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.”

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.

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.
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 Catarzano, 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.

<table>
<thead>
<tr>
<th>Original</th>
<th>More descriptive</th>
<th>Include a metaphor</th>
<th>Include the word “park”</th>
<th>Include the word “balloon”</th>
<th>More comic</th>
<th>More melodramatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>There, in the middle of the street, stood an old man in a weatherbeaten brown coat.</td>
<td>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.</td>
<td>There, in the middle of the street, stood an old man whose face resembled the weatherworn bark of a gnarled tree.</td>
<td>There, in the middle of Central Park, stood an old man in a weatherbeaten brown coat.</td>
<td>There, in the middle of the street, stood an old man with several colourful balloons tied to the straps of his coat.</td>
<td>There, in the middle of the street, stood an old man in a raggedy brown coat that was much too big for him.</td>
<td>There, in the middle of the street, stood a grizzled old man, the light of life faded from his sunken eyes.</td>
</tr>
</tbody>
</table>

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). 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

1Hosted by Luo et al. (2019).
and dialog data and Wikipedia. The dataset was
tokenized into 2.81T BPE tokens with a Sentence-
Piece vocabulary size of 32K (Kudo and Richar-
donson, 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 mod-
els (see Table 8). For LLM and GPT-3, we use the
prompts shown in Figure 1 (see 7a for the unab-
abbreviated prompts). For LLM-Dialog, the prompt
is formulated as a conversation between one agent
who is requesting rewrites and another who is per-
foming the rewrites (see Table 7b in the appendix.)

4 Results

4.1 Non-Standard Styles

For our six non-standard styles, we asked six pro-
fessional 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 rat-
ings 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 Ap-
pendix. 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 transfor-
mations 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
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 evalua-
tion was also possible. We found that the “balloon”
and “park” transformations successfully inserted
the target word 85% of the time. For “more descrip-
tive” 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 out-
puts 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 Sud-
hakar et al. (2019), we perform automatic evalua-
tion 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 con-
straints. We automatically evaluate (1) transfer
strength using a sentiment classifier from Hug-
augmented zero-shot learning substan-
tially outperforms vanilla zero-shot learning and
almost reaches the accuracy of five-shot learning.

(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 perplex-
ity 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 lan-
guage 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 sub-
stantially 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 qual-
itative 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

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
UI to write a 100-300 word story, collecting
333 rewrite requests in total. Table 3 shows a sub-
et 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.

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


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 that the output cannot be automatically parsed into usable answers. For example, 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 unparsable 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.
Table 4: Comparing variations of augmented zero-shot learning prompt wording for sentiment style transfer.

<table>
<thead>
<tr>
<th>Model / prompt wording</th>
<th>Acc</th>
<th>Bleu</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“more positive/negative”</td>
<td>76.3</td>
<td>14.8</td>
<td>180</td>
</tr>
<tr>
<td>“happier/sadder”</td>
<td>62.6</td>
<td>15.5</td>
<td>173</td>
</tr>
<tr>
<td>“more optimistic/pessimistic”</td>
<td>69.7</td>
<td>14.1</td>
<td>143</td>
</tr>
<tr>
<td>“more cheerful/miserable”</td>
<td>74.5</td>
<td>15.7</td>
<td>186</td>
</tr>
<tr>
<td>LLM-Dialog</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“more positive/negative”</td>
<td>90.5</td>
<td>10.4</td>
<td>79</td>
</tr>
<tr>
<td>“happier/sadder”</td>
<td>85.9</td>
<td>9.6</td>
<td>90</td>
</tr>
<tr>
<td>“more optimistic/pessimistic”</td>
<td>85.8</td>
<td>10.2</td>
<td>79</td>
</tr>
<tr>
<td>“more cheerful/miserable”</td>
<td>88.8</td>
<td>11.4</td>
<td>93</td>
</tr>
</tbody>
</table>

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...
You honor me,” said Bruno. As he spoke, he began to laugh. His laugh was merry sounding and melodious. He took a deep breath. When he spoke again, his voice was low and strong. “I take this task on and with my brothers shall lead us on to victory!

*Rewrite it to be about mining.*

I reckon,’ smiled Toe, ‘you’d have to say the same for icicles, wouldn’t you?’

*Rewrite it to be more interesting.*

with its southern perfume

*Rewrite it to be more descriptive*

But Sally remained unmoved—and lonely.

*Rewrite it to be more emotional*

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

<table>
<thead>
<tr>
<th>Input</th>
<th>Model suggestion</th>
</tr>
</thead>
<tbody>
<tr>
<td>You honor me,” said Bruno. As he spoke, he began to laugh. His laugh was merry sounding and melodious. He took a deep breath. When he spoke again, his voice was low and strong. “I take this task on and with my brothers shall lead us on to victory!</td>
<td>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 task on and with my brothers shall delve into the mines!’</td>
</tr>
<tr>
<td>‘I reckon,’ smiled Toe, ‘you’d have to say the same for icicles, wouldn’t you?’</td>
<td>‘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.</td>
</tr>
<tr>
<td>with its southern perfume</td>
<td>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.</td>
</tr>
<tr>
<td>But Sally remained unmoved—and lonely.</td>
<td>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.</td>
</tr>
</tbody>
</table>

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.
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.

Here is some text: {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 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: {they asked loudly, over the sound of the train.}. Here is a rewrite of the text, which is more intense.

Here is some text: {next to the path}. Here is a rewrite of the text, which is about France.

