A Recipe For Arbitrary Text Style Transfer with Large Language Models

Anonymous ACL submission

Abstract

In this paper, we leverage large language models (LMs) to perform zero-shot text style transfer. We present a prompting method that we call *augmented zero-shot learning*, which frames style transfer as a sentence rewriting task and requires only a natural language instruction, without model fine-tuning or exemplars in the target style. Augmented zero-shot learning is simple and demonstrates promising results not just on standard style transfer tasks such as sentiment, but also on arbitrary transformations such as "make this melodramatic" or "insert a metaphor."

1 Introduction

Text style transfer is the task of rewriting text to incorporate additional or alternative stylistic elements while preserving the overall semantics and structure. Although style transfer has garnered increased interest due to the success of deep learning, these approaches usually require a substantial amount of labeled training examples, either as parallel text data (Zhu et al., 2010; Rao and Tetreault, 2018) or non-parallel text data of a single style. (Li et al., 2018; Jin et al., 2019; Liu et al., 2020; Krishna et al., 2020). Even bleeding-edge approaches that tackle the challenging problem of label-free style transfer are limited in that they require at least several exemplar sentences that dictate a given target style (Xu et al., 2020; Riley et al., 2021). Hence, recent survey papers have identified a need for new methods that both reduce the training data requirements and expand the scope of styles supported (Jin et al., 2020; Hu et al., 2020).

In this work, we present *augmented zero-shot learning*, a prompting method that allows large language models to perform text style transfer to arbitrary styles, without any exemplars in the target style. Our method builds on prior work showing that sufficiently large LMs such as GPT-3 can perform various tasks ranging from classification to translation, simply by choosing a clever prompt to prepend to the input text for which the model is asked to continue (Brown et al., 2020; Branwen, 2020). Using a single prompt that provides several demonstrations of sentences being "rewritten" to meet a desired condition, language models can extrapolate and rewrite text in unseen styles. We are thus able to perform style transfer to arbitrary styles such as "*make this sentence more comic*" or "*include the word balloon*."

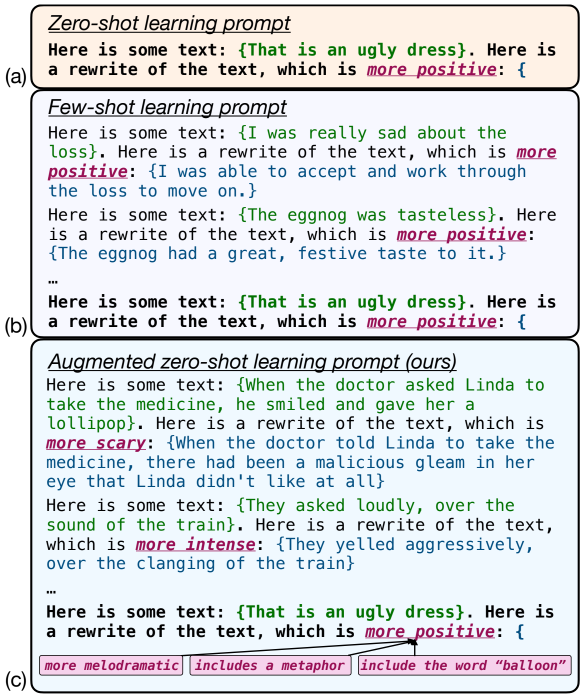

Figure 1: Zero-shot, few-shot, and augmented zero-shot prompts for style transfer. The boldface text is the zero-shot prompt, and the plain text is the additional priming sequence. The full prompts used in this paper are shown in Table 7. We encourage readers to examine the outputs of our model at https://bit.ly/3fLDuci.

Augmented zero-shot learning is simple and facilitates the application of style transfer to a wider range of styles than existing work. Our contributions are the following.

1. We propose a recipe for style transfer using large LMs that is label-free, training-free, and intuitively controllable.
2. Via human evaluation, we find that our method achieves strong performance on both standard and non-standard style transfer tasks. We also compare our approach for sentiment transfer with prior methods using automatic evaluation.
3. We explore real-world desired style transfers generated from users of a text editing UI that implements our method.

2 Augmented zero-shot prompting

Although large LMs are trained only for continuation, recent work has shown that they can perform a variety of NLP tasks by expressing the task as a prompt that encourages the model to output the desired answer as the continuation (Puri and Catanzaro, 2019; Weller et al., 2020; Brown et al., 2020; Schick and Schütze, 2021, *inter alia*; see Liu et al. (2021a) for a survey). The simplest approach, **zero-shot prompting**, directly uses natural language to ask the large LM to perform a task, as shown in Figure 1a. Zero-shot learning, however, can be prone to failure modes such as not returning well-formatted or logical outputs (see Appendix §A). **Few-shot prompting**, as shown in Figure 1b, has been shown to achieve higher performance, but requires exemplars for the exact task that we want the model to perform. Such few-shot examples can be easily obtained if the desired style transformation is known ahead of time, but this ultimately limits style transfer to a set of pre-specified styles.

