Style Knowledge Graph: Augmenting Text Style Transfer with Knowledge Graphs

Martina Toshevska, Slobodan Kalajdziski and Sonja Gievska
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
{martina.toshevska, slobodan.kalajdziski, sonja.gievska}@finki.ukim.mk

Abstract

Text style transfer is the task of modifying the stylistic attributes of a given text while preserving its original meaning. This task has also gained interest with the advent of large language models. Although knowledge graph augmentation has been explored in various tasks, its potential for enhancing text style transfer has received limited attention. This paper proposes a method to create a Style Knowledge Graph (SKG) to facilitate and improve text style transfer. The SKG captures words, their attributes, and relations in a particular style, that serves as a knowledge resource to augment text style transfer. We conduct baseline experiments to evaluate the effectiveness of the SKG for augmenting text style transfer by incorporating relevant parts from the SKG in the prompt. The preliminary results demonstrate its potential for enhancing content preservation and style transfer strength in text style transfer tasks, while the results on fluency indicate promising outcomes with some room for improvement. We hope that the proposed SKG and the initial experiments will inspire further research in the field.

1 Introduction

Text style transfer (TST) is the task of modifying particular stylistic features of a text while preserving its original meaning. The task involves rewriting a text to match several stylistic attributes such as sentiment, formality, or politeness without changing the semantic meaning. With the emergence of large language models (LLMs), their application for TST gained attention primarily focused on prompting techniques (Reif et al., 2022; Suzgun et al., 2022) that reduce the need for extensive parallel datasets. Other approaches like fine-tuning (Mukherjee and Dušek, 2023), reinforcement learning (Deng et al., 2022), knowledge augmentation (Zong et al., 2024), and others (Lai et al., 2024; Pan et al., 2024) have also inspired recent research.

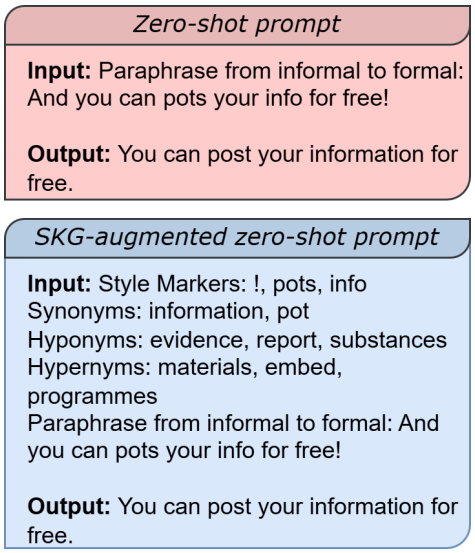

Figure 1: Examples for zero-shot and SKG-augmented zero-shot prompts for text style transfer that were used to evaluate and compare the proposed style knowledge graph.

Knowledge graphs (KGs) provide a structured representation of knowledge that enables efficient organization and retrieval across various domains. By integrating structured knowledge from KGs, LLMs can provide more accurate and contextually relevant outputs. We believe that combining both structured knowledge representation in KGs and the generative capabilities of LLMs has the potential to improve text style transfer tasks. While augmentation with KGs has been explored for many tasks, to the best of our knowledge, its application in text style transfer remains relatively understudied. Existing research is primarily focused on integrating knowledge base information to provide particular words for the desired style (Xu et al., 2022), similar sentences to the input to provide context (Toshevska and Gievska, 2024) or guidelines for the desired style (Zong et al., 2024).

Proceedings of the Generative AI and Knowledge Graph Workshop (GenAIK), pages 123–135
January 19, 2025.

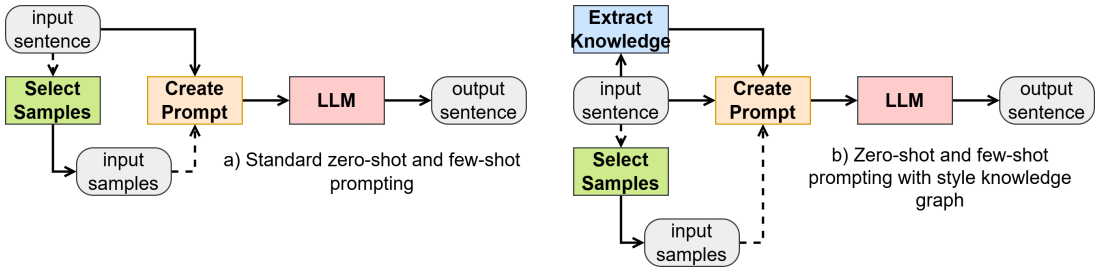

Figure 2: Overview of the prompting strategies. a) Standard prompting. b) Prompting augmented with SKG.

To combine the advantages of both approaches and facilitate further research in the field, we propose a Style Knowledge Graph (SKG) for text style transfer. The SKG is designed to capture words, their attributes, and relations for various styles with the aim of providing a source of knowledge that can enhance text style transfer. To evaluate the effectiveness of the proposed SKG we perform several prompting experiments where parts of the proposed SKG, that are relevant to the particular input sentence, are provided in the prompt. An example of the used prompts is shown in Table 1. We hope that the proposed SKG and the preliminary experiments will motivate further research.

