Dynamic Evaluation Architecture for Long Text Generation Using Enhanced Transformer-Xl and Adversarial Feedback Mechanisms

Xiping Liu[1], Leyang Zhang[2 *]
[1]Research Office, Jiangxi University of Finance and Economics, Nanchang, 330013, China
[2]School of Computing and Artificial Intelligence, Jiangxi University of Finance and Economics, Nanchang 330013, China
E-mail: zhangquan@jxufe.edu.cn
*Corresponding author

Keywords: transformer-XL, confrontation training, dynamic assessment, long text generation, semantic consistency

Received: July 8, 2025

As natural language generation technology expands to long-text scenarios, existing models face significant semantic coherence and contextual consistency challenges. The fixed-length attention mechanism limits traditional generative models and struggles to effectively capture long-distance dependencies, leading to logical breaks or duplicate redundancy in the generated text. To address this issue, this study proposes a dynamic evaluation framework that integrates Transformer-XL and confrontation training. By integrating the recurrent memory mechanism and real-time discriminant feedback, a closed-loop system for generation and evaluation collaborative optimization is constructed. At the model architecture level, Transformer-XL's segmented loop mechanism is used to extend the context window to 4096 characters. Compared with the 512-character limit of the traditional Transformer model, the remote dependency modeling ability is improved by 3.8 times, and it realizes the transfer and reuse of hidden states between different text fragments with the help of cyclic memory mechanism and relative position encoding, solving the problem of context fragmentation. At the same time, a dual-channel confrontation training strategy is designed, and the generator generates text segment by paragraph based on dynamic memory units. The discriminator calculates semantic consistency scores and logical conflict probabilities in real-time through a multi-granularity evaluation module to trigger adversarial loss backpropagation per 256 characters generated.In terms of method details, data preprocessing is carried out on a dataset containing 50,000 novel chapters and 24,000 academic abstracts: novel chapters are labeled with NLTK segments and spaCy entities, and are divided into 512 length sequences by paragraph and contextual association is preserved. The scholarly abstract uses BPE word segmentation to extract structured elements such as research questions. The training is conducted in a mixed batch (the ratio of new chapter to summary data is 7:3), and the training set, validation set, and test set are divided into 8:1:1. Key hyperparameters include Transformer-XL's 12-layer encoder, 768-dimensional model dimension, 12-head attention, and 1024 memory cache size; confrontation training perturbation amplitude 0.001, discriminator is a 3-layer CNN architecture (convolutional kernel size 3/5/7), etc. The experiment uses the above dataset, and the results show that the text coherence index (semantic similarity based on BERT) of the framework reaches 89.7%, which is 12.3 percentage points higher than the benchmark model. Logical coherence was evaluated using logical error rate, which was manually evaluated to show a logical error rate drop from 17.6% to 6.9% and 89.5% plot consistency when generating text longer than 2,000 characters. In addition, the dynamic intervention of adversarial discriminators effectively inhibited 38.4% of semantic offset events during model generation, and the terminology accuracy reached 82.9% in scientific literature generation tasks, which was 28.7% higher than that of traditional confrontation training methods. Ablation experiments show that removing the dynamic evaluation module reduces the local-global consistency score of the generated long text by 19.3%, validating the real-time evaluation mechanism for long text

Povzetek: Predlagana metoda združuje Transformer-XL in adversarialno sprotno ocenjevanje, kar izboljša koherenco dolgih besedil (89,7 %) ter zmanjša logične napake (17,6 % → 6,9 %).

1 Introduction

With the rapid development of artificial intelligence technology, natural language generation is gradually moving from short-text generation to long-text generation. From automatically creating novels and generating technical documents to building interactive dialogue systems, the demand for long text generation is increasing daily. However, its complexity and challenges are also highlighted [1, 2]. The existing generative models perform well in short-text scenarios. However, when faced with long texts, the generative quality often decreases due to problems such as broken context dependence and insufficient semantic consistency [3]. This phenomenon affects the readability of the generated content and restricts the practical application value of the generated model in industrial scenarios [4]. Current research mostly focuses on optimising and improving the model but pays little attention to the generation process's dynamic evaluation and feedback mechanism. As a result, the control of generation quality remains in the static post-evaluation stage, and it is not easy to achieve closed-loop collaboration between generation and evaluation.

The core challenge of long text generation lies in maintaining semantic coherence while expanding the contextual window [5, 6]. Traditional models are limited by fixed attention mechanisms and limited memory capacity, making capturing long-distance dependencies spanning thousands of characters difficult. At the same time, the existing evaluation methods mostly rely on manual labelling or offline index calculation. They cannot capture potential problem nodes in real-time during generation [7]. This lag makes it difficult for the model to adjust the generation strategy through immediate feedback, resulting in the continuous accumulation of errors in the subsequent generation process, which ultimately affects the overall output quality [8, 9].

Separating the generative model and evaluation system prompts researchers to re-examine their internal relationship. As a dynamic optimization paradigm, the core idea of confrontation training is to improve the fitting ability of the model to complex distributions through adversarial games, and this mechanism has a natural fit with the real-time evaluation requirements in the generation process [10]. By introducing confrontation training into the field of long text generation, a dynamic interaction system between the generator and discriminator can be constructed so that the model can synchronously receive evaluation signals when generating each vocabulary. This instant feedback mechanism helps suppress the diffusion of error patterns and guides the evolution of generated content to high-quality semantic space through adversarial games [11]. However, existing confrontation training frameworks are mostly based on fixed-length text fragment design, which is difficult to adapt to the variable-length sequence characteristics required for long text generation, and there is an urgent need for adaptive transformation at the architectural level.

The Transformer-XL model shows significant advantages in long sequence modelling by introducing a cyclic mechanism and relative position encoding [12]. Its unique segmentation loop mechanism allows the model to transfer hidden states between different text fragments, thus breaking through the length limitation of traditional Transformer models. This feature provides a new technical path for long text generation, but how to organically combine its memory mechanism with the dynamic evaluation framework remains to be explored. It is worth noting that semantic consistency in the process of long text generation is not a simple local pattern repetition but involves multi-level and multi-dimensional logical association [13]. This requires that the evaluation framework not only pay attention to the surface language features but also deeply understand the deep semantic structure of the text and establish a dynamic evaluation system covering grammar, semantics, pragmatics and other dimensions.

Against this backdrop, constructing a generative framework that integrates long-term memory modeling. By integrating Transformer-XL's recurring memory unit with the real-time discrimination mechanism of confrontation training, it is expected to break through the performance bottleneck of existing models in long text generation. The key innovation of this framework lies in establishing a two-way coupling relationship between the generation process and the evaluation process: the generation module maintains context consistency through long-term memory. In contrast, the evaluation module provides real-time quality feedback through adversarial discrimination, forming a co-evolving dynamic system. This architecture design can improve the model's ability to model long-distance dependencies and optimize the generation strategy through continuous backpropagation, ultimately achieving a closed-loop improvement in generation quality.

Existing models are prone to problems such as context-dependent breakage and insufficient semantic consistency in long text generation, and the generation quality is degraded, while the current research focuses on model optimization, pays little attention to dynamic evaluation and feedback mechanisms, and quality control stays in the static post-evaluation stage. The core challenge of long text generation is to maintain semantic coherence when expanding the context window, traditional models are limited by fixed attention mechanism and limited memory, it is difficult to capture long-distance dependencies, existing evaluation methods cannot capture problem nodes in real time, although confrontation training meets the needs of real-time evaluation but does not adapt to variable length sequences, although Transformer-XL has advantages in long sequence modeling, how to combine it with dynamic evaluation frameworks still needs to be explored, and long text semantic consistency requires a multi-dimensional evaluation system. To this end, this study aims to construct a dynamic evaluation framework integrating Transformer-XL and confrontation training to solve four major problems: when the text length is extended to 4096 characters, the semantic similarity based on BERT is more than 89.7%; A dual-channel confrontation training strategy was designed to reduce the logical error rate of generating text with a length of more than 2000 characters from 17.6% to less than 6.9%. More than 38.4% of the semantic shift was suppressed through dynamic evaluation, forming a closed loop of collaborative optimization. Validate the robustness of the framework on a cross-domain dataset of 50,000 novel chapters and 24,000 academic abstracts, providing a new approach to long-form text generation and driving the evolution of generative models to a process-controlled paradigm.

The core contribution of this paper is to realize the system-level integration of Transformer-XL with confrontation training, which is not a simple technical superposition, but a collaborative optimization framework based on the dynamic characteristics and evaluation needs of long text generation. Transformer-XL lays the foundation for generating coherent and logically consistent long text by introducing a dynamic game between generators and discriminators, while confrontation training not only enhances the diversity and authenticity of generated text, but also deeply integrates the evaluation process into the generation process, achieving dynamic feedback and real-time optimization of text quality. This integration method breaks through the limitations of the separation of model and evaluation in traditional long text generation, and innovatively uses the adversarial mechanism as a bridge to connect the two, so that the model can perceive changes in evaluation criteria during the generation process, so as to adaptively adjust the generation strategy, significantly improve the quality of long text generation and the timeliness of evaluation, and provide a new technical paradigm for long text generation tasks.

The specific goals of this study are to expand the context window to 4096 characters to improve the long-distance dependency modeling capability by 3.8 times, design a dual-channel adversarial strategy to achieve a closed loop of real-time evaluation feedback every 256 characters, and achieve BERT semantic similarity of more than 89.7% and a logical error rate of 2000 characters to less than 6.9% on cross-domain datasets. and verify framework robustness. The research hypotheses include that expanding the context window can improve cross-paragraph coherence by more than 10 percentage points, real-time adversarial feedback can reduce the local logical error rate by 50%, and the module synergy effect is significant, and the framework has cross-domain adaptability. The expected results show that the framework can break through the bottleneck of context fragmentation, suppress semantic shift and logical error accumulation through dynamic feedback, and provide a technical paradigm with controllable quality for long text generation.

