

## A Appendix

### A.1 Reduction-and-Synthesis

Given the source text  $X$ , the expected inference  $Y$  with the target style  $s$ , we assume that a neutral text  $C$  sharing the same semantic information with  $X$  entails the style-free content which is preserved during transferring from  $X$  to  $Y$ . The SST task can be further decomposed as Eq. 4:

$$\begin{aligned}
 P(Y|X, s) &= \frac{P(Y, X, s)}{P(X, s)} \\
 &\geq \frac{P(Y, X, C, s)}{P(X, s)} \\
 &= \frac{P(Y, X, C, s)}{P(X) P(s)} \\
 &= \frac{P(X, C)}{P(X)} \cdot \frac{P(Y, X, C, s)}{P(X, C) P(s)} \\
 &= \frac{P(X, C)}{P(X)} \cdot \frac{P(Y, X, C, s)}{P(X, C, s)} \\
 &= \underbrace{P(C|X)}_{\text{reduction}} \underbrace{P(Y|X, C, s)}_{\text{synthesis}} \quad (4)
 \end{aligned}$$

### A.2 Hyperparameter

Considering the time and computing cost, We choose the LLaMA2-13B (et al, 2023) as the backbone during inference. The model is experimented with Pytorch on one NVIDIA A6000 GPU (48GB memory). The main hyper-parameters are shown in Table 4. For a fair comparison with related work, we utilized the same version of the Yelp and Amazon datasets cleaned by Suzgun et al. (2022).

Name	Value
max sequence length	1,024
max generation length	96
max batch size	4
the value of top_p	0.9
the value of temperature	0.6

Table 4: Hyper-parameter setting for LLaMA-2-13B during inference.

### A.3 Additional Experimental Results

Table 5 illustrates the performance with different LLMs for both transfer directions ( $neg \rightarrow pos$ , and  $pos \rightarrow neg$ ) on Yelp dataset. We explored the experiments with three popular open-source LLMs (Mixtral, Gemma, and LLaMA with the same 7B size). For a fair comparison, we use the Ollama<sup>6</sup>, a tool for running LLMs in local, to infer

all results. As shown in Table 5, the overall performance obtained by the baseline is the worst among the three models. In contrast, our BL+RS shows the improvement except for **r-sB** and **s-sB** in both  $neg \rightarrow pos$  and  $pos \rightarrow neg$ .

Table 6 shows the results obtained by our reduction-synthesis (RS) method and baseline (BL) in four challenging SST cases. The examples shown in Table 6 are randomly selected from the challenging cases on the Yelp dataset.

We also conducted the experiments by using the Amazon dataset. Table 7 and 8 show the comparison with the baseline and the distribution of the style of input/output at each phase, respectively.

### A.4 Prompt Templates

Three types of prompt templates, i.e., generation, feedback, and refine on the Yelp dataset are illustrated in Figures 3 ~ 11. Figures.3, 4, and 5 indicates the Self-Refine baseline. Figures.6, 7, and 8 refer to reduction phase, and Figures.9, 10, and 11 shows synthesis phase.

