Transformer and Its Application and Advances in Text Generation

Tong Wu[1, *] and Xiangzhe Xu[2]

[1] School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an, 710061, China
[2] College of Automation, Hangzhou Dianzi University, Hangzhou, 310018, China
atharp1@my.apsu.edu

Abstract. The Transformer model holds a pivotal position in NLP (natural language processing), demonstrating significant advantages in text generation tasks. This paper explores its applications and recent advances across domains. First, multimodal and syntax-controlled text generation approaches are introduced. These include the VCT (Vision-enhanced and Consensus-aware Transformer), integrating visual information and consensus knowledge, and the syntax-guided model GuiG, both enhancing semantic and syntactic quality through cross-modal interactions and dynamic attention mechanisms. Second, a comparative analysis of LLMs (large language models) like BART (Bidirectional and Auto-Regressive Transformer), T5 (Text-To-Text Transfer Transformer), and PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) validates the Transformer's superiority in handling long-term dependencies and consistency for automatic summarization. Domain-specific applications are further examined. Biomedical text generation models (e.g., BioGPT) improve accuracy via domain adaptation, while task-oriented dialogue systems (e.g., MegaT) optimize practicality through hybrid architectures. Addressing security challenges, a hybrid detection framework combining bidirectional LSTM (long short-term memory) and attention mechanisms achieves high-precision classification of AI-generated text. This study systematically reviews the technological evolution and boundaries of Transformer models, offering insights for advancing text generation technologies, refining domain-specific designs, and mitigating security risks posed by generative content.

Keywords: Transformer model, text generation, multimodal approaches, domain-specific applications.

1 Introduction

In the area of NLP, Transformer, as a new deep learning architecture, has been widely used in a variety of tasks, including text classification, text summarization, machine translation, text generation, etc. Compared with traditional sequential processing models, Transformer analyzes the entire sentence at once by using parallel computation, instead of the previous word order analysis. Additionally, Transformer

J. Iqbal (ed.), *Proceedings of the 2025 2nd International Conference on Mechanics, Electronics Engineering and Automation (ICMEEA 2025)*, Advances in Engineering Research 272,
https://doi.org/10.2991/978-94-6463-821-9_44

analyzes the relationship between words and sentences through the self-attention mechanism to help understand the meaning of sentences, further improving the efficiency of processing text data. To highlight the significance of this study, the following contributions are outlined. This article explores how the Transformer model can be applied and optimized in text generation to enhance semantic coherence and syntactic accuracy. This work explores how the Transformer model can be used for specific applications in some areas of progress, such as Medical and Biomedical. This study looks to the future of AI-generated text and delves into some of the potential challenges.

This research paper provides a comprehensive examination of Transformer and its application and development in text generation in the second part, we mainly propose Transformer-based models as well as text generation methods, which are VCT and GuiG, respectively. In addition, we analyze the effectiveness of various LLMs for text summarization in different scenarios, including BART, T5, PEGASUS, GPT (Generative Pre-Trained Transformers) and BERTSum (Bidirectional Encoder Representations from Transformers for Summarization), and test the functions of related Transformer models on automatic summarization. In the third part, we explore some applications of Transformer to Medical and Biomedical Text Generation, looking for task-oriented conversational and hybrid architectures, MegaT and BERT-GPT-4. In the fourth part, we believe that Hybrid Transformer would be used to detect AI-generated text in LLMs with high accuracy, while GPT-2 and LSTM networks being used for NLP. At the end of this paper, the conclusions about Transformer and its application and development in text generation are given.