To remove the need for these labeled exemplars in few-shot prompting, we propose *augmented zero-shot learning*. Instead of using exemplars of the exact style transfer task we wish to perform, we prompt the model via related style transfer tasks in the same format, as shown in Figure 1c. This intuition is inspired by Reynolds and McDonell (2021)'s observation that successful prompts constrain the behavior of the large LM away from failure modes—in our case, we aim to preserve the flexibility of a zero shot prompt while encouraging the model to produce outputs of a specific template. We keep the exemplars constant and insert any desired transformation in the final continuation, e.g., "*more melodramatic,*" "*insert a metaphor,*" or "*include the word balloon.*" This augmented zero-shot formulation works for a range of arbitrary styles.

Original	`There, in the middle of the street, stood an old man in a weatherbeaten brown coat.`
More descriptive	`The man's brown trenchcoat was coated with snow, and as he leaned against the telephone booth, he cast a long shadow across the road in the late afternoon sun.`
Include a metaphor	`There, in the middle of the street, stood an old man whose face resembled the weatherworn bark of a gnarled tree.`
Include the word "*park*"	`There, in the middle of Central Park, stood an old man in a weatherbeaten brown coat.`
Include the word "*balloon*"	`There, in the middle of the street, stood an old man with several colourful balloons tied to the straps of his coat.`
More comic	`There, in the middle of the street, stood an old man in a raggedy brown coat that was much too big for him.`
More melodramatic	`There, in the middle of the road, stood a grizzled old man, the light of life faded from his sunken eyes.`

Table 1: Example style transfer outputs from augmented zero-shot learning for non-standard styles.

3 Experimental Setup

Style transfer tasks. We consider six style transfer tasks that we deem non-standard, listed in Table 1. These styles were chosen to be representative of most frequent style adjustments made by users of an AI-assisted text editor that employs our method (discussed further in §5). As source sentences, we use 50 sentences randomly drawn from the Reddit Writing Prompts validation set (Fan et al., 2018), excluding those that already clearly exhibited one of the styles or were ungrammatical/incoherent. We use human evaluation for these styles, since not all styles have readily available classifiers.

We also evaluate our method on two standard style transfer tasks: sentiment and formality. We use the Yelp polarity dataset (Zhang et al., 2015) for sentiment and Grammarly's Yahoo Answers Formality Corpus (GYAFC) dataset for formality (Rao and Tetreault, 2018).[1] These datasets allow us to evaluate performance of augmented zero-shot learning in the context of prior supervised methods which have been used on these tasks.

Model. Augmented zero-shot learning requires a large language model. We use two dense left-to-right decoder-only transformer language models (Vaswani et al., 2017), each with a non-embedding parameter count of 137B. The first model, which we refer to as *LLM*, was trained on a corpus comprising public web documents, including forum

[1] Hosted by Luo et al. (2019).

and dialog data and Wikipedia. The dataset was tokenized into 2.81T BPE tokens with a SentencePiece vocabulary size of 32K (Kudo and Richardson, 2018). The second model, which we refer to as *LLM-Dialog*, was the result of finetuning *LLM* on a curated, high-quality subset of data identified to be in a conversational format. All generation was done with top-k=40 as the decoding strategy.

To show that the success of augmented zero-shot learning is not restricted to these two large LMs, we also perform an experiment using GPT-3 models (see Table 8). For *LLM* and GPT-3, we use the prompts shown in Figure 1 (see 7a for the unabbreviated prompts). For *LLM-Dialog*, the prompt is formulated as a conversation between one agent who is requesting rewrites and another who is performing the rewrites (see Table 7b in the appendix.)

4 Results

4.1 Non-Standard Styles

For our six non-standard styles, we asked six professional raters (see Appendix § E) to assess <input sentence, target style, output sentence> tuples. For each style, we compare outputs from our method plus the three baselines for 50 sentences.

Each tuple was scored by three raters (3,600 ratings total) on the following three axes which are standard to textual style transfer (Mir et al., 2019): **(1) transfer strength** (the amount that the output actually matches the target style), **(2) semantic presentation** (whether the underlying meaning of the output text, aside from style, matches that of the input), and **(3) fluency** (whether the text is coherent and could have been written by a proficient English speaker). Following Sakaguchi and Van Durme (2018), transfer strength and semantic preservation were rated on a scale from 1–100. A screenshot of the evaluation UI is shown in Figure 5 in the Appendix. We use *dialog-LLM*, and compare it with three other methods: **(1) zero-shot** (a baseline), **(2) paraphrase** (our normal augmented zero shot prompt, but with the target style of *"paraphrased"*, as a control) and **(3) human** (ground-truth transformations written by the authors).