The main contributions of the paper are: (1) We propose a knowledge graph for text style transfer, which we refer to as a Style Knowledge Graph (SKG). (2) We evaluate the effectiveness of augmenting text style transfer with SKG via prompting. (3) We analyze the influence of various parts of the SKG on the text style transfer task.

The rest of the paper is organized as follows. A brief introduction of previous text style transfer methods and knowledge augmentation is presented in Section 2. The definition and creation process of SKG is provided in Section 3. The preliminary experiments and baseline results are presented in Section 4 and Section 5, respectively. Section 6 concludes the paper.

2 Related Work

Before the advent of LLMs, TST methods commonly employed encoder-decoder architectures (Sutskever et al., 2014), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), and Reinforcement Learning (RL) (Williams, 1992). The methods based on encoder-decoder comprise an encoder to produce a style-neutral representation and a decoder to generate a sentence in the desired style, often augmented by additional components such as style classifiers (Lample et al., 2019; Xu et al., 2019; Cheng et al., 2020), and style embeddings (Li et al., 2018). GAN-based approaches use a generator to produce a sentence in the target style trained with adversarial objectives (Hu et al., 2017; Shen et al., 2017; Fu et al., 2018). RL-based approaches use a reward-based system to generate sentences in the desired style, by using multi-part rewards combining content preservation, style change, and fluency (Luo et al., 2019).

Prompting techniques are among the first approaches for text style transfer with LLMs, that explore zero-shot and few-shot techniques. Augmented zero-shot (Reif et al., 2022) explores a vanilla prompt that specifies the target style augmented with a single set of exemplars within the prompt to include a variety of sentence rewriting operations instead of exemplars specific to the target style. In addition to the vanilla prompt, Prompt&Rerank (Suzgun et al., 2022) explores a contrastive prompt to specify both the source and the target style that create a clear contrast between them, and two negation prompts to specify the target style as a negation of the source style and vice versa. Several approaches focus on editing the input sentence via prompting. PromptEdit, assesses TST as a text classification task with the goal of generating candidate sentences with an edit-based search algorithm that employs insertion, deletion, and replacement as edit operations, and then determining a style score for them with an LLM (Luo et al., 2023). PEGF utilizes two-way prompting that first identifies stylistic words as words with a score higher than a particular threshold via an initial prompt and then edits those stylistic words via implicit or explicit masking with a second prompt (Liu et al., 2024).

Continuing the research in the prompting directions, our proposed method introduces a style knowledge graph to augment text style transfer by including relevant parts of the graph in the prompt. Unlike previous research that relies primarily on

Dataset	Style 1 (s1)	Style 2 (s2)	Parallel?	# Samples	Task
Yelp	negative	positive	✗	428,632	sentiment transfer
Politeness	neutral	polite	✗	371,018	politeness transfer
GYAFC	informal	formal	✓	330,060	formality transfer
WNC	biased	neutral	✓	111,006	neutralizing subjective bias
Shakespeare	modern	Shakespearean	✓	42,150	personal style transfer
ParaDetox	toxic	neutral	✓	31,302	detoxification

Table 1: Statistics for the text style transfer datasets.

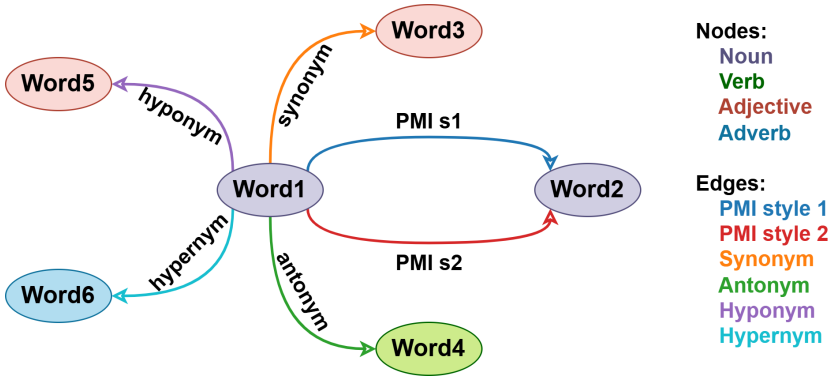

Figure 3: A visual representation of the style knowledge graph. The graph contains four types of nodes: nouns, verbs, adjectives, and adverbs; and six types of edges: PMI from style 1, PMI from style 2, synonyms, antonyms, hyponyms, and hypernyms.

generative capabilities or edit-based techniques, our approach aims to provide structured knowledge in the form of word suggestions to assist the LLM in the word choices for the output sentence. The combination of knowledge graphs and LLMs opens new directions for further research in the domain.

3 Style Knowledge Graph

3.1 Text Style Transfer Datasets

Text style transfer methods were evaluated using many parallel and non-parallel datasets. We selected a set of datasets that were mostly used for assessing text style transfer methods which we believe have the potential to foster further research in the field. The total number of datasets is six: Yelp[1], Politeness (Madaan et al., 2020), GYAFC (Rao and Tetreault, 2018), WNC (Pryzant et al., 2020), Shakespeare (Xu et al., 2012; Xu, 2014), and ParaDetox (Logacheva et al., 2022). Their statistics are summarized in Table 1.

