# Comparative Performance Analysis of Neural Architectures for Poem Generation

Bharathi Mohan G.
*Department of Computer Science and Engineering*
*Amrita School of Computing*
Amrita Viswa Vidyapeetham
Chennai, India
g_bharathimohan@ch.amrita.edu

Prasanna Kumar R.
*Department of Computer Science and Engineering*
*Amrita School of Computing*
Amrita Viswa Vidyapeetham
Chennai, India
r_prasannakumar@ch.amrita.edu

Shabarirajan KJ
*Department of Computer Science and Engineering*
*Amrita School of Computing*
Amrita Viswa Vidyapeetam
Chennai, India
shabarirajan77@gmail.com

Logeshwar B S
*Department of Computer Science and Engineering*
*Amrita School of Computing*
Amrita Viswa Vidyapeetham
Chennai, India
logeshlogesh1272004@gmail.com@gmail.com

Aadhithyan D
*Department of Computer Science and Engineering*
*Amrita School of Computing*
Amrita Viswa Vidyapeetham
Chennai, India
aadhithyan.d@gmail.com

Elakkiya R
*Department of Computer Science*
*BITS Pilani*
Dubai Campus
Dubai, United Arab Emirates
elakkiyaceg@gmail.com

***Abstract*—Recent advancements in artificial intelligence have significantly impacted various sectors, including the creative domain of poetry generation. This paper examines the effectiveness of neural architectures, specifically Bi-directional Long Short-Term Memory (Bi-LSTM), Vicuna-1B, and GPT-2, in generating poetry. By initiating these models with detailed prompts encompassing genre and theme, we aim to explore their capacity for producing poetry that resonates with human emotion and stylistic precision. Our methodology involves comparing the poems generated by these models against a curated dataset of existing poems, employing BLEU and ROUGE scores for a comprehensive evaluation. This study seeks to shed light on the potential of each model in the field of computational creativity, offering insights into their applicability in poetry generation.**

***Index Terms*—Text generation, vicuna-1b, Bi-lstm,Gpt-2, Rogue score**

## I. Introduction

The exploration of how artificial intelligence can be blended with artistic expression, particularly poetry, is not only fascinating but increasingly relevant in today's digital age. With advancements in neural network architectures like Bi-LSTM, Vicuna-1B, and BERT, we are poised to explore new horizons in creative writing. Bi-LSTM, for example, enhances traditional LSTM networks by interpreting sequences from both ends, potentially offering a deeper understanding of poetic context and structure. Vicuna-1B, designed specifically for creative applications, may offer unique insights into genre-specific language patterns, while BERT's transformative approach to contextual understanding could redefine syntactic and semantic nuances in poetry.

In this study, we employ a rigorous methodological framework to dissect how these models interact with and interpret poetic prompts, aiming to generate works that not only mimic human poetic flair but also push the boundaries of what can be considered 'creative' from an AI perspective. We will conduct experiments that not only apply traditional metrics like BLEU and ROUGE for evaluating textual similarity but also incorporate newer, more nuanced measures that assess creativity and emotional resonance.This facet of the study aims to highlight the potential of AI in adhering to traditional poetic forms while possibly creating new ones. Through these insights, we hope to enrich the dialogue between technology and the arts, underscoring the transformative potential of AI in expanding the tools available to human creativity. This investigation thus serves as a cornerstone for future studies in computational creativity and its implications for the intersection of technology and human artistic endeavor.

## II. Related Works

[1] A detailed survey on how topic modeling techniques, especially Latent Dirichlet Allocation (LDA), are applied in text summarization. They focus on multi-document summarization, showing how LDA improves summary quality by identifying key topics, thereby enhancing coverage and reducing redundancy. Other topic modeling methods like Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), and lda2vec are discussed, but LDA is highlighted for its effectiveness in capturing the thematic essence of texts. [2] The authors evaluate GPT-2 and BERT models for text summarization within big data analytics. Using ROUGE metrics on a dataset of 100 news articles, they compare these transformer-

based models' performance in multi-document summarization. The results show BERT outperforming GPT-2 in accuracy, making it a more suitable model for summarization tasks.

