Style-Specific Neurons for Steering LLMs in Text Style Transfer

Wen Lai[1,2]**, Viktor Hangya**[2,3]**, Alexander Fraser**[1,2]

[1] School of Computation, Information and Technology, Technical University of Munich, Germany
[2]Munich Center for Machine Learning, Germany
[3]Center for Information and Language Processing, LMU Munich, Germany
{wen.lai, alexander.fraser}@tum.de, hangyav@cis.lmu.de

Abstract

Text style transfer (TST) aims to modify the style of a text without altering its original meaning. Large language models (LLMs) demonstrate superior performance across multiple tasks, including TST. However, in zero-shot setups, they tend to directly copy a significant portion of the input text to the output without effectively changing its style. To enhance the stylistic variety and fluency of the text, we present *sNeuron-TST*, a novel approach for steering LLMs using style-specific neurons in TST. Specifically, we identify neurons associated with the source and target styles and deactivate source-style-only neurons to give target-style words a higher probability, aiming to enhance the stylistic diversity of the generated text. However, we find that this deactivation negatively impacts the fluency of the generated text, which we address by proposing an improved contrastive decoding method that accounts for rapid token probability shifts across layers caused by deactivated source-style neurons. Empirical experiments demonstrate the effectiveness of the proposed method on six benchmarks, encompassing formality, toxicity, politics, politeness, authorship, and sentiment[1].

1 Introduction

Text style transfer (TST; Jin et al., 2022; Hu et al., 2022) aims to transform text from a source style to a target style while maintaining the original content and ensuring the fluency of the generated text. Given any text x in an original style s_1, the objective of TST is to transform x into a new text $\hat{x}$ in a different style s_2 ($s_2 \neq s_1$), ensuring that the content remains unchanged despite the shift in style. Large language models (LLMs; Minaee et al., 2024) exhibit exceptional performance across various NLP tasks (Chang et al., 2024), including TST (Ostheimer et al., 2023; Chen, 2024). However, existing LLMs (e.g., LLaMA-3 Meta, 2024)

tend to prioritize preserving the original meaning over enhancing stylistic differences in TST. Our analysis reveals that 34% of the outputs generated by LLaMA-3 are identical to the input text when tasked with transferring polite text to impolite text (Section 6.2). Enhancing the generation of words that align with the target style during the decoding process remains a significant challenge in TST.

Recent LLMs have been successfully applied to TST, broadly categorized into two approaches: (i) employing single-style or parallel-style text data for either full-parameter or parameter-efficient fine-tuning (Mukherjee et al., 2024c,a), and (ii) leveraging the robust in-context learning capabilities of LLMs to create specialized prompts for zero-shot or few-shot learning (Chen, 2024; Pan et al., 2024). However, (i) typically requires substantial data and computational resources to achieve good results, while (ii) is highly sensitive to prompts, where even minor changes can significantly impact the outcomes (Chen et al., 2023).

Neuron analysis (Xiao et al., 2024), which aims to identify and understand the roles of individual neurons within a neural network, is a crucial method for enhancing the interpretability of neural networks and has garnered increasing attention in recent years. By identifying neurons associated with specific attributes such as language (Zhao et al., 2024), knowledge (Niu et al., 2024), and skill (Wang et al., 2022), neuron analysis can boost performance on targeted tasks. Recent research has demonstrated that focusing on language-specific neurons can markedly enhance the multilingual capabilities of LLMs during the decoding stage (Kojima et al., 2024; Tan et al., 2024). However, the exploration of style-specific neurons remains relatively underexplored until now.

Thus motivated, we raise the following two research questions:

Q1: Do LLMs possess neurons that specialize in processing style-specific text?

Q2: If such neurons exist, how can we optimize their utilization during the decoding process to steer LLMs in generating text that faithfully adheres to the target style?

To address these research questions, we introduce *sNeuron-TST*, a novel framework designed to steer LLMs in performing TST by leveraging style-specific neurons. Initially, we feed both source- and target-style texts into the LLM to identify neurons that exclusively activate in each style based on their activation values. We distinguish neurons active in both styles as overlapping neurons. Notably, eliminating these overlapping neurons during style-specific neuron selection is crucial as their presence can hinder the generation of text in the target style. Our experiments highlight that deactivating neurons specific solely to the source style (excluding those active in both source and target styles) improves style transfer accuracy while impacting sentence fluency. Furthermore, to improve the fluency of generated text, we adapt the state-of-the-art contrastive decoding algorithm (Dola; Chuang et al., 2024) for optimal performance in TST tasks. Our empirical findings (detailed in Section 3.3.2) reveal that layers primarily responsible for style-related outputs are concentrated in the model's latter layers, termed as *style layers*. This indicates that the determination of style-specific words predominantly occurs in these style layers. More precisely, we refine the probability distribution of generated words by comparing logits from these style layers with the final layers, which exert significant influence on style-related outputs.

We conduct a comprehensive evaluation to verify the efficacy of our approach across six benchmarks: formality (Rao and Tetreault, 2018), toxicity (Logacheva et al., 2022), politics (Voigt et al., 2018), politeness (Madaan et al., 2020), authorship (Xu et al., 2012) and sentiment (Shen et al., 2017). Each benchmark contains two distinct styles, resulting in a total of 12 TST directions. Experimental results demonstrate that our method generates a higher proportion of words in the target style compared to baseline systems, achieving superior style transfer accuracy and fluency, while preserving the original meaning of the text.

In summary, we make the following contributions: **(i)** To the best of our knowledge, this is the first work on using style-specific neurons to steer LLMs in performing text style transfer tasks. **(ii)** We emphasize the significance of eliminating overlap between neurons activated by source

and target styles, a methodological innovation with potential applications beyond style transfer. **(iii)** We introduce an enhanced contrastive decoding method inspired by Dola. Our approach not only increases the production of words in the target style but also ensures the fluency of the generated sentences, addressing issues related to direct copying of input text in TST.

2 Related Work

Text Style Transfer. Recently, LLMs have shown promising results in TST through additional fine-tuning (Mukherjee et al., 2024c,b,a; Dementieva et al., 2023), in-context learning (Chen, 2024; Zhang et al., 2024; Pan et al., 2024; Mai et al., 2023) techniques or prompt-based text editing approaches (Luo et al., 2023; Liu et al., 2024). However, these methods often require either extensive computational resources or sensitive prompts, impacting their practicality. In this paper, we focus on a novel decoding approach to guide LLMs for TST using fixed prompts and therefore it does not require significant computational consumption and ensures stable outputs.