[1]https://www.yelp.com/dataset, last visited: 05.09.2024

3.2 Style Knowledge Graph Creation

We create a heterogeneous graph for each text style transfer dataset which we refer to as **Style Knowledge Graph (SKG)**.[2] The graph contains four types of nodes: nouns, verbs, adjectives, and adverbs; and six types of edges: PMI from style 1, PMI from style 2, synonyms, antonyms, hyponyms, and hypernyms. A visual representation of the graph is shown in Figure 3.

3.2.1 Nodes

Nodes in the SKG are derived from the words that are present in the corresponding text style transfer dataset as follows. For each word in the sentences, its grammatical category is determined using the Part-of-Speech (PoS) tagger available in the NLTK Python library[3]. Based on the determined category, the words are grouped into four node types. Considering that a word may have a different category in a different sentence, the same word may be present as two different nodes. Each pair of words and categories that appear at least once is part of the *candidate node set*. For each node (word and

[2]The official GitHub repository for this paper is: https://github.com/mtoshevska/SKG

[3]https://www.nltk.org/, last visited: 05.09.2024

	Yelp	Politeness	GYAFC	WNC	Shakespeare	ParaDetox
Nouns	32,715	14,255	5,391	9,426	2,077	2,343
Verbs	10,453	5,622	3,789	4,435	1,336	864
Adjectives	15,646	5,020	3,025	6,839	732	807
Adverbs	2,306	891	889	1,467	287	174
# Nodes	61,120	25,788	13,094	22,167	4,432	4,188
PMI s1	136,107	95,284	27,734	400,521	3,966	3,614
PMI s2	573,226	241,565	18,540	367,898	3,213	1,417
Synonyms	14,202	10,763	5,731	18,443	2,448	585
Antonyms	523	668	392	1,576	116	13
Hyponyms	16,637	16,962	7,705	26,961	3,169	234
Hypernyms	14,042	16,129	6,538	25,840	2.834	219
# Edges	754,737	381,371	66,640	841,239	15,746	6,082

Table 2: Graph statistics (number of nodes and edges) for the six style knowledge graphs.

its grammatical category), we calculate the polarities (Li et al., 2018) in both styles using the Eq. 1:

$$p(w, s_i) = \frac{count(w, D_{s_i}) + \lambda}{count(w, D_{s_j}) + \lambda} \quad (1)$$

where $count(w, D_{s_i})$ is the number of times a word w appears in the set D_{s_i} of sentences with style s_i, and λ is the smoothing parameter. Then the absolute difference between both polarities is computed. The first 20% of the words with the highest polarity difference compose the *final set of nodes* for the graph.

3.2.2 Edges

The edges in the graph belong to two categories based on the creation technique: edges based on the information extracted from the corresponding text style transfer dataset and edges based on WordNet (Miller, 1995) semantic relations. Edges are created only between nodes in the final set.

To create edges based on the text style transfer datasets, we calculated point-wise mutual information (PMI) in a particular style s_i for a pair of nodes m and n (Yao et al., 2019) using the Eq 2:

$$PMI(m, n, s_i) = log \frac{p(m, n, s_i)}{p(m, s_i) \cdot p(n, s_i)} \quad (2)$$

$$p(m, n, s_i) = \frac{\#W(m, n, s_i)}{\#W_{s_i}} \quad (3)$$

$$p(m, s_i) = \frac{\#W(m, s_i)}{\#W_{s_i}} \quad (4)$$

where $\#W(m, n, s_i)$ is the number of sliding windows that contain both words m and n, $\#W(m, s_i)$ is the number of sliding windows that contain word m, and $\#W_{s_1}$ is total number of sliding windows in the corpus. A positive PMI value implies a high semantic correlation of words in the dataset and therefore we add an edge between a pair of nodes for which the PMI value in the corresponding style is greater than 0. Two sets of edges are created for the two styles.

We used the WordNet implementation in the NLTK Python library to extract semantic relations between words. For each word in the final set of nodes, we extracted four semantic relations: synonyms, antonyms, hyponyms, and hypernyms. The grammatical category of the word is also considered when extracting the semantic relations. Since our nodes set contains only a subset of the total words, an edge is added only if the two nodes are already part of the graph. The statistics for the six SKGs are summarized in Table 2.

4 Baseline Experiments

4.1 Text Style Transfer Tasks and Datasets

We performed prompting experiments to evaluate the effectiveness of augmenting text style transfer tasks with a style knowledge graph. For the preliminary results, we evaluate the approach on four text style transfer tasks using the parallel datasets and the created SKGs described in the previous section: *formality transfer* with the **GYAFC** dataset, *neutralizing subjective bias* with the **WNC** dataset, *personal style transfer* with the **Shakespeare** dataset, and *text detoxification* with **ParaDetox** dataset.

