Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Research and Analysis on Text Interaction Methods Based on Large Language Models

Shangze Yu *

Information Engineering College, Wuhan College, Wuhan, China

* Corresponding Author Email: 23202030136@whxy.edu.com

Abstract. Large Language Models (LLMs), as a cutting-edge technology in the field of artificial intelligence, have revolutionized human-computer interaction paradigms by integrating natural language processing, Transformer architecture, and reinforcement learning techniques, thereby achieving in-depth understanding and generation of linguistic logic and emotions. Their technological core lies in the Transformer architecture trained on massive datasets, which can capture complex semantics and contextual associations, optimize generation quality through reinforcement learning, and enhance interpersonal-like interaction via affective computing. This paper systematically reviews the technical framework and application scenarios of LLMs: in fields such as intelligent customer service, educational assessment, and game narration, these models demonstrate application values such as multi-turn dialogue, personalized learning path planning, and dynamic plot generation. Meanwhile, it delves into the challenges faced by technological development, including ethical risks arising from biases in training data, deployment cost issues due to high computational power requirements, and deficiencies in the controllability of generated content. In response to these issues, collaborative solutions such as multimodal data fusion, lightweight model deployment, and interdisciplinary ethical governance are proposed. Research indicates that LLMs are at a critical stage of transitioning from technological breakthroughs to large-scale applications, and their sustainable development necessitates the construction of a technological governance framework to achieve social value balance, on top of algorithm optimization and computational power enhancement. Future research should focus on enhancing model interpretability, exploring green computing pathways, and promoting the virtuous development of technology through human-machine collaboration mechanisms.

Keywords: Large language model; Natural language processing; Affective computing; Reinforcement learning; Transformer.

1. Introduction

In recent years, Large Language Models (LLMs) have emerged as a revolutionary technology in the field of Natural Language Processing (NLP). These models, based on generative architectures and deep neural networks such as Transformer, automatically capture grammatical, semantic, and implicit logical patterns in language through pre-training on massive datasets. For instance, models like BERT and GPT series have demonstrated exceptional performance in language understanding and generation tasks through this pre-training mechanism. With the continuous advancement of pre-training technologies, LLMs not only excel in traditional tasks such as text generation, translation, and question answering but also further optimize generation strategies through reinforcement learning, thereby enhancing the quality and intelligence of textual interactions with users [1].

Among these advancements, few-shot learning and prompt engineering have enabled models to possess strong generalization capabilities when faced with new tasks. Early research indicated that through carefully designed prompts, models could accomplish complex tasks with minimal fine-tuning. Subsequent work introduced Chain-of-Thought Prompting, which requires the model to output a series of intermediate reasoning steps before generating the final answer, significantly improving generation quality in logical reasoning and multi-step problem-solving [2]. Meanwhile, Reinforcement Learning from Human Feedback (RLHF) has been widely adopted to adjust the

outputs of LLMs, aligning them more closely with human expectations and demonstrating higher interactive intelligence in dialogue generation and task execution [3].

In addition to these technological breakthroughs, prompt programming strategies have also played a crucial role in the textual interactions of LLMs. The prompt programming method proposed by Reynolds and McDonell enhances the model's ability to efficiently respond to diverse user needs through systematic design and optimization of prompts [4]. These studies not only provide theoretical support for the interaction mechanisms of LLMs but also lay a solid foundation for practical problem-solving in applications such as dialogue systems and virtual assistants.

Despite the significant progress in text generation and interaction, LLMs still face numerous challenges in practical applications. Firstly, the pre-training and fine-tuning processes require substantial computational resources, and the increasing model size drives up training and deployment costs. Secondly, due to the diverse and vast sources of training data, biases in the data may be inadvertently learned and amplified by the model, affecting the fairness and reliability of its outputs. Furthermore, ethical controversies are increasingly prominent, with issues such as privacy breaches and the generation of harmful information posing significant barriers to the widespread adoption of LLMs.

