Text Style Transfer Evaluation Using Large Language Models

Phil Ostheimer, Mayank Nagda, Marius Kloft, Sophie Fellenz

University of Kaiserslautern-Landau

How to Standardize Automated Text Style Transfer Evaluation?

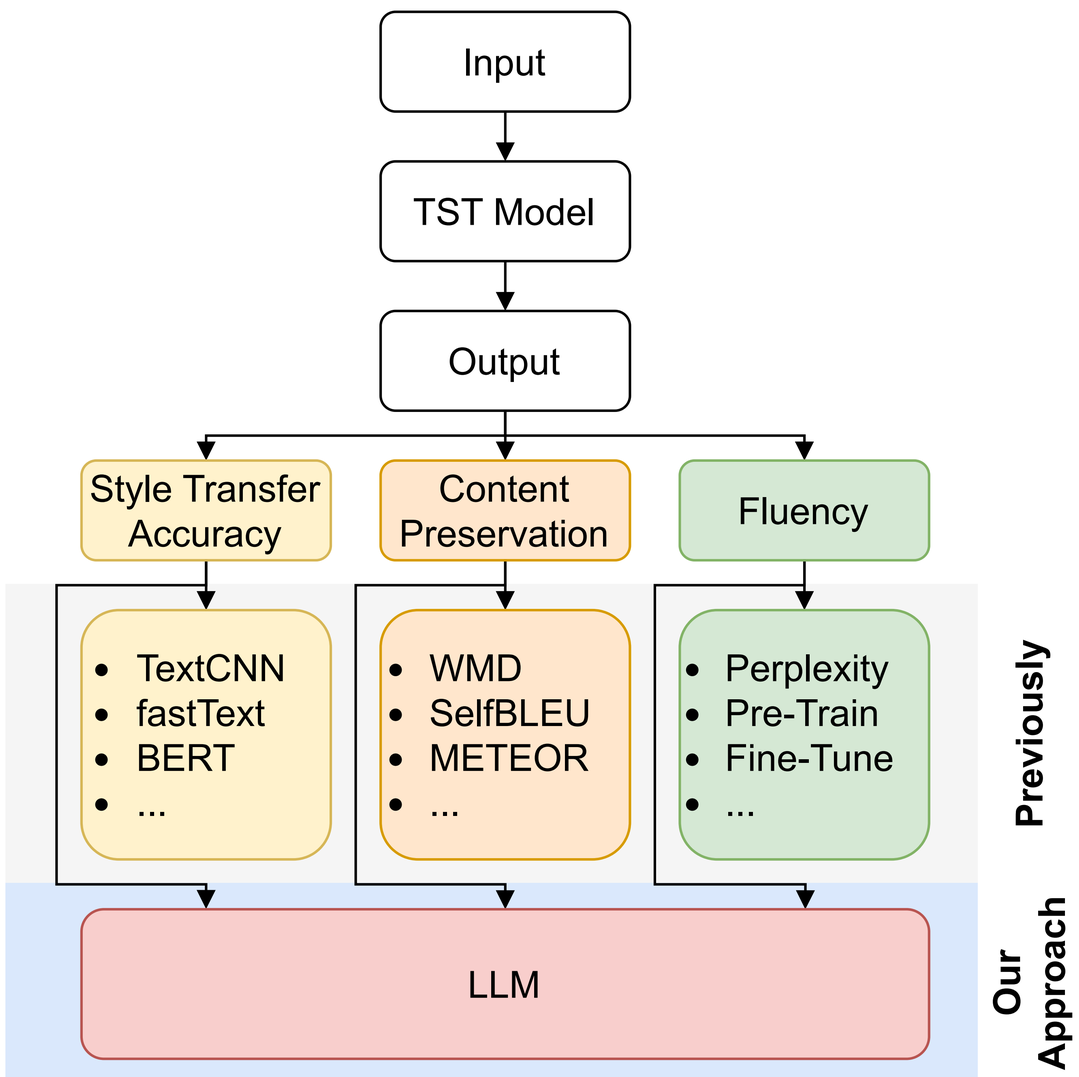

Figure 1. Shown is the multitude of automated text style transfer evaluation metrics. We set out to replace all of them with LLM evaluation.

