	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
Style Transfer Accuracy												
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	80.00	11.20	47.67	29.04	35.50	48.20	79.50	14.80	63.80	43.80	76.40	52.80
APE	74.00	12.20	47.57	28.44	**40.90**	44.80	77.10	18.20	55.80	44.60	78.90	48.00
AVF	76.00	12.40	47.57	28.44	38.80	44.20	77.90	18.70	55.60	44.40	**79.20**	47.90
PNMA	73.85	8.70	42.43	23.79	35.57	37.05	72.84	14.16	53.74	37.58	75.39	41.71
Our	**80.80**	**14.40**	**55.36**	**31.98**	37.81	**50.30**	**80.63**	**23.27**	**73.40**	**45.14**	77.93	**54.73**
Content Preservation												
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	**85.95**	74.71	73.54	82.71	82.48	75.77	75.32	**89.14**	78.75	**62.28**	76.17	**74.47**
APE	76.72	85.06	**76.72**	83.00	**87.99**	**82.21**	76.80	87.89	80.07	57.61	**76.52**	73.53
AVF	75.21	84.53	76.63	**83.57**	86.92	80.68	**76.94**	87.32	**80.94**	58.98	76.15	73.95
PNMA	75.52	84.11	75.67	82.54	86.79	80.67	76.04	86.93	79.22	57.42	75.04	72.67
Our	85.84	**86.28**	75.85	80.10	82.32	74.96	75.65	82.47	77.19	60.92	75.25	74.21
Fluency												
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
LLaMA-3	92.53	87.69	113.84	191.30	88.22	68.49	105.35	92.34	197.62	136.03	177.01	125.98
APE	94.27	89.93	133.12	188.34	88.51	69.06	108.24	95.17	250.65	133.92	**151.06**	126.73
AVF	96.63	89.36	131.10	191.29	87.93	75.94	112.67	97.50	220.30	**126.42**	151.33	130.17
PNMA	103.61	90.85	136.27	194.71	96.31	77.95	111.77	101.61	260.52	135.00	154.85	129.49
Our	**90.79**	**81.46**	**85.65**	**172.26**	**85.28**	**66.68**	**104.92**	**83.36**	**151.71**	134.86	174.46	**110.48**

Table 2: **Main Results:** Style transfer accuracy (higher values are better; ↑), content preservation (↑) and fluency (↓) on 6 datasets across 12 transfer directions. Best results are highlighted in bold.

vation is largely attributable to the copy mechanism, i.e., the generated text tends to prioritize maintaining the original semantics, thereby neglecting the stylistic differences. A detailed discussion on this can be found in Section 6.2. Another potential explanation is the semantic gap, which varies significantly between sentences of different styles, and for which no effective metric currently exists to fully measure this gap. For example, when transferring text from an informal to a formal style, the original text *"Sorry about that."* and the target text *"I apologize for the inconvenience caused."* are stylistically aligned, but they diverge significantly in semantic space. This is reflected in a low cosine similarity score of 0.447 between them.

Different Directions. We observe significant performance discrepancies when transferring between different directions within the same task. For example, transferring from impolite to polite achieves a style accuracy of nearly 80%, whereas the reverse direction achieves only about 12%. This disparity can be attributed to the training data of LLMs, which predominantly consist of positive corpora (e.g., polite, neutral, formal), with inadequate representation from negative corpora. Additionally, LLMs have a tendency to generate safer responses (Touvron et al., 2023), which can compromise the utility of tasks involving style transfer.

6 Analysis

In this section, we conduct an ablation study to verify the criticality of eliminating overlap between source- and target-side style neurons, alongside the importance of neuron deactivation and contrastive decofing (Section 6.1). Subsequently, we conduct a detailed analysis of the copy problem in the TST task (Section 6.2). Finally, we delve into several other significant factors inherent to our approach (Section 6.4).