2 Fundamental Applications and Optimization of Transformer Models in Text Generation

2.1 Multimodal and Syntax-Controlled Text Generation

A VCT that creates image descriptions by combining consensus knowledge with visual information. In this paper, VCT, which employs external knowledge for word inference based on self-attention, is suggested to use three essential elements to capitalize on both visual information and consensus knowledge for image captioning: a consensus-aware knowledge representation generator, a consensus-aware decoder, and an encoder with vision enhancement. A vision-enhanced encoder is proposed to model visual relationships, comprising a memory-based attention module and a visual perception module. The memory-based attention module allows for a thorough exploration of the interaction between regions and an image. The visual perception module then uses human-like information capture to gain the visual enhanced features. One graph convolutional network is made to extract the consensus knowledge from sentence corpora in order to take use of it.

Graph convolution, word embedding, and idea selection are all steps in the consensus knowledge acquisition process, as seen in Fig. 1. A word correlation graph is created using the co-occurrence relation between concepts. For picture captioning,

researchers suggest a VCT that performs cross-modal interactions. The comprehensive experimental findings on the Flickr30k and MSCOCO datasets show that the suggested model performs at the cutting edge. Additionally, in the online leaderboard of the MSCOCO test server, VCT can perform better than other models from published studies

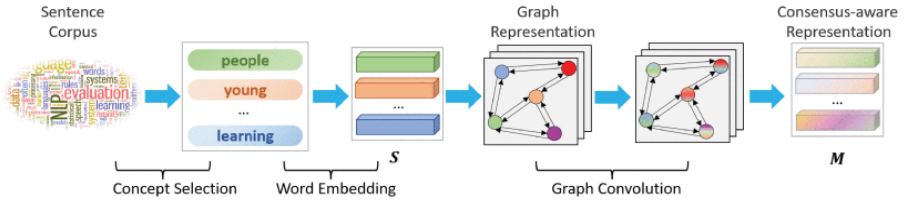

Fig. 1. LSTM schematic diagram [1].

GuiG Syntax-Controlled Text Generation with Multi-Encoder Attention and Path-Aware Transformers. The proposed approach presents GuiG, a Transformer-based method for syntactically guided text generation that leverages explicit constituency parse trees to enhance both semantic fidelity and syntactic alignment [2]. A template parse tree that illustrates some of the upper tiers of a complete tree is provided by the authors. Their pipeline is shown in Fig. 2. Among the primary innovations are: (1) a multi-encoder attention technique that successfully handles the fusion of variable-length inputs by allowing the dynamic integration of information from syntax and text encoders; (2) a path attention mechanism that restricts node attention to ancestral and descendant nodes within syntax trees, prioritizing high-level syntactic patterns over superficial token mappings; and (3) a node-level tree linearization format, which compresses parse trees into compact node-level sequences, reducing sequence length by approximately 67% compared to traditional bracketed formats and improving computational efficiency. Evaluated on controlled paraphrasing tasks, GuiG achieves a BLEU score of 48.03 (vs. SCPN's 23.23) when guided by full target parse trees. Its semantic quality improves by 122.1%, and TED (tree edit distance) decreases to 6.23 (vs. 6.55 for SCPN). When using partial template trees, GuiG attains a BLEU score of 26.27, surpassing SCPN (11.83) by 122%. Human evaluations confirm significant enhancements in semantic and syntactic quality, with average scores improving by 1.13 and 0.62 points (5-point scale), respectively. These results demonstrate GuiG's effectiveness in balancing syntactic control with semantic preservation, offering a robust framework for applications like style transfer and data augmentation.

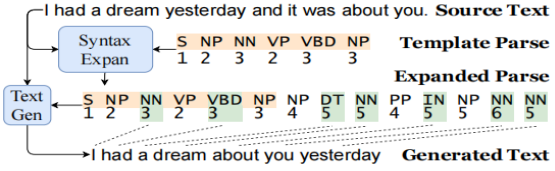

Fig. 2. The syntactically directed paraphrasing process pipeline to produce target text that is semantically consistent with the source text, the template parse—which is the uppermost tier of a parse tree—is extended into a complete tree [2].

