Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377–0430
Vol. 5, No. 4, 2025

Technical Principles of Large Language Models: From Transformer Architectures to Future Challenges

Miles Whitaker
University of Missouri–Kansas City, Kansas City, Missouri, USA
miles.whitaker@umkc.edu

Abstract:

With the rapid development of artificial intelligence technologies, Large Language Models (LLMs) have demonstrated exceptional capabilities in language understanding and generation, becoming a central focus in the field of Natural Language Processing (NLP). This paper systematically reviews the fundamental theories and key technical principles of LLMs, with an emphasis on Transformer architectures and self-attention mechanisms. It further analyzes major pretraining methods such as masked language modeling and causal language modeling, and discusses the critical role of fine-tuning and alignment techniques in practical applications. In addition, it introduces parameter expansion and compression strategies, inference optimization methods, and recent advances in enhancing safety and robustness. Finally, the paper explores the future trends of LLMs in multimodal integration, personalized intelligence, and autonomous agent development, and identifies the major efficiency, robustness, and ethical challenges ahead. This review aims to provide a systematic theoretical reference and technical guide for researchers and practitioners, promoting sustained innovation and application expansion of LLM technologies.

Keywords:

Large Language Models; Transformer; Pretraining Methods; Fine-tuning and Alignment; Inference Optimization; Safety; Multimodality; Future Trends

1. Introduction

With In recent years, with the rapid development of artificial intelligence, Large Language Models (LLMs) have become one of the core research directions in Natural Language Processing (NLP). By conducting self-supervised pretraining on massive text corpora, LLMs have demonstrated outstanding capabilities in language understanding and generation, achieving unprecedented performance across various downstream tasks. The widespread application of the Transformer architecture, in particular, has enabled substantial expansion of model scale, driving advancements in applications such as intelligent dialogue, automatic programming, and complex reasoning [1].

The rise of LLMs is attributed not only to algorithmic and hardware advancements but also to the coordinated evolution of pretraining methods, fine-tuning techniques, inference optimization, and model alignment. This review systematically outlines the technical principles of LLMs, covering basic theories, pretraining and fine-tuning methods, inference optimization strategies, and future challenges, aiming to provide researchers and engineering practitioners with a clear theoretical framework and technical guidance.

2. Fundamental Theories and Technical Principles

The development of LLMs is fundamentally supported by deep neural network theories. Early NLP methods primarily relied on Bag-of-Words models and statistical language models, but these approaches struggled to capture complex linguistic structures and semantic associations. Later, distributed representations, such as Word2Vec and GloVe [2], partially addressed these limitations. With the rise of deep learning, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks became mainstream, enabling the modeling of variable-length sequential data and temporal dependencies [3]. However, RNN-based models faced issues like vanishing gradients and computational bottlenecks, limiting their scalability.

The introduction of the Transformer architecture marked a fundamental shift in sequence modeling. Transformers are entirely based on self-attention mechanisms, enabling each input position to establish global dependencies efficiently without relying on recursion [4]. The encoder-decoder structure, composed of stacked self-attention modules and feed-forward networks, greatly enhances model expressiveness and training efficiency. The multi-head attention mechanism further enables learning diverse relational patterns across different subspaces, enhancing the modeling of complex linguistic phenomena.

On top of the Transformer architecture, pretraining language models typically adopt two main objectives: Masked Language Modeling (MLM) and Causal Language Modeling (CLM) [5]. MLM strategies, used in models like BERT and RoBERTa, mask random portions of the input and train the model to predict the masked tokens, encouraging deep bidirectional contextual representation learning. In contrast, CLM strategies, adopted by the GPT series, train the model to predict the next token based only on past context, making them well-suited for generation tasks.

Studies have shown that as model parameters, training data volume, and computational resources scale up, model performance exhibits predictable improvements, known as scaling laws [6]. This finding has driven the development of ever-larger models such as GPT-3 by OpenAI, PaLM by Google, and LLaMA by Meta [7]. These large models not only achieve breakthroughs on benchmark tasks but also exhibit emerging abilities like zero-shot learning, few-shot learning, and in-context learning.