4.2 SKG-augmented Prompting

Two prompting strategies were explored for text style transfer across our selected tasks. An example

Model	Technique	rBLEU↑	sBLEU↓	Acc↑	PPL↓	GM_2↑	GM_3↑
Standard prompting							
$T5_{small}$	0-shot	12.7	52.4	49.4	**185.4**	25.0	21.6
$T5_{base}$	4-shot	10.9	**13.9**	27.8	262.8	17.4	16.6
FLAN-$T5_{small}$	1-shot	38.7	40.9	71.0	396.8	52.4	34.0
FLAN-$T5_{base}$	1-shot	36.6	47.2	73.9	229.7	52.0	34.8
Prompting augmented with SKG							
$T5_{small}$	1-$shot_{SKG}$	10.8	27.5	17.2	395.3	13.7	13.9
$T5_{base}$	4-$shot_{SKG}$	12.4	17.8	18.5	456.6	15.1	14.8
FLAN-$T5_{small}$	1-$shot_{SKG}$	46.8	23.6	73.7	461.9	58.7	36.4
FLAN-$T5_{base}$	1-$shot_{SKG}$	**48.9**	25.8	**88.0**	201.4	**65.6**	**40.9**

Table 3: Zero-shot and few-shot performance with standard prompting and prompting augmented with SKG for formality transfer on the GYAFC dataset. Only the best result per model is shown. rBLEU - reference-BLEU. sBLEU - self-BLEU. Acc - Accuracy. PPL - Perplexity. GM_2 - Geometric Mean (rBLEU and Acc). GM_3 - Geometric Mean (rBLEU, Acc, and PPL). The best value is **bold** and the second best is underlined.

of the two prompting strategies is shown in Figure 1. The first approach employs a simple prompt that specifies only the input and desired target style designed following the recommendations from related research that evaluate prompting techniques for text style transfer. It was used as a baseline for comparison. The second approach integrates style-relevant semantic information from a style knowledge graph to enrich the prompt with contextually relevant alternatives that guide the model toward generating outputs in line with the desired style attributes. Beginning with identifying the top three words in the input sentence with the highest target style polarity, a corresponding subgraph is extracted from the style knowledge graph for each of the three words. The semantic relations of the top three words (synonyms, antonyms, hyponyms, and hypernyms) are added to the prompt to enrich the input with contextual clues. For both prompting strategies, we experimented with zero-shot and few-shot settings. An overview of the two strategies is displayed in Figure 2.

4.3 Evaluation metrics

Evaluation has been performed across three dimensions to comply with the previous research in the field. The *semantic content preservation* was evaluated with the BLEU (Papineni et al., 2002) metric. The Prompt-and-Rerank (Suzgun et al., 2022) method proposed using self-BLEU (sBLEU) to measure the degree to which the model directly copies the input sentence and reference-BLEU (rBLEU) to measure the distance from the ground-truth references. We also report on these two metrics. *Style transfer strength* was calculated with the accuracy of a pre-trained DistilRoBERTa (Sajjad et al., 2020) model on the style detection task as a percentage of the generated sentences labeled with the target style by the model. To measure the *fluency*, the perplexity of the generated sentences with a pre-trained GPT-2 (Radford et al., 2019) model was computed. Several studies (Li et al., 2018; Luo et al., 2019, 2023) use the geometric mean of rBLEU and accuracy to compute a single *joint* metric. While we calculated the two-fold joint metric, we also included the inverse perplexity value to compute a three-fold joint metric that integrates the three evaluation dimensions. The inverse perplexity was computed using the Eq. 5:

$$PPL_{inv} = \frac{1}{1 + ln(PPL)} \tag{5}$$

4.4 Implementation Details

For both prompting strategies, we assess the performance of multiple LLMs that encompass different parameter sizes: T5 (Raffel et al., 2020) and FLAN-T5 (Chung et al., 2022). Following the previous studies, we experiment with zero-shot and few-shot settings. For the few-shot setting, we explored with 1, 2, 3, and 4 demonstrations. We have used PyTorch implementation of the models available in the HuggingFace Transformers library[4] and evaluation metrics available in the HuggingFace Evaluate library[5].

[4]https://huggingface.co/docs/transformers/en/index, last visited: 15.09.2024

[5]https://huggingface.co/docs/evaluate/en/index, last visited: 15.09.2024

5 Results and Discussion

In this section, we present the main results for the proposed approach of augmenting text style transfer with SKG which we hope to serve as a baseline for comparison of further research in the field. Due to space limitations, we present only the results for formality transfer. For the results on the other datasets and results with other LLMs for the formality transfer task on the GYAFC dataset, we encourage the reader to refer to the Appendix.

5.1 Main Results

Table 3 presents the evaluation results of standard prompting and prompting augmented with SKG. Both prompting strategies were evaluated on four models: $T5_{small}$, $T5_{base}$, FLAN-$T5_{small}$, and FLAN-$T5_{base}$. A total of five experiments were performed for each model and prompting strategy. For brevity, only the best-performing one in terms of geometric mean is shown.

The evaluation results suggest that SKG-augmented prompting improves content preservation for formality transfer, as demonstrated by higher rBLEU and lower sBLEU scores when compared with standard prompting. A possible reason may be the structure of the prompt that provides specific word choices. This approach includes specific word suggestions as part of the input prompt that help the model to choose particular words for the output sentence.

For style transfer strength, FLAN-T5 achieved higher accuracy with SKG-augmented prompting, while T5 achieved higher accuracy with standard prompting. FLAN-T5, which is an instruction-tuned LLM, may benefit more from structured prompts that offer more context for the word choices that align with the desired target style. The prompt design closely resembles the instruction setting that was used for training. Both geometric mean scores further confirm this hypothesis.