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**

Figure 2. Shown is one example prompt per evaluation aspect.

Prompt Ensembling for Higher Correlations

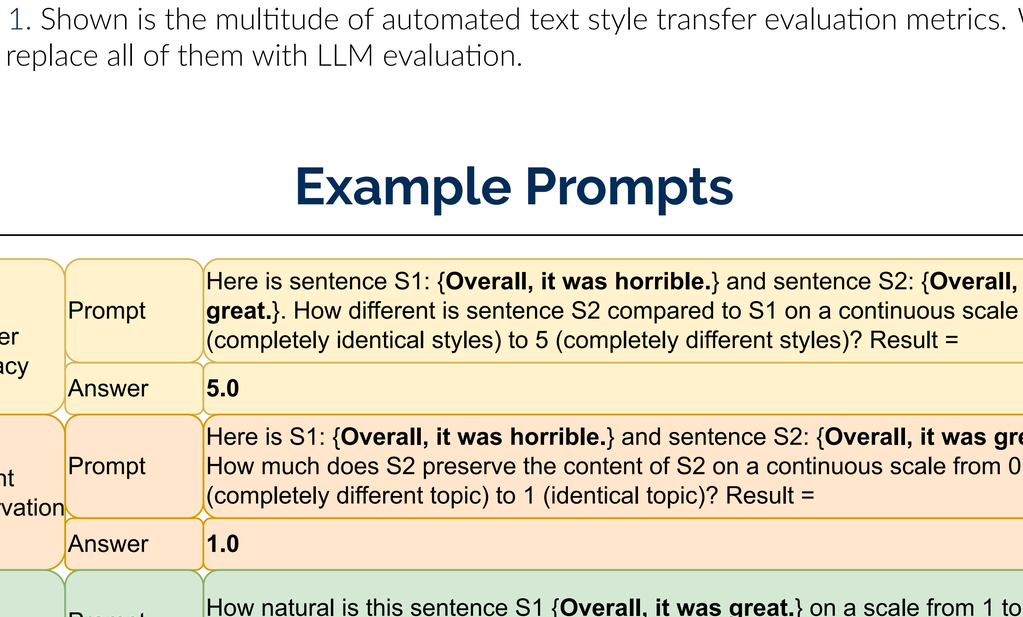

Figure 3. Shown is the correlation between the scores of each prompt and human evaluations for style transfer accuracy. Dashed lines indicate ensembled prompts.

