Style-Specific Neurons for Steering LLMs in Text Style Transfer

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

[1] Technical University of Munich, [2] Munich Center for Machine Learning,

[3] Center for Information and Language Processing, LMU Munich

{wen.lai, alexander.fraser}@tum.de, hangyav@cis.lmu.de

EMNLP 2024

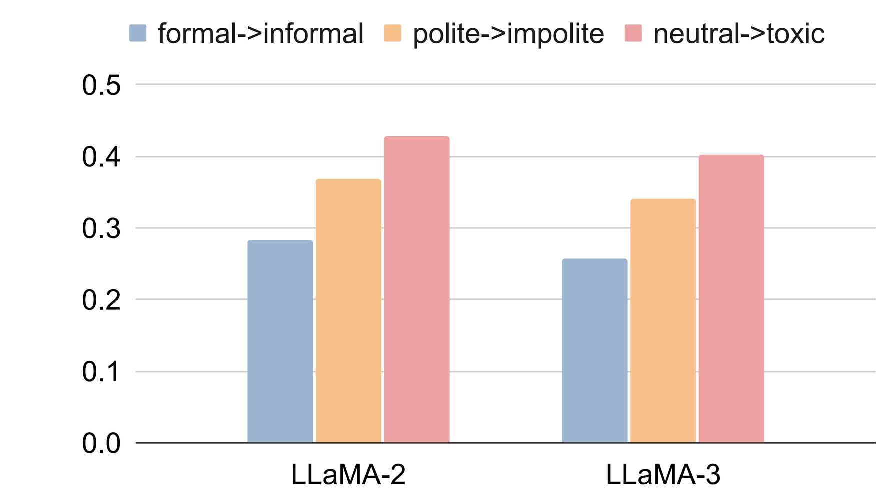
Background

Task Definition: 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.

Motivation:

- LLMs tends to directly *copy* a significant portion of the input text to the output without effectively changing its style.
- Recent research (Tang et al., 2024) has demonstrated that language-specific neurons exists in LLMs, however, deactivating such neurons leads to a remarkable degradation in the model's understanding and generation abilities for that language.

Copy Ratio in LLMs on TST Task



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?

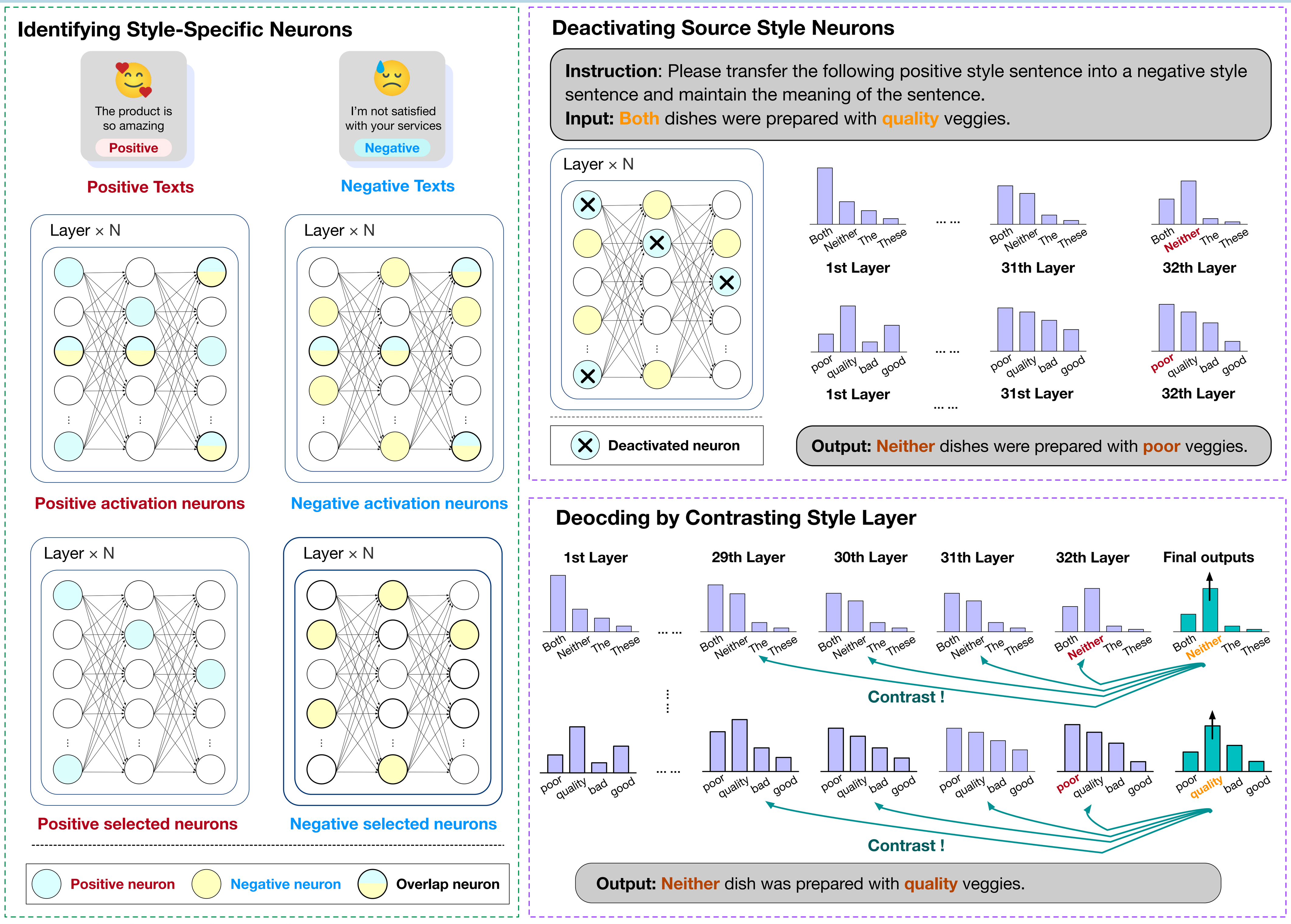
Goal: Enhancing the generation of words that align with the target style during the decoding process.