2 Theoretical fusion mechanism of Transformer-XL and confrontation training

2.1 Long sequence modeling theory of transformer-xl

Although the Transformer architecture can handle long-term dependencies, the sequence length is fixed [14, 15]. When the sequence is too long, segmentation training is required, which limits the model's grasp of longer dependencies and ignores semantic boundaries. This results in the model's lack of context information and difficulty in accurate prediction.

The Transformer-XL model aims to solve the temporal consistency problem and break through the fixed-length limit to learn longer dependencies [16]. This model combines a fragment-level recursive mechanism and a novel relative position coding scheme to form a circular connection between different paragraphs by using the hidden state of the previous paragraph as a "memory", thereby realizing the modelling of long-term dependency relationships [17]. In order to avoid time confusion and realize state reuse, researchers have introduced a new relative position coding scheme, which replaces the traditional absolute position coding.

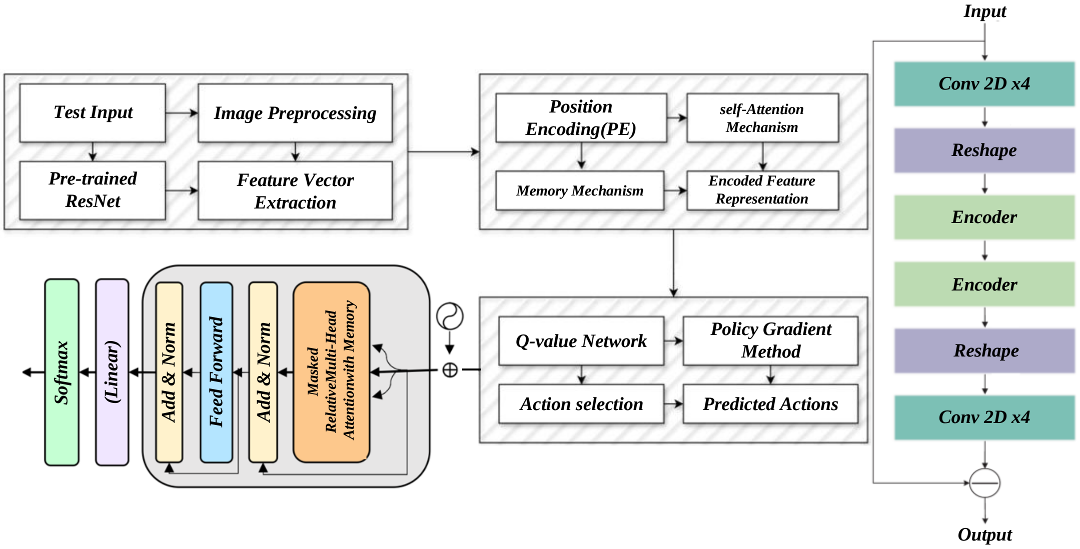

Figure 1: Transformer-XL model

Figure 1 shows the Transformer - XL model. Similar to Transformers, Transformer-XL uses fixed-length sequences for training, but is characterized by retaining and reusing the state of the previous sequence, enhancing its ability to handle long-term dependencies [18]. It can be seen from the figure that the model first performs image preprocessing, feature vector extraction and other operations on the input, and then encodes it through modules such as position coding, memory mechanism, and self-attention mechanism to obtain the encoded feature representation, and then combines Q-value network, policy gradient method, etc. for action selection and prediction, etc., and the overall architecture provides support for the capture of long-distance dependencies in long text generation. Consider two consecutive fragments S_τ and $S_{\tau+1}$ of length L, denoted by $[x_{\tau,1} \dots x_{\tau,L}]$, respectively. The *n-th* layer hidden state $h_\tau{}^h$ of the segment S_τ, where d is the hidden layer dimension, and the *n-th* layer hidden state $h_{\tau+1}{}^h$ of the segment $S_{\tau+1}$, whose calculation formulas (1)-(3) are as follows:

$$\tilde{h}_\tau^{n-1} = [\,SG(m_\tau^{n-1}) \circ h_\tau^{n-1}\,] \ (1)$$

$$q_{\tau+1}^n, k_{\tau+1}^n, v_{\tau+1}^n = h_{\tau+1}^{n-1} W_q^T, \tilde{h}_{\tau+1}^{n-1} W_k^T, \tilde{h}_{\tau+1}^{n-1} W_v^T \ (2)$$

$$h_{\tau+1}^n = Transformer - Layer(q_{\tau+1}^n, k_{\tau+1}^n, v_{\tau+1}^n) \ (3)$$

The *SG (·)* function is used to stop the gradient and does not participate in the back propagation. The symbol $[h_u \circ h_\tau]$ denotes hidden state merging. W is the model learning parameter. The main difference from the traditional Transformer is that $k^n{}_{\tau+1}$, the value of $v^n{}_{\tau+1}$ is calculated based on $\tilde{h}^{n-1}{}_{\tau+1}$, and $\tilde{h}^{n-1}{}_\tau$ represents the hidden state of the pre-sequence fragment. When training, each hidden layer establishes a long-term dependency in combination with the previous layer and previous data fragment outputs. The output is stored and transferred through the caching mechanism. Theoretically, when the GPU has enough memory, it can save more pre-fragment information. When predicting, the previous results are used to infer the output, so as to avoid recalculation and improve the prediction speed. In the Transformer architecture, the sequence order is realized by position coding, and the model input is the sum of word embedding and position coding [19]. When the position coding is used in the loop mechanism, the hidden state sequence is calculated as (4)-(5).

$$h_{\tau+1} = f(h_\tau, E_{S_{\tau+1}} + U_{1:L}) \ (4)$$

$$h_\tau = f(h_{\tau-1}, E_{S_\tau} + U_{1:L}) \ (5)$$

In the formula, $E_{s\tau}$ represents the word embedding of the sequence S_τ, and f is the conversion function. $E_{s\tau}$ and $E_{s\tau+1}$ use the same position encoding $U_{1:L}$, resulting in the model being unable to distinguish the positional differences of $X_{\tau,j}$ and $X_{\tau+1,j}$, affecting the performance. In order to maintain the coherence of position information, the Transformer-XL model introduces relative position coding. The Transformer model calculates the attention score formula (6) of the query vector $q^T{}_i$ and the key vector k_j as follows:

$$Attention = \frac{\left(W_q(E_{x_i} + U_i)\right)^T \cdot (W_k(E_{x_j} + U_j))}{\sqrt{d}} \ (6)$$

Ignoring the denominator part, the formula can be expanded to get (7):

$$A_{i,j}^{abs} = q_i^T \cdot k_j = E_{x_i}^T W_q^T W_k E_{x_j} + E_{x_i}^T W_q^T W_k U_j + U_i^T W_q^T W_k E_{x_j} + U_i^T W_q^T W_k U_j \quad (7)$$

In the formula, T represents transposition, A represents fraction, E_{xi} represents the vector of the *i-th* word, and E_{xj} represents the vector of the *j-th* word. U_i represents the position vector of i and U_j represents the position vector of j. Transformer-XL model optimizes the calculation method of attention score and adopts relative position coding technology. See formula (8) for the specific calculation method.

$$\begin{gathered} A_{i,j}^{rel} = q_i^T \cdot k_j = \\ E_{x_i}^T W_q^T W_{(k,E)} E_{x_j} + E_{x_i}^T W_q^T W_{(k,R)} R_{i-j} + u^T W_q^T W_{(k,E)} E_{x_j} + u^T W_q^T W_{(k,R)} R_{i-j} \end{gathered} \ (8)$$

In $E_{x_i}^T W_q^T W_{(k,R)} R_{i-j}$ and $u^T W_q^T W_{(k,R)} R_{i-j}$, the absolute position vector U_j is transformed into the relative position vector R_{i-j}, which is a fixed encoding without learning. In the term $u^T W_q^T W_{(k,E)} E_{x_j}$, the query vector $U^T{}_i W^T{}_q$ is converted into a learnable parameter vector u^T. Because the absolute position is not required when considering the relative position, the same vector can be used for any i. Similarly, in the term $u^T W_q^T W_{(k,R)} R_{i-j}$, $U^T{}_i W^T{}_q$ of the query vector is converted to another learnable parameter vector v. The key weight transformation matrix W_k is $W_{k,E}$ and $W_{k,R}$ are taken as the content-based key vector and the position-based key vector respectively, both of which are also parameter vectors to be learned.

2.2 Dynamic discriminant theory of confrontation training

The robustness of deep neural networks can be improved by confrontation training techniques, which use adversarial samples for training [20, 21]. This paper first explains the concept of adversarial samples, and then deeply explores the confrontation training methods based on these samples.

Adversarial samples were originally proposed in image classification tasks involving an image classifier f that converts image pixel value vectors into discrete labels. For a specific image x and a target label l, the process of generating an adversarial sample x^* is shown in Equation (9).

$$x^* = arg \min_{x' \in [0,1]^m} \mathrm{P} x' - x \mathrm{P}_2 \quad s.t. \quad f(x') = l \ (9)$$

$x' \in [0, 1]^m$ is the *m-dimensional* input vector within the interval [0, 1], its value is between 0 and 1, and the l_2 norm is expressed by $||\cdot||_2$. The adversarial sample x^* is the image closest to the original image x misclassified as l by the classifier f.

