Distilling Text Style Transfer With Self-Explanation From LLMs

Chiyu Zhang[1,2,⋆]   **Honglong Cai**[2]   **Yuezhang (Music) Li**[2]   **Yuexin Wu**[2]
Le Hou[2]   **Muhammad Abdul-Mageed**[1,3]
[1] The University of British Columbia   [2] Google
[3]Department of NLP & ML, MBZUAI
chiyuzh@mail.ubc.ca, {honglongcai, lyzmuisc}@google.com

Abstract

Text Style Transfer (TST) seeks to alter the style of text while retaining its core content. Given the constraints of limited parallel datasets for TST, we propose CoTeX, a framework that leverages large language models (LLMs) alongside chain-of-thought (CoT) prompting to facilitate TST. CoTeX distills the complex rewriting and reasoning capabilities of LLMs into more streamlined models capable of working with both non-parallel and parallel data. Through experimentation across four TST datasets, CoTeX is shown to surpass traditional supervised fine-tuning and knowledge distillation methods, particularly in low-resource settings. We conduct a comprehensive evaluation, comparing CoTeX against current unsupervised, supervised, in-context learning (ICL) techniques, and instruction-tuned LLMs. Furthermore, CoTeX distinguishes itself by offering transparent explanations for its style transfer process.

1   Introduction

TST aims to rephrase a source text s with the desired style τ while preserving its core meaning and ensuring fluency of the generated text t (Jin et al., 2022). The term "style" can encompass the personal characteristics of an author, such as age, and pragmatic use like formality or toxicity. To develop TST systems using supervised methods, several human-annotated datasets have emerged (Rao and Tetreault, 2018). For instance, Rao and Tetreault (2018) introduced a corpus for formality style transfer, transforming informal language to its formal counterpart and vice versa. Nonetheless, supervised parallel data, crucial for training deep neural networks, is scarce and costly to obtain. Hence, unsupervised methodologies (Shen et al., 2017; Liu

et al., 2021) have been proposed to manage stylistic attributes without relying on parallel data. Liu et al. (2022) and Zhang et al. (2020) create pseudo-parallel data from unlabeled samples via diverse data augmentation with task-specific knowledge. Works by Gong et al. (2019); Wang et al. (2019a); Reid and Zhong (2021) employ an auxiliary style classifier to steer the transfer direction. Meanwhile, Krishna et al. (2020) and Hallinan et al. (2023b) deploy multiple style-specific models to produce various styles individually. Of late, LLMs have demonstrated exceptional prowess across diverse NLP tasks. Studies like Reif et al. (2022); Pu and Demberg (2023) have found that extremely large LMs, with over 100B parameters, are adept at TST with ICL. Drawing from these findings, our paper uses LLMs to generate pseudo-parallel data and distills the TST skills of the LLM into a compact student model. Moreover, we enhance distillation and efficiency using CoT prompting.

LLMs have demonstrated impressive performance across various tasks and reasoning capabilities. CoT prompting (Wei et al., 2022) is a promising technique that extracts these reasoning skills and enhances accuracy in target tasks. However, deploying these enormous LLMs poses computational and practical challenges. Recent studies (Huang et al., 2022; Wang et al., 2023a; Hsieh et al., 2023) have thus turned to offline knowledge distillation (KD) (Hinton et al., 2015) to condense these reasoning capabilities into a smaller model. Using CoT rationales can also increase distillation efficiency with less data (Li et al., 2022; Shridhar et al., 2023). Concurrently, Saakyan and Muresan (2023) examine CoT prompting combined with domain expert feedback for improved formality transfer. Nevertheless, the potential of CoT prompting and KD to enrich a broader range of TST tasks remains underexplored.

In this paper, we present CoTeX framework, using CoT prompting to improve TST. It identifies
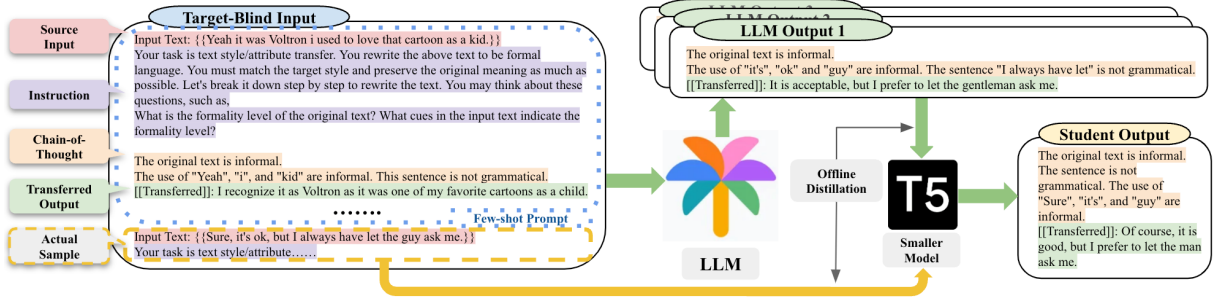