[3] The discussion centers around text generation in NLP, emphasizing Long Short-Term Memory (LSTM) networks over its variants like Gated Recurrent Unit (GRU) and peephole-connected LSTMs. Despite the rise of transformer models, LSTMs are shown to excel in capturing long-term dependencies, crucial for generating coherent texts. [4] A novel Deep Neural Network (DNN) model for sentiment analysis is proposed, addressing the challenges of analyzing sentiments from unstructured social media data. This hybrid model integrates sentiment lexicons, BERT for sentiment-enhanced word embeddings, BiLSTM for understanding text's sequential and contextual semantics, and an attention mechanism to focus on significant features. Additionally, CNNs are used to distill crucial local features.

In this paper [5], the authors explore the use of BERT-based language models for detecting misinformation, a critical issue in the age of social media. By fine-tuning BERT models on datasets related to rumors and fake news, the study shows these models surpass traditional machine learning methods in identifying misinformation. [6] The challenges in natural language generation (NLG), particularly due to limited training data and the quality of generated text, are addressed through a novel language model named EDA-BoB. This model uses text augmentation and a knowledge understanding mechanism to improve data quality and expand datasets without significant increases in computational resources.

[7] A new approach to generating classical Chinese poetry using a Transformer-XL model is introduced. This model, with a multi-head self-attention and segment-level recurrence mechanism, effectively captures the complex relationships among Chinese characters and long-term dependencies. An automatic evaluation model, including a BERT-based fluency checker and a tone-checker module, validates the poems' quality. [8] The Bairon system is presented for automatically generating poetry in the style of famous poets like Shakespeare and Cummings, based on user-provided initial lines. Using fine-tuned GPT-2 and T5 models, the system is evaluated through a Poetry Turing Test, demonstrating its ability to mimic poets' styles convincingly, especially Cummings.

[9] A novel approach for enhancing Vietnamese text generation by incorporating topic information into the neural network model is introduced. This leads to significant improvements in text relevance and coherence, with experimental results showing at least a 23 percent improvement in BLEU scores over baseline models. [10] A hybrid text summarization method is proposed, combining Latent Dirichlet Allocation (LDA) for extracting semantically rich sentences with transformer-based models for abstractive summary generation. This innovative approach demonstrates improvements in precision and F1 scores, effectively capturing the original documents' essence while covering a wider range of semantics.

# III. METHODOLOGY

## A. Dataset Preparation

The Poetry Foundation data is a dataset which is designed for training models to generate poems and for other language tasks [16]. It includes about 13,854 poems from the Poetry Foundation's website. The collection is diverse, covering different genres. Each poem comes with its title, the author's name, and tags that describe what it's about. The main goal of this dataset is to help improve how well models can create poetry and assist in other language-related projects by providing a wide range of poems.

## B. Models Under Study

The research primarily aimed to assess and improve how well two pre-trained models, Vicuna-1B and GPT-2, along with a bi-LSTM model specifically trained for creating poetry, performed in poetry generation tasks. These models were chosen because of their success in generating text and their potential to handle the unique challenges of poetry. Vicuna-1B and GPT-2 [11] have been widely used for creating text that's coherent and relevant to its context, which is why they were considered a good fit for this study.

1) GPT-2: GPT-2 is a tool made by OpenAI that's really good at writing text that sounds like it was written by a person. It uses a special method to figure out how to put words together in a way that makes sense and fits the topic you ask about. There are different sizes of GPT-2; the biggest ones can do more complex tasks. People use it for things like writing articles, making summaries, and other writing jobs.

TABLE I
HYPERPARAMETERS OF GPT-2 MODEL

| **Hyperparameters** | ***Value*** |
|---|---|
| Model type | gpt2 |
| Architecture | GPT2LMHeadModel |
| Attention Heads | 12 |
| Hidden layers | NULL |
| Layers | 12 |
| n_positions | 1024 |

There was a lot of talk about making sure it's used carefully. TableI gives details about the hyperparameters of the model. GPT-2 model's performance has to be improved by fine-tuning,since the poem generated by this base model is qualitatively and quantitatively got a low score.