<sup>6</sup><https://github.com/ollama/ollama>

Model		$neg \rightarrow pos$					$pos \rightarrow neg$				
		Acc $\uparrow$	r-sB $\uparrow$	s-sB $\uparrow$	r/s-sB $\uparrow$	t-PPL $\downarrow$	Acc $\uparrow$	r-sB $\uparrow$	s-sB $\uparrow$	r/s-sB $\uparrow$	t-PPL $\downarrow$
<b>Mistral-7B</b>	BL	82.0	<b>14.1</b>	<b>15.9</b>	0.883	28	95.6	<b>14.2</b>	<b>19.9</b>	0.715	46
	RS	74.8	11.9	15.0	0.789	30	93.8	11.0	14.8	0.742	58
	BL+RS	<b>86.4</b>	13.7	15.3	<b>0.897</b>	<b>27</b>	<b>97.0</b>	<b>14.2</b>	19.4	<b>0.730</b>	<b>45</b>
impv. (%)		+5.4	-2.8	-3.8	+1.6	+3.6	+1.5	+0	-2.5	+2.1	+2.2
<b>Gemma-7B</b>	BL	52.4	<b>21.2</b>	<b>44.3</b>	0.479	38	88.8	<b>22.6</b>	<b>44.2</b>	0.513	61
	RS	33.6	17.1	38.9	0.438	28	81.0	20.6	38.4	0.536	56
	BL+RS	<b>58.6</b>	20.5	42.3	<b>0.485</b>	<b>38</b>	<b>92.6</b>	22.3	41.8	<b>0.534</b>	<b>59</b>
impv. (%)		+11.8	-3.3)	-4.5	+1.3	+7.9	+4.3	-1.3	-5.4	+4.1	+3.3
<b>LLaMA-7B</b>	BL	75.4	<b>7.9</b>	<b>8.8</b>	0.901	23	97.0	<b>7.9</b>	<b>9.8</b>	0.806	<b>32</b>
	RS	76.0	7.8	<b>8.8</b>	0.888	26	97.6	7.1	9.3	0.763	41
	BL+RS	<b>81.8</b>	7.7	8.2	<b>0.938</b>	<b>22</b>	<b>98.6</b>	7.8	9.7	<b>0.809</b>	<b>32</b>
impv. (%)		+8.5	-2.5	-6.8	+4.1	+4.3	+1.6	-1.3	-1.0	+0.4	0

Table 5: Comparison with the Self-Refine (baseline, represented with BL) on Yelp dataset by different LLMs. The RS indicates the plug-and-play method, and the BL+RS is the method augmenting the BL with RS, i.e., replacing the incorrect output of BL with the generation of RS. The **bold** font refers to the best performance of each metric. The "impv." means the improvements of BL+RS against the baseline.

$neg \rightarrow pos$	
Case 1	you won't find a better selection in scottsdale.
Human	you won't find a better selection in arizona.
BL (✗)	You won't find a better selection in scottsdale.
RS (✓)	You will find the best selection of things in scottsdale.
Case 2	then he proceeded to tell me how to fix it.
Human	He told me kindly how to fix it.
BL (✗)	He told me how to fix it.
RS (✗)	Then he told me how to fix the problem, and I did.
$pos \rightarrow neg$	
Case 3	a great stop if you want nice things at a steal.
Human	a great stop if you want bad things at scammy prices
BL (✗)	a great stop if you want nice things at a steal.
RS (✓)	a terrible place to shop if you want to buy things at a good price.
Case 4	if you're young and are into sports this is the place for you.
Human	no matter your age, you won't like it
BL (✗)	if you're old and are into sports this is the place for you.
RS (✗)	This is a place for young people who are into sports.

Table 6: The examples obtained by our reduction-synthesis (RS) method and baseline (BL) for challenging SST cases in the  $neg \rightarrow pos$  and  $pos \rightarrow neg$  tasks. The ✓ and ✗ indicate the text is transferred correctly and incorrectly, respectively.

Model	$neg \rightarrow pos$					$pos \rightarrow neg$				
	Acc $\uparrow$	r-sB $\uparrow$	s-sB $\uparrow$	r/s-sB $\uparrow$	t-PPL $\downarrow$	Acc $\uparrow$	r-sB $\uparrow$	s-sB $\uparrow$	r/s-sB $\uparrow$	t-PPL $\downarrow$
BL	30.4	<b>32.5</b>	<b>64.4</b>	0.505	63	34.0	<b>40.8</b>	<b>76.5</b>	0.533	71
RS	32.6	30.6	58.6	0.526	60	<b>37.8</b>	31.4	57.4	0.547	51
BL+RS	<b>38.2</b>	31.1	60.7	<b>0.513</b>	<b>58</b>	<b>45.4</b>	38.7	70.1	<b>0.552</b>	<b>62</b>
impv. (%)	+25.7	-4.3	-5.7	+2.0	+7.9	+33.5	-5.1	-8.4	+5.5	+12.7