Figure 2 shows these results. We found that the outputs of our method were rated almost as highly as the human-written ground truth for all three evaluations. The zero-shot baseline performed the worst in all categories: 25.4% of the time, it did not return a valid response at all (see Appendix §A), compared with 0.6% for augmented zero shot. The

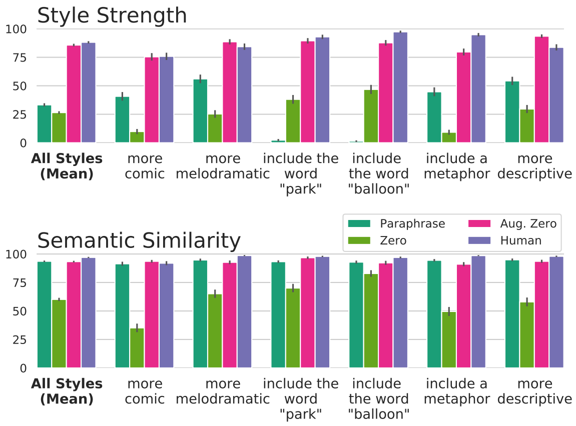

Figure 2: Human evaluation of style transfer for six atypical styles. Our method is rated comparably to the human-written ground truth. Error bars show Standard Error of the Mean. Evaluation of fluency is shown in Figure 4 in the Appendix.

strong performance of the paraphrase baseline at fluency and semantic similarity shows that large LMs are capable of generating high quality text that remains true to the input sentence's meaning.

For a subset of the tasks, some automatic evaluation was also possible. We found that the "*balloon*" and "*park*" transformations successfully inserted the target word 85% of the time. For "*more descriptive*" and "*include a metaphor*" the transformed text was, as expected, longer than the original (by 252% and 146% respectively, compared with 165% and 146% for human baselines).

4.2 Standard Styles

To better contextualize the performance of our method with prior methods, we also generated outputs for two standard style transfer tasks: sentiment and formality. Figure 3 shows human evaluations (same setup as before) for our outputs as well as the outputs from two popular prior style transfer methods, Unsup MT (Prabhumoye et al., 2018) and Dual RL (Luo et al., 2019). The outputs from our method were rated comparably to both human generated responses and the two prior methods.

Furthermore, following Li et al. (2018) and Sudhakar et al. (2019), we perform automatic evaluation for sentiment style transfer since there are classifiers available for these styles. We note that although automatic evaluations can diverge from human ratings, they can still be a good proxy as we could not perform human evaluation against every prior method due to time and resource constraints. We automatically evaluate **(1) transfer strength** using a sentiment classifier from Hug-

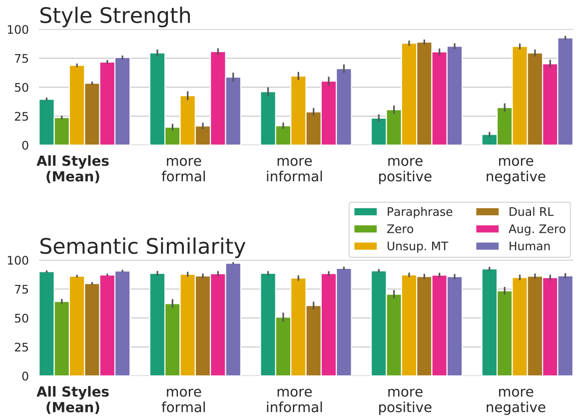

Figure 3: Human evaluation of sentiment and formality transfer. Our method is rated comparably to human-written ground truth as well as prior methods. Error bars show Standard Error of the Mean. Unsup. MT is Prabhumoye et al. (2018); Dual RL is Luo et al. (2019).

gingFace Transformers (Wolf et al., 2020), **(2) semantic similarity** to human examples provided by Luo et al. (2019) via BLEU score, and **(3) fluency** via perplexity, as measured by GPT-2 (117M).

Table 2 shows these automatic evaluations, with four main takeaways. First, augmented zero-shot prompting achieves high accuracy and low perplexity compared with baselines. The BLEU scores, however, are low, which we believe is because it tends to add additional information to generated sentences (see Appendix C for a deeper analysis). Second, we apply augmented zero-shot learning to GPT-3 175B; these results indicate that augmented zero-shot learning generalizes to another large language model. Third, we vary model size for GPT-3 models, finding that larger size greatly improves style transfer. Fourth, for *LLM* and *LLM-dialog*, we find that augmented zero-shot learning substantially outperforms vanilla zero-shot learning and almost reaches the accuracy of five-shot learning.

