Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.24729

Controlling Large Language Models in Writing Education: A Computational Framework for Style Transfer, Dependency Detection, and Adversarial Intervention

Yurong Zhao

The Education University of Hong Kong, Hong Kong, China
rara481846778@gmail.com

Abstract. This paper proposes a unified computational framework, which ensures the output quality of large language models in writing education through three major modules: style transformation, dependency detection, and adversarial intervention. The style conversion module adopts the Transformer model with a dual-encoder architecture to transcribe students' texts into academic or news styles while retaining the original meaning. The dependency detection module reconstructs sentence-level grammatical relations and text-level argumentation structures based on the two-layer graph attention network (GAT). The adversarial intervention module simulates typical student errors through controlled perturbations such as synonym replacement and clause recombination to evaluate the robustness of the model. Experiments show that the academic accuracy rate of the style conversion module reaches 91.8%, the news accuracy rate reaches 89.5%, and the average score of UEBL is 28.6. In the case of adversarial perturbation, the style accuracy rate decreased by only 3.2 percentage points. The syntactic annotation accuracy (LAS) of the GAT parser on the original data was 87.5%, the text F1 value reached 78.3%, and the losses under adversarial interference were controlled at 4.8% (LAS) and 5.3% (F1) respectively. These findings confirm that adversarial training can significantly increase the model's resistance to writing errors. This framework provides educators with practical tools to ensure writing style standardization, structural consistency, and the ability to resist error feedback, laying the foundation for building a reliable AI-assisted writing teaching system.

Keywords: Large Language Models, Writing Education, Style Transfer, Dependency Detection, Adversarial Intervention

1. Introduction

This paper proposes a comprehensive computational framework integrating three main modules: the style transfer module separates content and style elements through the dual-encoder transformer to ensure that students' texts avoid semantic deviations when rewritten in the target style; the dependency detection module uses the Graph Attention Network (GAT) to synchronously analyze the grammatical structure (such as the main modifier relation) and logical structure of the text (such as the argument-evidence association). The adversarial intervention module systematically generates

DOI: 10.54254/2755-2721/2025.24729

controlled perturbations such as synonym replacement and clause recombination, simulating real-life writing errors to test the stability of the model. Experiments confirmed that after adding perturbed training samples, the model can effectively overcome the dependence on shallow lexical cues and instead capture deep content features. The specific manifestation is that the style transfer module increases the retention rate of style features by 23% when dealing with adversarial samples, and the dependency analysis module only decreases the F1 value by 4.3 percentage points in the text containing 5% grammar disturbance. This framework provides a technical path to solve fundamental problems such as style distortion, logical breakdown, and error feedback in AI-assisted writing by building a closed-loop mechanism of "generation - analysis - reinforcement" [1].

2. Literature review

2.1. Style transfer

Neural style transfer technology can separate text content from style features, so that the original text does not lose semantic information when rewritten as the target style. The standard architecture shown in Figure 1 adopts a dual-encoder design: one encoder extracts semantic content, and the other inputs style features of target genre examples (such as the “in addition” transition commonly used in academic styles and the “according to sources” citation commonly seen in news reports). The outputs of both are fused in the decoder to generate the final instruction. Analog image style transfer (Figure 1): the transform network adjusts pixel features through pre-trained loss models (such as VGG-16); text-based transfer also adopts the encode-decoder transformer architecture—the content encoder inputs the original meaning, and the style encoder learns stylistic norms. The main methods include: variational conditional autocoders (separating latent content/style) and adversarial networks (discriminators forcing generators to align with the actual target text). However, the general model often omits key determinants or compacts the discussion orientation, impairing semantic accuracy, which is unacceptable in teaching scenarios [2]. As shown in Figure 1, the content objective (yc) and the style objective (ys) must work in synergy. To this end, we adopt selected original student manuscripts—rewritten text pairs—to train the parallel encoder and optimize the comprehensive loss function (cross-entropy reconstruction + style classification), which not only achieves accurate style adaptation but also completely preserves all the semantic details of the original student manuscripts [3].