Although SKG-augmented prompting improves content preservation and in some cases improves style transfer strength, this approach fails to retain the fluency for three out of four models. As indicated by the higher perplexity, we observe a decrease in fluency in the SKG-augmented setting. A possible reason may be the word suggestions in the prompt that potentially increase the complexity and result in less fluent outputs. By adding particular word suggestions, sometimes the model tends to copy and include them in the output which has a potential negative impact of fluency. To address this limitation a future direction would be to experiment with LLMs with more parameters, as these models are typically more fluent.

	rBLEU	sBLEU	Acc	PPL
0-shot	18.4	72.2	22.6	4935.6
1-shot	**48.9**	**25.8**	**88.0**	**201.4**
2-shot	<u>44.9</u>	<u>29.7</u>	<u>74.8</u>	338.2
3-shot	42.1	32.3	69.3	<u>318.6</u>
4-shot	40.0	32.7	61.8	635.2

Table 4: Comparison of different number of demonstrations in the prompt for our best-performing approach for SKG-augmented prompting (FLAN-$T5_{base}$). rBLEU - reference-BLEU. sBLEU - self-BLEU. Acc - Accuracy. PPL - Perplexity. The best value is **bold** and the second best is <u>underlined</u>.

5.2 Effect of Number of Demonstrations in the Prompt

Our best-performing model for formality transfer is FLAN-$T5_{base}$ augmented with SKG in a one-shot setting. For further analysis, we use only this model. Next, we explore how the different number of demonstrations in the prompt affects the performance. Following the prior studies, we experimented with 0-4 demonstrations. The results are summarized in Table 4. The results suggest that the best performance is achieved with a single demonstration i.e. in a one-shot setting. One-shot prompting yielded the highest rBLEU and accuracy, and the lowest perplexity and sBLEU.

The zero-shot approach showed the worst performance across all metrics, with a significant decrease in fluency. We hypothesize that the absence of demonstrations negatively impacts the capability of generating fluent outputs in the desired target style. While there is an improvement with the switch from a zero-shot to a one-shot setting, further increasing the number of demonstrations also results in lower performance. In contrast to previous findings, in our approach, adding additional examples in the prompt may introduce complexity that reduces overall performance.

5.3 Comparison with Pre-LLM and LLM-based methods

Table 5 shows the performance of our best model against previous LLM and pre-LLM approaches. In comparison with pre-LLM unsupervised methods, SKG-augmented prompting showed competitive

Approach	rBLEU↑	sBLEU↓	Acc↑	PPL↓
Pre-LLM approaches				
CAAE (Shen et al., 2017)	17.9	-	75.3	-
DeleteOnly (Li et al., 2018)	29.2	-	18.8	-
DeleteAndRetrieve (Li et al., 2018)	21.2	-	55.2	-
MultiDecoder (Fu et al., 2018)	12.3	-	17.9	-
StyleEmbedding (Fu et al., 2018)	7.9	-	22.7	-
DualRL (Luo et al., 2019)	41.9	-	71.1	-
LLM approaches				
P&R (Suzgun et al., 2022)	36.4	49.6	*85.0*	68.0
PromptEdit (Luo et al., 2023)	37.7	50.2	81.0	87.0
PEGF (Liu et al., 2024)	38.2	46.4	**88.0**	**31.0**
SKGPrompt (Ours)	***48.9***	***25.8***	***88.0***	*201.4*

Table 5: Comparison of our best-performing approach for SKG-augmented prompting with previous pre-LLM and LLM-based approaches for formality transfer on the GYAFC dataset. rBLEU - reference-BLEU. sBLEU - self-BLEU. Acc - Accuracy. PPL - Perplexity. The best value is **bold** and the second best is underlined. The results for previous approaches were obtained either from the original papers that introduce the particular approach or, if an approach was not initially designed for formality transfer, from other studies that re-ran those approaches for comparison. Results for DeleteOnly, DeleteAndRetrieve, MultiDecoder, StyleEmbedding, and CAAE were obtained from (Luo et al., 2019). Results for P&R and PromptEdit were obtained from (Liu et al., 2024). For all other approaches the results were obtained from their original paper.

performances, surpassing them in content preservation and accuracy, despite not being trained or fine-tuned on the task.

When compared with previous LLM prompting-based approaches, SKG-augmented prompting achieves the overall best performance for content preservation, as suggested by its highest rBLEU score. Demonstrated by similar accuracy scores, we observe that our approach matches the PEGF (Liu et al., 2024) approach for style transfer strength. Both approaches share a similar idea of identifying stylistic words. PEGF identifies stylistic words via prompting and replaces them with a second prompt, while our approach utilizes style polarities to determine stylistic words and provides word suggestions based on semantic relations.

As indicated by the higher perplexity score, our approach demonstrates worse performance in fluency. Its outputs are less fluent compared to other LLM-based prompting methods. One possible reason could be the fact that these models utilize LLMs with more parameters that are considered to generate more fluent outputs.

5.4 Ablation Experiments

To analyze the contribution of different parts of the SKG to the text style transfer task, we perform ablation experiments. The experiments were performed for our best-performing model (FLAN-T5$_{\text{base}}$ in a one-shot setting): **no SR** - prompt without semantic relations, **all SR** - prompt with all semantic relations, **o/ Rel** - prompt with a single semantic relation *Rel*, and **w/o Rel** - prompt with all semantic relations except *Rel*. The results are summarized in Table 6

	rBLEU	sBLEU	Acc	PPL
no SR	48.3	26.6	86.1	306.6
o/ Syn	48.5	26.4	87.3	232.7
o/ Ant	48.3	26.5	86.4	349.4
o/ HypR	48.5	26.6	87.1	232.9
o/ HypO	48.5	26.5	87.9	285.7
w/o Syn	48.8	26.0	87.7	234.3
w/o Ant	48.8	26.0	87.7	228.0
w/o HypR	**48.9**	**25.8**	**88.1**	230.8
w/o HypO	**48.9**	25.9	87.9	243.6
all SR	**48.9**	**25.8**	87.9	**201.4**

Table 6: Results of the ablation experiments for our best-performing approach for SKG-augmented prompting for formality transfer on the GYAFC dataset. rBLEU - reference-BLEU. sBLEU - self-BLEU. Acc - Accuracy. PPL - Perplexity. The best value is **bold** and the second best is underlined.