However, merely scaling up model size encounters diminishing returns and sharply increases training and inference costs. Consequently, research has shifted toward more efficient pretraining methods, multitask and multimodal learning frameworks, and parameter-efficient fine-tuning (PEFT) techniques [8]. The Transformer and its derivative architectures remain the theoretical cornerstone for the ongoing evolution of LLMs and serve as the foundation for exploring more intelligent and reliable language models in the future.

3. Development of Pretraining Methods

Pretraining serves as the core phase in the training of large language models and has evolved from single-objective tasks to multitask and multimodal approaches. Initially, pretraining primarily relied on simple self-supervised tasks, such as the Masked Language Modeling (MLM) objective proposed by BERT [5]. By conducting MLM training on large-scale unlabeled text corpora, models were able to learn rich contextual semantic information, providing powerful feature representations for downstream tasks. BERT and its improved version, RoBERTa, further enhanced pretraining performance by expanding the training data scale and removing the Next Sentence Prediction task [9].

Unlike the BERT series, the GPT series models employ a Causal Language Modeling (CLM) strategy, emphasizing unidirectional sequence generation [10]. This design enables GPT models to excel at generation tasks, producing coherent and logically consistent long texts. GPT-2 first demonstrated that massive autoregressive pretraining alone could enable zero-shot or few-shot transfer to multiple downstream tasks. GPT-3 significantly expanded the parameter scale to 175 billion, showcasing new capabilities such as in-context learning [7].

In terms of pretraining objectives, new task designs such as Span Prediction (e.g., SpanBERT) and Permutation Language Modeling (e.g., XLNet) have emerged. These methods aim to enhance the model's understanding of sentence-level, paragraph-level, and even document-level structures [11]. For example, XLNet captures long-distance dependencies more effectively by learning all possible permutations of token sequences.

As pretraining methods evolved, multitask learning became an important approach to improving model generalization. T5 (Text-to-Text Transfer Transformer) unified all tasks into a text-to-text format, encompassing classification, question answering, summarization, and more within a single modeling framework [12]. This approach not only simplified downstream adaptation but also significantly enhanced task transferability. Similarly, UL2 (Unifying Language Learning Paradigms) proposed a framework that combines various pretraining objectives, including MLM, CLM, and Span Infilling, to further boost comprehensive performance [13].

Beyond unimodal text pretraining, multimodal pretraining has become a prominent research direction. CLIP (Contrastive Language-Image Pretraining) employed contrastive learning to embed images and texts into the same vector space, enabling cross-modal understanding and generation [10]. Models like ALIGN and Florence expanded the scale and coverage of multimodal datasets, allowing models to simultaneously process text, images, and audio. Multimodal large models such as Flamingo and Gemini have further advanced capabilities in visual reasoning, cross-modal retrieval, and more, demonstrating the early potential of Artificial General Intelligence (AGI) [14].

Overall, the development of pretraining methods can be summarized along three main lines: designing richer self-supervised tasks to enhance language understanding and generation, employing multitask and unified modeling to improve generalization, and transcending single-modality boundaries to build models that integrate multiple forms of perception. These technological advances have laid a solid foundation for the success of large language models across diverse application scenarios.

4. Fine-tuning and Alignment Techniques

Although pretraining grants large language models powerful general capabilities, fine-tuning and alignment processes are essential to optimize their performance for specific application scenarios. Traditional supervised fine-tuning typically involves training on small-scale, high-quality labeled datasets to adjust model parameters for specific tasks. However, as model sizes increase, simple supervised fine-tuning often fails to fully exploit the potential of pretrained models and is prone to overfitting and catastrophic forgetting [15].

To better align model outputs with human expectations, Reinforcement Learning with Human Feedback (RLHF) was introduced and widely applied in models like GPT-3.5 and GPT-4 [11]. RLHF consists of three main steps: first, supervised fine-tuning (SFT) is performed on high-quality data; second, a reward model (RM) is trained to evaluate output quality; third, reinforcement learning algorithms such as Proximal Policy Optimization (PPO) are used to optimize the language model to maximize the reward signal. This process significantly improves model performance in dialogue systems by enhancing the coherence, relevance, and safety of generated texts.

In addition to RLHF, Instruction Fine-tuning has become an important technique. Research exemplified by InstructGPT demonstrates that fine-tuning on large-scale human-written instruction datasets enables models to better understand and execute natural language instructions [16]. This approach not only improves controllability but also enhances zero-shot and few-shot generalization abilities, promoting the development of instruction-centered human-computer interaction paradigms.