⋆Work done during internship at Google.

200

Figure 1: Overview of CoTeX framework. We use few-shot CoT prompting to generate reasoning paths and transferred texts from an LLM and then train a smaller task-specific model with generated data.

cues for TST and clarifies the rewriting process (§ 2). We then distill the reasoning and style transfer abilities of LLMs into compact models. We exploit CoT prompting to enhance TST, applicable to scenarios both with and without parallel data, and show the effectiveness of CoTeX in low-resource settings. Our ***primary findings*** include: **(1)** Our target-blind CoTeX (CoTeX-TB) substantially boosts data efficiency for training smaller student models for TST. **(2)** The target-aware CoTeX (CoTeX-TA) consistently outperforms SFT and conventional KD across various datasets and training data sizes. **(3)** Our CoTeX-TB outperforms state-of-the-art (SoTA) unsupervised and ICL methods on three TST datasets. **(4)** Leveraging CoT rationales, our distilled student models can elucidate the rewriting procedure.

2 Method

2.1 Data Generation

We employ CoT combined with instruction prompting to extract rationales from LLMs regarding the TST process. We have two different settings (target-blind and target-aware) to generate rationales.

Target-Blind (TB). We first explore our method in the target-blind setting where we only give a source text and the name of the desired target style. This setting can be adaptable to a broader range of style transfer directions. As shown in the left side of Figure 1, each input example is constructed using an instruction template, p_{tb}. This template encompasses a source input s_i, a task instruction, the target style τ, and a CoT trigger phrase: "Let's break down the rewriting process step by step." LLM is tasked with producing the CoT, c_i, pertaining to the text rewriting process and the resultant transferred text, $\hat{t}_i$. To distinguish between the CoT and the transferred text, we instruct the model to

initiate the transferred text on a new line, prefixed with a special token '[Transferred]:'. To facilitate the LLM's adherence to the desired output structure, we present m examples created by humans as context before the actual input. In our implementation, we employ three manually crafted examples as few-shot prompts.

Target-Aware (TA). For datasets with supervised parallel data, we use the instruction template p_{ta}.[1] As Figure 2 shows, this template p_{ta} integrates a source text s_i, its corresponding human-annotated target text t_i, and the target style τ. The LLM is then prompted to explain how s_i is transformed into t_i, leading it to produce a CoT, c_i. This generated CoT is prefixed with a distinct token '[EXPLANATION]:'. To ensure LLMs produce outputs in the desired format, we also employ m guiding examples, $m = 3$ in our experiments.
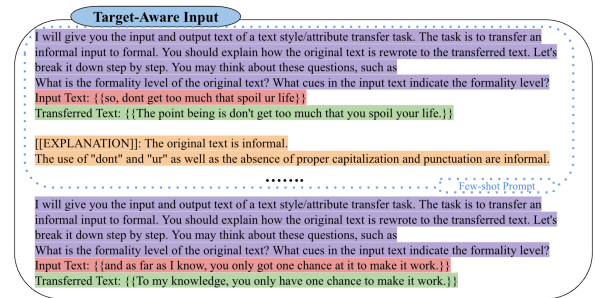

Figure 2: Few-shot chain-of-thought prompting for data generation with supervised data (target-aware setting). We use the few-shot prompts that include a few examples to guide LLM to generate desired outputs in a standard format.