Neuron Analysis. Neuron analysis (Xiao et al., 2024) has emerged as a powerful method for elucidating the inner workings of neural network models, offering deeper insights into their behaviors and attracting growing interest in recent years. The common practice is to associate neuron activation with learned knowledge, demonstrating effectiveness in tasks such as knowledge enhancement (Li et al., 2024), sentiment analysis (Tigges et al., 2023) and multilingualism in LLMs (Kojima et al., 2024; Tan et al., 2024). Motivated by the promising outcomes of neuron analysis in enhancing multilingual capabilities of LLMs, this paper posits the presence of style-specific neurons, identifies them, and integrates neuron activation and deactivation seamlessly into the decoding process.

3 Method

Our goal is to identify style-specific neurons to steer LLMs towards generating vocabulary tailored exclusively to a target style, while maintaining fluent text generation in a zero-shot setting. To accomplish this, we first identify style-specific neurons based on their activation values and demonstrate the necessity of eliminating source- and target-style neurons to avoid overlap (Section 3.1). Then, we deactivate neurons associated solely with the
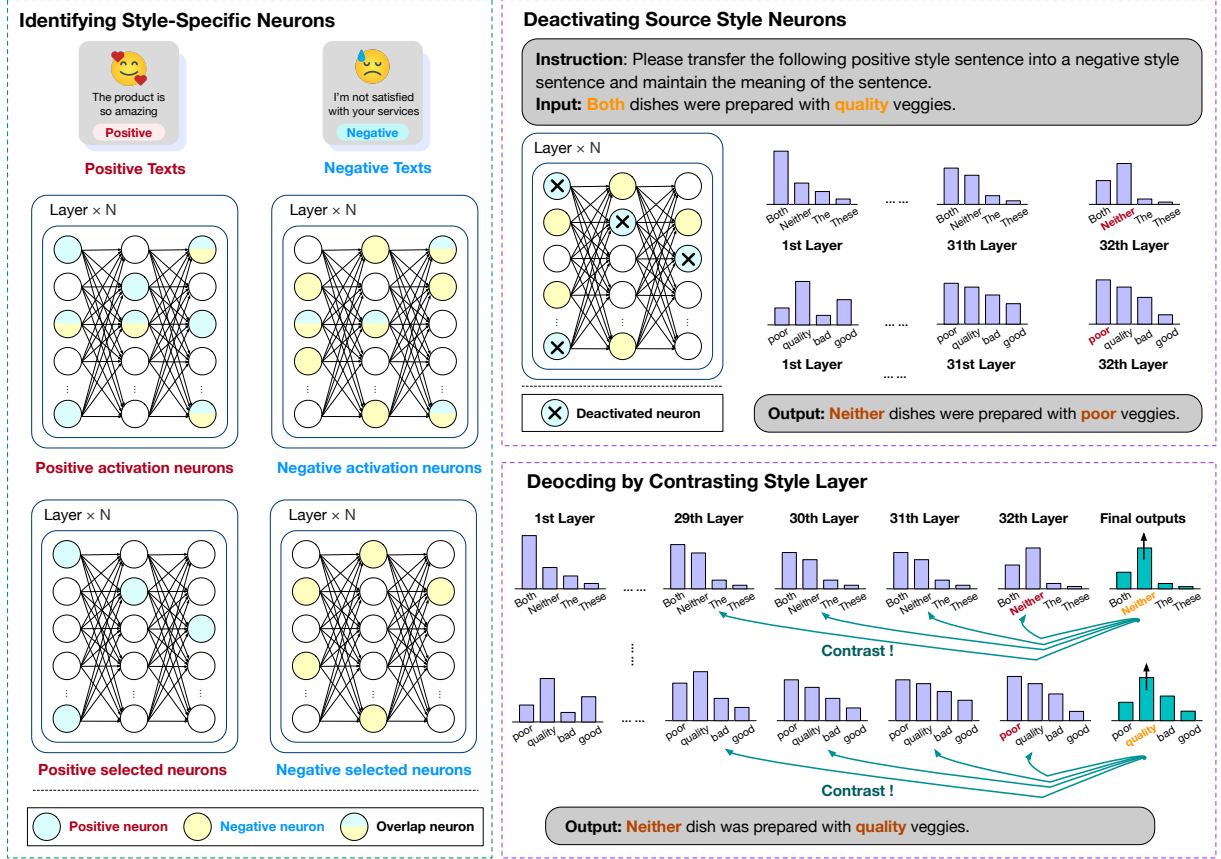
Figure 1: Method overview. The whole framework consists of three parts: identifying style-specific neurons, deactivating source style neurons, and decoding by contrasting style layer. The histogram represents the probability distribution of each word across different layers. When source style neurons are deactivated, LLMs tend to generate all target-style words, such as "Neither" and "poor". By employing contrastive decoding, LLMs take fluency into account and reduce the probability of generating "poor".

source style, observing an increased probability of generating words aligned with the target style, albeit at the expense of fluency (Section 3.2). Finally, we adapt the recent contrastive decoding approach Dola (Chuang et al., 2024) to TST, ensuring the fluency of generated sentences (Section 3.3). Figure 1 illustrates the framework of our approach.

3.1 Identifying Style-Specific Neurons

Neurons are commonly perceived as feature extractors that map neural networks to human-interpretable concepts (Dreyer et al., 2024). However, neurons can exhibit polysemy, where a single neuron may encode multiple features (e.g., formal and informal styles), thereby complicating their interpretability. To selectively modify specific features of LLMs without unintended changes, it becomes imperative to identify and remove unambiguous neurons.

3.1.1 Neurons in LLMs

The dominant architecture of LLMs is the Transformer (Vaswani et al., 2017), characterized by

multiple layers of multi-head self-attention and feed-forward network (FFN) modules. FFNs contain 2/3 of the model's parameters and encode extensive information, which is crucial for multiple tasks (Yang et al., 2024). Moreover, the activation or deactivation of neurons within the FFN can exert significant influence on the model's output (Garde et al., 2023). Inspired by this, we aim to identify neurons in the FFN modules of LLMs that are dedicated to specific styles.