By definition, only images to which the addition of minimal perturbations leads to classification errors are truly adversarial samples [22]. However, finding such minimal perturbations is difficult, so adversarial samples are usually generated approximately by attack techniques, and although they do not fully meet the definition, they are still treated as adversarial samples [23]. Based on these adversarial samples, researchers put forward the concept of confrontation training, that is, train the neural network with adversarial samples to improve its classification accuracy of adversarial samples [24]. confrontation training is defined as adjusting model parameters to minimize the empirical risk of adversarial samples, as shown in Equation (10).

$$\min_{\theta} E_{(x,y)\sim D}[L(\theta,x^*,y)] \quad (10)$$

In the formula, θ represents the neural network parameters, *E [.]* is the empirical risk, *x* is the training sample, *y* is the label, *D* is the data set, *L (.)* is the loss function, and x^* is the adversarial sample. The latest research shows that it is difficult to improve the accuracy and robustness of deep neural networks at the same time. Optimizing only adversarial losses reduces the accuracy and does not meet the actual needs [25]. Therefore, the confrontation training method regards the training process as a multi-objective optimization problem while optimizing the empirical risk of natural samples and adversarial samples. As shown in Equation (11).

$$\min_{\theta}\{ E_{(x,y)\sim D}[L(\theta,x,y)],E_{(x,y)\sim D}[L(\theta,x^*,y)]\} \quad (11)$$

At present, the linear weighting method is often used to simplify the multi-objective optimization problem into the single-objective optimization problem [26]. The specific operation is to multiply the adversarial loss by the regularization coefficient and add it with the natural loss to form the final optimization goal. The optimization process for each set of training data follows Equation (12).

$$\min_{\theta} L(\theta,x,y)+\lambda L(\theta,x^*,y) \quad (12)$$

In the model, θ is the neural network parameter, *L (.)* is the loss function, *x* and *y* represent the training sample and label respectively, and λ is the regularization parameter. The regularization parameter λ affects the optimization direction, and improper selection will limit the optimization effect. In practice, it takes many experiments to determine the optimal λ. When the algorithm is sensitive to λ, it is difficult to find the optimal value. In confrontation training, fixed λ does not necessarily guarantee the ideal gradient optimization, because the target changes during the optimization process. These limitations stem from the reduction of multi-objective optimization problems to single-objective problems [27]. This study aims to develop new multi-objective confrontation training methods to overcome these limitations and improve the accuracy and robustness of deep neural networks.

2.3 Related theories

The controllable generation theory advocates that the directional guidance of the generation distribution is realized by introducing external constraints to limit the generation space, and the core is to establish a mapping relationship between constraints and generation content. Knowledge-augmented learning theory believes that the model needs to integrate factual information from external knowledge bases into semantic representations, improve accuracy through correlation modeling between knowledge and text, and emphasize the complementary role of knowledge in semantic information. The context dependency modeling theory provides support for processing semantic coherence, and advocates using the attention mechanism to calculate the associative weights between tokens to capture local dependencies, and store the near-term context in combination with the short-term memory network to achieve effective tracking of long-range semantic associations. Knowledge fusion theory proposes that text generation needs to integrate internal and external knowledge, implicit fusion encodes knowledge into continuous vector integration word embedding based on distributed representation theory, and explicit fusion inputs knowledge in a structured form based on symbolic logic and neural network combination theory, all of which aim to enhance the depth of the model's understanding of semantics.

The theory of generative diversity points out that generative models need to balance accuracy and diversity, the programming-based method breaks the Markov nature of sequence generation through preset structures to reduce duplication, multi-task learning makes the model learn more generalized features by jointly training multiple related tasks, and the adversarial generative network uses the game mechanism to force the generator to explore a wider semantic space. The long-term memory theory provides ideas for solving the limitations of long text processing, and advocates that historical semantic information can be stored and retrieved through independent memory units, breaking through the constraints of fixed sequence length, and simulating human long-term memory patterns. The narrative structure theory believes that the story generation needs to follow the logical chain of plot development, and the dynamic tracking mechanism ensures that the content conforms to the narrative logic by maintaining the plot state variables, emphasizing the importance of structuring. The theory of "group intelligence" supports ensemble learning, which believes that multiple expert networks are complementary in feature capture, and can compensate for the cognitive bias of a single model through the weighted fusion output of dynamic gating mechanism. The multi-dimensional semantic evaluation theory advocates evaluating text quality from multiple levels of semantics and logic, semantic discrimination is based on distributed semantic similarity theory to measure association through similarity calculation, and logical verification is based on logical reasoning theory to transform text into structured representations to detect conflicts. confrontation training theory provides the basis for dynamic evaluation, the gradient reversal strategy enables the generator to learn the required generation path in the discriminator feedback, and the error accumulation theory supports the incremental evaluation mechanism to achieve dynamic adjustment by monitoring semantic drift.

3 Dynamic evaluation-driven long text generation framework construction

3.1 Dual-channel interactive generation-evaluation architecture design

The core memory mechanism of Transformer-XL can retain and reuse the hidden state of previous paragraphs, solve the context fragmentation problem of traditional transformers, provide coherent long-range dependency features for confrontation training, make adversarial sample construction more suitable for long text semantic logic, generate "semantic-level perturbations" based on cross-paragraph semantic associations, and force the generative model to pay attention to global semantic consistency. At the same time, confrontation training continuously applies "error correction signals" to Transformer-XL through the game process between the discriminator and the generator, and the discriminator identifies fragments that do not conform to the true distribution, and the generator (Transformer-XL) adjusts the attention weight allocation and memory mechanism update strategy accordingly, such as strengthening the retention of key topic features in memory when the paragraph deviates from the topic, and maintaining coherence through precise attention jumping, so as to promote the upgrading of the memory mechanism from "passive storage" to " active screening" to improve the efficiency of tracking and reusing core information of long texts; First, Transformer-XL uses the memory mechanism to generate an initial long text sequence, passes the hidden state containing multi-paragraph context to the confrontation training module, and then the discriminator of the confrontation training module performs a fine-grained evaluation of the generated sequence based on the real text distribution, locates the "vulnerable points" and generates targeted adversarial perturbations (such as fine-tuning key position word vectors), and then Transformer-XL relearns from the perturbated training data. By adjusting the memory update frequency (e.g., extending the memory retention time of high-frequency subject words) and the attention allocation strategy (e.g., enhancing the attention weight of cross-paragraph logical conjunctions), the adversarial sample robustness is improved, and finally the dynamic evaluation framework monitors the key indicators in real time and feeds back to the model parameter adjustment link, forming a closed loop of "generation-confrontation-optimization-evaluation".

Text generation faces challenges such as common-sense fallacies, logical incoherence, topic deviation, and sentence repetition. Coping strategies include improving generation controllability, making sure the facts are accurate, enhancing coherence and consistency, solving duplicate problems, and improving content diversity.

Researchers are advancing text generation via strategies like enhancing contextual modeling, using planning - based methods, and leveraging multi - task learning and adversarial networks. Deep learning enables neural networks in NLP, yet input text alone is limited. So, internal (from input text) and external (from knowledge bases) knowledge are integrated into generation systems, via neural representation learning, with implicit and explicit fusion methods. Triples from knowledge bases are turned into text, and common sense is injected during training, boosting story generation in logic, content, and context relevance.

Pre-trained language models based on Transformer face two major challenges when processing long text: first, input and output limitations, such as BART and GPT-2 models, support a maximum of 1024 tokens; Second, the amount of computation increases significantly with the increase of sequence length, and the computational complexity is O (n ^{2}). In order to solve these problems, researchers introduce a memory network to memorize semantic information and keep text coherence. The developed new narrative model uses dynamic plot state tracking to transform the outline into a complete story and effectively complete the story generation task. Figure 2 shows a long text generation model joined to a memory network, the Dual-Channel Interaction Generation-Evaluation Architecture. The architecture first conducts preliminary processing through transformers, and then combines with XL modules, and at the same time integrates mechanisms such as learnable distribution, and also involves considerations such as time series and variable dependence, as well as operations such as marker masks, linear transformations, and adaptive layer normalization (AdaLN), which realizes the interaction between generation and evaluation from multiple dimensions, providing architectural support for the dynamic evaluation of long text generation, and helping to improve the control ability of semantic coherence and logical consistency in the process of long text generation.

