Mind the Style Gap: Meta-Evaluation of Style and Attribute Transfer Metrics

Amalie Brogaard Pauli[1] **Isabelle Augenstein**[2] **Ira Assent**[1]
[1]Department of Computer Science, Aarhus University, Denmark
[2]Department of Computer Science, University of Copenhagen, Denmark
{ampa,ira}@cs.au.dk, augenstein@di.ku.dk

Abstract

Large language models (LLMs) make it easy to rewrite a text in any style – e.g. to make it more polite, persuasive, or more positive – but evaluation thereof is not straightforward. A challenge lies in measuring *content preservation:* that content not attributable to style change is retained. This paper presents a large meta-evaluation of metrics for evaluating style and attribute transfer, focusing on content preservation. We find that meta-evaluation studies on existing datasets lead to misleading conclusions about the suitability of metrics for content preservation. Widely used metrics show a high correlation with human judgments *despite* being deemed unsuitable for the task – because they do not abstract from style changes when evaluating content preservation. We show that the overly high correlations with human judgment stem from the nature of the test data. To address this issue, we introduce a new, challenging test set specifically designed for evaluating content preservation metrics for style transfer. We construct the data by creating high variation in the content preservation. Using this dataset, we demonstrate that suitable metrics for content preservation for style transfer indeed are style-aware. To support efficient evaluation, we propose a new style-aware method that utilises small language models, obtaining a higher alignment with human judgements than prompting a model of a similar size as an autorater.

1 Introduction

Large language models allow for rewriting text in a variety of styles or alter any text attribute, without the need for training data. Examples are to make a text more formal, polite, simpler, or to change its sentiment. However, evaluating style and attribute transfer is still challenging as it lacks validation and standardization (Ostheimer et al., 2023; Briakou et al., 2021b).

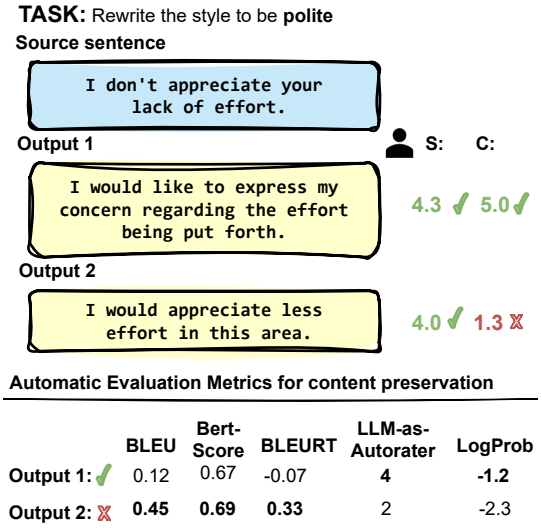

Figure 1: Sample from our new test set with human annotations. S: style strength, C: content preservation. Metrics for content preservation: Bold indicates the highest rated output, and the checkmark indicates successful transfer.

Traditionally, style and attribute transfer are evaluated via: 1) style strength/shift; 2) fluency and 3) content preservation (Jin et al., 2022). Style transfer papers heavily rely on automatic evaluation metrics (Ostheimer et al., 2023). To assess which metrics are most suitable, meta-evaluation efforts measure the correlation of metrics with human judgments.

Measuring content preservation is particularly difficult for style and attribute transfer, as this requires the distinction between content preservation and changes to style or attributes. Still, the status quo is to use metrics not designed for the task of style transfer (most commonly BLEU, Ostheimer et al. (2023); Jin et al. (2022)), which assess the lexical or semantic similarity of rewrites to the source sentence, not considering style change. Moreover, prior work criticises the use of similarity-based metrics for content preservation (Mir et al., 2019;

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21550–21564
November 4-9, 2025

Lai et al., 2024; Scialom et al., 2021b; Logacheva et al., 2022a). Intuitively, measuring the similarity between source and output is not suitable for the style transfer task, because the more the style or attribute in the output changes, the more dissimilar the sentences become. Hence, metrics for content preservation should be style-aware (Mir et al., 2019). Empirically, however, similarity-based metrics show good correlation with human judgments. In fact, approaches using LLMs only show comparable, and not superior, results to those of semantic similarity metrics (Lai et al., 2023; Ostheimer et al., 2024). However, this is even though the LLM-based approaches are a better fit for the style transfer task, as they can be style-aware, as opposed to the similarity-based metrics.

Fig. 1 illustrates a case where similarity-based metrics between source and output fail to correctly identify which output best preserves content, as opposed to style-aware approaches (the last two). In this paper, we study the research question: **Why do metrics that are theoretically unsuitable for content preservation in style transfer show a high correlation with human judgments?** as illustrated in the first two columns in Fig. 2.

To answer this question, we conduct a large-scale meta-evaluation study with a particular focus on metrics for content preservation. Our key finding is that current meta-evaluation efforts for content preservation have misleading conclusions and overestimate the performance of similarity-based metrics. Consequently, the use of such metrics leads to misleading evaluation results of style transfer methods.

We identify the data used for meta-evaluation as the primary reason for the unreliable correlation results. To overcome this data bias, we propose and construct a test set that offers carefully designed test cases – featuring errors such as substituted, deleted and fabricated information, which are unrelated to style, but change the content (see Figure 1).

Our aggregated results (Figure 2) demonstrate that similarity-based metrics for content preservation show low to negative correlations with human judgment on our test set (last column), but exhibit overly high correlations to human judgment on existing data sets (first two columns). Style-aware approaches show positive correlations to human judgments on all data types (all columns). Note, our dataset is designed to stress-test the metrics, and we therefore recommend assessing all data types for evaluating the metrics.

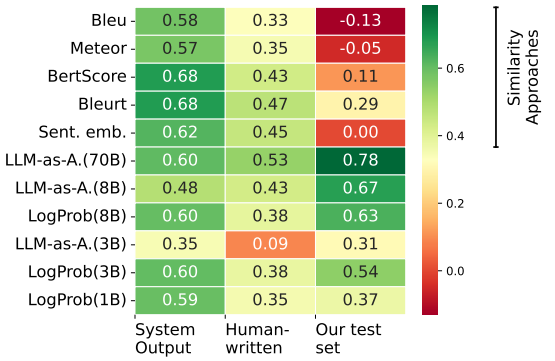

Figure 2: Average Spearman correlation between metrics and human judgments for content preservation, grouped by whether test data originates from system-generated output (total of 5 sources), human-written (total of 3 sources) or our constructed test set. Which metric is deemed best changes depending on the test data, and only our test set correctly finds similarity-based metrics to be unsuitable.

We propose the style-aware method **LogProb** to evaluate style strength and content preservation using estimated token probabilities. LogProb offers better performance on smaller language models (1B,3B) than a strong baseline of using similar-sized LLMs as autoraters.