To reduce the computational cost of fine-tuning large models, Parameter-Efficient Fine-Tuning (PEFT) methods have emerged. Techniques such as Adapters, LoRA (Low-Rank Adaptation), and Prefix Tuning involve updating only a small number of additional parameters without modifying the entire model, significantly reducing fine-tuning's computational and storage overhead [17]. LoRA, in particular, applies low-rank decomposition to weight matrices during fine-tuning, enabling large models to adapt to new tasks at low cost and has become a mainstream industrial solution.

Alignment techniques also encompass safety and ethical alignment. For example, introducing refusal mechanisms reduces harmful responses to sensitive or illegal queries, and bias mitigation techniques aim to reduce biases in sensitive attributes such as gender and race [18]. Emerging approaches like Constitutional AI attempt to embed rule-based systems into model reasoning and generation processes, representing a significant trend in alignment research.

Overall, fine-tuning and alignment are not only key steps for adapting pretrained large models to practical tasks but also crucial for improving model safety, reliability, and social acceptability. In the future, optimizing fine-tuning efficiency, enhancing alignment quality, and achieving unified alignment in multimodal, multilingual, and multitask environments will remain important research directions.

5. Parameter Expansion and Compression Techniques

As large language models continue to scale up, with parameter counts growing from hundreds of millions to hundreds of billions, managing model size efficiently while maintaining performance becomes a major challenge. Early research indicated that coordinated scaling of parameters, data, and compute leads to systematic performance improvements, a phenomenon known as scaling laws [6]. Models such as OpenAI’s GPT-3, Google’s PaLM, and Anthropic’s Claude series have achieved remarkable performance through large-scale expansion. However, the computational and energy costs associated with scaling have driven extensive research into model compression and efficient modeling methods.

For parameter expansion, model parallelism and data parallelism techniques are widely used. Model parallelism partitions large models across multiple computing devices, while data parallelism accelerates training by processing different data batches across devices. Optimizers like ZeRO (Zero Redundancy Optimizer) further distribute optimization states across devices, significantly reducing memory overhead [19]. Frameworks such as Megatron-LM and DeepSpeed integrate multiple parallelism strategies, supporting the training of models with hundreds of billions of parameters and becoming essential tools for large-scale pretraining.

To lower the inference and deployment costs of large models, various parameter compression techniques have been developed. Model pruning removes redundant connections or neurons to reduce model size and computation; knowledge distillation trains smaller models to mimic the outputs of larger models, achieving lightweight models [20]. Models like TinyBERT and DistilBERT maintain high accuracy while greatly reducing inference latency and storage requirements.

Recently, low-rank approximation methods have become a hot topic in parameter compression research. LoRA (Low-Rank Adaptation) applies low-rank decomposition to weight matrices during fine-tuning, significantly reducing computational and storage overhead [17]. QLoRA further combines this with quantization techniques, compressing weight matrices to 4-bit representations to minimize fine-tuning and inference costs [21]. These methods not only facilitate efficient fine-tuning but also significantly reduce model size during deployment.

Quantization techniques compress floating-point weights into low-bit integer formats such as INT8 or INT4, effectively reducing storage needs and inference latency. Methods like GPTQ and AWQ achieve efficient inference of large models while maintaining accuracy, enabling consumer-grade hardware to run models with tens of billions of parameters [22].

Additionally, Mixture of Experts (MoE) models propose sparse activation mechanisms, where only a subset of submodels is activated during each inference step, significantly reducing computational costs. Models like GShard, Switch Transformer, and GLaM achieve nearly linear growth in inference cost relative to model size [23]. MoE methods provide a new direction for scaling future ultra-large models.

Overall, parameter expansion and compression techniques form the technical foundation for the sustainable development of large language models. Future research will continue to focus on maintaining performance while improving the energy efficiency of training and inference.

6. Inference Optimization and Deployment Strategies

The inference process of large language models involves massive parameters and complex computations, making inference optimization a critical step for practical deployment. Inference optimization aims to reduce inference latency, memory footprint, and energy consumption without significantly degrading model performance, thereby meeting the diverse needs of various application scenarios.

