The 2024 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results

Anya Belz, Craig Thomson
ADAPT Research Centre, Dublin City University
Dublin, Ireland
{anya.belz,craig.thomson}@dcu.ie

Abstract

This paper presents an overview of, and the results from, the 2024 Shared Task on Reproducibility of Evaluations in NLP (ReproNLP'24), following on from three previous shared tasks on reproducibility of evaluations in NLP, ReproNLP'23, ReproGen'22 and ReproGen'21. This shared task series forms part of an ongoing research programme designed to develop theory and practice of reproducibility assessment in NLP and machine learning, against a backdrop of increasing recognition of the importance of reproducibility across the two fields. We describe the ReproNLP'24 shared task, summarise results from the reproduction studies submitted, and provide additional comparative analysis of their results.

Keywords: Reproducibility, Shared Task, Evaluation.

1. Introduction

Reproducibility continues to be a problem in search of a solution in the Natural Language Processing (NLP) field (Belz et al., 2021a, 2023). We still do not understand well enough what makes system evaluations, both human and metric-based, easier or harder to reproduce, while a growing number of reproduction studies have revealed alarmingly poor degrees of reproducibility and numerous issues with current evaluation practices (Belz et al., 2023).

The aim of this fifth reproduction-focused shared task in NLP, following REPROLANG'20 (Branco et al., 2020), ReproGen'21 (Belz et al., 2021b), ReproGen'22 (Belz et al., 2022), and ReproNLP'23 (Belz and Thomson, 2023), is generally to continue to add to the body of reproduction studies in NLP and machine learning (ML), and more specifically, to produce and analyse multiple reproductions of shared original evaluations, thereby creating more reliable reproducibility results for individual evaluations and evaluation methods, given that the evidence is that multiple reproductions rarely produce the same reproducibility results.

The 19 new reproduction studies that make up ReproNLP'24 add a good number of further data points available for investigating reproducibility, and help to continue identifying properties of evaluations that are associated with better reproducibility.

We start in Section 2 with a description of the organisation and structure of the shared task, along with track details. Next, we summarise results at the level of individual experiments, in terms of the reproduction task, and different degree-of-reproducibility assessments, first for Track B (Section 3), then Track A (Section 4).

In Section 5, we look at the quality criteria assessed in evaluations and other properties of the ReproNLP evaluation studies in standardised terms as facilitated by HEDS datasheets, and explore if any of these show signs of affecting degree of reproducibility (Section 5). We conclude with some discussion (Section 6) and a look to future work (Section 7).

2. ReproNLP 2024

ReproNLP 2024[1] consisted of two tracks, one an 'unshared task' in which teams re-run their own or any other previous work (Track A), the other a standard shared task in which teams re-run one of a set of organiser-selected experiments (Track B):

- **A Open Track:** Repeat any previously reported work developing and evaluating systems, and report the approach and outcomes. Unshared task.
- **B ReproHum Track**: For a shared set of selected evaluation studies (listed below) from the ReproHum Project, participants repeat one or more of the studies and compare results, using the information provided by the ReproNLP organisers only, and following a common reproduction approach.

Track B forms part of the ReproHum project[2] and the studies offered in it were selected according to criteria of suitability and balance to form part of a larger coordinated multi-lab multi-test reproduction study, as described in detail elsewhere (Belz et al., 2023).

[1] All information and resources relating to ReproNLP are available at https://repronlp.github.io/.

[2] https://reprohum.github.io/

Proceedings of the HumEval @LREC-COLING 2024, pages 91–105
Turin, Italy. May 21st, 2024.

An overview of the papers we selected experiments from, and the complete studies the latter formed part of, is presented below. Note that we only include here the original papers for which we received submissions; there were 21 papers offered in the track in total (the full list can be found on the ReproNLP website).

The information provided for each study below covers whether the assessment of systems was *relative* to other systems or *absolute* without comparitors; what the language*(s)* of the systems were; how many *datasets* were used; how many *systems* were evaluated and by how many *evaluators*; and whether the evaluation was run on a *crowdsourcing* platform.

