Exploring Large Language Models Text Style Transfer Capabilities

Weijie Li[a,b,1], **Zhentao Gu**[a,1], **Xiaochao Fan**[a,*],
Wenjun Deng[a], **Yong Yang**[a], **Xinyuan Zhao**[a], **Yufeng Diao**[c] **and Liang Yang**[d,e]

[a]College of Computer Science and Technology, Xinjiang Normal University, China
[b]School of Software, Xinjiang University, China
[c]College of Computer Science and Technology, Inner Mongolia Minzu University, China
[d]School of Computer Science and Technology, Dalian University of Technology, China
[e]Key Laboratory of Social Computing and Cognitive Intelligence, Ministry of Education, Dalian, China

Abstract. The emergence of Large Language Models (LLMs) provides a new solution to text generation tasks that involve high complexity, such as text style transfer (TST) tasks. However, previous studies have not fully explored the TST capabilities of different LLMs, and have faced issues with a lack of uniform standards in the human evaluation stage. This makes the results of human evaluation difficult to reproduce and less credible. To address this, this paper designs a prompt template to guide the cutting-edge LLMs to perform effective text style transfer and carries out an in-depth comparative analysis of various small-scale language models. In the stage of human evaluation, this paper eschews the conventional rating system, opting instead for a comparative human assessment methodology, which we refer to as duel-ranking. This method determines the relative ranking of models through mutual comparison, serving as an alternative to direct scoring. Detailed evaluation instructions are provided herein, to enhance the reproducibility of this method and ensure consistency throughout the evaluation process. This manual evaluation process reveals that GPT-3.5 and GPT-4 exhibit excellent performance in the TST tasks.

1 Introduction

Text style transfer (TST) is the process of automatically converting text from a source style domain to a target style domain while preserving its deeper content meaning [35]. This involves adjusting various stylistic dimensions such as formality, sentiment polarity, and tone [4, 11]. TST can enhance the flexibility, diversity, and adaptability of text, improve the expression of ideas, and facilitate effective communication across a wide range of applications.

The TST models can be classified based on the generation mechanism into the simple reconstruction model [34], the model with embedded style classifiers [38] and the adversarial generation network model [21]. However, due to the underutilization of contextual information, the high dependence on the accuracy of the labelled data, and the problem of stability during training, the performance of these methods on the TST task is not yet satisfactory. Large Language Models (LLMs), on the other hand, have the ability to learn from massive language data in the pre-training stage, which allows them to deeply understand the deep semantic and stylistic features of text. This provides new possibilities for TST tasks. Although the current research on the performance of LLMs on TST is relatively limited, the huge potential it contains is expected to be further explored and practically verified.

As one of the important research directions in the field of natural language generation, the manual evaluation phase of TST tasks faces a series of challenges [9, 20]. Human evaluation is more comprehensive and accurate in evaluating content preservation, fluency, and transfer strength of text content compared to automatic evaluation. Nonetheless, there is a lack of uniform and recognized evaluation standards for manual evaluation, and the standard process and specific details of manual evaluation are not well documented in the literature. This makes it difficult to reach consistent and reproducible evaluation conclusions between different studies, weakening the reliability of the results of manual evaluation and the comparability of the results between different experiments [23]. Therefore, it is essential to establish uniform evaluation standards and documented processes for manual evaluation to ensure the reliability and comparability of the results between different experiments.

In this study, we design a template for TST prompt through several iterations. This template is designed to guide LLMs in completing TST tasks. We have also analyzed the TST ability of multiple LLMs by combining automatic and human evaluation methods. During the human evaluation stage, we have provided detailed information about the annotators and suggested a duel-ranking method for evaluation. This approach addresses the challenge of quantifying the subtle differences between model-generated utterances. It also makes the annotation rules more intuitive and easier to understand, resulting in more accurate and feasible evaluation. [2]

2 Related Work

2.1 The Deep Learning Based Approach to TST

Deep learning-powered TST methods primarily rely on deep neural network technologies. The fundamental process of such approaches can be summarised into two stages: 1) Learning representations by

* Corresponding Author. Email: fxc1982@xjnu.edu.cn.
[1] Equal contribution.

[2] Code is available at: https://github.com/SilverBeats/llm-tst-code

training deep models to capture and understand the textual content and its corresponding stylistic characteristics; and 2) Reconstructing new sentences that embody the intended target style while retaining the original meaning [35]. Currently, research strategies for TST vary, with some studies achieving simple style transfer by directly rewriting input sentences [17, 34]. Others incorporate additional style classifiers within the encoding-decoding process to guide the generation towards the desired style [5]. More innovative methods draw inspiration from adversarial learning principles, ensuring that the generated text remains faithful to the original intent and closely adheres to the specified stylistic requirements [6]. With the growing popularity and advancements in LLMs, TST techniques are entering a new era. It is now feasible to execute style transfer effectively without fine-tuning specific datasets. However, further research is necessary to determine the actual efficacy and potential of the LLMs in the task of TST.

2.2 Application of Prompt Methods to LLMs

LLMs have tens of billions of parameters or more. Due to their significantly increased parameter size, they possess capabilities that are not present in smaller models. The emergence of LLMs such as GPT-4 [1] has led to a paradigm shift in natural language processing. Research focus has now shifted from traditional tasks such as translation [26], sentiment analysis [8], and Q&A [27] towards more complex and intelligent decision-making related tasks. For instance, LLMs can now simulate tool use ability [30], role-playing ability [30], evaluation tools [14] and much more. The purpose of role-playing is to enable LLMs to simulate various roles or character identities with unique attributes and dialogue styles, providing users with a more immersive and approachable experience. The RoleLLM framework proposed by [40] aims to standardize testing, stimulation, and reinforcement assessment of role-playing abilities for LLMs. Additionally, [16] built ChatHaruhi, which allows LLMs to play real characters from anime, TV, or other works in dialogue with impressive results. The Character-LLM proposed by [32] can also make LLMs imitate specific people. In conclusion, role-playing enriches the functionality and expressiveness of LLMs, promotes their performance optimization, and expands their deep application in specific fields.