2.2 Comparative Analysis of LLMs and Automatic Summarization

Various LLMs for text summarization. Researchers reported their work analyzes the performance of various LLMs, including BART, T5, PEGASUS, GPT, and BERTSum. Text summarization mainly falls into two categories: extractive summarization and abstractive summarization. So the research paper investigated various LLMs for text summarization, analyzing their effectiveness across different scenarios. Both model performance tables and human assessments revealed superior performance from BART, PEGASUS, and T5.

According to this comparison, BART and T5 utilize an encoder-decoder framework. PEGASUS emphasizes gap-sentence reconstruction during pre-training. Conversely, GPT uses a decoder-only transformer, while BERTsum employs an encoder-decoder setup with a BERT encoder. In terms of speed, GPT is the fastest, followed by T5, BART, and PEGASUS, while BERTsum's speed varies from time to time. T5 and PEGASUS show superior fine-tuning performance across tasks, with BART and GPT also performing well. BERTsum stands out in extractive summarization. Despite differences, all models exhibit strong transfer learning capabilities.

This paper conducts a comparative study of LLMs and transformers for text summarization in NLP. Additionally, based on various characteristics such as model architecture, pre-training objective, inference speed, fine-tuning performance, transfer learning capabilities, and accessibility, this work compares various models using ROUGE scores, key characteristics, and strengths and weaknesses of the models. Besides, it provides insights into LLM models effectiveness and guides future NLP research for text summarization. The paper discusses challenges in comparing LLMs and outlines future directions to enhance their capabilities, interpretability, and efficiency for text summarization tasks. The research concludes with a final comparison of transformer-based LLM models for text summarization (Table 1).

Table 1. Comparing LLM models using keycharacteristics [3]

Key Characteristics	LLMs				
	BART	T5	PEGASUS	GPT	BERTsum
Model Architecture	Encoder-decoder with bidirectional attention	Encoder-decoder with masked self-attention	Encoder-decoder with bidirectional attention	Decoder-only transformer with self-attention	Encoder-decoder with BERT encoder
Pre-training Objective	De-noising auto-encoding for text	Text-to-text format transformations	Gap-sentence reconstruction	Masked language modeling	Extractive summarization with BERT
Pre-trained on Specific Task	No	No	Can be fine-tuned for	Yes (language modeling)	Yes (summarization)

			specific domains		
Inference Speed	Faster than GPT	Faster than BART	Faster than GPT	Fastest	Depends on implementation
Fine-Tuning Performance	Good performance on various tasks	Very good performance on various tasks	Excellent performance on summarization	Good performance on text generation	Good performance on summarization
Transfer Learning Capabilities	Strong	Strong	Strong	Strong	Moderate
Availability and Accessibility	Open-source	Open-source	Open-source	Limited open-source availability	Open-source implementations
Recent Advancement	BART-LARGE model with improved performance	Focus on making T5 more efficient	New variants with improved summarization quality	Focus on explainability and reducing biases	New approaches for abstractive summarization

Transformer model, an automatic summarization system that integrates deep learning models. This paper aims to explore and practice how deep learning technology can play a role in the challenging task of automatic summary generation of English reading materials, and to build a comprehensive and systematic performance evaluation system. This paper constructs an automatic summarization system of integrated deep learning models. To guarantee the caliber and variety of training data, we first gathered and sanitized a sizable collection of English reading materials using a carefully thought-out data preprocessing procedure. Then, the performance of RNN, LSTM, integrated attention mechanism model and Transformer model on automatic summarization task is systematically compared. The Transformer-based model's supremacy in this task is validated by the experimental findings, which show that it has outstanding skills in terms of accuracy, consistency, and thorough coverage of the generated summaries.