1. **Reif et al. (2022)**: *A Recipe for Arbitrary Text Style Transfer with Large Language Models*: https://aclanthology.org/2022.acl-short.94

   Absolute evaluation study; English; 3 quality criteria; 3 datasets; varies between 4 and 6 systems and between 200 and 300 evaluation items per dataset-criterion combination; crowdsourced.

2. **Liu et al. (2021)**: *DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts*: https://aclanthology.org/2021.acl-long.522

   Relative evaluation study; English; 3 quality criteria; 2 datasets; varies between 5 and 6 systems and between 960 and 1200 evaluation items per dataset-criterion combination; crowdsourced.

3. **Atanasova et al. (2020)**: *Generating Fact Checking Explanations*: https://aclanthology.org/2020.acl-main.656

   Absolute evaluation study; English; 1 quality criterion; 1 dataset; 3 systems and 240 evaluation items. Relative evaluation study; English; 4 quality criteria; 1 dataset; 3 systems and 40 evaluation items per criterion.

4. **August et al. (2022)**: *Generating Scientific Definitions with Controllable Complexity*: https://aclanthology.org/2022.acl-long.569

   Absolute evaluation study; English; 5 quality criteria; 2 datasets; 3 systems and 300 evaluation items per dataset-criterion combination; some crowdsourced.

5. **Hosking et al. (2022)**: *Hierarchical Sketch Induction for Paraphrase Generation*: https://aclanthology.org/2022.acl-long.178

   Relative evaluation study; English; 3 quality criteria; 1 dataset; 4 systems and 1800 evaluation items per criterion; crowdsourced.

6. **Yao et al. (2022)**: *It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books*: https://aclanthology.org/2022.acl-long.54

   Absolute evaluation study; English; 3 quality criteria; 1 dataset; 3 systems and 361 evaluation items per criterion.

7. **Feng et al. (2021)**: *Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization*: https://aclanthology.org/2021.acl-long.117

   Absolute evaluation study; English; 3 quality criteria; 2 datasets; 7 systems and varies between 70 and 700 evaluation items per dataset-criterion combination.

8. **Gabriel et al. (2022)**: *Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines*: https://aclanthology.org/2022.acl-long.222

   Absolute evaluation study; English; 3 quality criteria; 1 dataset; 3 systems and 588 evaluation items per criterion; crowdsourced.

9. **Kasner & Dusek (2022)**: *Neural Pipeline for Zero-Shot Data-to-Text Generation*: https://aclanthology.org/2022.acl-long.271

   Absolute evaluation study; English; 5 quality criteria; 2 datasets; 6 systems and 600 evaluation items per dataset-criterion combination.

10. **Shardlow & Nawaz (2019)**: *Neural Text Simplification of Clinical Letters with a Domain Specific Phrase Table*: https://aclanthology.org/P19-1037

    Relative evaluation study; English; 1 quality criterion; 1 dataset; 4 systems and 100 evaluation items; crowdsourced.

11. **Castro Ferreira et al. (2018)**: *NeuralREG: An end-to-end approach to referring expression generation*: https://aclanthology.org/P18-1182

    Absolute evaluation study; English; 3 quality criteria; 1 dataset; 6 systems and 144 evaluation items per criterion; crowdsourced.

In the ReproHum multi-lab multi-test study (for which the above papers were selected), rather than attempt to repeat entire studies, we decided to use our limited resources to repeat assessments of individual quality criteria on individual datasets (which is what we mean by a single 'experiment'), with specific properties so as to have equal numbers of assessments with the specific properties the ReproHum study is designed to compare. Some of the properties of these individual experiments are given in Table 2 alongside the (single) quality criteria they assess.

Each of these experiments is being re-run in two separate reproduction studies in ReproHum. Those that have completed in the current batch are

included here in the ReproNLP'24 report. All experiments from the current and preceding batch (the latter reported in ReproNLP'23) were also open to all other ReproNLP'24 participants.

Note that non-ReproHum participants were free to include more than the ReproHum experiment in their reproduction study, and some did (Section 4).

