ReproHum #0669-08: Reproducing a Recipe for Arbitrary Text Style Transfer with LLMs

Saad Mahamood
Shopware
Düsseldorf, Germany
saad@saad.me.uk

Abstract

We describe our attempt to reproduce a single human evaluation quality criterion that was conducted in the paper "Reproducing a Recipe for Arbitrary Text Style Transfer with LLMs". This paper describes the approach and challenges involved in reproducing the human evaluation as done by the original authors. In particular, we describe negative results obtained during the reproduction, and we compare our results with an earlier reproduction for the same experiment. Finally, we describe the insights we gained from attempting this particular reproduction and the barriers that remain in attempting successful reproductions. The results and insights presented will hopefully enable the broader NLP research community to improve both how human evaluations are conducted and enable better reproducibility of NLP experiments in the future.

1 Introduction

A significant challenge in Natural Language Processing (NLP) is the difficulty in reproducing human evaluations. Initiatives like ReproNLP[1] are investigating the reasons behind this lack of reproducibility, which is particularly concerning because human evaluations are often considered the best measure of NLP system performance (Belz and Reiter, 2006). However, there are significant barriers to replicate previously reported results due to technical issues, resource limitations, flawed user interfaces, data handling errors, reporting problems, and ethical concerns (Belz et al., 2021b, 2023b; Thomson et al., 2024).

The ReproNLP project has over many years published the results of reproduction results of selected papers (Belz et al., 2021c, 2022; Belz and Thomson, 2023, 2024; Belz et al., 2025, 2026). Past results from previous ReproNLP experiments have found a tendency, where one reproduction would give contradictory results across the two reproduction experiments. One reproduction experiment would agree strongly with the original reproduction, while the other equally strongly disagreeing (Belz et al., 2025). Huidrom and Belz (2025) experiment using LLMs-as-a-judge and found that LLM evaluators would tend to agree strongly with one of experiments and likewise strongly disagree with the other reproduction result. The use of LLMs was adopted in the previous iteration of ReproNLP as another tool to provide an additional perspective in results obtained by researchers in their reproductions.

Reproductions can differ from the original experiments for many reasons. Whilst the ReproNLP organisers have made great efforts to standardise how reproductions are carried out and reported, differences in results between the either the original experiment or even between reproductions remain. Analysis of past results have found that different cohorts can impact the reproducibility of a given experiment (Belz et al., 2021a). Additionally, NLP evaluation interfaces that don't respect fundamental human-centred interaction principles (Calò et al., 2025) can increase cognitive load for evaluators thus impacting reproducibility (Belz et al., 2022). Finally, bugs and/or missing resources in the original experiment can hinder reproduction attempts by labs seeking to generate comparable results (Belz et al., 2023a).

We describe in our reproduction effort the process we used to reproduce the paper "A Recipe For Arbitrary Text Style Transfer with Large Language Models" by Reif et al. (2022). This report follows the earlier published reproduction effort by Onderková et al. (2025). However, both reproduction efforts were done parallel to each other despite the differences in publication date. In section 2 we describe the reproduction experiment with a particular focus on any differences between the original experiment and reproduction. Section 3

[1] ReproNLP - https://repronlp.github.io

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1127–1132

the methodology used and what challenges were encountered when performing the reproduction. The results from this reproduction are reported in section 4, where we report our results and compare them to original experiment and to that of the results obtained by Onderková et al. (2025). Finally, in section 5 we conclude with main findings from this reproduction effort.

2 Reproduction Experiment

In the original paper Reif et al. (2022) presents a series of experiments exploring the use of language models (LMs) to perform zero-shot text style transfer with the aim of transforming the stylistic attributes of a given text while persevering the semantic intent. The authors, in particular, present a new type of zero-shot prompting called *augmented zero shot learning* that allows a given LM to perform text style transfer to different styles without any exemplars in the target style. Leveraging the Reddit Writing Prompts validation dataset (Fan et al., 2018), 50 sentences are transferred 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*). Reif et al. (2022) leverage three different prompting methods and a manually written baseline to comparatively evaluate their approach for style transfer:

- **human** – Set of ground truth transfers written by the authors of the original study (Reif et al., 2022).
- **zero-shot** – A default approach using a base prompt without any examples.
- **augmented zero-shot** – Version of the prompt that includes seven exemplars of different style transfers.
- **paraphrase** – Similar to the zero-shot prompt, but a paraphrased version of the target style as a control.

