

## Natural Language Processing with AI: Progress, Models, and Emerging Directions

**Syeda Hina Shah** (Corresponding Author)

Harbin Institute of Technology, Shenzhen, China

Email: engrsyeda8@gmail.com

**Atif Rehman**

Department Of Computer Science, Muslim Youth University, Islamabad

Email: dcs232006@myu.edu.pk

**Kamal Khan**

Department Of Computer Science, University Of Makran

Email: kamalkhan@uomp.edu.pk

**Summayya Shabbir Baloch**

Department Of Computer Science, University Of Makran

Email: summayyabaloch634@gmail.com

**Shahnaz Muhammad**

University Of Engineering And Technology Khuzdar

Email: shahnazmuhammad2003@gmail.com

**Ghulam Yasin**

Department of Computer science, University of Makran Panjgur

Email: yasinarman38@gmail.com

### ABSTRACT

Natural Language Processing (NLP) has experienced a profound transformation with the integration of Artificial Intelligence (AI), particularly deep learning and attention-based architectures. This review provides a comprehensive and structured overview of the evolution of NLP, tracing its progression from rule-based and statistical approaches to modern deep neural networks and Large Language Models (LLMs). We examine foundational architectures, including recurrent neural networks and their limitations, before detailing the transformative impact of the Transformer architecture and self-attention mechanisms that enabled large-scale parallelization and contextual modeling. The paper further explores pre-training paradigms such as BERT and GPT, highlighting their roles in transfer learning, few-shot learning, and the emergence of general-purpose language intelligence. Recent advances in scaling laws, instruction tuning, and parameter-efficient fine-tuning techniques including adapters, prefix tuning, and low-rank adaptation are critically reviewed as solutions to computational

and deployment challenges. Emerging research directions are discussed with emphasis on multimodal large language models, domain-specific specialization, and architectural optimizations aimed at efficiency and generalization. Finally, the review addresses ethical challenges associated with LLMs, including bias, fairness, transparency, and governance, presenting current mitigation strategies and Responsible AI frameworks. By synthesizing architectural, methodological, and ethical perspectives, this review provides a unified reference for researchers and practitioners seeking to understand the current state, limitations, and future trajectories of AI-driven NLP systems.

## Introduction

Natural Language Processing (NLP), the discipline dedicated to facilitating meaningful interaction between computers and human languages, has undergone a revolutionary transformation with the integration of Artificial Intelligence (AI), particularly deep learning (Torfi et al., 2020). The field's rapid acceleration, particularly since the mid-2010s, is attributed to the confluence of technological advancements, notably the availability of massive training datasets and the parallel processing capabilities provided by modern Graphical Processing Units (GPUs) (Coates et al., 2018; Liddy, 2001). Deep learning methods and sophisticated neural models have advanced nearly all core NLP tasks and applications, ranging from text classification and machine translation to complex reasoning and generation (Torfi et al., 2020). The efficacy of these models, particularly the Transformer architecture and its derivatives, necessitates a systematic review to categorize, analyze, and compare the state-of-the-art approaches (Carolan et al., 2024). This paper provides an expert-level review of the progress of NLP with AI, tracing its evolution from traditional methods to modern Large Language Models (LLMs), examining key architectures, detailing emerging directions in efficiency and multimodality, and critically analyzing the vital domain of Responsible AI (Ferdaus et al., 2024). The subsequent sections of this report are structured to first establish the historical context of NLP, followed by a detailed analysis of foundational deep learning models. The discussion then progresses to the contemporary challenges of scaling, model efficiency, and the development of multimodal systems (Carolan et al., 2024). Finally, the review culminates in an essential examination of the ethical and governance frameworks required for trustworthy AI deployment (Sreerama & Krishnamoorthy, 2024).

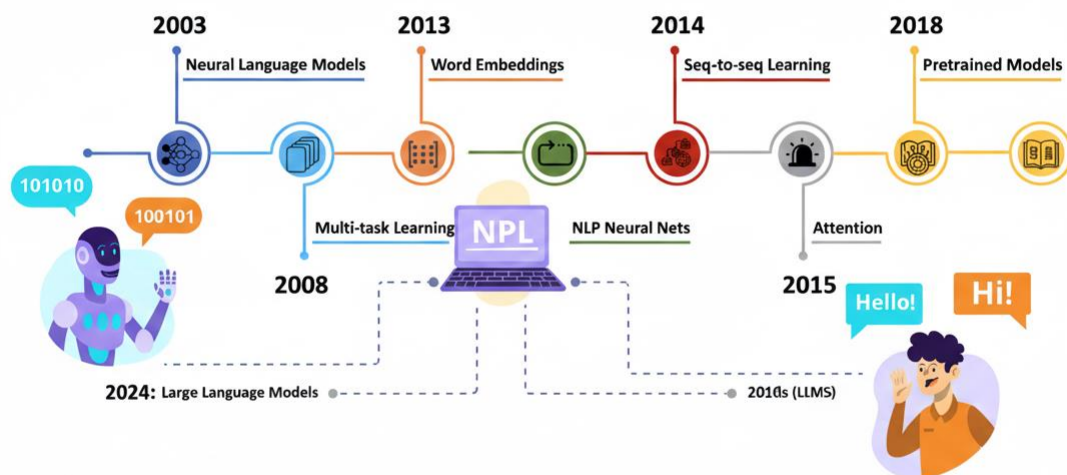
The history of NLP, spanning decades, has seen multiple dominant approaches, characterized initially by rule-based systems and subsequently by statistical machine learning (Jones, 1994). Traditional machine learning, including techniques such as Hidden Markov Models (HMMs), Conditional Random Fields (CRFs), Naïve Bayes, and Support Vector Machines, underpinned data-driven computation through statistics and probability throughout the late 20th and early 21st centuries (Qatawneh et al., 2025). The advent of deep learning instigated a significant paradigm shift a wholesale transformation where these traditional methodologies were either entirely replaced or significantly enhanced by neural models (Coates et al., 2018). This transition was not without its trade-offs, particularly in initial applications like neural machine translation (NMT). While NMT models proved superior to their statistical and rule-based counterparts, they demanded significantly more data and demonstrated poor