Formally, the activation values of layer j in a network are defined as:

$$a^{(j)} = \text{act\_fn}(W^{(j)}a^{(j-1)} + b^{(j)}) \qquad (1)$$

where $W^{(j)}$ and $b^{(j)}$ are the weights and biases of layer j, while $a^{(j-1)}$ is the activation values of the previous layer and $\text{act\_fn}(\cdot)$ denotes the activation function (e.g., GLU; Shazeer, 2020 used in LLaMA). The ith neuron of the layer is considered to be active when its activation value $a_i^{(j)} > 0$.
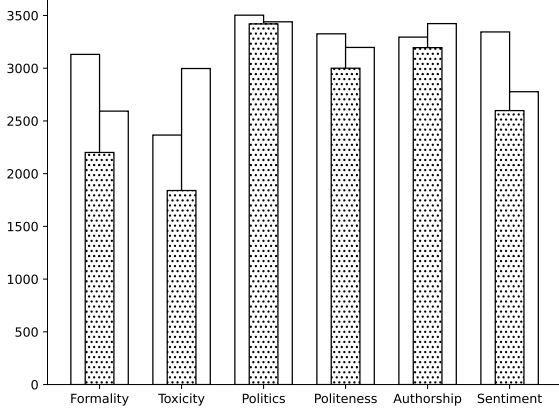
Figure 2: Overlap statistics of style-specific neurons identified using the method of (Tang et al., 2024) on six benchmarks.

3.1.2 Neuron Selection

Recently, Tang et al. (2024) introduced a method for identifying language-specific neurons and demonstrated a significant overlap among neurons across different languages, such as an approximate 25% overlap between Chinese and English neurons. However, their study did not evaluate the performance implications of these overlaps. We measure the overlap of style-specific neurons by applying the method of Tang et al. (2024) directly to a style-specific corpus. As illustrated in Figure 2, we observe a higher overlap among style-specific neurons. For instance, in the Politics benchmark, nearly 95% of neurons overlap between "democratic" and "republican" styles. Moreover, we demonstrate that this substantial overlap negatively impacts the performance of TST (Section 6.1).

To eliminate the overlap between neurons of different styles, we identify style-specific neurons and their intersection. Formally, suppose we have two distinct styles, denoted as A and B. We feed the corpora of the two styles to an LLM separately, to obtain the activation values of the neurons in the FFN layers for both styles, as described in Eq (1). We then select the neurons whose activation value exceeds zero, forming two sets denoted as S_A and S_B, respectively. Subsequently, we sort the activation values within S_A and S_B in descending order and select the neurons with the top k values ($k = 500n, n \in \{1, 2, 3, \ldots, 20\}$ tuned on the validation set), resulting in S'_A and S'_B. Finally, we identify the neurons associated with strictly one of the styles by computing the disjoint sets of the two smaller sets: $N_A = S'_A \setminus S'_B$ and $N_B = S'_B \setminus S'_A$.

		Style Accuracy			
		Formality		**Politeness**	
Source	**Target**	informal	formal	impolite	polite
✗	✗	80.00	11.20	79.50	14.80
✓	✗	**80.53**	**13.63**	**80.06**	**19.37**
✗	✓	76.25	8.51	65.50	9.27
✓	✓	78.42	9.27	73.48	10.36
		Fluency			
		Formality		**Politeness**	
Source	**Target**	informal	formal	impolite	polite
✗	✗	**92.53**	**87.69**	**105.35**	**92.34**
✓	✗	104.17	96.83	127.26	105.12
✗	✓	113.14	106.23	136.10	112.51
✓	✓	108.22	100.79	131.22	108.64

Table 1: Experiments for deactivating neurons on formality and politeness benchmarks. ✓ means the neuron is deactivated, while ✗ means the neuron is activated. "Source" and "Target" denotes the neuron sides. The indicated style (e.g. formal) within a task (e.g. Formality) indicates the source, and its pair is the target style. Style accuracy and fluency are defined in Section 4.4.

3.2 Deactivating Source Style Neurons

After identifying neurons associated with a particular style, a common practice (Tang et al., 2024) is to deactivate these neurons by setting their activation values to zero during the model's forward pass. However, neurons are sensitive components in neural networks; thus, deactivating a neuron associated with a specific feature (e.g., formal style) can lead to significant performance deterioration (Morcos and Barrett, 2018). To investigate the effects of deactivating source- and target-style neurons in TST task, we conduct experiments focusing on formality and politeness transfer tasks.

From Table 1, we observe that: **(1)** Deactivating the source-style neurons while keeping the target-style neurons active improves the accuracy of generating the target style. Conversely, deactivating the target-style neurons, regardless of the state of the source-style neurons, leads to a decrease in the accuracy of generating the target style. This occurs because deactivating the target-style neurons impairs the ability of LLMs to generate target-style words during decoding, resulting in lower accuracy. On the other hand, deactivating the source-style neurons allows LLMs to focus more on generating target-style words, thus improving target style accuracy. This finding aligns with related work on language-specific neuron deactivation (Tang et al., 2024; Zhao et al., 2024). **(2)** Fluency decreases whenever neurons are deactivated, whether they

are source-style or target-style neurons. This is mainly due to the significant impact that deactivating neurons has on the word distribution during decoding. Specifically, the model tends to generate words of the non-deactivated style with a higher probability, leading to generated texts that are simply a concatenation of non-deactivated style words, thereby compromising fluency. As illustrated in Figure 1, after deactivating the source-style neurons, the generated text includes both "Neither" and "quality"— two target-style words without maintaining sentence fluency.

3.3 Contrastive Decoding for TST

Contrastive decoding (CD; Li et al., 2023), which adjusts the probability of predicting the next word by comparing the outputs of a LLM with a weaker, smaller model, has been proven effective in enhancing fluency and coherence. More recently, Chuang et al. (2024) proposed Dola, a CD approach that achieves excellent results by comparing outputs between the final layer and the early layers. We adapt Dola to TST to mitigate the fluency issues observed during neuron deactivation.