2.3 Human Evaluation Methods

In the TST research field, human evaluation is a widely used quality assessment strategy [33, 17]. Compared to automated evaluation methods, manual evaluation is considered more accurate and reliable in terms of judging results in the academic community [3]. However, most of the existing research literature focuses on the assessment sample capacity, the number of experts involved in the evaluation, and the specific dimensions of the evaluation during the manual evaluation stage and provides insufficient information on specific operational details and processes in the implementation of manual evaluation [23]. In practice, a graded scale is used to quantify the results of the evaluation, such as the common 3-point system [6, 25], 4-point system [33], 5-point system [17, 12, 22, 18, 43, 15, 41], and even the more nuanced 7-point [19] and 10-point [42]. However, it is worth noting that there is a lack of uniform specification for these scoring systems, and the boundaries between the scoring levels are not clearly and consistently defined. This phenomenon largely undermines the reproducibility and stability of manual evaluation results and consequently limits the credibility of assessment conclusions and the validity of cross-sectional comparisons.

3 Dataset

In this paper, we have used four non-parallel corpora that are commonly used in TST studies for experimental purposes. The details of the data division and task types of these corpora are presented in Table 1.

YELP [45] The first corpus is YELP, which consists of business reviews on Yelp. Each review is labeled with positive or negative sentiment, and the task type is to convert between positive and negative sentiment.

CAPTIONS [7] The second corpus is CAPTIONS, which contains picture captions labeled as factual, romantic, or humorous. The task type is to convert factual statements to either romantic or humorous style.

GENDER [37] The third corpus is GENDER, which consists of Yelp food reviews labeled with reviewer gender. The goal is to change the review style from one gender to another.

POLITICAL [24] The fourth corpus is POLITICAL, which contains Facebook comments made by members of the U.S. Senate and House of Representatives. The comments are labeled as Republican or Democratic, and the task type is to convert between both styles.

Table 1: Dataset statistics

Dataset	Style	Train	Dev	Test
YELP	Positive	266,041	2,000	500
	Negative	177,218	2,000	500
CAPTIONS	Romantic	6,000	300	0
	Humorous	6,000	300	0
	Factual	0	0	300
POLITICAL	Democrat	268,961	2,000	500
	Republican	268,961	2,000	500
GENDER	Male	267,230	2,246	500
	Female	1,334,592	2,246	500

4 Methodology

4.1 The Prompt Template

When creating a prompt template, it is important to include various essential components such as clear instructions for task orientation, comprehensive background information that is relevant to the context, specific instructions for processing the data, and standardized specifications for the desired output format [29]. Additionally, according to [31], by defining the role or domain identity that the model should follow in the prompt, the model's performance in that particular domain can be significantly improved. After multiple attempts, we have developed a template structure that comprises four core components: 1) instructions for setting the role; 2) details of the task; 3) requirements for delivering the output; and 4) input data. The prompt template we designed is presented below:

You are a linguist. You need to complete a text style transfer task. I will give you a {} style sentence, please change it to a {} style sentence. Please give me the revised sentence directly, without explaining the revision process. Sentence: {}

We explored several roles before deciding to use a large language model as a linguist. A linguist is someone who studies human language as an object of study and focuses on various aspects of language such as its structure, usage, social functions, historical development, and other relevant issues. They are expected to demonstrate a certain level of proficiency in these areas. We believe that this role is adaptable and can be tailored to different styles of TST tasks.

4.2 Automated Evaluation Methods

[23] identified various automatic evaluation settings for TST. All evaluation settings aim to measure content preservation, style transfer strength, and fluency of generated utterances.

To evaluate the extent of content preservation, BLEU methodology is widely used in the industry. In this paper, we have adopted the content preservation evaluation system of [15] and employed a set of BLEU-derived metrics and BERTscore [44] to measure the degree of content preservation of the generated sentences and their semantic similarity with human rewriting results. We have introduced three BLEU metrics: 1) source-BLEU (**s-BLEU**), which measures the BLEU score of the generated sentence compared to the source sentence, 2) reference- BLEU (**r-BLEU**) which evaluates the BLEU score between the generated sentence and the manually rewritten text, and 3) **g-BLEU** $= \sqrt{\text{s-BLEU} \times \text{r-BLEU}}$, which reflects the balanced performance of the generated text in terms of word usage.

To measure the style transfer ability of the model, we adopt the approach of [34] by training FastText [13] as a style classifier on the dataset. This allows us to gauge the TST ability of the model. The accuracy of the classifier on the validation set is presented in Table 2.

To evaluate the fluency of the sentence after the style transfer, we use [15]'s measure of fluency. We compute data fluency (**d-PPL**), general fluency (**g-PPL**), and composite fluency **t-PPL** $= \sqrt{\text{d-PPL} \times \text{g-PPL}}$. We measure g-PPL using a pre-trained GPT-2-large [3], while data fluency is measured using GPT-2 [4] fine-tuned on the dataset. The perplexity of GPT-2 on the validation set after fine-tuning is also presented in Table 2.

Table 2: Acc is the accuracy of Fasttext on the validation set. PPL is the perplexity score of GPT-2-small on the validation set after fine-tuning.

Dataset	Acc ↑	PPL↓
YELP	97.3%	14.24
CAPTIONS	77.2%	29.51
GENDER	82.4%	17.02
POLITICAL	83.0%	29.61

4.3 Duel-ranking Human Evaluation Methods

The purpose of this paper is to manually evaluate the utterances generated by LLMs using three dimensions: fluency, degree of content preservation, and transfer strength. In the text generation domain, there are two common manual evaluation schemes: quantitative scoring and contrastive preferences. These schemes have been used in previous research studies such as [46, 28, 39, 47] to evaluate text generation models.

