Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/154/2025.TJ23212

# *Comparative and Performance Analysis of Different Deep Learning Models in Text Generation*

**Jiawei Xing**

*Beijing-Dublin International College, Beijing University of Technology, Beijing, China*
*jiawei.xing@ucdconnect.ie*

***Abstract:*** Deep learning-based text generation plays a critical role in modern natural language processing. It supports applications such as automatic translation, storytelling, and content creation. Recent research has explored various neural network architectures for generating coherent and contextually relevant text. This paper systematically reviews and compares four major architectures: Recurrent Neural Networks, Variational Autoencoders, Generative Adversarial Networks, and Transformer-based models. The comparative analysis uses standard performance metrics and human evaluations to identify each model's strengths and limitations. Results show that Transformer-based models outperform others in fluency, coherence, and context awareness, especially in tasks requiring long-range text understanding. However, controllability, interpretability, and ethical concerns still pose significant challenges. Future research directions include developing more efficient models, improving controllability mechanisms, and creating reliable evaluation methods. Overcoming these challenges will expand the practical use of text generation models in diverse real-world scenarios.

***Keywords:*** Deep Learning, Text Generation, Natural Language Processing, Transformers, Neural Networks

## 1. Introduction

Deep learning-based text generation technology plays a key role in modern natural language processing, enabling machines to produce human-like text for a wide range of applications. From automatic translation summaries to story generation, the ability to generate coherent and logical text is fundamental to human-computer interaction. Deep learning models have significantly advanced this field by learning complex speech patterns from large data sets, allowing systems to create more coherent text. This progress has driven many real-world applications that have high requirements for text fluency and content relevance, including chatbots, automatic content creation, and assisted writing systems.

To solve the problem of text generation, many deep learning architectures have been explored, each with unique advantages. A study showed that recurrent sequence-to-sequence models, such as long short-term memory neural networks, have achieved state-of-the-art results on text generation tasks such as machine translation, far surpassing previous rule-based methods. This breakthrough demonstrated the effectiveness of deep neural networks in capturing linguistic text [1]. To further improve the quality and diversity of generated text, researchers introduced variational autoencoders (VAEs) and generative adversarial networks (GANs). VAEs contain latent representations that

DOI: 10.54254/2755-2721/154/2025.TJ23212

enhance output diversity, and GAN-based methods adversarially train a generator and discriminator to produce more realistic text. However, adversarial text generators often suffer from mode collapse, such as repeated outputs, so VAEs are generally more reliable than GANs [1]. Despite some challenges, performance improvements have been observed. For example, the later LeakGAN model achieved a BLEU-4 score that is more than twice that of the earlier SeqGAN model (0.437 vs 0.178), which shows an improvement in output quality [1]. Several studies have evaluated these models on standard benchmarks using evaluation metrics such as BLEU, ROUGE, and perplexity, as well as human judgment of output quality. In fact, given the open-ended nature of language, human evaluation remains the gold standard for measuring generated text. For example, a dialogue system can generate multiple plausible responses to the same input, a flexibility that is difficult to capture with a single quote metric [2]. Despite this, automated metrics are still widely used for efficient comparison of model performance [2]. Recent studies have also demonstrated the superior performance of Transformer-based language models. For example, the large pre-trained model GPT-2 can generate fluent and coherent text that is often difficult to distinguish from human-written content [3]. These advances show that the latest models represented by the Transformer-based architecture are now state-of-the-art, surpassing some earlier methods in their ability to generate natural and plausible text.

This paper systematically compares the performance of Recurrent Neural Networks (RNNs), VAE, GAN, and Transformer architectures in the field of text generation by analyzing their performance in different tasks. The methodology adopted in this paper includes reviewing previous research and comparing model outputs using standard evaluation metrics supplemented by human judgment to obtain consistent evaluation results. The goal is to identify the relative strengths and weaknesses of each model, guide practitioners to choose the right technology for specific text generation scenarios, and propose directions for future improvements.