3.3.1 Dola

Given a sequence of tokens $\{x_1, x_2, \ldots, x_{t-1}\}$ and the total number (N) of layers in LLMs, the probability of the next token x_t in j-th transformer layer can be computed in advance (known as *early exit*; Schuster et al., 2022) as:

$$p^j(x_t \mid x_{<t}) = \text{softmax}\big(\phi(h_t^{(j)})\big)_{x_t} \quad (2)$$

where h_t is the hidden states obtained from the embedding layer. $\phi(\cdot)$ is the vocabulary head used to predict the probabilities of the tokens.

Dola aims to contrast the information of the final layer and a set of early layers ($\mathcal{J} \subset \{0, \ldots, N-1\}$) to obtain the next-token probability as:

$$\hat{p}(x_t \mid x_{<t}) = \text{softmax}\big(\mathcal{F}\big(p^N(x_t), p^M(x_t)\big)\big)_{x_t} \quad (3)$$

where $\mathcal{F}(\cdot)$ is the function used to contrast between the output distributions from one premature layer M and the final layer by computing the log-domain difference between two distributions (Li et al., 2023) as follows:

$$\mathcal{F}\big(p^N(x_t), p^M(x_t)\big) = \begin{cases} \log \dfrac{p^N(x_t)}{p^M(x_t)}, & \text{if } x_t \in \Phi, \\ -\infty, & \text{otherwise.} \end{cases} \quad (4)$$

where Φ is defined as whether or not the token has high enough output probabilities from the mature layer as:

$$\Phi(x_t \mid x_{<t}) = \Big\{ p^N(x_t) \geq \max_w p^N(w) \Big\} \quad (5)$$

Layer M, the *premature layer*, is selected dynamically at each time step by taking the layer with the largest Jensen-Shannon Divergence (JSD; Menéndez et al., 1997) to contrast output distributions from the final and the set of early candidate layers.

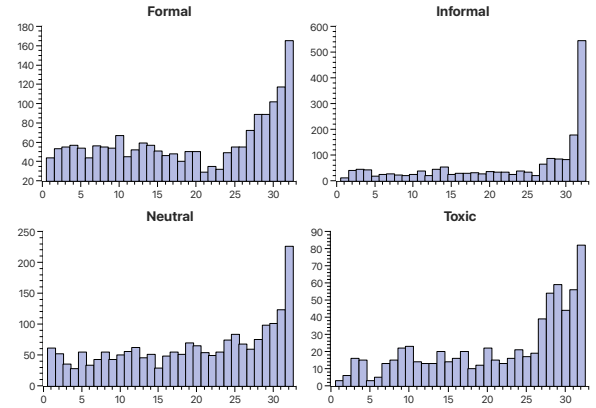#### 3.3.2 Our adaptation to TST

Figure 3: Statistics of the number of style-specific neurons in each layer in LLaMA-3 on formality and toxicity benchmarks.

Candidate layer selection. To better adapt Dola to TST, we select candidate layers for comparison based on the criterion that these layers should contain more style information. To this end, we measure the amount of style-specific neurons across each layer. As shown in Figure 3, the last few layers, particularly the final layer, contain significantly more style neurons compared to the earlier layers. Therefore, we select the last few layers (4 in our experiments) as our candidate layers.

Next-token prediction. After deactivating the source-style neurons, LLMs tend to generate target-style tokens. However, we need to determine whether the appearance of these target-style tokens is due to their consistently high probability from the early layers to the final layer or due to a probability shift caused by neuron deactivation in the last few layers. If the probability of tokens at a given time step remains consistent from the first layer to the final layer, it indicates that these tokens are style-independent (typically function words) and are retained in the output of the final layer by Eq. (3). Conversely, if these words have a low

probability in the early layers (typically target-style words) and only exhibit a probability "mutation" in the last few layers due to the deactivation of source-style neurons, we then select the layer with the maximum JSD distance from the candidate layers as our premature layer M and adjust their probability distribution according to Eq. (3).

4 Experiments

4.1 Datasets

We evaluate our approach on six typical TST tasks: formality, toxicity, politics, politeness, authorship, and sentiment on GYAFC (Rao and Tetreault, 2018), ParaDetox (Logacheva et al., 2022), Politeness (Madaan et al., 2020), Shakespeare (Xu et al., 2012) and Yelp (Shen et al., 2017). The statistics of the datasets can be found in Appendix A.

4.2 Baselines

We compare our approach with the following baselines: (1) **LLaMA-3:** We use LLaMA-3 (Meta, 2024) without additional fine-tuning as the vanilla baseline system. (2) **APE:** Using activation probability entropy to identify the style specific neurons (Tang et al., 2024). (3) **AVF:** Using activation value frequency and set a threshold to identify the style neurons (Tan et al., 2024). (4) **PNMA:** Finding neurons that activate on the source style sentences but do not activate on target style sentences (Kojima et al., 2024). Note that (2), (3), and (4) from the original paper focus on identifying language-specific neurons to enhance the multilingual capabilities of LLMs, and we extend these methodologies to our style-related corpus. For (4), it requires the use of parallel data from both source and target texts to identify neurons, whereas (2), (3), and our method does not require the use of parallel data. Additionally, after identifying the neurons, we deactivate the source-style neurons in (2), (3), and (4). For a detailed comparison of various decoding strategies, please refer to Appendix G.

4.3 Implementation

We use the 8B model of LLaMA-3, available in the HuggingFace repository[2] in zero-shot setting. To further assess the scalability of our method, we also employ the 70B LLaMA-3 model (Appendix D). For each baseline system, we use the same hyper-parameters (e.g., threshold) as the original paper.

---

[2] https://github.com/huggingface/transformers

4.4 Evaluation Metric

We evaluate our approach using three metrics commonly employed in TST tasks. **Style Accuracy.** Accuracy of labels predicted as correct by a style classifier. Please refer to Appendix B for more details. **Content Preservation.** Cosine similarity between the embeddings of the original text and the text generated by the model, using LaBSE (Feng et al., 2022) to obtain sentence embeddings as our primary metric. Additionally, we employ BLEURT metrics (Sellam et al., 2020) for comparison, as recent studies indicate strong correlations between BLEURT assessments on TST and human evaluation results (Appendix F). **Fluency.** Perplexity of the generated sentences using GPT-2 (Radford et al., 2019).