Quantitative scoring mechanisms often use a Likert scale to measure the performance gap between different models. However, this approach has a limitation as it can be challenging for assessors to distinguish subtle quality differences between neighbouring scores accurately. This reduces the accuracy and credibility of the results to some extent. On the other hand, the contrastive approach is more intuitive, requiring the evaluator to choose the best item from several alternatives without specific scoring criteria. However, this method produces a sparse matrix of data, which is not suitable for exploring specific differences between the performance of each model in-depth.

This paper proposes a new evaluation methodology called duel-ranking that combines the advantages of two mainstream manual evaluation strategies. The methodology determines the relative order of merit of each model's output through direct duelling and then uses this order as a quantitative score. The algorithm pseudocode 1 shows the implementation steps in detail. The aim is to create a set of human evaluation processes that are both easy to implement and accurately reflect deep performance comparisons.

In the manual evaluation process, we randomly select 100 samples (50 in each transformation direction) from the dataset i, denoted as $X_i = \{x_{i,1}, \ldots, x_{i,j}, \ldots, x_{i,100}\}$, as well as the results of the rewritings of the different models for $x_{i,j}$, denoted as $T_{i,j} = \{t^1_{i,j}, \ldots, t^k_{i,j}, \ldots, t^n_{i,j}\}$, where $t^k_{i,j}$ denotes the rewriting result of model k on $x_{i,j}$, and n denotes the total number of models to be manually evaluated. In this paper, we ask the annotator to follow Algorithm 1 to rank $T_{i,j}$. After that, we get the final ranking of the models by Eq. 1, where $r^k_{i,j}$ denotes the ranking of model k on $x_{i,j}$, and R^k_i denotes the ranking of model k on dataset i.

$$R^k_i = \text{Average}\left(\sum_{j=1}^{100} r^k_{i,j}\right) \tag{1}$$

In this study, three graduate students with a specialization in linguistics and their second year of masters's degree were chosen as professional annotators after a rigorous selection process. Each annotator independently performed a comprehensive sequential annotation procedure for each of the textual materials that needed evaluation. We selected 100 samples from each of four test sets, totaling 400 samples for evaluation. To ensure the quality of the manual evaluation, before the evaluation, 10% of these samples were used to train annotators, helping them become familiar with the rules, thus ensuring accuracy and consistency in the evaluation. After that, an in-depth discussion and analysis of the initial annotation results was organized to establish unified annotation principles and standards, as well as to refine and improve the original annotation process. Following these standardization steps, the annotation team systematically annotated the data to be assessed according to the revised and improved annotation procedures.

5 Experiments

5.1 Model Setup

In order to measure the performance of LLMs on TST tasks, this paper selects representative LLMs, including [Qwen-7b-chat, Qwen-14b-chat] [2], [LLaMA2-7b-chat, LLaMA2-13b-chat] [36], Mistral-7b-instruct [10], GPT-3.5-turbo, and GPT-4 [1]. The baseline models include CrossAligned [33], [DeleteOnly, DeleteAndRetrieve] [17], [StyleEmbedding, MutliDec] [6], BT [24], DualRL [22], IMaT [12], StyTrans [5], [B-GST, G-GST] [34], StyIns [43], and TSST [41].

5.2 Analysis of Automated Assessment Indicators

The outcomes of the automated assessment can be found in Tables 4 to 7.[5] It is worth mentioning that due to the ability of LLMs to detect and eliminate sensitive words, style transfer procedures may fail in certain instances. For our analysis, only successful style transfer outcomes are taken into account.

On the content preservation dimension, although most of the LLMs show significantly lower BLEU scores compared to the small-scale

[3] https://huggingface.co/openai-community/gpt2-large

[4] https://huggingface.co/openai-community/gpt2

[5] As GENDER and POLITICAL did not manually modify the results, r-BLEU, g-BLEU, and BERTscore values are not present in the results.

Algorithm 1: Pseudocode of human evaluation rules.

Input : Source style sentence x, rewriting results for n models for x sampleList $= [y_1, \ldots, y_n]$, source style s, target style s'
Output : Manually sorted sampleList by transfer strength, content preservation, and fluency

scoresList ← []
foreach $y_i \in$ *sampleList* **do**
    // Transfer Strength
    // CountStyleWord: Counts the number of words in a sentence that specify a style.
    if *transfer success* **then**
        strengthScore ← CountStyleWord(y_i, s')
    else if *transfer failed* **then**
        strengthScore ← −CountStyleWord(y_i, s)
    else
        // not sure
        strengthScore ← 0
    // Content Preservation
    // ContentPreservationJudgment: Judging the degree of content preservation is subjective and difficult to quantify.
    if y_i *is related to* x **then**
        contentScore ← ContentPreservationJudgment(x, y_i)
    else if y_i *is not related to* x **then**
        contentScore ← −1
    else
        // not sure
        contentScore ← 0
    // Fluency
    // FluencyJudgment: Assign levels to sentences according to Table 3.
    fluencyScore ← FluencyJudgment(y_i)
    add (y_i, *strengthScore*, *contentScore*, *fluencyScore*) to *scoresList*
end
strengthSortedList ← *scoresList* in descending order of *strengthScore*
contentSortedList ← *scoresList* in descending order of *contentScore*
fluencySortedList ← *scoresList* in descending order of *fluencyScore*
return *strengthSortedList, contentSortedList, fluencySortedList*

Table 3: Human evaluation of fluency score criteria

FluencyScore	Criterion
6	Rich and varied grammatical structures
5	Accurate use of complex lexical features and rich vocabulary
4	Coherent and without grammatical errors
3	A trivial clerical error
2	A grammatical error
1	Multiple grammatical errors; Incoherent sentences

models, their performance on the BERTscore is comparable to that of the small-scale models. This is mainly because the BLEU metric focuses on exact matching at the n-gram level, whereas small-scale models tend to optimize directly on the training set, resulting in higher BLEU scores. In contrast, LLMs generate style-shifted text directly on the test set using cueing techniques that do not rely on the training corpus. With their broader linguistic knowledge, LLMs are better able to retain and transform semantic information when generating text, thus exhibiting better performance under the BERTscore rating system that reflects semantic coherence and similarity.