In addition the above four approaches, the authors also added two additional popular prior baselines style transfer methods, Unsup MT (Prabhumoye et al., 2018) and Dual RL (Luo et al., 2019). Outputs were evaluated using six professional annotators. The annotators evaluated the outputs from the above four prompt approaches and two baselines on a 1-100 scale evaluating three core aspects:

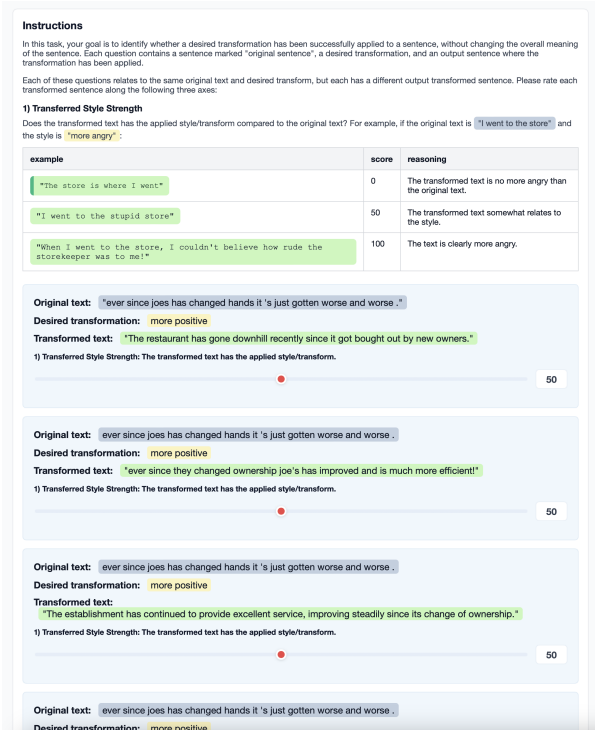

Figure 1: Evaluation interface used for rating the degree of clarity of a text containing generated referring expressions (highlighted in yellow).

- **transfer strength** - the amount that the output actually matches the target style.
- **semantic preservation** - whether the meaning of the text matches irrespective of style.
- **fluency** - the degree to which the text is coherent and could have been written by an English speaker.

For the reproduction experiment we were tasked to evaluate semantic preservation for the *more positive* standard style. This was done with the four prompting methods proposed by Reif et al. (2022).

3 Methodology & Challenges

In the original experiment the human evaluators were six professional annotators. Since these annotators were not available to us for the reproduction experiment six participants were recruited from Prolific[2] as stipulated by the ReproNLP organisers. Whilst compensation was not mentioned in the original experiment, for the reproduction participants were paid he equivalent of the UK living wage[3] of

[2]Prolific - https://www.prolific.com
[3]UK Living Wage - https://www.livingwage.org.uk

Aspect	Original	Reproduction
Number of Items	300	300
Number of Systems	6	4
Number of Participants	6	6
Participants per Item	3	3
Items per Participant	150	150
Recruitment Platform	*unknown*	*Prolific*
Compensation	*unknown*	*£12.60 per hour equivalent*
Participation controls	*unknown*	*native English speakers*

Table 1: Methodological similarities & differences between the original and reproduction human evaluations.

£12.60 per hour. Table 1 details the methodological and participatory similarities and differences between the two experiments.

Data for experiment came from the authors of the original experiment and was reused for the reproduction. However, the interface was not made available for the reproduction as this was an internal interface. Therefore, for the reproduction the interface was reproduced from the image provided in the original paper using GPT-4 (Achiam et al., 2023) to recreate the web-based evaluation interface as closely as the experiment interface image in the original experiment as possible. Figure 1 shows the experimental interface that was used in this reproduction. The code for running the experiment was adapted from a previous reproduction effort (Mahamood, 2024) and incorporated the new generated evaluation interface.

In addition to setting up the reproduction experiment by using the original experiment's codebase a Human Evaluation Datasheet (HEDS) (Shimorina and Belz, 2022) was also completed[4]. The HEDS form records in a standardised way the properties of human evaluations to support comparability, meta-evaluation, and reproducibility of human evaluations.

4 Results

We present our results against the original study in table 2. The results from the original study were presented as bar plots and thus making it challenging to determine the exact numerical values. The values shown in table 2 are those estimated by Onderková et al. (2025) using a pixel counting estimation approach. However, the large differences between results obtained in our reproduction for the *paraphrase* and *zero-shot* prompting approaches is very abundant irrespective of the potential deviations introduced by the estimation method performed by Onderková et al. (2025). For *augmented zero-shot* and *human* the differences are smaller comparatively, with the *human* baseline results show better reproducibility than any of the prompting approaches. Nevertheless, we don't see the *augmented zero-shot* having a comparable score to the baseline *human* outputs, contrary to the original experiment results.