2.2 Training Student Models

We leverage the LLM-generated data to finetune smaller, task-specific student models. For the data

---

[1]We manually tune both instruction templates and find the optimal templates used in this paper.

generated in the target-blind setting, we utilize the instruction template p_{tb}, which includes source text s_i and target style τ as input. The corresponding supervision $\hat{y}_i$ for training is the generated CoT c_i combined with the synthetic transferred text $\hat{t}_i$, i.e., $\hat{y}_i = c_i \oplus \hat{t}_i$. When employing the data generated from the target-aware setting, we also adopt the template p_{tb}. From this, we derive the generated CoT c_i and merge it with the gold target text t_i. This composite then serves as the supervision $\hat{y}_i$ (i.e., $\hat{y}_i = c_i \oplus t_i$) for training a student model. A student model is trained with $\hat{y}_i$ employing the conventional cross-entropy loss.

3 Experiments

3.1 Datasets and Metric

We employ four public datasets across three style transfer directions, chosen for their inclusion of human-annotated parallel data in both training and evaluation sets. This facilitates direct comparisons between different settings.

Formality Transfer. We use GYAFC dataset from Rao and Tetreault (2018) and focus on the *informal to formal language* transfer direction. GYAFC dataset includes two domains, Family & Relationships (F&R) and Entertainment & Music (E&M). **Detoxification.** ParaDetox (Logacheva et al., 2022b) is a parallel dataset for text detoxification. **Shakespeare to Modern English.** Xu et al. (2012) introduce a human-annotated dataset for translating text between William Shakespeare's plays and their modernized versions.

Low-Resource Training. Our method offers advantages in low-resource settings, as the CoT is poised to enhance the learning efficiency of student models and bolster their generalizability. Thus, we create smaller training sets by randomly sampling training data, ranging from 1K to 20K.

Evaluation Metric. We report BLEU, leveraging the Sacre-BLEU Python library (Post, 2018), as main metric for evaluation.

3.2 Model Comparison.

In low-resource settings, CoTeX is compared to **(1) SFT:** conventional supervised fine-tuning using parallel data, **(2) teacher LLM:** the teacher model evaluated on the Test set via few-shot ICL, i.e., using the three-shot prompt and template p_{tb} described in Section 2, and **(3) Distill:** traditional offline knowledge distillation, which relies solely on LLM-generated pseudo-parallel data without a CoT path.

For comprehensive evaluations, CoTeX is further compared with **(1) Prompt&Rank:** a SoTA in-context learning method for TST (Suzgun et al., 2022), and **(2) instruction-tuned LLMs:** open-source LLMs assessed through three-shot ICL using the same prompt and template described in Section 2; these LLMs include Alpaca 7B (Taori et al., 2023), Vicuna 7B (Chiang et al., 2023), LLaMA2-Chat 7B (Touvron et al., 2023), and FlanT5-XL (Chung et al., 2022) (with 3B parameters). Additionally, for each dataset, comparisons are made with existing *dataset-specific* unsupervised and supervised methods. **Unsupervised** methods include DualRL (Luo et al., 2019), STRAP (Krishna et al., 2020), DLS (He et al., 2020), and TSST (Xiao et al., 2021) for formality transfer; Mask&Infill (Wu et al., 2019) and CondBERT (Dale et al., 2021) for detoxification; and STRAP and TSST for modernizing Shakespearean text. **Supervised** methods include Multi-NMT (Niu et al., 2018), GPT-CAT (Wang et al., 2019b), and SemiFST (Liu et al., 2022) for formality transfer; ParaDetox (Logacheva et al., 2022a) for detoxification; and PointerS2S (Jhamtani et al., 2017) for modernizing Shakespearean text.

3.3 Implementation

We employ PaLM2 Unicorn (Anil et al., 2023) as our LLM for data generation. In the target-blind setting, we generate a CoT path and a transferred text.[2] For the target-aware approach, we solely produce a CoT path. Both approaches use a temperature of 0.7. Afterward, we finetune a T5-large model (with 770M parameters) (Raffel et al., 2020) with the curated dataset.[3] We finetune T5 for 2,000 steps with a learning rate of $1e-3$ and batch size of 128. We evaluate validation performance every 16 steps and report test result of the best step.[4]

3.4 Hyperparameter for Training Student Model

We set the maximal input and output sequence lengths to 512 and 256, respectively. To optimize the T5 model's finetuning, we search both the learning rate and batch size within specified search spaces: $lr \in \{1e-3, 5e-4, 1e-5\}$ and

---

[2]Our ancillary study also examines the generation of multiple pairs of CoT paths and transferred text.

[3]We provide a concise experiment of using T5-XL model in Appendix B.

