Artificial Intelligence Applications And Innovations Vol. 1 No. 4 (2024)

Article history:
Received 7 June 2024
Revised 25 July 2024
Accepted 02 September 2024
Published online 01 October 2024

From Statistical Models to LLMs: A Comprehensive Survey of Language Model Evolution

Maryam. Majidi[1], Hamid. Hassanpour[1*]

[1] Department of Computer Engineering and IT, Shahrood University of Technology, Shahrood, Iran

*** Corresponding author email address**: h.hassanpour@shahroodut.ac.ir

A r t i c l e I n f o

Article type:
Review Article

How to cite this article:
Majidi, M., & Hassanpour, H. (2024). A Novel U-Net Architecture with Attention Mechanism for Image Denoising. *Artificial Intelligence Applications and Innovations, 1*(4), 55-75.
https://doi.org/10.61838/jaiai.1.4.5

A B S T R A C T

The evolution of language models marks one of the most transformative trajectories in the history of Natural Language Processing (NLP). This survey aims to provide a structured overview of key developments, tracing the progression from early statistical models to deep learning approaches, and culminating in the rise of Transformer-based architectures and Large Language Models (LLMs). We categorize and synthesize key contributions based on algorithmic paradigms, performance metrics, and systemic challenges. Specifically, we examine contributions from foundational models such as n-gram and Hidden Markov Models (HMMs), advances enabled by Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, and the paradigm shift introduced by self-attention mechanisms in Transformer architectures. Additionally, the survey discusses how LLMs have expanded the capabilities of NLP systems in tasks including text generation, translation, and dialogue modeling. Alongside these achievements, we critically highlight ongoing challenges, including model bias, interpretability, computational costs, and environmental impacts, drawing on recent literature and evaluation frameworks. Emerging trends toward improving model efficiency, fairness, and societal alignment are also explored. By mapping historical progress and identifying open questions, this article offers a comprehensive reference for researchers and practitioners interested in the evolving landscape of language models.

Keywords: *Language Modeling, Natural Language Processing, Statistical Language Models, Recurrent Neural Networks, Transformer Models, Large Language Models.*

1. Introduction

Language models have emerged as a foundational technology in the field of Natural Language Processing (NLP), driving innovations across a diverse range of applications—from machine translation and sentiment analysis to conversational agents and content generation. At their core, language models aim to capture the probabilistic structure of language, enabling machines to generate and understand human-like text. In recent decades, the development of language models has progressed from early statistical methods to powerful deep learning architectures, culminating in the rise of Large Language Models (LLMs) that now define the state-of-the-art in NLP.

Despite these advancements, the task of language modeling presents several enduring challenges. Early models such as n-grams and Hidden Markov Models

(HMMs) struggled with data sparsity and limited contextual understanding [1, 2]. Later, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [3] improved sequential learning but encountered issues with long-range dependencies and scalability. The introduction of the Transformer architecture [4] in 2017 revolutionized the field by introducing self-attention mechanisms, enabling parallel processing and more robust context handling. These developments eventually gave rise to LLMs, which possess billions of parameters and have demonstrated remarkable capabilities in generation, reasoning, and adaptation across domains. However, their increasing complexity has also brought new challenges in terms of interpretability, fairness, computational cost, and environmental impact.

This paper aims to offer a comprehensive and structured survey of the evolution of language modeling, with a focus on identifying major paradigm shifts, current challenges, and emerging directions. Unlike previous surveys that often emphasize either statistical foundations [5] or recent neural architectures [6-8] in isolation, this work provides a unified perspective that traces the progression from classical statistical models to contemporary LLMs. In doing so, it highlights key contributions from seminal models, draws comparisons between different modeling paradigms, and discusses ongoing limitations and ethical considerations.

To structure our review, we examine language modeling across three major thematic trajectories: algorithmic evolution—from statistical to neural to transformer-based models; performance evaluation—covering standard metrics such as perplexity and BLEU as well as emerging task-based and human-aligned benchmarks; and systemic challenges—including interpretability, efficiency, and societal impact. Accordingly, this survey addresses the following research questions: What historical transitions have defined the development of language models? How do current models perform across different evaluation frameworks? What critical limitations remain, and which directions offer the most promising solutions?

By examining state-of-the-art metrics for accuracy, robustness, and fairness; proposing a taxonomy of model evolution stages; and identifying open questions, this paper seeks to offer both a historical synthesis and a forward-looking roadmap. The result is a resource designed to support researchers, practitioners, and policymakers in navigating the rapidly evolving landscape of language modeling.

Therefore, the contributions of this survey are threefold: (1) It provides a historical and technical synthesis of major language modeling approaches, bridging gaps between generations of models. (2) It offers a critical analysis of the capabilities and limitations of LLMs in both academic and applied settings. (3) It outlines future directions by identifying underexplored areas and recommending research pathways based on current trends and limitations.

The remainder of the paper is organized as follows: Section 2 provides a historical overview of language modeling, beginning with statistical approaches and their transition to early neural models, including RNNs and LSTMs. Section 3 examines the rise of attention-based architectures and the transformative role of Transformer models, followed by an in-depth discussion of LLMs, their breakthroughs, limitations, and future prospects. Section 4 evaluates language models through the lens of current metrics and assessment methodologies. Section 5 critically analyzes key technical, ethical, and societal challenges in modern language modeling. Finally, Section 6 offers a forward-looking discussion on applications, research frontiers, and Section 7 presents the conclusion.

2. Background

The evolution of language modeling has been marked by a sequence of transformative developments, reflecting the interplay between theoretical advances and practical demands in NLP. Before the emergence of neural approaches, statistical models formed the backbone of computational linguistics, offering a probabilistic framework to analyze and generate human language. These early methods, while limited in expressiveness, laid the groundwork for many of the concepts and challenges that persist in modern language modeling. To understand the trajectory of progress in this field, we begin by exploring the statistical foundations and the critical insights they introduced. Despite the remarkable capabilities of contemporary LLM, several fundamental challenges remain and have not been fully resolved—such as improving data efficiency, enhancing reasoning abilities, and overcoming memory and computation constraints—many of which have roots traceable to the limitations observed in early statistical models.

2.1. Statistical Models and Their Foundations

The statistical era of language modeling, spanning from the late 1940s to the early 2010s, established the foundational principles upon which modern NLP systems are built. This period marked the shift from rule-based systems to probabilistic, data-driven approaches, capturing the probabilistic relationships between words based on large corpora of text.

A pivotal contribution came from Shannon's seminal work on information theory [9], which introduced the concept of entropy to quantify uncertainty in linguistic signals. This laid the foundation for probabilistic models of language, which estimate the likelihood of a word sequence based on its past occurrences in a corpus. These probabilistic perspectives became central to the development of more sophisticated NLP models.

Among the core models of this period were n-gram models and HMMs. N-gram models estimate the probability of a word given a fixed number of its predecessors. While effective for local context modeling, these models were constrained by a fixed context window, making them ineffective for capturing long-range dependencies. Additionally, data sparsity emerged as a significant challenge, with many word sequences either unseen or rare in the training data, leading to poor generalization and unreliable estimates [1]. From a technical standpoint, n-gram models approximate the joint probability of a sentence $w_1, w_2, \ldots, w_n$ using the Markov assumption, as shown in Equation (1):

$$P(w_1^n) = \prod_{i=1}^{n} P(wi \mid w_{i-(n-1)} \ldots w_{i-1}) . \quad (1)$$

This leads to a combinatorial explosion in the number of possible sequences when the context size increases, approximately $O(|V|^n)$ where $|V|$ is vocabulary size. To address this, smoothing techniques such as Laplace, Kneser-Ney, and backoff were introduced, though they still could not fully overcome sparsity. In contrast, **HMMs** introduced latent states s_t, with separate transition $P(s_t \mid s_{t-1})$ and emission $P(w_t \mid s_t)$ probabilities. Inference in HMMs often relied on algorithms like **Viterbi** or **Forward-Backward,** allowing structured sequence modeling, yet still limited in capturing long-range semantic structure.

HMMs addressed some of these issues by incorporating latent states, providing a more structured approach to sequential data modeling. However, like n-grams, they struggled to capture long-term dependencies and complex semantic relationships. Despite their success in early speech recognition and tagging tasks [10], statistical models were ultimately limited by their focus on surface-level patterns and short-term dependencies.

2.1.1. Limitations of Statistical Approaches

Statistical models, while foundational, faced a range of significant limitations. The most pressing challenge was data sparsity. As the number of possible word sequences grows exponentially with the length of the context window, many sequences were either unseen or rare in the training data. This led to poor generalization, particularly when dealing with rare events [11]. Another key limitation was the short-term nature of these models. Since n-grams and HMMs only considered a fixed, short window of preceding words, they failed to model long-range dependencies or capture broader contextual information. This made them less effective for tasks that required maintaining coherence or meaning over longer sequences of text [12].

Furthermore, these models were focused primarily on surface-level syntactic patterns and word co-occurrence statistics, leaving them limited in their ability to model deeper semantic relationships. This focus on local context, while useful in some applications, prevented them from achieving more complex language understanding.