## 2. Theoretical principles

### 2.1. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs)

RNNs process sequences by maintaining a hidden state that is updated at each time node, allowing them to capture sequential patterns and dependencies in text [4]. However, basic RNNs struggle with long-term dependencies due to gradient explosion. LSTMs introduce a gating mechanism to regulate information flow, enabling long-range context retention and mitigating RNN training challenges [1]. This gated memory architecture enables LSTMs to outperform standard RNNs on tasks that require long-term sequence learning, making it the base model for early neural text generation systems.

### 2.2. Variational Autoencoders (VAEs)

Variational autoencoders are deep generative models that learn latent representations of data via an encoder-decoder architecture. The encoder network compresses the input (e.g., a sequence of text) into a continuous latent vector, and the decoder network reconstructs the original data from this latent code [1]. VAEs are trained by optimizing a loss function with two components. The first is a reconstruction loss, which ensures that the decoded output closely matches the input. The second is a regularization term, which encourages the latent variable distribution to align with a predefined prior, typically a Gaussian distribution [1]. This alignment is enforced through a Kullback-Leibler (KL) divergence penalty. It measures the difference between the encoder's output distribution and the chosen prior, thereby encouraging the latent space to capture generalizable features of the data. Since the encoder maps the input to a continuous distribution, new text can be generated by sampling a latent vector from the prior and feeding it to the decoder.

DOI: 10.54254/2755-2721/154/2025.TJ23212

## 2.3. Generative Adversarial Networks (GANs)

GANs employ an adversarial framework, where a generator produces text samples, and a discriminator assesses their authenticity (human-written or machine-generated) [1]. In the field of image synthesis, GANs have achieved remarkable realism with this approach. However, applying GANs to language poses additional challenges. The text consists of discrete tokens, making it non-differentiable and thus difficult to propagate gradients from the discriminator back to the generator [1]. Early text GANs approaches employed techniques such as policy gradients (reinforcement learning) or auxiliary teacher networks to align generated and real sequences. Even so, text GANs often suffer from problems such as mode collapse or producing repetitive short outputs. Therefore, GAN-based text generators have not yet reached the fluency and reliability of other deep generative models, and active research continues how to improve training stability [1].

## 2.4. Transformer-based models

Transformers revolutionized NLP by using self-attention to capture the context of the entire sequence without relying on recurrence [4]. In the Transformer model, each output token is generated by processing all input tokens (or previously generated tokens) through a series of learned attention weights. This architecture, proposed by Vaswani et al. (2017), allows highly parallel training and has achieved state-of-the-art performance in tasks ranging from machine translation to language understanding [4].

Generative Pre-trained Transformer (GPT) Models: OpenAI's GPT demonstrated the power of pre-training for large-scale text generation. GPT uses a unidirectional Transformer architecture and is trained with a simple objective: predict the next word given all previous words [3]. Its successor, GPT-2, significantly increased the model size (about 1.5 billion parameters) and showed that scaling the Transformer leads to very fluent and diverse text generation. The GPT family of models is purely generative and unsupervised, generating text in a free-form manner after being initialized with a prompt.

Conditional Transformer Language Model (CTRL) and Grover: Researchers developed Transformer-based generators such as CTRL and Grover. CTRL is a 1.63-billion-parameter Transformer language model trained on data labeled with various control codes (e.g., News, Reviews, Dialogue), allowing controllable text generation without model fine-tuning [3]. Grover, another large Transformer trained specifically on news data, uses a GPT-2-like architecture optimized for realistic news article generation. It effectively generates credible news texts and serves to detect machine-generated fake news, illustrating domain-specific training advantages [3].

