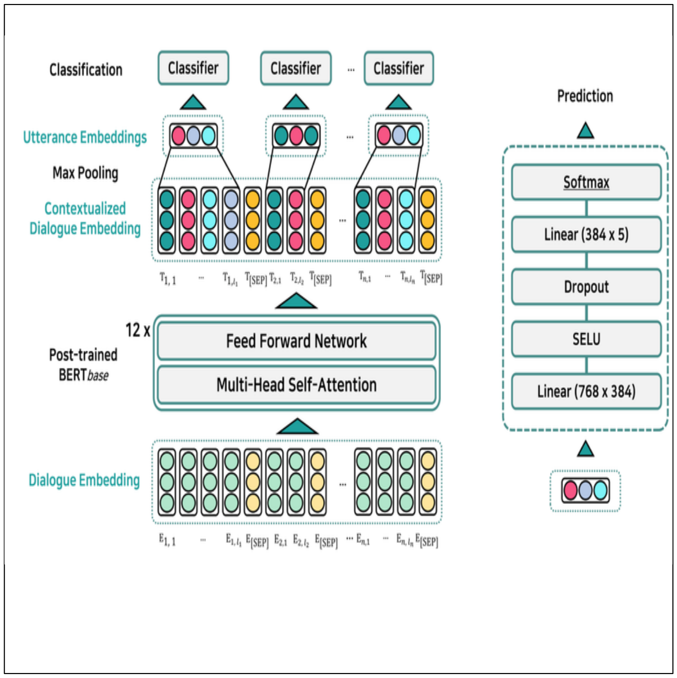

Classification
Classifier
Classifier
...
Classifier
Utterance Embeddings
Max Pooling
Contextualized Dialogue Embedding
$T_{1,1}$ ... T_{1,l_1} $T_{[SEP]}$ $T_{2,1}$ T_{2,l_2} $T_{[SEP]}$ $T_{n,1}$ ... T_{n,l_n} $T_{[SEP]}$
12 x
Post-trained BERTbase
Feed Forward Network
Multi-Head Self-Attention
Dialogue Embedding
$E_{1,1}$ ... E_{1,l_1} $E_{[SEP]}$ $E_{2,1}$ E_{2,l_2} $E_{[SEP]}$ ... $E_{n,1}$ ... E_{n,l_n} $E_{[SEP]}$
Prediction
Softmax
Linear (384 x 5)
Dropout
SELU
Linear (768 x 384)

Fig. 6. EmotionX benchmark setup and architecture comparison: LSTM vs. Attention-based RNN for emotion detection in dialogues [6].

context modeling remained a challenge—single utterances were often ambiguous without prior turns in the dialogue.

Since EmotionX, deep transformer-based models like BERT, RoBERTa, and DeBERTa have further advanced the field. These models are often fine-tuned on emotion-labeled datasets like GoEmotions (by Google), ISEAR, or SEMEVAL tasks, achieving high F1 scores even on multi-class classification tasks. Some recent models also incorporate external affective lexicons and sentiment knowledge graphs to improve semantic richness.

Despite improvements, emotion recognition still suffers from key challenges. Contextual ambiguity, sarcasm, irony, and domain mismatch (e.g., from Twitter vs. movie reviews) make generalization difficult. Additionally, emotion labels are inherently subjective—what feels neutral to one user may be perceived as sad by another. To mitigate this, some researchers explore multi-annotator agreement metrics and distributional labeling over hard categories.

New frontiers in this area include multimodal emotion recognition using video/audio/text data and emotion-aware dialogue generation in conversational AI. Emotion classification is also central to applications like mental health monitoring, where real-time emotion detection can assist in early diagnosis and intervention.

VIII. Question Answering in NLP – UnifiedQA

Question Answering (QA) is a central challenge in NLP, involving the ability to locate or generate answers from a given context, often requiring reasoning and language understanding. QA models must handle a variety of formats, such as extractive answers from passages, multiple-choice options, and free-form generative responses. Traditional approaches have involved task-specific architectures fine-tuned for each QA format. However, this siloed approach limits generalization and scalability.