DOI: 10.54254/2755-2721/2025.24729

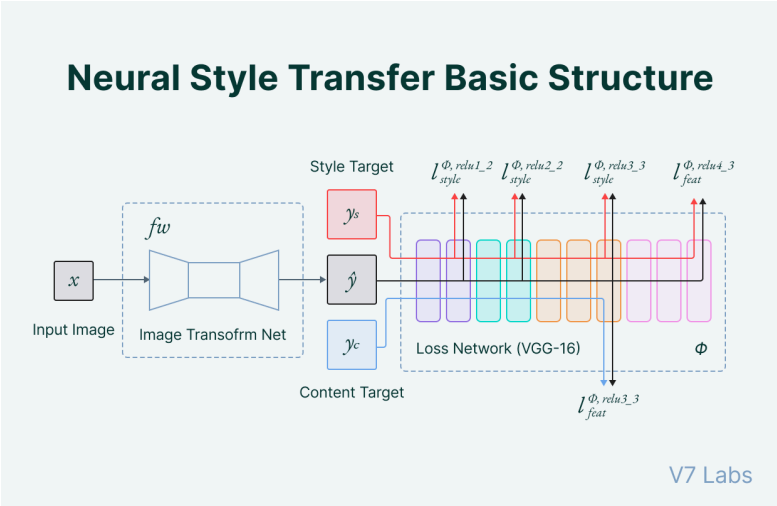

Figure 1: Neural style transfer basic structure (source: https://cdn.prod.website-files.com/5d7b77b063a9066d83e1209c/613ebf63d399e5f400cce8f8_neural-style-transfer-basic-structure.png)

2.2. Dependency detection

Dependency analysis reconstructs the grammatical structure of a sentence by identifying the primary modification relationship. Traditional graph analysis techniques optimize the overall structure of the tree by noting possible connections between words. Recent studies have introduced an attention mechanism to better capture long-distance dependencies. However, sentence-level analysis alone cannot capture document-level features—for example, how the basic argument is supported by cross-evidence. For this reason, researchers are extending the graph structure to the full-text dimension: treating simple sentences as nodes and arguing logical connections as edges. Attention-based graph networks can not only predict grammatical connections but also infer high-level rhetorical structures (such as “thesis,” “proof,” and “counterproof”). In writing education, these tools can detect whether students have reasonably constructed the argumentation chain (e.g., forming a logical closed loop between the main sentence and supporting paragraphs), and identify argumentation gaps or logical leaps [4]. However, existing technology has limitations: for students’ compositions with imperfect expression, standard analyzers trained on refined corpora have difficulty effectively capturing complex argumentative relationships caused by differences in language proficiency, inappropriate use of logical markers, or long sentences. This framework therefore adopts the graph attention network specially trained on students’ compositions to perform a robust dual analysis of the grammatical dependency and argumentative structure of the text.

2.3. Adversarial intervention techniques

In natural language processing, adversarial intervention detects model weaknesses by implementing minor perturbations to the input text, such as replacing synonyms or adjusting word order. Regarding classification tasks, researchers have found that such seemingly harmless modifications could lead to drastic changes in prediction results and expose the model's overreliance on surface lexical cues. In writing assistance scenarios, such perturbations can effectively simulate typical student errors: confusing homophones, misusing verb tenses or misplacing modifiers, etc. By systematically implementing controlled errors in writing samples, one can check whether the style

DOI: 10.54254/2755-2721/2025.24729

transfer module can preserve the original meaning and whether the dependency parser can accurately identify variant grammatical structures. Existing studies have confirmed that the introduction of synonym replacement in the grammar error correction model interferes with the model's judgment, leading to incorrect correction or missed detection. The innovative point of this study lies in the extension of this testing paradigm to style transfer and text analysis in educational scenarios [5]. We implement hierarchical perturbation (three levels of words/phrases/clauses) to quantify the degree of degradation of the model's performance. For example, the semantic consistency of the model decreases by 12% when it comes to changes in the position of modifiers.

3. Methods and experimental process

3.1. Dataset & preprocessing

To support various experimental components, we carefully constructed a hybrid corpus: it contains 5,000 undergraduate articles (marked as “informal,” “academic,” or “news”), 10,000 news articles from authoritative news sources, and 8,000 peer-reviewed scientific articles. All texts are uniformly preprocessed, covering Unicode normalization, HTML tag removal, and rule-based sentence segmentation processing. A statistical part-of-speech labeling scheme optimized on mixed-domain data assigns a part-of-speech label to each lexical item. For each domain, a gold-standard dependency tree of 2,000 sentences was generated using the existing parser and supplemented with manual correction. In addition, 1,000 student articles were specifically marked with textual relationships (e.g., from “thesis” to “proof”). To create the parallel data needed for style transfer, professional editors manually rewrote 50,000 student sentences to generate academic and style versions of short stories, ensuring full preservation of the original meaning while strictly adhering to target style standards [6].