2) Bi-LSTM Network: Our model architecture, structured as a Sequential model in Keras, comprises several key components: an Embedding layer converting input sequences into dense vectors of size 300, with a maximum sequence length of 99 words and 3,524,100 trainable parameters; a Bidirectional LSTM layer consisting of two LSTM layers processing input sequences in both forward and backward directions,

each with 150 units and a total of 541,200 trainable parameters; a Dropout layer with no units dropped out; a unidirectional LSTM layer with 100 units and 160,400 trainable parameters; and two Dense layers, the first with 5873 units and the second with 11,747 units, totaling 69,301,873 trainable parameters. This architecture encompasses a total of 73,820,751 trainable parameters, suggesting its potential for tasks such as sequence classification or generation due to the combination of bidirectional and unidirectional LSTM layers followed by dense layers for classification [17]. The following Figure 1 explains the Architecture and parameters used and Figure 2 represents loss and accuracy achieved by Bi-lstm when trained.

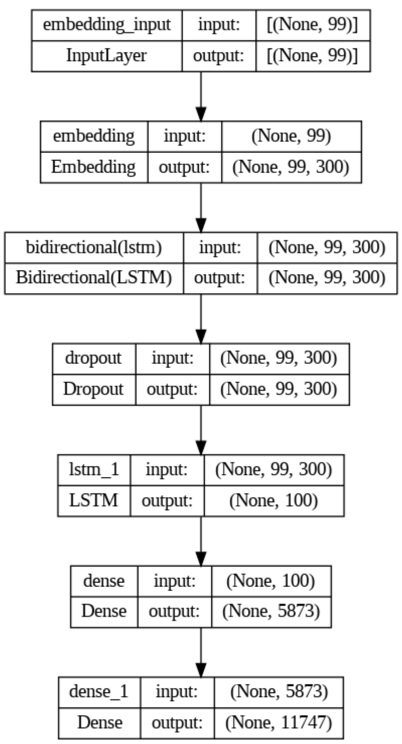


Fig. 1. Bi-lstm Network Architecture

3) Tiny-Vicuna-1B: This model represents an enhanced version that has been fine-tuned using the TinyLlama architecture [12], specifically trained on the WizardVicuna Dataset. It's designed to be fully compatible with the Vicuna-v1.5 series. This compatibility suggests that it inherits the core strengths of the Vicuna architecture, potentially offering improved performance or specialized capabilities in certain tasks, while also benefiting from the specific nuances and strengths of the TinyLlama model. The fine-tuning process likely aims to leverage the unique features of the WizardVicuna Dataset to refine the model's ability in particular areas of interest or application.Table II shows the hyperparameters of the model.

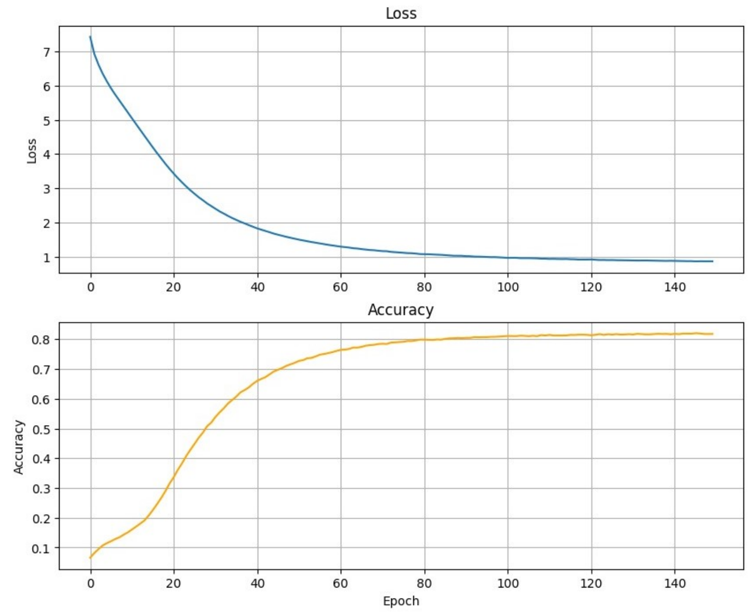


Fig. 2. Loss and Accuracy of Bi-lstm

TABLE II
HYPERPARAMETERS OF TINY-VICUNA-1B

| **Hyperparameters** | **Value** |
|---|---|
| Model type | Llama |
| Architecture | LlamaForCasualLM |
| Attention Heads | 32 |
| Hidden layers | 22 |
| Max position Embeddings | 2048 |
| Hidden Activation function | Silu |

This model needs only 700Mb of RAM, despite it's size it reply well to human language.When tasked with generating a poem and given a specific genre, the model was able to produce a sentence that was meaningful and relevant to the request. However, it fell short in terms of quality. This suggests that while the model can grasp the basic requirements of the task and align its output with the given genre, it might struggle with capturing the depth, emotional resonance, or artistic intricacies typically associated with high-quality poetry. This could be due to limitations in its training data, the model's inherent capabilities, or the complexity of accurately mimicking the nuanced art of poetry generation.