We obtained agreement from the original authors to use their experiments in the ReproHum project and provided very detailed information about the experiments which were shared with all participants.

2.1. Participation

There were three submissions for Track A and 15 for Track B. One submission in Track A did not meet our quality threshold and was rejected. The ReproHum partners reporting in Track B are listed in Table 1. The non-ReproHum participating labs were University of Bucharest (Florescu et al., 2024) in Track B, and Heriot-Watt University (Sasidharan Nair et al., 2024) and ADAPT Centre / Dublin City University (Lorandi and Belz, 2024) in Track A (see Sections 4 and 3.2, respectively).

2.2. Approach to reproduction and reproducibility assessment

We encouraged all participants to complete a HEDS datasheet (Shimorina and Belz, 2022) in the ReproHum version,[3] and to follow the ReproHum Common Approach to reproduction laid out in Appendix A which includes QRA++ (Belz, 2022; Belz and Thomson, 2023), an approach to measuring how close results from two evaluations are, and how reproducible evaluation measures are, in a way that accommodates multiple reproduction studies of the same original work and is comparable across different such sets of reproductions.

In this report we analyse all submissions in terms of QRA++ measures recomputed by us to facilitate comparison across submissions. In brief summary, QRA++ distinguishes four types of results commonly reported in NLP and ML papers:

1. Type I results: single numerical scores, e.g. mean quality rating, error count, etc.
2. Type II results: sets of related numerical scores, e.g. set of Type I results .
3. Type III results: categorical labels attached to text spans of any length.
4. Type IV results: Qualitative findings stated explicitly or implied by quantitative results in the original paper.

The above are quantitatively assessed as follows:

1. Type I results: Small-sample coefficient of variation CV* (Belz, 2022).
2. Type II results: Pearson's r, Spearman's ρ.
3. Type III results: Multi-rater: Fleiss's κ; Multi-rater, multi-label: Krippendorff's α.
4. Type IV results: Proportion of findings that are / are not confirmed by the repeat experiment. To obtain comparable results we restrict ourselves to pairwise system ranks as findings.

In the submissions analysed in this paper we have Type I, II and IV results, and therefore apply the corresponding quantitative measures above. CV* plays a central role in our analyses, and is a version of the standard coefficient of variation corrected for small samples (Belz, 2022).

The ReproHum reproduction studies were strictly controlled to be comparable to each other and the original work. However, there was a difference between the studies reported in 2023 and 2024 in this respect. For the earlier batch, our aim was to achieve maximum similarity between original and reproduction studies, and we strove to resolve every last bit of lack of clarity. In the batch reported here, we abandoned this ultimately infeasible approach, recognising that evaluation experiments should be robust to minor differences. As a result, when there was insuffidient clarity about how an aspect of an experiment was implemented, partner labs drafted solutions which were moderated by the ReproHum project team to provide an agreed solution that both partner labs reproducing the same experiment then used. For more details on such cases, please see the individual submission reports in this volume.

Finally, we have by now gathered a sufficient number of reproduction studies reporting CV* values to support the following categorisation for *human evaluations*: we refer to any CV* from 0 to around 10 as indicating a good degree of reproducibility, between 10 and around 30 as medium, and anything above that as poor.

Note that high CV* scores indicate poor reproducibility, and vice versa.

3. Track B

The subsections within Sections 3.1 and 3.2 each report the results from all reproduction studies for one of the Track B experiments, Sections 3.1 as conducted by ReproHum partners, in Sections 3.2 by other ReproNLP participants.