In sum, our **contributions** are:

- We construct a new test set[1] (500 human-annotated samples) with carefully designed test cases as a new resource for evaluating content preservation metrics in style transfer;
- We conduct a large meta-evaluation of content preservation spanning 9 metrics/approaches on a variety of tasks (9) and data (7), offering recommendations on metrics for measuring content preservation and style strength;
- We show that existing meta-evaluations may lead to misleading conclusions for metrics for content preservation, namely that widely used semantic similarity metrics are not a good fit – we recommend to discontinue their use for evaluating style transfer;
- We propose an efficient, zero-shot method LogProp for evaluating style transfer, adopting small language models (3B,1B), accounting for content preservation and style strength. On small model sizes, LogProb outperforms our LLM-as-Autorater baseline.

[1] github.com/AmaliePauli/style_transfer_evaluation

2 Background and Related Work

Style and Attribute Transfer. We use the working definition from Jin et al. (2022) of *style transfer as a generation task aiming to control specific attributes of the text*. Note that this definition of style transfer is data-driven and different from a linguistic definition. We likewise note that 'attribute' is a broader category that covers changes that affect semantics, where a linguistic understanding of 'style' would not, e.g., constitute a change of sentiment. Hence, the aim is to change the attribute/style while preserving the *contextual or semantic* content.

Evaluation. Traditionally, style transfer systems have been evaluated along three dimensions: 1) style strength – did the style successfully shift?; 2) content preservation – is the content otherwise the same?; 3) fluency – is the rewrite fluent and grammatically correct? (Jin et al., 2022).[2] Some work on style transfer systems includes small-scale human evaluations of the systems, but papers heavily rely on automatic metrics for evaluation (Ostheimer et al., 2023):

Style. The de facto evaluation approaches, also in recent work, use either a classifier, e.g. RoBERTa, trained on a specific style (Lai et al., 2024; Hallinan et al., 2023; Han et al., 2024; Mukherjee et al., 2024; Liu et al., 2024; Luo et al., 2023; Zeng et al., 2024), or a regressor (Briakou et al., 2021a; Lai et al., 2022; Pauli et al., 2025). However, both require training data for the specific style, limiting their scope and scale. Further, Briakou et al. (2021a) criticises the use of binary classifiers, as this approach does not capture that a style can shift to differing degrees. In this paper, we focus on evaluating arbitrary styles without trained classifiers or regressors, and we focus on approaches that output scores to assess different degrees of style changes.

Content Preservation. Many different metrics are used to evaluate style transfer for content preservation (Ostheimer et al., 2023). The most commonly used is lexical similarity using BLEU (Ostheimer et al., 2023; Jin et al., 2022), but widely used ones include BertScore, cosine distance between embeddings, and BLEURT - also in recent works (Lai et al., 2024; Hallinan et al., 2023; Han et al., 2024; Mukherjee et al., 2024; Liu et al., 2024; Zhang et al., 2024; Luo et al., 2023). These metrics originate from other natural language generation tasks, e.g., translation, to measure the similarity between output and gold references. However, in the absence of such references, many style transfer papers use these metrics to compare source and output sentences. We evaluate these approaches using source sentences compared to output [**S**ource-], and references compared to output [**R**eference-] where available.

Prior work has criticised the use of similarity-based approaches for measuring content preservation in style transfer. Mir et al. (2019) highlight the flaw of metrics like source BLEU, which cannot distinguish between style and content changes, thus penalising style edits as lower content preservation. Lai et al. (2024) note that similarity-based metrics favour copying over paraphrasing, disadvantaging certain transfer models. Cao et al. (2020) further argue that style transfer evaluation with these metrics for content preservation can be gamed by simply appending style words to the source sentence.

A proposed solution to this criticism is to make metrics for content preservation style-aware. Mir et al. (2019) proposes to use a lexicon of style words to mask the sentence in a sentiment task. In a follow-up study, Yu et al. (2021) extract a style lexicon automatically. More recent work prompts an LLM for evaluation (LLM as an autorater, Zeng et al. (2024); Lai et al. (2023); Ostheimer et al. (2024)). For other natural language generation tasks, LLM-as-a-judge approaches are deemed better than similarity-based metrics when tasks are semantically nuanced (Li et al., 2024).

Still, despite the established criticism of the use of semantic similarity metrics for content preservation in style transfer, these metrics often show reasonable correlation with human judgment on content preservation in style transfer. Prior meta-evaluation efforts (Lai et al., 2023; Ostheimer et al., 2024) show that using LLMs for evaluating content preservation is only on par rather than superior to using semantic similarity-based metrics.

We investigate the discrepancy between what would conceptually be a suitable metric and what meta-evaluation results show. In addition, we propose a new target-style-aware metric with improved efficiency compared to an LLM-as-a-judge approach.

Meta-evaluation. In order to determine suitable metrics for evaluating style transfer, prior work has examined how metrics correlate with human judgments. Human judgments are often collected as multiple annotations, for style, fluency, and content preservation separately, e.g., asking how successful

[2]Note that fluency is not covered in this work

the style change is on a scale from 1 to 5.

Examples of meta-evaluation studies on different style transfers include sentiment (Yu et al., 2021; Mir et al., 2019; Ostheimer et al., 2024), formality (Briakou et al., 2021a; Lai et al., 2022), simplification (Scialom et al., 2021b; Alva-Manchego et al., 2021; Cao et al., 2020), and detoxifying (Logacheva et al., 2022a).

Criticism. Prior studies have voiced different criticisms about the meta-evaluations themselves: For simplicity transfer, Scialom et al. (2021b) find that using system-generated rewrites for meta-evaluations leads to spurious correlation due to inter-correlation between dimensions, e.g. rewrites low on fluency also tend to be low on content and simplicity. Instead, they propose using human-written data for meta-evaluation purposes. In our broader study, we also group data into system-generated and human-written outputs. We further construct a more challenging dataset, since we argue that existing human-written test sets do not include instances of low preserved content. Also, for a simplicity task, Devaraj et al. (2022) examine faithfulness mistakes in training data and in rewrites by systems. They show that metrics have more difficulty locating some error types related to low faithfulness – we take inspiration from these error types when constructing our test set. Logacheva et al. (2022a) conclude, on detoxifying transfer, that automatic metrics are less reliable for high-performing systems, as the correlation between metrics and human judgment is lower for some systems' output, thereby pointing out an issue of meta-evaluation itself.

In our paper, we address the problem of meta-evaluation across a wide range of style transfer tasks. We propose a new test set, which is constructed to show the ability of metrics to detect low content preservation when the style or attribute change.