Multilingual BERT (mBERT) and Variants: Transformer architectures extend beyond text generation to classification and representation learning. The multilingual variant, mBERT, pre-trained via masked language modeling on diverse languages, creates a shared cross-lingual embedding space beneficial for multilingual tasks like classification and named entity recognition [4]. DistilBERT, derived through knowledge distillation, retains BERT's performance with roughly half the parameters, providing efficient inference with minimal performance loss [4].

IndicBERT: Focused specifically on Indian languages, IndicBERT is based on a simplified ALBERT architecture and trained on extensive multilingual Indian corpora. This specialization helps it capture language-specific features effectively, particularly suited for low-resource language tasks [4]. These examples highlight Transformer flexibility across diverse NLP scenarios, tailored by altering architecture and training regimes.

DOI: 10.54254/2755-2721/154/2025.TJ23212

## 3. Comparative analysis

### 3.1. Authorship attribution for neural text generation

Uchendu et al. evaluate the effectiveness of generative models such as GPT, GPT-2, CTRL, and GROVER in distinguishing machine-generated text from human-written text. The study finds that while most neural generators still have significant differences from human writing, advanced models (GPT-2, GROVER) produce more realistic text, which often confuses automatic classifiers [3]. By stylistic measures, the outputs of GPT-2 and GROVER are closer to human text than those of earlier models, indicating that these generators effectively mimic human writing style [3].

Classifiers can achieve high accuracy in authorship attribution. In a 9-class attribution test (1 human vs. 8 generators), the automated models overall achieve a macro F1 score of nearly 90% [3]. The more formulaic generators (such as GPT, CTRL) are the easiest to spot, while GROVER's human-like text is the hardest to spot (only about 0.55 F1, much lower than the others) [3]. Therefore, as machine-generated content becomes increasingly difficult to identify, the new generators' strong ability to mimic human output becomes a weakness for detection.

### 3.2. Exploring controllable text generation techniques

Prabhumoye et al. examine how different models support controllable text generation in tasks such as story writing, conversation, and email composition. Traditional RNN-based models can enforce desired properties by modulating control signals, such as adjusting the personality or politeness of conversational responses [5]. Similarly, story generation can be guided by specific plot points or endings, and even email composition systems can change the form or tone of messages as needed [5]. These examples show that, under the right conditions, generative models can integrate style or content constraints into their output.

However, model architectures differ significantly in supporting controlled text generation for long sequences. LSTM-based generators typically require retraining for each new attribute and struggle to maintain coherence, achieving lower coherence scores (BLEU ~23.7) compared to GPT-2 (BLEU ~34.5) in narrative text generation [5]. GAN-based models frequently encounter instabilities such as mode collapse, producing repetitive or truncated outputs and reduced text diversity (self-BLEU scores up to 0.67) [1]. In contrast, large pre-trained Transformers like GPT-2 exhibit superior fluency, maintaining topic coherence approximately 1.6 times better than LSTM generators, and can effectively incorporate lightweight controllers without significant quality loss [5]. For example, a plug-and-play approach can guide the topics of GPT-2 through auxiliary models while minimizing the loss of coherence [3]. Overall, when paired with these techniques, modern pre-trained generators offer better controllability and quality than earlier models.

### 3.3. Telugu language hate speech detection using transformer models

Khanduja et al. focus on the classification task of detecting hate speech in Telugu, a low-resource language, using fine-tuned transformer models. Multilingual pre-trained transformers (mBERT, DistilBERT, and IndicBERT) are fine-tuned on a labeled corpus of Telugu tweets [4]. Despite limited data, these models leverage their pre-training capabilities to achieve strong hate-speech classification, highlighting the value of transfer learning in low-resource scenarios.

In terms of performance, all three Transformer models achieve very high accuracy. The fine-tuned mBERT model achieves an accuracy of approximately 98.2%, followed by IndicBERT and DistilBERT (both approximately 98%), and all of these models significantly outperform the traditional RNN+LSTM baseline model (approximately 91%) [4]. Notably, the compact DistilBERT is almost on par with mBERT, which is valuable for deployments with limited resources [4].