5 Potential of Arbitrary Styles

One promising application of augmented zero-shot learning is an AI-powered writing assistant that can allow writers to transform their text in arbitrary ways that the writer defines and controls. As a qualitative case study to explore what arbitrary re-write styles may be requested, we built an AI-assisted story-writing editor with a "rewrite as" feature that uses our augmented few-shot method. Our editor has a freeform text box for users to specify how they would like a selection of their story to be rewritten (see Figure 6 in the Appendix). We asked

	Acc	BLEU	PPL
SUPERVISED METHODS			
Cross-alignment (Shen et al., 2017)	73.4	17.6	812
Backtrans (Prabhumoye et al., 2018)	90.5	5.1	424
Multidecoder (Fu et al., 2018)	50.3	27.7	1,703
Delete-only (Li et al., 2018)	81.4	28.6	606
Delete-retrieve (Li et al., 2018)	86.2	31.1	948
Unpaired RL (Xu et al., 2018)	52.2	37.2	2,750
Dual RL (Luo et al., 2019)	85.9	55.1	982
Style transformer (Dai et al., 2019)	82.1	55.2	935
INFERENCE-ONLY METHODS			
GPT-3 ada, aug zero-shot	31.5	39.0	283
GPT-3 curie, aug zero-shot	53.0	48.3	207
GPT-3 da vinci, aug zero-shot	74.1	43.8	231
LLM: zero-shot	69.7	28.6	397
five-shot	83.2	19.8	240
aug zero-shot	79.6	16.1	173
LLM-dialog: zero-shot	59.1	17.6	138
five-shot	94.3	13.6	126
aug zero-shot	90.6	10.4	79

Table 2: Comparing augmented zero-shot prompting with supervised style transfer methods on the Yelp sentiment style transfer dataset using automatic evaluation. Acc: accuracy; PPL: perplexity. The inference-only table shows our method applied to 3 different sizes of GPT-3, plus our own LLM.

to be a little less angsty • to be about mining • to be better written • to be less diabolical • to be more absurd • to be more adventurous • to be more Dickensian • to be more emotional • to be more magical • to be more melodramatic • to be more philosophical • to be more revolutionary • to be more surprising • to be more suspenseful • to be more technical • to be more whimsical • to be warmer • to fit better grammatically with the rest of the story • to make more sense

Table 3: Requests in the form of "*Rewrite this...*" made by real users to a large LM-powered text editor. For the full set of unique requests, see Table 5 in the Appendix.

30 people from a creative writing group to use our our UI to write a 100-300 word story, collecting 333 rewrite requests in total. Table 3 shows a subset of these, which were as diverse as asking for the text "*to be about mining*" or "*to be less diabolical*."

6 Conclusions

We introduced augmented zero-shot learning, which we find shows shows strikingly promising performance considering its simplicity. This prompting paradigm moves the needle in text style transfer by expanding the range of possible styles beyond the currently limited set of styles for which annotated data exists. More broadly, we also hope that the strategy of prompting a large LM with non-task specific examples can inspire new inference-only methods for other NLP tasks.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Gwern Branwen. 2020. GPT-3 creative fiction.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhiqiang Hu, Roy Ka-Wei Lee, and Charu C. Aggarwal. 2020. Text style transfer: A review and experiment evaluation. *CoRR*, abs/2010.12742.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised text attribute transfer via iterative matching and translation. 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)*, pages 3097–3109, Hong Kong, China. Association for Computational Linguistics.

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)*, pages 737–762, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

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, New Orleans, Louisiana. 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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021b. A transformer-based framework for neutralizing and reversing the political polarity of news articles. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

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.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, 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*, pages 1869–1881, Online. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. *CoRR*, abs/1904.02295.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models. *arXiv preprint arXiv:1912.10165*.

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.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David C. Uthus, and Zarana Parekh. 2021. Textsettr: Label-free text style extraction and tunable targeted restyling. *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

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*, volume 30. Curran Associates, Inc.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "Transforming" delete, retrieve, generate approach for controlled text 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)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia. Association for Computational Linguistics.

Peng Xu, Yanshuai Cao, and Jackie Chi Kit Cheung. 2020. On variational learning of controllable representations for text without supervision. *Proceedings of the International Conference on Machine Learning (ICML)*, abs/1905.11975.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Proceedings of the Conference on Neural Information Processing Systems*.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

Appendix

A Limitations and Failure Modes

This section details several qualitative limitations with our method.

Unparsable answers A frequent problem that arises when using large LMs for other NLP tasks is their outputs cannot be automatically parsed into usable answers. For example, when given a prompt like `"Here is some text: that is an ugly dress. Here is a rewrite of the text, which is more positive"` *LLM-Dialog* might return something like `"Sounds like you are a great writer!"` Similar error modes exist for *LLM*, which might output something like `"Here are more writing tips and tricks."` Other times, the response contains correct information, but it cannot be automatically parsed (e.g., `"a good rewrite might be to say that the dress is pretty."`) In hindsight, these outputs make a lot of sense: most of the training data of large LMs is not well-formatted pairs of inputs and outputs (Reynolds and McDonell, 2021). See §B for how we dealt with these issues.

Hallucinations Large LMs are known to hallucinate text content; we saw this happen frequently for style transfer. While this is an advantage in some contexts like creative writing, it is undesirable for applications like summarization.

Inherent style trends We also noticed that even our *"paraphrase"* baseline, where the model was simply asked to rewrite the input sentence, was rated highly for style strength for a few styles, including *"more formal"* and *"more melodramatic"*. This implies that our method's generations generally trend toward these styles. A direction for future work would be to see what styles and qualities of text our method (and large LMs in general) are inherently more likely to produce.