UnifiedQA, introduced by Khashabi et al. [7], addresses this problem by proposing a single model that works across multiple QA formats using a unified text-to-text framework. Built upon the T5 architecture, UnifiedQA reformulates all QA inputs as text prompts and expects textual outputs. For example, both SQuAD-style span extraction and multiple-choice tasks are treated identically from the model's perspective:

Input: ”question: What is the capital of France? context: France is a country in Europe. Its capital is Paris.”

Output: ”Paris”

The key innovation lies in multi-format training. UnifiedQA was trained on over 20 different QA datasets, including SQuAD, TriviaQA, RACE, BoolQ, and Natural Questions, each with varied task structures. The model learns to understand the task implicitly from the input prompt structure, without requiring architectural modifications.

This approach led to strong zero-shot and few-shot generalization, allowing the model to adapt to unseen QA formats simply by phrasing the prompt correctly. Fine-tuning on just a few examples of a new format allows it to perform competitively, a hallmark of instruction-aware modeling.

UnifiedQA demonstrates significant improvements over baseline T5 on several benchmarks, especially in cross-format transfer scenarios. On OpenBookQA, ARC, and CommonsenseQA, it outperformed task-specific models even without access to additional task metadata.

The paper also evaluates scalability by experimenting with different T5 sizes—from T5-base to T5-11B—confirming that larger models yield better generalization with less training. Furthermore, the authors highlight the importance of prompt engineering: subtle differences in input phrasing can drastically influence performance, an insight that paved the way for future prompt-based and instruction-tuned models like FLAN-T5.

Despite its strengths, UnifiedQA is not without limitations. It is sensitive to prompt formats, requires significant compute for large-scale inference, and still lacks reasoning depth in multi-hop or common-sense-based QA scenarios. Nonetheless, it represents a significant step toward universal QA systems that can handle the full diversity of human questioning.

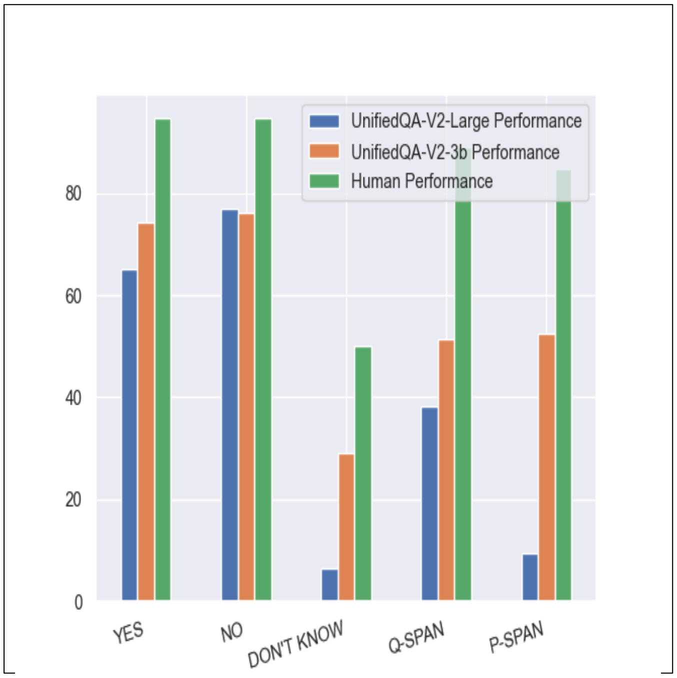


Fig. 7. UnifiedQA architecture and input formatting across multiple QA datasets using the T5 model [7].

IX. Biomedical NLP – BioBERT

The biomedical domain contains vast amounts of unstructured textual data—from clinical notes and scientific literature to electronic health records (EHRs). Extracting meaningful insights from this data requires specialized NLP models trained to understand domain-specific terminology, syntax, and structure. To this end, BioBERT, introduced by Lee et al. [8], represents a landmark in biomedical NLP by adapting the BERT architecture for medical and biological texts.

BioBERT is initialized with pretrained weights from BERT-base and then further pretrained on large biomedical corpora, specifically PubMed abstracts (4.5B words) and PMC full-text articles (13.5B words). This domain adaptation enables BioBERT to better capture the semantics of biomedical vocabulary and phraseology compared to generic language models.

The model architecture remains identical to BERT, using 12 transformer encoder layers, 768 hidden units, and 12 self-attention heads. However, the contextual embeddings produced after domain-specific pretraining exhibit significantly improved performance across several biomedical NLP tasks.