Within this architectural framework, W and b denote the trainable parameters, whereas ft signifies the obsolescence portal. The particulars necessitating refreshment are gauged via the ingress aperture, culminating in the derivation of the cell's prevailing state, denoted as to pertinent to the current chronological juncture. This intricate interplay ensures a judicious amalgamation of novel inputs with residual knowledge, facilitating an efficacious progression through successive temporal intervals. By comprehensively comparing RNN, LSTM, Attention Mechanisms and Transformer models, the study reveals the significant advantages of Transformer in handling such tasks. The results clearly show that Transformer-based automatic summarization

systems not only surpass other models in handling complex language structures and long-term dependencies, but also reach new heights in the quality of the generated summary-including content accuracy, text coherence, and comprehensiveness of the information (Table 2).

Table 2. LSTM schematic diagram [4].

Model	ROUGE-1	ROUGE-2	ROUGE-L
RNN	29.86	25.31	22.62
LSTM	30.55	26.82	23.54
Transformer	31.42	27.6	25.07

3 Domain-Specific Applications of Transformer Models

3.1 Medical and Biomedical Text Generation

Transformer-Based Medical Text Generation for Clinical NLP. Augmenting Datasets and Enhancing Downstream Task Performance, the present work proposes a Transformer-based text generation framework for augmenting clinical NLP datasets, addressing data scarcity caused by privacy constraints [5]. Leveraging discharge summaries from MIMIC-III, conditioned on structured patient data, the authors construct high-resource (MimicText-98, ~98M words) and low-resource (MimicText-9, ~9M words) datasets to evaluate vanilla Transformer and GPT-2 models (e.g., demographics, diagnoses, medications). Key contributions include: (1) A novel seq2seq generation framework integrating clinical context, where the vanilla Transformer outperformed GPT-2 in high-resource settings (BLEU: 4.76 vs. 0.06). (2) Enhanced downstream task performance: Augmented data significantly improved unplanned readmission prediction using BioBERT, achieving AUC 0.669 (vs. 0.606 baseline, $p<0.05$) and F1 0.312 (vs. 0.242, $p<0.1$), while GPT-2 showed robustness in low-resource scenarios (AUC: 0.586 vs. Transformer's 0.527). (3) Model comparison revealing domain-specific advantages: The encoder-decoder Transformer excelled in high-resource hierarchical generation; however, GPT-2's pretraining showed advantages in low-resource adaptability. BioBERT, pretrained on biomedical corpora, consistently surpassed BERT (F1 gains ~30%), emphasizing domain adaptation. Results demonstrate that synthetic data introduces beneficial noise for generalization in semantic-rich tasks (e.g., readmission prediction) but faces challenges in accuracy-sensitive tasks such as phenotype classification (best AUC: 0.774). The study releases open-source generation models and interfaces, offering a scalable solution for medical NLP data augmentation. It also highlights the critical role of domain-specific pretraining and task-aligned evaluation.

BioGPT: A Domain-Specific Generative Pre-trained Transformer for Biomedical Text Generation and Mining with Natural Language Task Adaptation. The present work presents BioGPT, a domain-specific generative pre-trained Transformer model for biomedical text generation and mining, addressing the limitations of existing

discriminative models (e.g., BERT variants) in biomedical generation tasks [6]. The authors employ soft prompts in prefix-tuning, inserting learnable virtual tokens between the source and target sequences to guide the model in a "source-prompt-target" structure during both training and inference, as shown in Fig. 3.

The core innovations include: (1) the introduction of the first GPT-style model tailored for biomedicine. It is pre-trained from scratch on 15 million PubMed abstracts to mitigate domain shift; (2) a novel natural language-based target sequence design (e.g., "rel-is" format) combined with soft prompt strategies, enhancing task adaptation efficiency. Evaluations across six biomedical NLP tasks demonstrate state-of-the-art performance. For end-to-end relation extraction, BioGPT achieves F1 scores of 44.98%, 38.42%, and 40.76% on BC5CDR, KD-DTI, and DDI datasets, outperforming REBEL by 8.28%, 8.03%, and 12.49%, respectively. In PubMedQA question answering, it attains 78.2% accuracy, surpassing BioLinkBERT by 6.0%. BioGPT also exhibits superior text generation capabilities, accurately linking biomedical terms like 'hydroxychloroquine' to specific treatments while outperforming GPT-2 in fluency and domain relevance. Ablation studies confirm the effectiveness of natural language target formats (2.1–2.6% F1 gains over structured formats) and soft prompts. By integrating domain-specific pre-training and task-adaptive design, BioGPT establishes a robust framework for biomedical text mining and generation, achieving significant advancements across multiple benchmarks.

