ReproHum #0669-08: Reproducing Sentiment Transfer Evaluation

Kristýna Onderková, Mateusz Lango, Patrícia Schmidtová and Ondřej Dušek
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{onderkova,lango,schmidtova,odusek}@ufal.mff.cuni.cz

Abstract

We describe a reproduction of a human annotation experiment that was performed to evaluate the effectiveness of text style transfer systems (Reif et al., 2022). Despite our efforts to closely imitate the conditions of the original study, the results obtained differ significantly from those in the original study. We perform a statistical analysis of the results obtained, discuss the sources of these discrepancies in the study design, and quantify reproducibility. The reproduction followed the common approach to reproduction adopted by the ReproHum project (Belz et al., 2025).

1 Introduction

Human evaluation is considered to be the gold standard for assessing natural language processing (NLP) systems, although many factors can affect its reliability. Subjectivity in human ratings can make experiments difficult to reproduce (Belz et al., 2021); the definitions of the evaluated criteria are often inconsistent (Howcroft et al., 2020) and may confuse annotators (Hosking et al., 2024). Furthermore, external factors such as interface design can bias annotator behavior in unexpected ways (Calò et al., 2025). In some cases, issues such as unclear instructions, inappropriate dropping of outliers, or overlooked implementation bugs are only revealed during reproduction (Thomson et al., 2024). Therefore, efforts such as the ReproHum project (Belz and Thomson, 2023) help us identify these challenges and develop more robust and transparent evaluation practices.

In this report, we describe our reproduction study of human evaluation of sentiment transfer, originally performed by Reif et al. (2022). We focus on a single quality evaluated in the original experiment: semantic preservation, i.e., how much of the original meaning was preserved after performing the sentiment transfer. We also limit our evaluation to a single style: *more positive* (see Section 2).

The original experiment is described in Section 2. We reproduce the setting of the original study as closely as possible and describe this process in Section 3. The results of our human annotation are shown in Section 4. Section 5 describes how we compared key numerical results to assess reproducibility and compares the findings of our reproduction against the original study. Finally, in Section 6 we discuss reasons for differences between the original and reproduced results.

2 Original Experiment

The original study (Reif et al., 2022) presents a zero shot prompting method with large language models (LLMs) for text style transfer. The text style transfer task transforms or adds stylistic attributes to a text while preserving the global structure, e.g. converting “It is a nice day.” to a more positive “It is a truly magnificent day!” (Hu et al., 2017; Prabhumoye et al., 2018). Reif et al. (2022)’s LLM prompting method can perform any arbitrary text transformation (e.g. "more melodramatic") without fine-tuning or presenting specific exemplars in the prompt.

The style is transferred for 50 randomly chosen sentences from the Reddit Writing Prompts validation set (Fan et al., 2018). The sentences are transformed into three standard styles (*more positive*, *more negative*, *more formal*) and six non-standard styles (*more melodramatic*, *more comical*, *include the word “baloon”*, *include the word “park”*, *include a metaphor*, *more descriptive*). The researchers compared the following six systems:

- **human** – ground truth transfers written by the authors of the original study (Reif et al., 2022)
- **zero-shot** – an approach using a base prompt with no examples: “*Here is some text: ... Here is a rewrite of the text, which is more positive:*”

July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

- **augmented zero-shot** – this version of the prompt additionally includes seven exemplars of different style transfers (e.g. more scary, intense, flowery, including "snow"...)

- **paraphrase** – an ablation using a zero-shot prompt which only specifies the target style as "paraphrase": "*Here is a rewrite of the text, which is paraphrased:*"

- **Unsup MT** (Prabhumoye et al., 2018) – an approach using translation into a second language and back to remove stylistic features, coupled with style-specific decoders trained using adversarial techniques.

- **Dual RT** (Luo et al., 2019) – a model for style transfer trained by reinforcement learning with two rewards, one for style accuracy and second for content preservation.

The prompts were executed with the LaMDA and LaMDA-Dialog language models (Thoppilan et al., 2022), as well as GPT-3 (Brown et al., 2020).

