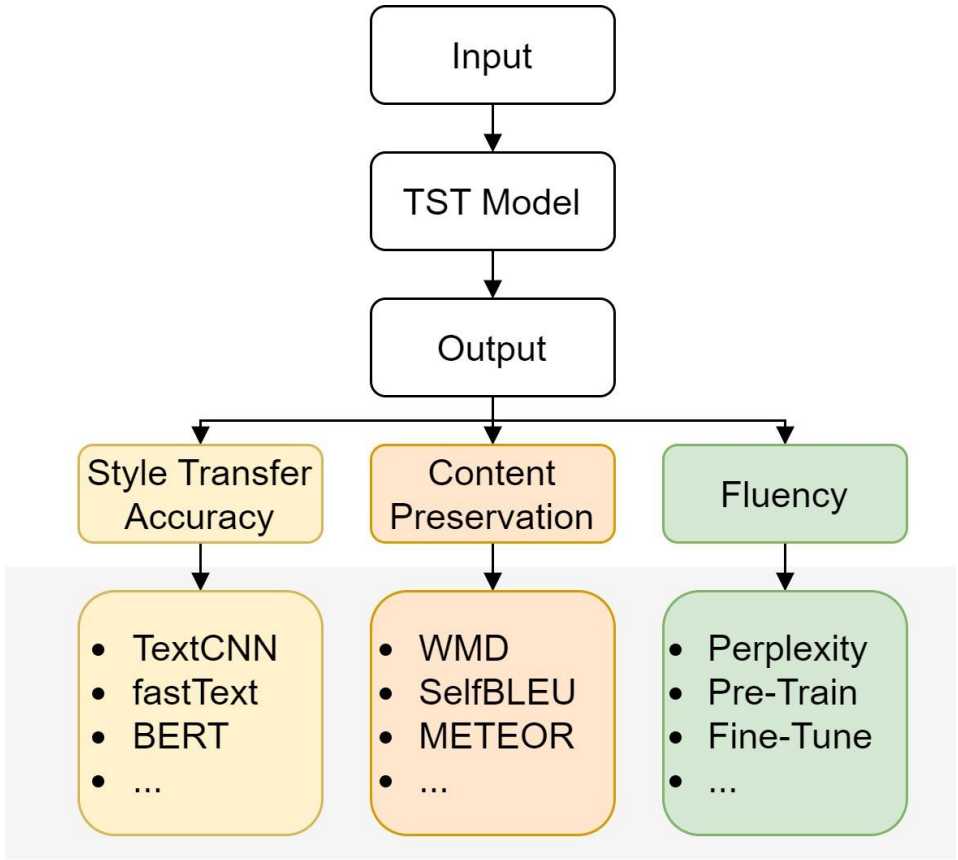
Text Style Transfer Evaluation Using Large Language Models

Phil Ostheimer, Mayank Nagda, Marius Kloft, Sophie Fellenz

LREC-COLING 2024

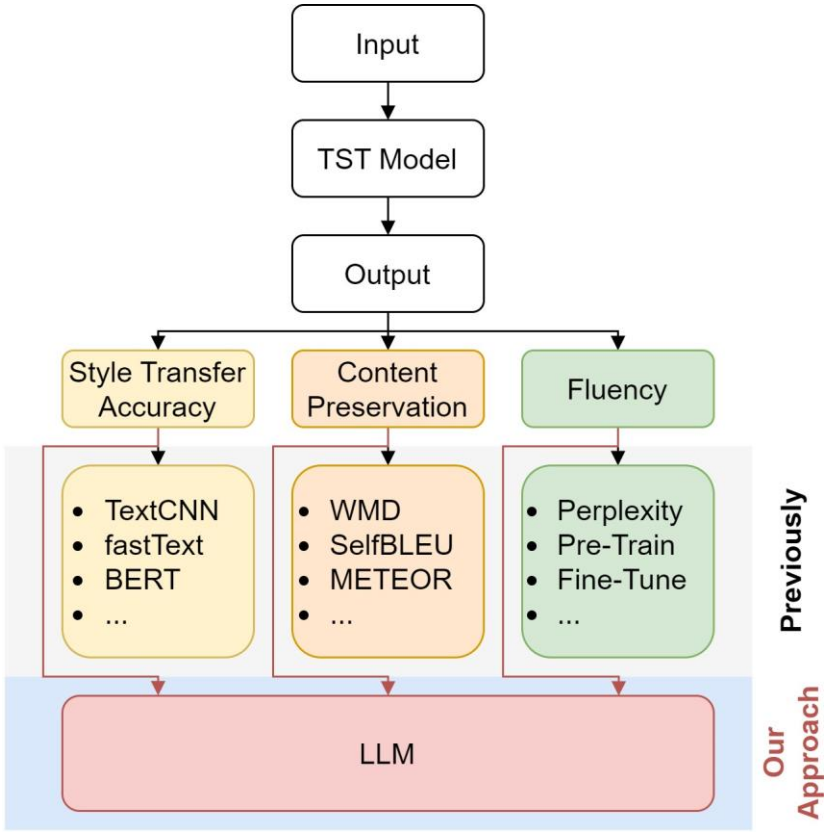
RPTU Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Many Automated Metrics for Text Style Transfer Evaluation