Less reliable than trained methods For style transfer tasks that have available training data, prior methods that either train or finetune on that data are going to be inherently more reliable at producing text that looks like their training data. This can be observed in the lower BLEU scores our method achieves than trained methods, despite comparable transfer accuracy (Section C). Thus, augmented zero-shot learning offers less fine-grained controllability in the properties of the style-transferred text than methods which see task-specific training data.

Large LM safety concerns Large LMs themselves come with their own host of difficulties, barriers to entry, and potential safety concerns as discussed by Bender et al. (2021), which are also valid for this style transfer method. However, we also think that this method can be a useful tool in exploring and exposing the safety and boundaries of these models themselves: what happens if we try to force the large LM to make a text "more racist", "more sexist", or "more incendiary"? It is important to keep pushing these models to their boundaries to see where they fail and where problems arise, and specific use cases that show a broader range of the model's capabilities also show a broader range of its failure modes.

B Prompt Selection

A promising new area of prompt engineering has arisen to address the failure modes discussed above, specifically the invalid or unparseable answers. Reynolds and McDonell (2021) find that prompting a model for a task is more akin to locating an already-learned task than truly learning a new one. Moreover, they emphasize that that prompt engineering is mostly about avoiding various failure cases such as those described above. In this work, we use delimiters ("{" and "}") to help avoid these types of errors, giving scores of zero when there was no valid responses with such delimiters. There are other delimiters that could be used (e.g., quotes, "(" and ")", "<" and ">", newlines with a colon (as used by GPT-3), etc. We chose curly braces as they were 1) likely to occur in the training data as delimiters in other contexts and 2) not frequently part of the input sentence itself. We also use a second person prompt template for the dialog, which yielded better results as it was more similar to the training data. Exploring these options more quantitatively would be an interesting direction for future work. Because the performance of prompting can vary depending on the exact language of the prompt (Reynolds and McDonell, 2021), we compare four variations of prompts for sentiment: "*more positive/negative*," "*happier/sadder*," "*more optimistic/pessimistic*," and "*more cheerful/miserable*." As shown in Table 4 in the Appendix, performance differed across the four prompts, but we found them comparable.

Model / prompt wording	Acc	Bleu	PPL
LLM			
"more positive/negative"	76.3	14.8	180
"happier/sadder"	62.6	15.5	173
"more optimistic/pessimistic"	69.7	14.1	143
"more cheerful/miserable"	74.5	15.7	186
LLM-Dialog			
"more positive/negative"	90.5	10.4	79
"happier/sadder"	85.9	9.6	90
"more optimistic/pessimistic"	85.8	10.2	79
"more cheerful/miserable"	88.8	11.4	93

Table 4: Comparing variations of augmented zero-shot learning prompt wording for sentiment style transfer.

into paragraphs • to be a bit clearer • to be a little less angsty • to be a word for a song • to be about mining • to be about vegetables • to be better written • to be less descriptive • to be less diabolical • to be more absurd • to be more adventurous • to be more angry • to be more cheerful • to be more descriptive • to be more Dickensian • to be more emotional • to be more fancy • to be more flowery • to be more interesting • to be more joyful • to be more magical • to be more melodramatic • to be more philosophical • to be more revolutionary • to be more scary • to be more subtle • to be more surprising • to be more suspenseful • to be more technical • to be more violent • to be more whimsical • to be warmer • to fit better grammatically with the rest of the story • to make more sense • to use a more interesting word • with a few words

Table 5: Full results for requests in the form of "*Rewrite this...*" made by users to a large LM-powered text editor.

C Low BLEU for *LLM* Outputs

As we saw in 2, the outputs of our model had low BLEU scores with respect to human generated outputs, while simultaneously having high semantic similarity in human evaluations. Based on qualitative examination of outputs, we believe that this is because model outputs often, despite having high semantic similarity with the source sentence, used different language from human annotations. For instance, for transferring the sentiment of "*ever since joes has changed hands it's just gotten worse and worse*" to positive sentiment, our augmented zero-shot learning model outputted "the establishment has continued to provide excellent service, improving steadily since its change of ownership." This will have low BLEU with the ground truth with respect to human references, which is simply "*ever since joes has changed hands it's just gotten better and better.*"

Though we do not see this as an inherent problem, increasing the BLEU for the purposes of comparison can be done in an easy way via candidate selection, as our model returns sixteen possible continuations. In applications for which we prefer model outputs to have high lexical similarity to the source sentence, we could select the candidate of the sixteen with the highest BLEU score compared with the original source sentence. We find that this candidate selection step can substantially improve the BLEU score with the ground truth target sentences, as we show in Table 8.

D Further Related Work

Style transfer has gained increasing attention in the NLP landscape, for which neural models have been trained to perform style transfer for styles including sentiment, formality, politeness, gender, and political slant (Prabhumoye et al., 2018; Madaan et al., 2020; Liu et al., 2021b). We will briefly summarize the primary approaches to style transfer here, and refer the involved reader to either (Jin et al., 2020) or (Hu et al., 2020) for a survey.