3 Methodology: Test Set

We propose a new dataset to test metrics on content preservation used in style transfer, inspired by several shortcomings we observe in existing meta-evaluation studies. Namely, meta-evaluation studies (both prior studies described in Section 2 and our study in Section 6), show e.g. that similarity-based metrics highly correlate with human judgments and even outperform LLM-based approaches. However, we hypothesise that this is a misleading result caused by shortcomings in how the meta-evaluations are conducted. The reason is that conceptually, similarity-based metrics for content preservation between source and output sentences are unsuitable for the task, because they favour verbatim repetitions between source and output sentences, and the more the style changes, the more the similarity drops despite the content being preserved (Section 2).

We illustrate the shortcomings of similarity-based metrics with an example (Fig. 1). Suppose the task is to rewrite *I don't appreciate your lack of effort* to be more polite. Consider these rewrites:

- Output 1: *I would like to express my concern regarding the effort being put forth*
- Output 2: *I would appreciate less effort in the area*

Human annotators rate style shift on a 1–5 scale (5 = very good). Both outputs are seen as successfully shifting to a more polite tone, but Output 2 scores poorly on content preservation. Yet, similarity-based metrics wrongly rate Output 2 as better at preserving content than Output 1, while style-aware metrics correctly assign it a lower score. The employed metrics are formally introduced in Section 5.

Figure 1 is an example of how we can construct more challenging data to test the abilities of content preservation metrics to detect low content preservation under style shifts.

We aim to create a challenging test set for meta-evaluating metrics for content preservation. Our core principle is to create data that succeeds in style transfer but has a high variation in content preservation. We create variation in content preservation by deliberately creating mistakes in the content, such that we can assess the metrics' suitability to detect low content preservation. We supply the data with human judgments for both style strength and content preservation. For each source sentence, we construct two rewrites:

1) one where we preserve the content, and shift the style to a large degree, with more variation in the rewrite;

2) one where we shift style to a lesser degree, staying closer to the wording in the source sentence but, introducing a content error.

The methodology for adding content errors is inspired by Devaraj et al. (2022) – they analyse text simplification by categorising errors as substitution, deletion and insertion. For various style transfer tasks, we construct errors with these categories in

mind – we substitute or swap key information, drop key information, or fabricate additional information not supported by the source sentence.

We aim to construct data where all output sentences succeed on style transfer, but for each source sentence, only one output preserves the content well, creating variation in content preservation quality.

Other examples from our test set include the task of rewriting the following sentence to be detoxified; *yeah but she is making me feel like shit*; with the two rewritings:

- Output 1: *Yes I know, but she is really making me feel terrible*
- Output 2: *yeah but she is making me feel*

Another example is a neutral headline from WikiNews to be framed more positively; *President of China lunches with Brazilian President*; with the rewritings;

- Output 1: *The Great Presidents of China and Brazilian strengthen important ties over lunchs*
- Output 2: *The President of China enjoys lunches with the Brazilian first lady*

In total, our constructed test data covers six style/attribute tasks and consists of 500 samples, 100 of which are manually created and the remaining generated using an LLM for scale.

Manually created. We construct 100 transfer samples from a source sentence by manually creating two output sentences as described. We create: Task 1) on sentiment (positive/negative), using headlines from Wiki News[3] as source sentences with minor modifications; Task 2) on detoxifying, using toxic sentences from Logacheva et al. (2022b). We supply each sample with a gold-truth reference sentence to enable a small evaluation of reference-based methods.

LLM generated. We use GPT-4o-mini[4] to create style transfer pairs in a multi-step process:

1. Generate source sentences.
2. Generate two rewrites for each source:
    a. *a lot more* in the target style
    b. *a bit more* in the target style with a content error of substitution, deletion or fabrication of information.

We construct the LLM-generated part of the test set with four different style transfer tasks: i) a headline to be more catchy; ii) an impolite sentence from an email to be polite; iii) a persuasive request to be more persuasive; and iv) a sentence with informal language, with grammatical mistakes and internet slang to be formal. Prompt details are included in App. B, where we also include an automatic assessment of the 'linguistic acceptability' of the samples.

[3]Wikinews.org
[4]openai.com

Task		#	C (IAA)	S>=3 (%)
sentiment	ma	50	0.676	88
detoxify	ma	50	0.758	96
catchy	LLM	100	0.806	90
polite	LLM	100	0.645	100
persuasive	LLM	100	0.80	88
formal	LLM	100	0.817	99

Table 1: Stats on our test set. ma: manually created, LLM: LLM-generated. C: content preservation, S: style strength. All samples are manually annotated by three raters, and IAA shows the inter-annotator-agreement using Krippendorff's alpha.

3.1 Human Annotations

We multi-annotate our samples and obtain a good level of agreement: Three workers annotate each sample in batches, grouped by style task, with a total of five different workers. Annotators rate, on a 5-point Likert scale, how well the style or attribute change is achieved, and how well the meaning/content unrelated to the style/attribute change is preserved, following previous work (Mir et al., 2019; Ziegenbein et al., 2024). We use the scale

```
1: Very poor, 2: Poor, 3: Fair,
   4: Good, 5: Very Good.
```

We achieve a high level of human agreement on content preservation with a Krippendorff's Alpha (Krippendorff, 2011) of 0.768, and a good level of success in style transfer, with an average score on the 5-point Likert scale of 4.27. Detailed results per task in Table 1, where we report the percentage of samples that at least obtain an average rating of 3 (fair). Most sentences successfully transfer style.

The workers are recruited via a crowdsourcing platform (prolific.com). The five different workers who participate are from the UK, hold a BA in Arts and are experienced on the platform. Workers are paid a fixed amount per batch at a rate, which the platform considers a 'great' hourly pay. Details on annotation guideline, setup and payment are provided in App. C.

4 Methodology: LogProb

We propose a novel approach for evaluating style transfer with the following properties: 1) no training data required on style; 2) no need for gold truth / reference sentences; 3) improved feasibility in the form of model size (e.g. 1B,3B) with improved performance compared to our baseline of prompting an LLM-as-Autorater. For content preservation, we additionally aim for: 4) accounting for changes in target style – ensuring theoretical suitability as discussed in Section 2. For style strength, beyond properties 1,2 and 3, we aim for: 5) scoring on a range rather than a binary score – motivated by the fact that styles can shift to various degrees as discussed in Section 2 and recommended by Briakou et al. (2021a).

Similar to the approach in Jia et al. (2023), who evaluate the faithfulness of summaries, we use token likelihood estimates to evaluate a style transfered sentence. The idea in our work is to provide *different* rewrite instructions as part of the context *before* the rewrite sentence, e.g. with and without the target style. See Table 2 for specific implementation. This yields token likelihood estimates that vary based on whether a style change is expected. We use these estimates to measure both style strength and content preservation. Using the likelihood estimates of smaller models as opposed to generated answers offers more robustness, since prompting may result in inconsistent scores depending on the prompt, as well as inconsistent adherence to output formats, complicating postprocessing. Accessing the likelihoods avoids these issues.