As text style classifiers (Wolf et al., 2020; Sudhakar et al., 2019) are not available for all target styles, the researches relied on human evaluation with six professional annotators. The annotators evaluated three aspects on 1-100 scale: **(1) transfer strength** – to what extent the output matches the target style; **(2) semantic preservation** – how well the output preserves the meaning of the input, excluding the style change; **(3) fluency**. To achieve good inter-annotator agreement, the researchers run an initial calibration session where annotators rated a small subset of data (excluded from the main results) and asked clarifying questions about the instructions. Each triple of input-transformation-output was rated by three annotators.

Target styles commonly used in research for style transfer (positive and negative sentiment and formality), where data are available, are also evaluated on the Yelp polarity dataset (Zhang et al., 2015) and Gramarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018). Those are also evaluated with automatic metrics: the HuggingFace Transformers sentiment classifier (Wolf et al., 2020) for transfer strength, semantic similarity to human examples from (Luo et al., 2019) through the BLEU score, and fluency as measured by GPT-2's (Radford et al., 2019) perplexity.

3 Reproduction Study

We reproduced the human annotation of a single style transfer transformation – *more positive* – and one evaluation aspect – *semantic preservation*. We followed the original experiment as closely as possible. However, instead of using internal annotators which are not available to us, we recruited annotators from the Prolific crowdsourcing platform.[1] In effect, we could not perform the initial calibration session. The setup of the reproduction is based on the original study's design and the ReproHum guidelines (Belz et al., 2025):

Datasets We use the same 50 sentences from the Reddit Writing Prompts as the original study, the *more positive* transformation, and the outputs of all the six systems that were compared.

Evaluated quality factors The original annotation included three quality factors described above – transfer strength (dubbed *transferred style strength*), semantic preservation (dubbed *meaning*) and *fluency*. Our reproduction included only the semantic preservation.

Annotation interface The original annotation interface was an internal system of the researchers, which we could not reuse. Therefore, we recreated the interface using Google Apps Script[2] (see Appendix B). The interface shows six system outputs for one input on a page, together with a slider from 0-100% for one rated aspect (*meaning*). One annotator rates 25 inputs as in the original study, each on a different page. Each page includes a collapsible instructions panel for easy reference to the guidelines.

Annotators The annotators were recruited on Prolific by using the following filters:

1. devices: tablet, desktop (no mobile phones);
2. region control: UK, USA, Australia, Canada;
3. number of previous submissions: 200–10000;
4. approval rate: 99–100%.

Remuneration Based on the ReproHum project rules (Belz et al., 2022), the annotators were compensated using the UK living wage of 12.60 GBP per hour.

[1] https://app.prolific.co/

[2] https://developers.google.com/apps-script

	Original	Reproduction	Confidence interval	CV*	Krippendorff's α
Paraphrase	90.29	45.55	(38.64, 52.47)	65.66	0.040
Zero-shot	69.71	49.66	(42.27, 57.06)	33.48	0.087
Unsup. MT	86.76	73.39	(68.58, 78.20)	16.64	-0.129
Dual RL	85.29	68.29	(61.63, 74.95)	22.07	0.077
Augmented zero-shot	86.47	64.99	(58.46, 71.52)	26.78	0.125
Human	85.29	74.76	(69.40, 80.12)	13.11	-0.073
Average	83.97	62.78			

Table 1: The results of our reproduction – average semantic preservation on a 0-100 scale – compared to those from the original study. Additionally, 95% confidence intervals, inter-annotator agreements, and coefficient of variation values are reported.

Annotation guidelines The annotation guidelines are the same as the original ones, with omitted instructions and examples for *transferred style strength* and *fluency*, which were not measured. They can be found in the annotation interface depicted in Appendix B.

4 Main Results

The results of our reproduction are presented in Table 1. For each output, we averaged the annotated meaning preservation values and then computed an overall average for each system. During post-processing of the data, we discovered that the annotations for nine instances had been corrupted when they were saved. Still, all instances had at least one annotation, and this error only had a minimal impact on the overall results, increasing the standard deviation of the reported mean scores by approximately 3.5%.[3]

In the original study, the results were presented as bar plots, which meant that the precise numerical values were not directly available. To enable a comparison with our results, we estimated the original values by measuring the number of pixels between the top of each bar and the end of the scale, then calculating the corresponding proportion relative to the full 0–100 scale (also measured in pixels). Based on this approach, one pixel corresponded to a score of 0.2941 (0.3%), which is the accuracy of our estimation.