On the transfer strength dimension, LLMs generally perform lower than small-scale models, but the difference is not significant. Notably, LLMs excel in transforming objective factual statements in CAPTIONS into humorous or romantic style statements, indicating their proficiency in learning and mastering rhetorical skills. Thus, LLMs display superior adaptability and transformation effects for TST tasks involving complex rhetorical changes. We notice that LLMs have a high transfer success rate (Acc) for converting sentiment polarity and factual to romantic or humorous styles. However, they have a lower Acc when it comes to tasks involving political stance and gender tone switching. LLMs rely heavily on lexical, syntactic, and general context understanding to perform tasks such as sentiment polarity and humorous romantic styles. These skills are more common in large-scale pre-training datasets where LLMs can capture regular changes in affective tendencies and literary styles. However, tasks such as shifting party positions and gender tones are more complex and require more in-depth understanding of social and cultural connotations, political concepts, and gender culture. These elements are often difficult to express and quantify, resulting in difficulties in capturing complex social attributes and individual identity characteristics. Consequently, LLMs perform poorly when dealing with tasks requiring deep cultural background knowledge.

On the fluency dimension, most LLMs perform significantly better than small-scale models. This is because LLMs are capable of learning complex linguistic structures and expression patterns, which allows them to simulate the fluency and coherence of natural language more effectively when generating text. Experimental results have shown that this is the case.

5.3 Analysis of Human Evaluation Results

The manual evaluation results are presented in Table 8, which displays the mean and standard deviation of the model ranking, with a lower score indicating better performance. [6] In this study, we compared the latest baseline model to the large language model for each dataset.

In terms of content preservation, the small-scale models have achieved relatively high rankings, consistently staying within the top 50% across all datasets. This indicates a high level of consistency among the annotators. Among the LLMs, the GPT-3.5 and GPT-4 models show the best performance, holding the top 2 positions in all the TST tasks.

In terms of transfer strength, the small-scale model, TSST, is relatively lagging. GPT-3.5 and GPT-4 models perform better and are at the top of the list on all datasets among the LLMs.

In terms of fluency, the small-scale models have consistently performed poorly on all datasets. Among the LLMs, LLaMA2 has done well in the sentiment transfer task but not so well in the other three tasks. On the other hand, GPT-3.5 has performed best in the non-sentiment transformation task.

After conducting a comprehensive evaluation of the selected LLMs in this paper, it can be concluded that the GPT-3.5 and GPT-4 models perform remarkably well on all manual evaluation metrics. However, the Qwen model's overall performance is relatively low, especially

[6] It is worth noting that the Acc metrics are not included for the GENDER, POLITICAL, and CAPTIONS datasets, as evaluating the strength of TST requires specialized, in-depth knowledge.

Table 4: Results of the auto-evaluation metrics on the YELP dataset. The best results are bolded and the second best results are marked by *.

Model	s-BLEU	r-BLEU	g-BLEU	BERTscore	Acc(%)	d-PPL	g-PPL	t-PPL
CrossAligned	20.74	7.44	12.42	87.02	73.80	72.98	377.20	165.91
StyleEmbedding	**67.43**	17.71	34.56	89.37	8.90	57.57	321.76	136.10
MutliDec	40.07	12.18	22.09	87.39	51.37	142.19	629.98	299.29
D&R	36.75	13.44	22.22	88.21	88.70	57.73	272.96	125.53
StyTrans(Multi-Class)	63.12*	23.59*	38.59*	90.77	85.30	60.10	327.91	140.38
StyTrans(Conditional)	53.32	20.17	32.79	90.02	90.20	85.19	396.28	183.74
StyIns	53.10	20.76	33.20	90.01	91.40	59.26	333.67	140.61
DualRL	45.64	**48.85**	**47.22**	**94.67**	88.40	41.05	216.82	94.34
IMaT	16.92	8.66	12.10	88.92	**93.20**	**13.55**	115.12	**39.50**
G-GST	45.62	19.20	29.59	89.89	76.70	60.96	280.01	130.65
B-GST	45.22	19.24	29.49	90.48	85.50	37.98	201.11	87.40
TSST(dense)	59.21	23.01	36.97	90.64	91.70*	52.92	286.51	123.13
TSST(sparse)	58.90	22.81	36.66	90.63	90.60	50.70	283.77	119.95
Qwen-7b-chat	29.46	14.20	20.45	90.83	71.61	28.82*	94.84	52.28*
Qwen-14b-chat	26.85	13.47	19.02	90.66	66.90	31.31	94.44	54.38
LLaMA2-7b-chat	12.20	5.98	8.54	88.92	83.03	74.27	**55.87**	64.41
LLaMA2-13b-chat	14.69	7.54	10.52	89.26	80.67	58.50	67.81*	62.98
Mistral-7b-instruct	29.73	12.76	19.48	90.37	63.20	37.47	90.47	58.23
GPT-3.5-turbo	36.02	17.22	24.90	91.43*	73.80	36.71	132.82	69.83
GPT-4	31.45	16.07	22.48	91.22	81.00	47.43	160.54	87.26

Table 5: Automated evaluation metrics results for LLMs on the CAPTIONS dataset. The best results are bolded and the second best results are marked by *.

