Style-Specific Neurons for Steering LLMs in Text Style Transfer

Wen Lai[1,2], Viktor Hangya[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

12th November, 2024

Background (I)

- Text style transfer (TST) 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.

Style Features	Example	
Formality	Formal:	Would you like to get a drink?
	Informal:	Wanna get a drink?
Politeness	Polite:	I see your point, but I have a different perspective.
	Impolite:	You're wrong.
Sentiment	Positive:	I'd enjoy something more exciting.
	Negative:	This is boring.

Background (II)

- LLMs tends to directly copy a significant portion of the input text to the output without effectively changing its style.

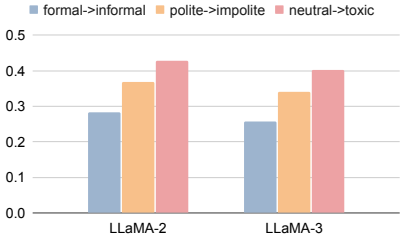


Fig. Copy ratio for LLMs (e.g., LLaMA-2 and LLaMA-3) on TST task.

Enhancing the generation of words that align with the target style during the decoding process remains a significant challenge in TST.

Background (III)

- Neuron analysis aims to identify and understand the roles of individual neurons within a neural network.

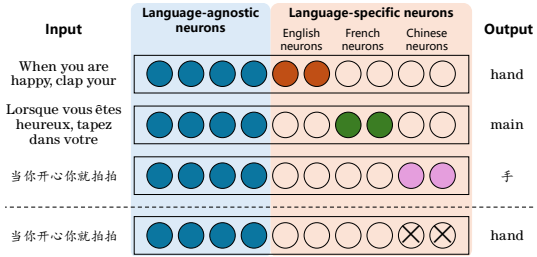


Figure directly taken from Tang et al., 2024

- LLMs exist language-specific neurons that can process the vocabulary, grammar, and idioms of a particular language.
- Deactivating the language-specific neurons leads to a remarkable degradation in the model's understanding and generation abilities.

Research Question

- **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?

Text Style Transfer

- Traditional TST task: parallel data, non-parallel data (Jin et al., 2022)
- Recent work on using LLMs
  - Fine-tnuning (Mukherjee et al., 2024; Dementieva et al., 2023)
  - In-context Learning (Chen et al., 2024; Zhang et al., 2024; Pan et al., 2024)
  - Prompt-based text editting (Liu et al., 2024; Luo et al., 2023)

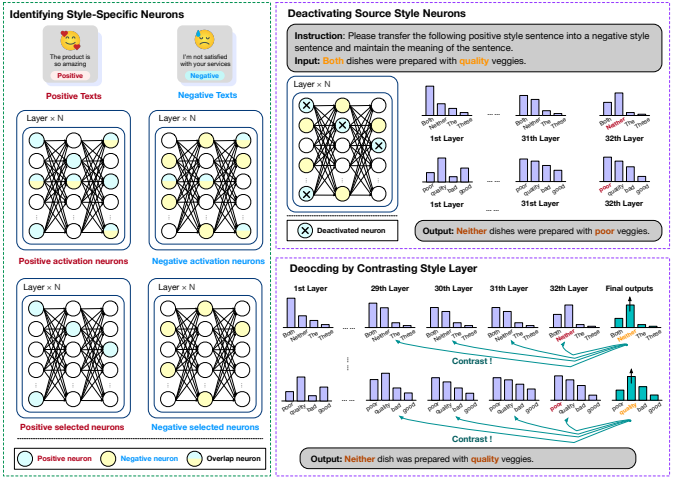
In this paper, we focus on a novel **decoding approach** to guide LLMs for TST using fixed prompts and therefore it does not require significant computational consumption and ensures stable outputs.

Neuron Analysis

- Multilinguality analysis (Kojima et al., 2024; Tan et al., 2024; Tang et al., 2024)
- Knowledge enhancement (Li et al., 2023)
- Sentiment analysis (Tigges et al., 2023)

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.