DOI: 10.54254/2755-2721/154/2025.TJ23212

Compared to traditional RNN and LSTM models, pre-trained Transformer models have clear advantages in Telugu hate speech detection. They effectively capture long-range context and semantic nuances. Transformers also reduce training and inference time, improving both efficiency and accuracy [4].

## 4. Challenges and outlook

### 4.1. Controllability

Deep learning-based text generation models often lack fine-grained controllability, making it difficult to steer outputs toward specific content, style, or constraints. Ensuring that a model's output meets user or domain requirements is crucial yet challenging in practice. For instance, one automotive service system used NLP and deep learning to filter out vague or misleading free-text inputs, thereby controlling input quality and improving the reliability of downstream results. Another example is a multilingual image captioning system that lets users choose the output language. The generator adapts accordingly, demonstrating controlled generation aligned with user preferences. Despite such advances, achieving precise control (e.g., enforcing factual accuracy or a particular tone) remains an open challenge.

### 4.2. Interpretability

Most deep text generation models operate as "black boxes" with internal reasoning that is difficult to explain [6]. This lack of transparency hinders debugging and trust, as errors or biases in the generation process may go undetected [6]. Although recent research explores explainability techniques (such as attention-weight visualization or attribution methods) to illuminate how a model produces text, practical interpretability is still limited. Without clearer model explanations, users in critical domains (e.g., healthcare or customer service) may remain wary of fully relying on generated outputs.

### 4.3. Innovation

Deep learning continues to drive innovation in text generation, expanding the technology's adaptability and reach. A common approach is to leverage large pre-trained language models and fine-tune them on the target domain or language, which does improve adaptability [7]. This strategy has enabled systems to perform well even in specialized or multilingual settings. For example, a recent cross-modal system generates image descriptions as spoken captions in multiple languages, achieving high bilingual evaluation scores (BLEU-1 $\approx$ 0.48) and effectively bridging language barriers [8]. In specialized text domains, advanced models can even surpass human performance. One model for vehicle service report analysis validated requests with 18% higher accuracy than experienced technicians. Ongoing research is exploring greater creativity and context-awareness in generation, aiming to produce more novel and insightful outputs beyond what is seen in training data.

### 4.4. Ethical issues

Deep generative models pose several ethical challenges. They can inadvertently learn and amplify biases from training data, leading to outputs that are prejudiced or offensive [7]. Another concern is the generation of plausible-sounding yet incorrect information, which can propagate misinformation if not detected. Additionally, many models underperform for less-represented languages or dialects, raising fairness issues about unequal access to the benefits of text generation technology [9].

DOI: 10.54254/2755-2721/154/2025.TJ23212

### 4.5. Outlook

Researchers are actively investigating new architectures and training paradigms to overcome current limitations. One emerging line of work is diffusion-based text generators, which borrow techniques from image generation to improve text output incrementally—these models aim to improve the diversity and controllability of generated text. Another promising trend is the rise of large-scale multimodal systems, which integrate text with other modalities such as images, audio, or video [6]. By building on the linguistic foundations in other data sources, multimodal models can produce more contextually relevant and factual descriptions. Early studies of multimodal deep learning have highlighted their potential to revolutionize content creation, although they have also introduced new challenges in aligning cross-modal representations. In summary, as researchers address current limitations, the next generation of text generation models will likely be more controllable, more interpretable, more adaptable, and more responsibly designed. Future work will bring us closer to truly reliable and general text generation systems that apply to a wide range of languages and domains by improving training stability, evaluation protocols, and ethical standards [5, 6]. Recent advancements demonstrate that combining domain-specific NLP preprocessing techniques with deep neural architectures significantly enhances accuracy and reliability in real-world scenarios, such as automated vehicle diagnostics based on free-text customer reports [10].