Model	s-BLEU	r-BLEU	g-BLEU	BERTscore	Acc(%)	d-PPL	g-PPL	t-PPL
CrossAligned	2.12	1.82	1.96	88.45	80.33	**12.78**	99.54	**35.67**
StyleEmbedding	32.04	8.80	16.79	88.70	54.83	78.69	443.19	184.53
MultiDec	23.65	6.65	12.54	88.67	70.83	46.58	288.28	115.88
DeleteOnly	38.75	11.99	21.56	89.38	78.67	47.85	277.94	115.32
D&R	32.07	12.00	19.62	89.52*	**96.67**	23.77	187.78	66.81
G-GST	48.09*	15.42*	27.22*	89.08	65.98	34.44	112.16	62.15
B-GST	**63.06**	**19.08**	**34.69**	**90.71**	69.83	24.89*	153.42	61.80
Qwen-7b-chat	4.61	2.04	3.07	87.17	88.83*	43.26	44.21*	43.73
Qwen-14b-chat	5.18	2.53	3.62	87.57	84.44	53.51	56.93	55.19
LLaMA2-7b-chat	3.45	1.33	2.14	84.98	86.05	52.61	**29.09**	39.12*
LLaMA2-13b-chat	9.47	3.45	5.72	87.18	79.66	64.73	48.00	55.74
Mistral-7b-instruct	13.37	5.21	8.34	88.07	76.67	47.83	63.93	55.30
GPT-3.5-turbo	5.72	2.58	3.84	87.24	73.17	70.93	70.16	70.55
GPT-4	3.24	1.64	2.31	87.13	71.83	114.66	117.21	115.93

Table 6: Results of the auto-evaluation metrics for the LLMs on the GENDER dataset. The best results are bolded and the second best results are marked by *.

Model	s-BLEU	Acc(%)	d-PPL	g-PPL	t-PPL
B-GST	**62.08**	52.30*	41.43	119.10	70.24
Qwen-7b-chat	43.51	33.57	**29.61**	73.18	46.55*
Qwen-14b-chat	49.65*	33.20	34.80	78.33	52.21
LLaMA2-7b-chat	18.48	**52.61**	31.69*	**48.51**	**39.21**
LLaMA2-13b-chat	15.46	51.77	45.57	52.69*	49.00
Mistral-7b-instruct	42.13	27.93	31.87	81.06	50.82
GPT-3.5-turbo	30.38	41.50	44.32	92.99	64.20
GPT-4	27.22	44.60	40.96	104.83	65.53

Table 7: Automated evaluation metrics results for LLMs on the POLITICAL dataset. The best results are bolded and the second best results are marked by *.

Model	s-BLEU	Acc(%)	d-PPL	g-PPL	t-PPL
BT	7.48	**80.30**	43.25	137.74	77.18
G-GST	51.56*	56.20	127.48	255.64	180.52
B-GST	**56.92**	60.70*	69.87	163.08	106.74
Qwen-7b-chat	17.18	43.70	38.74	55.04	46.18
Qwen-14b-chat	14.29	31.12	39.80	57.38	47.79
LLaMA2-7b-chat	10.09	29.55	34.42*	53.48*	42.91*
LLaMA2-13b-chat	9.64	30.81	**31.26**	**48.47**	**38.93**
Mistral-7b-instruct	21.51	26.90	40.10	68.55	52.43
GPT-3.5-turbo	14.58	29.60	62.32	90.05	74.92
GPT-4	17.69	35.00	87.91	117.69	101.72

when it comes to processing English corpus, which may be due to its limited adaptability and processing capability.

5.4 Comparative Analysis of Automated and Human Evaluations

In the evaluation of TST tasks, both the automated and manual evaluation systems have shown that small-scale models are better than LLMs in maintaining the original content. The conclusions obtained from the two evaluation methods are consistent. However, there is a significant difference between the automated and manual evaluation results in terms of the ability to switch styles. The automated assessment data indicates that small-scale models have an advantage in this area, but the manual review finds that LLMs perform better, showing inconsistent results between the two assessment methods. As for the criterion of text fluency, the two evaluation approaches together confirm that LLMs are better at generating coherent and naturalistic target texts.

In summary, automatic and manual evaluations can maintain consistency in terms of content preservation, but they show more noticeable differences when measuring transfer strength and fluency. Automatic evaluation technology is still unable to capture people's subjective feelings brought by the model in practical application scenarios. Therefore, relying solely on automatic evaluation methods may not be sufficient to comprehensively and accurately evaluate the performance of models in TST tasks. The duel-ranking method is an important supplementary evaluation method that can effectively reduce deviation caused by different subjective standards of evaluators through relative comparison, rather than giving absolute scores directly, thus improving the fairness and reliability of the evaluation. Duel-ranking emphasizes the distinction of small differences between model outputs and simplifies the evaluation process into sorting behavior, which is convenient for implementation and forming common judgments among multiple evaluators, thus improving the stability and reliability of the evaluation. Especially in examples like affective style transformation, duel-ranking vividly reflects the degree of adaptation of the model in the real situation, and the evaluators rely on the actual effect rather than isolated quantitative data to determine which model is closer to the ideal goal in affective style transformation. Furthermore, this approach can reveal deeper textual characteristics that automatic evaluation tools are not aware of. In the complex process of style transformation, while automatic evaluation systems may focus only on lexical consistency, contrast ranking rules can be sensitive to identify works that have successfully achieved a more subtle style

Table 8: Manual evaluation results. "Acc" is the success rate of style conversion, "Cont" is content preservation, and "Flu" is fluency. The best results are bolded and the second best results are marked by *.