6.1 Ablation Study

We conduct an ablation study, detailed in Table 3, to evaluate the effectiveness of removing overlapping source- and target-style neurons. The results demonstrate a considerable advantage in eliminating such overlap compared to allowing mixed patterns of neuron activation. As highlighted by the statistics in Section 3.1, there is a substantial 95% overlap in most neurons, indicating that source style neurons largely coincide with target style neurons, meking them nearly indistinguishable when directly decoding using LLMs.

Additionally, Table 4 presents the results of ablating neuron deactivation and contrastive decoding

Style		without	with
Formality	informal→formal	74.00	**79.40**
	formal→informal	12.20	**13.63**
Toxicity	toxic→neutral	47.57	**49.78**
	neutral→toxic	28.44	**29.82**
Politics	democratic→republican	**40.90**	37.51
	republican→democratic	44.80	**49.70**
Politeness	impolite→polite	77.10	**80.10**
	polite→impolite	18.20	**21.73**
Authorship	shakespeare→modern	55.80	**63.00**
	modern→shakespeare	44.60	**45.42**
Sentiment	positive→negative	78.90	**79.75**
	negative→positive	48.00	**51.70**

Table 3: **Ablation study:** Style transfer accuracy on removing overlap between source- and target-side style neurons in six benchmarks. "with" indicates the removal of overlap.

	Deactivate	Contrastive	Toxicity		Authorship	
			toxic	neutral	shakespeare	modern
#1	✗	✗	47.67	29.04	63.80	43.80
#2	✓	✗	52.63	31.07	68.39	44.71
#3	✗	✓	46.82	28.31	63.23	43.16
#4	✓	✓	**55.36**	**31.98**	**73.40**	**45.14**

Table 4: **Ablation study:** Style transfer accuracy for neuron deactivation and contrastive decoding on the toxicity and authorship tasks. "✓" means the inclusion of the neuron deactivation or contrastive decoding steps, while "✗" means they are turned off. #1 indicates the results from baseline LLaMA-3 model, which do not use the deactivation nor the contrastive steps.

(CD). Our findings are as follows: **(1)** Comparing #1 and #2, we observe a significant impact of deactivating neurons on the final results. This is because deactivating neurons on the source side encourages the LLMs to generate words in the target style. **(2)** Comparing #1 and #3, we find that using CD alone does not significantly improve and may even degrade the results. This is attributed to the fact that style-related information is processed in later layers, and simply comparing these layers does not yield substantial improvements. Without deactivating neurons, the target style words are not effectively generated, resulting in minimal JSD distance between the style layers and the final layer, thereby reducing the effectiveness of CD. **(3)** Experiment #4 demonstrates that optimal performance is achieved when both deactivating source-side style neurons and employing CD. Deactivating neurons enhances the probability to generate target style vocabulary, as discussed in Section 3.2, albeit at the cost of fluency in generated sentences. Therefore, CD proves crucial in further enhancing the fluency of sentences.

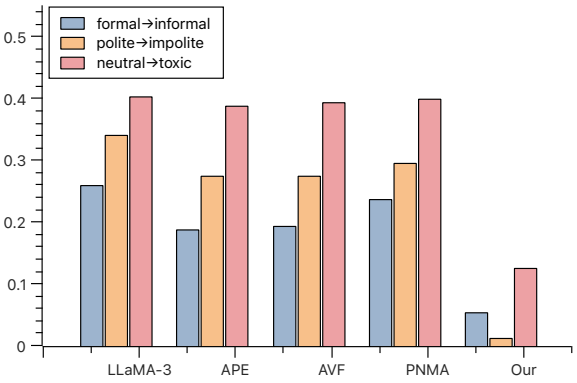

Figure 4: Copy Ratio on three selected TST tasks. Lower value indicates better performance of the model.