Figure 3 provides a methodology diagram for this paper

The fine-tuning of Vicuna-1B and GPT-2 was essential to adapt these models for poetry generation, enhancing their understanding of the stylistic and thematic elements specific to various poetry genres. [13] We implemented full parameter tuning, with the creation of genre-specific prompt templates

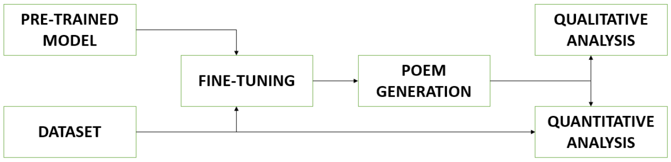


Fig. 3. Methodology

playing a pivotal role. These templates guided the models to more authentically replicate the characteristics of the desired genres, thereby improving the authenticity of the generated poems.

Our methodology utilized the Low-Rank Adaptation (LoRA) approach to refine the models' poetic capabilities. We carefully configured training parameters, including a per-device batch size of 2, using gradient accumulation to effectively manage hardware limitations and employing paged adam 32bit as the optimizer to balance memory efficiency and convergence speed. The training regimen also incorporated a constant learning rate scheduler with a warm-up phase to stabilize training dynamics. The use of a custom collator and the SFT-Trainer class facilitated the integration of LoRA, optimizing the models' ability to generate coherent, genre-specific poetry. This approach demonstrates a novel strategy to enhance the creative generative potential of language models in the domain of poetry.

### C. Evaluation Metrics

The evaluation framework was based on ROUGE scores for quantitative, industry-standard metrics for assessing the quality of text generation. These metrics allow for a quantitative comparison of the generated poems against a reference set, providing insights into the models' ability to replicate genre-specific stylistic and thematic elements.ROUGE scores emphasize recall, offering a comprehensive view of the models' performance.

1) Response Collection: The models were tasked with generating poems for each of the 13 genres, using the specially designed prompt templates. The responses were systematically recorded in an Excel sheet, for a structured comparison.
2) Comparative Analysis: The generated poems were then compared to the reference corpus for each genre within our dataset. This comparison, grounded in ROUGE scores, enabled a detailed assessment of how closely the models' outputs align with the genre-specific characteristics of the dataset. This rigorous analytical process was designed to quantify the effectiveness of the fine-tuning and prompt template strategies in enhancing the poetic quality of the generated content.But these are not recommended when you generate a novel poem, so having a qualitative analysis over quantitative is advisable.

## IV. RESULT

### A. Model Performance Overview

The refined Vicuna-1b model and the bi-lstm exhibited clearer results compared to a fine-tuned GPT-2 model. Despite having fewer parameters, GPT-2 struggled to perform adequately on this dataset even after fine-tuning. Figure 4, 5, and 6 display poems generated by each model, respectively. Through both quantitative and qualitative analysis, insights were gained on the reception of these poems by the general public and their potential future use.