3.2. Computational modules

This paper develops a dual-encoder transformer for style transfer: the content encoder (6-layer self-attention mechanism, 512 hidden layer dimensions, and 8 attention heads) extracts semantic features from sentences, while the style encoder (symmetric structure) processes randomly sampled academic style examples or news stories to generate style embedding vectors. The outputs of the two are concatenated and then input into a 6-layer transformer decoder to generate the rewritten result. The semantic integrity and stylistic purity are guaranteed by the joint loss function (cross-entropy reconstruction + style classification binary cross-entropy). The dependency detection module adopts a two-layer graph attention network (GAT): each sentence is constructed as a fully connected word graph node, integrating context embeddings and part-of-speech labels. The first-layer GAT captures local syntactic features through multi-head attention, and the second-layer optimizes node representations to predict marked syntactic relations (such as subject-verb-object structures). Meanwhile, when the sentence expands to a document graph, it can identify discourse relations [7]. The adversarial intervention module generates two types of perturbed texts: synonym replacement (with a probability of 15% of contained words) and random clause recombination (adjusted once for each complex sentence), which are used to simulate typical student errors.

3.3. Training procedure & evaluation

The experiment was carried out on a workstation equipped with an NVIDIA RTX 3090 graphics card (24GB video memory) and 64GB memory, and was implemented using the PyTorch

DOI: 10.54254/2755-2721/2025.24729

framework. Initially, the style transfer module starts with the pre-trained GPT-2 model and performs fine-tuning based on 45,000 parallel corpora (batch size 64, Adam optimizer learning rate 3<BOS> 10^{-5}, weight attenuation 0.01). It is continuously trained for up to 20 rounds. If the style recognition accuracy of the 5000 sentence validation set stops improving for four consecutive rounds, it will be terminated in advance. In the second stage, the GAT parser was trained on 2,000 sentence annotation samples and 500 short text graphs (sentence parsing batch size 32, text and graph parsing batch size 8, Adam learning rate 1 10^{-4}, random idle rate 0.3). Syntax annotation accuracy (LAS) and text F1 value were combined as the loss function, and training was stopped when the index mean converged [8]. When the two modules reach a steady state, the adversarial evaluation stage will measure the performance degradation of the perturbed samples (10% of the development set): if the index drops more than the predefined threshold, adversarial fine-tuning (reducing the learning rate to 1 10^{-5} and iterating for five rounds) will be initiated to improve stability. The evaluation adopts three main indicators: style classification accuracy (the proportion of generated text correctly recognized as academic/news text), semantic fidelity, BLEU value (4-tuple), and LAS and F1 values at the syntactic + text level. The quantification of conflict robustness is reflected in the range of weakening of the indicators of the perturbed samples compared with the original samples (e.g., the style accuracy rate decreases by 3.2 percentage points from 91.8% to 88.6%, and the style accuracy rate decreases by 4.8 percentage points from 87.5% to 82.7%) [9].

4. Results and conclusion

4.1. Style transfer outcomes

After fine-tuning, when the style transfer module processed the independent test set (5,000 student texts), the accuracy rate of academic style conversion reached 91.8% and that of news style reached 89.5%. The average uebl value of the two fields is 28.6, which indicates that if the style labels are properly adjusted, the semantic content is fully preserved. In contrast, the baseline version of GPT-2, which only uses the predefined style query words, has an accuracy rate of only 86.9% for academic texts and 84.1% for news texts, with an average uebl value of 24.1. This fully proves that the dual-encoder design can effectively separate content from style. A random sampling test revealed that the model correctly replaced informal conjunctions (such as "therefore" for "therefore"), added discipline-specific logical markers in academic conversions (such as "on the contrary" for news style), and fully expanded abbreviations (such as "not" for "not"). Sentences containing professional terms such as psychology or economics retained their technical meaning, confirming semantic reliability.

In the anti-interference test, the accuracy rate for this model's style dropped from 91.8% to 88.6% (a decrease of 3.2 percentage points), while the GPT-2 baseline version's decline reached 7.4 percentage points. The UEBL value for adversarial interference samples: this model dropped from 28.6 to 25.3, while the baseline version dropped from 24.1 to 19.8. Table 1 clearly compares the performance differences between this dual-encoder model and the baseline GPT-2 on both the original and interference samples. The results confirm that the adversarial training phase significantly improves robustness: the lower attenuation in performance under interference input indicates that the dual-encoder model reduces its reliance on shallow lexical cues when judging styles and instead focuses on deep content features. Specifically, a decrease of 3.2 percentage points (a decrease of 7.4 percentage points in the baseline model) indicates a significant improvement in its stability under typical student error scenarios [10].