Key Contributions

- 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.
- We emphasize the significance of *eliminating overlap* between neurons activated by source and target styles, a methodological innovation with potential applications beyond TST.
- 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.

Paper Code Contact

Methods



Identify Style-Specific Neurons

- High overlap among style-specific neurons when using neuron selection from (Tang et al., 2024).
- We remove the overlap between source and target style neuron.

Deactivating Source Style Neurons

- Deactivate which side? source or target?
- Deactivate source-style neurons improves the accuracy but decreases the fluency.

Decoding by contrasting Style Layer

- Motivated by Dola (Chuang et al., 2024), which amplify the factual knowledge in higher layers.
- We adapt Dola to TST, i.e., amplify the style specific words during decoding.

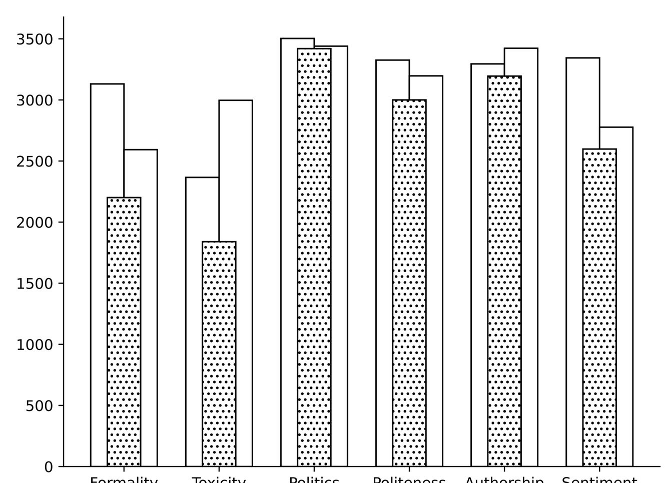


Fig. Overlap statistics in (Tang et al., 2024)

Style Accuracy

Source	Target	Formality: informal	Formality: formal	Politeness: impolite	Politeness: 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: informal	Formality: formal	Politeness: impolite	Politeness: 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. Which side should we deactivate?