Model quantization is one of the most common inference optimization techniques. By compressing weights and activations from 32-bit floating-point (FP32) to lower-bit formats such as FP16, INT8, or INT4, quantization significantly reduces model size and accelerates inference. Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT) correspond to post-inference and training-time quantization strategies, respectively, with QAT better preserving model accuracy during compression [24]. Techniques like LLM.int8, GPTQ, and AWQ leverage advanced quantization strategies to enable inference of billion-scale models on consumer hardware.

Inference pruning dynamically skips redundant computation units during inference to further reduce computational costs. Early methods such as Dynamic Sparsity and Block Sparse Attention sparsify attention computations, effectively speeding up inference for large models [25]. Token Pruning and Layer Skipping strategies have also been proposed to dynamically adjust the computation path based on the input during inference, enhancing responsiveness.

Caching and reuse strategies are particularly important in long text generation and continuous dialogue scenarios. By caching intermediate computation results (such as Key-Value Cache) of previous layers, repeated computations are avoided, significantly reducing inference latency [26]. Caching mechanisms have become key technologies for improving interactive experiences in applications like chatbots and intelligent assistants that require long-context management.

Parallel and distributed inference techniques are also widely adopted. Techniques like pipeline parallelism, tensor parallelism, and tensor slicing efficiently distribute inference workloads across multiple devices, supporting real-time deployment of ultra-large models. Inference acceleration frameworks such as DeepSpeed-Inference and FasterTransformer integrate multiple optimization methods, including quantization, pruning, and parallel inference [19].

To adapt to the computational resource constraints of mobile and edge devices, researchers have proposed edge inference and tiny model distillation methods. By incorporating efficient architecture designs such as MobileBERT and TinyBERT during pretraining and optimizing through quantization and distillation afterward, models can achieve satisfactory performance while being suitable for edge deployment [20].

In summary, the continuous development of inference optimization and deployment strategies enables large language models to achieve efficient inference across servers, personal devices, and edge environments, greatly expanding their application boundaries. In the future, inference optimization will continue to evolve toward goals such as lower latency, higher throughput, and lower energy consumption, driving the widespread adoption of large language models.

7. Safety, Robustness, and Ethical Issues

Although large language models have achieved remarkable progress in natural language processing, they still face significant challenges in safety, robustness, and ethics. The text generated by LLMs may contain false information, harmful content, biased expressions, or even be exploited maliciously, severely impacting their usability and social acceptability. Therefore, enhancing the safety and robustness of LLMs and ensuring ethical compliance have become indispensable aspects of research and application.

Safety issues mainly include harmful content generation, sensitive information leakage, and adversarial attacks. LLMs may inadvertently generate violent, pornographic, or hateful content, posing potential risks [27]. Additionally, sensitive personal information contained in training data may be memorized and leaked by models, raising serious privacy concerns. To address these issues, researchers have proposed multiple protection strategies, such as filtering and debiasing training data to reduce exposure to harmful content, introducing refusal generation mechanisms during inference to proactively block sensitive queries, and applying differential privacy techniques to lower the probability of training data leakage [28].

Robustness issues manifest in the degradation of model performance when faced with adversarial inputs, noise perturbations, or out-of-distribution data. Adversarial attacks can use small perturbations to induce incorrect or inappropriate model outputs, revealing vulnerabilities [29]. To enhance robustness, researchers have explored adversarial training, perturbation detection, and robust inference methods. By introducing perturbed samples during training or designing robust optimization objectives, models can maintain high reliability under various input variations.

Ethical issues encompass bias, discrimination, transparency, and explainability. Due to historical biases embedded in training data, LLMs may produce unfair outputs related to sensitive attributes such as gender, race, and religion [30]. To mitigate bias, methods such as data rebalancing, bias detection and mitigation, and fairness optimization have been proposed. Furthermore, improving model decision-making transparency, for example, through attention visualization or generative explanations, helps enhance user understanding and trust in model behavior.

Alignment techniques are effective in improving safety and ethical compliance. Reinforcement Learning with Human Feedback (RLHF) and Constitutional AI embed human values and rule systems into model behavior, mitigating harmful generation and ethical bias issues [18]. Additionally, multiple rounds of safety evaluation and red teaming have become important steps before deployment to systematically assess model behavior under extreme inputs and define safety boundaries.

