CAT-LLM: Style-enhanced Large Language Models with Text Style Definition for Chinese Article-style Transfer

ZHEN TAO, Renmin University of China, China
DINGHAO XI*, Shanghai University of Finance and Economics, China
ZHIYU LI*, Institute for Advanced Algorithms Research, China
LIUMIN TANG, Renmin University of China, China
WEI XU, Renmin University of China, China

Text style transfer plays a vital role in online entertainment and social media. However, existing models struggle to handle the complexity of Chinese long texts, such as rhetoric, structure, and culture, which restricts their broader application. To bridge this gap, we propose a **C**hinese **A**rticle-style **T**ransfer (**CAT-LLM**) framework, which addresses the challenges of style transfer in complex Chinese long texts. At its core, CAT-LLM features a bespoke pluggable **T**ext **S**tyle **D**efinition (**TSD**) module that integrates machine learning algorithms to analyze and model article styles at both word and sentence levels. This module acts as a bridge, enabling large language models (LLMs) to better understand and adapt to the complexities of Chinese article styles. Furthermore, it supports the dynamic expansion of internal style trees, enabling the framework to seamlessly incorporate new and diverse style definitions, enhancing adaptability and scalability for future research and applications. Additionally, to facilitate robust evaluation, we created ten parallel datasets using a combination of ChatGPT and various Chinese texts, each corresponding to distinct writing styles, significantly improving the accuracy of the model evaluation and establishing a novel paradigm for text style transfer research. Extensive experimental results demonstrate that CAT-LLM, combined with GPT-3.5-Turbo, achieves state-of-the-art performance, with a transfer accuracy F1 score of 79.36% and a content preservation F1 score of 96.47% on the "Fortress Besieged" dataset. These results highlight CAT-LLM's innovative contributions to style transfer research, including its ability to preserve content integrity while achieving precise and flexible style transfer across diverse Chinese text domains. Building on these contributions, CAT-LLM presents significant potential for advancing Chinese digital media and facilitating automated content creation. Source code is available at GitHub[1].

CCS Concepts: • **Computing methodologies** → **Natural language generation**; • **Information systems** → *Data stream mining*; • **Applied computing** → *Text editing*;

Additional Key Words and Phrases: Text Style Transfer, Large Language Models, Chinese Article, Text Style Definition

*Corresponding authors.
[1]https://github.com/TaoZhen1110/CAT-LLM

Authors' Contact Information: Zhen Tao, Renmin University of China, Beijing, China, taozhen@ruc.edu.cn; Dinghao Xi, Shanghai University of Finance and Economics, Shanghai, China, xidinghao@mail.shufe.edu.cn; Zhiyu Li, Institute for Advanced Algorithms Research, Shanghai, China, lizy@iaar.ac.cn; Liumin Tang, Renmin University of China, Beijing, China, tangliumin@ruc.edu.cn; Wei Xu, Renmin University of China, Beijing, China, weixu@ruc.edu.cn.

ACM 1557-735X/2025/8-ART111
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Text style transfer is a pivotal advancement within the domain of Natural Language Processing (NLP), serving as a bridge between content fidelity and stylistic transformation. The primary objective of this technique is to modify the stylistic attributes of a given text while meticulously preserving its original semantic content, enabling a seamless transition between different stylistic renditions of the same information [30, 35]. The applications of text style transfer are diverse, with notable progress in areas such as fake news detection [21], where it identifies stylistic inconsistencies, and social media [33], where it enhances user engagement by tailoring content to audience preferences. In addition, it plays a key role in cultural preservation by modernizing classical texts while retaining their stylistic essence [37], and in customer communication, improving businesses to adapt messaging for different audiences to improve user experience [13].

Before the advent of large language models (LLMs), the frameworks for text style transfer predominantly rely on small deep learning models. In accordance with the operational process of text transfer, these models can be broadly classified into two main categories: end-to-end models and two-stage models. Despite their innovative approaches, these earlier models often exhibited lower transfer accuracy and were primarily effective for relatively straightforward style transfer tasks, such as emotion [35], politeness [6], and etiquette [25]. A significant limitation in previous research was the absence of large-scale parallel corpora, which necessitated the use of non-parallel datasets to train classifiers for evaluating the transfer accuracy of unsupervised models. This constraint led to notable evaluation errors, as an author's style is frequently tied to specific thematic elements. For instance, the term "Camel Xiangzi" predominantly appears in the work of Lao She, making accurate evaluation challenging without parallel datasets. Moreover, the majority of existing studies have focused on style transfer at the sentences level in English, with limited research dedicated to longer Chinese texts. Given the distinct linguistic structures, idiomatic expressions, and cultural nuances inherent in Chinese, there is a compelling need to design models specifically tailored for the style transfer of extended Chinese texts. Figure 1 illustrates an example of Chinese article-style transfer, which transforms ordinary vernacular into a fragment of the article "The Scream".

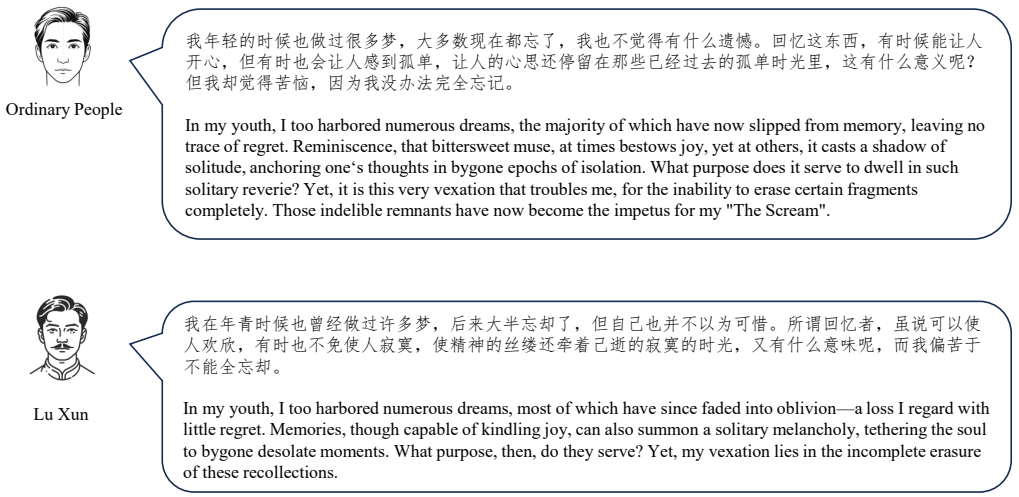


Fig. 1. An example that transfer styleless text to "The Scream" style text.

Recently, LLMs have demonstrated the ability to handle more complex NLP tasks such as summarization generation and role-playing through zero-shot transfer without the need for dedicated task-specific training [2, 28], attracting significant attention from both academia and industry. Given the strong capabilities of LLMs in semantic understanding and text generation, they can effectively address challenges faced by small models of style transfer, such as the lack of parallel data and low transfer accuracy. Currently, several studies have made significant headway in text style transfer leveraging LLMs. Lai et al. [14] employ ChatGPT as a multidimensional evaluator, assessing the style strength, content preservation, and fluency of text generated by various small-scale text style transfer models. Shanahan [23] and Wang et al. [33] investigate the use of LLMs for role-playing in conversational agents, where models like ChatGPT are tasked with generating text or dialogue in a specific role's style. A significant limitation of these methods is their difficulty in accurately mimicking the writing style of relatively unknown authors or anonymous works, often resulting in comprehension challenges and potential hallucinatory issues. In addition, the diverse and intricate expression methods in Chinese, such as idioms, rhetorical devices, and literary styles, pose significant challenges. Directly inputting works from a specified author into LLMs not only consumes more tokens, affecting inference speed, but also can lead to insufficient understanding of the author's writing style if only text segments are input. Moreover, this approach often lacks interpretability and guidance, making it difficult to capture the rich and diverse expressions inherent in Chinese article styles. Without clear insights into how stylistic features are identified and transferred, the outputs risk deviating from the intended style, underscoring the need for more nuanced methods that ensure fidelity while enhancing understanding of the transformation process.

To overcome the above challenges, we propose a **C**hinese **A**rticle-style **T**ransfer (**CAT-LLM**) framework based on LLMs. At the heart of CAT-LLM is a pluggable **T**ext **S**tyle **D**efinition (**TSD**) module that utilizes machine learning algorithms to analyze and model article styles at both the word and sentence levels. This module not only enables large language models to better capture the intricacies of Chinese article styles but also supports the dynamic expansion of internal style trees, facilitating seamless integration of new and diverse style definitions. Leveraging this comprehensive style definition, we design style-enhanced prompts to guide LLMs in generating high-quality, style-consistent text. Furthermore, we constructed large parallel datasets using ChatGPT, generating styleless text from ten distinct types of Chinese long texts to enable a more precise performance evaluation. These datasets represent a significant step forward in establishing a robust benchmark for text style transfer research. Extensive experiments demonstrate that our framework surpasses state-of-the-art research in both style accuracy and content preservation. The primary contributions are summarized as follows:

- We propose CAT-LLM, a pioneering Chinese Article-style Transfer framework to address the complexities of style transfer in Chinese long texts. Extensive experiments demonstrate that CAT-LLM achieves state-of-the-art performance in style accuracy and content preservation.
- We design a pluggable Text Style Definition (TSD) module that integrates various small models to analyze and model text styles at both the word and sentence levels. This module enhances LLMs' style transfer capabilities and allows dynamic expansion of style trees for seamless adaptation to diverse and evolving definitions.
- We create ten parallel datasets covering distinct types of Chinese long texts to facilitate robust and precise evaluation of style transfer models. This contribution establishes a novel paradigm for evaluating text style transfer and paves the way for future research.

2 Related Work

2.1 Large Language Models

The advent of Large Language Models (LLMs) has significantly transformed the landscape of Natural Language Processing (NLP), shifting the focus from traditional tasks like translation and classification to more sophisticated applications such as role-playing and dialogue systems [1, 4]. This marks a substantial enhancement in the scope and capability of NLP technologies, allowing for more complex language comprehension and generation tasks. LLMs like Baichuan [34] and ChatGLM [8] exemplify the cutting-edge applications of these technologies. Baichuan, developed to support a wide range of Chinese language tasks, excels in understanding and generating complex Chinese texts, making it a versatile tool for applications such as translation, sentiment analysis, and dialogue systems. It leverages extensive training on diverse datasets to provide high accuracy and contextual relevance, catering to the unique linguistic and cultural nuances of Chinese. ChatGLM is specifically designed for conversational AI, making it ideal for applications like customer service and virtual assistants. It generates human-like responses by ensuring coherence, relevance, and contextual accuracy, which enhances user engagement and satisfaction.

In terms of the openness of the LLMs source code, it can be broadly categorized into open-source and closed-source models [18]. Open-source models, such as LLaMA and its variants, have made high-performing NLP models accessible to a wider range of researchers and developers. LLaMA models range from 7B to 70B parameters and are trained on vast datasets, demonstrating competitive performance across various benchmarks [31]. Closed-source models, represented by OpenAI's GPT series, particularly GPT-4, push the boundaries of what LLMs can achieve. GPT-4, with over a trillion parameters, excels in diverse tasks including complex reasoning, coding, and multilingual understanding [1]. One of the critical innovations in leveraging LLMs is the design of effective prompts to guide their outputs. Shao et al. [24] propose Prophet to prompt LLMs with answer heuristics for knowledge-based Visual Question Answering. Jang et al. [10] investigate the effectiveness of negated prompts in directing LLMs, revealing limitations in their ability to follow such prompts accurately. Singh et al. [27] introduce a procedural LLMs prompt structure that enable LLMs to directly generate sequences of robot operations during task planning. Additionally, Li et al. [17] propose a directed stimulus prompt to guide LLMs in achieving specific desired outputs. In summary, the evolution of LLMs has opened new frontiers in NLP research, from enabling complex agent-level tasks to refining the art of prompt engineering. These advancements not only enhance the practical applications of NLP but also pave the way for more robust and interpretable AI systems.

2.2 Text Style Transfer

Text style transfer has been a focus in NLP research, attracting widespread attention in computer science and literature. Hu et al. [9] pioneer the introduction of the concept of style transfer from image to text, laying a cornerstone for subsequent explorations in text style transfer.

Recent advancements in NLP have led to the development of numerous text style transfer methods. These methods can be broadly categorized into two main families based on their transfer processes. The first family views content and style in a sentence as relatively independent elements, focusing on separating the original style from the content and replacing it with the target style. For example, Lee et al. [15] proposed a two-stage method: first, extracting and removing attribute markers using a pre-trained classifier, and then employing a Transformer-based generator to combine the target attribute with content tokens. Similarly, Tian et al. [15] utilized reverse attention to implicitly remove style information from each token, preserving the original content, and introduced conditional layer normalization to create content-dependent style representations. StoryTrans [39]

marked a significant advance in transferring long-form text. It used discourse representations to capture source content information and transform it into the target style through learnable style embeddings. To further enhance content preservation, the model employed a mask-and-fill framework, incorporating style-specific keywords from the source text into the generation process. However, these methods face challenges in maintaining the integrity of the original textual content, as certain tokens within sentences and paragraphs can embody both content and distinct stylistic characteristics. The second family focuses on designing end-to-end models for style transfer. The Style Transformer [5] leverages the robust capabilities of the Transformer architecture to handle long-term dependencies in text, directly providing the target style embedding to the decoder without the need for explicit disentanglement. StyIns [35] employs a generative flow technique to construct a distinctive and expressive latent style space, delivering robust style signals to the attention-based decoder supported by multiple style instances. Lyu et al. [19] proposed a diffusion-based language model capable of performing fine-grained text style transfer from scratch, utilizing a limited dataset without relying on external weights or tools. This model also supports multitasking and the composition of multiple fine-grained transfers. However, this family primarily addresses sentiment and tense transfer in individual English sentences, with limited research conducted on style transfer for longer Chinese texts.

Building on the advancements discussed, our work proposes a novel text style transfer framework based on large language models, specifically targeting long Chinese texts. This approach addresses the shortcomings of existing methods in terms of content preservation and stylistic accuracy.

3 Methodology

3.1 Overview of Task and Framework

To make our discussion clearer, we first define the task of Chinese article-style transfer. Suppose that there is a set of datasets $\{D_i\}_{i=1}^{K}$, each containing numerous textual fragments and sharing the distinctive style $\{v_i\}_{i=1}^{K}$. The task is to rewrite a article fragment x into a new fragment $x^{'}$ with the target style $v^{'}$. The process ensure that the new fragment $x^{'}$ possesses the target style $v^{'}$ while preserving as much information from the original fragment as possible.

Our proposed Chinese Article-style Transfer (CAT-LLM) framework is divided into eight stages. The specific structure of the framework is shown in Figure 2. In the first stage, we divide a certain corpus D_i into two parts: the style definition part D_{i1} and the style transfer part D_{i2}. In the second stage, we propose a Text Style Definition (TSD) module, which comprehensively considers style features based on the style definition part D_{i1} from the word and sentence levels. Given that the text acquisition in the style definition part is relatively random and extensive, we consider the summarized style features as the overall style features of the article. In the third stage, based on the text features extracted by the TSD module, we construct a comprehensive structured textual representation that encompasses features at both the word and sentence levels. In the fourth stage, we transform D_{i2} into styleless text $D_{i2}^{'}$ using ChatGPT to construct the parallel datasets. This strategy addresses issues such as missing labels and evaluation difficulties in unsupervised learning, providing an innovative approach to dataset construction in the era of large language models. In the fifth and sixth stages, we construct a style-enhanced prompt to guide LLMs in generating new text that is consistent with the original article style, while retaining styleless text information. In the final stages, newly generated texts $D_{i2}^{''}$ are compared with the original style transfer part D_{i2} and the corresponding styleless text $D_{i2}^{'}$ to comprehensively assess the performance of the style transfer framework, using both automatic evaluation and human rating. Our research primarily centers on TSD module design and prompt construction, which will be elaborated upon in the subsequent text.