5 Results

Table 2 shows the transfer performance (style accuracy, content preservation and fluency) of the six benchmarks in 12 directions.

Overall Performance. While the *APE*, *AVF*, and *PNMA* demonstrate strong performance in enhancing multilingual capabilities, they do not outperform the original LLaMA-3 model in the TST task, with the exception of the content preservation metric. This disparity arises primarily because language-specific properties can be identified using straightforward features, such as script differences. Consequently, the neuron selection methods of these baselines, despite their partial overlaps, have minimal impact on multilingual performance. However, text style represents a more complex attribute, requiring models to learn extensive knowledge and execute nuanced judgments at both the word and semantic levels. The overlap of neurons in baseline systems across source and target styles adversely affects the results, particularly in style accuracy. Furthermore, the baseline methods lack a contrastive decoding strategy, which compromises their fluency. Our method outperforms the baseline methods in terms of both accuracy and fluency, highliting the importance of eliminating overlapping style neurons and employing contrastive decoding.

Content Preservation. Interestingly, we observe that the original LLaMA-3 and other baseline systems exhibit strong performance in content preservation, which appears inconsistent with conclusions drawn from the other two metrics. Upon closer examination, we find that this content preser-

Style Transfer Accuracy

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	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

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	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

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	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 com-

promise 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
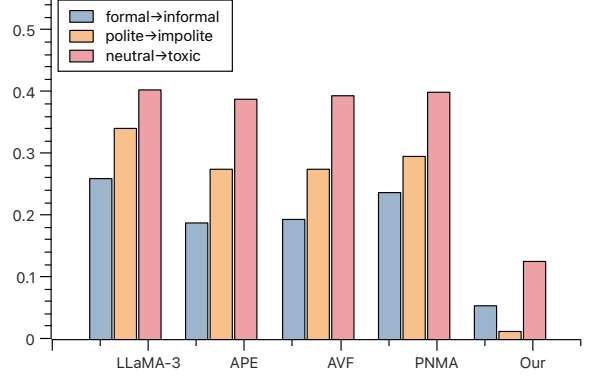

Figure 4: Copy Ratio on three selected TST tasks. Lower value indicates better performance of the model.

of sentences.

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

References

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Jianlin Chen. 2024. Lmstyle benchmark: Evaluating text style transfer for chatbots. *arXiv preprint arXiv:2403.08943*.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2023. On the relation between sensitivity and accuracy in in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 155–167, Singapore. Association for Computational Linguistics.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring methods for cross-lingual text style transfer: The case of text detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1083–1101, Nusa Dua, Bali. Association for Computational Linguistics.

Maximilian Dreyer, Erblina Purelku, Johanna Vielhaben, Wojciech Samek, and Sebastian Lapuschkin. 2024. Pure: Turning polysemantic neurons into pure features by identifying relevant circuits. *arXiv preprint arXiv:2404.06453*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Albert Garde, Esben Kran, and Fazl Barez. 2023. Deepdecipher: Accessing and investigating neuron activation in large language models. *arXiv preprint arXiv:2310.01870*.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter*, 24(1):14–45.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *arXiv preprint arXiv:2404.02431*.

Wen Lai, Alexandra Chronopoulou, and Alexander Fraser. 2022a. m^4 adapter: Multilingual multi-domain adaptation for machine translation with a meta-adapter. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4282–4296, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wen Lai, Alexandra Chronopoulou, and Alexander Fraser. 2023a. Mitigating data imbalance and representation degeneration in multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14279–14294, Singapore. Association for Computational Linguistics.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2023b. Extending multilingual machine translation through imitation learning. *arXiv preprint arXiv:2311.08538*.

Wen Lai, Jindřich Libovický, and Alexander Fraser. 2022b. Improving both domain robustness and domain adaptability in machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5191–5204, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Qingyi Liu, Jinghui Qin, Wenxuan Ye, Hao Mou, Yuxuan He, and Keze Wang. 2024. Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18689–18697.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

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

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.

Huiyu Mai, Wenhao Jiang, and Zhihong Deng. 2023. Prefix-tuning based unsupervised text style transfer. *arXiv preprint arXiv:2310.14599*.

ML Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Ari Morcos and David Barrett. 2018. Understanding deep learning through neuron deletion. Google Deepmind Blog. Https://deepmind.google/discover/blog/understanding-deep-learning-through-neuron-deletion.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr Ojha, John P McCrae, and Ondřej Dušek. 2024a. Text detoxification as style transfer in english and hindi. *arXiv preprint arXiv:2402.07767*.

Sourabrata Mukherjee, Atul Kr Ojha, Akanksha Bansal, Deepak Alok, John P McCrae, and Ondřej Dušek. 2024b. Multilingual text style transfer: Datasets & models for indian languages. *arXiv preprint arXiv:2405.20805*.

Sourabrata Mukherjee, Atul Kr. Ojha, and Ondřej Dušek. 2024c. Are large language models actually good at text style transfer? *arXiv preprint arXiv:2406.05885*.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? *arXiv preprint arXiv:2405.02421*.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2023. Text style transfer evaluation using large language models. *arXiv preprint arXiv:2308.13577*.

Lei Pan, Yunshi Lan, Yang Li, and Weining Qian. 2024. Unsupervised text style transfer via llms and attention masking with multi-way interactions. *arXiv preprint arXiv:2402.13647*.

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

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.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Shaomu Tan, Di Wu, and Christof Monz. 2024. Neuron specialization: Leveraging intrinsic task modularity for multilingual machine translation. *arXiv preprint arXiv:2404.11201*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hanyu Wang, Pengxiang Wu, Kevin Dela Rosa, Chen Wang, and Abhinav Shrivastava. 2024. Multimodality-guided image style transfer using cross-modal gan inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4976–4985.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiongye Xiao, Chenyu Zhou, Heng Ping, Defu Cao, Yaxing Li, Yizhuo Zhou, Shixuan Li, and Paul Bogdan. 2024. Exploring neuron interactions and emergence in llms: From the multifractal analysis perspective. *arXiv preprint arXiv:2402.09099*.

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

Shangshang Yang, Xiaoshan Yu, Ye Tian, Xueming Yan, Haiping Ma, and Xingyi Zhang. 2024. Evolutionary neural architecture search for transformer in knowledge tracing. *Advances in Neural Information Processing Systems*, 36.