Let X be the source sentence consisting of tokens $x_1, x_2, \ldots, x_m$, and let $\tilde{X} = \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ be the rewritten sentence that we want to evaluate with respect to some style change S. From an LLM, we estimate the probability of the output sentence when the model has seen the source sentence and some rewrite instruction as part of the context T, denoted as $P^{LM}(\tilde{X}|X, T)$.

Content Preservation Our requirement is to measure content preservation with respect to the style change, such that the content unrelated to the style should be preserved. Note that content preservation could also be high without a style change. Hence, the measure of content preservation should be high if one of the following conditions is likely: 1) the tokens of $\tilde{X}$ are likely a rewrite with respect to the style S; or 2) the tokens of $\tilde{X}$ are likely a paraphrase of X; or 3) the special case that the tokens of $\tilde{X}$ are likely a repetition of X. This implies that content *is not* well preserved if the likelihood of a token is low in all three cases. Hence, if a token can neither be attributed to repetition, paraphrasing, nor requested style change, then there must be a change in the text that does not preserve the context. To weigh these cases with tokens *not* preserving content, we take the log of the most likely token of our three cases to define our measure:

$$C = \frac{1}{n}\sum_{i=1}^{n} \log\left(\max\left(p_i^{t^s}, p_i^{t^{pa}}, p_i^{t^r}\right)\right) \quad (1)$$

$$p_i^{t^s} = p^{LM}(x_i|X, \tilde{x}_{<i}, t^s),$$
$$p_i^{t^{pa}} = p^{LM}(x_i|X, \tilde{x}_{<i}, t^{pa}),$$
$$p_i^{t^r} = p^{LM}(x_i|X, \tilde{x}_{<i}, t^r)$$

and t^s,t^{pa},t^r being instructions on rewrite with respect to style, paraphrase and repeat.

Style To assess how well style is transferred, we compare: 1) the likelihood of tokens in $\tilde{X}$ after seeing a context containing an instruction to rewrite to the target style and 2) the likelihood after seeing a context containing an instruction to paraphrase or repeat, without reference to style. The hypothesis is that tokens contributing to the target style transfer will have a higher likelihood when the context prior emphasises rewriting in this style. Thus, we measure style transfer success as the difference in token likelihood given different context instructions in 'style rewrite' and 'paraphrase/repetition'.

$$S = \frac{1}{n}\sum_{i=1}^{n} \left(p_i^{t^s} - \max\left(p_i^{t^{pa}}, p_i^{t^r}\right)\right) \quad (2)$$

We deploy three backbone model sizes—1B,3B and 8B parameters—using Llama 3 Instruct models (Dubey et al., 2024).[5] We deploy different instructions as part of the context as per Table 2. Further details in App. D.

5 Benchmarking

Many metrics are used to evaluate content preservation for style transfer. Here, we evaluate several recently and widely used ones (per Section 2). We group content preservation metrics into four categories: lexical similarity, semantic similarity,

[5] META-LLAMA/LLAMA-3.2-3B-INSTRUCT,META-LLAMA/LLAMA-3.2-1B-INSTRUCT and META-LLAMA/LLAMA-3.1-8B-INSTRUCT.

Contexts
X, t^{s}: *'Rewrite the following sentence to be S: X'*
X, t^{pa}: *'Paraphrase the following sentence: X'*
X, t^{r}: *'Repeat the following sentence: X'*

Table 2: Context with different instructions on the rewrite

factual-based, and style-aware. For the similarity-based metrics, if reference data is available, we test both source-based [S-](output vs. source) and reference-based [R-] (output vs. reference) approaches (For details and prompts, see App. D):

- Lexical similarity: **BLEU** (Papineni et al., 2002) and **Meteor** (Banerjee and Lavie, 2005) – both are n-gram match metrics originally developed for evaluating machine translation.
- Semantic similarity: token-based **BERTScore** (Zhang et al., 2019); **Cosine** similarity between sentence embeddings (Feng et al., 2022); learned BERT-based **BLEURT** (Sellam et al., 2020). **COMET** (Li et al., 2024), which uses both source and references to evaluate; however, COMET also have a reference-free version.
- Fact-based: **QuestEval** (Scialom et al., 2021a): uses question generation and question answering to evaluate answers on source and output; originally for factual consistency and relevance in summaries, later extended to faithfulness in simplification tasks (Scialom et al., 2021b).
- Style-aware: **LLM-as-Autorater**: Llama3 Instruct (Dubey et al., 2024) as an evaluator: prompting for a score on a 5-point scale given source sentence, style and output sentence for both content-preservation and style strength (70b parameter and 3b,8b); **LogProb**: our method using likelihood estimates, per Section 4.

We assess evaluation approaches on style strength using **LLM-as-Autorater** and our method **LogProb**, both of which as zero-shot approaches.

Datasets We benchmark nine metrics for evaluating content preservation in style transfer on seven different datasets covering nine styles. We use existing datasets already containing human evaluations of style strength and/or content preservation. The human evaluations are conducted on rewrites from system-generated output and/or human-written / gold reference output, using various scales (e.g., 5-point Likert or 1–100). These datasets cover rewriting tasks for simplifying, sentiment, formality, and making arguments more appropriate. The following datasets are used (abbreviation in brackets), in addition to our newly test set (Section 3) – further details in App. A:

- Lai et al. (2022) **[Lai]** formal/informal,
- Mir et al. (2019) **[Mir]** positive/negative,
- Alva-Manchego et al. (2020) **[Alv.]** simplifying,
- Scialom et al. (2021a) **[Sci.]** simplifying,
- Ziegenbein et al. (2024) **[Zie.]** appropriated arguments,
- Cao et al. (2020) **[Cao]** layman/expert.

We report Spearman rank correlation between metric scores and mean human judgement. We group test data as rewrites generated by systems, human-written gold-references, and our constructed test set. We summarise results in each group by reporting the average Spearman correlation across data sources (Avg), and the average ranking of the metric per data source by correlation score (Avg. rank).