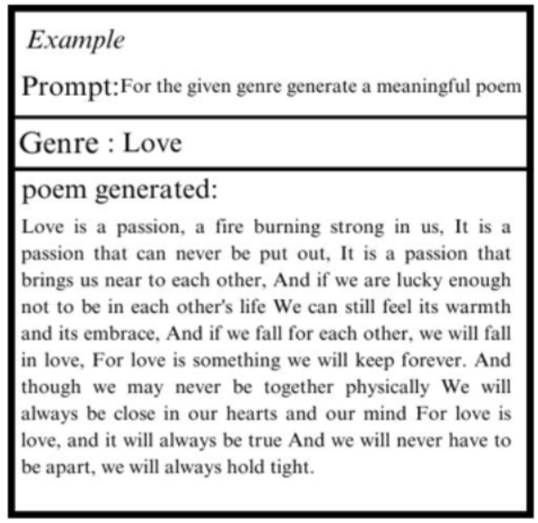


Fig. 4. Example result of Vicuna-1B

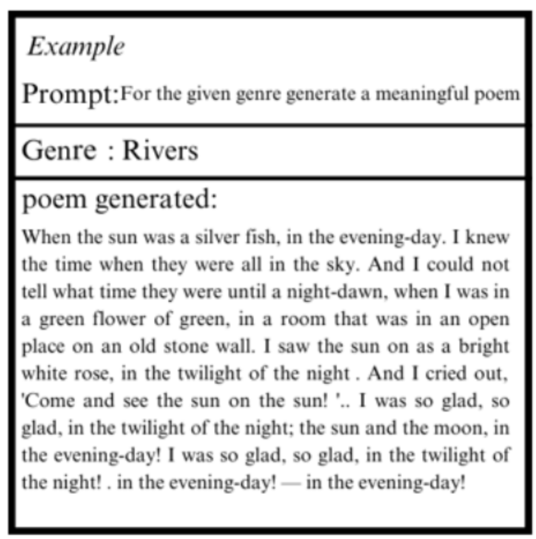


Fig. 5. Example result of GPT-2

### B. Quantitative Analysis

Quantitative analysis for poem generation involves using numerical methods, like ROUGE scores, to objectively evaluate the poems. ROUGE assesses the overlap of content between

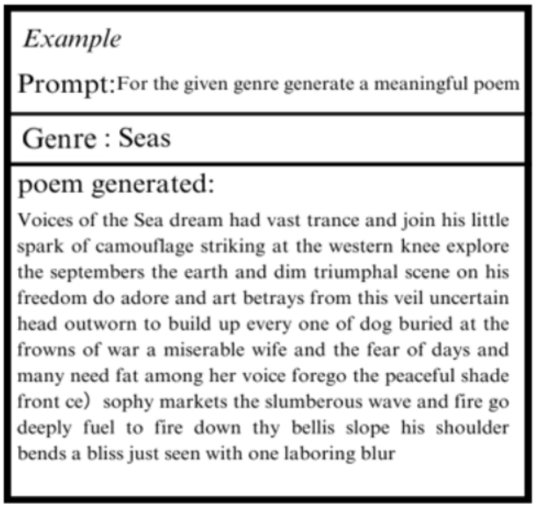


Fig. 6. Example result of Bi-LSTM

the generated and reference poems which has been explained [14], it is given by the formula:

$$ROUGE\text{-}N = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

This formula emphasizes the capture of essential themes. These metrics help compare models but have limits in assessing the artistic qualities of poetry, suggesting a blend of quantitative and qualitative evaluations for a fuller assessment. The tables below III give the max ROUGE scores. Some of the genres have been left out on purpose as we received zero scores for them. BLEU score of all three models was negligible, so it had been neglected as BLEU score cannot be used to determine their performance.

TABLE III
QUANTITATIVE EVALUATION OF FINE-TUNED VICUNA-1B, GPT-2 & BI-LSTM

| Genre | Vicuna-1B ROG Score | GPT-2 ROG Score | Bi-lstm Score |
|---|---|---|---|
| Nature | 0.125000 | 0.178344 | 0.141844 |
| Living | 0.163265 | 0.163934 | 0.197044 |
| Love | 0.175115 | 0.181818 | 0.102041 |
| Relationships | 0.164948 | - | 0.136364 |
| History & Politics | 0.125000 | - | 0.113402 |
| Religion | 0.112150 | 0.158730 | 0.115385 |
| Animals | 0.191617 | - | 0.191617 |
| Philosophy | 0.141593 | 0.158730 | 0.125561 |

### C. Qualitative Analysis

Thirty volunteers participated in the assessment of poetry generated by the three models. The survey was conducted online and it was purely anonymous in order to avoid bias for a particular type. The paper [15] conducts an analysis on the metrics and acquired great results. In this survey, they encountered a selection of poems displayed randomly, with two main requests:
a) They evaluated each poem's artistic merits, focusing on its readability, capacity to evoke emotions, and overall beauty.
b) They also guessed if each poem was written by a specific human poet or created by a computer program.
The responses of the online survey is given in the bar graph below in Figure 7, 8, 9.