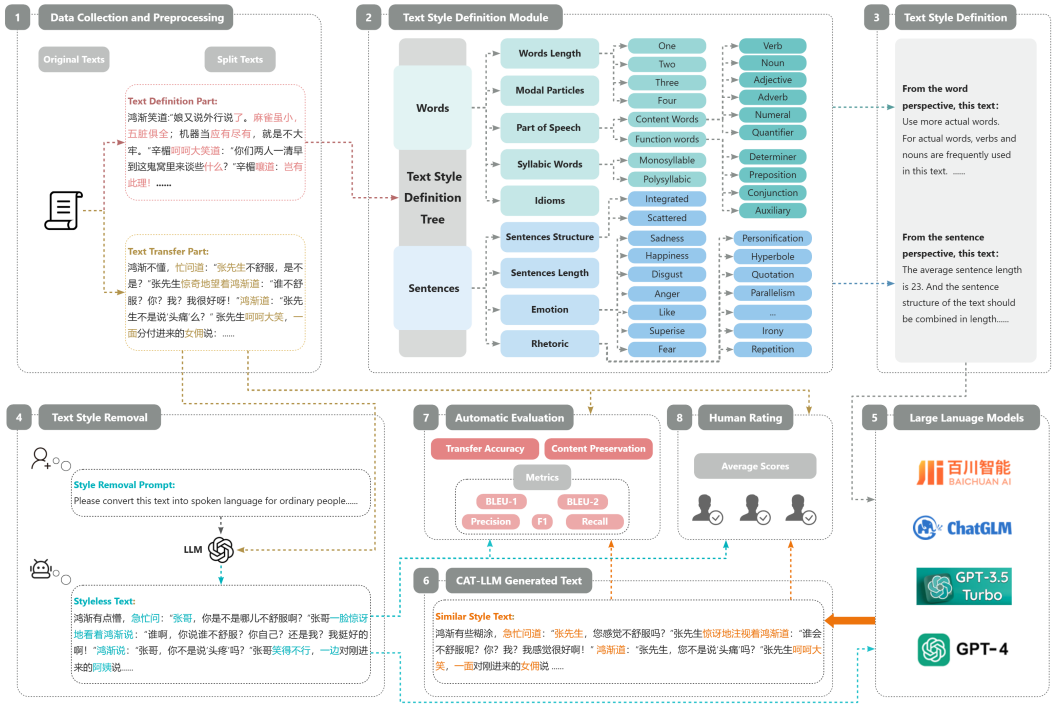


Fig. 2. Illustration of the proposed CAT-LLM framework. For the examples of Text Transfer Part, Styleless Text, and Similar Style Text in the figure, we have highlighted phrases that represent text styles with different colors to facilitate a better comparison of the text style transfer process.

3.2 Text Style Definition Module

As shown in Figure 2, we propose the Text Style Definition (TSD) module, which computes the style of the style definition part D_{i1} at both the word and sentence levels, providing precise style definitions. The TSD module is designed to enhance the accurate identification of text styles and to provide a clear and comprehensive prompt of styles for large language models (LLMs). Additionally, this module can be seamlessly integrated into various LLMs to optimize their performance in style transfer. Moreover, it offers interpretability for researchers, enabling them to gain deeper insights into the underlying mechanisms of text style recognition and transfer within the models.

3.2.1 ***Word Level.*** Words form the cornerstone of constructing text and play a crucial role in shaping language features [3, 22]. Quantitative analysis of the use of vocabulary in texts can reveal language style differences between different texts. In this section, we conduct an in-depth analysis of the words style of the article from multiple perspectives such as part of speech, words length and syllabic words, modal particles and idioms.

Part of Speech. To analyze the part-of-speech (POS) distribution in the text D_{i2}, we employ the *Cut* function from the *jieba*[2] library to tokenize the text and obtain POS annotations for each word. Let $W = (w_1, w_2, ..., w_i, w_{n_w})$ represent the sequence of the words, where n_w is the total number of words, and $T = (t_1, t_2, ..., t_i, t_{n_w})$ denote the corresponding sequence of POS tags. The POS tags are primarily categorized into content words, including nouns, verbs, and adjectives,

[2]https://github.com/fxsjy/jieba

and function words, such as adverbs, prepositions, and conjunctions. Through classification and statistical analysis of the vocabulary within an article, we can infer the distribution of part-of-speech categories.

To identify the predominant part-of-speech style of the article, we define the function $I_{ps}(\cdot)$ to count the occurrences of each POS type based on the tags T. The POS style f_{ps} of the article is determined by selecting the POS type with the highest frequency. This process is formalized as follows:

$$f_{ps} = \underset{i \in \{1,2,\ldots,n_w\}}{\text{argmax}} \{I_{ps}(w_i, t_i)\}, \tag{1}$$

where the function $I_{ps}(w_i, t_i)$ is defined as:

$$I_{ps}(w_i, t_i) = \sum_{j=1}^{n_w} \delta(t_j - t_i) \cdot \mathbb{1}(w_j = w_i), \tag{2}$$

and $\delta(t_j - t_i)$ is the Kronecker delta function which equals 1 if $t_j = t_i$ and 0 otherwise, and $\mathbb{1}(\cdot)$ is an indicator function that returns 1 when its argument is true and 0 otherwise. The goal is to maximize $I_{ps}(w_i, t_i)$, thereby determining the POS type t_i with the highest relative frequency in the text.

Word Length and Syllabic Words. The selection of word lengths by authors often reflects their personal style and linguistic preferences [16]. To elucidate the characteristic word length profile of a given text, we conduct a comprehensive statistical analysis of word lengths. This process is formalized as follows:

$$f_l = \underset{i \in \{1,2,\ldots,n_w\}}{\text{argTop}} K_l\{I_l(len(w_i))\}, \tag{3}$$

where $len(\cdot)$ is a function that computes the length of each word, and $I_l(\cdot)$ is an indicator function that enumerates the frequency of words with specific lengths. The top K_l word length categories, determined by their frequency, are then selected to represent the predominant word length characteristics of the text.

To further refine our analysis, we categorize words into monosyllabic and polysyllabic groups. We then identify the top K_{mo} and top K_{po} most frequently occurring words within these groups, respectively, to serve as representative syllabic word types. This is mathematically represented as:

$$f_{mo}, f_{po} = \text{argTop}K_{sy}\{freq(W_{mo}, W_{po})\}, \tag{4}$$

where W_{mo} and W_{po} denote the sets of monosyllabic and polysyllabic words, respectively, and the $freq(\cdot)$ is a function that computes the occurrence frequency of each word within these sets. By examining these distributions, we infer the author's stylistic inclinations toward the use of monosyllabic and polysyllabic words.

Modal Particles and idioms. The analysis of modal particles and idioms provides insights into the linguistic style and expressive tendencies of the text. Utilizing established modal particle[3] and idiom databases[4], we perform regular expression matching to extract the most frequently occurring modal particles and idioms. This method enables us to identify representative modal particles and idioms that characterize the text. The primary modal particles and idioms are formalized as follows:

$$f_{modal} = \underset{i \in \{1,2,\ldots,n_w\}}{\text{argTop}} K_{modal}\{freq(Re(w_i, D_{modal}))\}, \tag{5}$$

$$f_{idiom} = \underset{i \in \{1,2,\ldots,n_w\}}{\text{argTop}} K_{idiom}\{freq(Re(w_i, D_{idiom}))\}, \tag{6}$$

[3]https://baike.baidu.com/

[4]https://github.com/crazywhalecc/idiom-database

where D_{modal} and D_{idiom} represent the modal particle database and idiom database, respectively, and $Re(\cdot)$ denotes the regular expression operation used for matching. The function $freq(\cdot)$ calculates the frequency of each modal particle or idiom within the text. Identifying the top K_{modal} modal particles and the top K_{idiom} idioms based on their frequencies, we determine the predominant modal particles f_{modal} and idioms f_{idiom} that encapsulate the stylistic features of the text. This refined analysis not only quantifies the occurrence of modal particles and idioms but also provides a nuanced understanding of the author's linguistic choices and stylistic nuances.

After deriving various sub-features related to word characteristics, we integrate these features to construct a comprehensive representation of word attributes. This integrated approach allows for a holistic understanding of the textual style and linguistic nuances. To encapsulate the multi-faceted characteristics of words, we synthesize the previously computed sub-features into a unified feature representation. These features encompass part-of-speech, word length, syllabic words, modal particles, and idioms. This integration process is expressed as:

$$f_W = \text{Concat}\left(f_{ps} \oplus f_l \oplus f_{mo} \oplus f_{po} \oplus f_{modal} \oplus f_{idiom}\right), \tag{7}$$

where $\oplus$ denotes the concatenation operator. This detailed concatenation ensures that all relevant sub-features are comprehensively integrated into the final feature f_W.

3.2.2 ***Sentence Level.*** Compared to fine-grained analysis at the word level, a comprehensive examination of sentence style provides a more holistic understanding of the author's unique personality and characteristics in language expression [7, 11]. This approach contributes to a clearer perception of the overall style of the text, offering a broader perspective for a profound understanding of the article's stylistic nuances. Our analysis focuses on sentence style from the perspectives of sentence length, emotion, sentence structure, and rhetoric.

Sentences Length. To evaluate the average sentence length, we define sentence terminators as periods, question marks, and exclamation marks, using regular expressions to segment the text into a set of sentences $S = (s_1, s_2, ..., s_{n_s})$, where n_s denotes the number of sentences. Furthermore, we utilize regular expressions to match Chinese and English characters, yielding a valid character set C. By computing the lengths of the sentence set and the character set, the average sentence length of the article can be determined. This is formalized as:

$$f_{len} = len(C)/len(S), \tag{8}$$

where $len(\cdot)$ denotes the length of the respective sets.

Furthermore, to analyze the distribution of sentence lengths throughout the article, we classify each sentence as either short or long based on on a threshold parameter θ, which is set to 20 words following prior studies [12, 32]. This classification enables us to capture the tone conveyed by the sentence lengths. The classification process is represented as follows:

$$f_{ls} = \underset{i \in \{1,2,...,n_s\}}{\text{argmax}} \{I_{ls}(s_i)\}, \tag{9}$$

where $I_{ls}(\cdot)$ is a function that counts the occurrences of long and short sentences. The function $I_{ls}(s_i)$ is defined as:

$$I_{ls}(s_i) = \sum_{j=1}^{n_s} \delta(\text{len}(s_j) - \theta), \tag{10}$$

where $\delta(\text{len}(s_j) - \theta)$ is an indicator function that returns 1 if the length of sentence s_j exceeds θ words (classifying it as a long sentence), and 0 otherwise. If the article predominantly contains long sentences, it is classified as having a long sentence tone; conversely, if short sentences are more frequent, it is classified as having a short sentence tone. If the numbers are roughly equal, the article is considered to have a balanced combination of both long and short sentences.

Emotion. To analyze the emotional tone conveyed by the sentences, we utilize the ERNIE[5] pre-trained model, fine-tuned on the OCEMOTION dataset[6]. This model provides a more granular understanding of text semantics and emotional nuances by classifying emotions into seven fine-grained categories: *sadness*, *happiness*, *disgust*, *anger*, *like*, *surprise*, and *fear*. To improve the granularity and accuracy of emotion detection, we updated ERNIE's weights during fine-tuning on OCEMOTION. By aggregating the scores for each sentence, we determine the predominant emotion of the entire text. This process is formalized as follows:

$$f_{em} = \underset{j\in\{1,2,\ldots,7\}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n_s} I_{em}(\text{Emotion}_{\text{M}}(s_i, e_j)) \right\}, \tag{11}$$

where $\text{Emotion}_{\text{M}}(s_i, e_j)$ denotes the score of sentence s_i for the j-th emotion, M represent the ERNIE model,$I_{em}(\cdot)$ is a function that calculates the total score for each emotion category e_j. The summation $\sum_{i=1}^{n_s} I_{em}(\text{Emotion}_{\text{M}}(s_i, e_j))$ aggregates the scores of all sentences for each emotion. The emotion with the highest total score is identified as the dominant emotion of the text. By leveraging this comprehensive emotional analysis, we can gain deeper insights into the emotional tone and expressive nuances of the text, providing a robust foundation for understanding the author's stylistic tendencies.

Sentences Structure and Rhetoric. To evaluate the structure of integrated and scattered sentences, we employ a deep learning model to make classification decisions. Utilizing the powerful text generation capabilities of ChatGPT, we design prompts to generate a classification dataset containing integrated and scattered sentences. This dataset is then used to fine-tune the MacBERT model[7], resulting in a specialized Chinese sentence structure binary classification model $BERT_{st}$. Each sentence in the article is classified to assess the distribution of integrated and scattered sentences. If the number of integrated sentences exceeds that of scattered sentences, it can be inferred that the article predominantly consists of integrated sentences, otherwise it comprises more scattered sentences. This is formalized as:

$$f_{st} = \underset{i\in\{1,2,\ldots,n_s\}}{\operatorname{argmax}} \{I_{st}(BERT_{st}(s_i))\}, \tag{12}$$

where $I_{st}(\cdot)$ is a function that calculates the counts of integrated and scattered sentences.

Similarly, we classify the rhetorical devices used in sentences by generating sentences with rhetorical devices such as metaphor, personification, exaggeration, citation, parallelism, irony, rhetorical questioning, statement, and duality using ChatGPT to form a fine-tuning dataset. This leads to the development of the Chinese rhetorical multi-classification model $BERT_{rhe}$. The classification of rhetorical devices in each sentence is given by:

$$f_{rhe} = \underset{i\in\{1,2,\ldots,n_s\}}{\operatorname{argTop}} K_{rhe}\{I_{rhe}(BERT_{rhe}(s_i))\}, \tag{13}$$

where $I_{rhe}(\cdot)$ is utilized for the statistical counting of various types of rhetorical sentences. The top K_{rhe} rhetorical devices with the highest frequencies are identified as the main rhetorical features of the article. By leveraging these advanced models, we provide a detailed analysis of sentence structure and rhetorical usage, offering a comprehensive understanding of the text's stylistic attributes and enhancing the depth of linguistic analysis.

[5]https://github.com/PaddlePaddle/ERNIE
[6]https://aistudio.baidu.com/datasetdetail/100731
[7]https://github.com/ymcui/MacBERT

Subsequently, we integrate all sentence-level features to construct a comprehensive representation of sentence attributes. This integration is formalized as:

$$f_S = \text{Concat}\left(f_{len} \oplus f_{ls} \oplus f_{em} \oplus f_{st} \oplus f_{rhe}\right), \tag{14}$$

where $\oplus$ denotes the concatenation operator. Finally, we develop a holistic definition of the article's style by combining the features from both word and sentence levels. This comprehensive feature representation encapsulates the intricate stylistic elements of the text, enhancing our understanding of the author's unique linguistic expression. The final feature vector is constructed as:

$$F = \text{Concat}\left(f_W \oplus f_S\right), \tag{15}$$

where f_W represents the word-level features, f_S represents the sentence-level features.

3.3 Style-enhanced Prompt

At this stage, leveraging the article-style definition F generated by the Text Style Definition (TSD) module, we design a style-enhanced prompt to augment the zero-shot learning capability of large language models (LLMs) for Chinese article-style transfer.

The style-enhanced prompt comprises the task description, original text, style definition, and output indication. The complete format of the prompt is illustrated in Figure 3. Given that our text feature calculations primarily involve statistical analysis of vocabulary and syntactic structures, a mechanical enumeration of these structures may lead to misinterpretations by LLMs, resulting in unnecessary hallucinations. Therefore, we incorporate professional descriptions of feature structures to enable LLMs to better comprehend stylistic features. For instance, we might describe a style as containing many long sentences, which enrich the article with connotation, specificity in narrative, detailed reasoning, and emotional depth. For more specific details, please refer to the Appendix Figures A1-A10. Furthermore, considering the challenges LLMs face in understanding negative prompts, we meticulously design the task description to ensure that the generated text achieves an optimal balance between high transfer accuracy and content preservation, while avoiding the introduction of extraneous textual information. This approach helps maintain the integrity and authenticity of the original text, ensuring that the stylistic transfer is both accurate and contextually appropriate.

4 Experiment

4.1 Datasets

In our study, we carefully created diverse datasets consisting of texts from both literary and non-literary domains to ensure comprehensive validation of our framework.