In their evaluation, the authors fine-tuned BioBERT on three major tasks:

Biomedical Named Entity Recognition (NER) – Identifying names of diseases, genes, proteins, and chemicals in medical texts.

Relation Extraction (RE) – Detecting associations (e.g., drug–disease, protein–protein) within sentence-level context.

Biomedical Question Answering (QA) – Answering clinical and research questions using datasets like BioASQ and PubMedQA.

BioBERT achieved state-of-the-art performance across all these benchmarks, outperforming previous methods such as CNN, BiLSTM-CRF, and domain-specific embeddings like word2vec trained on PubMed. For instance, on the NCBI Disease and BC5CDR datasets for NER, BioBERT showed a significant boost in F1 scores compared to vanilla BERT.

An important contribution of BioBERT is its demonstration that continued pretraining on domain-specific text—even without architectural changes—can yield large improvements. This inspired the development of further models like ClinicalBERT, BlueBERT, and SciBERT, each targeting specific biomedical subdomains or document types.

However, BioBERT also inherits some limitations of BERT, including high computational costs and limited generative capacity. Additionally, the biomedical domain presents challenges in interpretability and ethical usage, especially when NLP outputs are used to assist diagnosis or treatment decisions.

Nevertheless, BioBERT represents a pivotal step in integrating transformer models into the biomedical research workflow, powering applications from drug discovery to literature mining.

X. Prompt Engineering and Instruction Tuning – FLAN-T5

The growing capabilities of large language models (LLMs) such as GPT-3 and T5 have catalyzed the shift from traditional fine-tuning to prompt-based learning, where models are conditioned to perform tasks by formatting input text as instructions. Building on this paradigm, FLAN-T5 (Fine-tuned LAnguage Net) introduced by Chung et al. [9], represents one of the most impactful efforts in instruction tuning, enhancing model alignment with user intent through diverse, carefully structured task instructions.

FLAN-T5 is based on the T5 encoder-decoder architecture, which is further trained on a large collection of instruction-formatted tasks covering domains such as translation, summarization, reasoning, question answering, and classification. The objective is to teach the model to follow human-like instructions instead of relying on implicit prompts or fine-tuning for each individual task.

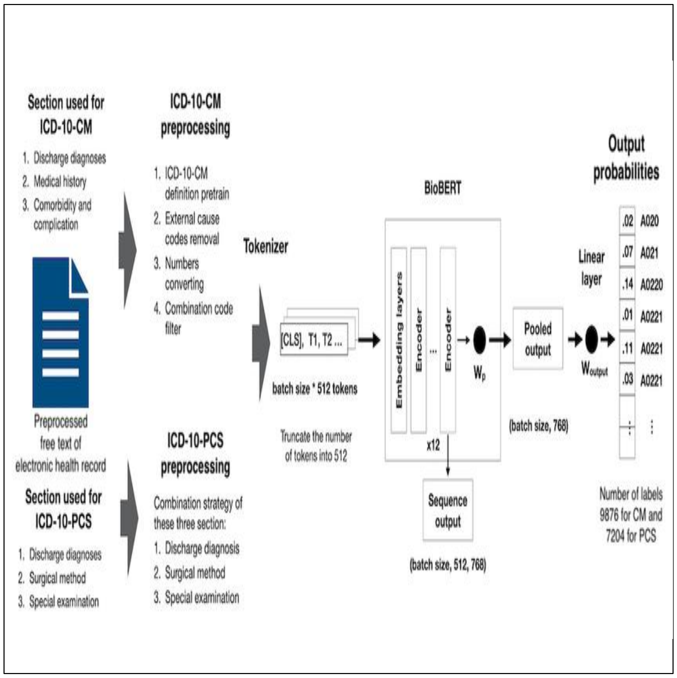


Fig. 8. BioBERT architecture and training pipeline: initialized from BERT and further pretrained on PubMed/PMC data [8].