In each such subsection, we start by giving a brief summary of the experiment. Next, we show the system-level evaluation scores from the original study and the either one or two reproduction studies, alongside the corresponding CV* value computed on all either two or three scores. We finish

[3] https://github.com/nlp-heds/repronlp2024

Original Study	Qual. Criterion	#ev-ors	#sys	items-per-sys	Labs reproducing study
Liu et al. (2021)	Fluency	varies	5	192	a) Heriot-Watt University b) U. de Santiago de Compostela
Hosking et al. (2022)	Preservation of meaning	varies	4	450	a) University of Illinois Chicago b) Edinburgh Napier University
Feng et al. (2021)	Informativeness	4	7	10	a) Bielefeld University b) Charles University
Atanasova et al. (2020)	Coverage	3	3	13	a) Manchester University b) Peking University
August et al. (2022)	Fluency	2	3	100	a) Tilburg University b) University of Groningen
Castro Ferreira et al. (2018)	Clarity	60	6	24	a) trivago
Kasner and Dusek (2022)	Number of redundancies per system	2	6	100	a) Technological University Dublin
Shardlow and Nawaz (2019)	Ease of understanding	varies	4	25	a) University of Groningen
Gabriel et al. (2022)	Social acceptability	varies	3	196	a) University of Cape Town

Table 1: ReproNLP experiments performed by ReproHum partner labs. All experiments were in the English language. The number of evaluators sometimes varies because some original studies did not control for this property, but rather allowed as many crowd-source participants to rate as many items as they wished. An item is defined as one system output evaluated absolutely, or a set of system outputs evaluated relatively.

by reporting the pairwise Pearson's r and Spearman's ρ correlation coefficients (Type II QRA) and the proportion of findings upheld (Type IV QRA). (See also Section 2.2.) In the present context, we consider each pairwise system ranking to be one finding. All scores are recomputed by us from the results reported in participants' papers, and those in the original studies.

As noted above, we report Type I, II, and IV QRA results only. This is because in most cases there are no Type III results, and in some cases where there are Type III results we do not have access to all of the raw annotations from the original studies (which would be needed in order to calculate Type III QRA).

3.1. Track B: ReproHum Partners

In this section, we summarise results from the reproduction studies performed by ReproHum partner labs reporting in Track B. We have five pairs of such studies, and four single studies where a second lab has either not yet completed and/or been assigned.

3.1.1. Liu et al. (2021)

In this experiment, participants were shown pairs of outputs from a new controlled text generation system (DExperts) and four different baselines. They were then asked which is more **fluent**. The following table shows the proportion of times DExperts was preferred over (>), considered equally good as (=), or dispreferred (<), over each of the four baselines, in original study (abbreviated *O*), reproduction 1 or *R1* (Dinkar et al., 2024), and reproduction 2 or *R2* (González Corbelle et al., 2024). The highest such proportion is highlighted in boldface. The last column shows the corresponding CV* (n=3) values for each row, finding overall a medium to poor degree of reproducibility.

System	O	R1	R2	CV*
DExperts > GPT-2	0.30	**0.39**	**0.35**	15.90
GPT-2 = DExperts	**0.40**	0.23	0.32	32.83
GPT-2 > DExperts	0.30	0.38	0.33	14.67
DExperts > DAPT	0.26	**0.42**	0.30	31.16
DAPT = DExperts	**0.39**	0.19	0.29	42.15
DAPT > DExperts	0.35	0.40	**0.41**	10.16
DExperts > PPLM	**0.37**	**0.47**	**0.39**	15.78
PPLM = DExperts	0.33	0.19	0.28	32.52
PPLM > DExperts	0.31	0.33	0.33	4.37
DExperts > GeDi	**0.36**	**0.45**	**0.36**	16.29
GeDi = DExperts	0.35	0.20	0.29	32.96
GeDi > DExperts	0.28	0.35	0.35	15.12
Mean CV*	–	–	–	21.99

In terms of Type II QRA, the correlations between each pair of columns above are as shown in the next table below. We can see that both r and ρ are negative for O and R1, and around 0 for O and R2.

In stark contrast, they are both medium to strong in the positive direction for R1 and R2.

Study A	Study B	r	ρ	Type IV
O	R1	-0.36	-0.18	2/4
O	R2	0.07	0.01	2/4
R1	R2	0.75	0.79	3/4

The above table also includes Type IV assessment, which assesses the proportion of times the pairwise system rank (e.g. DExperts was found to be better than PPLM) was upheld by a reproduction experiment. For this particular experiment we determined pairwise system rank as the relationship (>, <, =) that was selected most often by participants for a given pair of systems. In this way, we can see that both reproductions confirmed 50% (2/4) findings from the original experiment (the same two in both cases), while they agreed more with each other than the original study.