2.1.2. Transition to Neural Language Modeling

The limitations of statistical models drove the search for more powerful methods, leading to the rise of neural network-based models in the early 2010s. Unlike statistical models, which relied heavily on manually crafted features, neural networks offered a way to learn distributed, continuous representations of words and their relationships. This transition was pivotal in enabling the modeling of long-range dependencies and complex semantic patterns in language. The introduction of RNNs marked the beginning of this shift, providing a way to process sequences of arbitrary length by maintaining a hidden state that theoretically captured information from all previous time steps. However, RNNs still faced challenges in training on long sequences, primarily due to issues with vanishing and exploding gradients during backpropagation [13]. To better illustrate the key differences and limitations between statistical and early neural models, Table 1 provides a comparative overview of their main characteristics, highlighting the technical and representational shifts that marked the transition toward neural language modeling.

Table 1. Comparative summary of key characteristics and limitations across classical statistical models and early neural language models.

Model Type	Context Window	Handles Data Sparsity	Learns Semantic Representations	Captures Long-Term Dependencies	Memory Usage (MB)	Context Size	Trainable Params	Key Limitations
N-gram	Fixed (e.g., n =3)	Poor	No	No	Low	N-1 tokens	—	Data sparsity, limited context
HMM	Implicit via hidden states	Moderate	No	No	Low-Moderate	implicit	—	Weak semantic modeling, Markov assumption
RNN	Theoretically unbounded	Better	Partial	No (vanishing gradients)	High	theoretically ∞	1M–10M	Hard to train on long sequences
LSTM	Memory cells with gating	Better	Yes	Yes	Very High	∞	10M–100M	Sequential bottleneck, computational cost

2.2. *Recurrent Neural Networks: Principles and Challenges*

RNNs, which emerged in the 1980s but gained prominence in the 2010s, introduced a dynamic memory mechanism through recurrent connections. This allowed RNNs to process sequences and capture temporal dependencies, a significant advancement over previous models like n-grams and HMMs. RNNs are designed to process sequences by maintaining a hidden state that theoretically captures information from all preceding time steps, enabling the model to remember past events and use that information to predict future ones. From a formal perspective, the hidden state h_t in a simple RNN is computed as shown in Equation (2):

$$h_t = tanh\,(W_{xh}x_t + W_{hh}h_{t-1} + b_h)\,, \quad (2)$$

where x_t is the input at time t, W_{xh} and W_{hh} are weight matrices, b_h is the bias, and $tanh$ is the activation function. The output y_t is typically calculated by:

$$y_t = W_{hy}h_t + b_y\,. \quad (3)$$

These equations reflect the recurrent nature of the model, where the hidden state evolves over time based on both the current input and the previous hidden state. However, due to repeated multiplication through nonlinearities such as $tanh$, gradients computed via backpropagation through time tend to either vanish or explode for long sequences, leading to unstable training [13]. These challenges made it difficult to train RNNs effectively on long sequences, limiting their applicability in real-world language modeling tasks. In particular, gradients used for training would either shrink to zero or grow exponentially, hindering the learning process and making it nearly impossible to maintain useful information over long sequences. A simple diagram of a standard RNN cell is shown in Figure 1 to illustrate the basic mechanism of sequential information flow and hidden state propagation.

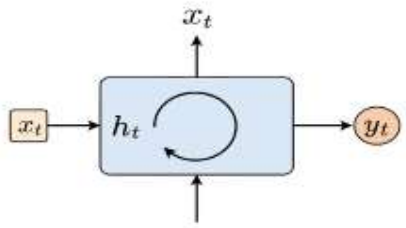

Figure 1. Architecture of a standard RNN cell with input, hidden state, and output.

2.2.1. *Advances Introduced by LSTMs*

To address the shortcomings of RNNs, LSTM networks were introduced [3]. LSTMs incorporated memory cells and gating mechanisms—input, output, and forget gates—that allowed the network to selectively retain and forget information over time. This innovation significantly improved the ability to capture long-term dependencies and sequence-level patterns in data, overcoming the vanishing gradient problem by allowing the flow of gradients across many time steps. Mathematically, an LSTM cell augments the standard RNN formulation by introducing gating mechanisms that regulate information flow. A schematic diagram of the internal structure of an LSTM cell is provided in Figure 2, illustrating how the gating mechanisms control information flow.

LSTMs [3] quickly became the state-of-the-art model in many NLP tasks, including machine translation [14], language modeling, and speech recognition [15]. Their

ability to store and update information across many time steps made them a powerful tool for handling complex sequential data, overcoming the limitations of earlier RNNs. The success of LSTMs can be seen in many applications, where they consistently outperformed traditional RNNs, especially in long-sequence modeling tasks [16].

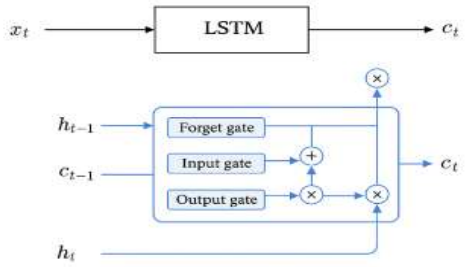

Figure 2. Architecture of an LSTM cell showing input, forget, and output gates along with the memory cell and hidden state transitions.

2.2.2. *Other Limitations of Early Neural Models*

Despite the advancements brought by LSTMs, early neural models still had limitations. One of the main challenges was their computational cost. Training RNNs and LSTMs required large datasets and substantial computational resources, often making them impractical for certain applications.

Another limitation was the sequential nature of RNNs and LSTMs. The sequential computation inherent in RNNs and LSTMs precludes parallelization during training, leading to inefficiencies on modern hardware like GPUs. These models processed sequences in a single direction, which made them less efficient in capturing global context across the entire sequence. The time complexity for processing a sequence of length T is $O(T)$, as each step depends on the output of the previous one. This unidirectional processing was a bottleneck, particularly for very long sequences or large datasets and tasks requiring deep contextual understanding or the ability to attend to any part of a sequence at any time. This limitation became particularly evident in tasks such as machine translation, where it was critical to understand the context of an entire sentence rather than just the immediate preceding words.

These shortcomings and challenges highlighted the need for architectures that could decouple positional dependence and allow parallel sequence processing—ultimately motivating the development of attention-based models such as Transformers, which would later revolutionize language modeling by overcoming many of the limitations faced by RNNs and LSTMs.

3. From Attention to Scale: The Rise of Transformers and LLMs

3.1. *Transformer Models and Attention-based Architectures*

The introduction of the Transformer model in 2017 by Vaswani et al. represented a transformative shift in the field of NLP [4]. Prior to this, models such as RNNs and LSTM networks processed data sequentially, meaning each token or word was processed one after another. In contrast, the Transformer leveraged a self-attention mechanism that facilitated parallel processing of entire input sequences. This innovation proved foundational in developing more efficient, scalable models capable of handling large datasets, ultimately becoming the core architecture for contemporary leading NLP models such as Bidirectional Encoder Representations from Transformers (BERT) [15], Generative Pre-trained Transformer (GPT) [17], and Text-to-Text Transfer Transformer (T5) [18].

3.1.1. *The Transformer Architecture*

The Transformer architecture represents a pivotal advancement in sequence modeling, built upon two principal components: the encoder and the decoder. The encoder processes the input sequence in parallel and generates a set of contextualized, attention-weighted representations. A schematic overview is shown in Figure 3, which depicts the encoder–decoder pipeline along with the flow of self-attention weights across tokens.

The decoder then consumes these encoder outputs to generate the target sequence step by step. What distinguishes the Transformer from earlier neural architectures is its self-attention mechanism, which dynamically computes the relevance of each token with respect to every other token in the sequence, regardless of their distance or positional order.

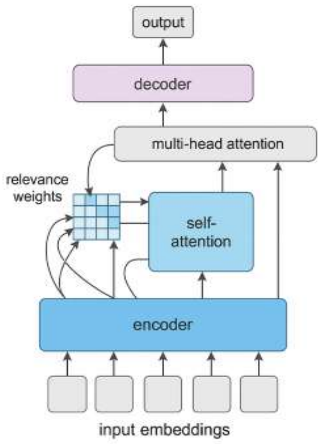

Figure 3. Diagram of the Transformer architecture showing the encoder-decoder structure and the flow of self-attention weights across input tokens.

This mechanism enables the Transformer to model long-range dependencies effectively—a challenge that recurrent architectures such as RNNs and LSTMs struggled with because of their inherently sequential nature.

Unlike those models, which process tokens one at a time, the Transformer processes the entire sequence simultaneously, enabling extensive parallelization during both training and inference [4].

Parallelization not only accelerates training but also makes the architecture highly scalable, supporting datasets of massive size and models containing billions of parameters [19].

Another essential component is multi-head attention, where several attention layers operate in parallel, each capturing different types of relationships among tokens. This design broadens the model's ability to encode subtle semantic and syntactic dependencies, thus enhancing its performance across diverse language tasks [4].

From an algorithmic perspective, the self-attention mechanism computes a weighted representation of each token by comparing it to all other tokens in the sequence. Given an input sequence of n tokens, each represented as a vector of dimension d, self-attention is computed using three learnable matrices: queries Q, keys K, and values V. These are derived as:

$$Q = XW^Q,\ K = XW^K,\ V = XW^V \quad (4)$$

where $X \in R^{n\times d}$ is the input matrix, and W^Q, W^K, W^V $R^{d\times d_k}$ are the projection matrices. The attention scores are computed as:

$$Attention\ (Q, K, V) = softmax\ (\frac{QK^T}{\sqrt{d_k}})V\ , \quad (5)$$

where the dot product QK^T measures pairwise similarity between tokens, scaled by $\sqrt{d_k}$ to control variance. This formulation enables the model to dynamically attend to relevant tokens regardless of position. The time complexity of computing self-attention is $O(n^2.d)$ which stems from the pairwise comparison of all token pairs—making it computationally expensive for long sequences, though highly parallelizable on modern hardware.