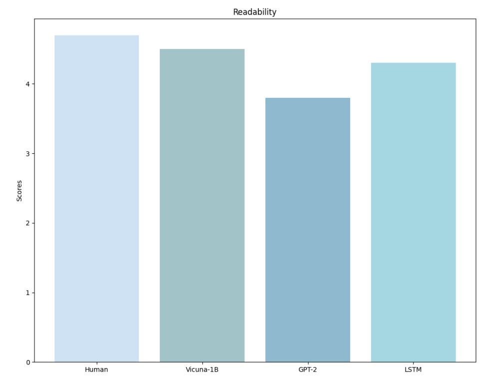


Fig. 7. Readability

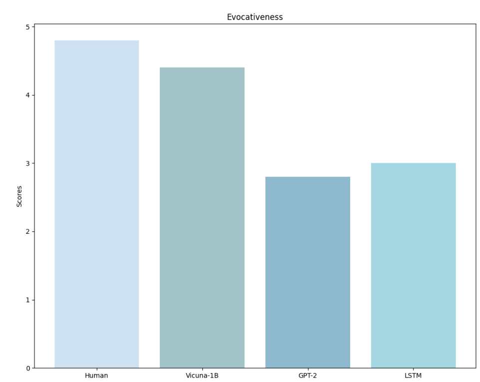


Fig. 8. Evocativeness

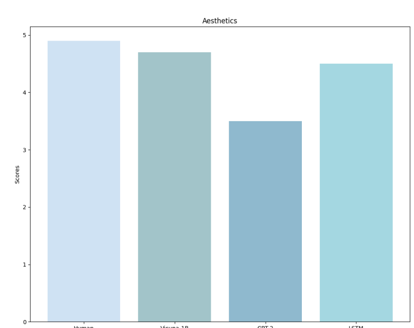


Fig. 9. Aesthetics

## V. CONCLUSION

This study investigates the transformative role of AI in creative writing, focusing on poetry generation. It critically assesses models like Vicuna-1B, GPT-2, and bi-LSTM, showing that with adequate fine-tuning and data, AI can produce poetry that rivals human efforts. The research employs a blend of quantitative metrics and qualitative feedback, underscoring the challenge of evaluating artistic quality through numerical scores alone. Notably, the difficulty participants experienced in distinguishing between AI-generated and human-created poems highlights the diminishing boundaries between human and machine creativity.

The findings of this research not only demonstrate the capabilities of AI in the arts but also provoke further discussion about the nature of creativity and the evolving relationship between technology and artistic expression. This work encourages a deeper exploration of how AI can augment and redefine creativity, suggesting a future where AI enhances the creative process, potentially leading to innovative collaborations in artistic fields. As we continue to integrate technology with traditional creative disciplines, it is essential to consider the implications and opportunities that arise from this convergence, setting a course for future research in the integration of AI with human creativity.

## REFERENCES


[1] Mohan, G.B., Kumar, R.P. (2022). A Comprehensive Survey on Topic Modeling in Text Summarization. In: Sharma, D.K., Peng, SL., Sharma, R., Zaitsev, D.A. (eds) Micro-Electronics and Telecommunication Engineering. ICMETE 2021. Lecture Notes in Networks and Systems, vol 373. Springer, Singapore. https://doi.org/10.1007/978-981-16-8721-1_22.

[2] Bharathi Mohan, G., Prasanna Kumar, R., Parathasarathy, S., Aravind, S., Hanish, K.B., Pavithria, G. (2023). Text Summarization for Big Data Analytics: A Comprehensive Review of GPT 2 and BERT Approaches. In: Sharma, R., Jeon, G., Zhang, Y. (eds) Data Analytics for Internet of Things Infrastructure. Internet of Things. Springer, Cham. https://doi.org/10.1007/978-3-031-33808-3_14.

[3] Li, Lifen, and Tianyu Zhang. 2021. “Research on Text Generation Based on LSTM.” In. https://api.semanticscholar.org/CorpusID:239938295.

[4] J. Khan, N. Ahmad, S. Khalid, F. Ali and Y. Lee, ”Sentiment and Context-Aware Hybrid DNN With Attention for Text Sentiment Classification,” in IEEE Access, vol. 11, pp. 28162-28179, 2023, doi: 10.1109/ACCESS.2023.3259107.