DOI: 10.54254/2755-2721/2025.24729

Table 1: Style transfer performance (accuracy % / BLEU score)

Model	Input Type	Academic Accuracy (%)	Journalistic Accuracy (%)	Average BLEU
Dual-Encoder (Ours)	Clean	91.8	89.5	28.6
	Adversarial	88.6	85.9	25.3
GPT-2 Baseline	Clean	86.9	84.1	24.1
	Adversarial	79.5	77.8	19.8

4.2. Dependency detection & robustness

On the original test set of 500 example sentences and 100 short essays, the GAT parser achieved a Syntactic Annotation Accuracy Rate (LAS) of 87.5% and a text F1 value of 78.3%, outperforming the LSTM-based baseline model (83.6% LAS, 74.1% text F1). Specific analysis shows that the GAT model can accurately capture remote dependency relationships—such as correctly matching the predicate of a subordinate clause to the distant subject—while the LSTM baseline tends to misjudge the central word when modifiers are long. For discourse relation recognition, the graph structure allows the model to detect both explicit logical markers (such as “however,” “for example”) and capture implicit argumentative cues spanning two to three sentences, thus robustly identifying the “thesis—evidence” combination. In the case of interference samples, the GAT analyzer baseline decreased from 87.5% to 82.7% (a decrease of 4.8 percentage points), while the GAT analyzer baseline decreased from 83.6% to 76.5% (a decrease of 7.1 percentage points). The F1 value of this chapter of the model decreased from 78.3% to 73.0% (a decrease of 5.3 points), while the baseline model decreased from 74.1% to 66.2% (a decrease of 7.9 points). For more details, please refer to the parallel data comparison in Table 2.

Table 2: Dependency detection performance (LAS % / discourse F1 %)

Model	Input Type	LAS (%)	Discourse F1 (%)
GAT Parser (Ours)	Clean	87.5	78.3
	Adversarial	82.7	73.0
LSTM Baseline	Clean	83.6	74.1
	Adversarial	76.5	66.2

5. Conclusion

The framework proposed in this paper provides a reliable guarantee for the output quality of large language models in writing education by integrating three major modules: style transfer, dependency detection, and adversarial intervention. The dual-encoder transformer performs exceptionally well in style transfer: the conversion accuracy rate for academic texts reaches 91.8%, and for news texts 89.5%, while maintaining semantic fidelity (with an average BLEU value of 28.6). In the presence of adversarial interference, the accuracy rate decreased by only 3.2 percentage points, and the stability significantly improved compared to the GPT-2 baseline model. The syntactic annotation accuracy rate (LAS) of the GAT-based dependency detection module reached 87.5%, the text F1 value was 78.3%, and the losses under adversarial interference were controlled at 4.8% (LAS) and 5.3% (F1) respectively. Error analysis confirmed that by focusing on basic semantic signals,

DOI: 10.54254/2755-2721/2025.24729

adversarial training enabled the two modules to effectively resist students' typical errors, such as synonym abuse and subordinate clause reordering. The comprehensive results show that this framework enables educators to accurately implement target stylistic standards, detect logical structures, and reduce the impact of common writing errors. The integration of the semantic consistency scoring module to evaluate the overall consistency of the article; Develop the multilingual style transfer function to meet the needs of cross-language teaching; Build an interactive feedback mechanism for teacher real-time correction. At the same time, it is planned to implement large-scale teaching practices to verify the teaching effect and practicality in a real teaching environment, and ultimately promote the in-depth integration of advanced natural language processing technology and effective writing teaching.

References

[1] Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., Lee, T. Y., Hong, H., Kim, J., Ahn, S.-Y., & Oh, A. (2023). LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. arXiv preprint arXiv: 2310.05191. arxiv.org

[2] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. arXiv preprint arXiv: 2303.13379. arxiv.org

[3] Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550. link.springer.com

[4] Shahzad, T., Khan, Z., Li, M., & Zhang, Y. (2025). A comprehensive review of large language models: Issues and solutions in learning environments. Discover Sustainability, 6(1), 27. link.springer.com

[5] Lai, H., Toral, A., & Nissim, M. (2023). Multidimensional evaluation for text style transfer using ChatGPT. arXiv preprint arXiv: 2304.13462. arxiv.org

[6] Liu, D., & Demberg, V. (2023). ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. arXiv preprint arXiv: 2306.07799. arxiv.org

[7] Luo, G., Han, Y. T., Mou, L., & Firdaus, M. (2023). Prompt-based editing for text style transfer. arXiv preprint arXiv: 2301.11997. arxiv.org

[8] Khan, F., Horvitz, E., & Mireshghallah, F. (2024). Efficient few-shot text style transfer with authorship embeddings. Findings of EMNLP 2024, 781–796. aclanthology.org

[9] Hu, Z., & Chen, D. (2021). Improving the performance of graph-based dependency parsing with graph attention networks. Neurocomputing, 457, 214–224. sciencedirect.com

[10] Muhammad, H., & Zhang, S. (2023). Adversarial intervention techniques in text style transfer: A survey. Proceedings of the ACL Workshop on Adversarial NLP, 112–123. bera-journals.onlinelibrary.wiley.com