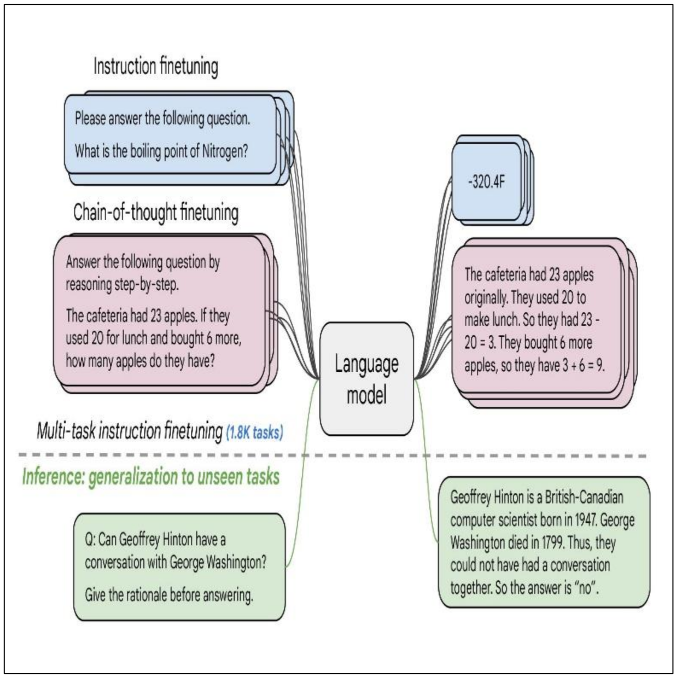

Instruction finetuning
Please answer the following question.
What is the boiling point of Nitrogen?
-320.4F
Chain-of-thought finetuning
Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?
The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.
Language model
Multi-task instruction finetuning (1.8K tasks)
Inference: generalization to unseen tasks
Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.
Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Fig. 9. Instruction tuning in FLAN-T5: training on a wide array of prompts improves performance and robustness across unseen tasks [9].

The key strategy in FLAN-T5 is supervised instruction tuning, where the model is trained on more than 1,800 tasks from various benchmarks (e.g., SuperGLUE, MMLU, BIG-Bench, and Natural Instructions). Each task is framed with diverse prompts like:

”Translate the following sentence into German: ...”

”Classify the sentiment of this tweet: ...”

”Answer the following question using common sense: ...”

This training regime improves zero-shot and few-shot generalization across unseen tasks and domains. FLAN-T5 models—from T5-Base to FLAN-T5-XXL—consistently outperform their non-instruction-tuned counterparts on multiple NLP and reasoning benchmarks, often rivaling or surpassing GPT-3 performance with fewer parameters.

One of the model’s major contributions is its instruction-following robustness. Unlike earlier models, FLAN-T5 performs well across varied phrasings of the same instruction, reducing sensitivity to prompt engineering and enhancing user-friendliness. Furthermore, FLAN-T5 shows strong abilities in cross-lingual generalization and logical reasoning, demonstrating broader capabilities than many task-specific models.

FLAN-T5 is also publicly released under a permissive license, making it widely adopted in research and industry. It serves as a foundation for Google’s own instruction-tuned models and has inspired further developments in chain-of-thought prompting and multi-task training.

However, limitations remain. While FLAN-T5 follows instructions better, it can still generate hallucinations or ambiguous answers, particularly in open-domain settings. It also inherits the computational intensity of large transformer models, which may hinder deployment in low-resource environments.

XI. Multimodal NLP – Flamingo: Vision-Language Models for Few-Shot Learning

Modern NLP systems are increasingly expanding beyond pure text to incorporate multimodal data, such as images, audio, and video. These capabilities are essential for applications like visual question answering (VQA), caption generation, and interactive agents. Flamingo, introduced by Alayrac et al. [10], is a state-of-the-art multimodal transformer designed for few-shot learning across vision-language tasks.

Flamingo builds upon pretrained language models and vision encoders, such as Chinchilla for language and Perceiver Resampler for images. It introduces a gated cross-attention mechanism, enabling visual inputs to condition the language model during generation. Unlike earlier multimodal models that require task-specific fine-tuning, Flamingo is trained in a few-shot regime and generalizes effectively to unseen tasks by simply conditioning on a few examples.

The architecture of Flamingo consists of a frozen vision encoder (e.g., CLIP or ViT), a frozen text decoder (Chinchilla or GPT-style), and learnable cross-attention modules inserted at specific layers. These cross-modal modules allow the model to align and integrate vision and language signals without retraining the entire network.

Flamingo was evaluated on 16 multimodal benchmarks, including VQAv2, OK-VQA, Image Captioning, SNLI-VE, and COCO. It achieved state-of-the-art few-shot performance, significantly outperforming previous models like CLIP, VisualBERT, and VinVL. In particular, it demonstrated strong performance with as few as 4–8 examples per task, making it both efficient and scalable.