To enrich the representational power, the Transformer uses multi-head attention, which runs h parallel self-attention mechanisms, each with separate parameter matrices, and concatenates their outputs. This allows the model to capture diverse types of relationships simultaneously and is formally expressed as:

$$MultiHead\ (Q, K, V) = Concat\ (head_1, \dots, head_h)\ W^O,$$
$$head_i = Attention\ (QW_i^Q, KW_i^K, VW_i^V)\ . \quad (6)$$

This structure enables the model to focus on multiple semantic or syntactic aspects of the sequence concurrently, providing richer contextual embeddings and significantly improving performance across a wide range of tasks.

3.1.2. *Key Innovations in Transformer Models*

The Transformer architecture introduced several innovations that have reshaped modern approaches to NLP. Foremost among these is the self-attention mechanism, which enables the model to dynamically evaluate the importance of each token in relation to others across the entire sequence. This mechanism solves a critical issue present in earlier models like RNNs, which had difficulty learning long-range dependencies [4]. By processing tokens in parallel, the Transformer significantly accelerates both training and inference times, making it particularly suited for environments where high computational power and large-scale datasets are required [15].

Another transformative feature of the Transformer is its scalability. Models such as GPT-3, which boasts over 175 billion parameters, are built upon the Transformer architecture and have set new benchmarks in language modeling [19]. These models leverage the inherent scalability of the Transformer, allowing them to achieve state-of-the-art performance across a variety of NLP tasks. Additionally, the pre-training and fine-tuning paradigm, another key innovation, has been instrumental in the success of Transformer models. During pre-training, models learn general language representations from vast

amounts of unsupervised text data. These models are then fine-tuned on specific downstream tasks, making them highly adaptable and capable of achieving exceptional results in applications ranging from machine translation to sentiment analysis and more [15].

3.1.3. Transformer Variants and Applications

The widespread success of the Transformer architecture has inspired the development of numerous specialized variants, each designed to optimize performance for particular NLP tasks or architectural limitations. A key structural aspect of these models is their encoder-decoder configuration. While the original Transformer employs both encoder and decoder stacks, later models adopt simplified or asymmetric designs. For instance, BERT exclusively uses the encoder stack to learn deep bidirectional representations of text, allowing it to consider both left and right contexts simultaneously [15]. This design is particularly effective in discriminative tasks such as question answering, sentiment classification, and named entity recognition.

In contrast, GPT utilizes only the decoder stack with a unidirectional (left-to-right) attention mechanism, optimized for autoregressive language modeling. This structure allows GPT to excel in generative applications such as open-ended dialogue systems and story generation, where contextual coherence over long sequences is critical [17-19].

A third major variant is T5, which adopts a full encoder-decoder architecture but introduces a unified text-to-text framework. T5 reframes every NLP task—classification, summarization, translation—as a sequence-to-sequence problem, thereby enabling task-agnostic learning [18]. This flexibility makes T5 exceptionally adaptable across domains and benchmarks.

These architectural differences reflect distinct underlying objectives: while BERT and T5 are primarily optimized for understanding tasks—such as classification, inference, and information extraction—GPT is fundamentally designed for generation, excelling in tasks that require fluent and coherent text continuation. This distinction is rooted in their directional attention mechanisms (bidirectional vs. unidirectional) and pretraining paradigms (masked language modeling vs. autoregressive decoding), which yield different capabilities and ideal applications [15, 17, 18].

Despite their differences, these models demonstrate the adaptability of the Transformer architecture across both interpretive and generative tasks. Comparative studies (such as those in [17] and [20]) have shown that while BERT often outperforms in token-level understanding, GPT and T5 are superior in tasks requiring text continuation or abstraction. However, a common limitation among these standard architectures is their quadratic complexity in self-attention, which becomes prohibitive for long inputs—a limitation addressed by recent optimized Transformer variants.

3.1.4. Challenges and Limitations

Despite their remarkable success, Transformer models face several key challenges. One major limitation lies in their **computational cost**, especially when training large-scale architectures. These models demand significant memory and processing resources, making them inaccessible for many organizations lacking high-performance computing infrastructure [21].

Moreover, although the self-attention mechanism efficiently models long-range dependencies, it incurs **quadratic memory and time complexity** with respect to sequence length. This becomes a bottleneck when handling long-form texts such as books or legal documents [22].

To address this, several efficient Transformer variants have been developed. **Longformer** employs sparse attention, combining local windows with selected global tokens, thus reducing complexity to linear scale while maintaining contextual coverage [23]. **Reformer** introduces Locality-Sensitive Hashing (LSH) to group similar tokens into buckets, restricting attention computations within these subsets and enabling sub-quadratic performance [24]. **Performer** approximates softmax attention using **random feature mappings**, projecting queries and keys into lower-dimensional spaces and leveraging kernel-based tricks to achieve linear complexity [25]. **Linformer** assumes that the attention matrix is low-rank and compressible, using linear projections of key and value matrices to reduce the attention computation from $O(n^2)$ to $O(n)$ [26]. These innovations significantly improve scalability, particularly for long-input or real-time applications.

Another persistent issue is **algorithmic bias**. Like other deep learning models, Transformers can absorb and reproduce societal stereotypes present in their training data, leading to **ethical concerns** in high-stakes contexts such as

hiring, criminal justice, and healthcare [27]. In parallel, the **lack of interpretability** in Transformer models remains problematic. Their internal decision-making processes are often opaque, complicating trust, verification, and debugging—especially in applications requiring transparency, such as legal or clinical decision support [28].

While these challenges are non-trivial, the Transformer architecture remains a transformative force within NLP. Its introduction of self-attention mechanisms, its parallel processing capabilities, and the pre-training and fine-tuning paradigm have set new standards for both language understanding and generation. The ongoing development of more efficient training methods, the exploration of ways to mitigate bias, and the pursuit of greater model interpretability all point to the continued evolution of Transformer-based models. Their scalability, flexibility, and state-of-the-art performance will likely continue to shape the future of NLP, addressing both current limitations and new challenges as the field progresses [4, 15, 18].

3.2. The Emergence of Large Language Models

The advent of LLMs has marked a transformative era in NLP, profoundly altering the capabilities and expectations of AI systems. These models, often comprising billions to trillions of parameters, are generally built upon the Transformer architecture and trained on massive corpora of unstructured text using self-supervised learning. Their unprecedented scale and architectural sophistication have enabled LLMs to perform a wide array of complex language tasks with remarkable fluency and generalization [19, 29].

3.2.1. Characteristics of LLMs

LLMs are primarily distinguished by their vast parameter counts and extensive training data, which endow them with the ability to capture deep semantic, syntactic, and pragmatic features of natural language. The scaling hypothesis—supported by empirical findings—suggests that as model size, dataset size, and compute increase, performance continues to improve in a predictable manner [30, 31]. This has motivated the development of ultra-large models such as GPT-3 (175B parameters) [19], PaLM (540B) [29], Claude (Anthropic) [32], and GPT-4 [33], which consistently push the boundaries of general-purpose NLP.

These models differ not only in scale but also in architecture, training objectives, and performance. For example, GPT-3 and GPT-4 (OpenAI) are autoregressive models optimized for few-shot generalization, while PaLM (Google) integrates pathway-based sparsity to optimize computational efficiency across tasks [29]. Claude (Anthropic) is specifically designed with a focus on constitutional AI principles, aiming for safer and more interpretable alignment [32]. Benchmark comparisons—such as MMLU, Big-Bench, and HELM—suggest that GPT-4 outperforms its predecessors and most contemporaries in a wide range of reasoning and multilingual tasks, though Claude demonstrates strengths in safety and instruction-following [29, 32, 33].

In contrast to proprietary models such as PaLM, Claude, and Gemini, which demonstrate remarkable capabilities but remain closed-source, a parallel movement has emerged around open-weight alternatives that prioritize transparency, accessibility, and reproducibility. A central pillar of this trend is Large Language Model Meta AI (LLaMA), developed by Meta, which catalyzed the open-weight ecosystem by offering high-performing models with significantly fewer parameters (7B–65B) than GPT-3 or PaLM, yet demonstrating competitive results due to efficient training and high-quality data curation [34]. LLaMA employs a decoder-only architecture with standard multi-head self-attention and autoregressive training; its successors, such as LLaMA 2, further enhance instruction-following and alignment capabilities, making them suitable backbones for instruction-tuned derivatives like Alpaca [35] and Vicuna [36].

Complementing LLaMA's foundational role, newer open-weight models such as Mistral and Falcon have introduced architectural innovations tailored for efficiency and scalability. Mistral, developed by Mistral AI, employs grouped-query attention (GQA) and sliding window attention to accelerate inference and reduce memory overhead, particularly in multilingual and resource-constrained settings [37]. Falcon, released by the Technology Innovation Institute, emphasizes deployment efficiency through multi-query attention (MQA) and streamlined training pipelines designed for commodity hardware [38]. Together, these models represent a significant shift in large-scale language modeling—from closed, monolithic architectures toward modular, open, and reproducible systems that empower academic and industrial research alike. Table 2 summarizes the architectural and

training design choices of several prominent Transformer-based LLMs, highlighting their respective strengths and limitations across generative and understanding tasks.