Chiyu Zhang, Honglong Cai, Yuexin Wu, Le Hou, Muhammad Abdul-Mageed, et al. 2024. Distilling text style transfer with self-explanation from llms. *arXiv preprint arXiv:2403.01106*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

A Datasets

All style data used for neuron identification are obtained from publicly available datasets. We applied the following preprocessing to the raw data: (1) removing sentences longer than 120 characters; (2) eliminating duplicate sentences; and (3) removing sentences containing a large number of special symbols. Table 7 provides detailed statistics of the preprocessed corpus.

B Classifiers used in Each Benchmark

To evaluate the accuracy of style transfer, we use open-source classifiers on the six benchmarks we evaluated. The sources of these classifiers are detailed in Table 8.

C JSD Distance between Layers

To verify whether the style-specific layers selected in Section 3.3 encode stylistic information, we calculate the Jensen-Shannon Divergence (JSD) distances between the final layer and all previous layers for the TST task of transfer from informal style text to formal style text. The results, shown in Table 9, led to the following findings: (1) For most of the early layers, from layer 0 to 26, the distances between the final layer and these layers remain almost constant or change very little, indicating that the information encoded in these layers is very similar. However, for the last few layers, from layer 27 to 31, the JSD distance from the final layer is smaller compared to the earlier layers, but the distance between different layers increases. This suggests that the last few layers are processing style-related information, consistent with the distribution characteristics of the style layers discussed in Section 3.3. (2) Some words associated with the formal style (target-side style), highlighted in bold in the Table 9, show a larger distance difference in the last few layers. This aligns with our expectation that words representing the target style are more likely to be activated in the style layer, increasing their probability of being selected as candidates for token generation in the style layer.

D Effectiveness of different model sizes

To verify the effectiveness of our method on a larger model, we conduct experiments using the 70B version of LLaMA-3. The results, presented in Table 10, indicate that our method is also effective on the larger model and consistent with the con-

	Style	Dola	Our
Formality	informal→formal	78.14	80.80
	formal→informal	12.63	14.40
Toxicity	toxic→neutral	49.25	55.36
	neutral→toxic	25.41	31.98
Politics	democratic→republican	36.26	37.81
	republican→democratic	46.25	50.30
Politeness	impolite→polite	76.58	80.63
	polite→impolite	20.57	23.27
Authorship	shakespeare→modern	65.87	73.40
	modern→shakespeare	42.43	45.14
Sentiment	positive→negative	73.12	77.93
	negative→positive	50.28	54.73

Table 6: Comparison of different layer selection strategies between Dola and our approach.

clusions drawn from the 7B model (See Table 2 in Section 5 for more details).

E Style Layers vs. Dola Layers

In Section 3.3, our method selects the style layers, specifically the last few layers of the LLMs, to decoding from contrasting against the final layer. In contrast, Dola selects the early layers to decode by contrasting the final layer. To verify the superiority of our selected style layers, we conduct a comparison experiment, the results of which are shown in Table 6. We can clearly observe the superiority of selecting the last few layers for contrastive decoding in the TST task.

F Different content preservation metrics

In Table 2, we find that our method is not optimal in content preservation. To verify whether this phenomenon occurs with other content preservation metric, we conduct a comparison experiment and present the results in Table 11. We observe the same conclusion as in Table 2, namely, our method is inferior to the baseline method in terms of meaning preservation. For a detailed analysis, please refer to Section 5.

G Different decoding strategy

In Section 3.3, we present a decoding strategy for contrasting style layers. To verify the advantages of this decoding strategy, we compare it with two additional decoding methods: nucleus sampling and contrastive search. As shown in Table 12, our decoding method outperforms the others. This is

primarily because contrastive search focuses on the isotropy of token representations during decoding, which means that the semantically similar words have less variation in the representation space and their probabilities should be increased. However, this does not align with the goal of the TST task, which aims to expose more target-style words. Source-style words and target-side style words are actually similar in representation. For example, in the emotion task, "like" and "hate" are semantically different but similar in the embedding space because both represent an emotion, making it difficult to distinguish between these words using isotropy at the representation level.

In addition, nucleus sampling (NP) is a decoding method by setting a threshold p and then restricting the sampling to the set of most probable tokens with cumulative probability less than p. NP is not suited for TST because after deactivating the style neurons at the source side, the probability distribution of the words is changed. The probability of all words in the target style becomes higher, resulting in candidate words predominantly being in the target style. This can cause issues with fluency, as words in the target style are not always meant to be revealed in every context.

Table 7: Data statistics on six benchmarks containing the size of train/valid/test set and transfer task we evaluated.

Benchmark	Dataset	Tasks	Size		
			train	vald	test
Politeness	Politness (Madaan et al., 2020)	impolite ↔ polite	100k	2000	2000
Toxicity	ParaDetox (Logacheva et al., 2022)	toxic ↔ neutral	18k	2000	2000
Formality	GYAFC (Rao and Tetreault, 2018)	informal ↔ formal	52k	500	500
Authorship	Shakespeare (Xu et al., 2012)	shakespeare ↔ modern	27k	500	500
Politics	Political (Voigt et al., 2018)	democratic ↔ republican	100k	1000	1000
Sentiment	Yelp (Shen et al., 2017)	positive ↔ negative	100k	1000	1000

Table 7: Data statistics on six benchmarks containing the size of train/valid/test set and transfer task we evaluated.

Benchmark	Source
Politeness	`https://huggingface.co/Genius1237/xlm-roberta-large-tydip`
Toxicity	`https://huggingface.co/s-nlp/roberta_toxicity_classifier`
Formality	`https://huggingface.co/s-nlp/xlmr_formality_classifier`
Authorship	`https://huggingface.co/notaphoenix/shakespeare_classifier_model`
Politics	`https://huggingface.co/m-newhauser/distilbert-political-tweets`
Sentiment	`https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english`

Table 8: Classifiers used to evaluate the accuracy of style transfer.

Instruction: Please transfer the following informal style sentence into a formal style sentence and maintain the meaning of the sentence.
Input: the movie The In-Laws not exactly a holiday movie but funny and good!
Output: **The** movie "The In-Laws" **is** not exactly a holiday movie, but **it is** funny and good!

