ReproHum #0744-02: A Reproduction of the Human Evaluation of Meaning Preservation in “Factorising Meaning and Form for Intent-Preserving Paraphrasing”

Julius Steen   Katja Markert
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
(steen|markert)@cl.uni-heidelberg.de

Abstract

Assessing and improving the reproducibility of human evaluation studies is an ongoing concern in the area of natural language processing. As a contribution to this effort and a part of the ReproHum reproducibility project, we describe the reproduction of a human evaluation study (Hosking and Lapata, 2021) that evaluates meaning preservation in question paraphrasing systems. Our results indicate that the original study is highly reproducible given additional material and information provided by the authors. However, we also identify some aspects of the study that may make the annotation task potentially much easier than those in comparable studies. This might limit the representativeness of these results for best-practices in study design.

1 Introduction

Reproducibility is a central requirement for human evaluation studies. Given the same data and setup, other researchers should be able to independently arrive at similar conclusion as the original works. However, in practice, reproducibility of human evaluation studies remains problematic in the field of natural language processing (Howcroft et al., 2020; Gehrmann et al., 2023). In this context, systematic reproductions of human evaluation studies play an important role in assessing the state of reproducibility in the field and in establishing best practices.

As part of the ReproHum project (Belz and Thomson, 2024; Belz et al., 2025), this report describes our effort to reproduce a human evaluation study of paraphrasing systems, originally conducted by Hosking and Lapata (2021). Based on information submitted by Hosking and Lapata to the ReproHum organizers, we attempt an otherwise independent reproduction that closely mirrors the original study. We also provide an HEDS (Shimorina and Belz, 2022; Belz and Thomson, 2024) form for our reproduction study, which is accessible in the shared ReproNLP repository.[1]

Our results indicate that the original study is highly reproducible, even in the face of a change in annotator recruitment and study scope.[2] However, we also find that this is in part due to the large quality differences in the systems in the study and not an exclusive consequence of the original design decisions.

2 Original Study

The basis of our reproduction study is an evaluation of paraphrasing systems conducted by Hosking and Lapata (2021). They propose a neural paraphrasing system that, given an input question, outputs a distinct paraphrase that conserves the original meaning. The system combines two encoder representations to generate the paraphrases: A continuous variational representation derived from the input to represent question semantics and a discrete syntactic representation to indicate the desired surface form of the paraphrase. In keeping with the original work, we refer to this system as *Separator*.

The study in question is part of the evaluation in Hosking and Lapata (2021) and focuses on comparing the newly introduced system *Separator* against competitors in three dimensions, which the authors describe as follows:

Fluency “Which system output is the most fluent and grammatical?”

Meaning “To what extent is the meaning expressed in the original question preserved in the rewritten version, with no additional information added? Which of the questions generated by a system is likely to have the same answer as the original?”

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

[2] All code used in the reproduction is available at https://github.com/julmaxi/reprohum_744 or in the project-wide repository.

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 568–575
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Dissimilarity "Does the rewritten version use different words or phrasing to the original? You should choose the system that uses the most different words or word order."

All dimensions were evaluated jointly in the same form by human annotators.

The goal of the original study was to demonstrate that the newly proposed approach preserves meaning and fluency while maintaining adequate dissimilarity to the input.

The following, more detailed description of the study is based both on the original paper, as well as on additional materials and resources that were obtained by the ReproHum organizers from the authors. At no point was there any direct interaction between the authors of the original study and the authors of this reproduction study.

2.1 Original Study Design

The original study was set up as a pairwise evaluation study between Separator and three competing systems, which were selected based on their performance in a previous automatic evaluation against reference paraphrases:

VAE is an ablation of Separator that computes a continuous representation from the input only, with no separation between syntactic and semantic representation.

LBoW (Fu et al., 2019) passes a bag-of-words content plan to the decoder, alongside an encoding of the input.

DiPS (Kumar et al., 2019) uses submodular functions during decoding of a paraphrasing model to encourage semantically similar and syntactically distinct candidate paraphrases.