For the literary works, we selected five seminal texts with distinct styles covering modern and contemporary China:

- **Fortress Besieged**, authored by Qian Zhongshu, is a modern novel characterized by humor and satire.
- **The Scream**, written by Lu Xun, exemplifies realism in modern Chinese literature.
- **The Fifteen Year of the Wanli Era**, by Huang Renyu, is a historical narrative.
- **The Family Instructions of Zeng Guofan**, crafted by Zeng Guofan, represents classical family instructions within ancient literature.
- **The Three-Body Problem**, authored by Liu Cixin, is a renowned science fiction novel.

To broaden the representativeness and diversity of our dataset, we incorporated five additional datasets from various non-literary domains:

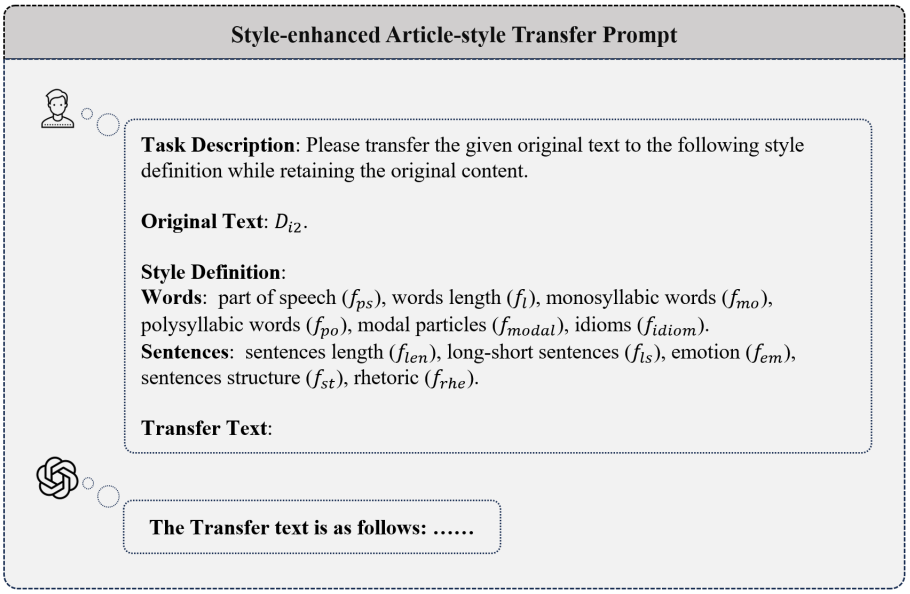


Fig. 3. The style-enhanced article-style transfer prompt.

- **News Dataset**: Sourced from *news2016zh*[8], this dataset includes modern Chinese news articles on topics such as politics, technology, and culture, representing formal and concise journalistic writing.
- **Encyclopedic Dataset**: The Chinese Wikipedia[8] corpus provides factual and objective content written in a formal, highly structured style typical of encyclopedic entries.
- **Social Media Dataset**: Sourced from Baidu's AI Studio platform[9], this dataset includes user-generated posts from Chinese social media platforms, reflecting informal, conversational, and often expressive styles.
- **Legal Documents Dataset**: The *CAIL2018*[10] dataset contains Chinese legal documents, such as court rulings, exemplifying technical and domain-specific writing.
- **Scientific Literature Dataset**: The *CSL3*[10] dataset includes abstracts and articles from scientific journals, showcasing a formal and technical writing style suited for academic communication.

We divide each type of text into two main parts: the style transfer part and the style definition part. We create ten parallel Chinese article-style transfer datasets by transferring the style definition part into styleless text using ChatGPT. The prompt used for ChatGPT is shown in Appendix Figure A11. The number of test samples for each book is approximately 400, with each sample consisting of 120 to 140 words. Table 1 presents detailed statistics of the datasets.

4.2 Implementation Details

For the algorithm parameters at the word level, we set $K_l, K_{sy}, K_{modal}, K_{idiom}$ to 4, 2, 6, 8, respectively. At the sentence level, we designate K_{rhe} as 2. The integrated and scattered sentences dataset we collected comprises 600 instances per class for binary classification, while the rhetoric dataset

[8]https://github.com/brightmart/nlp_chinese_corpus
[9]https://aistudio.baidu.com/datasetdetail/140467
[10]https://github.com/OYE93/Chinese-NLP-Corpus

Table 1. Statistical overview of ten different datasets.

Chinese texts	Numbers of style transfer examples	Numbers of style definition examples
Fortress Besieged	623	946
The Scream	217	284
The Fifteen Year of the Wanli Era	413	617
The Family Instructions of Zeng Guofan	302	1026
The Three-Body Problem	513	1762
News	413	516
Wikipedia	221	112
Social Media	330	324
Legal Document	439	455
Scientific Literature	401	407

consists of 600 instances per class for ten-class classification, encompassing nine rhetorical devices and a class without any rhetorical devices. We use these two datasets to fine tune BERT separately on four RTX 3090 GPUs. In the composition of text style definition F, the stylistic definition at the word level f_W is positioned before the sentences level style definition f_S. For the configuration of LLMs, the temperature parameter is set to 0.5, and the seed parameter is uniformly set to 42.

4.3 Evaluation Metrics

Our evaluation primarily focus on **style transfer accuracy** and **content preservation**. A high-quality Chinese article-style transfer model should balance these two aspects. We compare the generated text with the original text of the article to calculate the transfer accuracy. Simultaneously, comparisons between the generated text and the styleless text are conducted to assess the degree of content preservation. The evaluation metrics adopted for both perspectives are exactly the same. Specifically, we measure the lexical and semantic similarity between two texts using Bilingual Evaluation Understudy (BLEU)-n (n=1,2) [20] and BERTScore [36], where BERTScore mainly includes Precision, Recall, and F1 score. The specific calculation process is as follows:

BLEU: BLEU is a metric used to evaluate the quality of text by measuring the n-gram overlap between the generated text and reference texts. It is widely used for assessing the performance of machine translation and other text generation tasks, including style transfer. The general formula for BLEU is:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{16}$$

where BP (Brevity Penalty) is a factor that penalizes overly short translations. w_n is the weight for the n-gram precision, typically $w_n = 1/N$ when uniform weights are used. p_n is the precision for n-grams.

BERTScore: BERTScore leverages BERT embeddings to evaluate the similarity between generated text and reference text, capturing both lexical and semantic nuances. This metric is particularly suited for style transfer tasks where semantic fidelity is critical. The formula for BERTScore is:

$$\text{BERTScore} = \frac{1}{|T|} \sum_{t \in T} \max_{r \in R} \{\text{Sim}(\text{BERT}(t), \text{BERT}(r))\}, \tag{17}$$

where T is the set of tokens in the generated text, R is the set of tokens in the reference text, $\text{BERT}(t)$ represents the BERT embedding of token t, $\text{Sim}(a, b)$ is the cosine similarity between

Table 2. Specific parameter details for Chinese LLMs.

Chinese LLMs	Parm.	Type	Publisher	Release
Baichuan2	13B	Chat	Baichuan Inc.	2023.09
ChatGLM3	6B	Chat	Tsinghua	2023.10
GPT-3.5-Turbo	175B	Chat	OpenAI	2023.03
GPT-4-1106	NaN	Chat	OpenAI	2023.11

Note: NaN denotes no public data available.

vectors a and b, defined as:

$$\mathrm{Sim}(a, b) = \frac{a \cdot b}{\|a\|\|b\|}. \tag{18}$$

In addition, considering the randomness of LLMs generating text, we transfer each type of text ten times to calculate the final average metrics results.

4.4 Baselines

Considering that our algorithm primarily targets Chinese text, we have selected several state-of-the-art LLMs for generating Chinese text, including the GPT series, ChatGLM series, and Baichuan series. Detailed information on LLMs is shown in Table 2. The GPT series, developed by OpenAI, consists of a series of close LLMs, primarily including GPT-3.5 and GPT-4. These models are among the most widely used in the field. In this study, we utilized **GPT-4-1106-preview** and **GPT-3.5-turbo** as key components of our LLMs framework. The ChatGLM series of models were jointly developed by Tsinghua University and Zhichart AI Company. In this research, we utilized the **ChatGLM3-6B-chat** model to assist in generating Chinese text. The Baichuan series encompasses a range of advanced multilingual base language models developed by "Baichuan Intelligence". In this study, we employed the open-source **Baichuan-13B-chat** model.

In addition, to comprehensively evaluate our proposed framework, we conduct a comparative study on six state-of-the-art baseline models. **Style Transformer** [5] employs a transformer-based architecture to disentangle content and style, while **RACoLN** [15] utilizes reinforcement learning to balance style accuracy and content preservation. **StyleLM** [29] leverages a pre-trained encoder-decoder model fine-tuned with a denoising autoencoder loss to enable stylized text rewriting without parallel data. **SDR** [38] generates pseudo-dialogue pairs to train a stylized dialogue model for coherent, style-aligned responses. **DRAG** [26] introduces a director-generator framework to propose and refine style-aligned text variants, and **StoryTrans** [40] disentangles stylistic features from discourse representations, using a mask-and-fill framework to enhance content preservation. Through comparisons between the proposed CAT-LLM framework and these diverse state-of-the-art baselines, we can more accurately assess the strengths and weaknesses of CAT-LLM.

4.5 Automatic Evaluation Results

Our experimental section aims to evaluate the performance of different style transfer models, including the proposed CAT-LLM, models based on LLMs directly reading articles, role-playing under different LLMs, and six state-of-the-art baselines. The objective of these experiments is to explore how each model balances style transfer and content preservation and their ability to precisely control style transfer in generated text. To ensure the fairness of the experiments, we maintain consistency in the tokens used for reading articles in experiments involving LLMs directly reading the articles with the style definition tokens of CAT-LLM. This ensures a fair and reasonable

comparison between different models. As shown in Table 3, we present the experimental results of the models on three selected datasets: *Fortress Besieged*, *News*, and *Wikipedia*, chosen from the ten datasets. In addition, the experimental results of the LLMs on the selected *Social Media* and *Legal Document* datasets are presented in stacked bar charts shown in Figure 4, which provide a clear overview of the contributions of different metrics to the overall performance, as well as the overall performance of each model. The test results of all models on other datasets are detailed in Appendix Tables A1-A5. CAT-LLM, based on various LLMs, achieves a better balance between style transfer and content preservation in each LLM transfer, providing more precise control over style transfer in generated text. We used multiple datasets to evaluate the performance of the models, covering different styles and content to ensure the generalizability of the results. Evaluation metrics include the accuracy of style transfer, and the degree of content preservation.

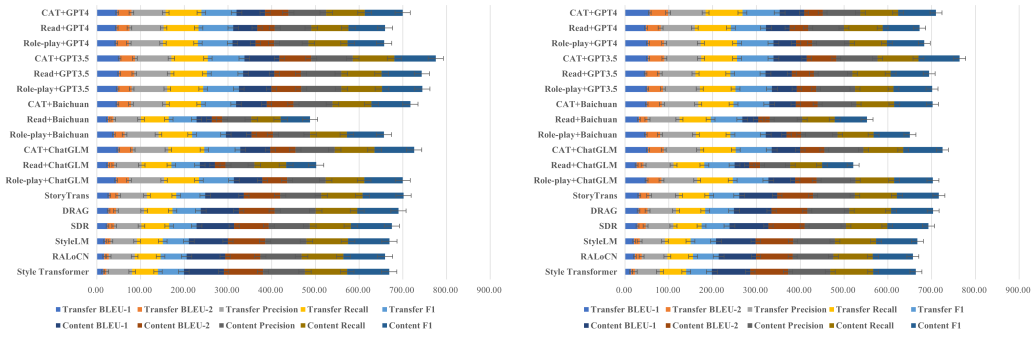
In terms of style transfer accuracy, the combined application of CAT+GPT-3.5-Turbo achieves the best performance on all datasets. In terms of lexical similarity evaluation index BLEU-1, CAT+GPT-3.5-Turbo has all reached about 50%, while maintaining a level of around 80% in semantic similarity. This success is not only attributed to the outstanding capabilities of GPT-3.5-Turbo but also highlights the effectiveness of the TSD module's style prompts, enabling CAT+GPT-3.5-Turbo to excel in article-style transfer. However, the experimental results of LLMs directly reading article fragments for style transfer remain unsatisfactory, with both style transfer accuracy and content preservation falling short of classic baselines such as Storytrans. This is mainly because the LLMs only summarize the style of certain fragments of the article and lack understanding of the overall style, illustrating the challenges LLMs face in handling more abstract and deeper semantic relationships, reasoning, and text comprehension. In terms of role-playing, LLMs excel at imitating writing styles of characters they are familiar with, while their performance significantly weakens when tasked with unfamiliar styles. As shown in Table 3, for news articles and Wikipedia texts, LLMs perform well in style transfer through role-playing. However, when asked to emulate the writing style of Qian Zhongshu's *Fortress Besieged*, the results of style transfer are notably diminished. Furthermore, the transfer accuracy of the six baseline models is relatively low. While we acknowledge the parameter discrepancies between LLMs and traditional methods, it is important to recognize that LLMs represent the latest advancements in NLP, and comparing them with older techniques underscores the progress made in the field. In the era of LLMs, the integration of such models into future developments has become a prevailing trend. Thus, comparing the transfer performance of both types of models remains of practical significance. Upon analyzing the text generated by the baseline models, we observe that these models demonstrate a tendency to replicate the input text, selectively substituting certain words, but fail to effectively capture the semantic and stylistic nuances of the original content. This is the primary reason for their lower transfer accuracy, despite maintaining relatively higher content preservation.

Additionally, as shown in Table 3 and Figure 4, GPT-4 does not perform as well as GPT-3.5-Turbo. For instance, in the style transfer of "Fortress Besieged", CAT+GPT-3.5-Turbo achieves a BLEU-1 score of 47.63%, whereas CAT+GPT-4 only achieves 35.12%. Similarly, in the "News" text, CAT+GPT-3.5-Turbo has a BERT-F1 score of 84.47%, while CAT+GPT-4 achieves 81.17%. This indicates that although GPT-4, as a more advanced model, is theoretically expected to perform better, its actual performance in style transfer tasks is not as good as anticipated, and it even falls short of GPT-3.5-Turbo. We speculate that this phenomenon may be due to several factors. Firstly, GPT-4, being a more advanced model, possesses higher intelligence and stronger language comprehension abilities. Therefore, during the generation process, it may adjust and optimize the input content to align with what it perceives as a more appropriate expression. This "creative freedom" enhances the readability and naturalness of the text but may lead to deviations from the required style definition in style transfer tasks, thereby affecting the accuracy of the transfer. Another possible reason lies in

Table 3. Automatic evaluation results on the "Fortress Besieged", "News", and "Wikipedia" datasets.