[5] R. Anggrainingsih, G. M. Hassan, and A. Datta, Evaluating BERT-based Pre-training Language Models for Detecting Misinformation. 2022. 10.21203/rs.3.rs-1608574/v1.

[6] L. Lei, Y. Sun, Y. Liu, R. Roxas, and R. Raga, “Research and Implementation of Text Generation Based on Text Augmentation and Knowledge Understanding,” Computational Intelligence and Neuroscience, vol. 2022, pp. 1–10, Sep. 2022, doi: 10.1155/2022/2988639.

[7] J. Zhao and H. J. Lee, “Automatic Generation and Evaluation of Chinese Classical Poetry with Attention-Based Deep Neural Network,” Applied Sciences, vol. 12, p. 6497, Jun. 2022, doi: 10.3390/app12136497.

[8] M. Badura, M. Lampert, and R. Dreżewski, “System Supporting Poetry Generation Using Text Generation and Style Transfer Methods,” Procedia Computer Science, vol. 207, pp. 3310–3319, 2022, doi: https://doi.org/10.1016/j.procs.2022.09.389.

[9] V. Hong and C. Le, “Topic-Guided RNN Model for Vietnamese Text Generation,” 2021, pp. 827–834. doi: 10.1007/978-981-15-7527-3_78.

[10] Bharathi Mohan Gurusamy, Prasanna Kumar Rangarajan, Partha Srinivasan. (2023). A hybrid approach for text summarization using semantic latent Dirichlet allocation and sentence concept mapping with transformer. International Journal of Electrical and Computer Engineering (IJECE). 13. 6663-6672. 10.11591/ijece.v13i6.pp6663-6672.

[11] Wei-Jen Ko, Junyi Jessy Li, “Assessing Discourse Relations in Language Generation from GPT-2”, 2020, doi:arXiv:2004.12506.

[12] Chiang, Wei-Lin and Li, Zhuohan and Lin, Zi and Sheng, Ying and Wu, Zhanghao and Zhang, Hao and Zheng, Lianmin and Zhuang, Siyuan and Zhuang, Yonghao and Gonzalez, Joseph E. and Stoica, Ion and Xing, Eric P.,Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, march, 2020, doi:https://lmsys.org/blog/2023-03-30-vicuna/

[13] Kai Lv, Yuqing Yang, Tengxiao Liu, Qinghui Gao, Qipeng Guo, Xipeng Qiu, ”Full Parameter Fine-tuning for Large Language Models with Limited Resources”, 16 jun 2023, doi:arXiv:2306.09782v1

[14] Max Grusky. 2023. Rogue Scores. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1914–1934, Toronto, Canada. Association for Computational Linguistics.

[15] Rodriguez Pascual, A. (2018). ”BACON: Deep-Learning Powered AI for Poetry Generation with Author Linguistic Style Transfer”. California Science and Engineering Fair, Los Angeles, CA, April 23-24, 2018. Campolindo High School.

[16] https://www.kaggle.com/datasets/tgdivy/poetry-foundation-poems

[17] J. Wei, Q. Zhou and Y. Cai, ”Poet-based Poetry Generation: Controlling Personal Style with Recurrent Neural Networks,” 2018 International Conference on Computing, Networking and Communications (ICNC), Maui, HI, USA, 2018, pp. 156-160, doi: 10.1109/ICCNC.2018.8390270

[18] N. Fatima, A. S. Imran, Z. Kastrati, S. M. Daudpota and A. Soomro, ”A Systematic Literature Review on Text Generation Using Deep Neural Network Models,” in IEEE Access, vol. 10, pp. 53490-53503, 2022, doi: 10.1109/ACCESS.2022.3174108.

[19] Calin, O. Statistics and Machine Learning Experiments in English and Romanian Poetry. Sci 2020, 2, 92. https://doi.org/10.3390/sci2040092

[20] R. Annamalai, S. Sudharson, T. Pratap and H. Kaushik, ”LSTM Based Monophonic Piano Melody Synthesis,” 2023 IEEE 7th Conference on Information and Communication Technology (CICT), Jabalpur, India, 2023, pp. 1-6, doi: 10.1109/CICT59886.2023.10455209.