The results of the ablation experiments demonstrate that semantic relations have a positive impact on performance. The best overall results are achieved when all semantic relations are used. Ex-

cluding specific semantic relations results in an increase in perplexity thus confirming that although the overall perplexity is relatively high, they have a positive impact.

Excluding hypernyms or hyponyms does not change the rBLEU score suggesting that these relations may not have a critical role in preserving the content. This is further confirmed by the increase in the accuracy score when hypernyms are excluded. On the contrary, excluding synonyms and antonyms negatively impacts the performance with an increase in perplexity and a slight decrease in accuracy and rBLEU.

6 Conclusion

In this paper, we proposed a Style Knowledge Graph for augmenting text style transfer using large language models. The SKG captures words, their attributes, and relations in a particular style to provide additional information for the task. We conducted preliminary experiments with prompting where the relevant part of the SKG was added as part of the prompt. The evaluation results demonstrated the potential of this method for enhancing content preservation and accuracy while highlighting areas for further improvement, particularly in fluency. We hope that this research will inspire further research in the field, extending beyond prompting to investigate new approaches and methodologies for text style transfer. SKGs have the potential to augment other text generation tasks beyond text style transfer, for example by guiding the model to generate more coherent summaries based on the selection of key parts of the input. We hope that future research will further explore this direction for SKGs for more context-aware and reliable text generation.

7 Limitations

Based on our experiments, we identified a few limitations. In some cases, parts of the instruction were returned as part of the output. This occurred more frequently with the T5 model. Since T5 is not an instruction fine-tuned model, challenges in distinguishing task instructions from the input content may be due to its lack of instruction-tuning. We observed lower performance for the SKG-augmented prompting when using smaller datasets. Since smaller datasets will lead to creating smaller SKGs we believe that this drop in performance is a direct result of the reduced richness and coverage of the SKG. A possible future direction to address this limitation may be to enrich the SKG with more information.

8 Ethical Considerations

As with other text generation tasks, our approach holds potential risks of misuse for malicious purposes, such as generating text that is negative, toxic, text that contains subjective bias, or text impersonating a specific author. Since we used existing text style transfer datasets to construct the SKG, any potential biases present in those datasets could be transferred and replicated in the SKG. Moreover, since LLMs are trained on datasets collected from the web, any biases present in the training data may be reflected in the outputs of our method. To address these risks it is crucial to raise awareness among researchers and users of such methods about the ethical implications and to promote responsible use for positive purposes.

References

Yu Cheng, Zhe Gan, Yizhe Zhang, Oussama Elachqar, Dianqi Li, and Jingjing Liu. 2020. Contextual text style transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2915–2924.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *CoRR*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 663–670.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering llms in text style transfer. *CoRR*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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, Volume 1 (Long Papers)*, pages 1865–1874.

Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15285–15295.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. 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.

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*. International Joint Conferences on Artificial Intelligence Organization.

Guoqing Luo, Yu Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5740–5750.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1869–1881. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Sourabrata Mukherjee and Ondřej Dušek. 2023. Leveraging low-resource parallel data for text style transfer. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 388–395.

Lei Pan, Yunshi Lan, Yang Li, and Weining Qian. 2024. Unsupervised text style transfer via llms and attention masking with multi-way interactions. *CoRR*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

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.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2020. On the effect of dropping layers of pre-trained transformer models. *CoRR*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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.

Martina Toshevska and Sonja Gievska. 2024. Large language models for text style transfer: Exploratory analysis of prompting and knowledge augmentation techniques. In *Intelligent Environments 2024: Combined Proceedings of Workshops and Demos & Videos Session*, pages 134–142. IOS Press.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

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

Wei Xu. 2014. *Data-driven approaches for paraphrasing across language variations*. Ph.D. thesis, New York University.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.

Wenda Xu, Michael Saxon, Misha Sra, and William Yang Wang. 2022. Self-supervised knowledge assimilation for expert-layman text style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11566–11574.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chang Zong, Yuyan Chen, Weiming Lu, Jian Shao, and Yueting Zhuang. 2024. Proswitch: Knowledge-guided language model fine-tuning to generate professional and non-professional styled text. *CoRR*.

A Appendix

A.1 Evaluation Results with Other LLMs

Apart from the main experiments with T5 and FLAN-T5 models, additional experiments with LLaMA and GPT models were performed on the formality transfer task with the GYAFC dataset. Only the zero-shot experiments were performed because of the time required for generating output sentences with these models. Their evaluation with few-shot prompting remains as future work.

