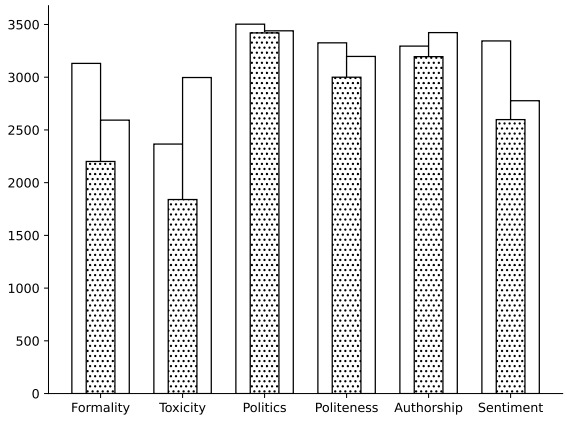

Figure 2: Overlap statistics of style-specific neurons identified using the method of (Tang et al., 2024) on six benchmarks.

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 $N_B = S'_B \setminus S'_A$.

Source	Target	Formality		Politeness	
		informal	formal	impolite	polite
Style Accuracy					
✗	✗	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					
✗	✗	**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:

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

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:

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

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

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

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

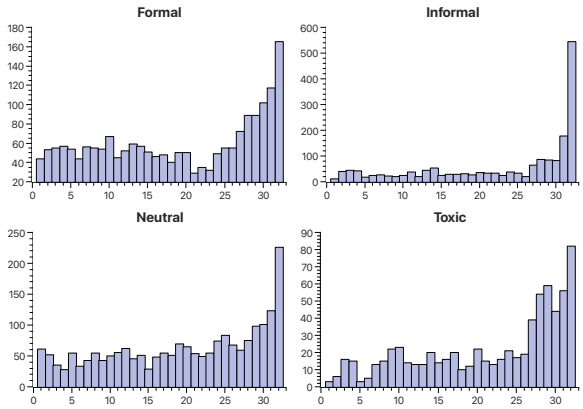

Figure 3: Statistics of the number of style-specific neurons in each layer in LLaMA-3 on formality and toxicity benchmarks.

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 hyperparameters (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-