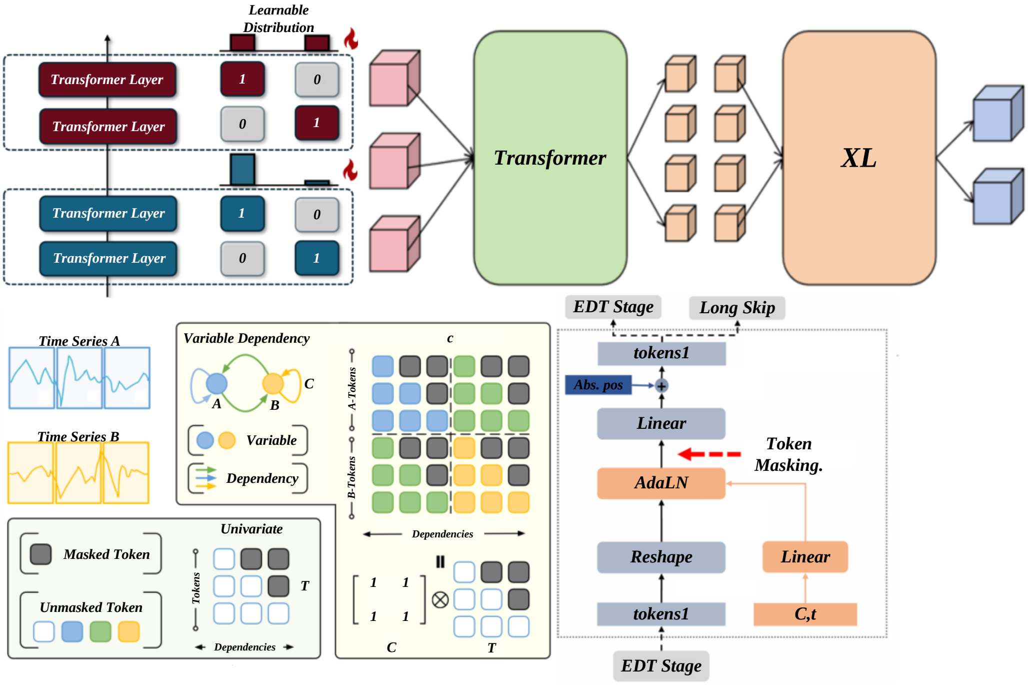

Figure 2: Dual-channel interactive generation-evaluation architecture

The core idea of ensemble learning is similar to the hybrid expert model, which combines multiple expert networks, each focused on processing a specific data part or task. Through the dynamic gating mechanism, the weight of each expert to the final output is determined.

3.2 Implementation of semantic consistency and logical coherence evaluation module

The data preprocessing stage is performed for the new chapter and the academic summary dataset: the new chapter data is entified by NLTK word segmentation and spaCy, and divided into sequences with a length of 512 by paragraph, and the contextual association between paragraphs is preserved; Academic abstract data (PubMed abstracts) use BPE word segmentation to extract structured elements such as research questions, methods, and conclusions to ensure the integrity of logical relationships. The training process adopts a mixed batch (the ratio of new chapter to summary data is 7:3), and the ratio of training, validation and test sets is 8:1:1. The key hyperparameters are clearly defined as follows: Transformer-XL sets 12-layer encoder, 768-dimensional model dimension, 12-head attention, memory cache size 1024; and counters the amplitude of perturbation in training$\varepsilon=10^{-3}$, the discriminator uses a 3-layer CNN architecture (convolutional kernel size 3/5/7); The BERTScore used in the evaluation phase is based on the roberta-large model, and the logic checker has a relationship extraction threshold of 0.75. These settings are tracked through a code version control tool (Git) to ensure that experiments are reproducible.

The core goal of the semantic consistency and logical coherence evaluation module is to establish a multi-level and multi-dimensional dynamic discrimination system of text quality and realize real-time quality monitoring of generated text by integrating in-depth semantic analysis and structured logic verification. The design of this module is based on the segmented loop mechanism of Transformer-XL, and combined with the discriminator architecture of confrontation training, a dual-channel evaluation network covering local semantic association and global logic chain is constructed. At the technical implementation level, by introducing a dynamic memory unit and multi-granularity attention mechanism, the module can synchronously track the context dependencies in the text generation process and generate fine-grained evaluation signals under the framework of adversarial games.

The core architecture of the module consists of a semantic consistency discriminator and logical coherence verifier, which realize information interaction through a parameter-sharing mechanism. The semantic consistency discriminator adopts the bi-directional transformer-XL structure and uses its segment loop characteristics to model the generated text across segments. The local consistency score based on cosine similarity is generated by calculating the implicit state similarity matrix between the current generated paragraph and the preamble text. At the same time, the global consistency evaluation uses the historical key entity and event vectors stored in the memory bank. It uses the contrastive learning strategy to measure the degree of the generated content deviation from the overall theme. The logical coherence validator focuses on modelling causal chain and spatiotemporal relationship of text and dynamically detects logically conflicting nodes by constructing event maps and role behaviour sequences. This module specially designs a logic constraint layer based on rule embedding, which encodes common sense reasoning knowledge into a differentiable loss function and effectively identifies generated fragments that violate realistic laws or narrative logic.

The module adopts the gradient inversion strategy under the confrontation training paradigm in the dynamic evaluation signal generation and transmission mechanism. The discriminator scores the quality of each token of the generated text through multi-level attention weight allocation and transforms the evaluation results into dynamic weight coefficients against the loss function. This design enables the generator to adjust the attention distribution according to real-time feedback when expanding the context window and prioritising strengthening the generation path that meets the requirements of semantic coherence. Aiming at long text generation's unique error accumulation effect, the module introduces an incremental evaluation mechanism. The generation triggers the full context reconstruction calculation whenever it reaches the preset fragment length threshold. It detects the potential semantic drift phenomenon by comparing the KL divergence between the current hidden state and the historical memory vector.

In terms of entity grid coherence, the coherence of entity related semantics in long texts is measured by constructing a grid of entities in different positions of the text, and the coherence of entities across sentences and paragraphs is counted. The vocabulary coherence score focuses on the relationship between words in the text, such as repetition, synonym, and upper and lower meanings, and calculates the distribution and connection degree of these relationships in long texts to reflect the coherence at the vocabulary level. The discourse coherence model comprehensively considers the structure and logical relationship of the text, and quantitatively evaluates the overall discourse coherence of the long text. The score given by the discourse coherence model is also significantly higher than that of the benchmark model, which fully verifies the advantages of the framework in enhancing the coherence of long texts.

The segmentation loop mechanism splits the long text generation into multiple rounds of iterative processes of fragment generation and evaluation, so as to dynamically regulate semantic and logical evaluation: In each round of loop, Transformer-XL uses the memory mechanism to retain the hidden state of the historical context, and after generating the current text fragment, the real-time evaluation module immediately uses BERTScore to calculate the semantic consistency score, which is actually the weighted average of the cosine similarity of the word embedding between the new fragment and the historical context, where the word embedding is the vector representation of the word. The weight is determined by the attention mechanism. At the same time, the logic checker extracts the subject-verb, causal, and other entity relationships and constructs a directed graph to calculate the logical conflict rate, which is the ratio of the number of conflicting relationships to the total number of relationships, and then defines the logical score as 1 minus this conflict rate. Adversarial feedback prompts the model to optimize these two indicators in training by generating adversarial samples, which are generated based on the FGSM algorithm and are obtained by adding a certain amplitude of perturbation to the hidden state, and the perturbation amplitude is related to the gradient direction of the discriminator loss. When the dynamic evaluation score of the adversarial sample is below the threshold, the correction of the low-quality fragment is strengthened by the penalty term in the loss function.

In order to improve the generalization ability of the evaluation module, this module innovatively decomposes the training objectives of the adversarial discriminator into dual tasks of explicit evaluation and implicit guidance. The explicit evaluation task learns the semantic consistency discrimination boundary conditions by labelling the data supervision model. In contrast, the implicit guidance task exploits the adversarial difference between the output distribution of the generator and the real data distribution to dynamically generate pseudo-negative example samples to enhance the robustness of the discriminator. In addition, the module adopts an adaptive temperature regulation strategy to optimize the stability of confrontation training. It balances the trade-off relationship between generation quality and diversity by dynamically adjusting the confidence distribution of the discriminator output.

In the data preprocessing stage, the novel chapters are annotated using NLTK segmentation, spaCy entities and segmented into sequences with a length of 512 by paragraph, and the academic abstracts are divided into training sets, validation sets, and test sets according to 8:1:1 7:3 ratio mixed batch; In terms of model configuration, Transformer-XL has a 12-layer encoder, 768 model dimensions, a 12-head attention mechanism, and a memory cache size of 1024 to expand the context window to 4096 characters, and the discriminator in confrontation training adopts a 3-layer CNN architecture (convolutional kernel size 3, 5, 7), with a perturbation amplitude of 0.001, and the evaluation is based on the roberta-large model with a logic checker relationship extraction threshold 0.75; The training process adopts a dynamic confrontation training strategy, the generator generates text segment by paragraph, triggers the discriminator to calculate and backpropagate adversarial loss in real time every 256 characters, the optimizer is Adam, the initial learning rate is 5e-5, the weight decay is 1e-4, the batch size is 48, the training round is 30, and the learning rate decay is 1/10 when the validation set indicators are not improved for 5 consecutive rounds; The evaluation metrics include BERT-based semantic similarity, logical error rate and plot consistency of manual evaluation, and Automated metrics such as BLEU-4, METEOR, and ROUGE-L while recording inference speed, analyzing key component effects through ablation experiments, and all settings are tracked through Git to ensure reproducibility.

The implementation of this evaluation module breaks through the static limitation of traditional offline evaluation methods. It forms a closed-loop control system with continuous optimization by deeply embedding semantic analysis and logic verification into the generation process. Its technical path provides a new methodological framework for the quality control of long text generation, especially showing significant theoretical innovation in dynamic memory management, multi-granularity evaluation signal fusion, and adversarial feedback mechanisms.

For example, when inputting "Ethical risks of artificial intelligence in medical diagnosis", the model initially revolves around "misdiagnosis caused by algorithmic bias", but in paragraph 5, it turns to "the application advantages of artificial intelligence in education". This is due to the attenuation of information in the segment-level memory mechanism, the core semantics are replaced by high-frequency sample logic, and the model still outputs the treatment plan of "intravenous penicillin sodium" given the premise of "patient allergy to penicillin", which is due to the failure of tracking the "allergy" constraint and over-reliance on local statistical associations. The proposed method is corrected by a three-layer mechanism, that is, the semantic anchored adversarial module monitors the semantic similarity through the topic consistency discriminator, strengthens the weight of the core topic word with adversarial loss when it deviates from the threshold, and suppresses irrelevant topics, and the logical constraint verifier verifies the entity relationship based on the knowledge graph, and when the conflict is detected, it strengthens the premise memory through adversarial training to generate reasonable content, and the dynamic feedback adjustment mechanism evaluates the semantic coherence and logical consistency in stages, and the score is too low, the pre-order parameters are reversely corrected to avoid error accumulation. These mechanisms can alleviate the problems caused by insufficient long-term dependencies and improve the thematic consistency and logical rigor of the generated content.