performance when applied outside the domain in which they were originally trained (Nanyonga et al., 2025; Kökver et al., 2025). Furthermore, early neural architectures struggled to handle rare words adequately and performed sub-optimally with sequences exceedingly approximately sixty words (Coates et al., 2018).

This dependence on vast data volumes and the lack of generalization highlighted a critical challenge: achieving state-of-the-art performance using deep learning often required sacrificing the data efficiency and domain robustness inherent in older, smaller models. The massive data dependency established during this initial shift continues to inform contemporary LLM research, emphasizing the ongoing necessity for research into domain adaptation techniques to ensure practical applicability across diverse, specialized fields (Jones, 1994).

**Figure 1.1 A Timeline of Evolutionary Shifts in NLP Modeling**

## The Evolution of Natural Language Processing



### The Foundational Architectures of Modern NLP

#### Rule-Based and Statistical Precursors

Before the ascendancy of deep neural networks, rule-based and statistical methods formed the backbone of computational linguistics (Jones, 1994). Rule-based systems, while relying on explicit linguistic knowledge and handcrafted rules, often exhibited excellent performance characteristics in narrow, well-defined domains (Liddy, 2001). For instance, specific rule-based NLP algorithms applied to clinical text analysis achieved exceptional performance, reporting sensitivity and specificity rates around 93.3 percent and 99.0 percent respectively, with F1-scores exceeding 0.90 at the record level (2025; Kökver et al., 2025).

In highly structured contexts, such as healthcare incident reporting, the combined identification strategy using traditional ICD codes alongside the rule-based NLP method provided the most comprehensive capture of specific episodes, such as fall-related events (Jones, 1994). The continued viability and specific success of these traditional, computationally simpler methods underscore a crucial point: high accuracy in critical, regulated applications often demands transparency and interpretability (Nanyonga et al., 2025; K  kver et al., 2025; Qatawneh et al., 2025). The inherent interpretability of rule-based systems serves as a necessary benchmark against the comparative opacity of deep learning models, where model decisions are less transparent, reinforcing the importance of clarity alongside performance in sectors like medicine and law (Sreerama & Krishnamoorthy, 2024).

### **The Deep Learning Transition: RNNs, LSTMs, and Limitations**

The initial transition to deep learning saw the widespread adoption of sequential models, such as Recurrent Neural Networks (RNNs) and their advanced variants, Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) (Torfi et al., 2020). Since language processing fundamentally depends on the order and sequence of words, RNNs were highly useful due to their internal mechanism allowing them to retain memory of previous elements while processing new ones (Coates et al., 2018).

Advanced sequential models, particularly LSTMs and GRUs, addressed the vanishing gradient problem inherent in basic RNNs, improving the ability to remember information from the distant past (Coates et al., 2018). These models established the state-of-the-art for sequence modeling and transduction problems, including language modeling and statistical machine translation (Vaswani et al., 2017).

However, the sequential processing nature of RNNs and LSTMs introduced a fundamental computational bottleneck (Zhao et al., 2025). Recurrent models inherently factor computation along the symbol positions of the input and output sequences, limiting the capacity for parallel processing (Nanyonga et al., 2025; K  kver et al., 2025). This dependence on sequential processing hindered the ability to efficiently model dependencies across very long distances within a text, paving the way for a more radically parallelizable architecture (Coates et al., 2018).

### **The Transformer: An Attention-Centric Revolution**

The limitations of recurrent architectures were fundamentally addressed by the introduction of the Transformer model, a state-of-the-art encoder-decoder architecture that relies exclusively on attention mechanisms (Vaswani et al., 2017). The Transformer was the first transduction model to entirely discard sequence-aligned RNNs and convolutions in favor of self-attention to compute representations of its input and output sequences (Zhao et al., 2025). This reliance on attention unlocked massive parallelization capabilities that were impossible under the sequential paradigm.