LLaMa and GPT model variants showed unexpectedly low BLEU scores for content preservation. We believe that the low BLEU scores are due to the fact that the choice of words made by these models differs from the words in the ground truth sentence. BLEU is a metric based on n-gram overlap and it is expected to obtain lower scores when there is a different choice of words in the generated and the expected sentence. However, by manual inspection, we noticed that the generated output sentences managed to preserve the content of the input sentence to some extent, but used different words and often added additional explanations. To evaluate content preservation, BERTScore (Zhang et al., 2019), which compares the semantic meaning based on embedding vectors, was computed instead of BLEU. In Table 7, example outputs from these models are shown, and the full evaluation results are summarized in Table 8.

The results indicate that T5 and FLAN-T5 models obtain the best overall results for content preservation as measured by the higher BERTScore for both standard zero-shot and SKG-augmented zero-shot prompting. In terms of style transfer strength measured by accuracy, LLaMA-based models significantly outperform the other approaches for SKG-augmented prompting. For standard prompting, GPT-Neo showed the best performance thus indicating that GPT-based models can better leverage structured prompts to generate a sentence in the desired style. These models demonstrated significantly better fluency than the T5-based and LLaMA-based approaches. A possible reason could be the setting for evaluating fluency which relies on calculating perplexity with GPT-2. Considering that GPT-2 is a part of the same family of models, output sentences generated by GPT-based models may be more naturally aligned with the evaluation metric therefore leading to significantly higher fluency scores.

A.2 Evaluation Results for Other Datasets

In the tables below we present the evaluation results for the remaining three text style transfer tasks with parallel datasets: neutralizing subjective bias on the WNC dataset (Table 9), personal style transfer on the Shakespeare dataset (Table 10), and text detoxification on the ParaDetox dataset (Table 11). Both prompting strategies were evaluated on four models: $T5_{small}$, $T5_{base}$, FLAN-$T5_{small}$, and FLAN-$T5_{base}$. A total of five experiments were performed for each model and prompting strategy. For brevity, only the best-performing one in terms of geometric mean is shown.

The results do not indicate a clear superiority of the SKG-augmented approach when compared against standard contrastive prompting among all three tasks. The significantly smaller number of samples in the corresponding datasets may be a

contributing factor for the lower performance. Considering that smaller datasets will lead to creating smaller SKGs the lower performance could be a result of the reduced richness and coverage of the SKG. To address this limitation, a possible future direction would be to enrich the SKG with more information. Across all experiments, the sBLEU scores are consistently lower suggesting that the SKG-augmented approach may favor generating more diverse outputs rather than simply copying the input.

Input sentence: *You can get almost anything on ebay!*
Ground truth output sentence: *You can get almost anything on ebay!*
Predicted output sentence: *It's possible to obtain nearly any item through eBays online auction service.*
Input: *Where in the world do you come up with these questions????*
Ground truth output sentence: *How do you come up with these questions?*
Predicted output sentence: *I am curious as to where your ideas for this quiz originated.*
Input: *i think pretty much they're all sort of humorous*
Ground truth output sentence: *I think they are all very funny.*
Predicted output sentence: *I believe that most are somewhat funny.*

Table 7: Example output sentences for the formality transfer task on the GYAFC dataset. The outputs were generated with the LLaMa-2_{13B} with zero-shot prompting augmented with SKG.

Model	BERTScore↑	Acc↑	PPL↓	GM_2↑	GM_3↑
Standard prompting					
$T5_{small}$	**99.0**	49.4	185.4	66.6	41.5
$T5_{base}$	91.7	37.2	340.3	18.5	17.1
FLAN-$T5_{small}$	92.2	23.1	805.4	46.2	30.3
FLAN-$T5_{base}$	<u>93.5</u>	24.5	318.9	47.9	32.4
LLaMA-2_{7B}	88.6	52.2	390.7	68.0	40.5
LLaMA-2_{13B}	81.2	36.5	<u>64.5</u>	54.4	38.6
LLaMA-2-$chat_{7B}$	88.5	83.3	548.1	<u>85.8</u>	<u>46.6</u>
LLaMA-2-$chat_{13B}$	88.0	<u>84.4</u>	610.2	**86.1**	46.4
GPT-J_{6B}	81.6	31.4	95.7	50.6	35.9
GPT-$Neo_{1.3B}$	87.2	**84.7**	**46.6**	85.0	**53.4**
Prompting augmented with SKG					
$T5_{small}$	88.8	21.3	330.2	43.4	30.3
$T5_{base}$	89.1	37.2	476.3	18.2	16.7
FLAN-$T5_{small}$	<u>91.7</u>	18.7	2081.6	41.5	27.1
FLAN-$T5_{base}$	**92.4**	22.6	4935.6	45.7	28.0
LLaMA-2_{7B}	86.3	<u>88.8</u>	182.0	<u>87.6</u>	49.8
LLaMA-2_{13B}	86.0	**93.4**	142.7	**89.6**	<u>51.3</u>
LLaMA-2-$chat_{7B}$	87.1	80.3	568.7	83.7	45.7
LLaMA-2-$chat_{13B}$	87.3	81.5	629.3	84.4	45.7
GPT-J_{6B}	81.8	36.9	<u>102.2</u>	54.9	37.7
GPT-$Neo_{1.3B}$	87.0	79.5	**55.9**	83.1	**51.6**

Table 8: Zero-shot performance with standard prompting and prompting augmented with SKG for formality transfer on the GYAFC dataset with T5, FLAN-T5, LLaMA-2, and GPT. BERTScore - reference-BERTScore. Acc - Accuracy. PPL - Perplexity. GM_2 - Geometric Mean (BERTScore and Acc). GM_3 - Geometric Mean (BERTScore, Acc, and PPL). The best value is **bold** and the second best is <u>underlined</u>.