One of the most impactful methodological innovations associated with LLMs is the pre-training and fine-tuning paradigm. During pre-training, models learn broad language representations from large-scale, unlabeled corpora via masked or autoregressive objectives [15, 17]. These representations are then adapted to specific downstream tasks through supervised fine-tuning. In some cases, task-specific fine-tuning is bypassed entirely using prompt-based learning, where models perform zero-shot or few-shot generalization simply by conditioning on task descriptions and examples [19].

These training strategies differ in terms of data requirements, generalization capabilities, and interpretability. Traditional fine-tuning enables precise control over downstream performance but requires large labeled datasets and retraining for each task. In contrast, prompt-based learning—especially in its in-context variant—relies on model internalization of task structures during pre-training, offering higher adaptability with minimal task-specific data. Studies have shown that GPT-3, when prompted effectively, can rival fine-tuned models in tasks like summarization and QA, while fine-tuned versions of PaLM and T5 maintain advantages in domain-specific tasks with structured outputs [19, 39].

Moreover, LLMs exhibit a strong ability to handle extended contexts and generate coherent, context-sensitive outputs. Their performance in summarization, dialogue systems, sentiment analysis, and textual entailment showcases their generalization power. Importantly, few-shot and zero-shot capabilities reduce reliance on costly annotated datasets, making these models flexible tools for low-resource or rapidly evolving domains [40].

Table 2. Comparison of prominent Transformer-based language models in terms of architecture, attention mechanisms, training objectives, and practical considerations.

Model	Architecture	Attention Type	Objective	Strengths	Limitations	Reference
BERT	Encoder-only	Bidirectional	Masked LM (MLM)	Understanding, classification	Not suited for generation	Devlin et al., 2019
GPT	Decoder-only	Unidirectional	Autoregressive LM	Generation, coherence	Weak bidirectional understanding	Radford et al., 2018
T5	Encoder-Decoder	Bidirectional + causal	Text-to-Text (Seq2Seq)	Multi-tasking, unified format	High compute cost	Raffel et al., 2020
PaLM	Decoder-only	Multi-head Attention	Autoregressive LM	Scaling to >500B, strong zero-shot	Expensive to train	Chowdhery et al., 2022
Claude	Decoder-only	RLHF-Aligned Attention	Constitutional LM	Alignment, safety, helpfulness	Closed weights	Anthropic, 2023
LLaMA	Decoder-only	Multi-head Attention	Autoregressive LM	Open weights, efficient scaling	Less instruction tuning (early versions)	Touvron et al., 2023
Mistral	Decoder-only	GQA + Sliding Window	Autoregressive LM	Fast inference, efficient scaling	Limited context window	Mistral AI, 2023
Falcon	Decoder-only	Multi-query Attention	Autoregressive LM	Optimized for deployment	Smaller training data	Almazrouei et al., 2023

3.2.2. Breakthroughs Enabled by LLMs

LLMs have catalyzed breakthroughs across both foundational and applied areas of NLP. In natural language generation, they produce text that is coherent, contextually appropriate, and often indistinguishable from human writing. GPT-3, for example, demonstrated proficiency in tasks ranging from story generation to programming code synthesis [19, 41].

These capabilities are corroborated by strong benchmark performance: for instance, GPT-3 achieved 76% accuracy on the LAMBADA task [19], PaLM 540B reached state-of-the-art results on BIG-bench Hard [29], and GPT-4 surpassed previous models on MMLU (86.4%) and HumanEval benchmarks [33], outperforming Claude and ChatGPT variants in most reasoning tasks [32].

In machine translation, models such as mBART and mT5 leverage multilingual pre-training to provide high-quality translations across a wide range of languages, often without requiring explicit alignment or task-specific adaptation [42, 43]. Similarly, LLMs have revolutionized question answering and information retrieval, enabling systems that can reason over large knowledge bases and provide detailed, human-like responses.

Applications in conversational AI have expanded significantly with the advent of models like ChatGPT, based on GPT-3.5 and GPT-4, which enable managing open-ended dialogues, multi-turn reasoning, and personalized user interactions [33]. Beyond traditional NLP tasks, LLMs have been integrated into specialized domains including code generation with Codex [41], scientific knowledge modeling through Galactica [44], and multimodal reasoning using models like Flamingo [45] and GPT-4V [33]. While these models share architectural roots, they differ significantly in specialization and design goals: Claude emphasizes safety through Constitutional AI [32], Galactica is tailored for scientific domains but faces reliability concerns [44], while multimodal models like Flamingo and GPT-4V are optimized for vision-language tasks. These models demonstrate the expanding versatility of LLMs in both textual and multimodal contexts.

Despite their impressive generalization capabilities, LLMs still exhibit notable limitations in complex reasoning and factual consistency. Tasks requiring multi-step mathematical deduction or symbolic manipulation—such as those evaluated in GSM8K [46], MATH [47], and BIG-Bench Hard—frequently expose models' tendencies to produce confident yet incorrect or hallucinated outputs. For instance, GPT-3 and GPT-4 often struggle with maintaining logical consistency over extended chains of reasoning, especially in problems involving arithmetic abstraction or formal logic. These shortcomings raise concerns about the deployment of LLMs in high-stakes domains such as scientific discovery, legal analysis, or education, where factual precision and deductive validity are critical.

3.2.3. *Challenges and Ethical Considerations*

Despite their advantages, LLMs present several technical, ethical, and environmental challenges. One major concern is the immense computational cost associated with training and deploying these models. The energy required to train a single LLM may emit hundreds of tons of CO_2, raising sustainability concerns [21, 47, 48].

Bias and fairness are also central issues. Since LLMs are trained on internet-scale data, they inevitably inherit and may amplify societal biases present in those datasets. This can result in outputs that reinforce harmful stereotypes or produce offensive content [49]. Mitigation techniques—such as bias-aware training, dataset filtering, and fairness auditing—are active areas of research but remain imperfect.

Interpretability and trustworthiness are additional challenges. LLMs operate as black boxes, making it difficult to trace the reasoning behind specific predictions. This lack of transparency is particularly problematic in high-stakes domains like healthcare, law, and finance, where explainability is essential for accountability and user trust [50].

3.2.4. *The Future of LLMs*

Looking ahead, research in LLMs is increasingly focused on achieving efficiency, controllability, and safety. New approaches such as parameter-efficient fine-tuning (e.g., Low-Rank Adaptation (LoRA) [51]), retrieval-augmented generation (e.g., Retrieval-Enhanced Transformer (RETRO) [52]), and Retrieval-Augmented Generation (RAG)-based models that combine pre-trained generators with external document retrieval systems to improve factual grounding and reduce hallucination [53], and sparsely activated models (e.g., Switch Transformers [54]) aim to reduce resource demands while maintaining performance [18]. At the same time, multimodal LLMs integrating text with images, audio, or structured data are broadening the range of supported tasks—from visual question answering to cross-modal reasoning [33, 45]. Alignment techniques such as instruction tuning and Reinforcement Learning from Human Feedback (RLHF) [55] are further enhancing model controllability and human preference alignment. As LLMs become more embedded in scientific, industrial, and social infrastructures, ensuring their development is equitable, transparent, and sustainable becomes imperative.

Yet, progress depends on tackling more subtle and emerging challenges. One such issue is verbatim memorization, which poses risks to originality and copyright. Empirical work has shown that LLMs can reproduce training set excerpts—such as code snippets or email addresses—especially under specific prompts. For instance, Carlini et al. (2023) demonstrated that GPT-3 could regenerate memorized content even without clear cues [56], raising both legal and creative concerns.

Another challenge lies in stylistic dissonance during co-authorship. While LLMs produce coherent texts autonomously, their outputs can clash in tone or vocabulary when mixed with human writing. To address this, researchers are exploring personalization techniques like LoRA [51] and Quantized Low-Rank Adapter (QLoRA) [57], which allow low-resource adaptation, and StyleVector

[58], which aligns generation with a user's stylistic signature via real-time latent feature extraction. These methods enable more adaptive, consistent, and personalized interactions over time.

Beyond writing, LLMs hold promise for education, especially in personalized tutoring. Their interactive and adaptive capabilities allow tailoring of content complexity, pace, and depth to learner needs, supporting inclusive instruction at scale [59, 60]. Such applications are particularly valuable where human tutoring is limited.

Ultimately, advancing LLMs requires more grounded, controllable, and user-aligned generation. Techniques such as in-context learning, instruction tuning [61, 62], and RLHF [55], combined with continual learning and user-in-the-loop feedback, are pushing LLMs toward being more responsive, reliable, and stylistically cohesive in dynamic use cases.

4. Evaluation of Models and Metrics

Evaluating language models is essential not only for benchmarking performance but also for identifying limitations, informing development, and ensuring safe deployment in real-world scenarios. As language models have evolved from statistical n-gram systems to massive pre-trained transformers, evaluation methodologies have struggled to keep up with their diverse capabilities and deployment contexts. In this section, we survey classical metrics, discuss modern evaluation paradigms suited for LLMs, and critically examine their quantitative performance, semantic fidelity, and ethical adequacy.

4.1. Classical Metrics and Their Limitations

Traditional metrics such as perplexity have long been used to evaluate language models based on next-token prediction. While it remains a fast and interpretable proxy for internal uncertainty [63], perplexity lacks semantic grounding and is not comparable across models with different tokenization schemes. For example, a model with lower perplexity might still produce incoherent or irrelevant outputs when evaluated qualitatively.