6 Results

Meta-evaluation is highly impacted by the type of test samples. Overall metric performance (correlation with human judgements) varies greatly depending on whether the human-rated outputs are system-generated, human-written, or from our newly constructed test set (Fig. 2). **On system-generated data** (Table 3): Many metrics show relatively high correlations across tasks. For all similarity-based metrics except COMET, the versions using source sentences [S-] result in higher correlations to human ratings than the versions using references [R-]. The best metrics overall are S-BertScore and S-BLEURT (avg+avg rank). **On gold reference data** (Table 3): Here, LLM-as-Autorater 70B performs best (avg+avg rank), but semantic-similarity based metrics still perform second best (S-BertScore and S-BLEURT). The ranking of best metrics changes when the underlying test data with human ratings originates from human-written/gold reference outputs or system-generated outputs. This is also discussed by Scialom et al. (2021a) for simplicity transfer. This is likely due to the nature of the system-generated output used, where verbatim repetitions happen to correlate with human judgments of content preservation. **Our test set** (Table 4): Several metrics show low or no correlation, or even significant negative correlation for

	System-Generated ouput							Gold-reference output				
	Mir	**Lai**	**Zei.**	**Cao.**	**Alv.**	**avg**	**avg rank**	**Lai**	**Zei**	**Sci.**	**avg.**	**avg. rank**
S-Bleu	0.51	0.52	0.63	0.58	0.66	0.58	6.4	0.44	0.38	0.17*	0.33	9
R-Bleu	–	0.2	0.26	–	–	–						
S-Meteor	0.49	0.48	**0.65**	0.59	0.64	0.57	6.9	0.45	0.44	0.15*	0.35	8.17
R-Meteor	–	0.24	0.24	–	–	–						
S-BertScore	**0.52**	**0.67**	0.74	0.65	0.81	0.68	**1.6**	0.54	0.44	0.31	0.43	4.8
R-BertScore	–	0.36	0.38	–	–	–						
S-Bleurt	0.51	**0.67**	**0.65**	**0.71**	**0.86**	**0.68**	2.0	0.48	0.52	0.41	0.47	3.7
R-Bleurt	–	0.4	-0.05*	–	–	–						
S-Cosine	**0.52**	0.57	**0.65**	0.61	0.74	0.62	4.0	**0.61**	0.48	0.26	0.45	3.5
R-Cosine	–	0.27	0.27	–	–	–						
S-Comet	0.21	0.31	-0.13	0.04	0.31	0.15	12.8	0.01*	0.19	0.09*	0.10	12.0
RS-Comet	–	0.46	0.25	–	–	–						
QuestEval	0.26	0.2	0.49	0.48	0.61	0.41	11.3	0.25	0.35	0.01*	0.20	11.7
LLM-as-A 70b	0.35	0.66	0.63	0.63	0.75	0.60	4.7	0.56	**0.57**	**0.46**	**0.53**	**1.3**
LLM-as-A 8b	0.3	0.5	0.38	0.54	0.67	0.48	9.5	0.45	0.5	0.33	0.43	4.8
LLM-as-A 3b	0.26	0.44	0.11	0.47	0.45	0.35	12.3	0.33	0.15	-0.21*	0.09	11.7
LogProb 8b	0.48	0.64	0.57	0.56	0.74	0.60	6.8	0.49	0.48	0.18*	0.38	5.5
LogProb 3b	0.49	0.65	0.58	0.56	0.73	0.60	6.3	0.51	0.47	0.15*	0.38	6.3
LogProb 1b	0.5	0.65	0.56	0.54	0.71	0.59	7.0	0.51	0.42	0.12*	0.35	7.8

Table 3: Spearman rank correlation between metric and human judgment split on data from system-generated or gold-reference rewrites. 'R-': reference-based, 'S-': source-based. Average correlation across datasets [avg.], average rank of metrics per dataset [avg. rank]. *not significant (level 0.05). Best metric per dataset in bold; second best underlined.

	ALL	**Sentiment**	**Detoxify**	**Catchy**	**Polite**	**Persuasive**	**formal**	**avg rank**
S-Bleu	-0.13	-0.79	-0.51	-0.14*	-0.07*	-0.15*	-0.1*	12.7
R-Bleu		-0.49	-0.51	–	–	–	–	
S-Meteor	-0.05*	-0.6	-0.44	-0.02*	0.06*	-0.13*	-0.03*	11.6
R-Meteor		-0.18*	-0.43	–	–	–	–	
S-BertScore	0.11	-0.44	-0.19*	-0.22*	0.35	0.05*	0.26*	9.3
R-BertScore	-0.15*	-0.33	–	–	–	–		
S-Bleurt	0.29	0.06*	0.37	-0.02*	0.33	0.22	0.46	7.3
R-Bleurt		0.19*	0.29	–	–	–	–	
S-Cosine	0.00*	-0.5	-0.3	-0.23*	0.1*	-0.07*	0.17*	11.0
R-Cosine		-0.17*	-0.28*	–	–	–	–	
S-Comet	0.29	0.4	0.44	0.29	0.17*	0.33	0.17*	6.8
RS-Comet		-0.17*	0.01*	–	–	–	–	
QuestEval	0.22	0.3	0.26*	0.36	0.19*	0.32	0.11*	6.6
LLM-as-A. 70b	**0.78**	**0.72**	**0.69**	**0.81**	**0.82**	**0.82**	**0.74**	**1.0**
LLM-as-A. 8b	0.67	0.41	0.64	0.75	0.64	0.67	0.7	2.5
LLM-as-A. 3b	0.31	0.29	0.24*	0.45	0.19*	0.39	0.34	6.3
LogProb 8b	0.63	0.12*	0.49	0.72	0.7	0.51	0.72	3.2
LogProb 3b	0.54	-0.08*	0.32	0.59	0.6	0.38	0.71	5.0
LogProb 1b	0.37	-0.18*	0.11*	0.32	0.51	0.23	0.52	7.2

Table 4: Spearman correlation between metrics and human judgment on **our testset**. 'R-': reference-based, 'S-': source-based. Correlation on entire testset [ALL], average rank per dataset [avg. rank]. *not significant (cannot reject null hypothesis of zero correlation) level 0.05. Bold best metric per dataset; second best underlined.

	Mir	**Lai**	**Zei.**	**Alv.**	**Sci.**	**avg.**
LLM-as-A. 70b	0.45	**0.6**	0.28	**0.6**	0.39	0.46
LLM-as-A 8b	0.34	0.55	0.18	0.32	0.2*	0.32
LLM-as-A. 3b	0.29	0.37	0.14	0.23	0.01*	0.21
LogProb 8b	**0.5**	0.28	**0.32**	-0.02*	**0.5**	0.32
LogProb 3b	0.46	0.28	0.29	-0.01*	0.45	0.29
LogProb 1b	0.41	0.31	0.27	-0.03*	0.48	0.29

Table 5: Style strength: Spearman correlation between metrics and human ratings. *not significant (level 0.05)

some metrics on the entire test set (BLEU) or on parts (Meteor, BertScore, Cosine).