Model	Technique	rBLEU↑	sBLEU↓	Acc↑	PPL↓	GM_2↑	GM_3↑
Standard prompting							
$T5_{small}$	0-shot	56.9	62.7	60.4	225.2	58.6	37.7
$T5_{base}$	0-shot	38.7	42.4	54.6	312.6	46.0	31.5
FLAN-$T5_{small}$	3-shot	<u>64.8</u>	70.5	<u>67.1</u>	**167.2**	<u>66.0</u>	<u>41.4</u>
FLAN-$T5_{base}$	4-shot	**77.9**	84.7	**71.0**	<u>187.2</u>	**74.4**	**44.6**
Prompting augmented with SKG							
$T5_{small}$	0-$shot_{SKG}$	30.6	33.4	64.8	368.1	44.5	30.6
$T5_{base}$	0-$shot_{SKG}$	20.2	<u>22.1</u>	49.8	534.0	31.7	24.0
FLAN-$T5_{small}$	0-$shot_{SKG}$	46.7	50.7	54.5	1028.6	50.4	31.8
FLAN-$T5_{base}$	0-$shot_{SKG}$	19.6	**21.3**	55.4	170753.6	33.0	20.3

Table 9: Zero-shot and few-shot performance with standard prompting and prompting augmented with SKG for neutralizing subjective bias on the WNC dataset. Only the best result per model is shown. rBLEU - reference-BLEU. sBLEU - self-BLEU. Acc - Accuracy. PPL - Perplexity. GM_2 - Geometric Mean (rBLEU and Acc). GM_3 - Geometric Mean (rBLEU, Acc, and PPL). The best value is **bold** and the second best is <u>underlined</u>.

Model	Technique	rBLEU↑	sBLEU↓	Acc↑	PPL↓	GM_2↑	GM_3↑
Standard prompting							
$T5_{small}$	0-shot	10.0	46.6	60.2	**152.9**	24.5	21.5
$T5_{base}$	0-shot	11.3	52.5	82.8	1272.1	30.6	22.6
FLAN-$T5_{small}$	1-shot	<u>14.8</u>	75.0	84.5	457.9	35.3	<u>26.0</u>
FLAN-$T5_{base}$	2-shot	**16.0**	83.5	<u>91.3</u>	<u>204.6</u>	**38.2**	**28.5**
Prompting augmented with SKG							
$T5_{small}$	0-$shot_{SKG}$	5.8	**24.1**	37.7	336.6	14.8	14.8
$T5_{base}$	0-$shot_{SKG}$	7.9	<u>32.5</u>	71.1	649.3	23.8	19.6
FLAN-$T5_{small}$	0-$shot_{SKG}$	12.7	72.5	87.8	797.4	33.4	24.4
FLAN-$T5_{base}$	0-$shot_{SKG}$	14.7	78.0	**92.3**	1190.5	<u>36.8</u>	25.6

Table 10: Zero-shot and few-shot performance with standard prompting and prompting augmented with SKG for personal style transfer on the Shakespeare dataset. Only the best result per model is shown. rBLEU - reference-BLEU. sBLEU - self-BLEU. Acc - Accuracy. PPL - Perplexity. GM_2 - Geometric Mean (rBLEU and Acc). GM_3 - Geometric Mean (rBLEU, Acc, and PPL). The best value is **bold** and the second best is <u>underlined</u>.

Model	Technique	rBLEU↑	sBLEU↓	Acc↑	PPL↓	GM_2↑	GM_3↑
Standard prompting							
$T5_{small}$	0-shot	23.5	48.4	71.5	**451.7**	<u>41.0</u>	**28.7**
$T5_{base}$	0-shot	<u>26.6</u>	53.2	63.5	2222.5	**41.1**	<u>26.9</u>
FLAN-$T5_{small}$	1-shot	19.5	26.8	39.9	19440.1	27.9	19.3
FLAN-$T5_{base}$	0-shot	**29.8**	46.5	54.2	5204.3	40.2	25.7
Prompting augmented with SKG							
$T5_{small}$	2-$shot_{SKG}$	12.5	22.2	<u>75.9</u>	<u>891.5</u>	30.8	23.0
$T5_{base}$	2-$shot_{SKG}$	9.9	16.9	**78.0**	1342.7	27.8	21.1
FLAN-$T5_{small}$	2-$shot_{SKG}$	8.1	**10.5**	48.7	56816.4	19.8	14.9
FLAN-$T5_{base}$	2-$shot_{SKG}$	8.9	<u>11.1</u>	50.7	36454.2	21.2	15.8

Table 11: Zero-shot and few-shot performance with standard prompting and prompting augmented with SKG for text detoxification on the ParaDetox dataset. Only the best result per model is shown. rBLEU - reference-BLEU. sBLEU - self-BLEU. Acc - Accuracy. PPL - Perplexity. GM_2 - Geometric Mean (rBLEU and Acc). GM_3 - Geometric Mean (rBLEU, Acc, and PPL). The best value is **bold** and the second best is <u>underlined</u>.