Other surface-based metrics such as BLEU [64] and ROUGE [65] measure lexical overlap between generated and reference texts. Although these scores have become standard for tasks like machine translation and summarization, they are brittle against paraphrasing and perform poorly in open-ended tasks like dialogue or storytelling. Studies have shown that BLEU may report high scores for grammatically awkward translations, while receiving low ratings in human evaluations [66]. For instance, in the WebNLG dataset, BLEU achieved a Pearson correlation of only 0.62 with human judgments, compared to 0.88 for BERTScore, demonstrating the limitations of n-gram overlap in capturing semantic adequacy [67].

4.2. Semantics-Aware Evaluation

To overcome surface-level limitations, embedding-based metrics such as BERTScore [67] and Sentence-BERT [68] have been proposed. These methods embed texts into contextual semantic spaces, measuring cosine similarity rather than n-gram overlap. In comparative benchmarks such as the WebNLG and DUC datasets, BERTScore demonstrated higher correlation with human judgments than BLEU and ROUGE [67]. In summarization tasks using the TAC2008 dataset, BERTScore achieved a Spearman correlation of 0.785 with human assessments, substantially higher than BLEU (0.617) and ROUGE-L (0.592), further validating the semantic alignment advantage of embedding-based methods [67].

However, these metrics are not without flaws. Embedding models themselves may encode societal biases and task-specific limitations. For example, in creative writing or metaphor-heavy content, cosine similarity may misjudge outputs that are semantically appropriate but lexically novel. Moreover, differences in sentence structure or idiomatic usage can yield misleading scores, even when human evaluators consider outputs valid.

4.3. Human-in-the-Loop and Qualitative Evaluation

Human evaluation remains indispensable for subjective criteria such as coherence, helpfulness, and creativity. A variety of human-centered protocols have emerged, including:

Likert scales: annotators rate individual outputs along dimensions (e.g., 1–5 for fluency).

Elo-style ranking: pairs of outputs are compared repeatedly, enabling skill-based ranking [69].

Best-Worst Scaling (BWS): annotators identify best and worst examples from a set, providing more stable signal [70].

Each method has trade-offs. While Likert scales are intuitive, they suffer from inter-annotator variance. BWS

improves annotation consistency but is harder to scale. In recent evaluations of models like GPT-4 and Claude, Elo-style and BWS yielded more stable preference signals than point-based ratings [33]. For example, GPT-4 received a human preference score of 85.5% over GPT-3.5's 72.3% in internal pairwise comparisons using Elo-style evaluation [33], highlighting the sensitivity of human judgment protocols to quality differences.

A critical limitation of human evaluation is cost and scalability. To address this, hybrid frameworks such as OpenAI's Evals protocol combine lightweight human feedback with automated heuristics, achieving partial automation while preserving judgment quality.

4.4. Fairness, Bias, and Ethical Evaluation

Large models often replicate harmful stereotypes present in their training data. This necessitates targeted fairness audits. Tools such as Winogender schemas [71], StereoSet [72], and CrowS-Pairs [73] quantify bias along axes of gender, race, and socioeconomic status. Comparative experiments show that models like GPT-2 and GPT-3 exhibit varying levels of stereotypical bias, with newer instruction-tuned models showing partial improvement [72]. On the StereoSet benchmark, GPT-2 showed a stereotype score of 59.4 (higher indicates more bias), while InstructGPT reduced this to 52.1, reflecting progress through alignment techniques [72].

Ethical evaluation has also begun incorporating counterfactual testing, where inputs are perturbed minimally (e.g., changing "John" to "Mary") to probe causal sensitivity. Metrics such as demographic parity, equalized odds, and equal opportunity are borrowed from fairness in classification tasks to quantify disparities in outputs. Despite growing attention, operationalizing fairness remains challenging due to group definition ambiguity, data imbalance, and cultural context variation. A comprehensive audit must triangulate multiple metrics and human judgment to avoid false confidence.

4.5. Real-World and Deployment-Oriented Evaluation

Deployment introduces new priorities: robustness, latency, task success, and user satisfaction. Models must maintain reliability under adversarial inputs, language variation, or domain shift. Evaluation platforms like HELM [70] and RAGAs [74] provide standardized multi-dimensional testing across tasks, languages, and stress conditions.

For example, HELM reports performance on factual QA under adversarial and multilingual stress tests, highlighting failure modes where token accuracy is high but factual reliability is low. In HELM's factual QA under multilingual adversarial stress, GPT-3.5 achieved 92.1% token accuracy but only 67.4% factual consistency, exposing vulnerabilities in model reliability under real-world perturbations [70]. RAGAs, in contrast, tests prompt inversion and hallucination resistance, helping benchmark generation stability.

In customer service deployments, KPIs such as time-to-resolution, sentiment delta, and escalation rate become more relevant than traditional NLP metrics. These application-aware indicators are critical for assessing downstream utility but lack cross-domain comparability.

As shown in Table 3, each evaluation method offers a unique lens into model behavior. No single metric suffices across all contexts; instead, practitioners must triangulate quantitative scores with qualitative assessments and stress testing. The lack of alignment between automatic metrics and human perception remains a persistent challenge, especially in open-ended generation, making hybrid approaches and ongoing benchmarks critical for robust LLM evaluation.

Table 3. Overview of major evaluation metrics used in LLM assessment, categorized by evaluation type, with representative strengths, limitations, and typical use cases.

Metric	Focus Area	Strengths	Limitations
Perplexity	Token-level prediction	Simple, efficient	Ignores semantics, tokenization-dependent
BLEU	Machine translation	Easy to compute, reproducible	Penalizes paraphrasing, weak correlation with human scores
ROUGE	Summarization (recall)	Captures content overlap	Fails on fluency and coherence
BERTScore	Semantic similarity (token)	Embedding-based, semantically aware	Sensitive to embedding biases
Sentence-BERT	Sentence-level semantics	Robust to surface variation	May misjudge creative or idiomatic language
Likert / BWS / Elo	Human preference	Holistic, task-specific	Costly, annotator bias, low scalability
Winogender / StereoSet	Bias detection	Socially critical, interpretable	Dataset-sensitive, narrow scope
HELM / RAGAs	Robustness, deployment	End-to-end, multi-dimensional	Context-specific, evolving standards

Table 4. Key benchmark datasets and leaderboards in language model evaluation.

Benchmark/ Dataset	Focus Area	Scale	Key Contribution
SQuAD [75]	MRC (Machine Reading Comprehension)	100k+ Q–A pairs	Established large-scale extractive QA; benchmark for early deep QA models
TriviaQA [76]	MRC / QA	650k Q–A pairs	Broadened QA sources (Wikipedia + web); tested generalization beyond SQuAD
Natural Questions [77]	Open-domain QA	300k real Google queries	Introduced naturally occurring search questions; realistic contexts
MS MARCO [78]	Passage retrieval + QA	1M queries	Connected QA with large-scale IR; real search-based evaluation
HotpotQA [79]	Multi-hop QA	113k Q–A pairs	Required reasoning across multiple documents; step beyond single-span extraction
TyDi QA [80]	Multilingual QA	200k Q–A pairs, 11 languages	Benchmarked cross-lingual generalization; diverse morphology
GLUE [81]	NLU (multi-task)	9 tasks, ~100k samples	Unified NLU evaluation across tasks; became a standard leaderboard
SuperGLUE [82]	Advanced NLU / reasoning	8 harder tasks	Raised evaluation difficulty; extended beyond GLUE
MMLU [83]	Knowledge + reasoning	57 domains	Tested general knowledge and professional reasoning at scale
BIG-bench [84]	Broad capability eval	204 tasks	Evaluated emergent properties of LLMs across adversarial/creative tasks
HELM [70]	Holistic evaluation	Multi-dimensional	Introduced multi-criteria evaluation (robustness, fairness, deployment readiness)
LMSYS Chatbot Arena [85]	Interactive eval	Continuous, crowd-sourced	First large-scale human preference leaderboard; pairwise model comparison

4.6. Benchmark Datasets and Leaderboards

Benchmark datasets and public leaderboards have played a pivotal role in shaping the development and evaluation of language models. They provide standardized tasks, enable reproducibility, and foster competitive improvements across the research community.

A large portion of early QA benchmarks were based on Machine Reading Comprehension (MRC), where models are required to extract answers directly from a passage. This paradigm began with SQuAD [75], which established extractive QA as a mainstream evaluation task. Follow-up datasets such as TriviaQA [76] and Natural Questions (NQ) [77] extended the challenge by incorporating longer contexts, diverse domains, and naturally occurring user queries. Similarly, MS MARCO [78] introduced real-world search queries paired with passage retrieval, bridging QA with information retrieval. For multi-hop reasoning, HotpotQA [79] required models to integrate evidence across multiple documents, moving beyond single-span extraction. Multilingual and cross-lingual evaluation was advanced by TyDi QA [80] and XQuAD [81], which tested generalization across languages with diverse morphology and syntax.

As models evolved, evaluation also moved beyond MRC. Benchmarks such as GLUE [86] unified Natural Language Understanding (NLU) tasks across sentiment, entailment, and similarity, while SuperGLUE [82] introduced harder reasoning-oriented challenges. MMLU [83] further expanded the scope by testing factual knowledge and reasoning across 57 academic and professional domains. BIG-bench [84] pushed the boundary with over 200 tasks, many adversarial or creative, highlighting emergent behaviors of LLMs. More recently, holistic benchmarks such as HELM [74] and interactive leaderboards like LMSYS Chatbot Arena [85] have emphasized multi-dimensional criteria, including robustness, fairness, and user preference under deployment scenarios.