Results on our test set show that similarity-based metrics for content preservation are unsuitable: Similarity-based metrics are deemed unsuitable for content preservation in style transfer by prior work (Section 2), and we can now support this claim by assessing the correlation of these metrics to human judgments on our test set. On our test set, similarity-based metrics show low or negative correlations with human judgment. We conclude that the high correlation scores previously reported on similarity-based metrics obtained on existing dataset are misleading. We hypothesise that misleading conclusions from prior meta-evaluations stem from the nature of the source data used: system-generated outputs happen to be very similar to source sentences lexically and semantically, whereas human-written/gold reference outputs do not, or at least rarely, contain samples where content not related to the style shift is poorly preserved or even fabricated – both limiting the cases where the metrics ability are tested.

We recommend that meta-evaluation efforts for content preservation go beyond measuring correlations with human judgments, and also account for the data source bias underlying those judgments. Specifically, content preservation should be benchmarked on diverse data sources – including those with intentionally low content preservation, such as our test set.

Content preservation: Content preservation metrics must be style-aware. Across all tasks and data sources, overall, the highest correlations between metric and human judgement are achieved by using LLM-as-Autorater using 70B Llama 3 Instruct. If compute is a constraint, we recommend the LogProb method with a smaller model. Both model sizes of 3B and 1B outperform 3B and 1B LLM-as-Autorater [6] (Figure 2). For a backbone model size of 8b, LLM-as-Autorater is recommended.

Style: LLM-as-Autorater 70B shows the best average performance in terms of correlations with human judgements. Still, our more efficient LogProb method scores higher than LLM-as-Autorater 70B on 3 out of 5 datasets, despite using a much smaller model (both 8B and 3B versions) (Table 5). The variation in results across datasets calls for a task-specific investigation of the best metrics.

[6]The LLM-as-Autorater 1b result is not reported as for over 50% of the cases, the output was not compliant with the expected output format

7 Conclusion

We conduct a large-scale meta-evaluation of metrics for style transfer, focusing on content preservation. We construct a new challenging test set for assessing metrics for content preservation with human judgment. We show that the meta-evaluation of metrics for content preservation is not straightforward. Previous meta-evaluation studies on existing datasets consisting of either automatically generated data from specific systems or human-generated data alike leads to misleading conclusions about best metrics, due to skewed test cases. We hypothesise that on existing system-generated data, human judgment on content preservation happens to correlate with verbatim repetitions, and on gold references/human-written data, the limits of the metrics are not tested, as cases of drastically low preserved or fabricated content are not or rarely present in the data. With our new test set, deliberately designed with variation in content preservation, we demonstrate empirically that similarity-based metrics are not suitable for content preservation for style transfer. Instead, metrics for content preservation should be style-aware. We recommend for meta-evaluation that a diverse set of data sources be used, including both system-generated test cases with human ratings, as well as our newly constructed test set, which tests the nature of content preservation metrics.

We propose a new efficient style-aware metric for content preservation and one for style strength, reusing the same computational effort (LogProb).

Overall, we find that a large LLM-as-Autorater (70B) achieves the highest correlation between content preservation scores and human judgments. However, if considering computational efficiency or feasibility in the form of a smaller model, then our method LogProb outperforms a LLM-as-Autorater baseline using a 3B parameter LLM.

8 Limitations

In evaluating style transfer, many different metrics are used for content preservation, leading to a need for standardization (Ostheimer et al., 2023). While there are more metrics we could have tested, we do test the more widely used ones, as well as metrics of different types: lexical similarity, semantic similarity, fact-based, and LLMs conditioned on style shift (style-aware). Especially for testing LLMs

within the LLM-as-a-Judge paradigm, our work is limited to testing one prompting approach, as our main focus has been to establish better meta-evaluation practices regarding content preservation in style transfer. We refer to a survey of different trends in the LLM-as-Judge paradigm in (Li et al., 2024). We provide general recommendations on metrics over a large variety of rewriting tasks; however, specific tasks may have specific requirements that are not covered in this paper. In this paper, we have not covered the evaluation of 'fluency' in rewrites, but mainly focused in-depth on content preservation and partly on style strength.

Acknowledgements

 This work was supported by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516). It was further supported by the European Union (ERC, ExplainYourself, 101077481), and by the Pioneer Centre for AI, DNRF grant number P1.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021a. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault, and Marine Carpuat. 2021b. A review of human evaluation for style transfer. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 58–67, Online. Association for Computational Linguistics.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. 2023. STEER: Unified style transfer with expert reinforcement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7546–7562, Singapore. Association for Computational Linguistics.

Jingxuan Han, Quan Wang, Zikang Guo, Benfeng Xu, Licheng Zhang, and Zhendong Mao. 2024. Disentangled learning with synthetic parallel data for text style transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15187–15201, Bangkok, Thailand. Association for Computational Linguistics.

Qi Jia, Siyu Ren, Yizhu Liu, and Kenny Zhu. 2023. Zero-shot faithfulness evaluation for text summarization with foundation language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11017–11031, Singapore. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 102–115, Dublin, Ireland. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multidimensional evaluation for text style transfer using chatgpt. *arXiv preprint arXiv:2304.13462*.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Qingyi Liu, Jinghui Qin, Wenxuan Ye, Hao Mou, Yuxuan He, and Keze Wang. 2024. Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38:17, pages 18689–18697.

Varvara Logacheva, Daryna Dementieva, Irina Krotova, Alena Fenogenova, Irina Nikishina, Tatiana Shavrina, and Alexander Panchenko. 2022a. A study on manual and automatic evaluation for text style transfer: The case of detoxification. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 90–101, Dublin, Ireland. Association for Computational Linguistics.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022b. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Guoqing Luo, Yu Han, Lili Mou, and Mauajama Firdaus. 2023. Prompt-based editing for text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5740–5750, Singapore. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. 2024. Are large language models actually good at text style transfer? In *Proceedings of the 17th International Natural Language Generation Conference*, pages 523–539, Tokyo, Japan. Association for Computational Linguistics.

Phil Ostheimer, Mayank Kumar Nagda, Marius Kloft, and Sophie Fellenz. 2023. A call for standardization and validation of text style transfer evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10791–10815, Toronto, Canada. Association for Computational Linguistics.