Given these challenges, a thorough analysis of the technical architecture of LLMs, a summary of their practical applications in textual interaction, and an examination of the various obstacles they face are of great importance for advancing this field. This paper aims to conduct research from the following aspects: Firstly, we will delve into the pre-training mechanisms based on the Transformer architecture and reinforcement learning fine-tuning methods, exploring how models learn language patterns from massive datasets and how they utilize human feedback to optimize generation strategies. Secondly, by reviewing cutting-edge literature including Chain-of-Thought, RLHF, and prompt programming, we will summarize the current research status and key technologies in textual interaction methods. Finally, in response to practical issues such as high computational costs, data biases, and ethical controversies faced by models, we will explore possible future directions, such as model compression, multimodal interaction, and safety-control technologies. Through a systematic exposition of the above content, this paper seeks to provide comprehensive references and insights for both theoretical research and practical applications of LLMs in the field of textual interaction.

In conclusion, as a quintessential representative of ultra-large-scale pre-trained models, LLMs are leading the transformation of artificial intelligence from statistical learning to intelligent interaction. This paper will systematically analyze the latest advancements in this field, focusing on technical architecture, application status, and challenges faced, and will provide forward-looking discussions on future development trends.

2. Overview of Related Technologies

2.1. Technological development

The development of natural language processing (NLP) technology has undergone a paradigm shift from rule-based systems to deep learning. Early statistical rule-based systems, such as n-gram models and Hidden Markov Models (HMM), relied on manually designed features and rules, processing linguistic sequences through word frequency statistics and probabilistic models. While these systems achieved initial success in tasks like machine translation and speech recognition, they struggled to capture long-range dependencies, had limited capabilities in handling rare words and complex contexts, and lacked contextual understanding. With the advent of deep learning, Recurrent Neural Networks (RNN) and their improved variants, Long Short-Term Memory Networks (LSTM), were introduced, partially addressing the limitations of traditional models through temporal modeling. However, their serial computational structures led to inefficient training, vanishing gradient problems restricted their ability to process long sequences, and they failed to effectively integrate global

semantic information, resulting in insufficient coherence and logic in generated text. This prompted researchers to explore more efficient architectures.

The revolutionary breakthrough of the Transformer architecture stems from its core self-attention mechanism. This mechanism computes the relevance weights between each word and all other words in the input sequence, enabling parallel modeling of global dependencies and breaking the sequential processing constraints. This significantly enhances the ability to capture long-range dependencies and supports efficient parallel computation. Introduced in 2017, the Transformer quickly became the cornerstone of NLP. In 2018, BERT, through its bidirectional encoder and masked language model (MLM) pre-training, pioneered learning bidirectional contextual semantics without task-specific annotated data, boosting accuracy on the GLUE benchmark by over 15%. The GPT series, on the other hand, achieved exponential scaling of parameters (from 117 million in GPT-1 to trillion-scale in GPT-4), enabling few-shot to zero-shot learning and expanding application boundaries through multimodal fusion. The flexibility and scalability of the Transformer laid the foundation for the development of Large Language Models (LLMs), fundamentally altering the technological trajectory of NLP.

The rise of large-scale pre-trained models marks a critical stage in LLM development. Take GPT-3 (2020) as an example, its 175 billion parameters enabled zero-shot learning for the first time, achieving 78% accuracy on the TriviaQA question-answering task without fine-tuning, demonstrating the significant impact of parameter scale on semantic understanding and generalization capabilities. The same year, the T5 model, through its unified "text-to-text" framework, achieved joint training across multiple tasks such as translation, summarization, and question-answering, further validating the role of model scaling in cross-task adaptability. Meanwhile, prompt engineering guides model output through natural language prompts, reducing the need for task-specific fine-tuning. Retrieval-Augmented Generation (RAG) integrates external knowledge bases to retrieve relevant information in real-time, reducing model "hallucinations" by approximately 30%, as experiments show. These techniques optimize the practicality and reliability of models, driving LLMs from closed pre-training to dynamic knowledge fusion paradigms, offering more flexible solutions for complex tasks.