## 5. Conclusion

The comparative analysis of deep learning models for text generation revealed clear performance distinctions. Transformer-based architectures consistently produced more fluent and coherent text than recurrent models like LSTMs or GRUs, excelling at capturing long-range context that led to lower perplexity and more relevant outputs. Recurrent networks performed well on short sequences but struggled to maintain context over longer passages. Convolutional sequence models offered faster parallel processing but underperformed on tasks requiring long-term coherence. Generative models with latent variables, such as variational autoencoders, improved output diversity, while adversarial approaches offered some creative variety at the cost of training stability.

These findings reinforce the current theory by confirming that attention mechanisms and large-scale representation learning improve sequence modeling. For the highest text quality, a transformer-based model should be chosen if computational resources allow. Simpler RNN models remain useful for applications with limited hardware or real-time requirements. Thus, deploying a text generation system involves balancing output quality and computational efficiency based on application needs. The evidence also shows that combining architectures or fine-tuning pre-trained models can significantly boost performance, providing a practical blueprint for system developers.

This review had a limited scope as it focused mainly on English text generation tasks. Its findings may not fully generalize to other languages or modalities. Performance comparisons relied on metrics from different studies. It emphasized trends over absolute scores to ensure fairness in these comparisons. Without a uniform evaluation framework, some subtle differences between models might be missed. Finally, deployment factors like inference speed and memory usage were only discussed qualitatively.

Text generation models have diverse applications and promising future prospects. They are being applied in content creation, conversational agents, storytelling, and assistive writing tools. Continued progress will enable more natural and context-aware AI writers. These models could transform creative industries, information access, and human-computer interaction in the coming years and beyond.

DOI: 10.54254/2755-2721/154/2025.TJ23212

## References


[1] *Iqbal, T., & Qureshi, S. (2022). Text generation models in deep learning. Journal of King Saud University – Computer and Information Sciences, 34, 2515–2528.*

[2] *Celikyilmaz, A., Clark, E., & Gao, J. (2021). Evaluation of text generation: A survey. arXiv preprint arXiv: 2006.14799.*

[3] *Uchendu, A., Le, T., Shu, K., & Lee, D. (2020). Authorship attribution for neural text generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 8384–8395). Association for Computational Linguistics.*

[4] *Khanduja, V., Sharma, G., & Pamula, R. (2024). Telugu language hate speech detection using deep learning transformer models. Procedia Computer Science, 233, 546–556.*

[5] *Prabhumoye, S., Black, A. W., & Salakhutdinov, R. (2020). Exploring controllable text generation techniques. arXiv preprint arXiv: 2005.01822.*

[6] *Sun, Z., Lin, M., Zhu, Q., Xie, Q., Wang, F., Lu, Z., & Peng, Y. (2023). A scoping review on multimodal deep learning in biomedical images and texts. Journal of Biomedical Informatics, 146, 104482.*

[7] *Ahmed, N., Saha, A. K., Al Noman, M. A., Jim, J. R., Mridha, M. F., & Kabir, M. M. (2024). Deep learning-based natural language processing in human–agent interaction: Applications, advancements and challenges. Natural Language Processing Journal, 9, 100112.*

[8] *Sangolgi, V. A., Patil, M. B., Vidap, S. S., Doijode, S. S., Mulmane, S. Y., & Vadaje, A. S. (2024). Enhancing cross-linguistic image caption generation with Indian multilingual voice interfaces using deep learning techniques. Procedia Computer Science, 233, 547–557.*

[9] *Ahmad, H. A., & Rashid, T. A. (2024). Planning the development of text-to-speech synthesis models and datasets with dynamic deep learning. Journal of King Saud University - Computer and Information Sciences, 36, 102131.*

[10] *Khodadadi, A., Ghandiparsi, S., & Chuah, C.-N. (2022). Natural language processing and deep learning-based model for automated vehicle diagnostics using free-text customer service reports. Machine Learning with Applications, 10, 100424.*