RPTU

Standardized Solution: LLM Evaluation



RPTU

Example Prompts

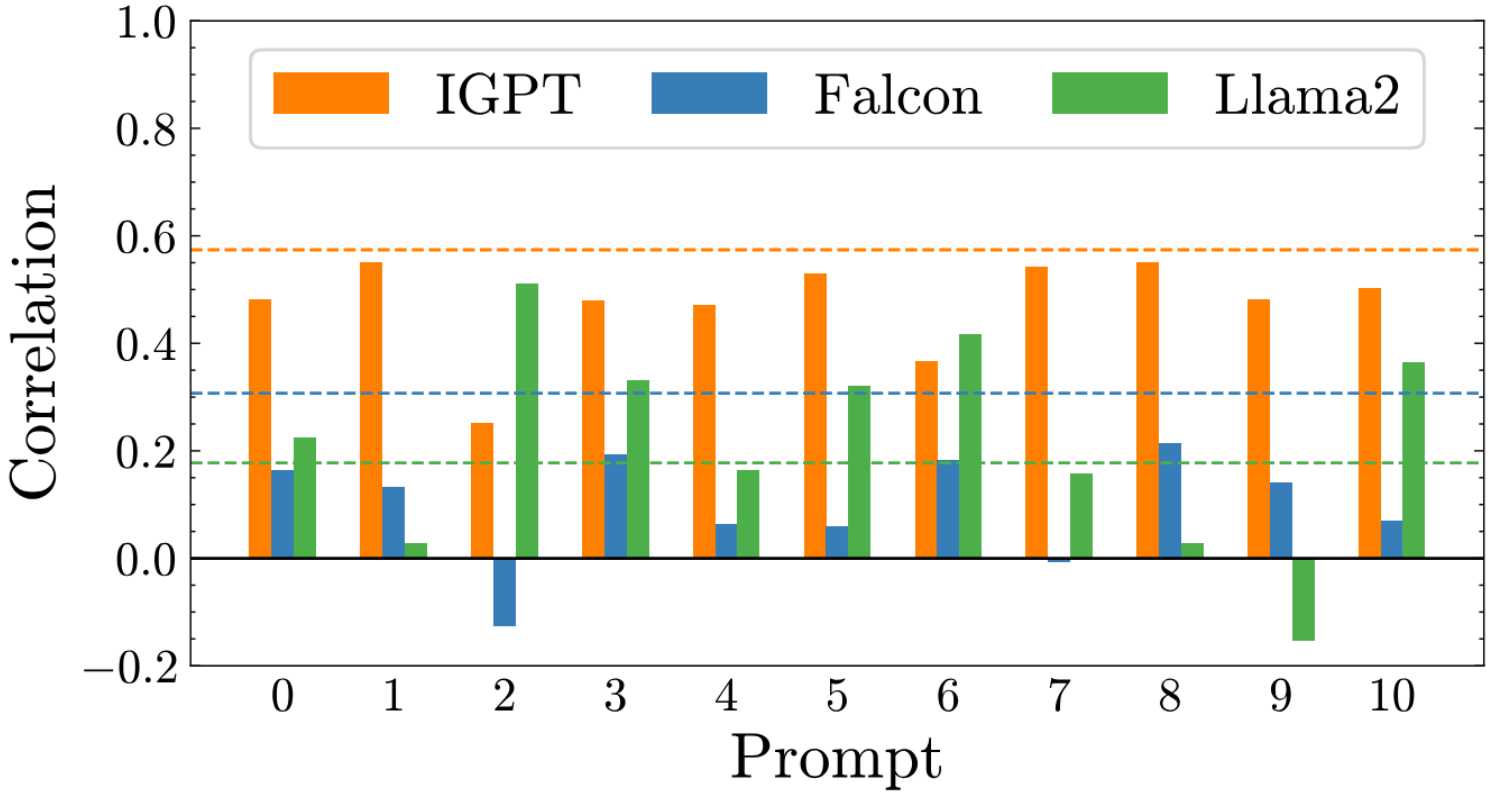
Style Transfer Accuracy	Prompt	Here is sentence S1: {**Overall, it was horrible.**} and sentence S2: {**Overall, it was great.**}. How different is sentence S2 compared to S1 on a continuous scale from 1 (completely identical styles) to 5 (completely different styles)? Result =
	Answer	**5.0**
Content Preservation	Prompt	Here is S1: {**Overall, it was horrible.**} and sentence S2: {**Overall, it was great.**}. How much does S2 preserve the content of S2 on a continuous scale from 0 (completely different topic) to 1 (identical topic)? Result =
	Answer	**1.0**
Fluency	Prompt	How natural is this sentence S1 {**Overall, it was great.**} on a scale from 1 to 5 where 1 (lowest coherent) and 5 (highest coherent)? Result =
	Answer	**5.0**