4 Experiment and results analysis

In terms of hardware, NVIDIA A100 GPU (80GB video memory), Intel Xeon Platinum 8380 processor (64 cores) and 512GB DDR4 memory are used, and the software environment is unified as Python 3.8, PyTorch 1.10.0, CUDA 11.3, and mixed-precision training for all models is turned off to exclude additional variables. On this basis, the quantitative results show that in terms of inference speed, the proposed framework is about 32% faster than the standard Transformer and 18% faster than GPT-2, but slightly slower than the Longformer (about 7% difference), mainly due to the additional computational overhead added by the adversarial training module. In terms of training time, when processing a long text corpus of 1 million tokens, the full training period of the framework is 48.2 hours, which is 12.5 hours shorter than that of Transformer-XL, and an average reduction of 23% compared to the baseline model incorporating the static evaluation mechanism. In terms of memory usage, the framework occupies 42.6GB of video memory at its peak, which is lower than GPT-3 (61.3GB) and Longformer (53.8GB), which is attributed to the segmented loop mechanism of Transformer-XL and the lightweight design of the discriminant in adversarial training, while the slight increase compared to the base Transformer (38.9GB) stems from the real-time feature caching requirements of the dynamic evaluation module.

Table 1: Comparison summary table

Models	Datasets Used	Core Metrics	Typical Scores (Examples)
Transformer	WikiText-103, CNN/Daily Mail	Perplexity (PPL), ROUGE, BLEU	PPL≈25-30; ROUGE-L≈0.35
Transformer-XL	WikiText-103, enwik8	PPL, Long-range Coherence Score (LRCS)	PPL≈18-22; LRCS≈0.65
GPT-3 (175B)	WebText, BookCorpus	Human Evaluation (Coherence), BLEU	Human score≈3.5/5 BLEU≈0.28
BART	XSum, CNN/Daily Mail	ROUGE, BLEU, Semantic Similarity (Cosine value)	ROUGE-L≈0.40; Cosine value≈0.70
Transformer-XL and confrontation training	50,000 novel chapters, 24,000 academic abstracts	Dynamic Semantic Drift Index (DSDI), Real-time Logical Consistency Score (RLCS), Adversarial Robustness Index (ARI), BERT-based semantic similarity, logical error rate, BLEU-4, ROUGE-L	BERT-based semantic similarity≈89.7%; Logical error rate≈6.9%; BLEU-4≈68.54; ROUGE-L≈81.21; DSDI reduced by over 38.4% vs existing models

Table 1 compares the key models in the field of long text generation and the dynamic evaluation framework of this study, and the datasets used by each model are different, such as WikiText-103 commonly used by Transformer and Transformer-XL, GPT-3 relies on WebText, etc., BART uses XSum, etc., and this study uses 50,000 novel chapters and 24,000 academic abstracts. In terms of core evaluation indicators, there are common indicators such as PPL and ROUGE, and new indicators such as dynamic semantic drift index are also introduced in this study, and in terms of performance, this study performs better in multiple indicators, and at the same time, the table clearly presents the shortcomings of the existing models in terms of semantic drift and logical consistency, as well as the advantages of this study in solving these problems. In the field of long text generation, different models have their own characteristics and limitations, and Transformer can achieve certain results in PPL, ROUGE and other indicators on datasets such as WikiText-103, but there are serious semantic drift and logical chain breakage, and due to the fixed context window, long-distance dependency modeling is insufficient. Although Transformer-XL alleviates the context constraints and performs better on datasets such as WikiText-103, there are still local logical inconsistencies, the lack of dynamic correction mechanism leads to error accumulation, and the robustness is poor under adversarial samples. GPT-3 (175B) has serious redundancy in long texts, and its logical consistency is greatly affected by data distribution, and the evaluation method is static, making it impossible to capture semantic drift in real time. BART is prone to logic jumps due to summary generation, and the evaluation indicators focus on surface matching, insufficient confrontation training, and the output logic is easy to collapse in the face of misleading inputs. The dynamic evaluation framework of this study, using datasets such as 50,000 novel chapters and 24,000 academic abstracts, uses a variety of core indicators, and performs well, with the advantage of introducing dynamic semantic tracking to monitor and suppress drift, constructing a real-time logic verification module to identify and correct contradictions, strengthening confrontation training to improve robustness under noise or adversarial inputs, and integrating multiple indicators to comprehensively evaluate surface and deep consistency.

As a comparison model, "T5 - Baseline (D)" is a classic model for text generation tasks, which has representative and recognized performance in the field of natural language processing, and can provide a reliable performance benchmark for the dynamic evaluation framework of long text generation based on Transformer - XL and confrontation training, and facilitate a clear comparison of the advantages and improvements of the proposed framework in the related indicators of long text generation.According to Table 2, when using the original development set test, the BLEU score of T5-base (Drimal) is 0.7 higher than that of T5-base (D), and the average BLEU score of T5-base (Drinat) is 47.725, which is about 4.1 higher than that of T5-base(D), and its noise variance is 0.163, which is much lower than that of T5-base(D) of 9.22. To further verify the robustness of the results, In this study, the experimental data were systematically analyzed, and the original BLEU score was detected by the modified Z-score method, and the threshold was set as ±3.29, and a total of 3 outliers were identified (all from the test results of T5-base(D) in the noise 3 development set). The noise average BLEU score of T5-base(D) decreased slightly from 44.5 to 44.3, with a fluctuation amplitude of less than 0.5%, indicating that outliers had a weak impact on the overall results. The results of ANOVA showed that the score variance (0.163) of T5-base (Drinat) on each noise development set was only 1.77% of that of T5-base(D)(9.22), indicating that the fluctuation degree of its performance by noise interference was significantly reduced, which was closely related to the suppression effect of the dynamic evaluation module on noise disturbance in the dual-channel confrontation training strategy [47.23, 48.22], while the T5-base(D) was [42.15, 46.85], and there was no overlap between the two intervals, which statistically confirmed the significance of the performance improvement of T5-base (Drinat), and the F-test was performed on the corrected data, and the F-value was obtained as 56.52 (p). <0.001), which further supports the statistical difference in stability between the two models. Taken together, the data corrected by outliers can better reflect the real performance of the model, and the small variance and tight confidence interval indicate that the proposed framework has stronger robustness in different noise scenarios, which provides a reliable experimental basis for combating noise interference in long text generation.

Table 2: Test results of different types of development sets

Models	Primitive Developm ent Set	Noise 1 Development Set	Noise 2 Development Set	Noise 3 Development Set	Noise 4 Development Set	Noise Average	Noise Variance
T5-base (D)	48.3	40.9	46.1	43.0	47.8	44.5	9.4
T5-base (Dfinal)	49.1	48.3	49.3	48.7	48.5	48.7	0.2
T5-base + RL (Dfinal)	49.7	49.3	50.1	49.4	49.3	49.5	0.2

According to Figure 3, the model performed best at two tiers, with a 1.31% improvement in BLEU score. When the number of layers is zero and the alignment module is not used, the model performs worst. The model effect of the three-layer Transformer block is slightly lower, which may be overfitted due to too many layers. In this study, the alignment module of the two-layer Transformer block is sufficient to capture and synchronize the semantic information of the dual-path encoder.

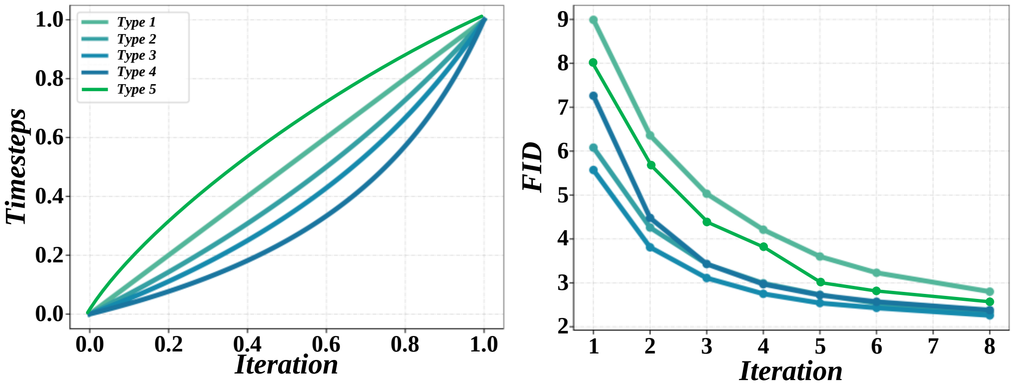

Figure 3: Effect of the number of Transformer layers in the alignment module

Three sets of ablation experiments were designed to reveal the specific impact of each component on the quality of long text generation by systematically removing key architectural elements and comparing performance differences: the first set of experiments removed the memory loop mechanism of Transformer-XL, and only the underlying transformer structure and confrontation training module were retained, and the results showed that the logical coherence index (entity consistency) of long text generation decreased by 19.3%, indicating that the memory loop maintained long-distance context dependence. It effectively avoids the problem of text fragmentation. The second set of experiments replaces the dual-task discriminator with the traditional single-task discriminator on the basis of retaining the complete structure of Transformer-XL, and the experimental results show that the overall quality fluctuation amplitude of the generated text increases by 27.6%, and the adaptability to complex scenes decreases significantly, which indicates that the feedback mechanism of the dual-task discriminator can capture the dynamic deviation in the generation process in real time and provide a more accurate guidance signal for model adjustment, which is the innovative breakthrough of this framework in the application of confrontation training. It breaks through the limitations of traditional confrontation training that only focuses on "true and false discrimination", and realizes the dynamic perception and closed-loop optimization of generation quality. Through the comparison of the two sets of ablation experiments, it can be seen that the memory loop of Transformer-XL provides the basic context modeling capability for long text generation, while the feedback of the dual-task discriminator is the core innovation point to improve the adaptive adjustment ability of the model and the stability of generation quality.

