Advancing Robustness and Instruction-following in LLM-Powered Multi-Style Text Rewritting

Yan Li Guangjun Wang Gang Qiao Zhenpeng Zhan
GBU,BaiduInc,China,518052
{liyan79,wanngguangjun,qiaogang01,zhanzhengpeng01}@baidu.com

Abstract

Now people frequently aim to polish their language or modify their text while using a mobile keyboard. However, leveraging large language models (LLMs) for multi-style text rewriting poses challenges, including semantic alteration, semantic addition, and task confusion, these challenges stem from the model's shortcomings in effectively following instructions and its lack of robustness. Particularly evident in handling non-standard, informal English rewrite requests, reflecting weaknesses in the model's robustness. To address these challenges, this study introduces a methodology that enriches training data with both positive and negative rewrite instances, accompanied by corresponding rationales. This approach aims to strengthen the model's discriminative capabilities. Additionally, adding Noisy Embedding during training is employed to enhance the model's robustness. Experiments validate the effectiveness of our methods in improving model directive adherence and robustness.

1 Introduction

In the era of mobile internet, social interactions on social apps are increasingly important. Expressing oneself with high emotional intelligence is a key aspect of social skills. Our analysis suggests that witty and creative language is well-received by people. With the success of GPT, domain-specific custom large language models are rapidly emerging in various fields, empowering numerous industries such as the internet, healthcare, and finance. Against this backdrop and considering the product form of mobile keyboard, we undertook a reevaluation and redesign of the keyboard functionality. Our analysis revealed that the younger generation of people not only has basic input needs but also a significant demand for personalized expression. Based on this insight and leveraging the technological advantages of LLM, we introduced the generation of long sentences or passage-level content into the keyboard functionality. This shift focuses on creatively rewriting input content to provide a more engaging and personalized input experience, satisfying the needs of the young people demographic.

Considering our research goals and challenges, we primarily approached the construction of the instruction dataset and training method. Addressing issues about semantic alteration, semantic addition, task confusion and poor generalization, we include both positive and negative rewrite examples in our own field-data, requiring the model to analyze the quality of examples and provide explanations. This is aimed at enhancing the self-developed large model's understanding of negative-class tasks, strengthening its ability to accurately follow instructions. Additionally, we introduce noise embedding during training to improve the model's robustness in handling misspelled words and non-formal English. We conducted experiments, evaluated by GPT-4 and three native English-speaking professionals, demonstrating a 2.3% reduction in the model's generation of incorrect rewrite results, and achieved a 0.25% improvement in rewriting quality.

2 Related Work

In the area of fine-grained text rewriting, Wu et al. [1] have made significant contributions with their work on a hierarchical reinforcement learning-based method for unsupervised text style transfer. Their approach, which employs a high-level agent to pinpoint locations for stylistic edits and a low-level agent to implement these modifications, provides a structured mechanism for altering text in line with specific style requirements without the need for parallel training data. Building on such unsupervised techniques, Raffel et al. [2] have demonstrated the adaptability of the T5 model to a range of natural language tasks, including style transfer. By pre-training on a large corpus and subsequently fine-tuning on texts representative of the target style, the T5 model effectively captures and generates text with the desired stylistic nuances. This showcases the model's potential for applications that aim to rewrite text to conform to various stylistic guidelines.

3 Method

In this section, we mainly introduce the attempts made to address the aforementioned rewriting issues, as well as our evaluation methods and results.

3.1 Challenges

We have identified several challenges in using state-of-the-art models like llama2 and GPT-3.5 for multi-style text rewriting tasks. Despite their advanced capabilities, these models often generate unsatisfactory results not only due to limitations in their inherent knowledge, but also due to various other rewriting issues. Specifically, we have observed the following problems:

1. Poor Generalization to non-formal English: Both models have difficulty with informal language, limiting their ability to generate natural-sounding text in less formal styles.

2. Semantic Alteration: In some instances, the rewritten output diverges significantly from the original meaning of the input, leading to an entirely different message being conveyed, which is not the intent of style transfer.

3. Semantic Addition: The models occasionally introduce unnecessary additional information in the latter part of the rewritten text, which, while related, was not present in the original input and is not required for the task at hand.