At the core of this revolution is the Self-Attention Mechanism, which allows the model to calculate dependencies between tokens regardless of their distance in the input or output sequence (Vaswani et al., 2017). An attention function operates by mapping a query (Q) and a set of key-value pairs (K, V) to an output vector

(Yaslioglu, 2025). The ability to model these dependencies globally, rather than sequentially, was a significant breakthrough (Vaswani et al., 2017).

Furthermore, the Transformer introduced Multi-Head Attention (Vaswani et al., 2017; Yaslioglu, 2025). This mechanism is crucial because it permits the model to jointly attend to information from different representation subspaces concurrently, addressing the issue where averaging across a single attention head might inhibit overall performance (Zhao et al., 2025). In addition to the attention mechanisms, Positional Encoding is a necessary component (Vaswani et al., 2017). Since the parallel architecture eliminates the inherent notion of word order present in RNNs, positional encodings are added to the input embeddings to inject sequence position information (Vaswani et al., 2017).

The explicit architectural decision to abandon recurrence and embrace parallel attention immediately transformed the computational landscape of NLP modeling. This shift fundamentally enabled the training of models with billions of parameters on vast datasets—the foundation of the LLM era which would have been computationally prohibitive using sequential architectures (Vaswani et al., 2017; Coates et al., 2018). The Transformer thus stands as the true catalyst for the current exponential scaling of AI in language processing (Yaslioglu, 2025).

**Table 2.1: Key Evolutionary Shifts in NLP Modeling**

| Era                    | Core Approach                        | Representative Model/Technique      | Key Limitation Addressed                                                      |
|------------------------|--------------------------------------|-------------------------------------|-------------------------------------------------------------------------------|
| Statistical/Rule-Based | Probabilistic Modeling               | Hidden Markov Models (HMMs), CRFs   | Data sparsity, lack of deep contextual understanding (Jones, 1994)            |
| Early Deep Learning    | Sequential Processing                | RNNs, LSTMs                         | Vanishing gradients, difficulty modeling long sequences (Coates et al., 2018) |
| Attention-Based        | Parallelized Attention               | Transformer                         | Sequential bottleneck of RNNs, distance limitations (Vaswani et al., 2017)    |
| Pre-trained Models     | Transfer Learning (Self-Supervision) | BERT (MLM), GPT (Unidirectional LM) | Need for heavily engineered task-specific architectures (Devlin et al., 2019) |

### Pre-training Paradigms and Model Families

#### Bidirectional Encoder Representations from Transformers (BERT)

The development of pre-trained language models marked the next critical advancement, establishing the "pre-train, fine-tune" paradigm (Devlin et al., 2019). The Bidirectional Encoder Representations from Transformers (BERT) model, introduced in 2018, leveraged the Transformer's encoder stack to achieve deep

language understanding (Devlin et al., 2019). BERT achieved remarkable state-of-the-art results across eleven NLP tasks, including pushing the GLUE benchmark accuracy to 80.4 percent and SQuAD v1.1 Test F1 score to 93.2 (Devlin et al., 2019; Khan et al., 2025).<sup>2</sup>

BERT's primary innovation was the utilization of deep bidirectional representations (Devlin et al., 2019). Unlike previous models that derived context-sensitive features from shallow concatenations of left-to-right and right-to-left language models, BERT was trained to condition jointly on both left and right context in all layers (Devlin et al., 2019). This bidirectionality was realized through the Masked Language Model (MLM) objective, which randomly masks a portion of the input tokens, requiring the model to predict the original vocabulary ID of the masked word based purely on its surrounding context (Zhao et al., 2025).

The MLM objective permits the pre-training of a deep, bidirectional Transformer model, overcoming the unidirectional constraint that limited the context fusion capabilities of earlier generative models (Devlin et al., 2019). Additionally, BERT utilized a Next Sentence Prediction (NSP) task, designed to model inter-sentence semantic relationships, enhancing its utility for tasks like question answering and semantic inference (Khan et al., 2025). The empirical success of BERT across highly diverse tasks confirmed the hypothesis that pre-training a general language representation significantly reduces the need for complex, heavily engineered task-specific architectures (Zhao et al., 2025). This success cemented the methodology of transfer learning as the dominant research trajectory for NLP understanding tasks.

## **Generative Pre-trained Transformers (GPT) and Unsupervised Learning**

While BERT focused on understanding via the encoder stack, the Generative Pre-trained Transformer (GPT) series, starting with GPT-2, focused on generation using the decoder stack and a unidirectional language modeling objective (Radford et al., 2019). GPT-2 demonstrated the potential for language models to operate as unsupervised multitask learners (Radford et al., 2019). These models exhibited the capability to perform downstream tasks in a zero-shot setting without any specific parameter or architecture modification achieving promising and competitive results across a wide range of tasks (Radford et al., 2019).