2.2. Core Method Analysis

2.2.1. Natural Language Processing (NLP)

The core principles of Natural Language Processing (NLP) encompass pre-trained language models and semantic analysis techniques. In the realm of pre-trained language models, BERT [5] captures contextual semantics through bidirectional encoding, while GPT [6] utilizes autoregressive generation for text continuation. Semantic analysis techniques include dependency parsing and named entity recognition. Dependency parsing [7] is employed to analyze the grammatical relationships among various components in a sentence, and named entity recognition [8] extracts entities such as person names, place names, and organization names from text, with an accuracy rate reaching 92%. The advantage of NLP lies in its multi-task generalization, meaning a single model can handle multiple tasks such as translation, summarization, and question-answering simultaneously, as demonstrated by the T5 model achieving a comprehensive score of 90.3 on the GLUE benchmark. Additionally, NLP's real-time optimization capability is reflected in incremental learning, enabling models to adapt to new domain data, such as certain customer service systems updating their corpora weekly. However, NLP also faces challenges, including gender and racial biases present in training corpora, which may lead to biased model outputs, such as gender recommendation biases of up to 18% in recruitment scenarios. Furthermore, handling complex contexts, such as recognizing irony, metaphor, and other non-literal expressions, poses significant difficulties, with misjudgment rates potentially reaching 25%.

2.2.2. Affective Computing (FAtiMA Toolkit)

The core principles of the FAtiMA Toolkit include three-dimensional affective modeling and a dynamic strategy engine. Three-dimensional affective modeling [9] quantifies users' emotional states through three dimensions: Valence, Arousal, and Dominance. The dynamic strategy engine [10] adjusts the system's response strategy based on users' real-time emotional states. For instance, when detecting user anxiety, the system automatically switches to a comforting mode. The advantages of affective computing lie in its ability to significantly enhance user satisfaction through emotional interactions, with a 25% increase in user satisfaction observed in psychological counseling robots. Additionally, integrating multimodal technologies such as speech tone recognition (e.g., pitch analysis) and facial expression detection can significantly improve the accuracy of emotion recognition, with an accuracy rate of up to 88%. However, affective computing also faces challenges, one of which is real-time bottlenecks. Due to the need for multimodal data processing, response delays may increase by approximately 30%, for example, from 100ms to 130ms. Furthermore, misjudgments of complex emotions still persist, particularly with a 15% misjudgment rate in irony recognition, necessitating the introduction of contextual reasoning mechanisms to improve recognition accuracy.

2.2.3. Model Architecture (Transformer)

The core principles of the Transformer model include the multi-head self-attention mechanism and positional encoding. The multi-head self-attention mechanism captures semantic features at different levels through parallel computation of multiple attention heads. For example, GPT-4 employs 128 attention heads to enhance semantic understanding. Positional encoding adds positional information to the input sequence to compensate for the self-attention mechanism's insensitivity to order, thereby preserving sequential information [11]. One of the advantages of the Transformer model is its powerful long-text modeling capability, achieving an accuracy rate of up to 85% in logical coherence analysis tasks, compared to only 72% for traditional RNN models [12]. Additionally, the Transformer architecture demonstrates strong multi-task compatibility, enabling the same model to handle various tasks such as machine translation, text generation, and summary extraction [13].

However, the Transformer model also faces challenges, one of which is high computational complexity. Due to its quadratic complexity with respect to the input sequence length n, processing a 1000-word text may require 50GB of video memory. Furthermore, the Transformer's performance in long-range coherence remains limited, particularly in multi-turn dialogues, where the loss rate of core intentions increases by 40% when the number of dialogue turns exceeds 10.

2.2.4. Reinforcement Learning Algorithm

The principle of Deep Q-Learning (DQN) is to approximate the Q-value function using a deep neural network, combined with Experience Replay and a Target Network to enhance learning stability. DQN has demonstrated outstanding performance in Atari games, achieving scores three times higher than traditional Q-learning [14]. The main challenge of DQN is that it only supports discrete action spaces, such as game buttons, and therefore cannot handle continuous control tasks, such as the movement of robot joints.

The principle of the Actor-Critic algorithm is that two networks work collaboratively: the Actor network generates action policies, while the Critic network evaluates the value of states. Through the Advantage Function, the algorithm optimizes the policy gradient [15]. In robot control tasks, the Actor-Critic algorithm can improve the precision of continuous actions to 95%. However, since the Actor and Critic networks need to be trained collaboratively, the convergence speed is relatively slow, and the training time usually increases by 50%.