Table 7: Comparison with the Self-Refine (baseline, represented with BL) on Amazon dataset. The RS indicates the plug-and-play method, and the BL+RS is the method augmenting the BL with RS, that is, replacing the incorrect output of BL with the generation of RS. The **bold** font shows the best performance for each metric. The "impv." means the improvements of BL+RS against the baseline.

Style	$neg \rightarrow pos$			$pos \rightarrow neg$		
	Reduction (%)	Synthesis (%)	Self-Refine (%)	Reduction (%)	Synthesis (%)	Self-Refine (%)
$s_o = neg$	199 (88.0)	88 (40.6)	101 (44.7)	71 (81.6)	90 (90.0)	82 (94.3)
$s_i = neg$ $s_o = neu$	21 (9.3)	33 (15.2)	29 (12.8)	12 (13.8)	4 (4.0)	4 (4.6)
$s_o = pos$	6 (2.7)	96 (44.2)	96 (42.5)	4 (4.6)	6 (6.0)	1 (1.1)
$s_i = neg$	<b>226</b>	<b>217</b>	<b>226</b>	<b>87</b>	<b>100</b>	<b>87</b>
$s_o = neg$	14 (7.7)	11 (5.7)	3 (2.2)	14 (6.9)	94 (40.9)	32 (15.8)
$s_i = neu$ $s_o = neu$	160 (87.9)	117 (60.6)	127 (93.4)	171 (84.7)	123 (53.5)	162 (80.2)
$s_o = pos$	8 (4.4)	65 (33.7)	6 (4.4)	17 (8.4)	13 (5.6)	8 (4.0)
$s_i = neu$	<b>182</b>	<b>193</b>	<b>136</b>	<b>202</b>	<b>230</b>	<b>202</b>
$s_o = neg$	4 (4.3)	2 (2.2)	0 (0.0)	15 (7.1)	63 (37.1)	81 (38.4)
$s_i = pos$ $s_o = neu$	12 (13.0)	2 (2.2)	1 (1.1)	47 (22.3)	8 (4.7)	8 (3.8)
$s_o = pos$	76 (82.6)	86 (95.6)	91 (98.9)	149 (70.6)	99 (58.2)	122 (57.8)
$s_i = pos$	<b>92</b>	<b>90</b>	<b>92</b>	<b>211</b>	<b>170</b>	<b>211</b>

Table 8: Distribution of the style of input and output pairs during every transfer phase on Amazon data. Self-Refine is the baseline that directly transfers the input to the target. The background   indicates the number and rate of correct results in each transfer phrase

```

###
Text: The chicken I ordered in this restaurant is tasteless.
Rewrite the text to express the content with positive emotions.
Rewrite: I went to the restaurant and ate some chicken, it is delicious.
###
Text: Salads are inappropriate for appetizers.
Rewrite the text to express the content with positive emotions.
Rewrite: Salads are a delicious way to begin the meal.
###

```

Figure 3: The generation prompt of the Self-Refine baseline. The task is  $neg \rightarrow pos$  transfer on Yelp data.

```

###
Text: The chicken I ordered in this restaurant is tasteless.
Rewrite the text to express the content with positive emotions.
Rewrite: I went to the restaurant and ate some chicken.
Does this rewrite meet the requirements?
Feedback: No, the rewrite just express the same content without positive emotions.
###
Text: Salads are inappropriate for appetizers.
Rewrite the text to express the content with positive emotions.
Rewrite: Salads are an appropriate way to begin the meal.
Does this rewrite meet the requirements?
Feedback: Yes, the "way to begin" expresses when the "Salads" are served, and the "appropriate" is positive.
###