	The	movie	"	The	In	-L	aws	"	is	not	exactly	a	holiday	movie	,	but	it	is	funny	and	good	!
0	2.35	2.93	3.18	3.70	3.80	3.53	5.44	3.76	3.59	3.92	3.70	4.02	4.02	4.29	3.86	4.03	3.95	4.28	3.89	4.37	4.12	3.04
1	2.35	2.93	3.18	3.68	3.79	3.52	5.42	3.74	3.58	3.91	3.70	4.02	4.01	4.28	3.86	4.02	3.93	4.27	3.87	4.36	4.11	3.04
2	2.35	2.93	3.18	3.68	3.79	3.52	5.43	3.74	3.58	3.91	3.70	4.02	4.01	4.28	3.86	4.01	3.92	4.26	3.87	4.36	4.10	3.04
3	2.35	2.93	3.18	3.68	3.79	3.52	5.43	3.74	3.58	3.91	3.68	4.02	4.01	4.27	3.87	4.01	3.92	4.26	3.87	4.36	4.10	3.04
4	2.35	2.93	3.18	3.68	3.79	3.52	5.44	3.74	3.59	3.92	3.68	4.01	4.02	4.27	3.87	4.01	3.92	4.26	3.87	4.36	4.10	3.04
5	2.35	2.93	3.18	3.67	3.79	3.52	5.44	3.75	3.59	3.90	3.68	4.02	4.02	4.28	3.87	3.99	3.91	4.26	3.87	4.37	4.10	3.05
6	2.35	2.92	3.18	3.67	3.79	3.52	5.42	3.74	3.60	3.91	3.68	4.01	4.01	4.27	3.87	3.99	3.91	4.26	3.88	4.35	4.10	3.04
7	2.35	2.92	3.18	3.67	3.78	3.50	5.43	3.75	3.58	3.90	3.67	3.99	4.01	4.28	3.87	3.99	3.90	4.24	3.87	4.33	4.10	3.04
8	2.36	2.92	3.18	3.67	3.79	3.50	5.42	3.76	3.58	3.90	3.67	3.98	4.01	4.27	3.86	3.99	3.91	4.24	3.89	4.35	4.11	3.06
9	2.36	2.92	3.18	3.67	3.77	3.50	5.42	3.75	3.58	3.90	3.68	3.99	4.01	4.26	3.86	4.01	3.91	4.24	3.87	4.33	4.11	3.05
10	2.35	2.91	3.17	3.66	3.77	3.49	5.45	3.74	3.58	3.90	3.68	3.98	4.01	4.26	3.86	3.99	3.90	4.23	3.89	4.34	4.11	3.05
11	2.35	2.91	3.16	3.65	3.76	3.48	5.44	3.75	3.58	3.89	3.68	3.97	4.00	4.27	3.86	3.99	3.89	4.23	3.89	4.35	4.12	3.05
12	2.36	2.93	3.16	3.65	3.76	3.49	5.44	3.74	3.58	3.90	3.67	3.97	4.00	4.26	3.85	3.99	3.89	4.22	3.89	4.36	4.12	3.05
13	2.35	2.93	3.17	3.66	3.76	3.50	5.44	3.73	3.58	3.89	3.68	3.97	4.02	4.27	3.89	3.98	3.89	4.23	3.91	4.35	4.12	3.06
14	2.34	2.91	3.15	3.64	3.76	3.50	5.46	3.71	3.58	3.87	3.67	3.98	4.01	4.27	3.86	3.96	3.87	4.21	3.90	4.33	4.10	3.05
15	2.34	2.90	3.14	3.62	3.78	3.50	5.44	3.71	3.55	3.66	3.66	3.97	4.01	4.26	3.84	3.93	3.87	4.20	3.91	4.31	4.11	3.04
16	2.34	2.87	3.11	3.62	3.74	3.49	5.39	3.72	3.52	3.81	3.61	3.92	3.96	4.23	3.81	3.87	3.84	4.18	3.87	4.24	4.05	3.03
17	2.32	2.87	3.08	3.61	3.71	3.46	5.39	3.71	3.53	3.79	3.60	3.89	3.93	4.21	3.79	3.74	3.81	4.17	3.85	4.22	4.02	2.99
18	2.31	2.84	3.02	3.56	3.64	3.45	5.39	3.67	3.46	3.68	3.54	3.83	3.86	4.16	3.75	3.71	3.77	4.12	3.80	4.13	3.96	2.97
19	2.30	2.80	3.00	3.53	3.61	3.42	5.38	3.62	3.41	3.62	3.47	3.78	3.84	4.13	3.71	3.67	3.68	4.07	3.77	4.09	3.92	2.93
20	2.26	2.77	2.96	3.50	3.55	3.39	5.36	3.60	3.37	3.63	3.43	3.73	3.80	4.09	3.68	3.58	3.62	4.01	3.76	4.04	3.89	2.91
21	2.23	2.74	2.92	3.46	3.50	3.39	5.33	3.60	3.30	3.50	3.39	3.65	3.78	4.04	3.62	3.44	3.58	3.95	3.72	3.93	3.84	2.87
22	2.19	2.68	2.87	3.40	3.45	3.35	5.31	3.49	3.25	3.36	3.25	3.58	3.50	3.96	3.56	3.35	3.40	3.86	3.62	3.87	3.74	2.84
23	2.14	2.57	2.80	3.33	3.35	3.33	5.27	3.44	3.15	3.28	3.11	3.47	3.34	3.88	3.49	3.25	3.28	3.73	3.54	3.77	3.61	2.81
24	2.10	2.43	2.73	3.27	3.25	3.30	5.26	3.39	3.06	3.14	2.96	3.36	3.22	3.72	3.42	3.08	3.14	3.61	3.36	3.71	3.53	2.75
25	2.07	2.37	2.60	3.22	3.16	3.25	5.24	3.33	2.96	3.02	2.77	3.22	2.71	3.65	3.34	3.03	3.00	3.54	3.20	3.60	3.38	2.70
26	2.06	2.29	2.56	3.18	3.14	3.17	5.19	3.31	2.88	2.93	2.65	3.14	2.59	3.41	3.30	2.91	2.93	3.45	3.09	3.52	3.28	2.66
27	1.98	2.15	2.46	3.13	3.09	3.15	5.16	3.20	2.72	2.80	2.57	2.98	2.46	3.24	3.11	2.77	2.80	3.26	2.96	3.39	3.17	2.56
28	1.94	2.07	2.36	3.09	2.96	3.05	5.17	3.08	2.53	2.72	2.60	2.84	2.68	3.15	2.98	2.50	2.69	3.07	2.87	3.22	3.04	2.46
29	1.85	1.95	2.13	2.86	2.81	2.79	5.09	2.91	2.30	2.54	2.32	2.52	2.49	3.05	2.72	2.25	2.48	2.84	2.78	2.92	2.84	2.26
30	1.84	1.93	1.99	2.80	2.52	2.41	4.87	2.87	2.19	2.34	2.17	2.32	2.35	3.04	2.55	2.09	2.31	2.74	2.62	2.77	2.74	2.21
31	1.57	1.69	1.62	2.30	2.24	2.23	4.51	2.31	1.94	2.10	2.05	2.19	2.27	2.74	2.08	1.92	2.05	2.46	2.30	2.10	2.56	1.86