Benchmark datasets and leaderboards thus act both as drivers (incentivizing new architectures tuned for competitive performance) and mirrors (revealing persistent weaknesses in reasoning, robustness, and fairness). Importantly, results on these datasets must be interpreted cautiously. For instance, models exceed 95% F1 on SQuAD yet remain brittle against adversarial phrasing and domain shift. Likewise, GLUE was saturated within a year of release, underscoring the need for evolving, realistic benchmarks.

Table 4 summarizes key datasets and leaderboards, highlighting their task focus, scale, and lasting contributions.

5. Challenges and Limitations

Despite the transformative advancements introduced by LLMs, their development and deployment present substantial challenges across various dimensions: computational efficiency, computational cost, interpretability, fairness, and ethical considerations. These challenges are crucial for ensuring responsible and sustainable progress in AI technologies.

One of the primary concerns is the computational cost of training and deploying LLMs. Models like GPT-4 and PaLM 2 require hundreds of billions of parameters and vast training datasets, which demand exascale-level computation [33, 87]. The training process often takes weeks using distributed systems of GPUs or TPUs, leading to significant financial and environmental costs. This creates a disparity, as only large tech corporations and well-funded research labs can afford the necessary infrastructure, contributing to an imbalance in AI research. To mitigate these issues, researchers have been exploring techniques such as sparse modeling, LoRA [51], and quantization-aware training. In parallel, retrieval-augmented approaches such as RAG [53] have been explored to offload knowledge representation onto external memory, reducing the need for parametric memorization and enabling smaller models to perform complex tasks efficiently. However, these solutions are still evolving and their scalability remains a topic of active research.

Beyond computational constraints, training efficiency is further complicated by the reliance on massive, often noisy, and inconsistently curated datasets. The sheer scale of data required for pretraining—often exceeding trillions of tokens—raises issues regarding data quality, representational balance, and potential contamination of evaluation benchmarks [88]. Moreover, the opaque nature of data collection pipelines means that harmful or private content can inadvertently be included in training corpora. While efforts like dataset filtering and controlled corpus construction have been proposed, there remains no standardized framework for ensuring data fidelity in LLM pretraining.

Another significant challenge lies in the interpretability of LLMs. These models operate as high-dimensional, non-linear function approximators, which makes it difficult to understand their decision-making processes. Despite advances in techniques like attention visualization and feature attribution methods such as Integrated Gradients [89], SHAP [90], explaining the reasoning behind a model's predictions in a human-interpretable way remains an unsolved problem. This is especially problematic in sensitive fields like healthcare, law, and autonomous systems, where model transparency and accountability are crucial. Approaches like integrating symbolic reasoning or developing hybrid neuro-symbolic systems are being explored to enhance interpretability, but these remain complex challenges.

The issue of bias and fairness in LLMs is also prominent. Training data typically come from large, uncurated web corpora that reflect societal biases. Consequently, LLMs often reproduce and even amplify these biases, leading to problematic outcomes [49, 91]. For example, gender, racial, and ideological biases can manifest in model outputs. Researchers are developing methods such as adversarial debiasing, counterfactual data augmentation, and fairness-aware loss functions [62, 92] to address these issues. However, achieving fairness across diverse demographic groups, while maintaining model performance, is an ongoing challenge.

Environmental sustainability is another pressing concern. Training large-scale models requires enormous computational power, which translates into high energy consumption and a significant carbon footprint. For instance, training a model like GPT-3 is estimated to emit as much CO_2 as several cars over their lifetimes [21, 48]. In response, approaches like energy-efficient transformers, sparsely activated models, and training-once-use-often paradigms are being explored to reduce the environmental impact. However, these solutions often come at the cost of performance and generalization, highlighting the trade-off between sustainability and capability.

LLMs also pose ethical risks due to their generative abilities. Their capacity to produce highly realistic and coherent text makes them powerful tools for malicious activities like generating disinformation, deepfake content, and phishing attacks. However, this ability to generate fluent text should not be conflated with factual reliability; despite surface-level coherence, LLMs frequently produce hallucinated or logically inconsistent outputs [93]. This dual nature poses challenges in distinguishing genuine fluency from misleading generation. Moreover, the risk of model inversion—where attackers can extract sensitive data from the model—raises serious concerns about data privacy [56, 94]. Research in differential privacy and secure model training aims to mitigate these risks, but ensuring the safe

deployment of LLMs is a complex, multifaceted issue that requires ongoing attention.

Therefore, despite their impressive fluency and linguistic capabilities, LLMs still exhibit limitations in reasoning and problem-solving. While they excel in pattern recognition and context-sensitive generation, they struggle with logical reasoning, causal inference, and complex decision-making tasks. Techniques like chain-of-thought prompting [44] and RAG [53] have been proposed to improve reasoning abilities. Unlike conventional prompting strategies, RAG combines external retrieval mechanisms with generation, helping to ground responses in factual sources and reduce hallucination. RAG also contributes to reducing model size while maintaining performance, and offers scalability benefits by decoupling memorization from inference. These solutions, however, are still in the early stages and often lack robustness in high-stakes environments like scientific discovery or legal analysis.

Overall, while LLMs have demonstrated exceptional capabilities, their challenges—ranging from computational inefficiencies to ethical dilemmas—must be addressed to ensure their responsible development and deployment. The next phase of research will need to focus not only on scaling model sizes but also on enhancing interpretability, fairness, sustainability, and ensuring these models are safely integrated into society.

6. Discussion and Outlook

With LLMs now embedded in high-impact applications across diverse fields, the landscape of NLP has expanded beyond textual understanding into collaborative problem-solving and human augmentation. LLMs are no longer confined to core NLP tasks but have become a transformative force in sectors ranging from healthcare and law to education, creative industries, and scientific research [95]. This section synthesizes key modern applications, emerging research trajectories, and critical considerations that define the current and future state of language modeling.

6.1. Modern Applications of LLMs: From Automation to Augmentation

LLMs such as GPT-4 [28], PaLM [87], and Claude [96] exhibit extraordinary generalization and contextual understanding, allowing them to perform tasks far beyond the traditional scope of NLP— for example, composing functional source code from ambiguous natural language prompts, generating scientific hypotheses from literature surveys, or assisting in legal document drafting with contextual sensitivity. In the domain of Conversational AI, these models power virtual agents that can engage in nuanced, multi-turn dialogue, maintain sentiment tracking, and retain contextual understanding [97]. Such capabilities are revolutionizing industries including mental health support, intelligent tutoring systems, and multilingual customer service, creating more responsive and interactive user experiences.

In scientific domains, models such as BioBERT [98], SciBERT [99], and Galactica [44] are helping researchers with tasks like finding relevant studies, generating new research ideas, and understanding complex biomedical information. Recent work has also demonstrated the utility of prompt-engineered LLMs like Gemma-7b-it in scientific writing assistance, where they are used to evaluate and iteratively refine abstracts based on clarity, coherence, and adherence to academic standards [100]. These models go beyond simple search tools by offering more intelligent and context-aware support for scientific discovery. Similarly, LegalBERT [101] and domain-specific transformer models are driving innovation in legal tech, enabling more accurate automated contract analysis, legal research, and compliance audits. These applications demonstrate the potential of LLMs to augment human expertise in specialized fields by automating tedious tasks and providing deep insights.

Creative industries have also been deeply impacted by the integration of LLMs. These models are now co-authoring screenplays, generating marketing campaigns, composing poetry, and even contributing to video game design. The rise of human-AI co-creativity offers a new paradigm for collaboration, where AI serves not only as a tool but also as a partner in creative processes [102]. This interaction raises important philosophical and practical questions about authorship, originality, and the nature of creativity itself, challenging traditional notions in the creative domain.

Beyond these domains, contemporary applications of LLMs have expanded into more operational and industry-facing roles. Patterns such as retrieval-augmented generation for document-grounded question answering, tool-augmented agents that can call APIs or external knowledge bases, and workflow orchestration in data analysis and automation pipelines illustrate how LLMs are being integrated as components of larger systems. At the same time, safety layers and content moderation

mechanisms are increasingly applied to ensure responsible use and mitigate harmful outputs. Domain-specific deployments span customer support automation, fraud detection and compliance monitoring in financial services, code generation and debugging in software engineering, clinical summarization in healthcare, and scientific discovery in fields like drug design and materials research. Taken together, these advances show how LLMs are pushing research frontiers across domains while fostering a timely and impactful convergence of academic innovation and industrial relevance.

6.2. Future Directions: Research Frontiers and Technical Imperatives

The future trajectory of language models is shaped by several critical research directions and technical challenges:

The development of multimodal systems such as GPT-4-Vision [33], Flamingo [45], and Gato [103] is pushing the boundaries of AI by integrating text, image, audio, and even robotic commands. These systems enable a more holistic understanding of the world, allowing for tasks such as image generation from text prompts and real-time interaction with physical environments. Research will continue to pursue generalist architectures that maintain high performance across diverse tasks while adapting to novel challenges.

Another significant direction involves efficient scaling and sparse architectures. Innovations such as Mixture-of-Experts (MoE) [104], sparse attention mechanisms, and LoRA [51] are becoming crucial to addressing the computational challenges associated with LLMs. These techniques aim to maintain high performance while drastically reducing the computational costs involved in training and inference, which is essential for the development of more sustainable AI systems.