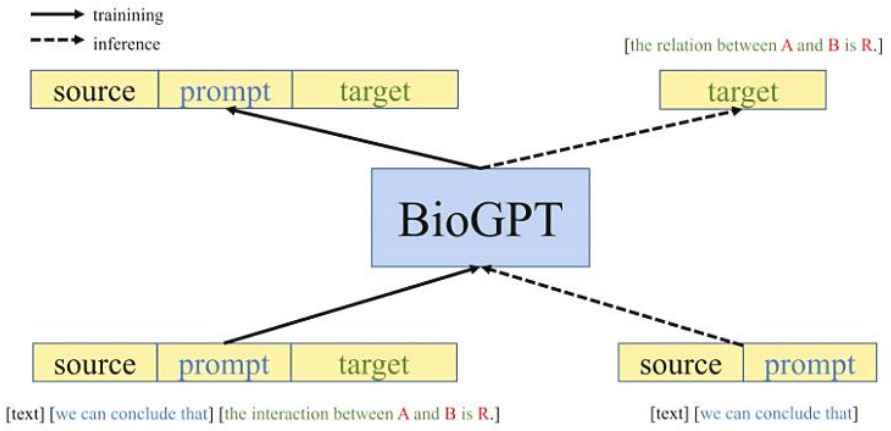

Fig. 3. Framework of BioGPT when adapting to downstream tasks [6].

3.2 Task-Oriented Dialogue and Hybrid Architectures

Using MegaT, based on the language processing model of the Transformer, a task-oriented dialogue is constructed. We suggest a novel language processing model called MegaT, which is an adaption of the LongT5 (Long Text-ToText-Transfer Transformer) architecture. According to experimental data, MegaT performs comparably better than the T5-based agent in handling longer text sequences and creating more interesting and informative responses, as well as in giving correct, fluent, and coherent answers to user questions [7]. The reduction of the Belief Loss and

Response Loss by at least 50 points overall is the most important improvement factor. By using the LongT5 model to generate high-quality task-oriented dialogues, this study offers insights into how to create conversational bots that are more successful. The pipeline technique and the end-to-end approach are the two basic TOD system architectures that have been put forth.

MegaT is very good at processing long input sequences with up to 16,000 tokens. To speed up calculation and solve the problems caused by lengthy input sequences in task-oriented dialog systems, it integrates the Transient Global Attention method. To sum up, this study concentrated on the creation of task-oriented dialogue systems, specifically examining how well two models—T5 and MegaT—performed on MultiWoz datasets. MegaT is superior at generating task-oriented talks, as evidenced by the fact that it regularly surpassed T5 in a variety of evaluation metrics. In addition, the study identified several key areas for future research in the field of task-oriented dialogue systems. These include improving multi-turn and contextual awareness, utilizing transfer learning techniques, investigating reinforcement learning for agent enhancement, optimizing Transformer designs especially for dialogue systems, and tailoring interactions according to user preferences.

BERT-GPT-4: Bidirectional Semantic Fusion with Dynamic Weighting for Coherent and Contextually Rich Text Generation. Present work introduces a novel hybrid model (BERT-GPT-4) that integrates BERT's bidirectional semantic encoding with GPT-4's autoregressive generation to enhance text coherence and semantic accuracy [8]. The core innovation lies in leveraging BERT to extract deep contextual representations from input text, which serve as semantic priors for GPT-4's generation process. A sigmoid-regulated dynamic weighting mechanism optimizes the influence of BERT-derived semantics, balancing semantic consistency and output diversity.