The observed differences between the original and reproduced results are significant. Our annotators seem to be more strict when assessing semantic preservation, as the overall average across all the methods is more than 20 percentage points lower than in the original study. All systems received lower scores, with the smallest drop for human-written outputs.

There are also substantial differences in the ranking of the evaluated methods. In the original study, the *paraphrase* method was ranked the highest, while human-written texts were outperformed – or scored the same – by four out of the five automatic methods. In the reproduction, the outputs of *paraphrase* method received the lowest score and humans outperformed all automatic methods. The rest of the systems receive similar ranks in both studies.

Inter-annotator agreement We measured the inter-annotator agreement of obtained annotations with Krippendorff's α (Krippendorff, 2006). To identify potential outliers, we also conducted ablation analyses by recalculating agreement scores after excluding each annotator's ratings in turn. The results are presented in Table 2.

According to (Marzi et al., 2024), the inter-annotator agreement obtained should be interpreted as poor. The original study did not report inter-annotator agreement, leaving it unclear whether our result is due to the lack of the initial annotator calibration session (conducted in the original experiment but omitted in the reproduction) or from the inherent difficulty of the annotation task.

Our ablation analysis in Table 2 revealed that some annotators had lower agreement with the rest. However, excluding none of the annotators exceeds the upper bound of the 95% confidence interval estimated via bootstrapping $(0.0316, 0.1930)$.

Statistical analysis Student's t-tests were performed to compare the meaning preservation scores obtained during reproduction with those obtained in the original study. The tests for all textual transfer methods rejected the null hypothesis that the

[3] Standard deviation of a mean is $\frac{\sigma}{\sqrt{n}}$. The relative increase in deviation caused by a smaller sample is $\frac{\sigma/\sqrt{140}}{\sigma/\sqrt{150}} = \sqrt{\frac{150}{140}} = 1,0351$. The observed differences from the original study are at least three times higher.

	Krippendorff's α
All annotators	0.103
w/o Annotator #1	0.037
w/o Annotator #2	0.189
w/o Annotator #3	0.130
w/o Annotator #4	0.058
w/o Annotator #5	0.172
w/o Annotator #6	0.066

Table 2: Inter-annotator agreement (Krippendorff's α) computed for all annotators as well as for all annotators excluding a selected one.

true mean of the reproduced scores was the same as the original mean. Table 1 shows the 95% confidence intervals for the reproduced scores; all values from the original study are well above our estimated upper bound. The paired Wilcoxon test comparing the ranks obtained by different systems also rejected the null hypothesis with $p = 0.031$.

5 Quantifying Reproducibility

The reproduction targets were determined based on the categories outlined in the ReproHum shared task guidelines (Belz et al., 2023, Sect. A5) and QRA++ (Belz, 2025). The targets in the following categories were identified:

- Type I – numerical scores: the average semantic preservation in texts generated by different text style transfer methods,
- Type II – sets of numerical values: the set of semantic preservation results for all the methods in the study,
- Type IV – findings stated explicitly or implied by quantitative results in the original paper.

Type I Following the quantified reproducibility assessment by Belz et al. (2022), we computed the small sample coefficient of variation (CV*) as a measure of the degree of reproducibility for numerical scores. The results are given in Table 1.

The values of CV* computed for the original study and the reproduction are in the range of 13-33, except for the substantially higher value for style transfer performed by paraphrasing.

Type II results are evaluated with Pearson and Spearman correlation (Huidrom et al., 2022), as well as with the root-mean-square deviations from the original results. The results are presented in Table 3. The values of Pearson and Spearman correlations are low. The statistical significance tests for both correlations, conducted at the standard signficance level $\alpha = 0.05$, were not able to reject the null hypothesis, i.e., that the correlation between the results of the original and the reproduced study is equal to zero.

	value	p-value
Pearson r	0.3063	0.5549
Spearman ρ	-0.2029	0.6998
RMSE	23.9575	-

Table 3: Statistics used to assess reproducibility of Type II results

Finally, the RMSE value of around 24 for a measurement on a scale from 0 to 100 confirms a large discrepancy between the results. It also reflects the general tendency of our annotators to rate meaning preservation lower than in the original study.