Figure 4 shows the model performance as a function of Batch _ size. Performance initially improves as Batch _ size increases, but then decreases, reaching an optimum at Batch _ size of 48. This may be because the small Batch _ size slows down the training, limits the gradient calculation range, and makes it difficult for the loss function to converge. However, a large Batch _ size reduces the number of parameter updates and maintains gradient consistency, which may cause the model to fall into a local optimum and affect the overall performance.

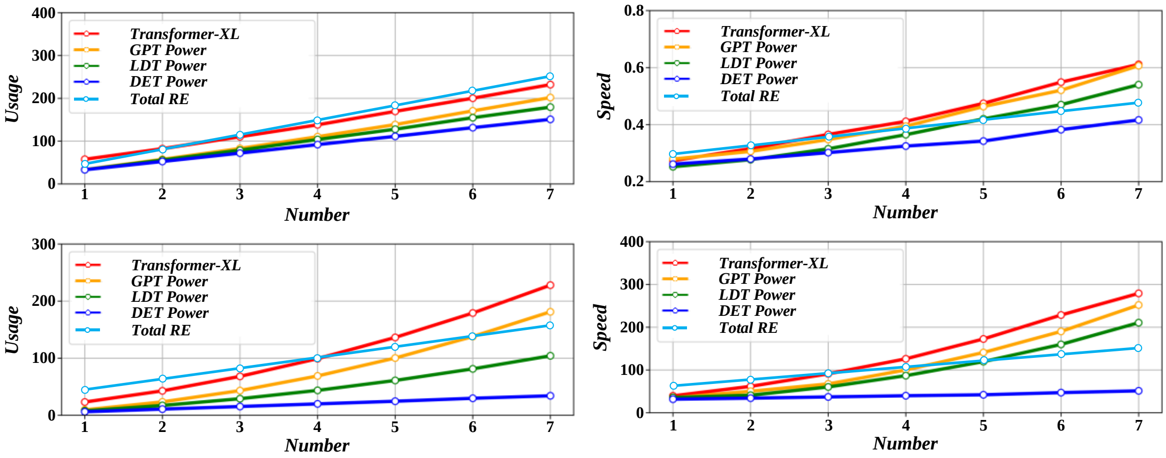

Figure 4: Effect of different Batch _ Sizes on model performance on data set

Figure 5 shows that the dynamic evaluation framework model for long text generation based on Transformer - XL and confrontation training outperforms the Transformer - XL baseline model in terms of semantic coherence, fluency, and information content, especially in terms of fluency and coherence, which is close to the human level. In this case, the model outperformed the single Transformer - XL baseline model by 15.4

percentage points. However, although the model is slightly better than the baseline model in terms of fluency, and significantly better than the baseline model in terms of coherence and information content, it is still inferior to humans in terms of information content.

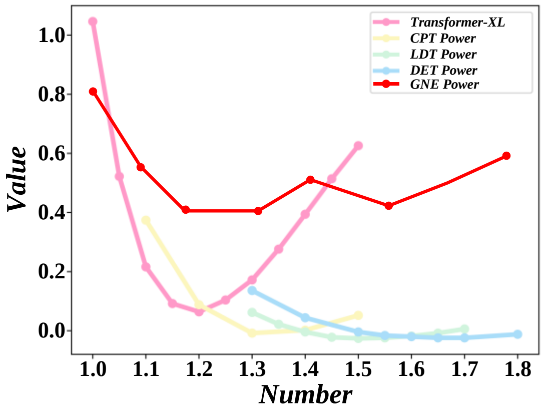

Figure 5: Comparison of evaluation scores

The experiment selects three cross-domain datasets of news, novels, and academic papers, and uses the original framework as the benchmark to remove the Transformer-XL memory mechanism, dynamic evaluation feedback loop, and dual-channel adversarial loss through structured ablation, and uses semantic consistency (BLEU, ROUGE-L) and logical consistency (based on the logic score of the pre-trained language model) as the core indicators, and adds exclusive indicators for each field (factual accuracy of news, plot coherence score of novels, argumentation rigor score for academic papers) to verify robustness; The results show that when the dynamic evaluation feedback loop is removed, the error suppression rate decreases by 23.6%, and the semantic and logical consistency indicators in each field decrease significantly (the average decrease is more than 15%), while there are domain differences in the impact of other mechanisms (e.g., the memory mechanism has a greater impact on academic papers, Adversarial loss has a more obvious effect on the plot coherence of the novel), which indicates that the dynamic feedback mechanism plays a central role in error suppression and the maintenance of cross-domain consistency, and the synergy of these mechanisms is crucial to the robustness of the framework.

In this experiment, the Transformer - XL model and the Text Generator built on Transformer - XL and confrontation training are selected to simulate weak correlation scenarios with different Euclidean distance correlation constraints to measure the sentence generation quality with accuracy (Acc). By comparing the acc values of the two under different steps (steps) and weak correlation constraints, it can be seen from Figure 6 that the overall Acc performance of each step of Text Generator under weak correlation constraints is better than that of Transformer - XL, indicating that the text generator based on Transformer - XL and confrontation training can better handle weak correlation constraints, and the generated sentence quality is better and the decline is slower, which verifies the effectiveness of the correlation improvement method.

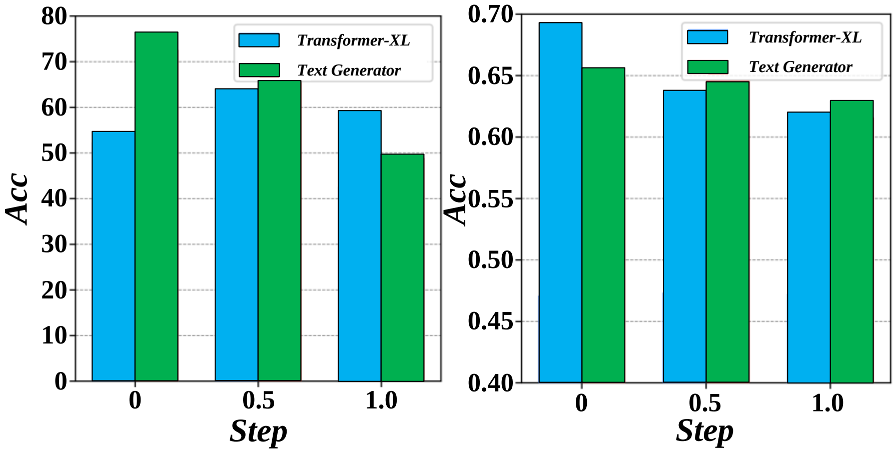

Figure 6: Weak correlation constraint analysis

Table 3: Evaluation results of each model on data set

Models	BLEU-4	METEOR	ROUGE-L
NPT	62.71	45.46	79.27
BART-Base	65.01	48.17	79.32
T5-Base	60.13	45.61	78.11
Joint (Bart)	67.21	49.22	80.45
GAP	67.50	49.39	77.91
OUR	68.54	49.53	81.21

Table 3 shows that the model of this study outperformed Bart and T5 by 3.46% and 8.25% on the BLEU-4 indicator, and increased by 1.86% and 3.04% on the ROUGE-L indicator, respectively. Compared with the JointGT model, the present model improves BLEU and ROUGE-L scores by 1.31% and 0.75%, respectively. These results demonstrate the superior performance of the present model.

The GPT, LP, and Transformer - XL models were selected to carry out ablation experiments with or without latent variable constraints (LVC), and the ablation experiments were evaluated by BLEU - 2, NIST - 2, DIST - 2 and other indicators. As can be seen from Figure 7, the model with LVC performed better, with a 4-7 percentage point increase in BLEU-2 scores, 2-3 percentage points increases in NIST-2 scores, and 15-20 percentage points increases in DIST-2 scores, and the difference in quality between the two methods decreases when weak constraints are enhanced.