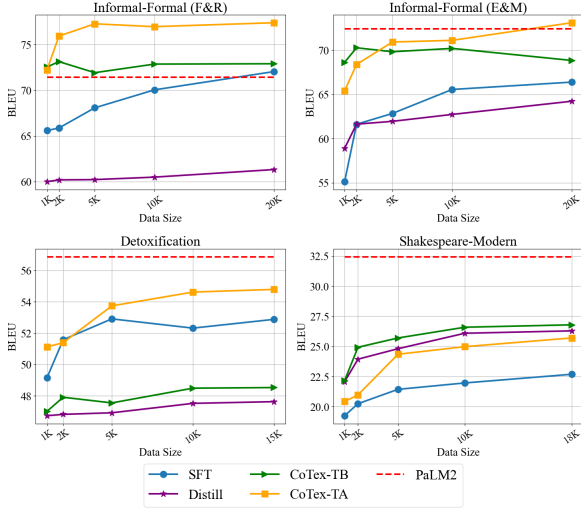[4]More details about hyperparameters are in Appendix 3.4.

Figure 3: Test results of low-resource settings.

	Method	BLEU	Method	BLEU
	Formality (F&R)		**Formality (E&M)**	
Unsup.	DualRL	53.01	DLS	23.09
	TSST	60.99	STRAP	31.39
ICL	Prompt&Rank	30.60	Prompt&Rank	30.96
	Alpaca	41.85	Alpaca	52.40
	Vicuna	37.09	Vicuna	46.47
	LLaMA2-C.	19.62	LLaMA2-C.	25.14
	FlanT5-XL	55.70	FlanT5-XL	42.58
Sup.	Multi-NMT[†]	75.35	Multi-NMT[†]	72.01
	GPT-CAT[†]	77.26	GPT-CAT[†]	71.39
	SemiFST	**80.32**	SemiFST	**76.87**
	SFT (ours)	77.12	SFT (ours)	73.01
	Distill (ours)	64.79	Distill (ours)	64.31
	CoTex-TB	<u>72.05</u>	CoTex-TB	<u>71.70</u>
	CoTex-TA	77.13	CoTex-TA	74.65
	Detoxification		**Modernizing Shake.**	
Unsup.	Mask&Infill[*]	44.77	DLS	12.85
	CondBERT[*]	48.89	STRAP	19.96
ICL	Prompt&Rank	11.06	Prompt&Rank	20.87
	Alpaca	24.32	Alpaca	24.33
	Vicuna	34.54	Vicuna	17.76
	LLaMA2-C.	14.65	LLaMA2-C.	25.19
	FlanT5-XL	<u>50.13</u>	FlanT5-XL	21.55
Sup.	ParaDetox	53.98	PointerS2S	**30.78**
	SFT (ours)	52.88	SFT (ours)	22.69
	Distill (ours)	43.97	Distill (ours)	22.88
	CoTex-TB	48.53	CoTex-TB	<u>26.79</u>
	CoTex-TA	**54.79**	CoTex-TA	25.70

Table 1: Comparing to previous methods. The best-performed method is in **bold**. The best method without utilizing a full parallel Train set is <u>underscored</u>. **Unsup.:** unsupervised, **Sup.:** supervised, [†]: Take from Liu et al. (2022). [*]: Utilize outputs from implementation of Logacheva et al. (2022b).

batch size $\in \{32, 64, 128\}$. We undertake hyperparameter tuning using formality (F&R) dataset. Based on the validation BLEU score, we identify the optimal hyperparameters are $lr = 1e - 3$ and batch size $= 128$. We finetune T5 for 2,000 steps, evaluate performance on the validation set every 16 steps, and report the test performance on the best step. All T5 models are trained on four V3 TPUs.

4 Results

We now present your experimental results. CoTeX-TB and CoTeX-TA denote models trained using datasets created through target-blind and target-aware methods, respectively.

Low-Resource Settings. We first examine Co-TeX's impact in low-resource context. Figure 3 shows CoTeX's performance in both target-blind and target-aware settings across varying training data sizes. In both formality transfer datasets, CoTeX-TB outperforms SFT-T5 and Distill-T5. This advantage is noticeable with limited data, specifically under 10K. For instance, using just 1K samples from the informal-formal (E&M) dataset, the BLEU scores for SFT, CoTeX-TB, and CoTeX-TA are 55.13, 68.62, and 65.40, respectively. We find that both CoTeX-TB and CoTeX-TA outperform or match the LLM's performance on the two formality datasets. In translating Shakespearean to modern English, CoTeX-TB exhibits significant superiority over SFT-T5 and Distill-T5 across all data sizes. We believe that such an enhancement can be attributed to the high quality of LLM generations. LLM with few-shot in-context learning obtains a BLEU score of 32.43. Though CoTeX-TB under-

performs SFT on detoxification, CoTeX-TA still outperforms SFT in most data sizes.