The study had 40 batches, each of which consisted of 30 head-to-head comparisons, plus two distractor questions, which we will discuss in Section 2.2. Figure 1a shows a screenshot the interface shown to annotators for each comparison. Each batch was constructed by comparing all six possible pairs of the four systems on five distinct input sentences. This resulted in a total of 200 distinct input sentences in the evaluation.

Each batch was evaluated by a set of three annotators, which were recruited via Amazon Mechnical Turk. Turkers were filtered to have an acceptance rate of >96% at >5000 accepted HITS and had to be located either in the United States or the United Kingdom. There was no limitation on the number of repeat annotations and annotators were paid 3 USD per batch according to communication between the authors and ReproHum.

2.2 Distractors

The original study employed distractor questions to identify and reject low-effort submissions. Two kinds of distractors were used in the original study:

- *Meaning* distractors consisted of a gold standard paraphrase and a gold standard paraphrase for a completely different input. Annotators had to correctly identify that the gold paraphrase is more semantically similar.
- *Input* distractors evaluated the input sentence against the gold standard paraphrase. Annotators had to correctly identify that the gold paraphrase is more dissimilar from the input.

Each batch in the original study contained one input and one meaning distractor. The authors reported in communication with the ReproHum organizers that all submissions with at least one failed attention check were rejected and resubmitted for annotation.

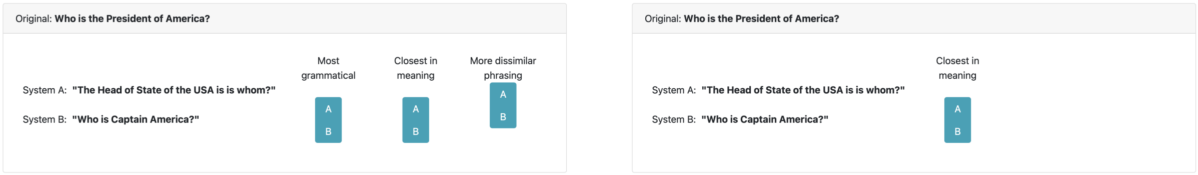
3 Reproduction Study

Following the guidelines of the ReproHum project, we reproduce the study as closely as possible following the setup described in Section 2.1. ReproHum organizers were able to obtain the original batches used in the study, as well as the original interface template. We employ both in our reproduction study.

However, we introduce two major deviations from the original study setup:

1. Following the guidelines of the ReproHum project, we only reproduce the *Meaning* criterion. Since in the original study, all three criteria were evaluated simultaneously in the same form, this requires us to modify the original interface.
2. We follow ReproHum reproduction guidelines in using Prolific[3] for crowd-worker recruitment, whereas the original study used Amazon Mechanical Turk. This additionally requires the use of a custom backend to replace the Mechanical Turk infrastructure and

[3] prolific.com



Original: **Who is the President of America?**

System A: **"The Head of State of the USA is is whom?"**

System B: **"Who is Captain America?"**

Original: **Who is the President of America?**

System A: **"The Head of State of the USA is is whom?"**

System B: **"Who is Captain America?"**

(a) Original

(b) Reproduction

Figure 1: Interface for a single comparison for the original and reproduction study.

changes in the way candidate annotators are screened. We elaborate on the latter in Section 3.1.

Both changes can potentially impact the results of the reproduction. Another threat to reproducibility is the relatively large time difference between the original study and the reproduction. While the original work was published in August 2021, with experiments likely concluded at least a few months before this date, all annotations in the reproduction study were elicited on May 3rd, 2024. This is particularly relevant, since there is some evidence that LLMs increasingly penetrate crowd-working platforms (Veselovsky et al., 2023a,b), which might alter annotator behavior.

3.1 Annotator Recruitment and Payment

We attempt to mirror the recruitment criteria of the original study with the built-in screeners available at Prolific while following the ReproHum project-wide guidelines. To mirror the acceptance rate requirement, we require workers to have an approval rate of 99-100%, with at least 200 previous submissions. This reflects both the smaller size of Prolific and their stricter requirements for rejection. We use the country of residence filter to limit participation to residents of one of four English-speaking countries: United Kingdom, United States, Canada, and Australia. The addition of Canada and Australia to the list of allowed countries compared to the original follows ReproHum project guidelines. While the original study did not control the number of batches each annotator was able to complete, we limit workers to a single batch, again following ReproHum guidelines.