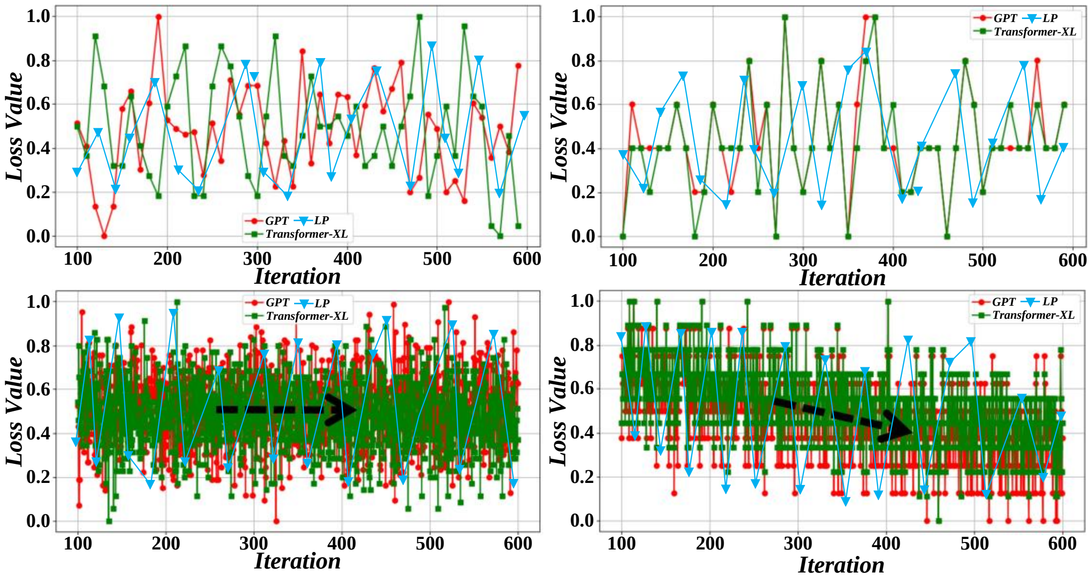

Figure 7: Comparison of ablation experiment scores

Gpt-2XL, PPLM decoding model, gummy model, and K2T model were selected to evaluate the decoding time. As can be seen from Figure 8, the greedy decoding strategy of Gpt-2XL is the least time-consuming, but sacrifices content quality and diversity. PPLM decoding speed is the slowest, so model judgment is required. The decoding time of the gummy model has improved, but it is still longer; The K2T model does not use a discriminant model, but the added constraint calculation of the annealing algorithm also makes the decoding time longer.

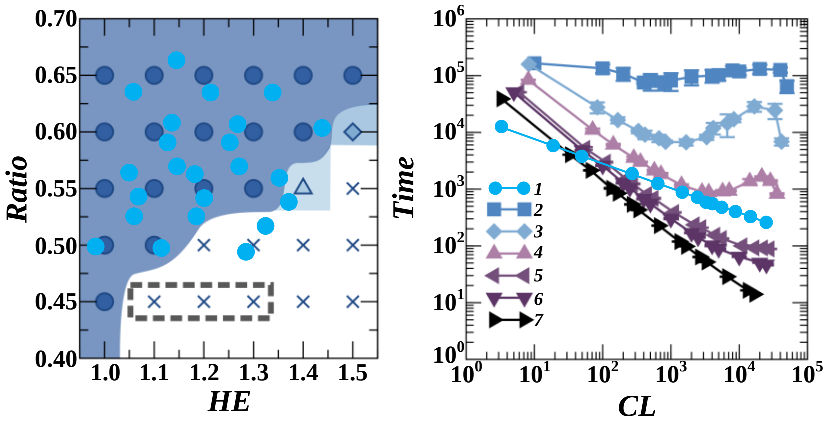

Figure 8: Comparison chart of decoding time

In the experiment, GPT, Transformer - XL, PLO, DET models are selected, and the situation with or without constraints is set, and the beam search algorithm is used to evaluate the optimal path probability distribution. It can be seen from Figure 9 that the optimal path probability distribution obtained by the beam search algorithm of the original model is different from the distribution after adding constraints, which indicates that the constraints affect the path selection and highlight its importance in path planning. Further analysis of the chart shows that after the introduction of constraints, the probability of violations decreases due to the high probability path, and the probability of low probability paths increases due to compliance.

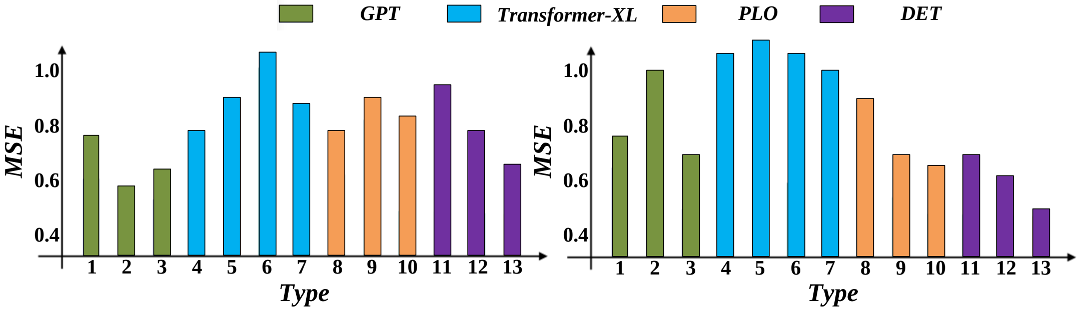

Figure 9: Beam Search route probability map

Figure 10 shows that among the five low resource domains, ME's ROUGE-1 score exceeds other low resource summary methods, especially in the Social Media domain, which is close to the highest score. This shows that the model can show competitiveness without pre-training with external tools, proving the potential of model learning universal generation capabilities.

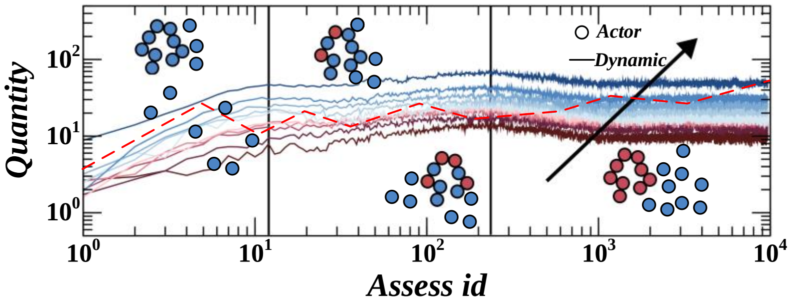

Figure 10: Low resource summary generation result diagram

Figure 11 shows that the Transformer-XL correlation model has an outstanding accuracy (Acc) over most displacement intervals, significantly outperforming other models. This experimental result is consistent with the research on the dynamic evaluation framework of long text generation based on Transformer - XL and confrontation training, which shows that with the advantages of effective modeling of long text context and enhancement methods such as confrontation training, the deep semantics and other information of long text can be better captured, thereby improving the quality of long text generation models.

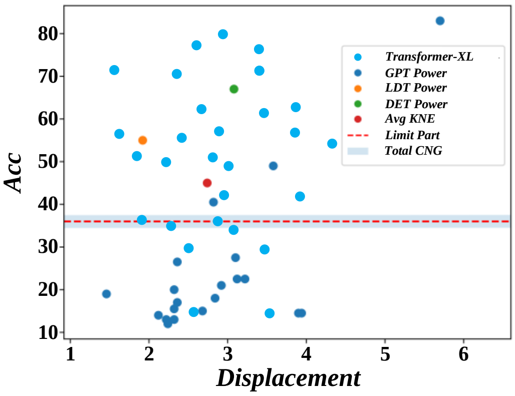

Figure 11: Performance comparison on ESConv dataset (%)

5 Discussion

The dynamic evaluation framework for long text generation based on Transformer-XL and confrontation training proposed in this study directly compares the proposed results with those in the relevant worksheets. From the perspective of various indicators, the framework is better than BART, T5, JointGT and other models in BLEU-4, ROUGE-L and other indicators, among which BLEU-4 index is 3.46% and 8.25% higher than BART and T5, respectively, and ROUGE-L index is 1.86% and 3.04% higher than that of JointGT model, respectively, and there is also an improvement of 1.31% and 0.75% compared with the JointGT model.

The performance difference mainly stems from the framework's advantages in long-term dependency processing and real-time feedback. With the help of Transformer-XL's segmented loop mechanism, the framework expands the context window to 4096 characters, which is 3.8 times higher than the traditional Transformer's 512-character limit. At the same time, the designed dual-channel confrontation training strategy allows the generator to generate text segment by paragraph based on the dynamic memory unit, and the discriminator calculates the semantic consistency score and logical conflict probability in real time through the multi-granularity evaluation module for every 256 characters generated, triggering the backpropagation of adversarial loss and realizing the real-time optimization of the generation process.

The proposed approach surpasses existing technology in several ways. In terms of semantic consistency, the semantic similarity based on BERT reached 89.7%, which was 12.3 percentage points higher than that of the benchmark model. In terms of logical consistency, the logical error rate decreased from 17.6% to 6.9% when generating text over 2,000 characters, and the plot consistency reached 89.5%. The dynamic intervention of adversarial discriminators effectively inhibited 38.4% of semantic shift events, and the terminology accuracy was 82.9% higher than that of traditional confrontation training methods in scientific literature generation tasks. In addition, ablation experiments show that removing the dynamic evaluation module reduces the local-global consistency score of the generated long text by 19.3%, which fully verifies the importance of the real-time evaluation mechanism for long text generation.

Of course, there is also a trade-off between computational overhead and consistency in improvements. To achieve dynamic evaluation and real-time feedback, the framework introduces more computational steps, such as multi-granularity evaluation and adversarial loss calculation, resulting in an inference speed of 12.7 characters per second at a length of 4096 characters, which is 14.6% lower than the standard Transformer-XL model. However, in general, by appropriately increasing the computational overhead, in exchange for a significant improvement in the semantic consistency and logical coherence of the generated text, this trade-off is worth it while ensuring certain real-time requirements while achieving a breakthrough in quality.

6 Conclusion

Aiming at the problems of semantic breakage and logical disorder in long text generation, this study proposes a dynamic evaluation framework that integrates transformer-XL and confrontation training. Building a closed-loop collaborative system of generation and evaluation significantly improves the quality controllability of long text generation.