In parallel, advancements in few-shot, zero-shot, and in-context learning are redefining how LLMs acquire new capabilities [19]. By leveraging prompt engineering and dynamic adaptation, these models can perform new tasks with minimal or no additional training data, reducing reliance on large labeled datasets. Future research is likely to explore hybrid approaches that combine RAG [53] and meta-learning strategies to further enhance their flexibility and performance.

The growing use of LLMs has also heightened concerns regarding ethics, fairness, and safety. As these models are deployed at scale, issues such as algorithmic bias, the risk of hallucinations, and the lack of transparency must be addressed [49, 105]. Techniques like RLHF [55] and Constitutional AI [32] show promise in aligning these models with ethical guidelines. However, these technical efforts need to be complemented by socio-technical frameworks to ensure responsible AI deployment in sensitive areas like healthcare, finance, and governance.

Despite their capabilities, LLMs remain poorly understood at a mechanistic level. Research into interpretability and theoretical understanding is a critical area of focus [106]. Exploring how these models internally represent syntax, semantics, and world knowledge could provide valuable insights. Probing methods and layerwise attribution are emerging as promising techniques for unveiling the inner workings of LLMs, enhancing their trustworthiness, debuggability, and transparency.

Finally, the vision for the future of language models involves human-AI collaboration and cognitive integration. Beyond automating tasks, LLMs are expected to act as partners in reasoning, creativity, and decision-making, augmenting human abilities. This shift demands not only better model architectures but also the development of interface design, interaction theory, and co-adaptive learning paradigms [107] that allow for seamless integration between human and AI systems.

7. Conclusion

This survey has traced the evolution of language models from early statistical methods to the rise of large-scale transformer-based architectures, offering a unified perspective across architectural innovation, training paradigms, evaluation strategies, and emerging challenges. We have highlighted how models have evolved from rule-based fluency to data-driven reasoning, and examined critical concerns such as scalability, interpretability, factuality, and ethical deployment.

While early probabilistic models such as n-grams and HMMs provided interpretable patterns of linguistic behavior, the advent of deep learning—through RNNs, LSTMs, and ultimately Transformers—enabled a paradigm shift in both performance and scope. As discussed in Section 3, LLMs like GPT-4 and PaLM have unlocked capabilities far beyond prior expectations, yet their emergence has exposed limitations in traditional benchmarks and surfaced significant risks regarding safety, fairness, and energy efficiency.

This survey finds that many current approaches—such as parameter-efficient training, retrieval-augmented generation, and alignment strategies—are promising but remain fragmented. A recurring theme across recent literature is the tension between scale and control: while scaling tends to improve generalization, it also complicates transparency and increases societal risk. By critically reviewing these trajectories, we identified a gap in unified frameworks that balance performance with interpretability and trust.

Going forward, progress in language modeling should not be measured by scale alone. As this study suggests, the future lies in integrating diverse disciplines—from cognitive science to HCI and social ethics—to co-design systems that are both powerful and aligned with human values. Innovations in neurosymbolic reasoning, modularity, and sustainable computation may drive the next wave of models—not merely as tools for automation, but as collaborators in knowledge creation.

Ultimately, this work underscores that the evolution of language models reflects deeper shifts in our aspirations for AI: from predictive accuracy to interpretive understanding, from mechanical outputs to human-centric interaction. To navigate this transition responsibly, the research community must continue to ask not just “what” these models can do—but “how”, “why”, and “for whom” they should do it.

Authors’ Contributions

All authors equally contributed to this study.

Declaration

None.

Transparency Statement

None.

Acknowledgments

None.

Declaration of Interest

The authors declare that they have no conflict of interest. The authors also declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

Not applicable.

References

[1] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language,* vol. 13, no. 4, pp. 359-394, 1999, doi: 10.1006/csla.1999.0128.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE,* vol. 77, no. 2, pp. 257-286, 2002, doi: 10.1109/5.18626.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[4] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. [Online]. Available: https://cir.nii.ac.jp/crid/1370849946232757637.

[5] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine,* vol. 13, no. 3, pp. 55-75, 2018, doi: 10.1109/MCI.2018.2840738.

[6] Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research,* vol. 57, pp. 345-420, 2016, doi: 10.1613/jair.4992.

[7] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 32, no. 2, pp. 604-624, 2020, doi: 10.1109/TNNLS.2020.2979670.

[8] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences,* vol. 63, no. 10, pp. 1872-1897, 2020, doi: 10.1007/s11431-020-1647-3.

[9] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal,* vol. 27, no. 3, pp. 379-423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[10] F. Jelinek, *Statistical methods for speech recognition*. MIT Press, 1998.

[11] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[12] J. Gao and H. Suzuki, "Long distance dependency in language modeling: an empirical study," in *International Conference on Natural Language Processing*, 2004, pp. 396-405, doi: 10.1007/978-3-540-30211-7_42.

[13] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks,* vol. 5, no. 2, pp. 157-166, 1994, doi: 10.1109/72.279181.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in*

Neural Information Processing Systems, 2014, vol. 27. [Online]. Available: https://proceedings.neurips.cc/paper/5346-sequence-to-sequence-learning-with-neural-.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[16] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645-6649, doi: 10.1109/ICASSP.2013.6638947.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog,* vol. 1, no. 8, p. 9, 2019. [Online]. Available: https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf.

[18] 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. [Online]. Available: http://www.jmlr.org/papers/v21/20-074.html.

[19] T. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems,* vol. 33, pp. 1877-1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_source=transaction&utm_medium=email&utm_campaign=linkedin_newsletter.

[20] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics,* vol. 8, pp. 842-866, 2021, doi: 10.1162/tacl_a_00349.

[21] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, 09 ed., pp. 13693-13696, doi: 10.1609/aaai.v34i09.7123.

[22] Q. Fournier, G. M. Caron, and D. Aloise, "A practical survey on faster and lighter transformers," *ACM Computing Surveys,* vol. 55, no. 14s, pp. 1-40, 2023, doi: 10.1145/3586074.

[23] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," in *arXiv preprint*, 2020, pp. 1-1, doi: 10.48550/arXiv.2004.05150.

[24] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *arXiv preprint*, 2020, pp. 1-1, doi: 10.48550/arXiv.2001.04451.

[25] K. Choromanski *et al.*, "Rethinking attention with performers," in *arXiv preprint*, 2020, pp. 1-1, doi: 10.48550/arXiv.2009.14794.

[26] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," in *arXiv preprint*, 2020, pp. 1-1, doi: 10.48550/arXiv.2006.04768.

[27] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 149-159. [Online]. Available: https://proceedings.mlr.press/v81/binns18a.html.

[28] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue,* vol. 16, no. 3, pp. 31-57, 2018, doi: 10.1145/3236386.3241340.

[29] A. Chowdhery *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research,* vol. 24, no. 240, pp. 1-113, 2023. [Online]. Available: http://www.jmlr.org/papers/v24/22-1144.html.

[30] J. Hoffmann *et al.*, "Training compute-optimal large language models," in *arXiv preprint*, 2022, pp. 1-1, doi: 10.48550/arXiv.2203.15556.

[31] J. Kaplan *et al.*, "Scaling laws for neural language models," *arXiv preprint,* pp. 1-1, 2020, doi: 10.48550/arXiv.2001.08361.

[32] Y. Bai *et al.*, "Constitutional ai: Harmlessness from ai feedback," in *arXiv preprint*, 2022, pp. 1-1, doi: 10.48550/arXiv.2212.08073.

[33] OpenAi, "GPT-4 Technical Report," 2023. [Online]. Available: https://openai.com/research/gpt-4.

[34] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," in *arXiv preprint*, 2023, pp. 1-1, doi: 10.48550/arXiv.2302.13971.

[35] R. Taori *et al.*, "Stanford Alpaca: An instruction-following LLaMA model," in *arXiv preprint*, 2023, pp. 1-1, doi: 10.48550/arXiv.2302.12025.

[36] W. L. Chiang *et al.*, "Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality," in *arXiv preprint*, 2023, pp. 1-1, doi: 10.48550/arXiv.2302.14603.

[37] A. Q. Jiang *et al.*, "Mixtral of experts," in *arXiv preprint*, 2024, pp. 1-1, doi: 10.48550/arXiv.2401.04088.

[38] E. Almazrouei *et al.*, "The Falcon series of open language models," in *arXiv preprint*, 2023, pp. 1-1, doi: 10.48550/arXiv.2311.16867.

[39] V. Sanh *et al.*, "Multitask prompted training enables zero-shot task generalization," in *arXiv preprint*, 2021, pp. 1-1, doi: 10.48550/arXiv.2110.08207.

[40] J. Wei *et al.*, "Emergent abilities of large language models," in *arXiv preprint*, 2022, pp. 1-1, doi: 10.48550/arXiv.2206.07682.

[41] M. Chen *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374,* 2021.

[42] Y. Liu *et al.*, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics,* vol. 8, pp. 726-742, 2020, doi: 10.1162/tacl_a_00343.

[43] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934,* 2020, doi: 10.1007/s11663-020-01867-z.

[44] R. Taylor *et al.*, "Galactica: A large language model for science," *arXiv preprint arXiv:2211.09085,* 2022. [Online]. Available: https://arxiv.org/abs/2211.09085.