We also compare the results across all three experiments using the Quantified Reproducibility Assessment (Belz and Thomson, 2026) for calculating the degree of convergence/divergence between multiple evaluation studies. In particular, we used the QRA tool[5] that was created by the ReproNLP organisers. Table 3 shows the results of this cross-assessment. The scores presented in table 3 shows a very similar pattern to that of in table 2. There are large differences between for *paraphrase* and *zero-shot* as seen by the CV* scores. However, there is greater convergence for both the *augmented zero-shot* and *human* results. This indicates a greater reproducibility for these system outputs in comparison to the other two types. However, the most reproducible score is that of the *human* outputs, beating the other three automatic approaches.

5 Conclusion

Like in the reproduction conducted by Onderková et al. (2025) we are also unable to validate the claim by Reif et al. (2022) that the *augmented zero-shot* show comparability to the *human* outputs. There could be several reasons for this discrepancy. Firstly, in the original study professional annotators were used as compared to crowd workers in both reproductions. It is plausible that these annotators

[4]ReproNLP 2025 HEDS forms - https://github.com/nlp-heds/repronlp2025

[5]QRA tool - https://github.com/DCU-NLG/qra

	Original	Reproduction	(mean) CV*
paraphrase	90.29	21.74	110.43
zero-shot	69.71	24.24	87.65
augmented zero-shot	86.47	67.31	22.37
human	85.29	79.33	6.5

Table 2: Clarity mean average results from both original and reproduction human evaluation. Unbiased coefficient of variation values (**CV***) calculated using the definition by Belz (2022). Original results are from Reif et al. (2022).

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 3$)		
				System level	QC level	Study level
Type I	semantic preservation – *more positive*	paraphrase	(mean) CV*↓	76.22	39.28	39.28
		zero-shot		54.86		
		augmented zero-shot		18.50		
		human		7.56		
Type II	semantic preservation – *more positive*	all	mean r ↑	n/a	0.513	n/a
		all	mean ρ ↑	n/a	0.067	
		all	W ↑	n/a	0.378	
Type IV	semantic preservation – *more positive*	all	P ↑	n/a	0.556	0.556

Table 3: QRA reproducibility assessment across three comparable experiments (n=3), Reif et al. (2022), Onderková et al. (2025), and *our reproduction*; n/a = measure does not apply at this level.

given their professional experiences would assess the outputs differently than compared to evaluators recruited from the public. Past reproduction results have shown that different cohorts can impact the reproducibility of a given experiment (Belz et al., 2021a). Another cause for this discrepancy could be due to the fact that the interface used for both reproduction efforts differed from the original in-house version. The fact that both reproduction efforts show similar results could be a strong indicator that these barriers mean that it is not possible to reproduce the results of the original study without a more complete set of resources. The results presented in this paper once again highlights the need for authors to document and retain non-research paper resources to enable lower barriers to reproduction and hopefully, better reproducibility.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th conference of the European chapter of the association for computational linguistics*, pages 313–320.

Anja Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021a. 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.

Anja 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.

Anya Belz. 2022. A Metrological Perspective on Reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021b. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021c. 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, Anastasia Shimorina, Maja Popović, and Ehud Reiter. 2022. The 2022 ReproGen shared task on reproducibility of evaluations in NLG: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 43–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia. ELRA and ICCL.

Anya Belz and Craig Thomson. 2026. Quantified reproducibility assessment for four common types of evaluation results in nlp/ml. *Computational Linguistics*, pages 1–10.

Anya Belz, Craig Thomson, and Javier González Corbelle. 2026. The shared task on reproducibility of evaluations in nlp (ReproNLP) 2026: Overview and results. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. Association for Computational Linguistics.

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 Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 1002–1016, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

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. 2023a. 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.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023b. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

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.

Rudali Huidrom and Anya Belz. 2025. Using LLM judgements for sanity checking results and reproducibility of human evaluations in NLP. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 354–365, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 5116–5112, Macao, China. International Joint Conferences on Artificial Intelligence.

Saad Mahamood. 2024. ReproHum #0124-03: Reproducing human evaluations of end-to-end approaches for referring expression generation. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 250–254, Torino, Italia. ELRA and ICCL.

Kristýna Onderková, Mateusz Lango, Patrícia Schmidtová, and Ondrej Dusek. 2025. ReproHum #0669-08: Reproducing sentiment transfer evaluation. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 601–608, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

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.

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.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*, pages 1–11.