Chinese Text	Models	Transfer Accuracy(%)					Content Preservation(%)				
		BLEU-1	BLEU-2	Precision	Recall	F1	BLEU-1	BLEU-2	Precision	Recall	F1
Fortress Besieged	Style Transformer	17.80	11.03	55.45	58.79	57.08	**95.34**	**90.93**	**96.87**	**96.50**	**96.69**
	RALoCN	22.86	17.01	58.52	58.71	58.70	92.27	88.07	95.34	95.84	95.59
	StyleLM	20.54	16.91	56.40	57.09	56.78	93.29	88.43	95.12	95.38	95.25
	SDR	30.11	21.68	60.92	62.17	61.66	91.31	80.79	93.02	92.27	92.74
	DRAG	33.20	22.45	63.95	64.92	64.46	91.18	86.33	94.92	95.20	95.03
	StoryTrans	**34.17**	**23.55**	**67.82**	**68.01**	**67.93**	90.02	84.37	94.73	94.98	94.86
	Role-play+ChatGLM	33.37	18.36	71.18	72.26	71.68	56.85	50.51	81.58	82.70	82.10
	Read+ChatGLM	21.68	7.36	60.96	62.66	61.77	28.01	14.39	63.72	64.69	64.15
	CAT+ChatGLM	43.81	25.40	77.06	78.93	77.97	82.80	77.15	94.33	94.37	94.34
	Role-play+Baichuan	35.13	16.54	72.18	73.88	73.01	85.86	82.24	87.45	87.15	87.29
	Read+Baichuan	17.71	6.91	60.23	63.91	62.00	29.91	21.92	64.27	67.34	65.74
	CAT+Baichuan	42.61	25.07	76.11	77.17	76.59	83.58	80.13	93.59	92.71	93.11
	Role-play+GPT3.5	38.57	20.67	74.81	76.70	75.73	64.40	53.79	88.32	88.63	88.46
	Read+GPT3.5	16.59	6.76	59.77	63.87	61.74	27.03	19.05	64.21	68.09	66.07
	CAT+GPT3.5	**47.63**	**27.05**	**78.74**	**79.93**	**79.36**	**88.96**	**85.16**	**96.24**	**96.62**	**96.47**
	Role-play+GPT4	27.10	10.85	69.40	71.64	70.48	32.39	16.73	74.57	75.82	75.17
	Read+GPT4	28.85	13.40	69.38	73.15	71.14	43.68	28.45	74.87	79.15	76.91
	CAT+GPT4	35.12	17.23	70.68	73.22	72.08	50.09	34.62	82.47	82.95	82.67
News	Style Transformer	11.82	7.47	50.06	51.71	50.87	**90.93**	**87.47**	92.06	88.88	91.07
	RALoCN	12.34	11.55	58.75	55.83	57.45	89.67	85.77	88.36	80.97	85.02
	StyleLM	17.75	10.08	61.31	62.32	61.87	87.45	82.15	92.03	86.69	89.09
	SDR	21.80	14.83	67.82	66.69	65.74	84.00	81.28	92.45	86.86	89.70
	DRAG	24.93	21.14	71.26	73.90	72.47	79.86	68.76	**96.51**	89.41	93.66
	StoryTrans	**29.48**	**22.48**	**75.41**	**78.11**	**77.07**	81.94	68.32	96.11	**94.43**	**95.27**
	Role-play+ChatGLM	50.88	30.49	83.06	80.15	82.08	61.41	48.18	88.73	86.16	87.43
	Read+ChatGLM	31.32	20.04	68.68	76.58	72.35	38.40	31.47	71.64	83.38	76.98
	CAT+ChatGLM	54.66	37.27	83.23	82.31	82.75	71.83	61.34	90.59	91.25	90.89
	Role-play+Baichuan	38.37	24.13	79.30	73.68	76.14	43.06	30.25	81.81	77.19	79.16
	Read+Baichuan	26.39	9.38	63.70	64.03	63.86	29.41	14.31	64.85	66.11	65.46
	CAT+Baichuan	53.47	35.96	82.86	81.37	82.08	72.11	62.15	90.36	90.26	90.29
	Role-play+GPT3.5	51.49	34.73	83.39	80.93	82.11	62.08	49.14	88.80	87.55	88.14
	Read+GPT3.5	53.55	36.30	82.91	81.38	82.12	70.69	60.34	90.41	90.25	90.32
	CAT+GPT3.5	**57.91**	**40.41**	**85.14**	**83.83**	**84.47**	**83.16**	**76.54**	**95.28**	**95.47**	**95.37**
	Role-play+GPT4	50.30	32.49	81.09	79.96	80.49	54.90	38.71	83.67	83.95	83.78
	Read+GPT4	50.59	33.85	78.49	82.48	80.41	60.63	45.52	84.18	86.35	85.24
	CAT+GPT4	53.73	35.53	80.93	81.44	81.17	61.63	50.43	83.12	89.52	86.17
Wikipedia	Style Transformer	10.41	9.75	58.93	59.12	59.03	**85.43**	**83.29**	**98.19**	**98.09**	**98.14**
	RALoCN	18.39	11.72	62.65	62.74	62.69	82.03	80.32	95.08	95.81	95.45
	StyleLM	21.77	13.46	68.01	67.29	67.66	85.06	80.92	93.88	95.18	94.79
	SDR	23.21	18.06	74.88	72.34	73.59	84.93	78.48	92.78	91.88	92.15
	DRAG	25.74	21.38	77.47	77.77	77.62	83.56	79.62	95.92	94.98	95.40
	StoryTrans	**28.01**	**24.49**	**80.92**	**81.46**	**81.21**	82.35	77.62	94.55	93.49	93.92
	Role-play+ChatGLM	50.36	37.06	83.52	82.65	83.06	66.51	53.38	89.66	88.84	89.23
	Read+ChatGLM	36.66	23.71	69.63	75.99	72.61	51.52	45.33	74.88	83.82	79.03
	CAT+ChatGLM	52.93	35.70	83.05	82.24	82.63	68.17	55.89	90.52	89.69	90.09
	Role-play+Baichuan	51.12	33.34	81.52	81.57	81.51	56.96	41.15	84.78	85.00	84.86
	Read+Baichuan	37.98	22.57	73.23	71.85	72.50	43.97	30.61	76.43	75.16	75.76
	CAT+Baichuan	51.50	34.42	82.11	81.28	81.61	64.85	52.71	88.90	87.95	88.40
	Role-play+GPT3.5	55.89	38.88	84.55	83.73	84.11	65.48	52.75	90.00	89.23	89.60
	Read+GPT3.5	52.51	35.83	82.90	81.53	82.19	67.06	56.40	90.29	88.84	89.53
	CAT+GPT3.5	**56.94**	**39.24**	**85.84**	**84.95**	**85.43**	**72.53**	**63.17**	**93.25**	**92.22**	**92.72**
	Role-play+GPT4	49.46	33.65	82.16	79.91	80.89	55.01	41.91	86.07	83.86	84.82
	Read+GPT4	42.32	27.29	74.79	81.63	77.99	48.23	35.60	78.68	86.64	82.39
	CAT+GPT4	55.76	37.98	83.79	83.84	83.80	61.73	46.76	87.45	87.64	87.52

the differences in training data and optimization objectives. GPT-4 may have been exposed to more diverse and complex text data during training, with broader optimization goals that go beyond



(a) Social Media. (b) Legal Document.

Fig. 4. Stacked bar chart representation of various models' performance on "Social Media" and "Legal Document" datasets.

merely preserving and transferring text style. This results in GPT-4 focusing more on diversity and innovation in text generation in practical applications, at the expense of the precision required for style transfer tasks. Therefore, GPT-4's poor performance in style transfer tasks may be due to its higher language comprehension abilities and greater generation freedom. While these abilities are advantageous in many language tasks, they can become a disadvantage in tasks that require a high degree of style consistency.

With regard to content preservation, CAT-LLM has also achieved competitive results in the style transfer of various LLMs. In the experiment of directly reading articles with LLMs, we find the problem of introducing a large number of unrelated words, indicating that LLMs may be prone to hallucinations when dealing with complex problems. It is worth noting that the combination of our CAT-LLM also outperforms the Role-play+GPT-3.5-Turbo, indicating that the proposed TSD module can develop the enormous potential of LLMs in article-style transfer. The effectiveness of CAT-LLM in maintaining high content preservation is particularly evident across different datasets. For example, in the case of "Wikipadia" dataset, CAT+GPT-3.5-Turbo achieves a remarkable BERT-F1 score of 92.72%, significantly higher than other models such as Role-play+GPT-3.5-Turbo, which scores 89.60%. This indicates that while Role-play models focus on mimicking specific styles, they may sacrifice content accuracy, whereas CAT-LLM maintains a better balance. Moreover, the experiments reveal that smaller models such as Style Transformer and RACoLN, although showing higher content preservation scores, tend to merely replicate input texts with minimal stylistic changes. For instance, in "Fortress Besieged", the Style Transformer achieves a content preservation BERT-F1 score of 96.69%, but its transfer accuracy remains low, suggesting that it prioritizes content retention over stylistic fidelity. This further highlights the superiority of CAT-LLM, which not only preserves content but also effectively transforms style, achieving a more holistic transfer. Additionally, the hallucination problem observed in direct LLM applications underscores the necessity for mechanisms like the TSD module. By providing structured prompts and guidance, the TSD module helps mitigate the risk of generating irrelevant content, thereby enhancing the reliability of style transfer tasks. This capability is crucial for practical applications where maintaining the integrity and coherence of the original content is as important as achieving the desired stylistic transformation.

4.6 Human Evaluation

Considering the inherent limitations of automatic evaluation methods, such as BERTScore and BLEU, which often exhibit mechanical assessment tendencies and fixed evaluation standards, these methods cannot fully capture the nuances of language style. To address this, we conducted a comprehensive human assessment of our proposed models and other baseline models. In collaboration

Table 4. Human evaluation results on the "Fortress Besieged", "News", and "Wikipedia" datasets.

Chinese Text	Models	Transfer Accuracy	Content Presevation	Fluency
Fortress Besieged	Style Transformer	40	87	65
	RALoCN	45	85	67
	StyleLM	50	84	68
	SDR	55	88	70
	DRAG	60	90	72
	StoryTrans	70	91	74
	Role-play+ChatGLM	78	82	77
	Read+ChatGLM	80	83	78
	CAT+ChatGLM	85	84	80
	Role-play+Baichuan	75	81	75
	Read+Baichuan	72	79	74
	CAT+Baichuan	88	83	76
	Role-play+GPT3.5	82	88	85
	Read+GPT3.5	85	90	85
	CAT+GPT3.5	**93**	**96**	88
	Role-play+GPT4	90	94	92
	Read+GPT4	92	95	93
	CAT+GPT4	91	94	**95**
News	Style Transformer	42	83	65
	RALoCN	47	82	67
	StyleLM	50	81	68
	SDR	55	85	70
	DRAG	60	87	72
	StoryTrans	68	90	74
	Role-play+ChatGLM	72	84	77
	Read+ChatGLM	75	85	78
	CAT+ChatGLM	80	86	80
	Role-play+Baichuan	73	82	75
	Read+Baichuan	70	80	74
	CAT+Baichuan	83	85	76
	Role-play+GPT3.5	87	89	85
	Read+GPT3.5	90	91	85
	CAT+GPT3.5	**95**	**97**	88
	Role-play+GPT4	92	95	92
	Read+GPT4	94	96	93
	CAT+GPT4	93	95	**95**
Wikipedia	Style Transformer	40	80	65
	RALoCN	45	78	67
	StyleLM	48	77	68
	SDR	52	81	70
	DRAG	57	83	72
	StoryTrans	65	86	74
	Role-play+ChatGLM	70	80	77
	Read+ChatGLM	75	81	78
	CAT+ChatGLM	80	82	80
	Role-play+Baichuan	73	79	75
	Read+Baichuan	70	77	74
	CAT+Baichuan	83	81	76
	Role-play+GPT3.5	87	86	85
	Read+GPT3.5	90	88	85
	CAT+GPT3.5	**96**	92	88
	Role-play+GPT4	92	93	92
	Read+GPT4	94	**94**	93
	CAT+GPT4	94	93	**95**

with three PhD students in literature, we assessed texts generated by 18 transfer methods across the three datasets, using criteria from [35]. The three datasets are consistent with those presented in the automatic evaluation, facilitating a comprehensive assessment of the models, with specific details shown in Table 4. Our evaluation focused on transfer accuracy, content preservation, and fluency, with potential scores from 0 (the worst) to 100 (the best). The human evaluation results of the transferred texts for all models on other parts of the dataset can be referred in Appendix Tables A6-A8.

In particular, CAT-based models (CAT+ChatGLM, CAT+Baichuan, CAT+GPT-3.5-Turbo, and CAT+GPT-4) exhibit superior performance across all metrics compared to their counterparts. For instance, CAT+GPT-3.5-Turbo achieves the highest transfer accuracy scores of 95 and 96 on "News" and "Wikipedia" datasets, respectively. This can be attributed to GPT-3.5-Turbo's advanced capability to understand and internalize the style definitions summarized by the TSD module, leading to a more accurate execution of the style transfer task. These models also excel in content preservation, with CAT+GPT-3.5-Turbo scoring about 90, indicating their effectiveness in maintaining the original content while accurately transferring the target style. Furthermore, these models achieve high fluency scores, with CAT+GPT-4 reaching up to 95, demonstrating their ability to generate coherent and fluent language outputs. This consistent outperformance highlights the robustness of the CAT architecture in handling complex style transfer tasks.

4.7 Ablation Study

To further investigate the role and interaction of word and sentence level style definitions in the TSD module, we conducted ablation experiments on the "The Scream" and "Scientific Literature" datasets using GPT-3.5-Turbo. As shown in Table 5, word level prompts have a more significant positive impact on LLMs' understanding of text style than sentence level prompts. This may be attributed to the fact that word level prompts provide more granular and localized style information, enabling LLMs to more precisely capture subtle semantic nuances and stylistic variations in the text. In contrast, sentence level prompts may be more macroscopic and struggle to convey the finer stylistic differences within the text.

The ablation study further reveals that combining both word level and sentence level prompts yields the best performance, with the "Words+Sentences" arrangement achieving the highest scores in both transfer accuracy and content preservation. Specifically, the "Words+Sentences" combination results in a transfer accuracy BERT-F1 score of about 80% and a content preservation BERT-F1 score of about 95%, outperforming other configurations. This suggests that the synergistic effect of initially focusing on fine-grained semantics through word level prompts, followed by sentence level prompts, enhances the model's ability to maintain both the style and content of the

Table 5. Ablation study results on the "The Scream", and "Scientific Literature" datasets.

Chinese Text	Style Arrangement	Transfer Accuracy(%)					Content Preservation(%)				
		BLEU-1	BLEU-2	Precision	Recall	F1	BLEU-1	BLEU-2	Precision	Recall	F1
The Scream	Sentence	15.33	10.41	53.14	52.80	52.97	**91.94**	**86.42**	95.30	95.79	95.54
	Word	18.78	12.95	56.78	58.16	57.46	89.62	83.03	92.95	92.68	92.81
	Sentence+Word	25.55	16.97	62.11	62.96	62.53	86.38	78.29	91.66	90.22	90.93
	Word+Sentence	**46.78**	**26.53**	**79.41**	**80.36**	**79.92**	88.89	84.59	**96.16**	**96.42**	**96.31**
Scientific Literature	Sentence	17.82	12.88	56.91	57.25	57.08	85.72	**81.39**	92.08	91.64	91.86
	Word	22.12	15.97	60.38	61.59	60.98	**87.14**	80.76	90.43	90.12	90.27
	Sentence+Word	30.27	24.84	70.03	70.46	70.24	81.52	76.14	89.31	87.79	88.55
	Word+Sentence	**51.38**	**35.55**	**84.31**	**85.13**	**84.70**	80.82	73.51	**95.17**	**94.64**	**94.89**

original text. Moreover, placing word level prompts before sentence level prompts yields better results in style transfer experiments compared to the reverse order. This may be because word level prompts can guide the LLMs to initially focus on fine-grained semantics and local stylistic information in the text. This arrangement can learn and retain the subtle semantic differences in the text more effectively, thereby more accurately conveying the required style and enhancing the sensitivity of the LLMs to the overall style of the text.

These findings underscore how crucial prompt arrangement is in style transfer tasks. By combining both granular word level and holistic sentence level style cues, we can achieve more accurate and effective results in text generation models. This approach leverages the detailed, fine-grained semantics provided by word level prompts and the broader context captured by sentence level prompts, resulting in a more nuanced and precise style transfer. Such a strategy not only enhances the model's ability to preserve content but also ensures that the stylistic nuances of the original text are maintained, leading to superior overall performance.

4.8 Case Study

In this case study, Figure 5 shows different types of text generated in the experiment of the "Fortress Besieged" dataset, including Styleless Text, Storytrans, Role-Play+GPT-3.5-Turbo, Read+GPT-3.5-Turbo, and CAT+GPT-3.5-Turbo, and Ground Truth. For the demonstration of case studies on other datasets, please refer to Figures A12-A17 in the Appendix. We marked the words and sentences that can represent the text style with a yellow background to facilitate our observation of the models' transfer effect. Our analysis focuses particularly on the performance of CAT+GPT-3.5-Turbo, which demonstrates exceptional accuracy in style transfer and preservation of original content.