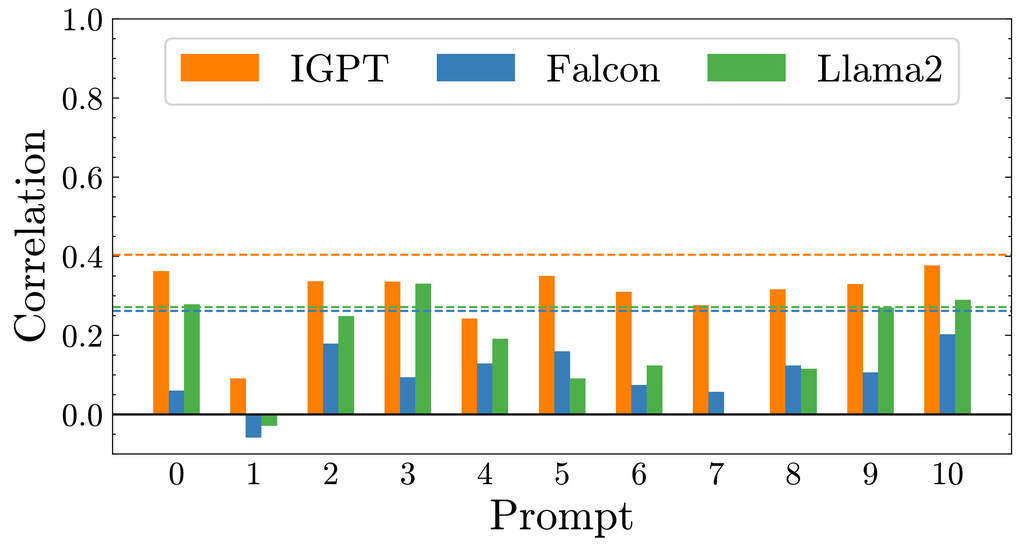

Figure 4. Shown is the correlation between the scores of each prompt and human evaluations for content preservation. Dashed lines indicate ensembled prompts.

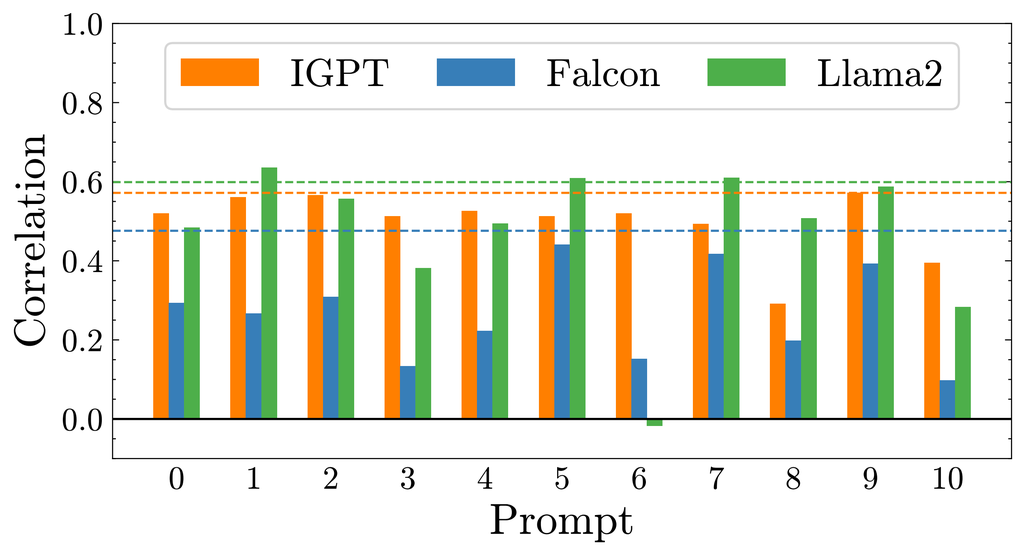

Figure 5. Shown is the correlation between the scores of each prompt and human evaluations for fluency. Dashed lines indicate ensembled prompts.

LLM Evaluation vs. Previous Automated Metrics

	ARAE	CAAE	DAR	All
Style Transfer Accuracy				
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-40b	0.206	0.389	0.313	0.307
Lla-70b	0.347	*0.075*	*0.077*	0.178
Content Preservation				
METEOR	**0.247**	**0.659**	**0.425**	**0.420**
WMD	0.240	0.615	0.361	0.377
IGPT	0.191	0.656	0.345	0.404
Fal-40b	0.167	0.386	0.240	0.262
Lla-70b	*0.104*	0.484	0.198	0.271
Fluency				
PPL PT	*0.076*	*0.044*	0.418	0.171
PPL FT	0.135	*0.120*	0.411	0.232
IGPT	0.518	**0.560**	**0.603**	0.571
Fal-40b	0.436	0.452	0.491	0.476
Lla-70b	**0.539**	0.551	0.602	**0.599**

Table 1. Shown are the Spearman rank correlations between human evaluations and the mentioned automated metrics, including InstructGPT (IGPT), Falcon (Fal), and Llama2 (Lla). All *italic correlations* have p>0.05.

SPONSORED BY THE
Federal Ministry of Education and Research