```

Figure 4: The feedback prompt of the Self-Refine baseline. The task is  $neg \rightarrow pos$  transfer on Yelp data.

```

###
Text: The chicken I ordered in this restaurant is tasteless.
Rewrite the text to express the content with positive emotions.
Rewrite: I went to the restaurant and ate some chicken.
Does this rewrite meet the requirements?
Feedback: No, the rewrite just express the same content without positive emotions.
Okay, let's try again. Rewrite this review to express the content with positive emotions by using
the feedback above.
Rewrite: I ate some noodles in this restaurant, it is tasteless.
Does this rewrite meet the requirements?
Feedback: No, the rewrite does not mention the taste of "chicken" which is the topic of the
text.
Rewrite: I went to the restaurant and ate some chicken, it is delicious.
###
Text: Salads are inappropriate for appetizers.
Rewrite the text to express the content with positive emotions.
Rewrite: Two staffs are serving for me, they are kind.
Does this rewrite meet the requirements?
Feedback: No, the "staffs are serving" is different from the topic about the taste of "Salads".
Okay, let's try again. Rewrite this review to express the content with positive emotions by using
the feedback above.
Rewrite: Salads are an inappropriate way to begin the meal.
Does this rewrite meet the requirements?
Feedback: No, the "way to begin" expresses when the "Salads" are served, but the "inappropri-
ate" is still negative.
Okay, let's try again. Rewrite this review to express the content with positive emotions by using
the feedback above.
Rewrite: Salads are an appropriate way to begin the meal.
###

```

Figure 5: The refine prompt of the Self-Refine baseline. The task is  $neg \rightarrow pos$  transfer on Yelp data.

```

###
Text: The chicken I ordered in this restaurant is tasteless.
Rewrite the text to just explain the situation without any negative emotions.
Rewrite: I went to the restaurant and ate some chicken.
###
Text: Salads are inappropriate for appetizers.
Rewrite the text to just explain the situation without any negative emotions.
Rewrite: Salads are served to begin the meal.
###

```

Figure 6: The generation prompt at the Reduction phase. The task is  $neg \rightarrow pos$  transfer on Yelp data.

```

###
Text: The chicken I ordered in this restaurant is tasteless.
Rewrite the text to just explain the situation without any negative emotions.
Rewrite: The chicken I ordered in this restaurant is tasteless.
Does this rewrite meet the requirements?
Feedback: No, the rewrite just duplicates the negative text, and “tasteless” represents negative sentiment.
###
Text: Salads are inappropriate for appetizers.
Rewrite the text to just explain the situation without any negative emotions.
Rewrite: Salads are served to begin the meal.
Does this rewrite meet the requirements?
Feedback: Yes, the rewrite expresses the content neutrally.
###

```

Figure 7: The feedback prompt at the Reduction phase. The task is *neg* → *pos* transfer on Yelp data.

```

###
Text: The chicken I ordered in this restaurant is tasteless.
Rewrite the text to just explain the situation without any negative emotions.
Rewrite: The chicken I ordered in this restaurant is tasteless.
Does this rewrite meet the requirements?
Feedback: No, the rewrite just duplicates the negative text, and “tasteless” represents negative sentiment.
Okay, let’s try again. Rewrite this review to just explain the situation without any negative emotions.
Rewrite: The chicken of the restaurant is not fresh.
Does this rewrite meet the requirements?
Feedback: No, the "chicken of the restaurant" express the same topic, but the "not fresh" is still negative.
Okay, let’s try again. Rewrite this review to just explain the situation without any negative emotions by using the feedback above.
Rewrite: I went to the restaurant and ate some chicken.
###
Text: Salads are inappropriate for appetizers.
Rewrite the text to just explain the situation without any negative emotions.
Rewrite: Two staffs are serving for me.
Does this rewrite meet the requirements?
Feedback: No, the "staffs are serving" is different from the topic about the "Salads".
Okay, let’s try again. Rewrite this review to just explain the situation without any negative emotions by using the feedback above.
Rewrite: Salads are served.
Does this rewrite meet the requirements?
Feedback: No, the rewrite is the same topic about "salads" but it does not mention when the "salads" are served.
Okay, let’s try again. Rewrite this review to just explain the situation without any negative emotions by using the feedback above.
Rewrite: Salads are served to begin the meal.
###