We set payment per batch at £2, based on an initial conservative estimate for the completion time of 10 minutes per batch. This results in a nominal rate of £12 per hour, satisfying both Prolific and ReproHum recommendations.

Prolific requires a short description of the study, which is shown to workers before they accept. We choose the following summary of the study:

> You will be shown a set of several items. Each item contains an original question, as well as two candidate paraphrases of the question. Paraphrases are generated by different automatic systems. You must select which paraphrase best captures the meaning of the original question.

3.2 Interface

While we have the original source code for the interface available, our focus on the *Meaning* criterion requires some modification to the original. Specifically, we:

1. Remove all buttons related to dissimilarity and grammaticality criteria.

2. Remove all instructions related to these criteria.

Figure 1 shows a direct comparison of the original and modified annotation interfaces. Since we remove the dissimilarity criterion, we also have to eliminate the *input* distractor. To maintain the length of each batch, we replace each *input* distractor with a randomly sampled *meaning* distractor from another batch.

In addition to the changes required by the difference in scope between the studies, we make some minor modifications to the instructions to comply with Prolific and ReproHum regulations:

1. The original study contains a remark that the study contains attention checks and that these checks will be used to reject low-effort submissions. However, since these checks are not instructional manipulation checks[4] (see

[4] I.e. a check that replaces the question for an instance with an explicit instruction to answer in a particular way (Oppenheimer et al., 2009).

Section 2.2), prolific guidelines[5] do not allow for rejection on grounds of a missed check. We thus remove this section of the original instructions.

2. We replace original contact information with our own contact information.

3. We exchange the word "HIT" with the word "study", which better follows Prolific terminology.

4. We add a more detailed informed consent section and required workers to explicitly indicate consent by clicking a checkbox.

We consider none of these modifications to be likely to have an impact on the results of our reproduction.

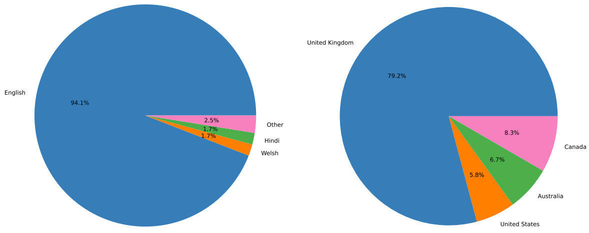
4 Study Statistics

Due to a bug in annotator assignment[6], we elicited a total of 121 batch annotations. Since we only require a total of 120 (= 3 repeat annotations × 40 batches), we randomly discard one repeat annotation from the over-annotated batch. We find only a single missed attention check in the entire annotation set. Due to the low prevalence of missed attention checks and the cost associated with resubmission, we opt to not discard the related submission in a slight deviation from the original protocol.

The median completion time, as measured from the time a study was accepted on Prolific to the time the annotator submitted the completion code to Prolific, is 7:16 minutes. This shorter than estimated completion time results in an average actual hourly pay of £16,51, well above our nominal target rate of £12,00.

4.1 Annotator Demographics

Prolific automatically provides self-reported demographic data about participants. This allows us to assess the effectiveness of the location filter. Additionally, we quantify possible differences between the original Mechanical Turk annotator pool and our Prolific annotators by studying the country of origin of the annotators. We report summary statistics for both language and country of residence in Figure 2. The self-reported language is overwhelmingly English, suggesting geographic filters work very well as a proxy for language.



(a) Language (b) Country of Residence

Figure 2: Distribution of self-reported Country of Residence and Language.

Interestingly, we find that there is a concentration of workers in the United Kingdom. While we do not have demographic data for the original study, this indicates potentially substantial demographic differences to the original study, since, on Mechanical Turk, most workers are from the United States (Difallah et al., 2018). While this is unlikely to affect rankings for meaning preservation, such systematic differences in annotator population might make reproduction more difficult for criteria such as grammaticality.