Model	YELP			CAPTIONS		GENDER		POLITICAL	
	Acc	Cont	Flu	Cont	Flu	Cont	Flu	Cont	Flu
GPT-3.5-turbo	$\mathbf{2.15_{0.25}}$	$\mathbf{1.81_{0.28}}$	$4.69_{1.02}$	$2.41_{0.42}$*	$\mathbf{2.99_{0.72}}$	$\mathbf{2.01_{0.18}}$	$\mathbf{3.14_{0.34}}$	$\mathbf{1.89_{2.42}}$	$\mathbf{2.36_{0.39}}$
GPT-4	$2.17_{0.31}$*	$1.88_{0.38}$*	$4.59_{1.15}$	$2.89_{0.04}$	$4.03_{0.75}$	$2.26_{0.34}$*	$3.29_{0.63}$*	$1.97_{0.16}$*	$3.02_{0.64}$*
LlaMA-7b-chat	$4.77_{0.58}$	$4.83_{0.99}$	$\mathbf{2.49_{0.09}}$	$5.21_{0.08}$	$3.57_{0.89}$*	$5.38_{0.23}$	$3.88_{0.72}$	$5.95_{1.36}$	$6.61_{0.21}$
LlaMA-13b-chat	$4.40_{0.67}$	$4.56_{0.89}$	$3.01_{0.03}$*	$4.94_{0.11}$	$3.65_{0.13}$	$5.50_{0.40}$	$4.86_{0.25}$	$5.59_{1.10}$	$6.03_{2.29}$
Mistral-7b-instruct	$3.39_{0.76}$	$2.55_{0.95}$	$4.02_{0.14}$	$3.10_{0.19}$	$4.50_{0.85}$	$2.75_{0.47}$	$4.37_{0.81}$	$3.82_{0.06}$	$3.36_{0.78}$
Qwen-7b-chat	$5.27_{0.26}$	$4.84_{1.11}$	$3.45_{0.66}$	$5.30_{0.04}$	$4.17_{0.56}$	$5.33_{0.48}$	$4.44_{0.93}$	$3.87_{0.08}$	$3.10_{0.88}$
Qwen-14b-chat	$5.25_{0.37}$	$4.90_{1.13}$	$3.18_{0.69}$	$5.18_{0.08}$	$4.04_{0.62}$	$5.33_{0.33}$	$4.38_{0.65}$	$3.95_{0.10}$	$3.12_{0.75}$
TSST	$3.46_{1.08}$	$2.40_{0.71}$	$4.79_{0.35}$	-	-	-	-	-	-
B-GST	-	-	-	$\mathbf{1.60_{0.18}}$	$6.02_{1.49}$	$2.89_{0.13}$	$4.28_{0.94}$	$2.20_{0.17}$	$4.22_{0.87}$

Table 9: Compare the manual evaluation results and the automatic indicator evaluation results on the YELP dataset, where the evaluation indicator results of the values in brackets and the values outside brackets are the model rankings.

Model	Generation	Human			Automatic		
Negative → Positive		Cont	Flu	Acc	Cont	Flu	Acc
Source	there is definitely not enough room in that part of the venue .						
Human	there is so much room in that part of the venue						
TSST	service is definitely worth enough room in that part of the venue .	$7(6.33_{0.94})$	$8(6.33_{2.36})$	$4(3.67_{1.25})$	3(58.41)	8(339.97)	7(0.00)
Qwen-7B-chat	there is definitely plenty of room in that part of the venue .	$1(2.00_{0.82})$	$5(3.67_{0.47})$	$3(3.67_{0.94})$	2(59.07)	3(71.57)	3(1.02)
Qwen-14B-chat	the taste was refreshingly light and not overpowering .	$8(8.00_{0.00})$	$7(5.67_{0.94})$	$8(6.00_{2.83})$	7(0.00)	4(93.27)	5(0.08)
Mistral-7B	there definitely is enough room in that part of the venue .	$1(2.00_{0.82})$	$6(5.00_{2.16})$	$5(4.00_{0.82})$	1(62.10)	5(125.45)	4(0.12)
LlaMA-7B-chat	there is an abundance of space available in that particular area of the venue , providing ample room for an exciting and memorable event .	$6(6.33_{0.47})$	$2(2.33_{0.94})$	$1(1.00_{0.00})$	7(0.00)	1(50.54)	6(0.02)
LlaMA-13B-chat	there is ample space in that part of the venue , ensuring a comfortable and enjoyable experience for all attendees .	$5(5.33_{0.47})$	$1(2.00_{0.00})$	$2(2.00_{0.00})$	6(25.54)	2(53.23)	7(0.00)
GPT-3.5-turbo	there is certainly ample space in that part of the venue .	$1(2.00_{0.82})$	$3(3.00_{1.63})$	$6(4.33_{1.25})$	4(49.64)	6(135.24)	1(1.84)
GPT-4	there is certainly ample space in that part of the venue .	$1(2.00_{0.82})$	$3(3.00_{1.63})$	$7(4.67_{1.25})$	4(49.64)	6(135.24)	1(1.84)

Table 10: Comparison of the generation results of GPT-3.5 and the baseline model on different style transfer tasks.

Model	Generation
YELP: Positive → Negative	
Source	homemade tortillas are so good !
Human	these homemade tortillas aren't good at all.
TSST	the tortillas are so bad !
GPT-3.5-turbo	homemade tortillas are not good at all .
CAPTIONS: Factual → Humour	
Source	a toddler sits with diapers spread around the floor .
Human	a baby is sitting on the floor surrounded by diapers and basket with a smiley face .
B-GST	a toddler sits with diapers spread around the floor looking for santa .
GPT-3.5-turbo	a toddler decides to have a little fashion show with diapers scattered all over the floor , because who needs a red carpet when you have a diaper runway ?
GENDER: Male → Female	
Source	the designers of this restaurant did a splendid job at mixing textures and colors to make this a highly stimulating location .
B-GST	the designers of this restaurant did a wonderful job at mixing textures and colors to make this a highly stimulating location .
GPT-3.5-turbo	the designers of this restaurant did an absolutely fabulous job at blending textures and colors , creating a tremendously captivating spot for everyone to enjoy .
POLITICAL: Republican → Democratic	
Source	tim hiser maybe you should actually read the entire bill before you start throwing stones .
B-GST	hiser sanders maybe you should actually read the bill before you start throwing stones .
GPT-3.5-turbo	tim hiser , perhaps it would be beneficial for you to thoroughly read the entire bill prior to casting criticisms .

transformation while maintaining semantic appropriateness.