Utilizing the Full Dataset. Training student models with CoTeX on all training samples of each dataset, we present comparative results in Tables 1.[5] Given that many unsupervised TST studies have not reported BLEU scores, we compute BLEU scores for their public outputs using our evaluation scripts to ensure a fair comparison. CoTeX-TB surpasses previous unsupervised methods, the SoTA ICL method Prompt&Rank, and instruction-tuned LLMs across both domains within the formality transfer dataset. Although CoTeX-TA does not exceed the performance of SoTA supervised methods, SemiFST, for formality transfer, it is noteworthy that our method does not depend on task-specific data augmentation strategies or knowledge, offering greater flexibility. In the detoxification task, our results are compared with the top-performing model from Logacheva et al. (2022b). CoTeX-TA outperforms previous supervised methods, while CoTeX-TB falls slightly

---

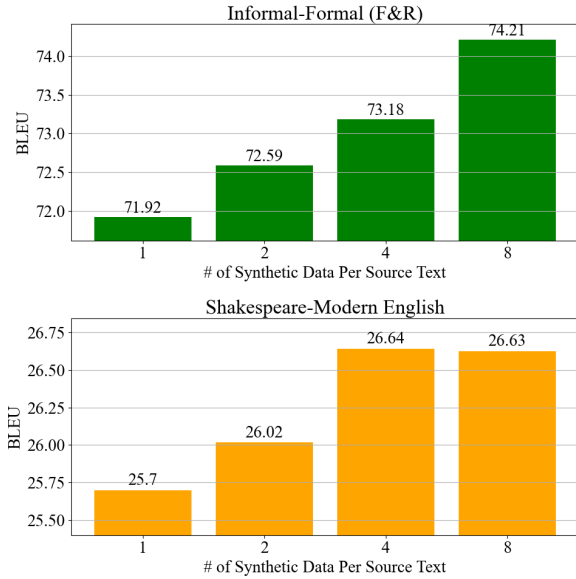[5]CoTeX-TB setting utilizes the source text from training sample while keeping the target undisclosed.

Figure 4: Scaling up synthetic data per source.

Source Text	just want to know HAVE YOU BEEN TO THE DOC YET
CoTeX Generation	The original text is informal. The use of all caps and the absence of punctuation are informal. The use of "DOC" is a misspelling of "doctor". [[Transferred]]: I just want to know if you have been to the doctor yet.
Reference	I want to know if you have been to the doctor yet.
Source Text	bembie hit the nail on the head
CoTeX Generation	The original text is informal. The use of "bembie" is a misspelling of "bembie". [[Transferred]]: Bembie hit the nail on the head.
Reference	Bembie reached the proper conclusion.

Table 2: Case study on CoTeX-TB generations.

short of CondBERT, which employs additional style-conditional LMs for transfer control. FlanT5-XL, an instruction-tuned LLM, leads in ICL performance with a BLEU score of 50.13. For translating Shakespearean to modern English, CoTeX-TB shows marked improvements over both unsupervised and ICL methods, attributed to the superior quality of LLM generations in this specific transfer task.

Increasing Synthetic Data per Source Text. For CoTeX-TB, we conduct an ancillary study to explore the benefits of employing multiple CoT paths with synthetic target texts for a source text. Given a source text s_i, the LLM generate q CoT paths, $\{c_{i,1}, c_{i,2}, \ldots, c_{i,q}\}$ and their corresponding synthetic target text $\{\hat{t}_{i,1}, \hat{t}_{i,2}, \ldots, \hat{t}_{i,q}\}$. We select a subset of 5K unique source texts as inputs and investigate the effect of q over a range of $\{2, 4, 8\}$. We experiment with two datasets, Formality (F&R) and Shakspeare-modern English. Table 4 shows a positive correlation between the student model's performance and increasing q values.