Table 9: JSD (scaled by 10^5) between the final layer and all previous layer in LLaMA-3. Each row represents the distance between all previous layers and the final layer, while each column corresponds to the token generated at each decoding step. Example taken from the TST task to transfer from informal style to formal style. The 0-th layer is the embedding layer.

Style Transfer Accuracy

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	81.45	11.73	48.57	30.57	35.97	49.43	81.21	15.05	64.55	44.67	77.06	53.08
APE	75.22	13.39	49.43	28.62	42.83	46.40	78.32	19.64	56.07	45.22	79.20	48.64
AVF	76.44	13.61	48.25	28.65	39.76	45.00	79.50	18.77	57.20	44.51	80.27	49.27
PNMA	74.10	10.60	43.87	24.51	35.60	38.24	74.23	15.19	55.29	38.30	75.43	42.95
Our	**83.83**	**16.27**	**57.28**	**33.09**	**43.69**	**51.26**	**82.16**	**24.94**	**74.91**	**46.40**	**82.39**	**55.43**

Content Preservation

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	**86.29**	76.54	73.85	**84.59**	83.01	77.49	76.03	**90.99**	79.89	**64.10**	76.63	74.89
APE	77.26	85.28	**78.18**	83.33	**89.48**	**83.52**	77.28	88.09	**81.72**	59.37	76.62	74.06
AVF	77.08	**85.73**	77.85	84.59	88.12	81.00	77.10	88.93	80.99	59.54	**78.05**	74.41
PNMA	77.01	85.12	76.27	83.67	87.77	82.13	76.98	88.06	79.52	57.90	75.28	72.91
Our	85.43	85.51	77.59	80.63	84.29	75.48	77.05	83.55	78.38	61.82	75.60	**75.79**

Fluency

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	87.93	86.58	112.55	190.29	86.41	63.86	101.40	90.29	196.47	135.93	172.68	122.16
APE	93.56	89.34	128.33	187.91	87.70	65.33	105.00	91.75	246.71	131.55	146.58	123.43
AVF	96.03	85.80	128.28	188.03	84.33	71.31	111.61	95.92	219.96	122.60	151.32	126.41
PNMA	103.47	90.65	131.33	190.10	91.86	77.24	108.24	99.59	256.96	132.28	154.48	126.53
Our	**87.16**	**76.93**	**80.75**	**171.37**	**81.08**	**62.28**	**100.91**	**81.45**	**146.72**	**113.43**	**140.09**	**107.88**

Table 10: **Main Results (70B model):** Style transfer accuracy (higher values are better; ↑), content preservation (↑) and fluency (↓) on 6 datasets across 12 transfer directions. Best results are highlighted in bold.

Content Preservation (Paraphrase Model)

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	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

Content Preservation (LaBSE model)

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	0.75	**0.90**	0.77	0.86	0.86	0.79	0.72	**0.92**	0.86	0.70	0.76	0.76
APE	0.75	0.88	0.78	0.87	**0.90**	0.85	0.74	0.90	0.86	0.66	0.75	0.75
AVF	0.74	0.88	0.78	0.87	0.89	0.83	0.75	0.90	0.87	0.67	0.75	0.76
PNMA	**0.79**	0.90	**0.79**	**0.89**	0.89	**0.87**	**0.86**	0.90	**0.89**	**0.74**	**0.82**	**0.81**
Our	0.74	0.89	0.74	0.81	0.84	0.75	0.69	0.90	0.84	0.54	0.75	0.64

Content Preservation (BLEURT)

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
LLaMA-3	**0.089**	**0.527**	0.132	0.345	0.304	0.039	**0.625**	0.040	0.307	0.320	0.136	0.084
APE	0.069	0.449	0.156	**0.376**	**0.488**	**0.262**	0.094	**0.535**	0.328	**0.461**	0.193	0.078
AVF	0.043	0.440	0.157	0.376	0.424	0.191	0.122	0.522	**0.344**	0.426	**0.207**	0.095
PNMA	0.002	0.433	0.139	0.360	0.399	0.181	0.074	0.513	0.334	0.417	0.197	0.085
Our	0.073	0.478	**0.157**	0.329	0.460	0.232	0.557	0.473	0.324	0.386	0.199	**0.133**

Table 11: Different content preservation metrics: sentence embedding model trained from paraphrase datasets, sentence embedding model from multilingual representation model and BLEURT metrics.

Style Transfer Accuracy

	Formality		Toxicity		Politics		Politeness		Authorship		Sentiment	
	informal	formal	toxic	neutral	democratic	republican	impolite	polite	shakespeare	modern	positive	negative
	→	←	→	←	→	←	→	←	→	←	→	←
NS ($p = 0.95$)	79.36	11.95	50.55	27.84	36.17	50.19	76.10	21.70	72.79	43.89	74.34	50.96
CS	79.46	12.40	54.05	30.12	36.04	48.39	79.61	22.62	72.18	43.94	76.54	52.82
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**

Table 12: Comparison of three different decoding methods: nucleus sampling (NP; p=0.95), contrastive search (CS) and our decoding method.