The scaling of the GPT family culminated in GPT-3 (Brown et al., 2020), which expanded the capacity and demonstrated unprecedented capabilities in few-shot learning (Brown et al., 2020). For these tasks, GPT-3 was applied without any gradient updates or fine-tuning; the task definition and few-shot demonstrations were provided purely via text interaction with the model a concept known as in-context learning (Brown et al., 2020; Radford et al., 2019). GPT-3 achieved strong performance on translation, question-answering, and cloze tasks, setting new state-of-the-art results on several datasets (Brown et al., 2020).

The spontaneous emergence of complex reasoning abilities, such as few-shot learning, solely through model scale and skillful textual interaction, suggested that general knowledge is encoded so effectively within these models that external fine-tuning becomes secondary for rapid task adaptation (Lane et al., 2025). This finding provided compelling evidence for the potential of massive scaling to yield general-purpose AI capabilities, shifting the focus from updating model weights to improving

prompt design and in-context strategies (Radford et al., 2019).

## **Large Language Models (LLMs): Scaling, Performance, and Efficiency**

### **The Phenomenon of Scaling and Few-Shot Learning**

The contemporary era of NLP is defined by Large Language Models (LLMs), which continue the scaling trend established by GPT (Luo et al., 2024). The immense success and advanced capabilities of these models such as textual understanding, natural language planning, and instruction following are directly attributed to their extensive number of parameters (Luo et al., 2024; Touvron et al., 2023). Key technical principles rely heavily on deep learning techniques, using the Transformer's self-attention framework to enable complex interactions among sequence elements during training on massive datasets (Liu et al., 2025).

While scaling provides capacity, a critical post-training technique, Instruction Tuning (IT), provides utility (Jiang et al., 2023). IT refers to the supervised training process on a dataset of (instruction, output) pairs (Liu et al., 2025). This process is essential for aligning the LLM's inherent next-word prediction objective with the user's objective, ensuring that the LLM adheres effectively to human instructions (Jiang et al., 2023). The simultaneous necessity of massive scaling to encode capacity and alignment techniques like IT to ensure useful behavior highlights a complex challenge: scaling alone provides the linguistic capacity, but alignment is required to ensure responsible and functional deployment (Wang et al., 2024b; Jiang et al., 2023). However, the "the bigger, the better" positive scaling law is not universal (McKenzie et al., 2023). Research into inverse scaling has identified scenarios where LLM performance on certain novel tasks decreases as model size increases or follows a non-monotonic trend (McKenzie et al., 2023). These phenomena are critical, as they reveal intrinsic limitations of current architectures and training paradigms, suggesting new research directions beyond simple scale maximization, focusing instead on structural improvements and prompting strategies like Chain-of-Thoughts (McKenzie et al., 2023).

### **Parameter-Efficient Fine-Tuning (PEFT) Strategies**

The continued growth in LLM parameter counts has made traditional full fine-tuning computationally prohibitive for most users and tasks (Khan et al., 2025). This challenge catalyzed the development of Parameter-Efficient Fine-Tuning (PEFT) methods, designed to adapt massive pretrained models to specific downstream tasks by updating only a small, dedicated set of parameters ( $\Phi$ ), while keeping the vast majority of the LLM's foundational parameters ( $\Theta$ ) frozen (Khan et al., 2025).

Several PEFT mechanisms have gained prominence due to their efficiency and effectiveness, particularly when full model updates are impractical or overly expensive (Khan et al., 2025). Prompt Tuning and Prefix Tuning are adaptation methods that introduce learnable parameters as contextual information (soft prompts or prefixes) (Khan et al., 2025). Prefix tuning updates a small section of parameters consistently across every layer of the model, enabling modification with minimal structural change (Khan et al., 2025).

In contrast, Adapter Training introduces small neural networks, termed adapters, strategically inserted between or within the layers of the pretrained model, where only the adapter parameters are updated during training (Khan et al., 2025). Furthermore,

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Low-Rank Adaptation (LoRA) employs low-rank factorization techniques to compress the update matrices effectively, resulting in significant reductions in computational requirements and the number of trainable parameters while preserving model accuracy (Khan et al., 2025).

A novel approach, Semantic Knowledge Tuning (SK-Tuning), utilizes the frozen LLM’s zero-shot capability to extract the semantic representation from a prompt composed of meaningful words instead of random tokens (Khan et al., 2025). This semantic representation is then processed by a small trainable adapter, and the resulting semantic projection is concatenated with the input text embedding (Khan et al., 2025). This tailored injection of semantic context is a sophisticated method for resource-efficient adaptation (Khan et al., 2025). The resource efficiency provided by these PEFT techniques is not merely an optimization but is the necessary enabling factor for the widespread deployment of LLMs, serving as the bridge between centralized model pre-training and decentralized, domain-specific application across various hardware constraints (Khan et al., 2025).