Overview



- Identify Style-Specific Neurons
  - Neuron Selection (remove overlap)
- Deactivating Source Style Neurons
  - Deactivate neurons from source side or target side
- Contrastive Decoding for TST
  - Adapt Dola (Chuang et al., 2024) to TST to mitigate the fluency issues observed during neuron deactivation

Identify Style-Specific Neurons

- Neurons in LLms

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

  - 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 act_fn$(\cdot)$ denotes the activation function (e.g., GLU used in LLaMA).
  - The i^{th} neuron of the layer is considered to be active when its activation value $a_i^{(j)} > 0$.

Neuron Selection

- Overlap Analysis

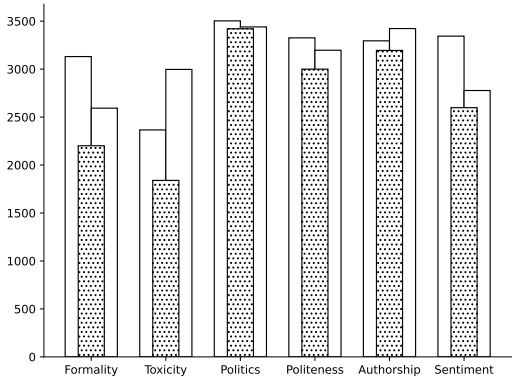


Fig. Overlap statistics of neuron using (Tang et al., 2024)

- Higher overlap among style-specific neurons when using traditional neuron selection method (e.g., 95% overlap in politics style).
- The overlap negatively impacts the performance of TST (details in ablation study).
- We **remove the intersection** between source style neuron and target style neuron.

Deactivating Source Style Neurons

- Deactivate the neurons by setting the activation values to zero during the model's forward pass.
- But, which side should we deactivate? Source or Target?

Style Accuracy					
Source	Target	Formality		Politeness	
		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					
Source	Target	Formality		Politeness	
		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

Tab. Experiments for deactivating neurons on formality and politeness benchmarks.

- Deactivating the source-style neurons while keeping the targetstyle neurons active improves the accuracy of generating the target style.
- Fluency decreases whenever neurons are deactivated, whether they are source-style or target-style neurons.

Contrastive Decoding for TST

- Overview of Dola (Chuang et al., 2024)

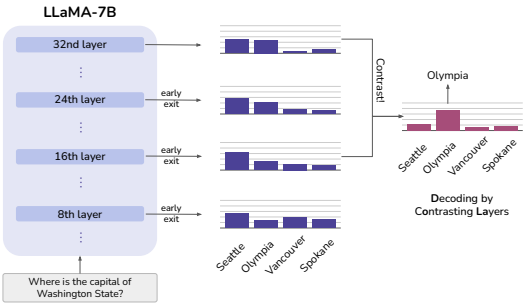


Fig. Figure directly taken from Chuang et al., 2024

- When predicting factual information, LLaMA tends to change the predictions in the higher layers. Otherwise, predictions usually have been decided by early layers.
- Three Steps:
  - Early exiting from all layers;
  - Pick a layer as "premature" layer, final layer as "mature" layer;
  - Subtract "premature" logits from "mature" logits in log domain.

Contrastive Decoding for TST

- Candidate layer selection

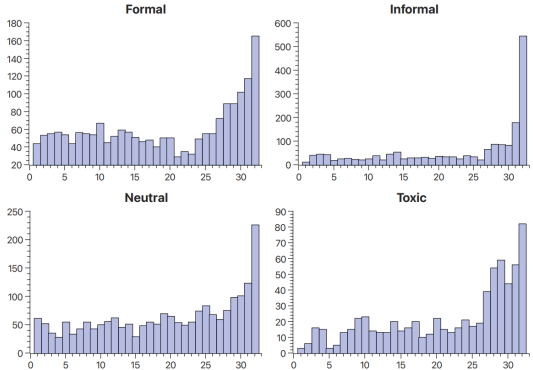


Fig. Statistics of the number of style-specific neurons in each layer

- The last few layers, particularly the final layer, contain significantly more style neurons compared to the earlier layers.
- We select the last few layers (4 in our experiments) as our candidate lay.