One of Flamingo's major strengths is task flexibility. It can perform a variety of vision-language tasks without explicit retraining, including:

Visual Question Answering: ”What is the boy doing in the image?”

Image Captioning: ”Describe this scene.”

Visual Reasoning: ”Which image best matches the given description?”

The paper also explores instruction-style prompting, where few-shot examples help condition the model to follow specific task formats. Flamingo's success highlights the potential of frozen backbone models with modular learning layers, which reduce computational overhead during training and improve generalization.

Challenges remain in multimodal reasoning, such as handling long-context dependencies across modalities, resolving ambiguity, and reducing hallucinations in image-grounded generation. However, Flamingo represents a powerful step toward unified AI systems that bridge perception and language understanding.

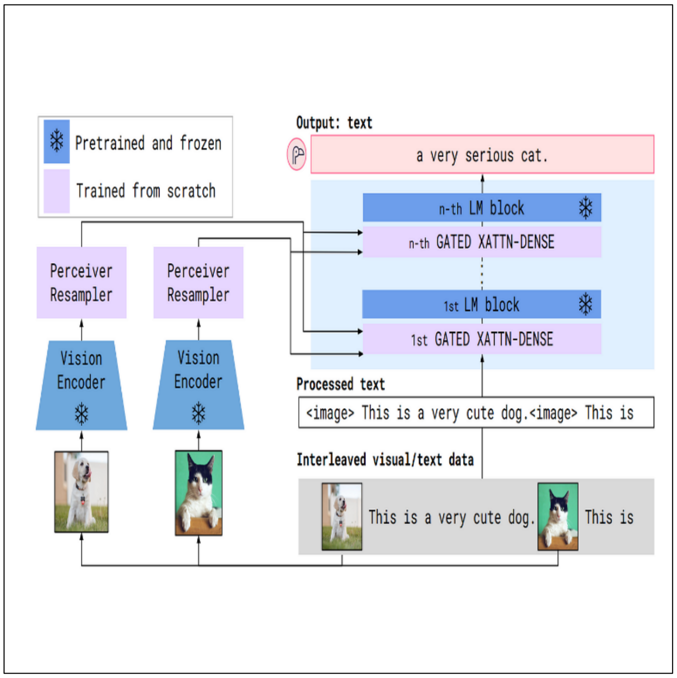


Fig. 10. Flamingo model architecture with vision encoder, text decoder, and gated cross-attention layers enabling few-shot multimodal learning [10].

XII. Ethics and Bias in NLP – The Stochastic Parrots Debate

As NLP systems scale in complexity, size, and influence, questions around ethical AI, bias, and responsible development have gained urgency. Large language models (LLMs), trained on web-scale corpora, inevitably inherit and amplify biases related to gender, race, culture, and socio-economic status. The landmark paper ”On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” by Bender et al. [12] initiated a critical dialogue within the NLP community regarding the social and environmental risks of scaling LLMs without accountability.

The term ”Stochastic Parrots” refers to the nature of these models as statistical systems that generate fluent, coherent text without understanding, often reproducing harmful stereotypes or misinformation. The paper argues that models like GPT-3 and its successors, when trained indiscriminately on massive, uncurated data, risk perpetuating biases and disinformation at scale.

Key ethical concerns highlighted include:

Bias and Toxicity: LLMs reflect the prejudices of their training data. This is particularly problematic in applications like hiring, law enforcement, and education, where such outputs may cause real-world harm.

Environmental Cost: Training massive models demands enormous computational resources, resulting in significant carbon emissions. For example, training a model like GPT-3 requires thousands of GPU hours, often powered by non-renewable energy.

Opacity and Accountability: LLMs are notoriously opaque, making it difficult to explain or trace their outputs, especially when used in high-stakes decision-making.

Data Consent and Ownership: Web-scraped data may include copyrighted or private material, raising questions about informed consent and data governance.

The authors call for transparency, documentation, and stakeholder engagement in model development. They propose principles such as data statements, model cards, and risk assessments to better understand and communicate what a model does and does not know. Furthermore, they stress the importance of interdisciplinary collaboration between technologists, ethicists, and affected communities.