**Table 4.1: Overview of Parameter-Efficient Fine-Tuning (PEFT) Techniques**

| Technique                  | Mechanism Summary                                                                   | Parameters Updated                  | Goal                                                                      |
|----------------------------|-------------------------------------------------------------------------------------|-------------------------------------|---------------------------------------------------------------------------|
| Adapter Training           | Insertion of small neural networks (adapters) within model layers                   | Adapter parameters only             | Domain specificity with minimal resource overhead (Khan et al., 2025)     |
| Prefix Tuning              | Addition of learnable parameters (prefixes) to input of every layer                 | Prefix tokens only                  | Task adaptation across multiple layers (Khan et al., 2025)                |
| LoRA (Low-Rank Adaptation) | Uses low-rank factorization to approximate weight matrix updates                    | Low-rank update matrices            | Efficient parameter compression and memory savings (Khan et al., 2025)    |
| Gradient Mask Tuning (GMT) | Selectively updates parameters based on gradient information (masking small values) | A sparse subset of model parameters | Reducing redundancy while preserving network structure (Luo et al., 2024) |

**Architectural and Training Optimizations**

The Transformer’s immense success is coupled with a persistent architectural challenge: the attention mechanism incurs a quadratic computational and memory cost  $O(n^2)$  relative to the input sequence length  $n$  (Luo et al., 2024; Yaslioglu, 2025). To mitigate this constraint, various optimization strategies have been developed, broadly categorized as efficient Transformers (Luo et al., 2024).

These techniques focus on reducing complexity, often aiming for linear cost  $O(n)$  or  $O(n \times \sqrt{n})$  (Yaslioglu, 2025). Examples include Sparse Transformers, which introduce fixed-pattern attention (strided and local attention), Linformer, which uses low-rank approximations of the attention matrix, and

Reformer, which utilizes locality-sensitive hashing (Luo et al., 2024). These methods primarily aim to approximate dense attention to preserve performance while substantially enhancing efficiency (Yaslioglu, 2025).

A significant finding demonstrates that structured sparsity in the attention mechanism provides benefits beyond computational savings (Yaslioglu, 2025). Research indicates that introducing high levels of sparsity does not degrade performance but rather acts as a powerful regularizer, resulting in significant improvements in model generalization and validation accuracy on tasks like SST-2 (Yaslioglu, 2025). This suggests that the traditionally dense attention mechanism may be prone to overfitting, and introducing intentional constraints forces the model to prioritize only the most salient connections, thereby yielding more robust representations (Yaslioglu, 2025).

Further training optimizations include techniques like Gradient-Mask Tuning (GMT) (Luo et al., 2024). GMT addresses the redundancy found in dense parameter updates by selectively updating parameters based on their gradient information (Luo et al., 2024). Specifically, after calculating gradients during backpropagation, GMT masks those parameters with small absolute gradient values (Luo et al., 2024). This method reduces redundancy while preserving the original network structure and exhibits computational efficiency comparable to vanilla training, ensuring that resources are only spent on optimizing the most influential parameters for the target task (Luo et al., 2024).

## **Emerging Directions: Multimodality and Specialized Models**

### **Multi-Modal Large Language Models (MM-LLMs)**

The evolution of LLMs has extended their capabilities beyond text, resulting in Multi-Modal Large Language Models (MM-LLMs) that integrate information from images, audio, and video (Carolan et al., 2024; Liu et al., 2025). This advancement enables a wide array of sophisticated applications, including image captioning, text-to-speech, and text-to-video synthesis (Carolan et al., 2024). MM-LLMs are generally developed through two main approaches: augmenting existing LLMs with multimodal capabilities or designing entirely new multimodal architectures (Carolan et al., 2024). The core technical challenge in MM-LLMs lies in aligning the features extracted by dedicated non-textual encoders (e.g., visual models like CLIP ViT-L/14) with the LLM's input token space (Carolan et al., 2024). This alignment is typically mediated by a projection layer or a more complex visual abstractor module (Carolan et al., 2024). For instance, in some models, a simple linear projection layer is trained to align frozen visual features with the frozen LLM, ensuring effective coordination between the two modalities (Carolan et al., 2024).

More advanced models often employ a visual abstractor module (Carolan et al., 2024). This module addresses the high computational complexity arising from integrating dense visual knowledge, which translates into lengthy token sequences, into the attention mechanism (Carolan et al., 2024). The abstractor compresses the extracted visual features into a few learnable tokens (Carolan et al., 2024). These compressed tokens are then seamlessly integrated with the word embeddings of the input sequence, reinforcing the ongoing importance of efficiency research at the interfaces of different modalities to maintain computational feasibility (Carolan et al., 2024).