Most text style transfer approaches fall in two categories. Early approaches tend to require *parallel* text data (Zhu et al., 2010; Rao and Tetreault, 2018), where every input in the source style has a corresponding output in the target style. Though this formulation elegantly fits the standard encoder–decoder paradigm, the availability of a parallel text corpus is a stringent requirement. Hence, recent text style transfer approaches have instead used *non-parallel* monostyle data (no one-to-one-mapping between instances in the source and target styles). Such methods include latent representation manipulation (Liu et al., 2020), prototype-based text editing (Li et al., 2018), and pseudo-parallel corpus construction (Jin et al., 2019). However, even non-parallel monostyle data can be hard to collect for arbitrary styles. As such, surveys have called for more research on approaches that expand the scope of supported styles and reduce the training data requirements for style transfer systems (Jin et al., 2020; Hu et al., 2020).

Several new methods tackle the challenging problem of *label-free* style transfer, which does not require a full corpus of labeled data, but rather just a few exemplars that define a style. Xu et al. (2020) use variational autoencoders for unsupervised learning of controllable representations for text. Riley et al. (2021) extract a style vector from a set of target texts and use this vector to condition the decoder to perform style transfer to a target style. These approaches have a similar goal to ours

Input	Model suggestion
You honor me," said Bruno. As he spoke, he began to laugh. His laugh was merry sounding and melodic. He took a deep breath. When he spoke again, his voice was low and strong. "I take this great task on and with my brothers shall lead us on to victory! **Rewrite it to be about mining.**	You honor me, said Bruno. As he spoke, he began to laugh. His laugh was deep and gravely, as if from the bowels of the earth. He took a deep breath. When he spoke again, his voice was deep and dark.'I take this great task on and with my brothers shall delve into the mines!"
I reckon,' smiled Toe, 'you'd have to say the same for icicles, wouldn't you? **Rewrite it to be more interesting.**	'I reckon,' smiled Toe, 'you'd have to say the same for icicles, wouldn't you?' These words were barely out of his mouth before the others realized exactly what they meant.
with its southern perfume **Rewrite it to be more descriptive**	A mixture of sweet honeysuckles and jasmine with a hint of tobacco and warm mesquite smoke. It was his southern perfume, with a hint of a memory of simpler times.
But Sally remained unmoved–and lonely. **Rewrite it to be more emotional**	Her tears rolled and rolled down her cheek, mixing with the saltwater around her, until the water was a rich mixture of the two. She wanted her poem to be cheerful like it used to be, but couldn't find the right words.

Table 6: Examples of users' arbitrary style transfer requests for which the model suggestion was accepted.

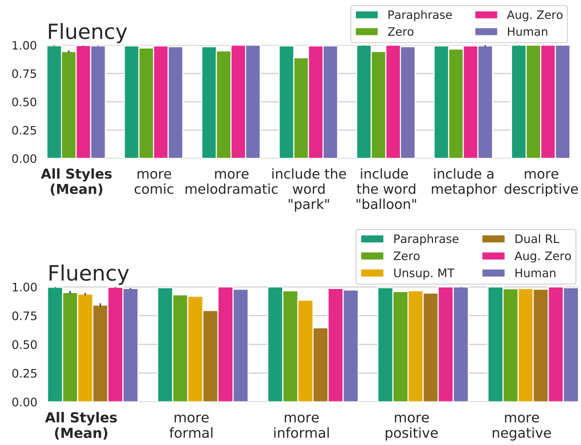

Figure 4: Human evaluation of fluency for style transfer for six atypical styles. Error bars show standard error of the mean.

in terms of expanding the scope of possible style transfers. However, they are different in two main ways. First, they require a fully specialized model, where our method can be applied out-of-the-box with something like GPT-3. This can either be a strength or weakness, depending on the availability of such a model. Second, they require exemplars to define a style rather than a plain text description.

E Rater education

These we used raters live in India and are fluent in English, and work full time labeling and evaluating data. To decrease inter-rater discrepancy and ensure that our instructions were clear, we had an initial calibration session where they test-rated a small portion of the data (around 10 datapoints which were then omitted from the results) and asked us any clarifying questions.