The paper was influential not only academically but also politically, sparking internal debate within tech companies and prompting institutions to revise their AI ethics policies. It underscores the need for value-aligned development, where technical advancement is balanced with social responsibility.

In today's era of powerful LLMs like GPT-4, Gemini, and Claude, the issues raised in this paper remain highly relevant. Researchers and developers are now integrating bias mitigation, fairness auditing, and explanation mechanisms as essential components of the NLP pipeline.

XIII. Conclusion and Future Directions

This review presents a comprehensive exploration of recent advances in Natural Language Processing (NLP), covering twelve distinct yet interconnected domains through the lens of landmark research contributions. From the foundational transformer architectures such as BERT, GPT, and T5, to modern innovations in instruction tuning, multilingual generalization, multimodal integration, and explainability, we observe a continuous evolution of NLP capabilities driven by scale, transfer learning, and architectural ingenuity.

Each topic highlighted in this paper reflects the field's trajectory toward more general-purpose, instruction-following, and domain-adaptable models. UnifiedQA and FLAN-T5

TABLE I
COMPARISON OF MAJOR NLP MODELS AND TOPICS COVERED IN THE REVIEW

No.	Topic / Model	Architecture	Task Type	Key Feature / Innovation	Strengths	Limitations
1	BERT	Encoder (Transformer)	Understanding	Bidirectional attention with MLM + NSP	Strong context modeling, universal base	Not generative, slow inference
2	GPT	Decoder (Transformer)	Generation	Autoregressive language modeling	Fluent generation, few-shot learning	Unidirectional, hallucinations
3	T5	Encoder-Decoder	Unified Text-to-Text	Reframes all tasks as text generation	High flexibility across tasks	Resource intensive
4	XLM-R	Encoder (RoBERTa)	Multilingual NLP	100-language MLM training with CC100 corpus	Strong zero-shot cross-lingual transfer	Token imbalance, low-resource issues
5	XNLP	Various	Interpretability	Post-hoc & intrinsic explanation methods	Improves trust & transparency	Lacks consistent evaluation standards
6	EmotionX	LSTM/Attention	Emotion Detection	Multilingual emotion classification in dialogues	Robust baseline for emotion modeling	Ambiguity, context-dependent outputs
7	UnifiedQA	Encoder-Decoder (T5)	QA (Multi-format)	Unified training across diverse QA datasets	Task generalization, format-agnostic	Prompt sensitivity, reasoning gaps
8	BioBERT	Encoder (BERT)	Biomedical NLP	Domain-specific pretraining on PubMed/PMC	Improved NER & QA in medical domain	High resource requirements
9	FLAN-T5	Encoder-Decoder	Prompt Engineering	Instruction-tuned on 1,800+ tasks	Follows human-like prompts effectively	May still hallucinate or misinterpret
10	Flamingo	Multi-modal	Vision + Language	Gated cross-attention, few-shot vision-language	Flexible multimodal learning	Lacks deep visual reasoning
11	EmpatheticDialogues	Seq2Seq / BERT	Dialogue Systems	Emotion-conditioned responses in conversations	Improved empathy & relevance	Limited domain, hard to scale emotions
12	Stochastic Parrots	N/A (Ethics Paper)	Ethical AI	Risks of scale, bias, opacity in LLMs	Highlights responsible AI development	Critique-based; not a technical model