Evaluated on the OpenAI GPT-3 dataset, the model outperforms state-of-the-art baselines (GPT-3, T5, BART, Transformer-XL, CTRL) with a perplexity of **15.8** (10.2% lower than CTRL) and a BLEU score of **29.6** (8.4% higher than CTRL), demonstrating superior fluency and semantic alignment with human language. Ablation studies confirm the efficacy of the dynamic weighting strategy in maintaining logical coherence while enabling flexible generation, particularly in complex semantic scenarios. This work establishes a robust framework for encoder-generator integration, offering significant advantages in long-text generation and contextually dense tasks. Future directions include optimizing computational efficiency and extending the architecture to multimodal inputs. These advancements pave the way for applications in dialogue systems, automated content creation, and personalized human-computer interaction.

4 Detection of AI-Generated Text and Challenges

4.1 Hybrid Transformer-Based Framework for High-Accuracy Detection of AI-Generated Text in LLMs

The present work proposes a Transformer-based deep learning framework for detecting AI-generated text from LLMs, integrating bidirectional LSTM, multi-head attention

mechanisms, and convolutional layers to achieve high-precision classification [9]. The core innovations include: (1) The framework includes a systematic text preprocessing pipeline, featuring Unicode normalization, lowercase conversion, non-alphabetic character removal, and punctuation standardization, all aimed at enhancing data consistency; (2) a hybrid neural architecture combining bidirectional LSTM for sequential feature extraction, Transformer blocks with multi-head attention to capture global dependencies, and 1D convolutional layers for local pattern recognition, augmented by residual connections to mitigate gradient vanishing. Trained on a balanced dataset of 1,378 human-written (labeled 0) and AI-generated (labeled 1) texts using the Adam optimizer and binary cross-entropy loss, the model demonstrates significant optimization: training loss decreases from 0.127 to 0.005, while validation accuracy improves from 94.96% to 99.8% over four epochs. Evaluation on the test set achieves 99% overall accuracy, with precision (0.99), recall (1.00), and F1-score (0.99), demonstrating robust discriminative performance. These results validate the efficacy of hybrid architectures in AI text detection, establishing a high-accuracy benchmark for future research. The framework's potential extends to practical applications in misinformation mitigation and content authenticity verification, offering a scalable solution to address challenges posed by LLM-generated content proliferation.

4.2 Utilizes two advanced NLP technologies: GPT-2 and LSTM networks

This study introduces an ensemble method leveraging two sophisticated NLP technologies: the GPT-2 and LSTM networks. The research began with acquiring an extensive, anonymized dataset from Circana, incorporating diverse real-world consumer data, including purchases, promotions, and e-commerce transactions across various retail sectors. This research aims to address these challenges through advanced machine learning techniques. It analyzes email subject lines using Generative AI GPT 2 model for generating text and then passes the generated text through an LSTM model for receipt identification [10]. The research is comprehensive, beginning with the extensive collection of E-mail messages, followed by the extraction of subject line and pre-processing to prepare the data for in-depth analysis. First, it illustrates the effectiveness of an ensemble GPT-2 and LSTM model in the domain of consumer communication analysis, proving its capability in accurately classifying and identifying emails. Second, it sheds light on the practical implications of our findings, emphasizing how this technological approach could transform email management systems through the automated recognition and sorting of transactional receipts. The results of this study suggest that while the combined use of GPT-2 and LSTM enhances the detection of transactional receipts, it also leads to an increased occurrence of false positives and higher computational demands.