6.2 Copy Problem

The "copy problem" arises when models simply reproduce the input text unchanged in the output, a challenge prevalent in multilingual machine translation (Lai et al., 2023a,b). Given the goal to maintain semantic consistency of the input sentences in TST, LLMs often resort to direct copying. To investigate this phenomenon, we analyze tasks related to *formality*, *politeness*, and *toxicity*. Figure 4 illustrates a significant number of copy instances in the original LLaMA-3, indicating a preference for preserving semantic meaning rather than incorporating stylistic variations in TST. Neuron-based approaches (*APE*, *AVF*, and *PNMA*) partially mitigate this issue by controlling neuron activation, thereby producing more target-style words during decoding, as evidenced in Section 3. However, these baselines suffer in performance due to their inability to fully eliminate overlap between source and target style neurons. In contrast, our approach achieves a reduced copy rate by deactivating source-side neurons and employing a novel decoding strategy.

6.3 Case Study

To further demonstrate the effectiveness of our approach, we conduct a case study on three style transfer tasks: informal to formal, impolite to polite, and negative to positive. As shown in Table 5, the original LLaMA model often retained a higher number of words from the source text, sometimes copying them verbatim—an issue discussed in Section 6.2. In contrast, by adjusting style neurons, we guided the model to generate more varied vocabulary aligned with the target style. For instance, the model produced the word "moist" as part of a

Prompt:	Please transfer the following {source_style} sentence into a {target_style} sentence and maintain the meaning of the sentence. {source_style} Sentence: {text}. Please only return the {target_style} sentence.	
Informal → Formal	**Informal Text:**	i don't know. and my exams r not over yet still 1 remaining i.e on 23-dec.
	LLaMA-8B:	I don't know and my exams are not over yet. I have one remaining exam on December 23rd.
	Our:	My exams are not over yet because I still have one (i.e December 23rd) left.
Impolite → Polite	**Impolite Text:**	It's hot, open the window.
	LLaMA-8B:	It's hot, please open the window.
	Our:	Do you mind if I open the window?
Negative → Positive	**Negative Text:**	it was super dry and had a weird taste to the entire slice .
	LLaMA-8B:	It was super dry and had a weird taste to the entire slice.
	Our:	The slice was moist and taste good.

Table 5: Case study on informal→formal, impolite→polite and negative→positive tasks.

positive style transformation.

6.4 Further Analysis

We conduct a comprehensive analysis of our method across various dimensions, including different model (Appendix D), layer selection strategies (Appendix E), content preservation metrics (Appendix F), and decoding strategies (Appendix G), yielding several key insights: **(1)** Our method consistently demonstrates effectiveness across diverse model sizes, including larger models like 70B. **(2)** Selecting the last few layers proves optimal compared to earlier layers. **(3)** Different strategies for preserving meaning yield similar outcomes, highlighting the importance of exploring innovative approaches in future research. **(4)** Contrastive decoding exhibits significant advantages over traditional decoding methods in the TST task, motivating our adoption of CD strategy.

7 Conclusion

We revisit the TST task in LLMs through a neuronal analysis perspective. Our study focuses on identifying style-specific neurons within LLMs, highlighting the critical importance of removing overlap between source- and target-side stylistic neurons. We find that deactivating source-specific neurons enhances the probability of generating target-style words but may compromise the fluency of generated sentences. To mitigate this issue, we adapt the state-of-the-art contrastive decoding method (Dola) for TST, ensuring both the fluency and effective style transformation of generated sentences. Experimental results across six benchmarks demonstrate the efficacy of our approach.

8 Limitations

This work has the following limitations: (1) We deactivate style-specific neurons across all layers; however, considering other layers may yield additional insights. For instance, Zhao et al. (2024) found that deactivating neurons in different layers (e.g., understanding layer or generating layer) can have subtle effects on experimental results. We will consider this as a direction for future research. (2) We evaluate our approach only on the text style transfer task; however, our method has the potential to be applied to other style-related tasks, such as image style transfer (Wang et al., 2024) and multilingual style transfer (Mukherjee et al., 2024b). Furthermore, our approach is task-agnostic, with significant potential to adapt to other tasks, such as identifying domain-specific neurons and applying them to domain adaptation tasks (Lai et al., 2022a,b).

Acknowledgement

The work was supported by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101113091) and by the German Research Foundation (DFG; grant FR 2829/7-1).