CAT+GPT-3.5-Turbo not only retains the relevant information from the Styleless Text but also closely matches the style of the Ground Truth. The Ground Truth passage features a dialogue between two characters, reflecting a humorous and light-hearted atmosphere. The text generated by CAT+GPT-3.5-Turbo accurately captures this atmosphere and maintains a very natural linguistic style. For instance, "Mr. Zhang looked at Hongjian with surprise and said: 'Who would be uncomfortable? You? Me? I feel very good!"' and "Mr. Zhang laughed and said to the maid who had just entered: 'Please go tell the wife and Miss that there are guests coming, and ask them to come out quickly. Hurry up!"' demonstrate the natural flow of conversation and the accurate portrayal of the characters' personalities. In contrast, texts generated by other models are somewhat lacking in maintaining the original intent and style. CAT+GPT-3.5-Turbo excels not only in precise word choice but also in better reflecting the tone and style of the original text, demonstrating a high level of text style transfer capability and completeness of original content. In general, CAT+GPT-3.5-Turbo showcases significant advantages in text style transfer in this case study, making it an excellent tool for research and practical applications.

4.9 Stylistic Feature Visualization

In our research on stylistic transfer, we follow the approach outlined by [39] to define and visualize stylistic features. Utilizing t-SNE for visualization, we provide a detailed comparative analysis of the stylistic features present in different sets of styleless texts and transferred texts, as shown in Figure 6. Specifically, we created four figures to illustrate these comparisons: (1) Styleless text and Ground truth text, (2) Styleless text and Role-play + GPT-3.5-Turbo text, (3) Styleless text and Read + GPT-3.5-Turbo text, and (4) Styleless text and CAT + GPT-3.5-Turbo text. The fourth figure, which compares the styleless text with the CAT + GPT-3.5-Turbo generated text, shows the closest resemblance to the Ground truth text. This suggests that the CAT + GPT-3.5-Turbo model does an excellent job of capturing and transferring the specific stylistic features of the source texts. In Figure 6(a) and 6(d), you can see that the stylistic elements in texts generated by CAT + GPT-3.5-Turbo

Styleless Text:
1鸿渐有点懵，急忙问：“张哥，你是不是哪儿不舒服啊？”2张哥一脸惊讶地看着鸿渐说：“谁啊，你说谁不舒服？你自己？还是我？我挺好的啊！”3鸿渐说：“张哥，你不是说‘头疼’吗？”4张哥笑得不行，一边对刚进来的阿姨说：“快去告诉太太和小姐，客人来了，让她们快点出来见见。快点快点！”5说着，右手的大拇指在中指上一弹，食指上‘啪’地一声响。

StoryTrans:
1鸿渐有点懵，急忙询问：“张哥，你是不是哪里不舒服？”2张哥惊讶地看着鸿渐说：“谁，你说谁不舒服？你自己？还是我？我挺好的！”3鸿渐说：“张哥，你不是说‘头痛’嘛？”4张哥一阵发笑，同时对刚进来的阿姨说：“快去告诉太太和小姐，客人来了，让她们快点出来见见。快点快点！”5说着间，右手的大拇指在中指上一弹，食指上‘啪’地一声响。

Role-Play+GPT-3.5-Turbo
1鸿渐略感困惑，急忙询问：“张先生，您是否感觉身体不适？”2张先生惊讶地望着鸿渐，答道：“何人感到不适？你？我？我身体状况良好！”3鸿渐继续道：“张先生，您不是提到了‘头痛’吗？”4张先生发出一阵笑声，同时对刚刚进来的女佣说道：“请速告知夫人与小姐，有客人到访，请她们尽快出来。速速！”5说着，他的右手大拇指在中指和食指之间轻轻一弹，发出了一声清脆的“啪”响。

Read+GPT-3.5-Turbo
1鸿渐有些愣住了，心跳不由得加快，急忙问：“张哥，你是不是哪里不舒服啊？”2张哥闻言，眼中带着一丝惊讶，愣了愣神，接着反问：“谁啊，你说谁不舒服？是你自己，还是我？我挺好的啊！”3鸿渐愣住了，心中有些茫然，皱着眉头：“张哥，你不是说‘头疼’吗？”4张哥听了，笑得几乎合不拢嘴，脸上流露出一丝戏谑，随即转向刚进门的阿姨，挥了挥手，像是没有半点犹豫：“快去告诉太太和小姐，客人来了，让她们快点出来见见。快点快点！”5话音未落，他那只修长的右手大拇指猛地在中指上一弹，食指一触即发，‘啪’地一声清脆的响声回荡在空荡的房间里，仿佛整个空气都为之凝固。

CAT+GPT-3.5-Turbo
1鸿渐有些糊涂，急忙问道：“张先生，您感觉不舒服吗？”2张先生惊讶地注视着鸿渐道：“谁会不舒服呢？你？我？我感觉很好啊！”3鸿渐道：“张先生，您不是说‘头痛’吗？”4张先生呵呵大笑，同时对刚进来的女佣说：“请速去告知太太和小姐，有客人来了，请她们迅速出来。快一点！”5说着，他的右手大拇指在中指和食指之间轻弹了一下，发出了‘啪’的一声。

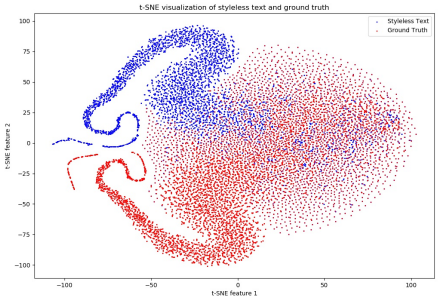
Ground Truth:
1鸿渐不懂，忙问道：“张先生不舒服，是不是？”2张先生惊奇地望着鸿渐道：“谁不舒服？你？我？我很好呀！”3鸿渐道：“张先生不是说‘头痛’么？”4张先生呵呵大笑，一面分付进来的女佣说：“快去跟太太小姐说，客人来了，请她们出来。makeitsnappy!”5说时右手大拇指从中指弹在食指上‘啪’的一响。

Fig. 5. Case study on the “Fortress Besieged” dataset.

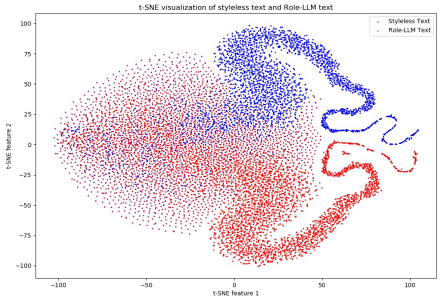
closely match those in the Ground truth texts. Texts generated by the Role-play + GPT-3.5-Turbo model also perform well, showing a slight difference compared to the CAT + GPT-3.5-Turbo model but still maintaining a high level of similarity to the Ground truth text. In contrast, the Read + GPT-3.5-Turbo model shows a more noticeable deviation in these stylistic features compared to the Ground truth texts. This indicates that while Role-play + GPT-3.5-Turbo can adapt styles almost as effectively as CAT + GPT-3.5-Turbo, the Read + GPT-3.5-Turbo model lags significantly behind. These findings are consistent with our automatic evaluation experiments, which also indicate that the CAT + GPT-3.5-Turbo model outperforms the other models in stylistic transfer.

4.10 Discussion

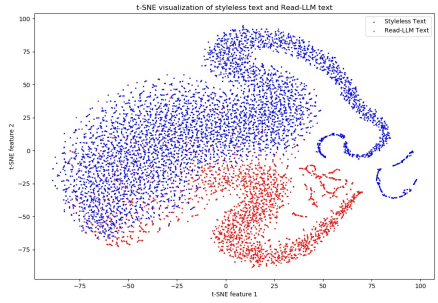
This study emphasizes the significant advantages of the CAT-LLM framework in achieving precise text style transfer, which comprehensively integrates word and sentence level style extraction algorithms. This clever combination enables the model to strike a delicate balance between preserving the original content and transforming its stylistic expression. By adopting this fine-grained approach, CAT-LLM is able to capture the inherent subtle differences in various Chinese text styles, including classical Chinese novels and modern social media texts. Through the meticulous design of prompts, CAT-LLM not only enhances stylistic consistency but also outperforms traditional methods and recent large language models in maintaining content integrity. The results of comprehensive evaluations, including automated assessments and human scoring, demonstrate the



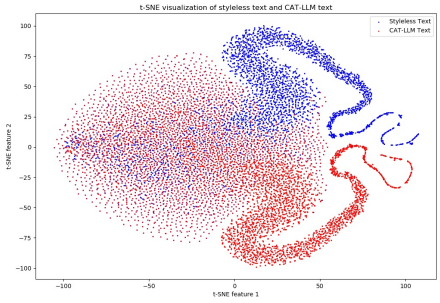
(a) Styleless text and Ground truth text.



(b) Styleless text and Role-play+GPT-3.5-Turbo text.



(c) Styleless text and Read+GPT-3.5-Turbo text.



(d) Styleless text and CAT+GPT-3.5-Turbo text.

Fig. 6. Stylistic features visualization of styleless text and transferred text on the "Fortress Besieged" dataset using t-SNE.

framework's ability to generate text that is stylistically accurate and content-preserving. These findings suggest that CAT-LLM holds considerable promise as a valuable tool for more sophisticated text generation tasks.

In addition to its strong performance in text style transfer, CAT-LLM is designed with real-world deployment and cost-efficiency in mind. Its modular architecture and the core Text Style Definition (TSD) module allow for flexible integration into various applications, such as digital media, content creation, and customer communication. The model's ability to deliver high-quality, stylistically consistent output can be leveraged across industries, from enhancing the productivity of journalists and writers to adapting educational content for modern audiences. Moreover, CAT-LLM is computationally efficient, relying on smaller, pretrained models like ChatGLM (6B) for fine-tuning, reducing the computational costs typically associated with large language models. This approach, combined with cost-effective third-party APIs like GPT-3.5-Turbo and GPT-4, ensures that CAT-LLM remains accessible to a wide range of industries. Looking ahead, future improvements such as knowledge distillation and lightweight fine-tuning will further enhance its scalability and affordability, making CAT-LLM a practical and sustainable solution for diverse text generation tasks.

While the CAT-LLM framework shows promising results in text style transfer, there are several limitations that warrant further exploration. One key limitation is that the current model is optimized for Chinese text, which restricts its applicability to other languages. Expanding its capabilities to support multilingual text styles would significantly broaden its potential. Additionally, although the model effectively preserves content integrity, challenges remain in minimizing the introduction of irrelevant or hallucinated content, especially when dealing with more complex stylistic nuances. Future research should also aim to refine evaluation metrics by incorporating factors such as readability and creativity, which would offer a deeper understanding of the generated text's quality. Finally, integrating domain-specific knowledge into style prompts could improve precision in specialized fields such as technical writing or literary analysis, improving the applicability of CAT-LLM to various tasks.

5 Conclusion and Future Work

In this paper, we propose the **C**hinese **A**rticle-style **T**ransfer (**CAT-LLM**) framework , a novel approach to addressing the challenges of style transfer in complex Chinese long texts. By integrating a specialized Text Style Definition (TSD) module, CAT-LLM effectively analyzes and adapts to the nuanced stylistic features of Chinese texts at both the word and sentence levels. This module bridges the gap between large language models (LLMs) and the intricate rhetorical, structural elements of Chinese writing. To facilitate robust evaluation, we constructed ten parallel datasets using ChatGPT and various Chinese texts, each corresponding to distinct styles of Chinese articles. These datasets provide a novel evaluation paradigm for style transfer research. Extensive experimental results show that CAT-LLM achieves a transfer accuracy F1 score of approximately 80% and a content preservation F1 score of around 95% across most datasets, significantly outperforming existing methods in balancing style transfer accuracy with content preservation.

Despite the promising results demonstrated by CAT-LLM, there are still several areas that require further refinement. Future work will focus on expanding the framework to support multilingual text style transfer, which would allow the model to process a wider range of languages and enhance its applicability. Additionally, addressing challenges related to content preservation while adapting more complex stylistic features, such as emotional tones and domain-specific attributes, will be a priority. To improve the evaluation process, future research will incorporate additional metrics, such as readability and creativity, to provide a more comprehensive assessment of the model's performance. Lastly, integrating domain-specific knowledge into style prompts will help to enhance the performance of CAT-LLM in specialized applications, such as technical writing or literary analysis, ensuring its effectiveness in a wider range of tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (72271233), the Fundamental Research Funds for the Central Universities (2024110591), Beijing Municipal Science and Technology Project (Z241100001324020), Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies (SZS2023007), Smart Social Governance Technology and Innovative Application Platform (YZCXPT2023101), and The Innovation System of the Integration between Industry and Education for Smart Governance (CJRH2024101).

References

[1] Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361* (2023).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural*

information processing systems 33 (2020), 1877–1901.
[3] Cristina Castillo and L. Tolchinsky. 2018. The contribution of vocabulary knowledge and semantic orthographic fluency to text quality through elementary school in Catalan. *Reading and Writing* 31 (2018), 293–323.
[4] WX Che, ZC Dou, YS Feng, et al. 2023. Towards a comprehensive understanding of the impact of large language models on natural language processing: challenges opportunities and future directions (in Chinese). *Sci Sin Inform* 53 (2023), 1645–1687.
[5] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5997–6007.
[6] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 250–259.
[7] S. Gajda. 2022. The faces of style and stylistics. *Journal of Linguistics/Jazykovedný casopis* 73 (2022), 7–26.
[8] Zhenyu Hou, Yiin Niu, Zhengxiao Du, Xiaohan Zhang, Xiao Liu, Aohan Zeng, Qinkai Zheng, Minlie Huang, Hongning Wang, Jie Tang, et al. 2024. ChatGLM-RLHF: Practices of Aligning Large Language Models with Human Feedback. *arXiv preprint arXiv:2404.00934* (2024).
[9] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward Controlled Generation of Text. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. 1587–1596.
[10] Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*. 52–62.
[11] G. Kakoma and Dr. Shira Tendo Namagero. 2020. A Stylistic Analysis of 'O Uganda, Land of Beauty' By Prof.George Kakoma. *International Journal on Studies in English Language and Literature* (2020).
[12] I Wayan Sidha Karya and Ida Bagus Adhika Mahardika. 2019. A STUDY ON HOW LONG AND SHORT SENTENCES SHOW THE STORY'S PACING IN ANTHONY HOROWITZ'S RAVEN'S GATE. *SPHOTA: Jurnal Linguistik dan Sastra* (2019).
[13] Amit Kumar Kushwaha and Arpan Kumar Kar. 2024. MarkBot–a language model-driven chatbot for interactive marketing in post-modern world. *Information systems frontiers* 26, 3 (2024), 857–874.
[14] Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multidimensional Evaluation for Text Style Transfer Using ChatGPT. *arXiv preprint arXiv:2304.13462* (2023).
[15] Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L. Zhang. 2021. Enhancing Content Preservation in Text Style Transfer Using Reverse Attention and Conditional Layer Normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 93–102.
[16] M. Lewis and Michael C. Frank. 2016. The length of words reflects their conceptual complexity. *Cognition* 153 (2016), 182–195.
[17] Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding Large Language Models via Directional Stimulus Prompting. *arXiv preprint arXiv:2302.11520* (2023).
[18] Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Zhaohui Wy, Dawei He, Peng Cheng, Zhonghao Wang, et al. 2023. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. *arXiv preprint arXiv:2311.15296* (2023).
[19] Yiwei Lyu, Tiange Luo, Jiacheng Shi, Todd C Hollon, and Honglak Lee. 2023. Fine-grained Text Style Transfer with Diffusion-Based Language Models. *arXiv preprint arXiv:2305.19512* (2023).
[20] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
[21] Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 490–497.
[22] M. Serrano, A. Flammini, and F. Menczer. 2009. Modeling Statistical Properties of Written Text. *PLoS ONE* 4 (2009).
[23] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature* (2023), 1–6.
[24] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14974–14983.
[25] Fadi Abu Sheikha and Diana Inkpen. 2010. Automatic classification of documents by formality. In *Proceedings of the 6th international conference on natural language processing and knowledge engineering (nlpke-2010)*. IEEE, 1–5.
[26] Hrituraj Singh, Gaurav Verma, Aparna Garimella, and Balaji Vasan Srinivasan. 2021. DRAG: Director-Generator Language Modelling Framework for Non-Parallel Author Stylized Rewriting. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 863–873.

[27] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2023. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11523–11530.

[28] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2998–3009.