Augmented Zero-shot Prompt: LLM

```
"Here is some text: {When the doctor asked Linda to take the medicine, he smiled and gave her a lollipop.}.
Here is a rewrite of the text, which is more scary.
{When the doctor told Linda to take the medicine, there had been a malicious gleam in her eye that Linda didn't
like at all.}
Here is some text: {they asked loudly, over the sound of the train.}. Here is a rewrite of the text, which is
more intense.
{they yelled aggressively, over the clanging of the train}
Here is some text: {When Mohammed left the theatre, it was already dark out}.
Here is a rewrite of the text, which is about the movie itself. {The movie was longer than Mohammed had expected,
and despite the excellent ratings he was a bit disappointed when he left the theatre.}
Here is some text: {next to the path}. Here is a rewrite of the text, which is about France.
{next to la Seine}
Here is some text: {The man stood outside the grocery store, ringing the bell.}. Here is a rewrite of the text,
which is about clowns.
{The man stood outside the circus, holding a bunch of balloons.}
Here is some text: {the bell ringing}. Here is a rewrite of the text, which is more flowery.
{the peales of the jangling bell}
Here is some text: {against the tree}. Here is a rewrite of the text, which is includes the word 'snow'.
{against the snow-covered bark of the tree}
Here is some text: {That is an ugly dress}. Here is a rewrite of the text, which is more positive."
```

Augmented Zero-shot Prompt: LLM-dialog

```
"Here is some text: {When the doctor asked Linda to take the medicine, he smiled and gave her a lollipop.}.
Rewrite it to be more scary.",
"{When the doctor told Linda to take the medicine, there had been a malicious gleam in her eye that Linda didn't
like at all.}",
"Here is some text: {they asked loudly, over the sound of the train.}. Rewrite it to be more intense.",
"{they yelled aggressively, over the clanging of the train}",
"Here is some text: {When Mohammed left the theatre, it was already dark out}.",
"Rewrite it to be more about the movie itself., {The movie was longer than Mohammed had expected, and despite the
excellent ratings he was a bit disappointed when he left the theatre.}",
"Here is some text: {next to the path}. Rewrite it to be about France.",
"{next to la Seine}",
"Here is some text: {The man stood outside the grocery store, ringing the bell.}. Rewrite it to be about
clowns.",
"{The man stood outside the circus, holding a bunch of balloons.}",
"Here is some text: {the bell ringing}. Rewrite it to be more flowery.",
"{the peales of the jangling bell}",
"Here is some text: {against the tree}. Rewrite it to be includes the word 'snow'.",
"{against the snow-covered bark of the tree}"
Here is some text: {That is an ugly dress}. Rewrite it to be more positive."
```

Table 7: The exact augmented-zero shot prompts used in our experiments. For *LLM-Dialog*, we replaced "Here is a rewrite of the text, which is" with "Rewrite it to be", and fed each line of the input to the model as individual dialog turns. The blue text is an example of a templated input text and style that would produce the final model output. Note that we can achieve high accuracy even though the prompt formulation resulted in some minor grammitical errors for some styles (e.g., "rewrite it to be include the word 'snow'")

	Acc	BLEU	PPL
LLM-128B			
Zero-shot	69.7	28.6	397
+ cand. select.	31.4	61.5	354
Five-shot	83.2	19.8	240
+ cand. select.	61.5	55.6	306
Augmented zero-shot	79.6	16.1	173
+ cand. select.	65.0	49.3	292
LLM-128B-dialog			
Zero-shot	59.1	17.6	138
+ cand. select.	46.8	24.2	166
Five-shot	94.3	13.6	126
+ cand. select.	81.3	47.6	345
Augmented zero-shot	90.6	10.4	79
+ cand. select.	73.7	40.6	184

Table 8: Sentiment style transfer results with candidate selection (cand. select.). Candidate selection means that of the sixteen examples returned by our model, we choose the one with the highest BLEU with the source sentence.

Instructions: In this task, your goal is to identify whether a desired transformation has been successfully applied to a sentence, without changing the overall meaning of the sentence. Each question contains a sentence marked "original sentence," a desired transformation, and an output sentence where the transformation has been applied.

Each of these questions relates to the same original text and desired transform, but each has a different output transformed sentence. Please rate each transformed sentence along the following three axes:

1) Transferred Style Strength: Does the transformed text has the applied style/transform compared to the original text? For example, if the original text is "I went to the store" and the style is "more angry":

example	score	reasoning
"The store is where I went"	0	The transformed text is no more angry than the original text.
"I went to the stupid store"	50	The transformed text somewhat relates to the style.
"When I went to the store, I couldn't believe how rude the storekeeper was to me!"	100	The text is clearly more angry.

2) Meaning: Does the transformed sentence still have the same overall meaning as the original? It is OK if extra information is added, as long as it doesn't change the underlying people, events, and objects described in the sentence. You should also not penalize for meaning transformations which are necessary for the specified transformation. For example, if the original text is "I love this store" and the style is "more angry":

example	score	reasoning
"it is raining today"	0	the transformed text is about something totally different. It would be hard to tell that the texts are related at all.
"they were out of chicken at the store"	50	The transformed text is mostly related to original-- some modifications of the meaning have been made but they are not egregious
"I adore the store." or "The store was really horrible; it took forever to do my shopping."	100	The text talks about the same concepts as the original, just with different or more words

3) Fluency: Is this sentence fluent english and does it make sense?

example	score	reasoning
"who said that? I thought we were going to go together!"	Yes	This text makes sense
"who, she said it up to me and to me together!"	No	The text is incoherent

Original text: "Everyone in my world had different eye colours."