RPTU

Chosen LLMs

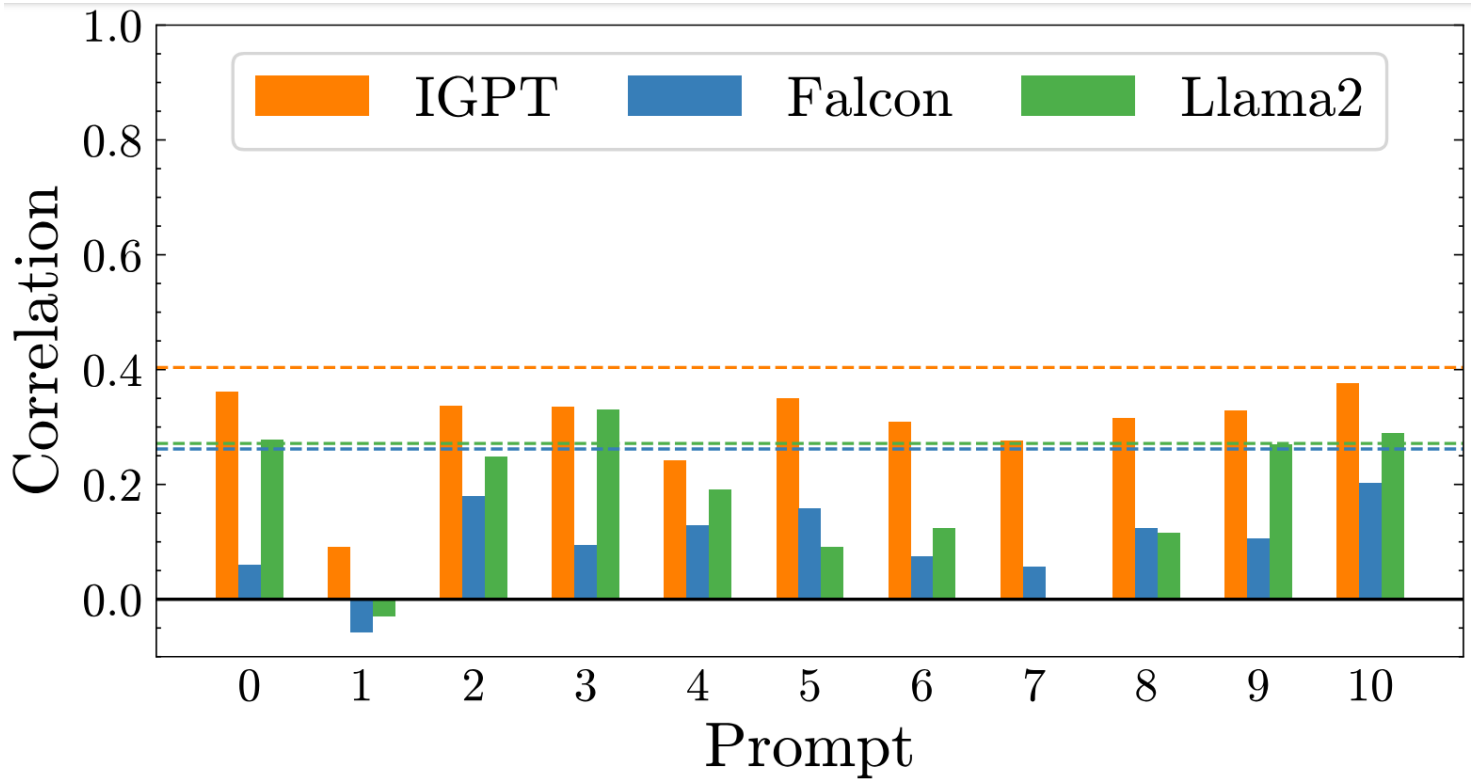
- Selection of recently published state-of-the-art LLMs
- Pre-trained: OPT, BLOOM, GPT3, Falcon, Llama2
- Pre-trained plus fine-tuned to follow instructions: InstructGPT, Falcon, Llama2