Level	Criteria
Rate A	• Valid, acceptable and satisfying (subject to the annotator) response; • Accurately identified the most cues for text style transfer; • The reasoning path can directly lead to the transferred text.
Rate B	• The response is acceptable but has minor errors that can be improved; • Mirror errors include out-of-context content, minimal factual errors, missing many cues for text style transfer, etc.
Rate C	• The response is relevant but it has significant errors in the content; • Cannot identify any correct cues for text style transfer. • The reasoning path cannot lead to the transferred text.
Rate D	• Invalid and unacceptable response; • Nothing related to the text style transfer task.
Instruction: This task is text styles transfer that transfers a {$source_style} source text to a target text with style {$target_style}. Each example includes a source text and the corresponding model-generated rationales of the rewriting process as well as the transferred text. You evaluate the rationales of the rewriting process and do not take the quality of the transferred text into account.	

Table 3: Human evaluation protocol and instruction. We adapt the evaluation criteria from Wu et al. (2023) and Wang et al. (2023b).

Qualitative Study. We now present a qualitative study to delve into rewriting rationales generated by CoTeX-TB. Examples are showcased in Table 2, derived from Test set of Formality (F&R) which transfers from informal to formal text. We sort generations by their BLEU scores against gold references and select random high and low-scoring samples. The first example, obtained BLEU of 100, correctly identifies informal components, fixes informal spellings, and yields a formal and grammatical sentence. The second example (BLEU=7.27) misses comprehending the idiom "hit the nail on the head" from the source, without translating it into a formal expression. Nevertheless, we note that the LLM (i.e., PaLM2) can appropriately adapt this idiom to "accurately identified the key point". This leads us to hypothesize that a smaller LM exhibits potential limitations in its ability to understand implicit style cues.
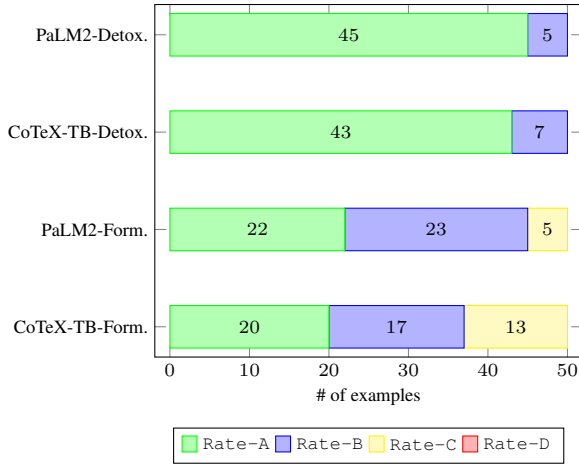
Figure 5: Human evaluation results of CoT reasoning paths of 50 samples. **Form.**: formality transfer, **Detox.**: Detoxification.

Human Evaluation on Generated Reasonings. To assess the quality of model-generated rationales (i.e., CoT path) for the rewriting process, we conduct a human evaluation. Following previous works (Wang et al., 2023b; Wu et al., 2023), we develop our evaluation protocol and instructions as shown in Table 3. We assemble a team of four human experts to undertake this evaluation. Each annotator was tasked with reviewing 50 generated rationales across different models and transfer tasks. For each evaluation, the dataset provided included the source text, a generated rationale for the rewriting process, and the resultant transferred text. As depicted in Figure 5, although CoTeX-TB lags behind the teacher model (PaLM2 Unicorn), 100% of its responses in the detoxification task and 74% in the formality transfer task are deemed acceptable.

5 Related Work

When parallel TST datasets are available, numerous studies (Rao and Tetreault, 2018; Shang et al., 2019; Chawla and Yang, 2020; Lai et al., 2021) have utilized a sequence-to-sequence framework for supervised training TST models. To improve model efficacy, multitask learning (Niu et al., 2018; Xu et al., 2019), lexically constrained decoding (Post and Vilar, 2018), and task-specific data augmentation (Zhang et al., 2020; Liu et al., 2022) have been incorporated. Addressing the scarcity of parallel data, unsupervised methods have been developed for TST, employing methodologies like disentanglement of latent representations (Liu et al., 2020; Nangi et al., 2021; Yi et al., 2021), proto-

type editing (Li et al., 2018), style rewriting using attribute-specific LMs (Krishna et al., 2020), and reinforcement learning (Luo et al., 2019; Hallinan et al., 2023a). Our CoTeX framework explores both parallel and non-parallel data landscapes. The advent of LLMs has introduced ICL for executing TST with few-shot prompts, bypassing the need for model parameter updates (Reif et al., 2022; Suzgun et al., 2022). Yet, these methods typically lack interpretability. In parallel, Saakyan and Muresan (2023) employ CoT prompting alongside domain expert feedback to enhance formality transfer and interpretability. Our CoTeX extends to broader range of TST directions, aiming to utilize CoT to provide rewriting explanations and minimize the requirement for human intervention.