Type IV Reif et al. (2022) summarises the findings from the original study as follows: "The outputs from our method were rated comparably to both human-generated responses and the two prior methods". However, these conclusions are not confirmed by our reproduction. As previously mentioned, human-written responses obtained the highest scores, with a difference of 9 percentage points to the approach proposed in Reif et al. (2022). In our study, this approach was outperformed by both baseline methods, but the difference to one of them was relatively small.

6 Discussion

One major difference between the original experiment (Reif et al., 2022) and the reproduction study is that the original experiment performed an annotation calibration procedure on 10 examples. These 10 examples were excluded from the evaluated data and allowed the authors to align their expectations with the annotators, who were free to ask questions during this process. We hypothesize that the absence of this calibration step affected the reproduction, especially since measuring meaning preservation in sentiment transfer is counterintuitive and requires clear guidance for consistent annotation.

Given that the original experiment was conducted in 2021 (i.e., before the introduction of LLMs to the general public), we also cannot rule out the possibility that people have increased their expectations of AI, leading to the lower scores we observed.

Acknowledgements

This research was co-funded by the European Union (ERC, NG-NLG, 101039303) and by Charles University projects GAUK 252986 and SVV 260 698. It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

References

Anya Belz. 2025. Qra++: Quantified reproducibility assessment for common types of results in natural language processing. *Preprint*, arXiv:2505.17043.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anya Belz, Craig Thomson, Javier González-Corbelle, and Malo Ruelle. 2025. The 2025 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM)*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, and 23 others. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Eduardo Calò, Lydia Penkert, and Saad Mahamood. 2025. Lessons from a user experience evaluation of NLP interfaces. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2915–2929, Albuquerque, New Mexico. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Rudali Huidrom, Ondřej Dušek, Zdeněk Kasner, Thiago Castro Ferreira, and Anya Belz. 2022. Two reproductions of a human-assessed comparative evaluation of a semantic error detection system. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 52–61, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Klaus Krippendorff. 2006. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5116–5122, Macao.

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. “transforming” delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279, Hong Kong, China. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in NLP. *Computational Linguistics*, 50(2):795–805.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, and others. 2022. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