Correlations with Humans – Style Transfer Accuracy



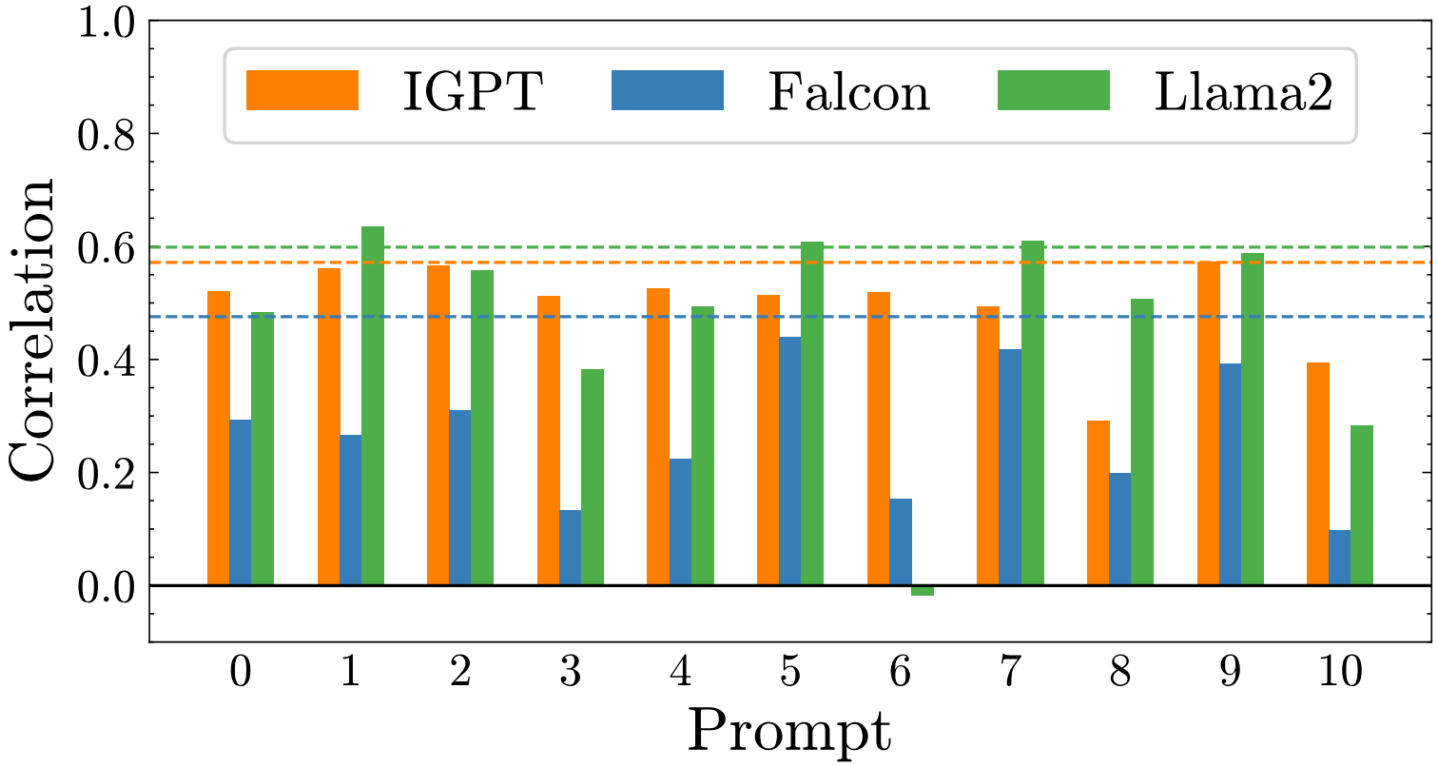
RPTU

Correlations with Humans – Content Preservation



RPTU

Correlations with Humans – Fluency



RPTU

Stronger Correlations Than Other Automated Metrics

Style Transfer Accuracy				
	ARAE	CAAE	DAR	All
fastText	0.498	**0.550**	0.332	0.473
TextCNN	0.512	0.525	0.331	0.458
BERT	0.513	0.559	0.408	0.497
IGPT	**0.618**	0.543	**0.584**	**0.574**
Fal-7b	*-0.027*	-0.219	*-0.118*	-0.131
Fal-40b	0.206	0.389	0.313	0.307
Lla-7b	*0.091*	-0.128	*-0.064*	*-0.039*
Lla-13b	*0.103*	*0.018*	*0.106*	*0.067*
Lla-70b	0.347	*0.075*	*0.077*	0.178

RPTU

Parsable Answers

	STA	CP	F
IGPT	100.0%	100.0%	100.0%
Falcon-7b	100.0%	99.8%	99.9%
Falcon-40b	92.1%	90.5%	89.8%
Llama2-7b	75.6%	59.5%	98.1%
Llama2-13b	85.3%	80.8%	99.0%
Llama2-70b	69.1%	71.4%	98.9%

In-Range Scores

	STA	CP	F
IGPT	100.0%	100.0%	100.0%
Falcon-7b	100.0%	99.9%	99.7%
Falcon-40b	100.0%	100.0%	99.0%
Llama2-7b	99.6%	96.3%	99.8%
Llama2-13b	99.5%	99.4%	99.8%
Llama2-70b	99.6%	99.0%	99.8%

Example Evaluations

Prompt	Here is sentence S1: {honestly they were down right rude .} and sentence S2: {honestly they were down right , friendly and fast !}. How different is sentence S2 compared to S1 on a scale from 1 (identical styles) to 5 (completely different styles)? Result =
IGPT	[...]5 (completely different styles)
Falcon-7b	[...]" 3 S1 is 4 (more similar styles) and S2 is 5 (completely different styles)"
Falcon-40b	[...]" 3 The difference between sentence S1 and S2 is a score of 3 out of 5."
Llama2-7b	[...]" To compare the styles of sentences S1 and S2, we need to analyze their grammatical structures, word choices, and tone. Here's a detailed analysis: 1. Grammar: Sentence S1: * Subject: They were * Verb: were * Ad"
Llama2-13b	[...]" Sure, I can help you with that! Sentence S1: {honestly they were down right rude .} Sentence S2: honestly they were down right, friendly and fast ! The difference between these two sentences is a 5"
Llama2-70b	[...]" The two sentences have different styles, but they share some similarities. S1 has a negative tone, using the word ""rude"" to describe someone's behavior. The sentence structure is also simple, with a single clause expressing a complete thought. S2, on the other hand"

RPTU

Conclusion

- Introduced standardized TST evaluation
- Can replace previous automated metrics for all three aspects
- Demonstrated validity whereas InstructGPT has the highest correlations
- Ensembling improves performance