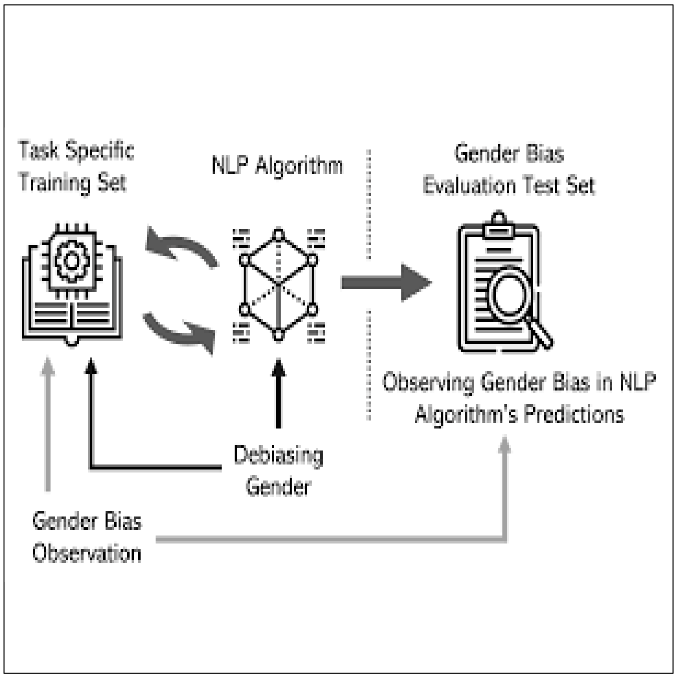


Fig. 11. Ethical concerns surrounding large-scale NLP systems, including bias, energy use, and lack of transparency [12].

demonstrate the power of formatting tasks as prompts, while BioBERT and XLM-R showcase the importance of domain and language-specific training. Flamingo exemplifies the growing convergence between vision and language, and EmotionX underscores the demand for emotionally intelligent systems. In parallel, the emerging subfield of Explainable NLP (XNLP) and the ethical scrutiny posed by ”Stochastic Parrots” emphasize the critical need for transparency, fairness, and responsible AI development.

Despite these advancements, several challenges remain. Large-scale models often lack interpretability, demand enormous computational resources, and risk encoding societal biases. Prompt sensitivity, hallucinations, and the lack of domain robustness still affect performance in practical scenarios. As models continue to scale and integrate into real-world systems, it becomes imperative to balance technical performance with ethical alignment, ensuring that NLP systems are not only powerful but also inclusive, transparent, and trustworthy.

Future directions in NLP may involve multi-modal grounding, human-in-the-loop learning, federated and privacy-preserving training, and universal representation frameworks that adapt seamlessly across languages, tasks, and modalities. Through this review, we hope to provide researchers, practitioners, and students with a consolidated perspective on the current landscape and emerging frontiers of NLP—equipping them to contribute responsibly to the next generation of language technologies.

ACKNOWLEDGMENT

The author wishes to express gratitude to the faculty members at Delhi Skill and Entrepreneurship University (DSEU) for their guidance and academic support. Special thanks to the global research community for open access to transformative NLP models and datasets, which made this survey possible.

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest with respect to the research, authorship, and/or publication of this review paper.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171-4186.

[2] T. Brown et al., ”Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[3] C. Raffel et al., ”Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1-67, 2020.

[4] A. Conneau et al., ”Unsupervised cross-lingual representation learning at scale,” in *Proc. ACL*, 2020, pp. 8440-8451.

[5] E. Mohammadi, K. Bhargava, and R. Gras, ”Explainable natural language processing: A comprehensive survey,” *ACM Computing Surveys*, vol. 55, no. 11, pp. 1-42, 2023.

[6] H.-H. Huang, C.-Y. Chen, K.-L. Liu, W.-B. Chen, B.-H. Huang, and H.-H. Chen, ”EmotionX-DLC: Self-attentive BiLSTM for detecting sequential emotions in dialogues,” in *Proc. SocialNLP Workshop*, 2018.

[7] D. Khashabi et al., ”UnifiedQA: Crossing format boundaries with a single QA system,” in *Findings of EMNLP*, 2020, pp. 1896-1907.

[8] J. Lee et al., ”BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020.

[9] H. W. Chung et al., ”Scaling instruction-finetuned language models,” arXiv preprint arXiv:2210.11416, 2022.

[10] J.-B. Alayrac et al., ”Flamingo: A visual language model for few-shot learning,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 23716-23736.

[11] H. Rashkin, E. Smith, M. Li, and Y. Choi, ”Towards empathetic open-domain conversation models: A new benchmark and dataset,” in *Proc. ACL*, 2019, pp. 5370-5381.

[12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, ”On the dangers of stochastic parrots: Can language models be too big?” in *Proc. FAccT*, 2021, pp. 610-623.

[13] A. Vaswani et al., ”Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998-6008.

[14] Y. Liu et al., ”RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

About the Author

Yash Rathore is currently pursuing his Diploma in Computer Engineering at Delhi Skill and Entrepreneurship University (DSEU), Delhi. His research interests include natural language processing, deep learning, multimodal learning, and ethical AI. He is particularly passionate about applying large language models to real-world problems and exploring the intersection of AI and human communication.