5 Reproduction Results

5.1 Agreement

System Pair	Agreement (%)
VAE/Sep.	80.0
VAE/LBoW	82.3
VAE/DiPS	84.0
Sep./LBow	81.0
Sep./DiPS	81.7
DiPS/LBow	79.0
Overall	81.3

Table 1: Empirical agreement for pairwise rankings overall and per system-pair.

While the original study does not report agreement, it is an important indicator for understanding the quality and difficulty of a study. We thus report overall agreement in pairwise decisions in Table 1. Additionally, we report detailed agreement figures per system-pair in the same table. Since the order of systems in a comparison is randomized and

[5] See https://web.archive.org/web/20240908022618/https://researcher-help.prolific.com/en/article/fb63bb.

[6] Unlike Mechanical Turk, Prolific does, at the time of the study, not have facilities to conduct a multi-batch study where each annotator may only annotate a single batch. This lead to a mismatch between our backend and Prolific when an assignment was returned.

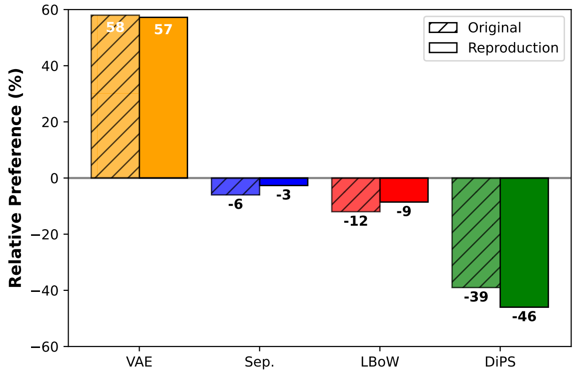


Figure 3: Original and reproduced relative preferences.

we elicit pairwise judgments, we find no justification to adjust for chance-agreement and report the empirical agreement directly.

We find overall moderate instance-level agreement. We note that agreement is notably higher for comparisons between VAE and DiPS, which also have the largest gap in meaning scores in both the original study and our reproduction.

5.2 Comparison of Results

Following the original study, we report relative preference values for each system. Relative preference is computed by assigning a value of $+1$ if a system wins a pairwise comparison, and a value of -1 if it loses a comparison and averaging these values. Figure 3 gives a direct comparison between the original and reproduced relative preferences. We find very similar trends across both. This matches findings by Arvan and Parde (2024); Watson and Gkatzia (2024), who independently reproduced a very similar human evaluation study of a successor paraphrasing system (Hosking et al., 2022) and also find high reproducibility. Compared to the original, the main deviation we find is that DiPS receives a lower preference score overall, profiting mainly Separator and Latent BoW.

In addition to the relative preferences, we also report the pairwise outcomes of each system pair in Figure 4. We find that win rates are mostly consistent with the overall ranking. Furthermore, all system pairs, with the exception of Separator and LBoW, have a $\geq 15\%$ margin in win rates, indicating large differences in system quality.

5.3 Detailed Assessment of Original Claims

Hosking and Lapata (2021) make two statements with regard to the result for the *Meaning* criterion:

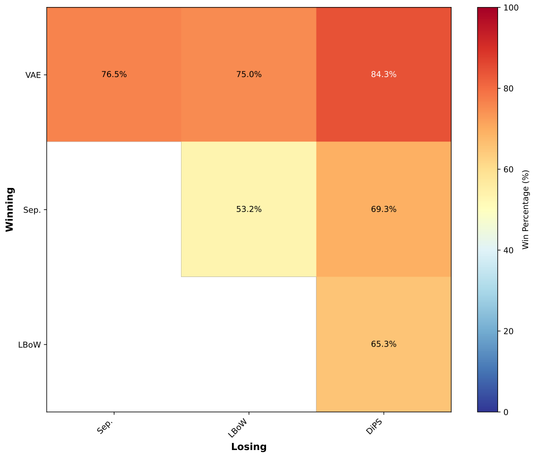


Figure 4: Pairwise win rates for each system pair. Each cell indicates the win rate of the system in the row against the system in the column.