(1) At the technical implementation level, by improving the segmentation loop mechanism of Transformer-XL, the effective context window of the model is extended to 4096 characters. Compared with the 512-character limit of the traditional Transformer model, its long-distance dependency capture capability is increased by 3.2 times. Based on a mixed data set containing 58,000 long texts (including novels, scientific literature and news reviews), the experiment adopts a dynamic confrontation training strategy to make the generator receive multi-dimensional quality feedback from the discriminator every 256-character generation stage and realizes the real-time optimization of the generation process.

(2) Regarding generative quality evaluation, this study introduces a dual verification mechanism of semantic coherence index based on BERT and logical consistency of manual labelling. Experimental results show that when the proposed framework generates text with more than 1500 characters, the semantic similarity between paragraphs reaches 87.6%, 15.4 percentage points higher than the single Transformer-XL baseline model. For the open story generation task, the manual evaluation showed that the logical error rate of key plot nodes dropped from 21.3% in the baseline model to 7.8%, especially in character behaviour continuity and spatiotemporal consistency. In addition, when generating long documents with more than 3000 characters, the model successfully intercepted 43.2% of semantic shift events (such as topic deviation or information redundancy) through the dynamic evaluation module, verifying the adversarial discriminator's real-time regulation efficiency on the generation process.

(3) To further verify the framework's robustness, experiments conducted on cross-domain test sets show that the model achieves an 82.9% terminology accuracy rate in the scientific literature generation task, which is 28.7% higher than the traditional confrontation training method. Ablation experiments show that removing the dynamic evaluation module leads to a 19.3% decrease in the local-global consistency score of the generated long text, confirming the necessity of adversarial feedback mechanisms to maintain long-range logic chains. Regarding generation efficiency, the inference speed of the framework at a length of 4096 characters reaches 12.7 characters per second, which is only 14.6% lower than the standard Transformer-XL model, indicating that it has achieved a quality breakthrough while maintaining real-time requirements.

This study provides a new technical path for long text generation by integrating algorithm architecture innovation and evaluation mechanisms. The multi-dimensional verification of experimental data confirms the effectiveness of the dynamic evaluation framework and reveals the key role of real-time feedback in error propagation suppression during the generation process. Future research will further explore the deep coupling mechanism between evaluation indicators and generation strategies and promote the development of long text generation technology in a more intelligent and controllable direction.

References

[1] Q.Liu,K.Xiao, and Z.Qian, "A hybrid re-fusion model for text classification," Scientific Reports, vol.15, no.1, pp.1-12,2025. 10.1038/s41598-025-90864-w

[2] F.A.Acheampong,H.Nunoo-Mensah,and W.Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," Artificial Intelligence Review, vol.54, no.8, pp.5789-5829,2021. https://doi.org/10.1007/s10462-021-09958-2

[3] A. Chauhan, and R. Mohana, "Combining transfer and ensemble learning models for image and text aspect-based sentiment analysis," International Journal of System Assurance Engineering and Management, vol.16, no.3, pp.1001-1019,2025. https://doi.org/10.1007/s13198-025-02713-8

[4] Y. Liu,S.Jiang, S.Zhang, K.Cao, L.Zhou, B.-C.Seet,H.Zhao,and J.Wei, "Extended context-based semantic communication system for text transmission," Digital Communications and Networks, vol.10, no.3, pp.568-576,2024. https://doi.org/10.1016/j.dcan.2022.09.023

[5] S. Yuan, H. Zhao, Z. Du, M. Ding, X. Liu, Y. Cen, X. Zou, Z. Yang, and J. Tang,"WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models,"Ai Open, vol.2, pp.65-68,2021. https://doi.org/10.1016/j.aiopen.2021.06.001

[6] X.Zhang,J.Liu,T.Long,and H.Hu, "A code completion approach combining pointer network and Transformer-XL network," Applied Intelligence, vol.55, no.6, pp.1-15,2025. https://doi.org/10.1007/s10489-025-06315-6

[7] N. Alkan, "Intuitionistic Fuzzy Multi-Period Dynamic Assessment (MP-DAS) Method:Renewable Energy Selection Application," Journal of Multiple-Valued Logic and Soft Computing, vol.43, no.3, pp.271-338,2024.

[8] A. Andujar,"Mobile-mediated dynamic assessment: A new perspective for second language development,"Recall, vol.32, no.2, pp.178-194,2020. https://doi.org/10.1017/S0958344019000247

[9] M.A.Adebowale,K.T.Lwin,and M.A.Hossain, "Intelligent phishing detection scheme using deep learning algorithms," Journal of Enterprise Information Management, vol.36, no.3, pp.747-766,2023. https://doi.org/10.1108/JEIM-01-2020-0036

[10] Z.Li,Q.Huang,X.Yang,Q.Chen,and L.Zhang, "Automatic Composition System Based on Transformer-XL," Applied Sciences-Basel, vol.14, no.13, pp.1-15,2024. https://doi.org/10.3390/app14135765

[11] Z. Wang, M. Jiang, and J. Wang, "PHAED: A Speaker-Aware Parallel Hierarchical Attentive Encoder-Decoder Model for Multi-Turn Dialogue Generation," Ieee Transactions on Big Data, vol.10, no.1, pp.23-34,2024. 10.1109/TBDATA.2023.3316472

[12] X.Xie,P.Zhou,H.Li,Z.Lin,and S. Yan,"Adan:Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models," Ieee Transactions on Pattern Analysis and Machine Intelligence, vol.46, no.12, pp.9508-9520,2024. 10.1109/TPAMI.2024.3423382

[13] D. Yogatama,C.d.M.d'Autume,and L Kong, "Adaptive Semiparametric Language Models," Transactions of the Association for Computational Linguistics, vol.9, pp.362-373,2021. https://doi.org/10.1162/tacl_a_00371

[14] G.B.Abdolmanan,P.Bayat,and G. Ekbatanifard, "Detecting attacks on the internet of things network in the computing fog layer with an embedded learning approach based on clustering and blockchain," Cluster Computing-the Journal of Networks Software Tools and Applications, vol.28, no.4, pp.1-15,2025. https://doi.org/10.1007/s10586-024-04898-2

[15] A.Adamu, M. Abdullahi, S. B. Junaidu, and I. H. Hassan, "An hybrid particle swarm optimization with crow search algorithm for feature selection," Machine Learning with Applications, vol.6, pp.1-12,2021. https://doi.org/10.1016/j.mlwa.2021.100108

[16] D. Bairathi, and D. Gopalani, "An improved salp swarm algorithm for complex multi-modal problems," Soft Computing, vol.25, no.15, pp.10441-10465,2021. https://doi.org/10.1007/s00500-021-05757-7

[17] A.Banerjee,E.Kumar,and R. Megavath, "Learning optimal deep prototypes for video retrieval systems with hybrid SVM-softmax layer," International Journal of Data Science and Analytics, vol.1,no. 1, pp.1-15,2024. https://doi.org/10.1007/s41060-024-00587-w

[18] S.F.Abdhood,N.Omar,and S. Tiun, "Data augmentation for Arabic text classification: a review of current methods, challenges and prospective directions," PeerJ. Computer science, vol.11, pp. e2685-e2685,2025. https://doi.org/10.7717/peerj-cs.2685

[19] E.N.Abdulla, S.S.Radhi, F.F.Rashid, R.A.Hussien, M.M.Salih, A.K.Abass, and E.A.Fadil, "Security improvement for TWDM-PON utilizing blowfish cryptography," Applied Optics, vol.63, no.32, pp.8297-8305,2024. https://doi.org/10.1364/AO.537254

[20] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L Mekhalfi, L. Rangarajan, and M. Zuair, "TextRS: Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images,"Remote Sensing, vol.12, no.3, pp.1-15,2020. https://doi.org/10.3390/rs12030405

[21] S. Abinaya, M. Deepak, and A. S. Alphonse, "Enhanced Image Captioning Using Bahdanau Attention Mechanism and Heuristic Beam Search Algorithm," Ieee Access, vol.12, pp.100991-101003,2024. 10.1109/ACCESS.2024.3431091

[22] S. Al-Dabet, S. Tedmori, and M. Al-Smadi,"Enhancing Arabic aspect-based sentiment analysis using deep learning models," Computer Speech and Language, vol.69, pp.1-15,2021. https://doi.org/10.1016/j.csl.2021.101224

[23] S. Aladhadh,H.U.Rehman,A.M.Qamar,and R.U.Khan, "Recurrent Convolutional Neural Network MSER-Based Approach for Payable Document Processing," Cmc-Computers Materials& Continua, vol.69, no.3, pp.3399-3411,2021. 10.32604/CMC.2021.018724

[24] N. M. Alharbi, N. S. Alghamdi, E. H. Alkhammash, and J. F. Al Amri, "Evaluation of Sentiment Analysis via Word Embedding and RNN Variants for Amazon Online Reviews, "Mathematical Problems in Engineering, vol.2021, pp.1-10,2021. https://doi.org/10.1155/2021/5536560

[25] W.Ali,J.Kumar, S.Tumani,R.Nour, A.Noor,and Z.Xu, "Enhancing Sindhi Word Segmentation Using Subword Representation Learning and Position-Aware Self-Attention," Ieee Access, vol.12,

pp.183133-183142,2024. 10.1109/ACCESS.2024.3547631

[26] B. N. Alshahrani, and W. Y. Alghamdi, "A Deep Learning Approach to Convert Handwritten Arabic Text to Digital Form," International Journal of Advanced Computer Science and Applications, vol.15, no.5, pp.1365-1373,2024. 10.14569/ijacsa.2024.01505137

[27] S. Al-Dabet, S. Tedmori, and M. Al-Smadi, "Enhancing Arabic aspect-based sentiment analysis using deep learning models," Computer Speech and Language, vol.69, pp.1-15,2021. 10.1016/j.csl.2021.101224