[29] Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9008–9015.

[30] Martina Toshevska and Sonja Gievska. 2021. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence* (2021).

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[32] Adrian Wallwork and Adrian Wallwork. 2016. Teaching Students to Recognize the Pros and Cons of Short and Long Sentences. *English for Academic Research: A Guide for Teachers* (2016), 69–77.

[33] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746* (2023).

[34] Cheng Wen, Xianghui Sun, Shuaijiang Zhao, Xiaoquan Fang, Liangyu Chen, and Wei Zou. 2023. Chathome: Development and evaluation of a domain-specific language model for home renovation. *arXiv preprint arXiv:2307.15290* (2023).

[35] Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2021. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 3801–3807.

[36] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

[37] Jie Zhao, Ziyu Guan, Cai Xu, Wei Zhao, and Yue Jiang. 2024. SC2: Towards Enhancing Content Preservation and Style Consistency in Long Text Style Transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9949–9960.

[38] Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2021. Stylized dialogue response generation using stylized unpaired texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14558–14567.

[39] Xuekai Zhu, Jian Guan, Minlie Huang, and Juan Liu. 2023. StoryTrans: Non-Parallel Story Author-Style Transfer with Discourse Representations and Content Enhancing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14803–14819.

[40] Xuekai Zhu, Jian Guan, Minlie Huang, and Juan Liu. 2023. StoryTrans: Non-Parallel Story Author-Style Transfer with Discourse Representations and Content Enhancing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14803–14819.

A Appendix

A.1 Experiment Results, Style Definitions, and Case studies

Table A1. Automatic evaluation results on "The Scream".

Chinese Text	Models	Transfer Accuracy(%)					Content Preservation(%)				
		BLEU-1	BLEU-2	Precision	Recall	F1	BLEU-1	BLEU-2	Precision	Recall	F1
The Scream	Style Transformer	16.89	10.22	59.58	60.71	60.13	**92.71**	**90.38**	**98.23**	**98.13**	**98.18**
	RALoCN	24.72	17.14	66.09	65.48	65.78	89.02	87.16	95.12	95.85	95.49
	StyleLM	19.34	16.18	55.61	56.05	55.86	92.31	87.81	93.92	95.22	94.82
	SDR	30.23	22.15	59.67	62.66	60.48	92.16	85.17	92.81	91.91	92.19
	DRAG	33.14	22.52	65.18	63.29	64.33	90.68	86.40	95.96	95.02	95.44
	StoryTrans	**33.33**	**24.42**	**67.81**	**69.09**	**68.45**	89.36	84.24	94.59	93.53	93.96
	Role-play+ChatGLM	38.67	19.87	76.89	77.12	76.99	75.62	68.16	88.64	89.56	89.06
	Read+ChatGLM	13.72	4.17	57.88	62.96	60.30	13.89	4.82	57.99	62.84	60.30
	CAT+ChatGLM	40.44	22.47	76.17	77.87	76.98	74.43	67.70	91.12	92.26	91.63
	Role-play+Baichuan	36.07	16.93	75.24	76.22	75.72	80.90	78.51	88.37	88.17	88.27
	Read+Baichuan	17.14	7.05	59.32	64.18	61.64	21.54	12.78	61.24	65.85	63.44
	CAT+Baichuan	42.55	24.31	77.32	77.65	77.45	83.33	79.16	94.03	93.13	93.54
	Role-play+GPT3.5	41.13	22.91	77.60	78.45	78.01	73.82	66.03	92.42	92.27	92.34
	Read+GPT3.5	19.76	8.94	60.91	64.75	62.74	29.70	22.23	64.94	68.59	66.68
	CAT+GPT3.5	**46.78**	**26.53**	**79.41**	**80.36**	**79.92**	**88.89**	**84.59**	**96.16**	**96.42**	**96.31**
	Role-play+GPT4	34.89	16.81	72.04	74.13	73.05	42.05	25.52	78.03	79.68	78.82
	Read+GPT4	30.33	14.14	73.95	73.64	73.78	37.98	23.25	79.90	78.62	79.23
	CAT+GPT4	36.92	18.16	73.83	75.25	74.31	55.76	39.96	87.46	87.38	87.42

Table A2. Automatic evaluation results on "Fifteen Year of the Wanli Era".

Chinese Text	Models	Transfer Accuracy(%)					Content Preservation(%)				
		BLEU-1	BLEU-2	Precision	Recall	F1	BLEU-1	BLEU-2	Precision	Recall	F1
Fifteen Year of the Wanli Era	Style Transformer	14.60	10.01	52.10	51.76	51.93	**93.82**	**88.18**	**94.36**	**94.84**	**94.60**
	RALoCN	18.62	13.58	58.01	58.41	58.31	92.89	86.79	89.49	88.92	89.20
	StyleLM	17.89	12.45	55.67	57.02	56.33	91.45	84.72	92.03	91.76	91.92
	SDR	24.33	16.32	60.89	61.73	61.15	88.14	79.89	90.75	89.33	89.95
	DRAG	26.72	18.45	63.14	64.06	63.72	89.66	81.32	92.49	91.78	92.13
	StoryTrans	**28.51**	**20.15**	**65.22**	**66.49**	**65.85**	90.52	85.74	93.11	92.45	92.88
	Role-play+ChatGLM	24.37	14.39	67.90	68.13	67.99	74.89	65.93	91.83	90.94	91.36
	Read+ChatGLM	13.09	3.33	54.40	57.37	55.83	16.67	6.04	57.69	60.76	59.17
	CAT+ChatGLM	43,62	25.37	78.63	79.60	79.08	85.54	79.85	93.98	93.97	93.97
	Role-play+Baichuan	37.33	21.88	74.99	79.18	76.84	70.46	68.75	83.85	88.65	85.79
	Read+Baichuan	18.04	6.20	59.25	62.25	60.69	26.51	16.60	62.50	65.39	63.89
	CAT+Baichuan	41.88	23.99	77.62	77.77	77.66	90.54	89.14	95.52	95.24	95.37
	Role-play+GPT3.5	44.71	26.24	79.35	80.24	79.78	84.38	78.70	95.67	95.51	95.58
	Read+GPT3.5	16.13	5.76	58.55	62.19	60.30	23.46	14.40	61.54	65.14	63.27
	CAT+GPT3.5	**46.27**	**27.94**	**80.78**	**80.66**	**80.73**	**93.63**	**90.14**	**96.17**	**96.99**	**96.56**
	Role-play+GPT4	30.50	14.31	69.09	71.87	70.41	40.29	23.79	74.25	77.11	75.60
	Read+GPT4	29.57	13.83	72.98	74.76	73.78	35.82	20.74	77.73	79.17	78.33
	CAT+GPT4	36.99	19.28	75.73	76.53	76.12	56.49	49.96	84.37	84.41	84.39

Table A3. Automatic evaluation results on "The Family Instructions of Zeng Guofan".

Chinese Text	Models	Transfer Accuracy(%)					Content Preservation(%)				
		BLEU-1	BLEU-2	Precision	Recall	F1	BLEU-1	BLEU-2	Precision	Recall	F1
The Family Instructions of Zeng Guofan	Style Transformer	12.98	8.42	56.76	59.42	58.07	**92.30**	**91.44**	**98.27**	**98.33**	**98.30**
	RALoCN	21.87	15.17	58.23	58.04	58.13	90.20	88.33	95.23	95.89	95.74
	StyleLM	19.75	12.35	57.21	57.82	57.51	91.55	89.72	96.21	96.53	96.34
	SDR	27.50	18.15	62.10	63.15	62.48	88.31	84.55	94.45	95.20	84.86
	DRAG	30.10	20.45	64.40	65.93	65.27	89.65	86.11	95.75	96.32	95.97
	StoryTrans	**31.45**	**22.12**	**66.75**	**67.80**	**67.14**	90.45	87.35	96.78	97.05	96.93
	Role-play+ChatGLM	40.30	21.56	75.38	74.90	75.11	66.19	55.29	87.97	86.97	87.43
	Read+ChatGLM	19.05	5.20	57.05	59.02	58.01	19.95	6.05	57.75	59.84	58.77
	CAT+ChatGLM	41.16	23.25	76.58	77.25	76.87	78.34	72.32	93.23	91.68	92.40
	Role-play+Baichuan	37.15	21.82	69.64	72.25	70.78	73.66	71.97	85.06	88.96	86.60
	Read+Baichuan	18.76	6.52	58.37	60.97	59.62	24.25	13.41	60.73	63.59	62.10
	CAT+Baichuan	41.97	23.61	76.74	76.43	76.56	89.14	87.61	95.09	94.81	94.94
	Role-play+GPT3.5	39.35	21.57	76.88	76.53	76.68	69.63	60.44	91.42	90.40	90.89
	Read+GPT3.5	14.72	4.82	56.63	60.39	58.44	15.44	5.77	57.03	60.82	58.85
	CAT+GPT3.5	**45.97**	**26.94**	**78.54**	**79.06**	**78.79**	**90.72**	**87.15**	**97.23**	**97.41**	**97.31**
	Role-play+GPT4	24.09	8.94	69.76	68.47	69.07	26.50	11.49	73.19	71.61	72.36
	Read+GPT4	32.11	14.97	71.88	72.14	71.97	47.09	32.57	78.56	83.83	81.03
	CAT+GPT4	33.55	18.24	73.30	74.35	73.84	40.84	25.14	77.33	77.40	77.32

Table A4. Automatic evaluation results on "The Three-Body Problem".

Chinese Text	Models	Transfer Accuracy(%)					Content Preservation(%)				
		BLEU-1	BLEU-2	Precision	Recall	F1	BLEU-1	BLEU-2	Precision	Recall	F1
The Three-Body Problem	Style Transformer	14.45	7.39	59.11	59.27	59.19	**92.00**	**90.55**	**97.20**	**97.91**	**97.05**
	RALoCN	16.56	9.67	59.01	59.46	59.24	87.60	80.25	94.42	96.16	95.80
	StyleLM	18.95	11.54	58.72	58.98	58.85	89.24	85.14	94.99	96.18	95.58
	SDR	24.35	15.31	62.35	63.45	62.85	85.55	79.47	93.49	94.17	93.86
	DRAG	26.80	17.26	64.01	65.19	64.78	87.16	81.74	94.18	95.35	94.81
	StoryTrans	**28.69**	**19.22**	**65.83**	**66.85**	**66.32**	88.23	83.26	95.31	93.91	94.62
	Role-play+ChatGLM	42.37	23.92	75.09	75.97	75.39	65.48	55.58	88.30	86.90	87.53
	Read+ChatGLM	15.06	5.78	58.18	62.95	60.46	19.37	11.52	59.55	64.30	61.81
	CAT+ChatGLM	45.89	25.79	75.90	76.80	76.16	79.71	74.13	94.07	92.48	93.21
	Role-play+Baichuan	42.29	25.52	73.87	76.43	74.99	77.57	75.20	89.51	92.05	90.48
	Read+Baichuan	25.59	12.80	64.39	67.90	66.06	40.21	32.91	70.79	73.80	72.21
	CAT+Baichuan	50.05	30.68	78.49	79.33	78.89	86.52	84.73	93.44	93.07	93.24
	Role-play+GPT3.5	45.52	27.39	77.02	77.90	77.43	68.09	58.08	88.82	88.82	88.80
	Read+GPT3.5	26.24	13.36	64.63	68.05	66.26	39.53	31.66	70.50	73.58	71.95
	CAT+GPT3.5	**50.28**	**30.80**	**79.24**	**80.46**	**79.83**	**88.16**	**83.80**	**96.13**	**96.39**	**96.25**
	Role-play+GPT4	35.34	18.01	70.07	73.98	71.95	40.89	24.14	74.07	77.74	75.84
	Read+GPT4	33.05	15.88	71.20	73.04	72.08	37.95	21.97	75.83	77.15	76.45
	CAT+GPT4	40.11	22.05	72.80	76.72	74.69	51.62	36.07	79.84	83.73	81.71

Table A5. Automatic evaluation results on "Scientific Literature".

Chinese Text	Models	Transfer Accuracy(%)					Content Preservation(%)				
		BLEU-1	BLEU-2	Precision	Recall	F1	BLEU-1	BLEU-2	Precision	Recall	F1
Scientific Literature	Style Transformer	15.33	10.41	53.14	52.80	52.97	**91.94**	**86.42**	**95.30**	**95.79**	**95.54**
	RALoCN	19.55	14.12	59.17	59.58	59.37	91.03	85.05	90.38	89.81	90.09
	StyleLM	18.78	12.95	56.78	58.16	57.46	89.62	83.03	92.95	92.68	92.81
	SDR	25.55	16.97	62.11	62.96	62.53	86.38	78.29	91.66	90.22	90.93
	DRAG	28.06	19.19	64.40	65.34	64.86	87.87	79.69	93.41	92.70	93.05
	StoryTrans	**29.94**	**20.96**	**66.52**	**67.82**	**67.16**	88.71	84.03	94.04	93.37	93.70
	Role-play+ChatGLM	47.74	32.16	81.22	82.74	81.94	67.85	54.59	89.42	89.47	89.43
	Read+ChatGLM	40.34	27.06	74.86	79.66	77.12	59.11	51.01	82.30	87.59	84.73
	CAT+ChatGLM	48.58	32.65	81.55	82.73	82.12	72.33	61.60	91.38	91.83	91.57
	Role-play+Baichuan	43.43	27.70	78.83	80.49	79.61	55.91	39.87	83.37	84.04	83.66
	Read+Baichuan	29.12	16.54	70.35	70.74	70.52	45.64	33.45	75.72	75.70	75.69
	CAT+Baichuan	45.62	30.00	80.29	81.13	80.68	72.98	63.18	91.20	90.75	90.94
	Role-play+GPT3.5	49.64	34.11	83.08	83.19	83.12	67.13	54.63	90.55	89.28	89.89
	Read+GPT3.5	48.49	32.89	82.01	82.10	82.04	74.22	65.54	92.60	91.26	91.91
	CAT+GPT3.5	**51.38**	**35.55**	**84.31**	**85.13**	**84.70**	**80.82**	**73.51**	**95.17**	**94.64**	**94.89**
	Role-play+GPT4	46.57	30.88	79.72	80.97	80.32	61.68	46.38	84.95	85.11	85.01
	Read+GPT4	39.20	25.44	73.91	81.24	77.38	57.37	45.60	80.62	88.26	84.24
	CAT+GPT4	48.70	32.66	81.12	83.23	82.14	63.68	48.24	87.04	88.26	87.63

Table A6. Human evaluation results on “Social Media” dataset.

Chinese Text	Models	Transfer Accuracy	Content Presevation	Fluency
Social Media	Style Transformer	59	92	55
	RALoCN	59	95	58
	StyleLM	60	94	59
	SDR	63	94	62
	DRAG	64	94	64
	StoryTrans	66	94	65
	Role-play+ChatGLM	80	88	75
	Read+ChatGLM	65	68	70
	CAT+ChatGLM	82	90	80
	Role-play+Baichuan	77	85	65
	Read+Baichuan	63	66	55
	CAT+Baichuan	79	89	75
	Role-play+GPT3.5	81	92	85
	Read+GPT3.5	83	92	80
	CAT+GPT3.5	**83**	**95**	90
	Role-play+GPT4	78	78	85
	Read+GPT4	78	84	80
	CAT+GPT4	81	86	**95**

Table A7. Human evaluation results on “Legal Document” dataset.