Contrastive Decoding for TST

- Next-token prediction
  - Select the layer with the maximum JSD distance from the candidate layers as our premature layer M and adjust their probability distribution.

Datasets

- We evaluate our approach on 6 TST benchmarks

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

Baselines

- We compare our approach with the following baselines:
  - **LLaMA-3:** We use LLaMA-3 (Meta 2024) without additional fine-tuning as the vanilla baseline system.
  - **APE:** Using activation probability entropy to identify the style specific neurons (Tang et al., 2024).
  - **AVF:** Using activation value frequency and set a threshold to identify the style neurons (Tan et al., 2024).
  - **PNMA:** Finding neurons that activate on the source style sentences but do not activate on target style sentences (Kojima et al., 2024).

Evaluation Metrics

- We evaluate our approach using three common metrics:
  - **Style Accuracy:** Accuracy of labels predicted as correct by a style classifier.
  - **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.
  - **Fluency:** Perplexity of the generated sentences using GPT-2 (Radford et al., 2019).
- Classifiers used to evaluate the accuracy of style transfer

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

Main Results

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.

- Overall Performance
  - All baseline systems do not outperform the original LLaMA-3 model in the TST task, with the exception of the content preservation metric.
  - Our method outperforms the baseline methods in terms of both accuracy and fluency.
- Content Preservation
  - copy mechanism, i.e., the generated text tends to prioritize maintaining the original semantics, thereby neglecting the stylistic differences.
- Different Directions
  - significant performance discrepancies when transferring between different directions within the same task.

Ablation Study

- Neuron overlapping ablation analysis

	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**

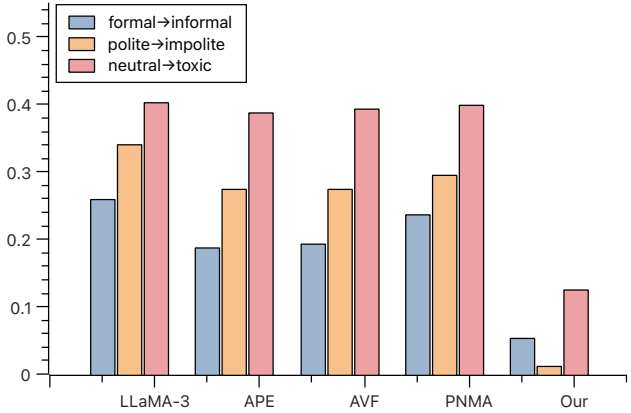
Table 3: **Ablation study:** Style transfer accuracy on removing overlap between source- and target-side style neurons in six benchmarks. "with" indicates the removal of overlap.

- Deactivation and contrastive decoding ablation analysis

	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.

Copy Problem



- Neuron-based approaches (APE, AVF, and PNMA) partially mitigate this issue by controlling neuron activation;
- our approach achieves a reduced copy rate by deactivating source-side neurons and employing a novel decoding strategy.

Further Analysis

- Different model
  - Our method consistently demonstrates effectiveness across diverse model sizes, including larger models like 70B.
- Layer selection strategies
  - Selecting the last few layers proves optimal compared to earlier layers.
- Content preservation metrics
  - Different strategies for preserving meaning yield similar outcomes, highlighting the importance of exploring innovative approaches in future research.
- Decoding strategies
  - Contrastive decoding exhibits significant advantages over traditional decoding methods in the TST task, motivating our adoption of CD strategy.

Conclusion

- **(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.

Thank You!

Email: wen.lai@tum.de
Homepage: https://wenlai-lavine.github.io
Address: Bildungscampus 3, 74076 Heilbronn, Germany

Paper

Code