Online ISSN

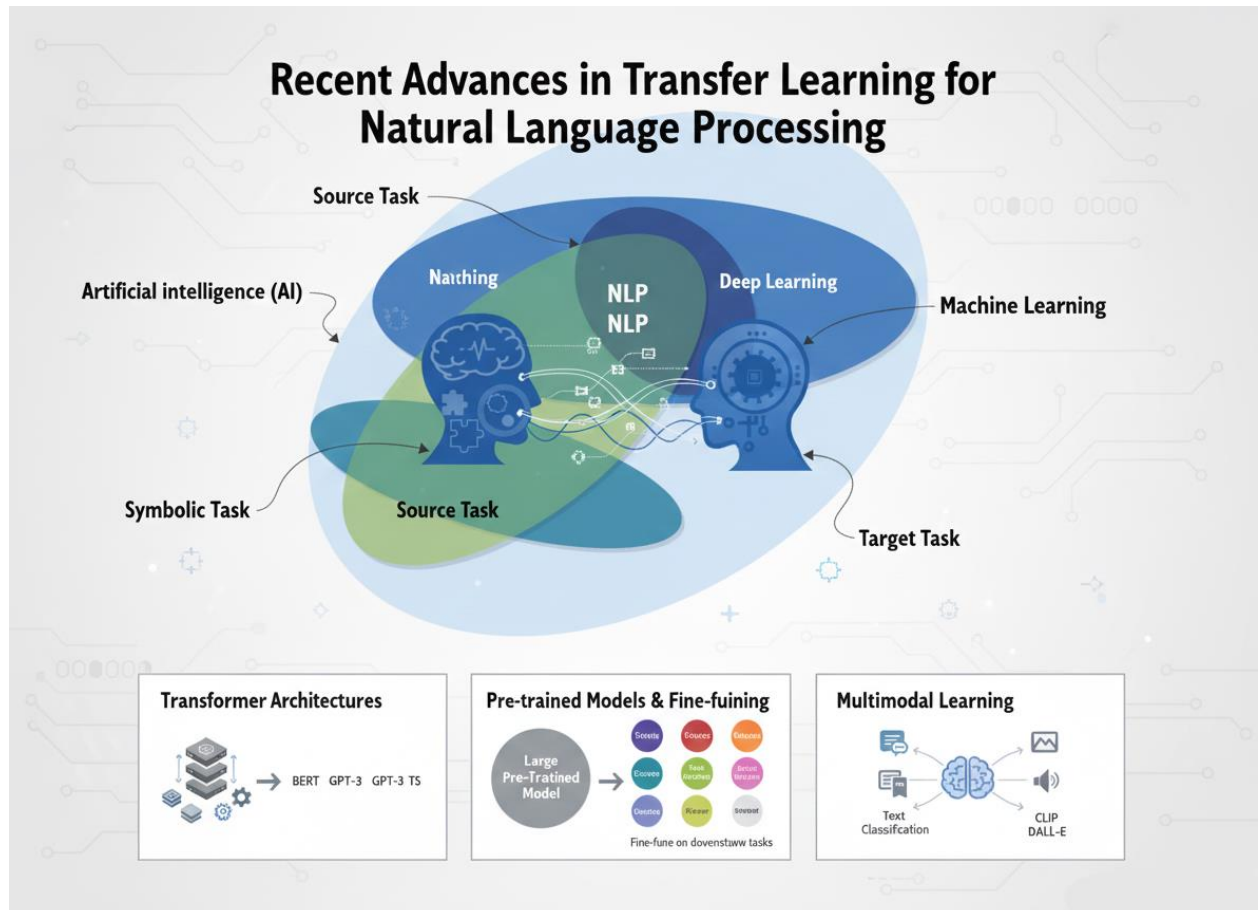
3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

## The Intersection of Transfer Figure Learning, Transformer Architectures, and Multimodality



### Domain-Specific Applications and Future Architectures

A growing trend in AI development is the customization and specialization of foundational models for expert domains (Liu et al., 2025). Specialized LLMs, such as AstroLLaMA and AstroLLaVA, are often built upon robust, general-purpose base architectures like LLaMA-2 or LLaMA-3 (Liu et al., 2025). These models are then extensively trained on domain-specific corpora, sometimes incorporating multimodality, as seen in AstroLLaVA, which integrates astronomical images alongside text through visual encoders (Liu et al., 2025).

The training data for these specialized models frequently includes texts from introductions and conclusions of relevant papers, alongside LLM-generated Question-Answering (QA) pairs derived from arXiv content, reflecting a shift toward interactive and precise domain tasks (Liu et al., 2025). The successful deployment of these expert models validates the core principle of transfer learning: a robust, general foundation model can be adapted quickly and efficiently to yield expert-level performance in niche fields (Devlin et al., 2019; Brown et al., 2020). This progression suggests that the economic and research trajectory for many industries will focus on

customizing existing models using PEFT and specialized datasets, rather than undertaking the resource-intensive process of training entirely new models from scratch for every application (Liu et al., 2025).

## Responsible AI: Ethical Challenges and Governance

### Sources and Quantification of Bias

As LLMs become ubiquitous, addressing the ethical liabilities inherent in large-scale model training is paramount (Ferdaus et al., 2024; Sreerama & Krishnamoorthy, 2024). LLMs inherit stereotypical social biases and potential toxicity from the vast, uncurated corpora on which they are built (Wang et al., 2019; Sreerama & Krishnamoorthy, 2024). Key ethical challenges that must be addressed include toxicity, stereotype bias, robustness, privacy risks, and fairness (Ferdaus et al., 2024). Evaluating the trustworthiness of LLMs requires assessing their alignment with social norms, interpretability, and reasoning capacity (Ferdaus et al., 2024).