Chinese Text	Models	Transfer Accuracy	Content Presevation	Fluency
Legal Document	Style Transformer	59	97	55
	RALoCN	60	92	58
	StyleLM	65	95	59
	SDR	70	94	62
	DRAG	75	96	64
	StoryTrans	80	96	65
	Role-play+ChatGLM	85	88	75
	Read+ChatGLM	70	73	70
	CAT+ChatGLM	90	91	80
	Role-play+Baichuan	85	82	70
	Read+Baichuan	75	74	55
	CAT+Baichuan	88	87	80
	Role-play+GPT3.5	90	88	85
	Read+GPT3.5	85	87	90
	CAT+GPT3.5	**92**	**94**	90
	Role-play+GPT4	90	86	90
	Read+GPT4	88	88	95
	CAT+GPT4	95	86	**95**

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中动词及名词使用较多。对于虚词来说，\
该文本中限定词及连词使用较多。同时，词语中占比最高的分别为词长为2以及词长为1。该文本中，\
也经常使用'了了，了，吧，呢，吗，啊'等语气词。在词语音节这一块，经常使用'也,在,就,说,人,不,都,有,但,要,这'等单音节词语，\
同时，经常使用'就是,一个,所以,曾国藩'等多音节词语。在成语使用上，\
经常使用'出人意料，天真烂漫，德才兼备，难能可贵，恃才傲物，赌咒发誓，飞黄腾达，分道扬镳'等成语。\
此外，该文本风格从句子角度来看，平均句长为35。且该文本句式长句较多，内涵丰富，叙事具体，说理周详，感情充沛。\
该作者在写作时，用问号较多。问句通常用于引导读者思考，激发其思考和参与，使文章更具互动性。\
大体看来，该文本主题情感基调主要是'喜爱'。文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。\
这种情绪可能通过赞美、推崇或表达满足感来体现。此外，该文本的情感基调还带有'厌恶'。文本可能会表达一种厌恶或反感的情绪，\
可能是对某些行为、现象或物体的不满和排斥。厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。\
对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。
"""
```

Fig. A1. The style definition of “Fortress Besieged”.

Table A8. Human evaluation results on “Scientific Literature” dataset.

Chinese Text	Models	Transfer Accuracy	Content Presevation	Fluency
Scientific Literature	Style Transformer	59	97	55
	RALoCN	60	92	58
	StyleLM	65	95	59
	SDR	70	94	62
	DRAG	75	96	64
	StoryTrans	80	96	65
	Role-play+ChatGLM	85	88	75
	Read+ChatGLM	70	73	70
	CAT+ChatGLM	90	91	80
	Role-play+Baichuan	85	82	70
	Read+Baichuan	75	74	55
	CAT+Baichuan	88	87	80
	Role-play+GPT3.5	90	88	85
	Read+GPT3.5	85	87	90
	CAT+GPT3.5	**92**	**94**	90
	Role-play+GPT4	90	86	90
	Read+GPT4	88	88	95
	CAT+GPT4	95	86	**95**

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中动词及名词使用较多。对于虚词来说，\
该文本中限定词及连词使用较多。同时，词语中占比最高的分别为词长为1以及词长为2。\
该文本中，也经常使用'了，了了，哦，呀，啊，呢'等语气词。在词语音节这一块，\
经常使用'阿,Q,又,这,人,便,都,在,说,也'等单音节词语，\
同时，经常使用'一个,因为,起来,什么,知道,没有,自己,他们,我们,七斤'等多音节词语。\
在成语使用上，经常使用'青面獠牙，阿弥陀佛，不以为然，怒目而视，易子而食，之乎者也，自言自语，自作自受'等成语。\
此外，该文本风格从句子角度来看，平均句长为24。且该文本句式长短结合。长短句结合常常用于表达复杂的情感，\
且可以营造出语言的节奏感和韵律感。该作者在写作时，用问号较多。问句通常用于引导读者思考，激发其思考和参与，\
使文章更具互动性。该作者在写作时，用感叹句较多。感叹句通常用于表达强烈的情感、赞美或惊讶，使文章更具情感色彩。\
大体看来，该文本主题情感基调主要是'喜爱'。文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。\
这种情绪可能通过赞美、推崇或表达满足感来体现。此外，该文本的情感基调还带有'厌恶'。文本可能会表达一种厌恶或反感的情绪，\
可能是对某些行为、现象或物体的不满和排斥。厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。\
对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。\
此外，文本还使用较多的反讽修辞：通过言辞上的反语或讽刺，对某一事物或观点进行批评或嘲笑，达到强调作者立场或调侃的目的。
"""
```

Fig. A2. The style definition of “The Scream”.

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中动词及名词使用较多。\
对于虚词来说，该文本中限定词及介词使用较多。同时，词语中占比最高的分别为词长为2以及词长为1。\
该文本中，也经常使用'嗯，嗯嗯，了，了了，呀，吗，哦，啊，吧，呢'等语气词。在词语音节这一块，
经常使用'在,着,不,和,中,但,有,上,也,都,这,就,说'等单音节词语，同时，经常使用'他们,一个,没有,我们,罗辑'等多音节词语。\
在成语使用上，经常使用'不可思议，微不足道，小心翼翼，不知不觉，不动声色，黄金时代，面目全非，一模一样'等成语。\
此外，该文本风格从句子角度来看，平均句长为32。且该文本句式长句较多，内涵丰富，叙事具体，说理周详，感情充沛。\
该作者在写作时，用问号较多。问句通常用于引导读者思考，激发其思考和参与，使文章更具互动性。大体看来，\
该文本主题情感基调主要是'喜爱'。文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。\
这种情绪可能通过赞美、推崇或表达满足感来体现。此外，该文本的情感基调还带有'厌恶'。\
文本可能会表达一种厌恶或反感的情绪，可能是对某些行为、现象或物体的不满和排斥。\
厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。对于整散句这一块，该文本的散句使用更多。\
散句结构相对简短，呈现出更为灵活、自由的写作风格，适合表达抒情、富有感情色彩的内容。\
此外，文本还使用较多的夸张修辞：夸张通过夸大事物的特征或程度，引起读者的注意，强调某种情感或观点，有时用于幽默或夸张的效果。
"""
```

Fig. A3. The style definition of “The Fifteen Year of the Wanli Era”.

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中名词及动词使用较多。\
对于虚词来说，该文本中限定词及连词使用较多。同时，词语中占比最高的分别为词长为2以及词长为1。\
该文本中，也经常使用'了了，了，吗，呢'等语气词。在词语音节这一块，经常使用'在,为,而,有,以,也,就,和'等单音节词语，\
同时，经常使用'这种,万历,没有,皇帝,他们,一个,张居正,李贽'等多音节词语。在成语使用上，\
经常使用'所作所为，不可收拾，励精图治，按部就班，九五之尊，事与愿违，水落石出，飞黄腾达'等成语。\
此外，该文本风格从句子角度来看，平均句长为33。且该文本句式长句较多，内涵丰富，叙事具体，说理周详，感情充沛。\
大体看来，该文本主题情感基调主要是'喜爱'。文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。\
这种情绪可能通过赞美、推崇或表达满足感来体现。此外，该文本的情感基调还带有'厌恶'。文本可能会表达一种厌恶或反感的情绪，\
可能是对某些行为、现象或物体的不满和排斥。厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。\
对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。\
对于文本句子的修辞手法，使用最多的是反讽修辞：通过言辞上的反语或讽刺，对某一事物或观点进行批评或嘲笑，\
达到强调作者立场或调侃的目的。
"""
```

Fig. A4. The style definition of “The Family Instructions of Zeng Guofan”.

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中动词及名词使用较多。对于虚词来说，\
该文本中限定词及介词使用较多。同时，词语中占比最高的分别为词长为1以及词长为2。该文本中，\
也经常使用'了，了了，啊，吗，呀，哦，呢，吧'等语气词。在词语音节这一块，经常使用'说,不,在,道,也'等单音节词语，\
同时，经常使用'没有,自己,小姐,鸿渐,什么,他们,知道,辛楣,孙小姐,柔嘉'等多音节词语。在成语使用上，\
经常使用'无论如何，问长问短，自言自语，岂有此理，总而言之，不识抬举，千方百计，一窍不通'等成语。\
此外，该文本风格从句子角度来看，平均句长为23。且该文本句式长短结合。长短句结合常常用于表达复杂的情感，\
且可以营造出语言的节奏感和韵律感。该作者在写作时，用问号较多。问句通常用于引导读者思考，激发其思考和参与，\
使文章更具互动性。该作者在写作时，用感叹句较多。感叹句通常用于表达强烈的情感、赞美或惊讶，使文章更具情感色彩。\
大体看来，该文本主题情感基调主要是'厌恶'。文本可能会表达一种厌恶或反感的情绪，可能是对某些行为、现象或物体的不满和排斥。\
厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。此外，该文本的情感基调还带有'喜爱'。\
文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。这种情绪可能通过赞美、推崇或表达满足感来体现。\
对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。\
对于文本句子的修辞手法，使用最多的是反讽修辞：通过言辞上的反语或讽刺，对某一事物或观点进行批评或嘲笑，\
达到强调作者立场或调侃的目的。
"""
```

Fig. A5. The style definition of “The Three-Body Problem”.

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中名词及动词使用较多。对于虚词来说，\
该文本中限定词及介词使用较多。\同时，词语中占比最高的分别为词长为2以及词长为1。该文本中，\
也经常使用'啊，吗，吧，了了，了，呢'等语气词。在词语音节这一块，经常使用'在,中,将,但,有,上,和,都,也,而'等单音节词语，\
同时，经常使用'球员,球队,已经,比赛,没有'等多音节词语。在成语使用上，\
经常使用'脱颖而出，出人意料，沸沸扬扬，轰轰烈烈，前所未有，引人注目，众所周知，按兵不动'等成语。\
此外，该文本风格从句子角度来看，平均句长为47。且该文本句式长句较多，内涵丰富，叙事具体，说理周详，感情充沛。\
大体看来，该文本主题情感基调主要是'喜爱'。文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。\
这种情绪可能通过赞美、推崇或表达满足感来体现。此外，该文本的情感基调还带有'厌恶'。\
文本可能会表达一种厌恶或反感的情绪，可能是对某些行为、现象或物体的不满和排斥。\
厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。\
对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。
"""
```

Fig. A6. The style definition of “News” dataset.

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中名词及动词使用较多。\
对于虚词来说，该文本中介词及连词使用较多。\
同时，词语中占比最高的分别为词长为2以及词长为1。该文本中，也经常使用'了'等语气词。\
在词语音节这一块，经常使用'和,年,在,有,等,或,为,中,与'等单音节词语，同时，经常使用'中国,一个,国家'等多音节词语。\
在成语使用上，经常使用'青天白日，不知所终，长生不老，成败得失，出类拔萃，出人意表，独来独往，独善其身'等成语。\
此外，该文本风格从句子角度来看，平均句长为38。且该文本句式长句较多，内涵丰富，叙事具体，说理周详，感情充沛。\
大体看来，该文本主题情感基调主要是'喜爱'。文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。\
这种情绪可能通过赞美、推崇或表达满足感来体现。此外，该文本的情感基调还带有'厌恶'。\
文本可能会表达一种厌恶或反感的情绪，可能是对某些行为、现象或物体的不满和排斥。\
厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。\
对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。
"""
```

Fig. A7. The style definition of “Wikipedia” dataset.

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中名词及动词使用较多。\
对于虚词来说，该文本中限定词及介词使用较多。\
同时，词语中占比最高的分别为词长为2以及词长为1。该文本中，也经常使用'啊，吗，吧，了，哦，呢'等语气词。\
在词语音节这一块，经常使用'在,和,将,上'等单音节词语，同时，文本中较少使用多音节词。\
在成语使用上，经常使用'安然无恙，并驾齐驱，力所能及，门外汉，宝刀不老，比比皆是，博采众长，不甘示弱'等成语。\
此外，该文本风格从句子角度来看，平均句长为43。且该文本句式长句较多，内涵丰富，叙事具体，说理周详，感情充沛。\
大体看来，该文本主题情感基调主要是'喜爱'。文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。\
这种情绪可能通过赞美、推崇或表达满足感来体现。此外，该文本的情感基调还带有'厌恶'。\
文本可能会表达一种厌恶或反感的情绪，可能是对某些行为、现象或物体的不满和排斥。\
厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。\
对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。
"""
```

Fig. A8. The style definition of “Social Media” dataset.

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中名词及动词使用较多。\
对于虚词来说，该文本中介词及限定词使用较多。\
同时，词语中占比最高的分别为词长为2以及词长为1。该文本中，也经常使用'吧，了'等语气词。\
在词语音节这一块，经常使用'年,月,在,与,后,将,其,并,某,和'等单音节词语，同时，\
经常使用'人民检察院,指控,被告人,被害人,某某,有限公司'等多音节词语。\
在成语使用上，经常使用'怀恨在心，拳打脚踢，信以为真，不省人事，改过自新，昏迷不醒，滥用职权，敲诈勒索'等成语。\
此外，该文本风格从句子角度来看，平均句长为91。且该文本句式长句较多，内涵丰富，叙事具体，说理周详，感情充沛。\
大体看来，该文本主题情感基调主要是'厌恶'。文本可能会表达一种厌恶或反感的情绪，可能是对某些行为、现象或物体的不满和排斥。\
厌恶的情绪可能通过描述令人不快的场景或使用贬义词汇来体现。此外，该文本的情感基调还带有'喜爱'。\
文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。这种情绪可能通过赞美、推崇或表达满足感来体现。\
对于整散句这一块，该文本的整句使用更多。整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。
"""
```

Fig. A9. The style definition of “Legal Document” dataset.