The principle of the Soft Actor-Critic (SAC) algorithm is to introduce entropy regularization into reinforcement learning to encourage more exploratory behavior and constrain moral hazard (such as compliance of generated content) through Lagrangian relaxation [16]. In game NPC dialogues, the SAC algorithm can increase the diversity of generated dialogues by 30 times while reducing the

proportion of non-compliant content by 37%. However, the SAC algorithm also faces some challenges, particularly the tendency to fall into local optima, leading to repetitive dialogue templates, and high training costs, typically requiring 1000 GPU hours.

3. Technical Summary

Table 1 illustrates the application scenarios, key advantages, and core challenges of six different technologies (NLP, FAtiMA Toolkit, Transformer, DQN, Actor-Critic, and SAC). These technologies span multiple domains ranging from language understanding and generation to emotional interaction and complex strategic decision-making. Each technology boasts unique strengths, such as multi-task generalization, real-time optimization, and global dependency capture, but also encounters challenges including data bias, complex context handling, and real-time latency.

Table 1. Technology Summary and Comparison.

Technology	application scenarios	key advantages	core challenges
NLP	Language Understanding and Generation	Multi-task generalization, real-time optimization	Data bias, complex context handling
FAtiMA Toolkit	Emotional Interaction	Emotion recognition, dynamic strategy	Real-time latency, sarcasm misjudgment
Transformer	Long-text Modeling	Global dependency capture, multi-task compatibility	High computational complexity, long-range coherence
DQN	Discrete Action Decision-Making	High-dimensional state processing, training stability	Limited to discrete actions, Q-value overestimation
Actor-Critic	Continuous Control and Complex Strategies	Flexible action space, direct policy optimization	Dual-network coordination difficulty, slow convergence
SAC	Morally Sensitive Scenarios	Balance of diversity and controllability	High computational resource consumption, local optima

4. Application Scenarios and Challenges

4.1. Application Scenarios

In the field of commercial services, an e-commerce platform constructed an intelligent customer service system by fine-tuning the GPT-3 model, achieving automated multi-round dialogues for return policy interpretation and promotion recommendations. This reduced human intervention rates by 40% and shortened average response times to 1.2 seconds. In the game development scenario, the technical demo of 1001 Nights integrated GPT-4 to dynamically generate nonlinear storylines, allowing players to trigger over 120 branching paths through natural language interactions. Third-party evaluations showed a 28% increase in player immersion scores and a 2.3-fold increase in story replay rates. In the realm of educational assessment, the ReaderBench system, based on the Transformer architecture, enabled automated essay grading, achieving a 92% accuracy rate in grammar correction and an F1 score of 0.87 in logical coherence analysis. The final scores showed an 85% consistency with teachers' manual evaluations, practically increasing teachers' grading efficiency by threefold and improving students' writing abilities by an average of 20%.

4.2. Technical Challenges

At the level of technological application, large language models (LLMs) face multiple real-world challenges: In terms of data dependency, the scarcity of professional corpora in vertical domains (such as healthcare and law) limits the models' generalization capabilities. For example, in healthcare scenarios, the rigorousness of medical terminology and the sensitivity of case data make models prone to misdiagnosis risks due to insufficient training, necessitating the supplementation of domain

knowledge through technologies like retrieval-augmented generation (RAG). From an ethical risk perspective, LLMs may generate misleading content such as fake news and forged contracts, and there are potential risks of misuse of user privacy data. For instance, a financial institution once leaked tens of thousands of user consultation records due to vulnerabilities in its AI customer service system. The issue of computational costs is particularly prominent. Taking GPT-3 as an example, its single training cost amounts to $4.6 million, and the training process consumes approximately 2840 MWh of computing power. The high resource and environmental costs severely constrain the technological accessibility for small and medium-sized enterprises.

5. Frontier Directions and Future Prospects

5.1. Technology Integration

Generative models based on the Transformer architecture achieve dynamic calibration and value alignment of generated content through the introduction of Reinforcement Learning from Human Feedback (RLHF). For example, in customer service dialogue scenarios, the model adjusts its response strategies by collecting real-time user satisfaction ratings, ensuring that the output aligns more closely with ethical norms and user expectations. Meanwhile, multimodal affective computing technology integrates vocal tone, facial expressions, and physiological signals to construct emotional state recognition models. In psychological counseling applications, the accuracy of emotional recognition has improved from 72% with single-modality to 89% with multimodality, significantly enhancing empathetic capabilities in interactive scenarios.