4. Task Confusion: There are cases where the models misconstrue the rewriting task as a different NLP task, such as responding to a prompt rather than rewriting it.

3.2 Data

In this study utilized three primary types of datasets. Firstly, we employed the official 52,000 utility instruction dataset from Alpaca, created using the Self-Instruct method introduced by Wang et al. [3] This method involves the construction of new instruction adjustment tasks from a small seed task set, followed by the filtration of low-quality tasks to improve dataset quality. However, as the Alpaca dataset alone did not fully meet the expression needs of young people who are currently good at symbolic expression and humorous expression, we augmented the fine-tuned instruction data with field-collected data which is our proprietary high-quality multi-style rewriting dataset. This additional dataset, comprising 32k entries, was curated by five professional English writers and covers various styles including humor, whimsy, emotion enhancement, and meme-based rewriting. Furthermore, to enhance the distinction between different rewriting styles and bolster the model's instruction compliance, we incorporated both positive and negative examples. These additional 5k samples were selected by professional English writers based on flawed rewriting results obtained from tests with GPT-3.5 and llama2_alpaca, in accordance with the Chain of Hindsight (CoH) proposed by Liu at al. [4], aiming to train the model to recognize and self-critique instances of semantic alteration, semantic addition, and task confusion in rewriting tasks. The composition of our training data is summarized in the table 1.

Table 1 dataset composition

Dataset	sample size
Alpaca	52k
Field-collected data	32k
Self-critique data	5k

3.3 Robustness

To enhance the robustness of the model, we conducted a sampling analysis of commonly used language among the current younger demographic. Among the 500 sampled inputs, we observed common occurrences of abbreviations, misspellings, and informal colloquial English expressions. The distribution of these language forms indicated that formal written English constituted only 55% of the inputs. Consequently, we decided to employ the Noisy Embedding Instruction Fine Tuning (NEFT) method [5] during the fine-tuning of the model. There is a common strategy to enhance generalization performance is to introduce random noise into the embedded vectors, this noise is generated by independently sampling from a uniform distribution between -1 and 1, and then multiplying the result by a scaling factor. Therefore, we adopted the NEFT method to optimize the rewriting effectiveness for informal English expressions.

3.4 Model and Baselines

In our study, we compared the performance of the llama2 model fine-tuned with Alpaca's dataset (llama2_alpaca), GPT-3.5 and llama2 fine-tuned with a combination of Alpaca and our field-collected data (llama2_alpaca_field-data), against a refined version of the llama2 model (ours), this refined version was fine-tuned using the NEFT method and augmented with additional positive and negative examples crafted through the CoH approach. Our aim was to demonstrate the effectiveness of these approaches in addressing rewriting quality and mitigating semantic alteration, semantic addition, and task confusion issues.

3.5 Evaluation Metrics

Our primary evaluation focused on two aspects: rewriting quality and the prevalence of semantic alteration, semantic addition, and task confusion issues. For rewriting quality, we assessed the level of interest, style coherence, appropriateness of generated emojis, and overall diversity, using a scoring system ranging from 1 to 5 to indicate satisfaction. Additionally, we measured the frequency of semantic alteration, semantic addition, and task confusion instances. We employed two evaluation methods: manual assessment by three native English-speaking professionals and an automated assessment using GPT-4. In both cases, the evaluators were tasked with scoring the rewriting results based on predefined quality criteria and identifying instances of semantic alteration, semantic addition, and task confusion. This comprehensive evaluation approach allowed us to effectively gauge the rewriting quality and the prevalence of undesirable rewriting behaviors.

4 Experiments

In order to facilitate multi-style text rewriting, we leverage the llama2-7b model for fine-tuning due to its notable balance between performance and efficiency. The llama2 model, available in 7 billion, 13 billion, and 70 billion parameter versions, offers a suite of options that cater to different computational power and complexity needs. Its Grouped-Query Attention (GQA) mechanism reduces the memory requirements of the Large Language Models (LLMs), lowering the computational cost per byte and allowing for the processing of more requests concurrently.