```
Style_Definition = """
该文本风格从词语角度来看，使用较多的实词。对于实词来说，该文本中名词及动词使用较多。\
对于虚词来说，该文本中介词及连词使用较多。同时，词语中占比最高的分别为词长为2以及词长为1。\
该文本中，也经常使用'了，呢'等语气词。在词语音节这一块，经常使用'和,在,中,与,对,并,为'等单音节词语，\
同时，经常使用'影响,分析,研究,方法,进行,通过,模型'等多音节词语。\
在成语使用上，经常使用'刻不容缓，讨价还价，不约而同，长治久安，聪明才智，错综复杂，大庭广众，代代相传'等成语。\
此外，该文本风格从句子角度来看，平均句长为75。且该文本句式长句较多，内涵丰富，叙事具体，说理周详，感情充沛。\
大体看来，该文本主题情感基调主要是'喜爱'。文本可能会描述一种积极向上的情感状态，表达对某人、某事或某物的喜爱和满意。\
这种情绪可能通过赞美、推崇或表达满足感来体现。此外，该文本的情感基调还带有'愉快'。\
文本可能会传达一种愉悦和幸福的情绪，可能包含快乐的事件、幽默的描写或令人感到温馨的故事。\
愉快的情绪通常会让读者感到轻松和愉快。对于整散句这一块，该文本的整句使用更多。\
整句结构使得文章更具逻辑性和结构性，各部分之间的连接更为紧密。\
对于文本句子的修辞手法，使用最多的是反讽修辞：通过言辞上的反语或讽刺，对某一事物或观点进行批评或嘲笑，\
达到强调作者立场或调侃的目的。
"""
```

Fig. A10. The style definition of “Scientific Literature” dataset.

Fig. A11. Text style removal prompt.

Styleless Text:
申阁老清楚，在这种情况下，如果他坚持每天早朝，肯定会让自己在朝廷中变得不受欢迎。公元1587年，也就是万历十五年，申时行已经52岁了，他感觉自己虽然年纪不大，却似乎提前衰老了。几年前，他的鬓角就已经开始泛白，现在作为朝中的首辅，地位崇高，但他还记得张居正的前车之鉴。为什么他还要对一切如此认真，不愿意随遇而安呢？

StoryTrans:
申阁老十分清楚，在这种局势下，如果他依然坚持每天参加早朝，必然会让自己在朝廷中变得不受待见。公历1587年，即万历十五年，申时行已经52岁了，虽然他觉得自己年纪不大，但却仿佛提前衰老。几年前，他的鬓角已经开始泛白，如今他是朝中的首辅，地位显赫，但他依然记得张居正的前车之鉴。为什么他还要如此执着，对一切都如此认真，不愿顺其自然呢？

Role-Play+GPT-3.5-Turbo
申阁老深知，若坚持每日早朝，在此情况下，必定招致朝廷中的不受欢迎。公历1587年，即万历十五年，申时行已年逾五旬，感觉自己虽未老，却似乎提前衰颓。数年前，鬓角已现斑白，如今身为朝中首辅，地位崇高，然仍忆起张居正的前车之鉴。为何对一切如此认真，不愿随波逐流呢？

Read+GPT-3.5-Turbo
公元1587年，即明万历十五年，申阁老已年届五十二，虽非垂暮之年，却常感力不从心。论年岁，尚属中年；但鬓边白发初现，身心皆觉疲惫，仿佛时光在他身上提前留下印记。作为当朝首辅，申时行的地位已至巅峰，然他心中清楚，位高未必无忧，前车之鉴犹历历在目。张居正之事尚未远去，昔日的权势滔天到最终的冷落遗弃，朝廷风云不过如此。他深知，如若每日早朝不曾有丝毫松懈，恐怕只会令自己在官场中徒增掣肘，甚至难免树敌。然而，令他不解的是，既知身在险地，为何仍要执着认真，凡事追求尽善尽美？世事难平，随遇而安或许本是为政之道，但对于申阁老而言，信念与责任却似一座无形的山，总在心头沉沉压下。这一年，北风料峭，京城街巷间的冰雪尚未融尽，而朝堂之上的暗流，却早已悄然涌动。

CAT+GPT-3.5-Turbo
申阁老深知，他的坚持每天早朝必定会让自己在朝廷中变得不受欢迎。公历1587年，时为万历十五年，申时行已经52岁了，他已感到未老先衰。几年前，他的鬓角就已经开始泛白，现在作为朝中的首辅，已感到未老先衰，但他还记得张居正的前车之鉴。为什么他还要对一切如此认真，不愿听其自然呢？

Ground Truth:
申阁老深知他在这种情形下，坚持早朝的不断举行，必将使自己成为朝廷上不受欢迎的人物。公历1587年，时为万历十五年，申时行52岁，他已感到未老先衰。几年之前，他即已鬓发苍苍，现在身为首辅，已感到未老先衰，又有张居正前车之鉴，为什么他对一切还要那么认真而不听其自然呢？

Fig. A12. Case study on the “The Fifteen Year of the Wanli Era” dataset.

Styleless Text:
2008年11月22日，中国羽毛球公开赛到了半决赛。国家羽毛球队的总教练李永波在参加“上海孙俊葛菲羽毛球俱乐部挂牌仪式”时，抽空接受了新浪体育的采访。他聊了聊北京奥运会男单冠军林丹带领的中国队整体表现，还第一次正式说队里还没收到谢杏芳的退役申请。他说年轻队员的表现对未来很关键，这次比赛上，中国的小将们，尤其是男女单打选手，让他看到了中国羽毛球的新希望。

StoryTrans:
2008年11月22日，中国羽毛球公开赛到了半决赛。国家羽毛球队的总教练李永波在出席“上海孙俊葛菲羽毛球俱乐部挂牌仪式”时，抽时间接受了新浪体育的采访。他聊到了北京奥运会男单冠军林丹带领的中国队整体表现，还第一次正式提及队里还没收到谢杏芳的退役申请。他说年轻队员的表现对未来很重要，这次比赛上，中国的小将们，尤其是男女单打选手，让他看到了中国羽毛球的新希望。

Role-Play+GPT-3.5-Turbo
2008年11月22日，中国羽毛球公开赛进入了半决赛。国家羽毛球队总教练李永波在参加“上海孙俊葛菲羽毛球俱乐部挂牌仪式”时，抽空接受了新浪体育的采访。他谈到了北京奥运会男单冠军林丹领导的中国队整体表现，同时首次正式提及队内尚未收到谢杏芳的退役申请。李永波表示，年轻队员的表现对未来发展至关重要，本次比赛中，中国的年轻选手，尤其是男女单打选手，给他带来了中国羽毛球前景的新希望。

Read+GPT-3.5-Turbo
2008年11月22日，中国羽毛球公开赛的半决赛中，国家羽毛球队总教练李永波在参加“上海孙俊葛菲羽毛球俱乐部挂牌仪式”时，抽出时间接受了新浪体育的采访。他谈到了北京奥运会男单冠军林丹领导的中国队整体表现，还首次正式提到队伍尚未收到谢杏芳的退役申请。他表示年轻队员的表现对未来至关重要，这次比赛中，中国的年轻选手，特别是男女单打选手，让他看到了中国羽毛球的新希望。

CAT+GPT-3.5-Turbo
2008年11月22日，中国羽毛球公开赛进入到半决赛阶段。国家羽毛球队的总教练李永波在出席“上海孙俊葛菲羽毛球俱乐部挂牌仪式”时，抽出时间接受了新浪体育的采访。他点评了北京奥运会男单冠军林丹领衔的中国队整体表现，同时首次正式透露队内尚未收到谢杏芳的退役申请。他表示年轻队员的表现对未来至关重要，这次比赛中，中国的小将们，尤其是男女单打选手，让他看到了中国羽毛球的新希望。

Ground Truth:
11月22日，2008年中国羽毛球公开赛进入到半决赛阶段，国家羽毛球队总教练李永波在出席“上海孙俊葛菲羽毛球俱乐部挂牌仪式”的间隙，接受新浪体育的采访时点评了由北京奥运会男单冠军得主林丹领衔的中国队整体发挥，并且首次正式表态队里至今没收到过来自谢杏芳的退役申请。年轻队员的表现对新周期的开展尤为重要，令人欣慰的是这次比赛上中国小将，特别是从男女单打选手身上更看到了中国羽毛球的新希望。

Fig. A13. Case study on the “News” dataset.

Styleless Text:
吴起是战国初期的一个很有名的军事家、政治家和改革家。他是卫国人，就是现在山东定陶县附近的人。吴起在鲁国、魏国和楚国都干过，他对兵法、法律和儒家思想都很有研究，在治理国家和打仗方面都特别厉害。在鲁国工作的时候，他曾经打退过齐国的进攻；在魏国的时候，他多次打败秦国，把秦国在黄河以西的地盘都占了，帮魏文侯实现了称霸的梦想；到了楚国，他还搞了一次很有名的改革，历史上叫“吴起变法”。

StoryTrans:
吴起，战国初期的有名的军事家、政治家和改革家。卫国人，就是现在山东定陶县附近的人。吴起在鲁国、魏国和楚国都工作过，他对兵法、法律和儒家思想都很有研究，在治国和打仗方面很厉害。在鲁国工作时，他曾经打退过齐国的进攻；在魏国时，他多次打败秦国，把秦国在黄河以西的地盘都占了，帮魏文侯实现了称霸；到了楚国，他进行了改革，历史上叫“吴起变法”。

Role-Play+GPT-3.5-Turbo
吴起，生活在战国时期的杰出军事家、政治家和改革家，其名声显赫。出生于卫国，即今山东定陶县一带。吴起曾在鲁国、魏国和楚国三国任职，对兵法、法律和儒家思想皆有深入研究，在国家治理和军事作战方面表现卓越。在鲁国任职期间，曾成功击退齐国的进攻；在魏国任职时，多次战胜秦国，夺取了黄河以西的领土，助魏文侯实现了称霸之志；后来在楚国，他进行了一场著名的改革，后人称之为“吴起变法”。

Read+GPT-3.5-Turbo
吴起，为战国时期著名军事家、政治家和改革家，出生于卫国，即今山东定陶县附近。吴起曾在鲁国、魏国和楚国任职，精通兵法、法律和儒家思想，在治国和作战方面表现出色。在鲁国任职期间，成功击退了齐国的进攻；在魏国时，多次战胜了秦国，夺取了黄河以西的领土，帮助魏文侯实现了称霸之志；在楚国，他进行了著名的改革，被称为“吴起变法”。

CAT+GPT-3.5-Turbo
吴起，战国初期军事家、政治家和改革家。卫国之人，即今山东定陶县附近之人。吴起曾游于鲁国、魏国和楚国，兵法、法律及儒家思想皆有深研。内政与战争之道，尤为娴熟。在鲁国任职之时，曾挫退齐国的入侵；在魏国之时，屡次击败秦国，占领黄河以西之地，成就魏文侯的霸业；至楚国，则推行改革，史称“吴起变法”。

Ground Truth:
吴起，中国战国初期军事家、政治家、改革家，兵家代表人物。衛国左氏（卫国）人。吴起历侍鲁、魏、楚三国，通晓兵家、法家、儒家三家思想，在内政、军事上都有极高的成就。仕鲁时曾击退齐国的入侵；仕魏时屡次破秦，尽取秦国的河西地区，成就魏文侯的霸业；仕楚时主持改革，史称吴起变法。

Fig. A14. Case study on the “Wikipedia” dataset.

Styleless Text:
微信和其他语音聊天工具的涌现，主要是因为大家越来越需要用手机上网聊天，而那些运营商却总是不给力，既不打折也不推出新的服务。虽然飞信后来也出了手机应用，但发个语音还要每分钟收一毛五到两毛五，这样一来，和微信比起来就完全没有竞争力了。

StoryTrans:
微信和其他语音聊天工具的涌现，主要是大家越来越需要用手机网络聊天，但那些运营商却总是既不打折也不推出新的服务。虽然飞信后来也推出了手机应用，但发个语音还要每分钟收一毛五到两毛五，这样，和微信比起来就毫无竞争力了。

Role-Play+GPT-3.5-Turbo
微信和其他语音聊天工具的出现，主要是因为大家越来越需要用手机上网聊天，而那些运营商却总是不给力，既不打折也不推出新的服务。虽然飞信后来也出了手机应用，但发个语音还要每分钟收一毛五到两毛五，这样一来，和微信比起来就完全没有竞争力了。 #手机上网# #语音聊天# #运营商# #微信# #飞信#

Read+GPT-3.5-Turbo
微信和其他语音聊天工具的涌现，主要是因为大家越来越需要用手机上网聊天，而那些运营商却总是不给力，既不打折也不推出新的服务。尽管飞信后来也推出了手机应用，但发个语音还要每分钟收一毛五到两毛五，这样一来，和微信比起来就完全没有竞争力了。

CAT+GPT-3.5-Turbo
微信和其他语音聊天工具的出现，离不开大家日益增长利用手机上网聊天的需求，而那些运营商却总是不给力，既不打折也不推出新的产品这样的矛盾。虽然飞信后来也开发了手机客户端，但发个语音还要每分钟收一毛五到两毛五，这样一来，和微信比起来就完全没有竞争力了。

Ground Truth:
微信等语音IM工具的出现，离不开一个矛盾：人民日益增长的移动网络沟通需求与运营商死皮赖脸不给优惠又不开发新产品之间的矛盾。虽飞信后来也开发了手机客户端，但是其发送语音还要每分钟收取1毛5到2毛5的费用，只这一项，就注定其在与微信的竞争中必须输。

Fig. A15. Case study on the “Social Media” dataset.

Styleless Text:
经过法院的审理，发现事实和证据跟原来的判决是一样的。原判决用的一些证据在第一次庭审的时候就已经展示和核实过了，法院也确认这些证据是真的。在二审期间，赵某和受害人赵某某就赔偿的事情达成了协议，受害人赵某某也表示理解赵某的情况。法院委托南乐县司法局对赵某进行了调查，调查结果认为赵某可以适用社区矫正。关于赵某上诉说原审判决的时候没有考虑到受害人的过错和他愿意赔偿损失的态度，认为判得太重了。

StoryTrans:
经法院审理，发现事实和证据跟原判决一致。原判决用的一些证据在第一次庭审的时候就核实过了，法院也确认这些证据是真的。二审期间，赵某和受害人赵某某就赔偿的事情签订协议，受害人赵某某也表示理解赵某的情况。本院委托南乐县司法局对赵某进行了调查，调查结果认为赵某适用于社区矫正。关于赵某上诉说原审判决的时候没有考虑到受害人的过错和他愿意赔偿损失的态度，认为判得严厉了。

Role-Play+GPT-3.5-Turbo
经法院审理查明，事实和证据与初审判决一致。初审所用证据已在初次庭审中展示核实，法院亦确认其真实性。在二审期间，赵某与受害人赵某某就赔偿事宜达成协议，受害人赵某某亦表示理解赵某之处。法院委托南乐县司法局对赵某展开调查，调查结果认定赵某适用社区矫正。关于赵某上诉称原审判决未考虑受害人过错及其愿意赔偿损失态度，认为判罚过重。

Read+GPT-3.5-Turbo
经过法院审查，发现事实和证据与原判决一致。原判决使用的证据在第一次庭审时已经出示和核实，法院也确认了这些证据的真实性。在二审期间，赵某和受害人赵某某就赔偿达成了协议，受害人赵某某也表示理解赵某的情况。法院委托南乐县司法局对赵某进行调查，调查结果认为赵某可以适用社区矫正。关于赵某的上诉称原审判决未考虑受害人的过错和他愿意赔偿损失的态度，认为判罚过重。

CAT+GPT-3.5-Turbo
经过本院的审理，发现事实和证据与原判决相同。原判决认定的一些证据在首次庭审时已经展示和核实，人民检察院也确认这些证据的真实性。二审期间，上诉人赵某与被害人赵某某就赔偿事宜达成了协议，被害人赵某某对赵某表示谅解。本院委托南乐县司法局对赵某进行了调查，调查结果认为赵某适用社区矫正。关于赵某上诉称原审判决未考虑到被害人的过错和他愿意赔偿损失的态度，认为量刑过重。

Ground Truth:
经本院审理查明的事实及证据与原判相同，原判认定的证据经一审当庭举证、质证，查证属实，经本院审核予以确认。二审期间，上诉人赵某与被害人赵某某就赔偿事宜达成民事调解协议，被害人赵某某对赵某表示谅解；本院委托南乐县司法局对上诉人赵某进行了调查，南乐县司法局出具的调查评估意见书认为，赵某适用社区矫正。关于上诉人赵某上诉提出原审量刑时未考虑被害人过错及其愿意赔偿损失，量刑过重的理由。

Fig. A16. Case study on the “Legal Document” dataset.

Styleless Text:
这篇文章研究了一种新的控制方法，把滑模控制和最优控制里的砰砰控制结合起来，用来控制建筑物的振动。因为建筑上用的那些控制设备大多数是液压的，它们不太容易在正负极限的控制力之间快速切换，所以我们改进了一下滑模砰砰控制的方法，用指数函数来代替原来的符号函数。考虑到传感器的数量有限，我们提出了一种只需要有限状态输出的滑模砰砰控制方法。

StoryTrans:
这篇文章说明了一种新的控制方法，将滑模控制和最优控制里的控制结合起来，用于控制建筑物的振动。因建筑上用的那些控制设备大多数是液压的，它们不易于在正负极限的控制力之间快速切换，所以这篇文章改进了一下滑模砰砰控制的方法，用指数函数来代替原来的符号函数。考虑到传感器的数量较少，提出了一种只需要有限状态输出的滑模控制方法。

Role-Play+GPT-3.5-Turbo
本文探讨了一种新的控制方法，将滑模控制与最优控制中的离散控制相结合，用于建筑结构的振动控制。鉴于建筑结构常采用液压控制设备，其在正负极限控制力之间切换困难，因此我们改进了滑模离散控制方法，引入指数函数取代原有符号函数。考虑到传感器数量有限，我们提出了一种仅需有限状态输出的滑模离散控制方法。

Read+GPT-3.5-Turbo
本研究探讨了一种新的控制方法，将滑模控制和最优控制中的控制策略相结合，用于调节建筑物的振动。由于建筑物上使用的控制设备主要是液压设备，其在正负极限控制力之间切换不够灵活，因此我们改进了滑模控制策略，采用指数函数替代传统的符号函数。考虑到传感器数量有限，我们提出了一种只需有限状态输出的滑模控制方法。

CAT+GPT-3.5-Turbo
本文研究了一项全新的控制方法，将滑模控制与最优控制中的砰砰控制相结合，用于管理建筑结构的振动。由于建筑结构中应用的控制设备大多为液压设备，它们很难适应正负极限控制力间的快速转换，因此本文对滑模砰砰控制方法进行了改进，采用指数函数取代了原有的符号函数。考虑到传感器数量的限制，我们提出了一种只需要有限状态输出的滑模砰砰控制方法。

Ground Truth:
本文将滑模控制和最优控制中的砰砰控制相结合，对建筑结构振动控制的滑模砰砰方法进行了研究。由于在建筑结构中应用的作动器多为液压形式，作动器很难适应正负极限控制力间的快速转换，为此本文给出了改进滑模砰砰控制律，用指数函数代替符号函数。考虑到安装传感器数量的限制，给出了有限状态输出的滑模砰砰控制方法。

Fig. A17. Case study on the “Scientific Literature” dataset.