5.2. Efficiency Optimization

To address the high computational complexity of the Transformer's self-attention mechanism, sparse attention technology enhances model inference speed by a factor of 3 and reduces memory consumption by 40% through dynamically masking redundant associations. Meanwhile, distributed training frameworks support the parallel training of models with hundreds of billions of parameters, maximizing resource utilization in GPU clusters through tensor slicing and pipeline parallelism strategies, thereby reducing training cycles by 60%.

5.3. Ethical Governance

Through explainability enhancement techniques (such as attention heatmap visualization), the basis for model decision-making is revealed, for example, in medical diagnosis scenarios, the medical literature support chain behind generated conclusions can be traced. The federated learning framework, on the other hand, enables privacy protection in multi-institutional data collaboration (such as cross-hospital medical record analysis) through localized model training and encrypted parameter aggregation (such as differential privacy mechanisms), reducing the risk of data leakage by 90%.

6. Conclusion

Large language models construct a text interaction paradigm shifting from "mechanical responses" to "intelligent co-creation" through semantic parsing capabilities in natural language processing, global dependency modeling in the Transformer architecture, dynamic strategy optimization in reinforcement learning, and empathetic interaction design in affective computing. Research focuses include optimizing the Transformer architecture, value alignment mechanisms in reinforcement learning, efficiency improvements in multimodal fusion, and ethical governance frameworks for privacy protection and bias detection. Through theoretical analysis and case validation, this research provides a systematic approach for the sustainable development of large language models.

The research summarizes dual advancements of large language models in technological breakthroughs and social applications, while also pointing out three major bottlenecks: generalization

ability, cost-effectiveness, and content controllability. Future efforts should focus on continuous improvement in technology optimization, governance innovation, and value orientation. This includes enhancing model efficiency and mobile deployment capabilities through knowledge distillation, quantization compression, and multimodal fusion, while also constructing interdisciplinary regulatory frameworks to strengthen privacy protection and content traceability mechanisms. Additionally, it is essential to clarify the auxiliary role of technology to develop fairness detection tools and eliminate social biases. Only by finding a dynamic balance between technological efficiency and ethical constraints can large language models truly empower social values such as educational equity and cultural preservation, avoid the risks of technological abuse, and become accelerators for the sustainable development of human civilization.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, ... & D. Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. (2020).

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, & D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*. (2022).

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, & R. Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. (2022).

[4] L. Reynolds, & K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv* preprint *arXiv:2112.07870*. (2021).

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

[6] A. Radford, K. Narasimhan, T. Salimans, & I. Sutskever. *Improving Language Understanding by Generative Pre-Training.* (2018)

[7] D. Chen, & C. D. Manning. *A Fast and Accurate Dependency Parser using Neural Networks.* (2014).

[8] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, & C. Dyer. *Neural Architectures for Named Entity Recognition.* (2016).

[9] S. Mascarenhas, J. Dias, R. Prada, & A. Paiva. *A dimensional model for cultural emotion expression in virtual agents.* Journal on Multimodal User Interfaces, 3(1-2), 29-38. (2010).

[10] R. Aylett, J. Dias, & A. Paiva. *An affectively driven planner for synthetic characters.* IEEE Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS), 2(1), 2-10. (2006).

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, & I. Polosukhin. *Attention is* All *You Need.* Advances in Neural Information Processing Systems, 30, 5998–6008. (2017).

[12] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, & R. Salakhutdinov. *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context.* In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2978–2988. (2019).

[13] N. Kitaev, Ł. Kaiser, & A. Levskaya. *Reformer: The Efficient Transformer.* In Proceedings of the 37th International Conference on Machine Learning (ICML), 7001–7011. (2020).

[14] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, & D. Hassabis. *Human-level control through deep reinforcement learning.* Nature, 518(7540), 529–533.(2015).

[15] V. Mnih, A. P. Badia,, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, & K. Kavukcuoglu. *Asynchronous* Methods *for Deep Reinforcement Learning.* In International Conference on Machine Learning (pp. 1928–1937). (2016).

[16] T. Haarnoja, A. Zhou,, P. Abbeel, & S. Levine.*Soft Actor-Critic: Off-Policy Maximum Entropy Deep* Reinforcement *Learning with a Stochastic Actor.* In International Conference on Machine Learning (pp. 1861–1870). (2018).