In summary, the safety, robustness, and ethical challenges of LLMs are highly complex and dynamic. Future research must establish a closed-loop mechanism across model design, training data construction, inference control, and post-deployment supervision to continuously monitor and optimize model behavior, ensuring safe, reliable, and responsible operation across diverse application environments.

8. Future Trends and Challenges

With the continuous evolution of large language model technologies, future development trends are expected to be multidimensional and multilayered, accompanied by numerous unresolved challenges. First, model efficiency and energy efficiency will become central issues. The current training and inference processes of large models consume massive computational resources and energy, raising concerns about sustainability and environmental impact [31]. In the future, low-resource efficient training methods (such as sparse training and self-distillation), lightweight inference techniques (such as 4-bit quantization and token merging), and the concept of Green AI are expected to become mainstream research directions.

Second, enhanced multimodal and interactive capabilities are critical for expanding the application boundaries of LLMs. Multimodal large models (Multimodal LLMs) integrate language with images, audio, video, and action commands, driving the evolution from single-modality language intelligence toward general perception and reasoning intelligence [14]. Models like Gemini and Flamingo have already demonstrated strong potential in complex multimodal tasks. In the future, multimodal LLMs will play key roles in fields such as autonomous driving, medical diagnosis, and human-computer interaction.

Third, adaptive and personalized learning capabilities will be crucial for improving user experiences. Current LLMs typically use unified parameters for all tasks and users, despite significant differences across application scenarios and individual needs. By incorporating mechanisms such as personalized fine-tuning, continual learning, and memory-augmented models, LLMs can dynamically adjust behavior based on user interaction history and context, achieving more accurate and natural interactions [32].

Fourth, the trend toward agentification is increasingly apparent. Future LLMs will not only act as passive text generators but will evolve into intelligent agents capable of autonomous planning, decision-making, and execution. Early explorations such as AutoGPT and BabyAGI have demonstrated the potential of LLMs in building autonomous agents by calling external tools, managing memory, and executing multi-step tasks [33]. Enhancing LLMs' capabilities in long-term reasoning, environmental modeling, and goal decomposition will be key to advancing agentification.

However, numerous challenges remain, including the diminishing returns of model scaling, limitations in long-context modeling capabilities, persistent threats to robustness and safety, and issues of ethical governance and regulatory compliance. Especially as LLMs are applied in high-risk fields such as decision support, education, and healthcare, balancing technological innovation with social responsibility becomes an urgent issue.

9. Conclusion

As a frontier area in current artificial intelligence development, large language models have demonstrated exceptional capabilities in language understanding and generation, driving revolutionary advances in natural language processing and related applications. This paper systematically reviewed the fundamental theories and technical principles of LLMs, covering the Transformer architecture, self-attention mechanisms, pretraining methods, fine-tuning and alignment techniques, parameter expansion and compression strategies, inference optimization methods, and the challenges of safety and ethics. Additionally, the paper explored the future development trends of LLMs in areas such as efficiency improvement, multimodal integration, personalized intelligence, and autonomous agent development.

Although LLMs have achieved impressive results in many fields, their sustainable development still faces a series of complex challenges. Future research needs to seek a balance between improving performance and reducing resource consumption, achieve synergy between expanding model capabilities and ensuring safety and reliability, and continuously explore the balance between technological advancement and ethical compliance. Through ongoing innovation and multi-stakeholder collaboration, large language models are expected to play an increasingly important role in building more intelligent, reliable, and responsible AI systems.

References

[1] Brown, T. B., Mann, B., Ryder, N., et al. Language Models are Few-Shot Learners. NeurIPS, 2020.

[2] Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient Estimation of Word Representations in Vector Space. ICLR, 2013.

[3] Hochreiter, S., Schmidhuber, J. Long Short-Term Memory. Neural Computation, 1997.

[4] Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is All You Need. NeurIPS, 2017.

[5] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL, 2019.

[6] Kaplan, J., McCandlish, S., Henighan, T., et al. Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361, 2020.

[7] Sun, Q., Duan, S. User Intent Prediction and Response in Human-Computer Interaction via BiLSTM. Journal of Computer Science and Software Applications, 5(3), 2025.