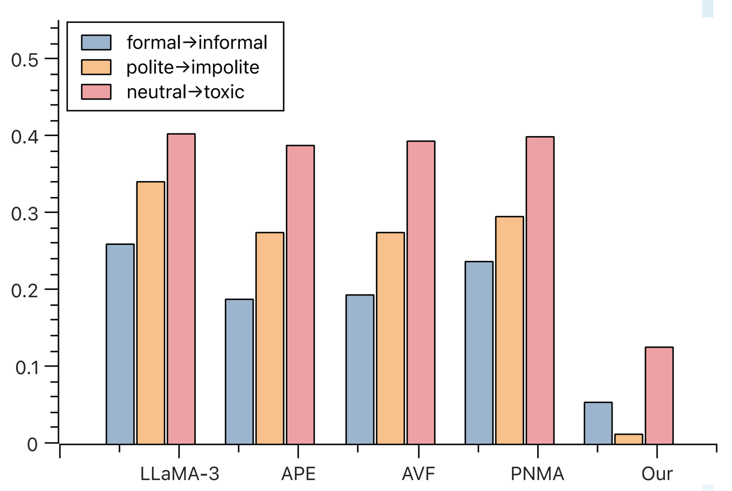


Fig. Copy ratio on three selected TST tasks

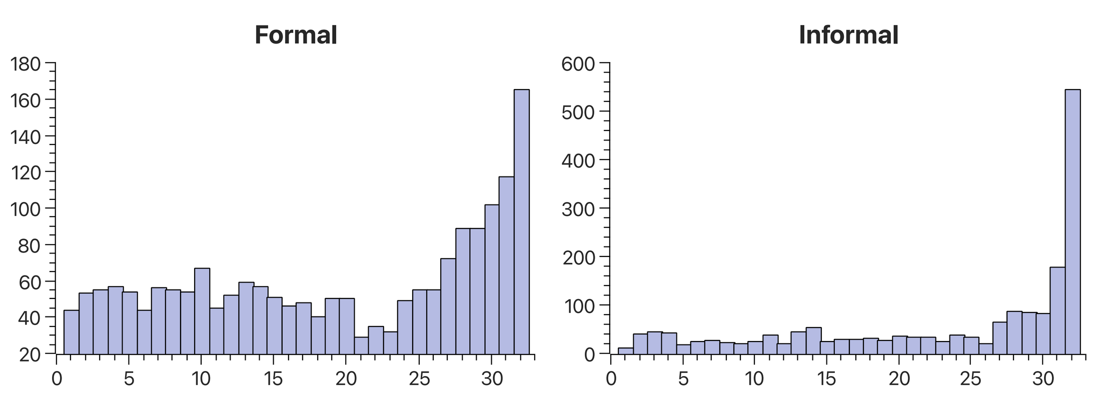


Fig. Statistics of style-specific neurons in each layer

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.

Tab. Case study

Results

Style Transfer Accuracy

	Formality: informal	Formality: formal	Toxicity: toxic	Toxicity: neutral	Politics: democratic	Politics: republican	Politeness: impolite	Politeness: polite	Authorship: shakespeare	Authorship: modern	Sentiment: positive	Sentiment: 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: informal	Formality: formal	Toxicity: toxic	Toxicity: neutral	Politics: democratic	Politics: republican	Politeness: impolite	Politeness: polite	Authorship: shakespeare	Authorship: modern	Sentiment: positive	Sentiment: 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: informal	Formality: formal	Toxicity: toxic	Toxicity: neutral	Politics: democratic	Politics: republican	Politeness: impolite	Politeness: polite	Authorship: shakespeare	Authorship: modern	Sentiment: positive	Sentiment: 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**

- We obtain the best results on style accuracy and fluency.
- Interestingly, our approach do not show advantages in content preservation (Section 5).
  - Attributable to the copy mechanism, i.e., the generated text tends to prioritize maintaining the original semantics, thereby neglecting the stylistic differences.

Ablation Study

- **Removing neuron overlapping?**

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**

- **Neuron Deactivation and Contrastive Decoding?**

	Deactivate	Contrastive	Toxicity: toxic	Toxicity: neutral	Authorship: shakespeare	Authorship: 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**

References

- Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models (Tang et al., ACL 2024)
- DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models (Chuang et al., ICLR 2024)