The methodology for quantifying bias has evolved considerably (Wang et al., 2019). For early, static word embeddings, bias was typically assessed using vector distances or similarities (Wang et al., 2019). However, this approach failed for modern, contextualized embeddings like BERT, which do not have a single static vector for a word (Wang et al., 2019). Researchers thus shifted to using Natural Language Inference (NLI)-driven probes to measure bias as a contextual behavior (Wang et al., 2019).

The NLI-driven approach measures bias by quantifying invalid inferences (Wang et al., 2019). Probes are constructed using inherently neutral sentence pairs that test for stereotypes (e.g., gender bias related to occupations) (Wang et al., 2019). The premise and hypothesis are designed such that the correct label should be neutral, yet a biased model representation might incorrectly predict entailment or contradiction (Wang et al., 2019). Aggregated metrics such as Net Neutrality and Fraction Neutral quantify the model's deviation from neutrality across these test sets (Wang et al., 2019). This shift to measuring bias through functional, task-specific behavior provides a more relevant metric for real-world evaluation and alignment (Ferdaus et al., 2024).

### Mitigation and Debiasing Techniques

Strategies for mitigating bias and promoting fairness involve a comprehensive approach covering data preprocessing, algorithmic transparency, and post-hoc debiasing methods (Sreerama & Krishnamoorthy, 2024). One established post-hoc technique for static embeddings (like GloVe) involves the projection method (Wang et al., 2019). This method identifies the dominant direction defining a specific bias (the vector difference between "he" and "she" for gender bias) and subsequently removes this derived subspace from all embedded vectors, effectively attenuating the bias (Wang et al., 2019).

Applying this projection method directly to all layers of deep contextualized models, however, proved ineffective (Wang et al., 2019). A critical breakthrough demonstrated that bias mitigation in models like BERT or ELMo must target the **static, non-contextual components** (Wang et al., 2019). Specifically, the projection mechanism is successfully applied only to the subword embeddings or the first, non-contextual layer of the model (Wang et al., 2019). By removing the bias direction

from this initial mapping of tokens to vectors, researchers found they could increase neutrality in the subsequent NLI predictions (Wang et al., 2019). This finding strongly suggests that the initial embedding layer is the primary point of bias injection, necessitating that future research prioritize low-level embedding manipulations before high-level contextualization propagates the bias throughout the deep Transformer layers (Sreerama & Krishnamoorthy, 2024).

## Responsible AI Frameworks and Governance

Given the increasing deployment of LLMs in high-stakes domains, establishing robust governance frameworks is non-negotiable (Sreerama & Krishnamoorthy, 2024). Responsible AI (RAI) practices mandate attention to qualities like fairness, security, sustainability, and privacy (Sreerama & Krishnamoorthy, 2024). Organizations are establishing comprehensive internal policies, often informed by standards such as the National Institute of Standards and Technology (NIST) Risk Management Framework, to guide the safe and impactful use of generative AI in research and application (Sreerama & Krishnamoorthy, 2024).

For educational and high-risk applications, robust mechanisms are needed for the automatic detection of biases and violations of responsible use (Ferdaus et al., 2024). One scalable method introduced is the LLM-as-a-Judge methodology (Ferdaus et al., 2024). This approach utilizes one LLM to automatically evaluate the outputs of other LLMs, assessing them for biased responses against synthetic question-based datasets (Ferdaus et al., 2024). The use of synthetic data mimicking student-chatbot interactions, for example, has demonstrated that some models exhibit more bias than others, emphasizing the need for careful selection and continuous evaluation (Ferdaus et al., 2024). The necessity of automated accountability measures like the LLM-as-a-Judge demonstrates that the complexity and scale of modern AI systems require quantifiable, machine-executable protocols to maintain public trust and achieve regulatory compliance (Sreerama & Krishnamoorthy, 2024).

## Conclusion

Natural Language Processing, fueled by deep learning and the Transformer architecture, has progressed from dependence on sequential recurrent models to the parallel, attention-only paradigm (Vaswani et al., 2017). Key milestones included the development of highly effective pre-trained representation models: the bidirectional encoder BERT for language understanding and the unidirectional decoder GPT for advanced generation and few-shot learning (Devlin et al., 2019; Brown et al., 2020). The success of scaling has enabled the emergence of LLMs, where capacity and linguistic knowledge are enhanced by sheer parameter count, yet alignment techniques like Instruction Tuning are crucial for translating capacity into utility (Luo et al., 2024; Jiang et al., 2023). Critical engineering advancements, such as Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA and Prefix Tuning, have addressed the computational burden of deployment, making LLMs accessible for domain-specific adaptation (Khan et al., 2025). Furthermore, architectural optimizations, including sparse attention and gradient masking techniques, are not only enhancing efficiency by addressing the quadratic complexity barrier but are also proving to act as powerful regularizers, yielding more robust model generalization (Yaslioglu, 2025;

Luo et al., 2024). Despite this rapid progress, unresolved challenges persist, particularly in understanding non-monotonic scaling behaviors and ensuring comprehensive, deep-seated mitigation of inherited social biases (McKenzie et al., 2023; Wang et al., 2019).