5.5 *Case Study*

Table 9 demonstrates a sample YELP dataset in the sentiment style transfer task along with its corresponding manual and automatic evaluations. It is worth noting that we use the g-BLEU indicator to capture aspects related to "Cont" and the t-ppl indicator to measure the performance of "Flu". The evaluation methods focus on measuring the consistency of the original text vocabulary before and after conversion for content preservation. The manual evaluation also takes into account the preservation of semantics and purpose along with the lexical-level preservation. However, the automatic evaluation metric only focuses on the degree of preservation at the lexical level. For fluency, the LLMs outperform the small-scale model, but there is some disagreement between the manual and automatic evaluation metrics. The manual evaluation prefers better results generated by GPTs, whereas the automatic evaluation shows higher quality generated by Qwen. For transfer strength, there is significant disagreement between the two evaluation metrics. The metric only considers the sentiment polarity of the sentence, and hence the automatic evaluation metric cannot effectively identify the actual change in the sentiment style of the converted text. It is worth noting that humans can easily find the style conversion of TSST to be successful, but the automatic evaluation metric cannot identify it effectively.

Table 10 shows a comparison between the best-performing LLM in manual evaluation, GPT-3.5, and the small-scale model. Both models performed well in the sentiment polarity conversion task, as the sentiment words were more pronounced. However, GPT-3.5's modifications were smaller than those of TSST, and therefore, GPT-3.5's modification results received a higher ranking in manual evaluation. In the factual to humorous style transfer task, GPT-3.5 generated content that enhanced the humor effect through clever metaphors and created a novel and imaginative situation. In the gender transfer task, GPT-3.5's generated sentences not only continued to praise the restaurant designer but also used words like "absolutely fabulous job" and "tremendously captivating spot", which are characteristic of the

female language style. [7] In the partisan style transfer task, the statements generated by the B-GST model did not sufficiently change the original strong, direct tone, while the statements generated by GPT-3.5 adopted a more euphemistic and rational way to make suggestions, which is more in line with the expression style of the Democratic Party.

In summary, manual evaluation is a more thorough and detailed way of assessing the effectiveness of text generation tasks, particularly when complex factors such as semantic, emotional, cultural, and social contexts are involved. Its accuracy and applicability are superior to the automatic evaluation system that relies solely on a single quantitative index. The duel-ranking method emphasizes the advantages of manual evaluation in evaluating NLP tasks, especially when they involve highly subjective tasks and require comprehensive judgment.

6 Conclusion

In this research, we investigate the ability of LLMs to perform zero-shot text style transfer on four popular non-parallel corpora. We conduct both an automated and manual evaluation to assess content preservation, transfer strength, and fluency, and compare the results to previous models. As opposed to prior human evaluation methodologies that relied on scoring, the duel-ranking approach adopted in this paper showcases greater resilience. To further fortify reproducibility, we have detailed comprehensive rules for conducting human evaluations. Our experiments show that the LLMs perform well in zero-shot text style transfer and can even outperform previous optimal models in some automated evaluation metrics. With the continued progress and optimization of LLM technology, we can expect the challenging problem of text style transfer might be solved in the future.

7 Limitations

We conducted a study to compare the performance of LLMs and small-scale models in TST. Previous studies have explored the effectiveness of small-scale models in this task, but they often lack source code and model outputs, making it difficult to replicate and verify results. To ensure an accurate and reliable comparison of previous research, we used published models that provide detailed source code and corresponding output data. We only used the test set portion of each dataset for comparison with existing models. As comprehensive testing was not carried out on all datasets, we plan to include the latest benchmark model in future research and expand the test data size for more representative and convincing results.

To evaluate the models, we introduced a comparative ranking strategy called duel-ranking. This helped reduce the training difficulty of taggers and obtain more objective and fair evaluation results. However, as the number of baseline models increases, the workload required for manual evaluation also increases, making it more challenging.

8 Ethics Statement

We make use of publicly available datasets and provide our computational evaluation metrics toolkit under the MIT license on GitHub. Our manual evaluations prioritize privacy protection and we ensure that no personal information is collected from the evaluator.

[7] This is a subjective judgment, and not all readers may agree that these words are feminine.

Acknowledgements

We appreciate the valuable comments of our anonymous reviewers. We are deeply appreciative of the support from the Natural Science Foundation of China, under Grants No. 62066044, 62167008, 62366040, and 62006130. Additional support was provided by Xinjiang Normal University's 2022 Young Top-Notch Talent Program (XJNUQB2022-23), the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2022D01A99), the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT24037), and the Fundamental Research Funds for the Central Universities(DUT24LAB123).

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] E. Briakou, S. Agrawal, K. Zhang, J. Tetreault, and M. Carpuat. A review of human evaluation for style transfer. In A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, and W. Xu, editors, *Proc. W-GEM*, pages 58–67, Online, Aug. 2021. Association for Computational Linguistics.

[4] Y. Cao, R. Shui, L. Pan, M.-Y. Kan, Z. Liu, and T.-S. Chua. Expertise style transfer: A new task towards better communication between experts and laymen. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proc. ACL*, pages 1061–1071, Online, July 2020. Association for Computational Linguistics.

[5] N. Dai, J. Liang, X. Qiu, and X. Huang. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proc. ACL*, pages 5997–6007, Florence, Italy, July 2019. Association for Computational Linguistics.

[6] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan. Style transfer in text: Exploration and evaluation. In S. A. McIlraith and K. Q. Weinberger, editors, *Proc. AAAI*, pages 663–670, New Orleans, Louisiana, USA, February 2018. AAAI Press.

[7] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. In *Proc. CVPR*, pages 3137–3146, Honolulu, HI, USA, July 2017.

[8] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444, 2023.

[9] A. Gatt and E. Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[11] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205, 2022.