Phil Sidney Ostheimer, Mayank Kumar Nagda, Marius Kloft, and Sophie Fellenz. 2024. Text style transfer evaluation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15802–15822, Torino, Italia. ELRA and ICCL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. Measuring and Benchmarking Large Language Models' Capabilities to Generate Persuasive Language. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10056–10075, Albuquerque, New Mexico. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021a. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Louis Martin, Jacopo Staiano, Eric Villemonte de La Clergerie, and Benoît Sagot. 2021b. Rethinking automatic evaluation in sentence simplification. *arXiv preprint arXiv:2104.07560*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Ping Yu, Yang Zhao, Chunyuan Li, and Changyou Chen. 2021. Rethinking sentiment style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1569–1582, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Biqing Zeng, Junjie Liang, Yining Hua, Ruizhe Li, Huimin Deng, Yihao Peng, and Ruitang Wang. 2024. The bat: Thoughts hierarchical enhancement beyond arbitrary text style transfer. In *International Conference on Intelligent Computing*, pages 376–388. Springer.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. Distilling text style transfer with self-explanation from LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211, Mexico City, Mexico. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Timon Ziegenbein, Gabriella Skitalinskaya, Alireza Bayat Makou, and Henning Wachsmuth. 2024. LLM-based rewriting of inappropriate argumentation using reinforcement learning from machine feedback. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4455–4476, Bangkok, Thailand. Association for Computational Linguistics.

A Data sets

In table 6, we list the stats of datasets with human annotations used for benchmarking in this paper.

The data from Mir et al. (2019) [Mir] is on the Yelp sentiment task and is annotated by 3 workers, where the mean ratings are released. Data is downloaded from `github.com/passeul/style-transfer-model-evaluation`. No licence.

The data from Lai et al. (2022) [Lai] supplied human annotations on system output on the formality task, using a continuous scale from 1-100. Download at `https://github.com/laihuiyuan/eval-formality-transfer` with MIT License.

The data from Scialom et al. (2021b) [Sci] is on human ratings for human-written output for a simplification sentence task. It complements the dataset from Alva-Manchego et al. (2020). Download using the URL in the paper. No license specified. We have filtered this data to obtain annotation in all three dimensions for the same data input (we check for an exact match on source sentence, rewrite, sentenceID), and we ended up with 65 samples annotated by 25 workers.

The data from Alva-Manchego et al. (2020) [Alv.] is system output on a simplification task. We use the resource released along with Scialom et al. (2021b), as it contains more metadata, such as system information and more annotations. We filter the data such that we have 135 samples with 11 annotations in all three dimensions because we favour more samples over the number of annotations per sample.

The data from Ziegenbein et al. (2024) [Zie.] is on rewriting inappropriate arguments to appropriate, download available at `https://github.com/timonziegenbein/inappropriateness-mitigation`.

The data from Cao et al. (2020) [Cao] is a human evaluation on a task of transferring different styles of expertise in the medical domain. The authors have kindly shared the data with human ratings.

B LLM-generated Test Part

A part of our construct test set is synthetically generated using GPT-4o-mini from OpenAI. We display the prompts for obtaining the samples from our stepwise approach; we show it for the subpart of politeness:

Generating source data prompt =*'Please give me* {number} *examples of impolite sentences from emails, and return only in json format with key "sentences"'*

Generate rewrites

1) *"Please rewrite the following* {number} *sentences to be very polite, return in JSON with*

Abb.	Style	Dim	Support	#Ann.	#Sys.	Ref.	Rating on ref.	Scale
Lai	formal, informal	S,C,F	640	2	8	✓	✓	1-100
Mir	positive, negative	S,C	2928	1*	12			1-5
Alva-M.	simplifying	S,C,F	135	11	6			0-100
Scialom	simplifying	S,C,F	65	25	1		✓	0-100
Ziegen.	appropriated arguments	S,C,F	1350	5	6	✓	✓	1-5
Cao	layman, expert	C	3800	1	5			1-5

Table 6: Stats on the dataset used for benchmarking style transfer metrics: **Dim**ensions which of (**S**tyle, **C**ontent preservation, **F**luency) are rated in the data, **Support**: numbers of rated samples, #**Ann**otators: number of annotations per sample, #**Sys**stems: number of different systems/settings (including references) used to produce the samples, **Ref**erence: is the data supplied with references to enable reference-based evaluation, **Rating on ref**erence: does the data have ratings on references. *Mir dataset is conducted with multiple annotators, but only the mean of the annotations is released.

key 'sentences:' {list of sentences}"

2) *"Please rewrite the following* {number} *sentences to be just a bit more polite, return in JSON with key 'sentences:'* {list of sentences}"

Generate one content error in the 2. rewrite

a) "*Please rewrite the following* {number} *sentences staying as close to the original wording as possible but make a mistake in the content by substituting some key information, return in json with key 'sentences:'* {list of sentences}

b) *"Please rewrite the following* {number} *sentences staying as close to the original wording as possible but make a mistake in the content by omitting some key information, return in json with key 'sentences:'* {list of sentences}

c) *"Please rewrite the following* {number} *sentences staying as close to the original wording as possible but make a mistake by adding a very short extra detail, return in json with key 'sentences:'* {list of sentences}

Automatic assessment of 'linguistic acceptability' To assess the quality of the LLM-generated test samples compared to the manually created ones, we assess the 'linguistic acceptability' of the samples. We use a pretrained binary classifier from HuggingFace.com [textattack/roberta-base-CoLA], trained on the CoLA dataset on grammatical acceptability (Warstadt et al., 2019). We report the percentage of the samples predicted as 'linguistically acceptable' grouped by LLM-generated and manually created sampels. We observe that most samples are predicted as grammatically acceptable, Table 7.

	Manually created	LLM-generated
source	94 %	93.5 %
rewrites	80 %	98.75 %

Table 7: Percentage of samples predicted as lingusitic acceptable

C Annotation Guide and Process

C.1 Annotation Procedure

We recruit annotators through the crowd-sourced platform prolific.com. We use workers who we have experienced delivering high-quality in a previous annotation study. In total, we use five different workers, all from the UK and holding a BA in Arts. The workers are paid a fixed amount per task, which Prolific considers a 'great' hourly pay. Considering the completion times, the average hourly pay to the workers was 16.2 GPB.

One of our subtasks involves detoxifying toxic content; we warn the workers of potentially disturbing content before they select the task.

Full annotation guidelines are below, and a screenshot of the annotation interface with a sample is in Figure 3. We use Google Forms as our annotation tool.

C.2 Annotation Guideline

Evaluating Rewrites to Change Style/Attribute

In this task, your goal is to help us evaluate sentences that have been paraphrased or rewritten to

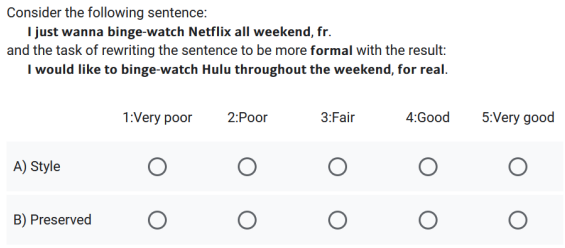

Figure 3: Screenshot of annotation tool, Google Forms

modify a specific style or attribute. These attributes may include tone, formality, positivity, politeness, or personalisation, among others. For example, a sentence might be rewritten to add a positive sentiment or to simplify the text for a younger audience.