```

Figure 8: The refine prompt at the Reduction phase. The task is *neg* → *pos* transfer on Yelp data.

###  
**Text:** The chicken I ordered in this restaurant is tasteless.  
**Content of the text:** I went to the restaurant and ate some chicken.  
Rewrite the text to express the content with positive emotions.  
**Rewrite:** I went to the restaurant and ate some chicken, it is delicious.  
###  
**Text:** Salads are inappropriate for appetizers.  
**Content of the text:** Salads are served to begin the meal.  
Rewrite the text to express the content with positive emotions.  
**Rewrite:** Salads are a delicious way to begin the meal.  
###

Figure 9: The generation prompt at the Synthesis phase. The task is  $neg \rightarrow pos$  transfer on Yelp data.

###  
**Text:** The chicken I ordered in this restaurant is tasteless.  
**Content of the text:** I went to the restaurant and ate some chicken.  
Rewrite the text to express the content with positive emotions.  
**Rewrite:** I ate some noodles in this restaurant, it is tasteless.  
Does this rewrite meet the requirements?  
**Feedback:** No, the rewrite does not mention the taste of “chicken” which is the topic of the text.  
###  
**Text:** Salads are inappropriate for appetizers.  
**Content of the text:** Salads are served to begin the meal.  
Rewrite the text to express the content with positive emotions.  
**Rewrite:** Salads are a delicious way to begin the meal.  
Does this rewrite meet the requirements?  
**Feedback:** Yes, the rewrite expresses when the "Salads" are served, the "they are delicious" are positive.  
###

Figure 10: The feedback prompt at the Synthesis phase. The task is  $neg \rightarrow pos$  transfer on Yelp data.

f ###

**Text:** The chicken I ordered in this restaurant is tasteless.

**Content of the text:** I went to the restaurant and ate some chicken.

Rewrite the text to express the content with positive emotions.

**Rewrite:** I ate some chicken in this restaurant.

Does this rewrite meet the requirements?

**Feedback:** No, the rewrite just expresses the same content without positive emotions.

Okay, let's try again. Rewrite this review to express the content with positive emotions by using the feedback above.

**Rewrite:** I ate some noodles in this restaurant, it is tasteless.

Does this rewrite meet the requirements?

**Feedback:** No, the rewrite does not mention the taste of "chicken" which is the topic of the text.

Okay, let's try again. Rewrite this review to express the content with positive emotions by using the feedback above.

**Rewrite:** I ate some chicken in this restaurant, it is tasteless..

###

**Text:** Salads are inappropriate for appetizers.

**Content of the text:** Salads are served to begin the meal.

Rewrite the text to express the content with positive emotions.

**Rewrite:** Two staff are serving for me, they are kind.

Does this rewrite meet the requirements?

**Feedback:** No, the "staff are serving" is different from the topic about the "Salads", although the "kind" is positive.

Okay, let's try again. Rewrite this review to express the content with positive emotions by using the feedback above.

**Rewrite:** Salads are delicious.

Does this rewrite meet the requirements?

**Feedback:** No, the rewrite is the same topic about "salads", but it does not mention when the "salads" are served.

Okay, let's try again. Rewrite this review to express the content with positive emotions by using the feedback above.

**Rewrite:** Salads are an appropriate way to begin the meal.

###

Figure 11: The refine prompt at the Synthesis phase. The task is  $neg \rightarrow pos$  transfer on Yelp data.