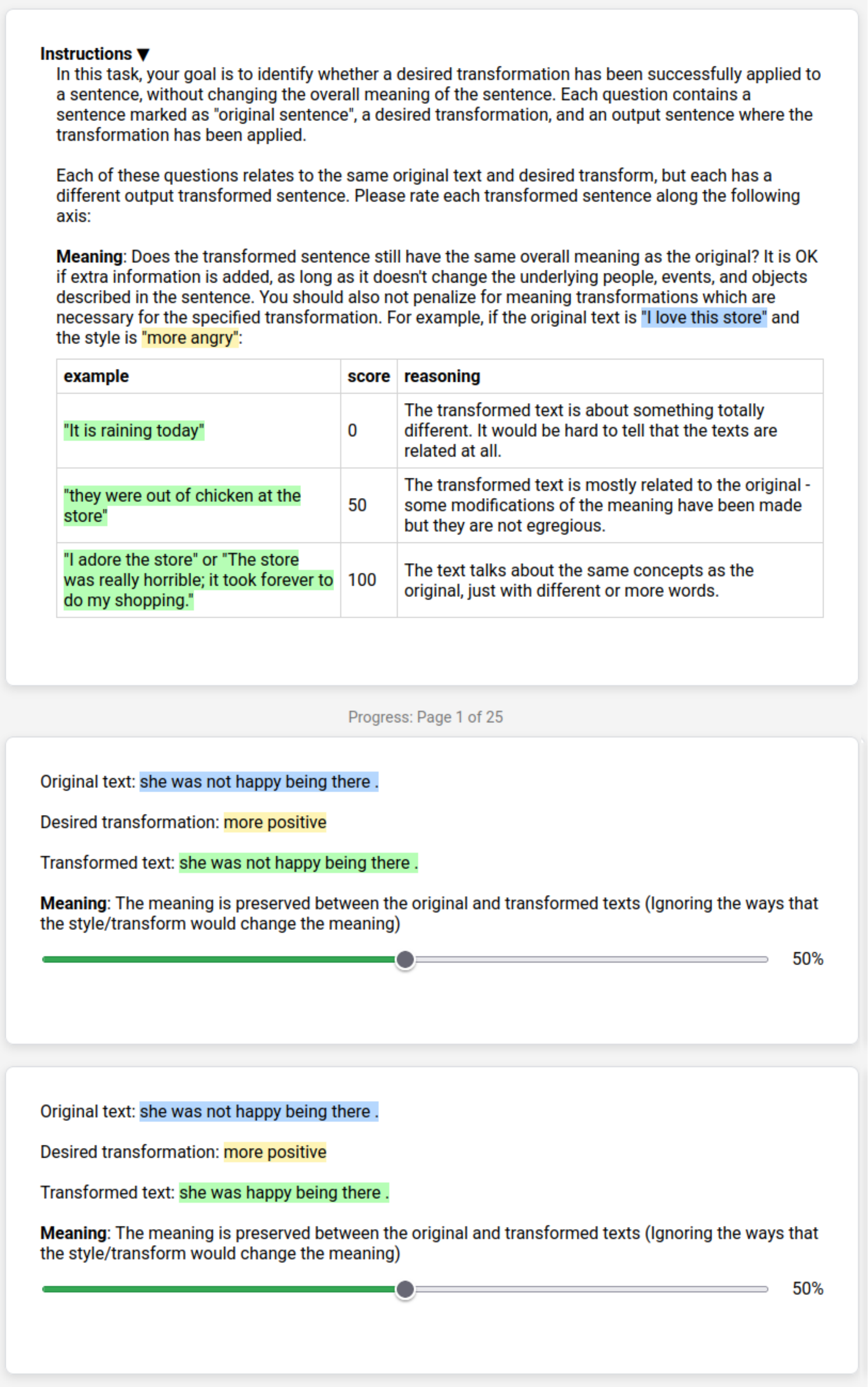
Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Human Evaluation Datasheet (HEDS)

Human Evaluation Datasheet (HEDS, Shimorina and Belz, 2022) for the main ReproHum reproduction (see Sec.) is provided in the ReproHum GitHub repository.[4]

B Annotation Interface

[4] https://github.com/nlp-heds/repronlp2025



Instructions ▼

In this task, your goal is to identify whether a desired transformation has been successfully applied to a sentence, without changing the overall meaning of the sentence. Each question contains a sentence marked as "original sentence", a desired transformation, and an output sentence where the transformation has been applied.

Each of these questions relates to the same original text and desired transform, but each has a different output transformed sentence. Please rate each transformed sentence along the following axis:

Meaning: Does the transformed sentence still have the same overall meaning as the original? It is OK if extra information is added, as long as it doesn't change the underlying people, events, and objects described in the sentence. You should also not penalize for meaning transformations which are necessary for the specified transformation. For example, if the original text is "I love this store" and the style is "more angry":

example	score	reasoning
"It is raining today"	0	The transformed text is about something totally different. It would be hard to tell that the texts are related at all.
"they were out of chicken at the store"	50	The transformed text is mostly related to the original - some modifications of the meaning have been made but they are not egregious.
"I adore the store" or "The store was really horrible; it took forever to do my shopping."	100	The text talks about the same concepts as the original, just with different or more words.

Progress: Page 1 of 25

Original text: she was not happy being there .

Desired transformation: more positive

Transformed text: she was not happy being there .

Meaning: The meaning is preserved between the original and transformed texts (Ignoring the ways that the style/transform would change the meaning)

50%

Original text: she was not happy being there .

Desired transformation: more positive

Transformed text: she was happy being there .

Meaning: The meaning is preserved between the original and transformed texts (Ignoring the ways that the style/transform would change the meaning)

50%

Figure 1: The annotation interface form with instructions