5 Conclusion

The Transformer model has shown remarkable potential and broad applicability in text generation tasks within NLP. This paper systematically investigates its key

technological advancements. First, multimodal and syntax-controlled generation models, such as VCT and GuiG, enhance semantic coherence and syntactic accuracy. They leverage cross-modal interaction and dynamic attention mechanisms, achieving state-of-the-art results in tasks like image captioning and controllable paraphrase generation. Second, comparative analyses of LLMs (e.g., BART, T5, and PEGASUS) in automatic summarization reveal the Transformer architecture's superior capability in handling long-range dependencies and maintaining generation consistency, while highlighting distinct characteristics across model architectures in terms of speed, adaptability, and task-specific performance. In domain-specific applications, biomedical generation models like BioGPT validate the necessity of domain-adapted architectures through pretraining and task-specific designs, improving generation accuracy and practicality. Task-oriented dialogue systems (e.g., MegaT) advance human-computer interaction naturalness and efficiency through enhanced context processing and hybrid architecture optimization. Hybrid frameworks combining bidirectional LSTM and Transformer effectively address AI-generated text detection, achieving high-precision classification for content security.

Future technological development should focus on five directions: (1) Optimizing multimodal fusion mechanisms to explore deep collaboration among visual, auditory, and other heterogeneous information sources; (2) Enhancing model interpretability while balancing generation diversity and controllability; (3) Moreover, reducing computational resource consumption in training and inference is essential to enable lightweight and real-time applications; (4) Strengthening cross-domain transfer learning capabilities to improve generalization in vertical scenarios such as healthcare and education; (5) Addressing security and ethical challenges of generated content through robust detection technologies and governance frameworks. Through these breakthroughs, Transformer models will continue to drive innovation in text generation technology, laying the foundation for more practical and trustworthy artificial intelligence systems.

Acknowledgments (Author contribution). All the authors contributed equally and their names were listed in alphabetical order.

References

1. Cao, S., An, G., Zheng, Z., Wang, Z.: 'Vision-Enhanced and Consensus-Aware Transformer for Image Captioning', IEEE Trans. Circuits Syst. Video Technol., **32**(10), 7005-7018 (2022)
2. Li, Y., Feng, R., Rehg, I., Zhang, C.: 'Transformer-based neural text generation with syntactic guidance', arXiv:2010.01737 (2020)
3. Kotkar, A.D., Mahadik, R.S., More, P.G., Thorat, S.A.: 'Comparative Analysis of Transformer-based Large Language Models (LLMs) for Text Summarization', Proc. Int. Conf. Advanced Computing and Emerging Technologies, Ghaziabad, India, 1-7 (2024)
4. Liu, Y.: 'Application and evaluation system of deep learning in automatic summary generation of English reading materials', Proc. Int. Conf. Computers, Information Processing and Advanced Education, Ottawa, Canada, 293-297 (2024)

5. Amin-Nejad, A., Ive, J., Velupillai, S.: 'Exploring transformer text generation for medical dataset augmentation', Proc. Int. Conf. Language Resources and Evaluation, Marseille, France, 4699-4708 (2020)
6. Luo, R., Sun, L., Ya, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: 'BioGPT: generative pre-trained transformer for biomedical text generation and mining', Brief. Bioinform., **23**(6), 1-9 (2022)
7. Petouo, F., Arafat, Y., Mushabbir, M., Hasan, K., Mahmud, H.: 'Dialog Generation with Conversational Agent in Task-Oriented Context using a Transformer Architecture', Proc. Int. Conf. Computer and Information Technology, Cox's Bazar, Bangladesh, 1-6 (2023)
8. Chen, J., Wang, S., Qi, Z., Zhang, Z., Wang, C., Zheng, H.: 'A Combined Encoder and Transformer Approach for Coherent and High-Quality Text Generation', arXiv:2411.12157, (2024)
9. Mo, Y., Qin, H., Dong, Y., Zhu, Z., Li, Z.: 'Large language model (llm) ai text generation detection based on transformer deep learning algorithm', arXiv:2405.06652 (2024)
10. Hirway, C., Fallon, E., Connolly, P., Flanagan, K.: 'An Ensemble Method to Enhance Receipt Identification in Consumer Communication Using Generative AI and LSTM', Proc. Int. Conf. Information Technology and Computing, Yogyakarta, Indonesia, 18-23 (2024)