[12] Z. Jin, D. Jin, J. Mueller, N. Matthews, and E. Santus. IMaT: Unsupervised text attribute transfer via iterative matching and translation. In *Proc. EMNLP-IJCNLP*, pages 3097–3109, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[13] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In M. Lapata, P. Blunsom, and A. Koller, editors, *Proc. EACL*, pages 427–431, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.

[14] H. Lai, A. Toral, and M. Nissim. Multidimensional evaluation for text style transfer using chatgpt. *arXiv preprint arXiv:2304.13462*, 2023.

[15] J. Lee. Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer. In *Proc. INLG*, pages 195–204, Dublin, Ireland, Dec. 2020. Association for Computational Linguistics.

[16] C. Li, Z. Leng, C. Yan, J. Shen, H. Wang, W. MI, Y. Fei, X. Feng, S. Yan, H. Wang, et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023.

[17] J. Li, R. Jia, H. He, and P. Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proc. NAACL*, pages 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[18] D. Liu, J. Fu, Y. Zhang, C. Pal, and J. Lv. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *Proc. AAAI*, pages 8376–8383, New York, NY, USA, February 2020. AAAI Press.

[19] R. Liu, C. Gao, C. Jia, G. Xu, and S. Vosoughi. Non-parallel text style transfer with self-parallel supervision. In *Proc. ICLR*, Virtual Event, April 2022.

[20] V. Logacheva, D. Dementieva, I. Krotova, A. Fenogenova, I. Nikishina, T. Shavrina, and A. Panchenko. A study on manual and automatic evaluation for text style transfer: The case of detoxification. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 90–101, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[21] L. Logeswaran, H. Lee, and S. Bengio. Content preserving text generation with attribute controls. pages 5108–5118, Montréal, Canada, Dec. 2018.

[22] F. Luo, P. Li, J. Zhou, P. Yang, B. Chang, X. Sun, and Z. Sui. A dual reinforcement learning framework for unsupervised text style transfer. In *Proc. IJCAI*, pages 5116–5122, Macao, China, 7 2019. International Joint Conferences on Artificial Intelligence Organization.

[23] P. Ostheimer, M. K. Nagda, M. Kloft, and S. Fellenz. A call for standardization and validation of text style transfer evaluation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proc. Findings-ACL*, pages 10791–10815, Toronto, Canada, July 2023. Association for Computational Linguistics.

[24] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black. Style transfer through back-translation. In I. Gurevych and Y. Miyao, editors, *Proc. ACL*, pages 866–876, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[25] A. Ramesh Kashyap, D. Hazarika, M.-Y. Kan, R. Zimmermann, and S. Poria. So different yet so alike! constrained unsupervised text style transfer. In *Proc. ACL*, pages 416–431, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[26] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

[27] P. K. Roy, S. Saumya, J. P. Singh, S. Banerjee, and A. Gutub. Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review. *CAAI Transactions on Intelligence Technology*, 8(1):95–117, 2023.

[28] S. Sabour, C. Zheng, and M. Huang. Cem: Commonsense-aware empathetic response generation. In *Proc. AAAI*, pages 11229–11237, Virtual Event, Feb. 2022. AAAI Press.

[29] E. Saravia. Prompt Engineering Guide. *https://github.com/dair-ai/Prompt-Engineering-Guide*, 12 2022.

[30] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[31] M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, pages 1–6, 2023.

[32] Y. Shao, L. Li, J. Dai, and X. Qiu. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, 2023.

[33] T. Shen, T. Lei, R. Barzilay, and T. S. Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Proc. NIPS*, pages 6830–6841, Long Beach, CA, USA, December 2017.

[34] A. Sudhakar, B. Upadhyay, and A. Maheswaran. "transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proc. EMNLP-IJCNLP*, pages 3269–3279, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[35] M. Toshevska and S. Gievska. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*, 3(5):669–684, 2022.

[36] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[37] R. Voigt, D. Jurgens, V. Prabhakaran, D. Jurafsky, and Y. Tsvetkov. RtGender: A corpus for studying differential responses to gender. In *Proc. LREC*, Miyazaki, Japan, May 2018.

[38] K. Wang, H. Hua, and X. Wan. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Proc. NeurIPS*, pages 11034–11044, Vancouver, BC, Canada, Dec. 2019.

[39] L. Wang, J. Li, Z. Lin, F. Meng, C. Yang, W. Wang, and J. Zhou. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proc. Findings-EMNLP*, pages 4634–4645, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[40] Z. M. Wang, Z. Peng, H. Que, J. Liu, W. Zhou, Y. Wu, H. Guo, R. Gan, Z. Ni, M. Zhang, et al. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023.

[41] F. Xiao, L. Pang, Y. Lan, Y. Wang, H. Shen, and X. Cheng. Transductive learning for unsupervised text style transfer. In *Proc. EMNLP*, pages 2510–2521, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[42] J. Xu, X. Sun, Q. Zeng, X. Zhang, X. Ren, H. Wang, and W. Li. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proc. ACL*, pages 979–988, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[43] X. Yi, Z. Liu, W. Li, and M. Sun. Text style transfer via learning style instance supported latent space. In *Proc. IJCAI*, pages 3801–3807, Online and Yokohama, Japan, 7 2020. International Joint Conferences on Artificial Intelligence Organization.

[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. In *Proc. ICLR*, Addis Ababa, Ethiopia, April 2020. OpenReview.net.

[45] X. Zhang, J. J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Proc. NIPS*, pages 649–657, Montreal, Quebec, Canada, Dec. 2015.

[46] C. Zheng, Y. Liu, W. Chen, Y. Leng, and M. Huang. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online, Aug. 2021. Association for Computational Linguistics.

[47] J. Zhou, C. Zheng, B. Wang, Z. Zhang, and M. Huang. CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proc. ACL*, pages 8223–8237, Toronto, Canada, July 2023. Association for Computational Linguistics.