1. The VAE baseline is best at preserving meaning of the questions.

2. Separator better preserves question intent than the remaining systems.

Our results confirm both statements.

Additionally, the authors conduct a one-way ANOVA with a post-hoc Tukey test to detect statistically significant differences in system scores. While they do not explicitly describe the aggregation they use for this test, we assume they conduct this analysis on the average system win-rates for each batch.

Using this procedure, we find that all pairwise differences are statistically significant ($p < 0.05$), with the exception of the difference in the *Meaning* scores of Separator and LBoW. The statistical analysis thus supports the first claim that the VAE baseline is best at meaning preservation, but not the second claim that there is a significant gap between LBoW and Separator.

5.4 Quantitative Reproducibility Assessment

To further quantify the degree of reproducibility of the original experiment, we conduct a quantified reproducibility assessment (Belz, 2025). We compute both the reproducibility of individual scores (Type I assessment), as well the reproducibility of the relative score differences between systems (Type II assessment).

For individual scores, we compute the bias-corrected coefficient of variation (CV*) (Belz,

2022) for each individual result as

$$CV^* = \frac{1}{4n} \cdot \frac{s^*}{|\bar{x}|}$$

where s^*, $\bar{x}$ are the bias-corrected estimate for sample standard deviation (assuming a normal distribution) and the sample mean respectively. n is the sample size. Since CV* is only meaningful on ratio scales, we transform the relative preferences into win rates for this analysis.[7]

System	Original	Reproduction	CV*
VAE	0.58	0.57	0.49
Sep.	-0.06	-0.03	3.47
LBoW	-0.12	-0.09	3.83
DiPS	-0.39	-0.46	12.14

Table 2: Original and reproduced relative preference values, as well as CV* values for all systems.

Table 2 shows CV* values for all scores. We note that CV* has limited expressivity considering the small sample size of only two studies (i.e. this study and the original). We include it for standardization of reporting in reproduction studies.

To test the reproducibility of relative score differences, we compute the correlation between original and reproduced scores. We can see directly that the Spearman correlation between original and reproduced scores is 1 ($p < 0.05$) and we compute the Pearson correlation between both as 0.994 ($p < 0.05$). Both values indicate high reproducibility.

6 Discussion

While the high degree of reproducibility of the original study is encouraging, we note that this result was achieved with access to information and resources that are not directly available from either the original publication or the associated repository. Specifically, neither the original batch assignment we use to reproduce results nor the original annotation interface are publicly available. Both were made accessible to the ReproHum project upon request.

We also highlight that, as an exact reproduction, we can only make statements about how well results reproduce under the same selection of documents for annotation. However, the claims of the original study should be replicable for *any* subset of the original dataset that is sampled according to the study description. Evaluating this goes beyond the scope of this study and it is reasonable to expect a higher variation if a different subset was sampled from the input.

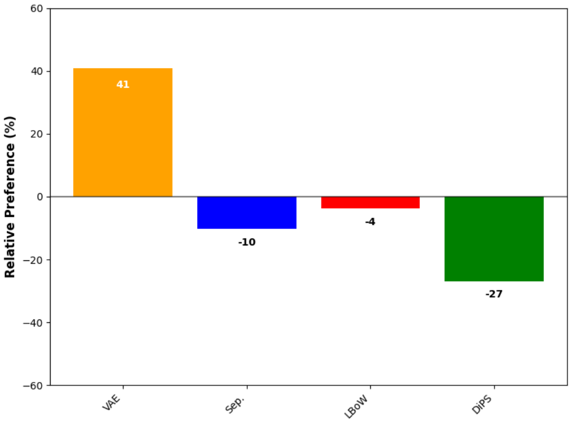


Figure 5: Relative preferences when assigning rankings based on overlap and presence of *Unk* tokens. We find that preferences are remarkably similar to the results of both studies considering their simplicity.