Here is some text: {next to la Seine}. Here is a rewrite of the text, which is about France.

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.

Here is some text: {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.

Here is some text: {against the tree}. Here is a rewrite of the text, which includes the word ‘snow’.

Here is some text: {That is an ugly dress}. Here is a rewrite of the text, which is more positive.

<table>
<thead>
<tr>
<th>Augmented Zero-shot Prompt: LLM</th>
<th>Acc</th>
<th>BLEU</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLM-128B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-shot</td>
<td>69.7</td>
<td>28.6</td>
<td>397</td>
</tr>
<tr>
<td>+ cand. select.</td>
<td>31.4</td>
<td>61.5</td>
<td>354</td>
</tr>
<tr>
<td>Five-shot</td>
<td>83.2</td>
<td>19.8</td>
<td>240</td>
</tr>
<tr>
<td>+ cand. select.</td>
<td>61.5</td>
<td>55.6</td>
<td>306</td>
</tr>
<tr>
<td>Augmented zero-shot</td>
<td>79.6</td>
<td>16.1</td>
<td>173</td>
</tr>
<tr>
<td>+ cand. select.</td>
<td>65.0</td>
<td>49.3</td>
<td>292</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Augmented Zero-shot Prompt: LLM-dialog</th>
<th>Acc</th>
<th>BLEU</th>
<th>PPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LLM-128B-dialog</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero-shot</td>
<td>59.1</td>
<td>17.6</td>
<td>138</td>
</tr>
<tr>
<td>+ cand. select.</td>
<td>46.8</td>
<td>24.2</td>
<td>166</td>
</tr>
<tr>
<td>Five-shot</td>
<td>94.3</td>
<td>13.6</td>
<td>126</td>
</tr>
<tr>
<td>+ cand. select.</td>
<td>81.3</td>
<td>47.6</td>
<td>345</td>
</tr>
<tr>
<td>Augmented zero-shot</td>
<td>90.6</td>
<td>10.4</td>
<td>79</td>
</tr>
<tr>
<td>+ cand. select.</td>
<td>73.7</td>
<td>40.6</td>
<td>184</td>
</tr>
</tbody>
</table>

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 grammatical errors for some styles (e.g., “rewrite it to be include the word ‘snow’”)

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":

<table>
<thead>
<tr>
<th>example</th>
<th>score</th>
<th>reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;The store is where I went&quot;</td>
<td>0</td>
<td>The transformed text is no more angry than the original text.</td>
</tr>
<tr>
<td>&quot;I went to the stupid store&quot;</td>
<td>50</td>
<td>The transformed text somewhat relates to the style.</td>
</tr>
<tr>
<td>&quot;When I went to the store, I couldn't believe how rude the storekeeper was to me!&quot;</td>
<td>100</td>
<td>The text is clearly more angry.</td>
</tr>
</tbody>
</table>

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":

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

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

<table>
<thead>
<tr>
<th>example</th>
<th>score</th>
<th>reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;who said that? I thought we were going to go together!&quot;</td>
<td>Yes</td>
<td>This text makes sense</td>
</tr>
<tr>
<td>&quot;who, she said it up to me and to me together!&quot;</td>
<td>No</td>
<td>The text is incoherent</td>
</tr>
</tbody>
</table>

Original text: "Everyone in my world had different eye colours."
Desired transformation: more melodramatic
Transformed text: "Everyone in my world had the most intensely 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.

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

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.
I first met Dean not long after my wife and I split up. I had just gotten over a serious illness that I won’t bother to talk about, except that it had something to do with the miserably weary split-up and my feeling that everything was dead.

and after I got over my illness, I felt dead on the inside, like something from a Greek Tragedy. So when Dean came by and asked if I was in, I felt the life come back into my heart, as if I’d been waiting for him and we’d spent an entire lifetime with our eyes on each other.

I first met Dean not long after my wife left me. I was just getting over a serious illness that had something to do with the miserably weary and painful split-up, and my feeling that everything was dead inside. And the sadness. So much sadness.

It was a dark and stormy night when I first met Dean.

The long, lonely walk home along the dark road is never far from my mind. As the black dog howls in the distance, I remember the bleak, dark mood that pervaded everything during that terrible time. I remember, too, the first time I met Dean...

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

<table>
<thead>
<tr>
<th>Style</th>
<th>Inputs</th>
<th>Aug. Zero</th>
<th>Zero</th>
<th>Human</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>more comic</td>
<td>75</td>
<td>116</td>
<td>63</td>
<td>97</td>
<td>87</td>
</tr>
<tr>
<td>more melodramatic</td>
<td>75</td>
<td>124</td>
<td>88</td>
<td>116</td>
<td>87</td>
</tr>
<tr>
<td>include the word “park”</td>
<td>75</td>
<td>124</td>
<td>72</td>
<td>94</td>
<td>87</td>
</tr>
<tr>
<td>include the word “balloon”</td>
<td>75</td>
<td>135</td>
<td>86</td>
<td>98</td>
<td>87</td>
</tr>
<tr>
<td>include a metaphor</td>
<td>75</td>
<td>110</td>
<td>74</td>
<td>110</td>
<td>87</td>
</tr>
<tr>
<td>more descriptive</td>
<td>75</td>
<td>190</td>
<td>105</td>
<td>124</td>
<td>87</td>
</tr>
<tr>
<td>Overall</td>
<td>75</td>
<td>133</td>
<td>81</td>
<td>107</td>
<td>87</td>
</tr>
</tbody>
</table>

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