[45] J. B. Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," *Advances in Neural Information Processing Systems,* vol. 35, pp. 23716-23736, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

[46] K. Cobbe *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168,* 2021. [Online]. Available: https://arxiv.org/abs/2110.14168.

[47] D. Hendrycks *et al.*, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv:2103.03874,* 2021. [Online]. Available: https://arxiv.org/abs/2103.03874.

[48] D. Patterson *et al.*, "Carbon emissions and large neural network training," *arXiv preprint arXiv:2104.10350,* 2021. [Online]. Available: https://www.kathimerini.gr/wp-content/uploads/2024/07/2104-10350_1.pdf.

[49] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* pp. 610-623, 2021, doi: 10.1145/3442188.3445922.

[50] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608,* 2017. [Online]. Available: https://arxiv.org/abs/1702.08608.

[51] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR,* vol. 1, no. 2, p. 3, 2022. [Online]. Available: https://arxiv.org/pdf/2106.09685v1/1000.

[52] S. Borgeaud *et al.*, "Improving language models by retrieving from trillions of tokens," 2022, pp. 2206-2240. [Online]. Available: https://proceedings.mlr.press/v162/borgeaud22a.html?utm_campaign=The%20Batch&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-8TZzur2df1qdnGx09b-Fg94DTsc3-xXao4StKvKNU2HR51el3n8yOm0CPSw6GiAoLQNKua.

[53] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems,* vol. 33, pp. 9459-9474, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[54] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research,* vol. 23, no. 120, pp. 1-39, 2022. [Online]. Available: http://www.jmlr.org/papers/v23/21-0998.html.

[55] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems,* vol. 35, pp. 27730-27744, 2022. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

[56] N. Carlini *et al.*, "Extracting training data from large language models," *In 30th USENIX Security Symposium (USENIX Security 21),* pp. 2633-2650, 2021. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[57] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *Advances in Neural Information Processing Systems,* vol. 36, pp. 10088-10115, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

[58] K. Konen *et al.*, "Style vectors for steering generative large language model," *arXiv preprint arXiv:2402.01618,* 2024. [Online]. Available: https://arxiv.org/abs/2402.01618.

[59] B. D. Nye, "Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context," *International Journal of Artificial Intelligence in Education,* vol. 25, no. 2, pp. 177-203, 2015, doi: 10.1007/s40593-014-0028-6.

[60] J. Roschelle, J. Lester, and J. Fusco, "AI and the Future of Learning: Expert Panel Report," *Digital Promise,* 2020, doi: 10.51388/20.500.12265/106.

[61] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems,* vol. 35, pp. 24824-24837, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[62] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang, "Gender bias in coreference resolution: Evaluation and debiasing methods," *arXiv preprint,* vol. arXiv:1804.06876, 2018, doi: 10.18653/v1/N18-2003.

[63] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker, "Perplexity-a measure of the difficulty of speech recognition tasks," *The Journal of the Acoustical Society of America,* vol. 62, no. S1, p. S63, 1977, doi: 10.1121/1.2016299.

[64] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," 2002, pp. 311-318, doi: 10.3115/1073083.1073135.

[65] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," 2004, pp. 74-81. [Online]. Available: https://aclanthology.org/W04-1013.pdf.

[66] P. Koehn and C. Monz, "Manual and automatic evaluation of machine translation between European languages," 2006, pp. 102-121, doi: 10.3115/1654650.1654666.

[67] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," *arXiv preprint arXiv:1904.09675,* 2019. [Online]. Available: https://arxiv.org/abs/1904.09675.

[68] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," *arXiv preprint arXiv:1908.10084,* 2019, doi: 10.18653/v1/D19-1410.

[69] D. M. Ziegler *et al.*, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593,* 2019. [Online]. Available: https://arxiv.org/abs/1909.08593.

[70] P. Liang, J. Wei, D. Schuurmans, M. Chen, S. Borgeaud, and L. Reynolds, *Holistic Evaluation of Language Models (HELM)*. 2022.

[71] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, "Gender bias in coreference resolution," *arXiv preprint arXiv:1804.09301,* 2018, doi: 10.18653/v1/N18-2002.

[72] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456,* 2020, doi: 10.18653/v1/2021.acl-long.416.

[73] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-pairs: A challenge dataset for measuring social biases in masked language models," *arXiv preprint arXiv:2010.00133,* 2020, doi: 10.18653/v1/2020.emnlp-main.154.

[74] S. Es, J. James, L. E. Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," 2024, pp. 150-158, doi: 10.18653/v1/2024.eacl-demo.16.

[75] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250,* 2016, doi: 10.18653/v1/D16-1264.

[76] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551,* 2017, doi: 10.18653/v1/P17-1147.

[77] T. Kwiatkowski *et al.*, "Natural Questions: A benchmark for question answering research," *Transactions of the*

Association for Computational Linguistics, vol. 7, pp. 453-466, 2019, doi: 10.1162/tacl_a_00276.

[78] P. Bajaj *et al.*, "MS MARCO: A human generated machine reading comprehension dataset," *arXiv preprint arXiv:1611.09268,* 2016. [Online]. Available: https://arxiv.org/abs/1611.09268.

[79] Z. Yang *et al.*, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint,* vol. arXiv:1809.09600, 2018, doi: 10.18653/v1/D18-1259.

[80] J. H. Clark *et al.*, "Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages," *Transactions of the Association for Computational Linguistics,* vol. 8, pp. 454-470, 2020, doi: 10.1162/tacl_a_00317.

[81] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," *arXiv preprint,* vol. arXiv:1910.11856, 2019, doi: 10.18653/v1/2020.acl-main.421.

[82] A. Wang *et al.*, "SuperGLUE: A stickier benchmark for general-purpose language understanding systems," *Advances in Neural Information Processing Systems,* vol. 32, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.

[83] D. Hendrycks *et al.*, "Measuring massive multitask language understanding," *arXiv preprint,* vol. arXiv:2009.03300, 2020. [Online]. Available: https://arxiv.org/abs/2009.03300.

[84] A. Srivastava *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research,* 2023. [Online]. Available: https://openreview.net/forum?id=uyTL5Bvosj&nesting=2&sort=date-desc.

[85] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *Advances in Neural Information Processing Systems,* vol. 36, pp. 46595-46623, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

[86] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint,* vol. arXiv:1804.07461, 2018, doi: 10.18653/v1/W18-5446.

[87] R. Anil *et al.*, "Palm 2 technical report," *arXiv preprint,* vol. arXiv:2305.10403, 2023. [Online]. Available: https://arxiv.org/abs/2305.10403.

[88] J. Dodge *et al.*, "Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus," *arXiv preprint,* vol. arXiv:2104.08758, 2021, doi: 10.18653/v1/2021.emnlp-main.98.

[89] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, 2017, pp. 3319-3328. [Online]. Available: http://proceedings.mlr.press/v70/sundararajan17a.html.

[90] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems,* vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[91] T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," *Advances in Neural Information Processing Systems,* vol. 29, 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[92] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, "Gender bias in neural natural language processing," *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday,* pp. 189-202, 2020, doi: 10.1007/978-3-030-62077-6_14.

[93] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys,* vol. 55, no. 12, pp. 1-38, 2023, doi: 10.1145/3571730.

[94] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 587-601, doi: 10.1145/3133956.3134077.

[95] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint,* vol. arXiv:2108.07258, 2021.

[96] Anthropic, "Claude: Constitutional AI and Helpful, Harmless, Honest Assistants," 2023. [Online]. Available: https://www.anthropic.com/news/introducing-claude.

[97] C. Zhou *et al.*, "Lima: Less is more for alignment," *Advances in Neural Information Processing Systems,* vol. 36, pp. 55006-55021, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

[98] 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, doi: 10.1093/bioinformatics/btz682.

[99] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," *arXiv preprint,* vol. arXiv:1903.10676, 2019, doi: 10.18653/v1/D19-1371.

[100] A. A. Kharazmi and H. Hassanpour, "Enhancing the Quality of Scientific Writing Using Advanced Language Models: Automated Evaluation and Proofreading," *Journal of AI and Data Mining,* vol. 13, no. 1, pp. 11-24, 2025. [Online]. Available: https://journals.shahroodut.ac.ir/article_3388.html.

[101] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," *arXiv preprint,* vol. arXiv:2010.02559, 2020, doi: 10.18653/v1/2020.findings-emnlp.261.

[102] Z. Ivcevic and M. Grandinetti, "Artificial intelligence as a tool for creativity," *Journal of Creativity,* vol. 34, no. 2, p. 100079, 2024, doi: 10.1016/j.yjoc.2024.100079.

[103] S. Reed *et al.*, "A generalist agent," *arXiv preprint,* vol. arXiv:2205.06175, 2022. [Online]. Available: https://arxiv.org/abs/2205.06175.

[104] N. Shazeer *et al.*, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint,* vol. arXiv:1701.06538, 2017. [Online]. Available: https://arxiv.org/abs/1701.06538.

[105] L. Weidinger *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint,* vol. arXiv:2112.04359, 2021. [Online]. Available: https://arxiv.org/abs/2112.04359.

[106] C. Olah *et al.*, "The building blocks of interpretability," *Distill,* vol. 3, no. 3, p. e10, 2018, doi: 10.23915/distill.00010.

[107] T. Schick *et al.*, "Toolformer: Language models can teach themselves to use tools," *Advances in Neural Information Processing Systems,* vol. 36, pp. 68539-68551, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.