System	Identical input/output pairs (%)
VAE	46
Sep.	7
LBoW	12
DiPS	2

Table 3: Percentages of outputs identical to the input for each system. VAE has by far the largest number of exactly identical inputs and outputs, explaining its high meaning preservation score.

Finally, if this study is interpreted to assess best practices for reproducible study design, we should be mindful that reproducibility is greatly aided by the clear-cut differences between systems. Specifically, we identify two properties of the dataset that likely make annotation much easier than for other, similar studies:

1. The low diversity of VAE generations leads to many instances where the VAE output is the same as the input (see Table 3). In these cases, the ranking is obvious unless the competitor system also exactly reproduces the input.

2. Both DiPS and LBoW have low linguistic quality as identified in the original study. In particular, we observe a high frequency of

[7] This is equivalent to computing CV* on relative preferences shifted to start at zero.

Unk tokens in DiPS, which removes important information from the question and likely also contributes to a set of easy annotation decisions.

To illustrate how these dataset properties make the task comparatively easy, Figure 5 shows the hypothetical relative preferences that systems would receive under the following deterministic annotation rules:

1. If both outputs are identical to the input, randomly choose one.
2. If only one of the outputs is identical to the input, choose that output.
3. If both outputs are different from the input, choose the one that does not contain an *Unk* token.
4. If none of the above rules apply, randomly choose one.

The resulting scores already closely resemble the original ranking. In particular, we can easily reproduce the very strong performance of VAE and the very weak score of DiPS. LBoW and Separator are close, just like in the original study, although the ranking is inverted. However, both manual inspection of the data and the original study scores for the *Fluency* score suggest that Separator is much more grammatical than LBoW, which sometimes outputs paraphrases that are difficult to parse. This is likely to skew results in favor of Separator as observed in the original study, but is not captured by our simple setup.

7 Conclusions

In this report, we have given an account of our reproduction attempts of a study of meaning preservation annotation in paraphrasing systems. Our results show an encouragingly high degree of reproducibility with the resources provided by the authors. In particular, the availability of original batches and interfaces makes it easy to design a highly similar setup to the original study. However, our analysis also shows that care needs to be taken not to over-interpret the outcomes of this study when it comes to making recommendations about best-practices in general. In particular, we find that the dataset contains many decisions which are likely to be very easy for annotators due to exact correspondence between inputs and outputs and the presence of obvious defects in some paraphrases, which make them unreadable. It thus remains unclear to which extent the results of this reproduction study are representative of more challenging annotation studies. This is particularly relevant, since current generations of NLG systems are well known to be much less prone to the kind of obvious mistakes present in the original study.

References

Mohammad Arvan and Natalie Parde. 2024. ReproHum #0712-01: Human evaluation reproduction report for “hierarchical sketch induction for paraphrase generation”. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 210–220, Torino, Italia. ELRA and ICCL.

Anya Belz. 2022. A metrological perspective on reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

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

Anya Belz and Craig Thomson. 2024. The 2024 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz and Craig Thomson. 2024. HEDS 3.0: The Human Evaluation Data Sheet Version 3.0. *arXiv e-prints*, arXiv:2412.07940.

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)*.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM ’18, page 135–143, New York, NY, USA. Association for Computing Machinery.

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase generation with latent bag of words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. *Journal of Artificial Intelligence Research*, 77:103–166.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1418, Online. Association for Computational Linguistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2022. Hierarchical sketch induction for paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2489–2501, Dublin, Ireland. Association for Computational Linguistics.

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.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. 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 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872.

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.

Veniamin Veselovsky, Manoel Horta Ribeiro, Philip Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2023a. Prevalence and prevention of large language model use in crowd work. *arXiv preprint*. ArXiv:2310.15683 [cs].

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023b. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *arXiv preprint*. ArXiv:2306.07899 [cs].

Lewis N. Watson and Dimitra Gkatzia. 2024. ReproHum #0712-01: Reproducing human evaluation of meaning preservation in paraphrase generation. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 221–228, Torino, Italia. ELRA and ICCL.