6 Conclusion

We introduced CoTeX, a novel approach for TST. Through CoT prompting, we elicit the rationals for the style rewriting process from LLMs and then distill both the TST and reasoning capabilities into smaller task-specific models. CoTeX demonstrated its efficiency and effectiveness with and without utilizing parallel data, especially in low-resource scenarios. The CoT reasoning from CoTeX bolstered the explainability of TST models.

7 Limitations

TST Directions. We incorporate three style transfer directions to enable a clear comparison between target-blind and target-aware CoTeX. Benefiting from the powerful capacity of LLMs, we believe that our method could be extended to a broader array of TST directions (e.g., sentiment transfer). We plan to explore more transfer directions in future work.

Model Selection. We only use T5-large as the student model in the paper. We also conduct a concise study to apply CoTeX to the T5-XL model. As results shown in Appendix B, our CoTeX-TA still outperforms SFT on ParaDetox dataset.

Evaluation Metrics. Unlike previous studies (Krishna et al., 2020; Liu et al., 2022), we abstain from using other automatic metrics (e.g., BERTscore for meaning preservation) to evaluate our models. Our decision is grounded in two main reasons: (1) While these automatic evaluations consider three facets, i.e., preservation of semantic meaning, accuracy of style transfer, and

fluency they lack an effective methodology for aggregating these metrics to convey the overall performance (Ostheimer et al., 2023); (2) Our preliminary experiments involving these automatic metrics revealed a misalignment between their outcomes and the BLEU score derived from human-annotated references. We thus opt to report the BLEU score in the paper. Detailed results from our preliminary tests are presented in Appendix A.

8 Ethical Consideration

The primary objective of training CoTeX model is to achieve more computationally efficient and effective models for TST. We focus on the positive TST directions, such as language detoxification. We use an LLM to generate rationales alongside transferred text, which are subsequently distilled into smaller LMs. It's important to acknowledge that the LLM's generation might encompass societal biases (Lucy and Bamman, 2021) or hallucinations (Zhang et al., 2023), and student models trained with this data could inherit these characteristics of the teacher LLM. Additionally, our CoTeX-TA relies on datasets from prior research. Thus, any biases present in the original annotation processes of these datasets might also be reflected in our trained models. We expect the ongoing work (Ouyang et al., 2022; Dev et al., 2022) of improving LM's social fairness, faithfulness, and trustworthiness could benefit both teacher and student models.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *Natural Language Processing and Information Systems*, pages 47–61, Cham. Springer Nature Switzerland.

Kunal Chawla and Diyi Yang. 2020. Semi-supervised formality style transfer using language model discriminator and mutual information maximization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2340–2354, Online. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An opensource chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Online only, November 20-23, 2022*, pages 246–267. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3168–3180, Minneapolis, Minnesota. Association for Computational Linguistics.

Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. 2023a. STEER: unified style transfer with expert reinforcement. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 7546–7562. Association for Computational Linguistics.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2023b. Detoxifying text with MaRCo: Controllable revision with experts and anti-experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–242, Toronto, Canada. Association for Computational Linguistics.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8003–8017. Association for Computational Linguistics.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *CoRR*, abs/2210.11610.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 737–762. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. *CoRR*, abs/2210.06726.

Ao Liu, An Wang, and Naoaki Okazaki. 2022. Semi-supervised formality style transfer with consistency training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4689–4701. Association for Computational Linguistics.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8376–8383.

Ruibo Liu, Chongyang Gao, Chenyan Jia, Guangxuan Xu, and Soroush Vosoughi. 2021. Non-parallel text style transfer with self-parallel supervision. In *International Conference on Learning Representations*.

Varvara Logacheva, Daryna Dementieva, Irina Krotova, Alena Fenogenova, Irina Nikishina, Tatiana Shavrina, and Alexander Panchenko. 2022a. 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. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022b. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5116–5122. ijcai.org.