3.1.2. Hosking et al. (2022)

Here, participants were shown pairs of outputs from paraphrase generation systems and asked which best **preserves the meaning** of the input text. The below table shows scores that represent the strength with which a system was (dis)preferred on a scale from -100 to +100 (negative meaning dispreferred), alongside the corresponding CV* (n=3) values, for *O* (the original study), *R1* (Arvan and Parde, 2024), and *R2* (Watson and Gkatzia, 2024), finding a good degree of reproducibility at the level of system scores, with uniformly low CV*.

System	O	R1	R2	CV*
VAE	36.00	37.04	23.00	7.24
Latent BoW	-16.00	-14.52	-8.67	5.45
Separator	-24.00	-29.78	-17.89	9.55
HRQ-VAE	4.00	7.26	3.56	2.35
Mean CV*	–	–	–	6.15

The correlations (Type II QRA) between all experiments are near perfect, and the pairwise ranks of systems (Type IV QRA) are confirmed in all cases:

Study A	Study B	r	ρ	Type IV
O	R1	0.99	1.00	6/6
O	R2	0.99	1.00	6/6
R1	R2	0.99	1.00	6/6

With all three QRA measures across both reproductions strongly confirming the original results, Hosking et al. has one of the three highest overall degree of reproducibility of any of the human evaluations in ReproNLP'24 (the other two being for Shardlow & Narwaz for ease of understanding, and Yao et al. for readability, below).

3.1.3. Feng et al. (2021)

For this experiment, participants were asked to rate system outputs on a scale of 1 (worst) to 5 (best) the **informativeness** of paragraph-sized summaries of multi-page meeting transcriptions. The below table shows the mean system scores from *O* (the original study), *R1* (Fresen et al., 2024), and *R2* (Lango et al., 2024), alongside the corresponding CV* (n=3) values, showing reproducibility for system scores across the board.

System	O	R1	R2	CV*
Golden	4.70	2.40	4.60	54.80
PGN	2.92	2.18	1.53	70.26
HMNet	3.52	2.20	2.68	45.37
PGN(DKE)	3.20	2.18	1.93	57.24
PGN(DRD)	3.15	3.00	1.90	49.56
PGN(DTS)	3.05	2.27	1.85	53.55
PGN(DALL)	3.33	2.52	1.85	57.83
Mean CV*	–	–	–	55.52

The next table below shows that despite much lower scores for all but the 'Golden' system, strong correlations are seen between the original study (O) and R2. However, correlations between O and R1, and between R1 and R2, are close to 0 (no correlation).

Study A	Study B	r	ρ	Type IV
O	R1	0.01	0.27	12/21
O	R2	0.99	0.85	18/21
R1	R2	-0.03	0.11	11/21

This picture is somewhat confirmed by the Type IV QRA scores which show best confirmation of results for R2 but interestingly also show that R1, despite the other QRA results above, still confirmed about half the findings from O.

3.1.4. Atanasova et al. (2020)

Here, participants were asked to rank the justifications generated by three different systems in terms of their **coverage** relative to an input claim. The below table shows the mean rank for each system from *O* (the original study), *R1* (Loakman and Lin, 2024), and *R2* (Gao et al., 2024), alongside the corresponding CV* (n=3) values. The latter show the degree of reproducibility of the mean system rank to be good to medium for the two Explain systems, but poor for the Just system.[4]

[4]Note that here we have a question mark over the accuracy of the scores reported in the original study. We had the raw responses from the original experiment available to us and both reproducing teams recalculated system scores on this basis, with neither team matching the original results (or each other).

System	O	R1	R2	CV*
Just	1.48	1.62	2.18	59.58
Explain-Extr	1.89	2.05	1.93	10.64
Explain-MT	1.68	1.78	1.62	14.25
Mean CV*	–	–	–	28.16

As the Type II/IV table below shows, strong correlations were found, and all findings were confirmed, between *O* (the original study) and *R1*. However, both QRA measures were very poor for O and R2, and also R1 and R2.