When a text is rewritten to modify its style or attribute, it is important that the original content unrelated to the intended change remains intact. For example, no important information should be dropped, new information should not be fabricated, and the rewritten sentence should not mix up or misrepresent facts—except where necessary to achieve the requested change.

We will provide you with an original sentence and a rewritten version, along with the intended style or attribute change. Note that the style/attribute may vary across different examples during the study, so please pay attention to the provided description for each pair.

For each sentence pair, you will answer two evaluation questions using a 5-point Likert scale from 'Very poor' to 'Very good':

B) Style/Attribute Change: Evaluate how well the intended style or attribute change is achieved in the rewritten sentence.

A) Content Preservation: Evaluate how well the meaning/content unrelated to the style or attribute change is preserved in the rewritten sentence.

The study will provide you with a link to a Google Form, where there will be 100 samples to evaluate. The study is estimated to take 35 minutes.

The annotation will be part of a research project for my PhD.

Thank you for considering participating.

D Methods Implementation

D.1 Our method: LogProb

We deploy the backbone models META-LLAMA/LLAMA-3.2-1B-INSTRUCT and META-LLAMA/LLAMA-3.2-3B-INSTRUCT and META-LLAMA/LLAMA-3.1-8B-INSTRUCT downloaded from https://huggingface.co/meta-llama.

We set the system prompt to *"You can repeat sentences, paraphrase sentences or rewrite sentences to change the style or certain attribute of the text while preserving non-related content and context. Your answers contain just the rewrite."*.

We run Python 3.13.2, and use the Transformer (4.51.3) library from Huggingface https://huggingface.co/docs/transformers/index. We have conducted the experiments on a machine with the following characteristics:

Intel Xeon Silver 4410T (40) @ 2.700GHz
NVIDIA GeForce RTX 4090 2 STK
252GB RAM,

running Ubuntu 22.04.5 LTS. Using this machine with one GPU, the average inference time to evaluate one sample is:

- **0.025** seconds per sample using the **1B** backbone model
- **0.041** seconds per sample using the **3B** backbone model
- **0.070** seconds per sample using the **8B** backbone model

calculated from time to evaluate our test set of 500 samples.

D.2 LLM-as-Autorater

We deploy LLM-as-Autorater using both the 70B parameter model and 8B,3B, and 1B versions of the Llama 3 Instruct models.

For the 70B version, we use API calls to replicate.com, and the model META-LLAMA-3-70B-INSTRUCT (replicate.com/meta/meta-llama-3-70b-instruct). We use the hyperparameter settings "temperature": 0.5, "top_p": 0.9, else default.

For the 1B,3B and 8B versions, we use the same machine as described in Section D.1. We set hyperparameters in order to lower computation at inference time: "temperature"=1.0,"do_sample"=False, "max_new_tokens" =20. The 1B version did not produce results that were compliant with the requested output format. The inference times for the 3B and 8B versions are:

- **0.20** seconds per sample using the **3B** backbone model

- **0.37** seconds per sample using the **8B** backbone model

calculated from time to evaluate our test set of 500 samples.

We use the following prompt for the 70b version:

- *"Evaluate the following completion of a task where a 'source sentence' has been rewritten to be more* {style} i*n the style, denoted 'target sentence', Ideally the context and content in the sentence which does not relate to the style should be preserved. Please evaluate on a Likert scale from 1-5 with 5 being the best: 1) how well the meaning is preserved and 2) how well the style is changed. Return in JSON format with the keys 'meaning' , 'style'. Given the 'source sentence':* {source sentence} *'target sentence':* {rewrite}"

On the 8b and 3b backbone model, we add an 'only' in the sentence 'Return only JSON format...'. It occurs rarely that the LLM does not provide an answer in the right format. In these cases, we provide the mean rating of the datasets. The number of occurrences where this happens is 0.3 % for the 70b, 0 % for the 8b, and 0.12 % for the 3b backbone model.

D.3 Metrics

BLEU (Papineni et al., 2002) we use the python package NLTK (3.9.1) implementations of BLEU with default settings.

Meteor (Banerjee and Lavie, 2005) we use the python package from Huggingface Evaluate (0.4.3) with default settings.

BertScore (Zhang et al., 2019) We use the implementation from https://github.com/Tiiiger/bert_score (MIT license) with the current recommended backbone model MICROSOFT/DEBERTA-XLARGE-MNLI.

BLEURT (Sellam et al., 2020) we use the python implemention from https://huggingface.co/Elron/bleurt-large-512 using Huggingface Transformer libary with the backbone model ELRON/BLEURT-LARGE-512.

Cosine similarity embeddings we use the SentenceTransformer (2.7.0) library with Labse embeddings SENTENCE-TRANSFORMERS/LABSE, (Feng et al., 2022).

QuestEval we use the implementations from https://github.com/ThomasScialom/QuestEval (MIT license).

	Lai	Alva-M.	Scia.	Zieg.
	spear.	spear.	spear.	spear.
ppl	0.45	0.36	0.24*	**0.23**
clf_cola	**0.52**	**0.43**	**0.49**	0.20

Table 8: Spearman's Rank correlation between human judgement and a method's predictions. * not significant.

COMET (Rei et al., 2020) we use the python package from Huggingface Evaluate (0.4.3) with default settings for the reference and source version [RS-COMET]. We use the guide and Python implementation from https://github.com/Unbabel/COMET with the recommended model [Unbabel/wmt22-cometkiwi-da] for reference-free evaluation.

E Fluency

The dimension of fluency is not covered in this work of meta-evaluating metrics for style transfer, as the main focus has been on content preservation and the discrepancy between what are suitable metrics and what the empirical results show. Furthermore, one of the main strengths of LLMs is their ability to generate very fluent sentences. However, we include a small assessment of fluency using the datasets, which has annotations on this dimension. We deploy two commonly used methods: perplexity using GPT2 (Radford et al., 2019) and classification on grammatical acceptability. To calculate perplexity using GPT2 (ppl), we utilise the Huggingface Library and the model available on [openai-community/gpt2]. To predict grammatical acceptability (clf_cola), we use a pretrained binary classifier from HuggingFace.com [textattack/roberta-base-CoLA], trained on the CoLA dataset on grammatical acceptability (Warstadt et al., 2019). Spearman's Rank correlation between the methods and human judgments is reported in Table 8. In general, we see the highest correlation with the Cola classifier to human judgment on fluency.

F Lincense

Our test set consists of sentences from Wiki News megarhyme.com/blog/wikinews-dataset/ and from the work of Logacheva et al. (2022b). These datasets, respectively, has the licenses Creative Commons Attribution 2.5 and CC0 1.0 Universal. We release our work (both code and dataset) under Creative Commons Attribution 4.0.