## Future Research Trajectories

Future research in NLP with AI is expected to pursue several key trajectories (Carolan et al., 2024). The trend toward multimodal LLMs (MM-LLMs) will continue, driven by the necessity of developing sophisticated feature alignment and compression mechanisms to efficiently integrate high-dimensional non-textual data, such as images and video (Carolan et al., 2024; Liu et al., 2025). This will push the boundaries of generalized intelligence and task complexity (Liu et al., 2025).

The pursuit of efficiency and specialization will also intensify (Khan et al., 2025). The confirmed utility of specialized models, built on robust foundational architectures and tailored with domain-specific data and PEFT, indicates a trajectory where customized, expert-level AI systems become the norm (Liu et al., 2025). Finally, the domain of Responsible AI will demand increasing rigor, necessitating continued effort to integrate formal governance frameworks (e.g., NIST RMF) and automated accountability measures, such as the LLM-as-a-Judge methodology, to ensure that the transformative capabilities of NLP are deployed ethically, securely, and fairly across all high-stakes applications (Sreerama & Krishnamoorthy, 2024; Ferdaus et al., 2024).

## References

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., et al. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Carolan, K., Fennelly, L., & Smeaton, A. F. (2024). A Review of Multi-Modal Large Language and Vision Models. *ArXiv*, abs/2404.01322.
- Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., & Andrew, N. (2018). Deep learning with cots hpc systems.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.
- Ferdaus, M. M., Abdelguerfi, M., Ioup, E., Niles, K. N., Pathak, K., & Sloan, S. (2024). Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models. *ArXiv*, abs/2407.13934.
- Jiang, J., He, S., Li, Y., et al. (2023). Instruction Tuning for Large Language Models. *ArXiv*, abs/2308.10792.
- Jones, K. S. (1994). Natural language processing: a historical review. In *Current Issues in Computational Linguistics: in Honour of Don Walker* (pp. 3–16). Springer.
- Khan, R. M. A., Hossain, M. S., Mukta, M., et al. (2025). Comparative Analysis of BERTurk and Gemma Models for Turkish Text Classification by LoRA Fine-Tuning. *Conference Paper*.
- Kökver, Y., Pektaş, H. M., & Çelik, H. (2025). Artificial intelligence applications in education: Natural language processing in detecting

- misconceptions. *Education and Information Technologies*, 30(3), 3035-3066.
- Lane, H., & Dyshel, M. (2025). Natural language processing in action. Simon and Schuster.
- Liddy, E. D. (2001). Natural language processing.
- Liu, J., Chen, K., Li, Y., Feng, P., Bi, Z., et al. (2025). Recent advances in astronomy-specific LLMs, highlighting representative models such as AstroLLaMA and AstroLLaVA.
- Luo, P., Wu, Q., Liu, B., et al. (2024). Gradient-Mask Tuning (GMT). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13), 14616-14624.
- McKenzie, R., Radev, D., et al. (2023). Inverse Scaling Prize.
- Nanyonga, A., Joiner, K., Turhan, U., & Wild, G. (2025). Applications of natural language processing in aviation safety: A review and qualitative analysis. In *AIAA SciTech 2025 Forum* (p. 2153).
- Qatawneh, A. M. (2025). The role of artificial intelligence in auditing and fraud detection in accounting information systems: moderating role of natural language processing. *International Journal of Organizational Analysis*, 33(6), 1391-1409.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.
- Sreerama, J., & Krishnamoorthy, G. (2024). Ethical Considerations in AI: Addressing Bias and Fairness in Machine Learning Models. *Journal of Knowledge Learning and Science Technology*, 1(1).
- Torfi, A., Shirvani, A., & Ahmad, M. (2020). Natural Language Processing Advancements By Deep Learning. *ArXiv*, abs/2003.01200.
- Touvron, H., Lavril, T., Izacard, G., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.13971.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems* 30.
- Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., & Xiong, C. (2019). A transparent framework for evaluating unintended demographic bias in word embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., & Xiong, C. (2024b). BERTurk model achieved 74.13% accuracy and 72.17% F1 score. *Conference Paper*.
- Yaslioglu, M. M. (2025). Attention is All You Need Until You Need Retention. *ArXiv*, abs/2501.09166.
- Zhao, D. (2025). The impact of AI-enhanced natural language processing tools on writing proficiency: An analysis of language precision, content summarization, and creative writing facilitation. *Education and Information Technologies*, 30(6), 8055-8086.