Sharmila Reddy Nangi, Niyati Chhaya, Sopan Khosla, Nikhil Kaushik, and Harshit Nyati. 2021. Counterfactuals to control latent disentangled text representations for style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 40–48, Online. Association for Computational Linguistics.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1008–1021. Association for Computational Linguistics.

Phil Ostheimer, Mayank Kumar Nagda, Marius Kloft, and Sophie Fellenz. 2023. A call for standardization and validation of text style transfer evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10791–10815, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1314–1324. Association for Computational Linguistics.

Dongqi Pu and Vera Demberg. 2023. Chatgpt vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1–18. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Machel Reid and Victor Zhong. 2021. LEWIS: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3932–3944, Online. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Arkadiy Saakyan and Smaranda Muresan. 2023. ICLEF: in-context learning with expert feedback for explainable style transfer. *CoRR*, abs/2309.08583.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4936–4945. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222, Abu Dhabi,

United Arab Emirates. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019a. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11034–11044.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023a. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019b. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3571–3576. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *CoRR*, abs/2304.14402.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization.

Fei Xiao, Liang Pang, Yanyan Lan, Yan Wang, Huawei Shen, and Xueqi Cheng. 2021. Transductive learning for unsupervised text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2510–2521. Association for Computational Linguistics.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *CoRR*, abs/1903.06353.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2899–2914. Indian Institute of Technology Bombay.

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*, pages 3801–3807.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3221–3228. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Appendices

A Preliminary Test on Evaluation Metrics

In our preliminary experiment, we evaluate model performance with several automatic metrics utilized by previous works (Krishna et al., 2020; Luo et al., 2019; Reif et al., 2022). These automatic metrics have been widely used in unsupervised TST due to their independence from human-labeled parallel data. However, we find that the outcomes from these metrics do not align with the reference-BLEU score derived from human-annotated references. These automatic metrics evaluate transferred text from three aspects:

1. Similarity: To evaluate the similarity between the source text and the transferred text, we employ BERTscore and self-BLEU. For BERTscore calculations, we use the SimCSE-large model (Gao et al., 2021) as the backbone.

2. Transfer Accuracy: To evaluate the efficacy of the style transfer, we employ a classifier (Babakov et al., 2023) to determine whether the transferred text successfully achieves the desired style.

3. Fluency: To access the fluency of the transferred text, we compute its perplexity using GPT. Additionally, we utilize a classifier trained on the Corpus of Linguistic Acceptability (CoLA) from (Krishna et al., 2020) to determine the grammaticality of the transferred text.

In this preliminary experiment, we conduct experiments using varying training sizes from the formality transfer (F&R) dataset. These experiments are carried out in a target-blind setting, where we finetune a T5-large model using the synthetic data generated from LLM. For assessing transfer accuracy, we employ a binary classifier introduced by Babakov et al. (2023), which is a RoBERTa-base model finetuned on the GYAFC's training set. This classifier achieves a test accuracy of 0.91. As Table 4 shows, the outcomes from these metrics did not correspond well with the reference-BLEU score. We thus opt to report the BLEU score in the paper.

B Experiment with T5-XL

We conduct a concise experiment to apply our Co-TeX to T5-XL (containing 3B parameters). As Table 5 shows, our CoTeX-TA outperforms SFT across all the data sizes.

# of data	Ref-BLEU	BERTScore	Self-BLEU	Tra. Acc.	PPL	CoLA
1000	72.54	0.96	45.34	0.94	61.15	0.95
2000	73.13	0.96	46.89	0.93	59.02	0.95
5000	71.92	0.96	50.71	0.90	65.54	0.94
10000	72.86	0.96	51.15	0.89	63.98	0.95
20000	72.90	0.96	49.89	0.90	64.66	0.94

Table 4: Preliminary result on GYAFC (F&R) for investigating evaluation metrics. **Tra. Acc.**: transfer accuracy, **PPL**: perplexity.

# Data	SFT	CoTex-TB	CoTex-TA
1000	49.15	46.64	**53.93**
2000	51.58	47.56	**54.26**
5000	52.91	47.92	**54.83**
10000	52.32	47.96	**55.13**
15000	52.88	48.47	**55.19**

Table 5: Finetuning T5-XL on detoxification dataset with our CoTeX or SFT.