Additionally, we ensured data balance across different text styles to prevent biases in rewriting results. This involved controlling the quantity of instructions for each style to achieve a balanced distribution, mitigating the risk of the rewriting leaning towards a particular style. Furthermore, as each style was crafted by multiple writers, repetitions of expressions were common, potentially compromising the diversity and generalization capability of the generated results. To address this, we employed n-gram to eliminate repetitive expression forms and enhance result diversity and generalization.

We trained the llama2-7b model for 2 epochs with a batch size of 4, utilizing a learning rate of $2*10^{-4}$ and the Adam optimizer, alongside linear scheduling with a warm-up rate of 0.1.

5 Results and Analysis

We randomly selected 531 common language samples used by the younger demographic on social media platforms and generated three rewriting results for each input. As shown in Tables 2 and 3, which present the ratings and assessments of our 1593 rewriting results by both GPT-4 and human evaluators, the inclusion of our proprietary rewriting data has significantly improved the quality of our rewriting compared to using llama2 fine-tuned only with Alpaca, and has slightly surpassed the performance of GPT-3.5. This underscores the effectiveness of our proprietary rewriting data. However, our proprietary data incorporates some commonly used phrases among the younger demographic, thereby addressing certain issues related to semantic alterations, semantic additions, and task confusion. However, it falls short compared to GPT-3.5.

Following the incorporation of the CoH data and the application of NEFT fine-tuning, our approach demonstrated significant improvements in addressing semantic alteration and task confusion when compared to other models, The rewrite failure rate has decreased by

2.3% overall compared to llama2_alpaca_field-data, resulting in a 0.25% improvement in rewrite quality. However, it appears that the issue of semantic addition did not exhibit noticeable optimization. This indicates that our approach can notably enhance the model's understanding of rewriting styles and semantic consistency before and after rewriting. However, due to the specific requirements of our rewriting task, which involve rewriting in a particular style and the addition of specific emoticons, determining whether semantic addition has occurred is challenging for the model.

Table 2 gpt-4 evaluate result

Models	Quality	Semantic Alteration(%)	Semantic Addition(%)	Task Confusion(%)
llama2_alpaca	3.23	23.53	13.35	8.68
Gpt-3.5	4.40	17.32	10.10	5.37
Llama2_alpaca_field-data	4.38	17.53	**9.87**	6.37
Ours	**4.41**	**16.49**	9.88	**5.10**

Table 3 human evaluate result

Models	Quality	Semantic Alteration(%)	Semantic Addition(%)	Task Confusion(%)
llama2_alpaca	3.19	21.37	12.76	8.92
Gpt-3.5	4.25	15.76	10.51	5.35
Llama2_alpaca_field-data	4.33	16.02	10.31	5.92
Ours	**4.34**	**15.37**	**10.30**	**5.30**

Furthermore, during the analysis of non-formal English cases, we observed that our method demonstrated improved handling of informal expressions, such as misspellings and colloquial language, when compared to llama2_alpaca_field-data and GPT-3.5, thus underscoring the effectiveness of the NEFT method.

6 Summary

Through our research, we utilized the llama2 model for multi-style text rewriting and introduced the CoH method and the NEFT method. Our experimental results demonstrated a significant improvement in rewriting quality after incorporating our proprietary rewriting data and applying the NEFT method, especially in handling colloquial expressions and misspellings, outperforming GPT-3.5. Moreover, by training the model with additional positive and negative examples, we enabled the model to learn self-criticism and recognize instances of semantic alteration, semantic addition, and task confusion in the rewriting task. Overall, our study has validated the effectiveness of the CoH method and the NEFT method in enhancing rewriting quality and model robustness, offering valuable insights for the task of multi-style text rewriting.

References

[1] Wu, Chen, et al. "A hierarchical reinforced sequence operation method for unsupervised text style transfer." arXiv preprint arXiv:1906.01833 (2019).

[2] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.

[3] Wang, Yizhong, et al. "Self-instruct: Aligning language model with self generated instructions." arXiv preprint arXiv:2212.10560 (2022).

[4] Chen, Kai, et al. "Gaining wisdom from setbacks: Aligning large language models via mistake analysis." arXiv preprint arXiv:2310.10477 (2023).

[5] Jain, Neel, et al. "NEFTune: Noisy Embeddings Improve Instruction Finetuning." arXiv preprint arXiv:2310.05914 (2023).