Desired transformation: more melodramatic

Transformed text: "Everyone in my world had the most intensly colorful eyes, and no one in this world can possibly understand how beautiful they were."

1) Transferred Style Strength: The transformed text has the applied style/transform.

50

2) Meaning: The meaning is preserved between the original and transformed texts (ignoring the ways that the style/transform would change the meaning)

50

3) Fluency: the transformed text is fluent English and it makes sense.

- Yes
- No

Figure 5: The rating UI used for human evaluation. The user may be shown a number of blue squares at once with the same original text and different outputs.

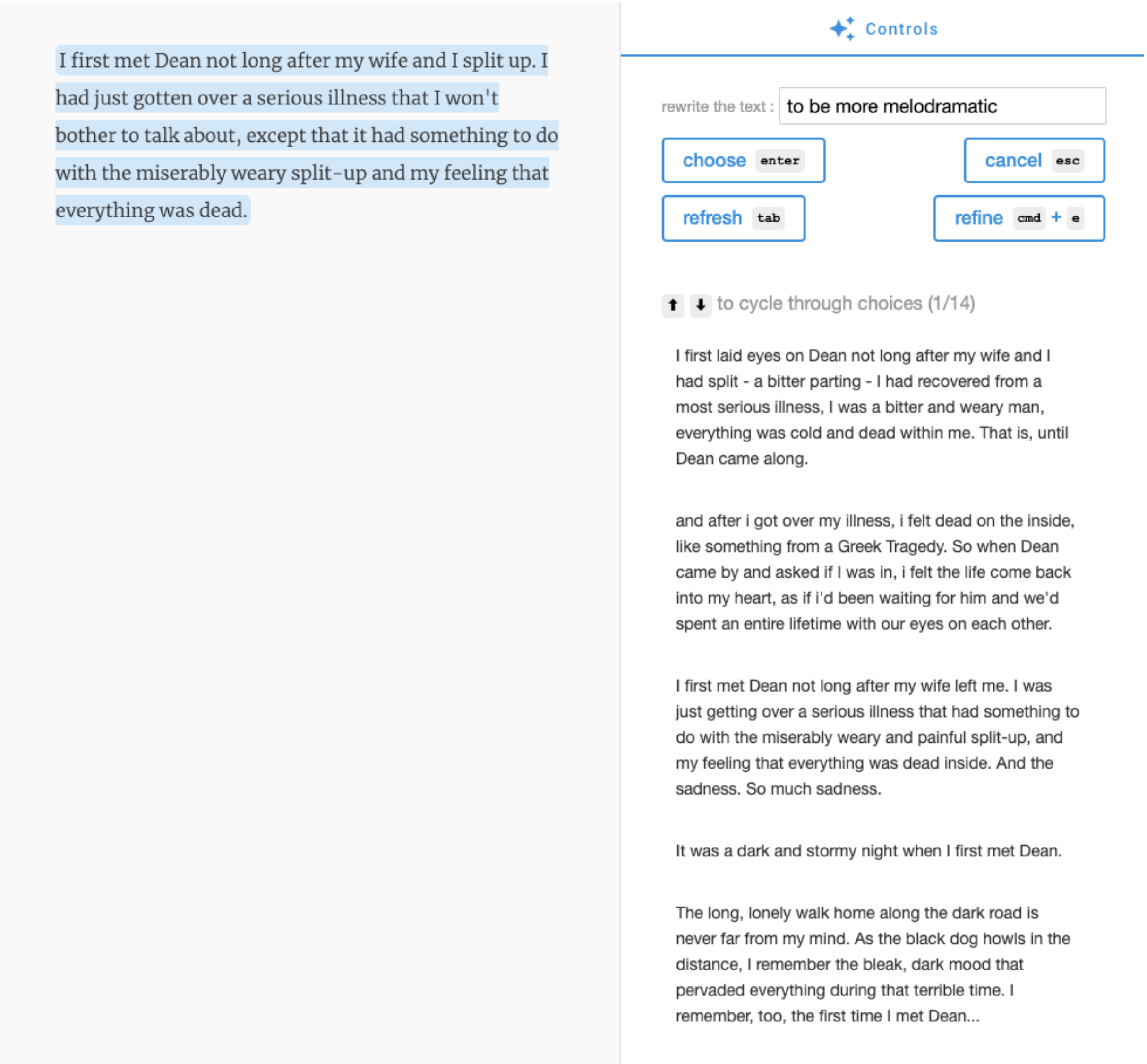

Figure 6: Screenshot AI-assisted editor with 'Rewrite as' feature.

Style	Inputs	Aug. Zero	Zero	Human	Paraphrase
more comic	75	116	63	97	87
more melodromatic	75	124	88	116	87
include the word "park"	75	124	72	94	87
include the word "balloon"	75	135	86	98	87
include a metaphor	75	110	74	110	87
more descriptive	75	190	105	124	87
Overall	75	133	81	107	87

Table 9: The mean length in characters of the inputs and outputs for our six atypical styles.