[8] Liao, X., Zhu, B., He, J., Liu, G., Zheng, H., Gao, J. A Fine-Tuning Approach for T5 Using Knowledge Graphs to Address Complex Tasks. arXiv preprint arXiv:2502.16484, 2025.

[9] Liu, Y., Ott, M., Goyal, N., et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692, 2019.

[10] Radford, A., Kim, J. W., Hallacy, C., et al. Learning Transferable Visual Models from Natural Language Supervision. ICML, 2021.

[11] Fang, Z., Zhang, H., He, J., Qi, Z., Zheng, H. Semantic and Contextual Modeling for Malicious Comment Detection with BERT-BiLSTM. arXiv preprint arXiv:2503.11084, 2025.

[12] Sun, D., He, J., Zhang, H., Qi, Z., Zheng, H., Wang, X. A LongFormer-Based Framework for Accurate and Efficient Medical Text Summarization. arXiv preprint arXiv:2503.06888, 2025.

[13] Tay, Y., Dehghani, M., Bahri, D., Metzler, D. UL2: Unifying Language Learning Paradigms. ICLR, 2023.

[14] Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. MIT Press, 2016.

[15] Wei, J., Tay, Y., Bommasani, R., et al. Emergent Abilities of Large Language Models. arXiv preprint arXiv:2206.07682, 2022.

[16] Hu, E. J., Shen, Y., Wallis, P., et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR, 2022.

[17] Bai, Y., Kadavath, S., Kundu, S., et al. Constitutional AI: Harmlessness from AI Feedback. arXiv preprint arXiv:2212.08073, 2022.

[18] Rajbhandari, S., Ruwase, O., Rasley, J., He, Y. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. SC, 2020.

[19] Jiao, X., Yin, Y., Shang, L., et al. TinyBERT: Distilling BERT for Natural Language Understanding. Findings of EMNLP, 2020.

[20] Dettmers, T., Pagnoni, A., Holtzman, A., et al. QLoRA: Quantization-Aware Low-Rank Adapter Tuning for LLMs. ICML, 2023.

[21] Frantar, E., Alistarh, D. Optimal Brain Compression: A Framework for Practical LLM Compression. NeurIPS, 2023.

[22] Lepikhin, D., Lee, H., Xu, Y., et al. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. ICLR, 2021.

[23] Banner, R., Hubara, I., Hoffer, E., Soudry, D. Post Training 4-bit Quantization of Convolutional Networks for Rapid-Deployment. NeurIPS, 2019.

[24] Child, R., Gray, S., Radford, A., Sutskever, I. Generating Long Sequences with Sparse Transformers. arXiv preprint arXiv:1904.10509, 2019.

[25] Shoeybi, M., Patwary, M., Puri, R., et al. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv preprint arXiv:1909.08053, 2019.

[26] Wu, L., Gao, J., Liao, X., Zheng, H., Hu, J., Bao, R. Adaptive Attention and Feature Embedding for Enhanced Entity Extraction Using an Improved BERT Model. In 2024 4th International Conference on Communication Technology and Information Technology (ICCTIT) (pp. 702-705). IEEE, 2024.

[27] Carlini, N., Tramer, F., Wallace, E., et al. Extracting Training Data from Large Language Models. USENIX Security, 2021.

[28] Wang, S., Liu, Z., Peng, B. A Self-Training Framework for Automated Medical Report Generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 16443-16449), 2023.

[29] Sheng, E., Chang, K. W., Natarajan, P., Peng, N. The Woman Worked as a Babysitter: On Biases in Language Generation. EMNLP, 2019.

[30] Strubell, E., Ganesh, A., McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. ACL, 2019.

[31] Madotto, A., Lin, Z., Wu, C. S., et al. Continual Learning in Task-Oriented Dialogue Systems. ICLR, 2021.

[32] Gao, J., Lyu, S., Liu, G., Zhu, B., Zheng, H., Liao, X. A Hybrid Model for Few-Shot Text Classification Using Transfer and Meta-Learning. arXiv preprint arXiv:2502.09086, 2025.

[33] Zhu, Z., Zhang, Y., Yuan, J., Yang, W., Wu, L., & Chen, Z. NLP-Driven Privacy Solutions for Medical Records Using Transformer Architecture.