Study A	Study B	r	ρ	Type IV
O	R1	0.99	1.00	3/3
O	R2	-0.43	-0.50	1/3
R1	R2	-0.31	-0.50	1/3

3.1.5. August et al. (2022)

For this experiment, participants were asked to rate the **fluency** of generated scientific definitions on a scale of 1 (not at all) to 4 (very). The below table shows the mean system scores, alongside the corresponding CV* (n=3) values, for *O* (the original study), *R1* (van Miltenburg et al., 2024), and *R2* (Li et al., 2024), finding a medium to borderline poor degree of reproducibility for all systems, albeit better for the SVM system.

System	O	R1	R2	CV*
SVM	3.71	3.12	3.02	19.96
GeDi	3.20	2.57	2.40	29.90
DExpert	2.33	2.28	1.81	30.76
Mean CV*	–	–	–	26.87

Correlations were very strong between all studies, with the order of pairwise ranks (Type IV) confirmed in all cases:

Study A	Study B	r	ρ	Type IV
O	R1	0.95	1.00	3/3
O	R2	0.99	1.00	3/3
R1	R2	0.99	1.00	3/3

3.1.6. Castro Ferreira et al. (2018)

Participants were shown outputs of a data-to-text system and asked to rate their **clarity** on a 1 (very bad) to 7 (very good) scale. The below table shows the mean system ratings, alongside the corresponding CV* (n=2) values, for *O* (the original study) and *R1* (Mahamood, 2024), finding an excellent degree of reproducibility across the board.

System	O	R1	CV*
OnlyNames	4.90	4.92	0.51
Ferreira	4.93	4.69	6.28
NeuralREG+Seq2Seq	4.97	4.97	0.00
NeuralREG+CAtt	5.26	4.97	7.03
NeuralREG+HierAtt	5.13	5.04	2.20
Original	5.42	5.22	4.62
Mean CV*	–	–	3.44

Correlations between R1 and O were medium-strong, and 80% (12/15) of pairwise system rankings were confirmed:

Study A	Study B	r	ρ	Type IV
O	R1	0.78	0.84	12/15

3.1.7. Kasner and Dusek (2022)

This experiment was originally an error analysis performed by the authors, although it fits the definition of a human evaluation used in the ReproHum Project. Participants (the two authors) were shown the input and outputs from data-to-text systems and asked to count the **number of repetitions** in the outputs. The below table shows repetition error counts for different systems and corresponding CV* (n=2) values for *O* (the original study) and *R1* (Klubička and Kelleher, 2024), finding extremely poor degrees of reproducibility at the level of system scores.

System	O	R1	CV*
Full-3-Stage	0	13	199.40
Full-2-Stage	1	11	166.17
Full-1-Stage	79	156	65.34
Filtered-3-Stage	0	9	199.40
Filtered-2-Stage	0	10	199.40
Filtered-1-Stage	41	84	68.59
Mean CV*	–	–	149.72

However, Pearson's is very nearly perfect,[5] with Spearman's a less strong 0.82, and 73% of pairwise system ranks confirmed:

Study A	Study B	r	ρ	Type IV
O	R1	0.99	0.82	11/15

We thus have a mixed picture here with system score level reproducibility extremely poor, about three quarters of findings confirmed, a reasonably strong rank correlation and near perfect product-moment correlation.

[5] It would round up to 1.00 but our rounding policy keeps it at 0.99 to avoid giving a false impression. See Appendix B.

3.1.8. Shardlow and Nawaz (2019)

In this experiment, participants were shown medical texts, from four text simplification systems, and asked to rank them from best to worst in terms of **ease of understanding**. The below table shows mean system rank for each system and the corresponding CV* (n=2) values for *O* (the original study) and *R1* (Li et al., 2024), finding a good degree of reproducibility for mean rank when considering the two *NTS* systems, and an excellent degree for the *ORIG* and *PTB* systems.

System	O	R1	CV*
NTS+PT	1.93	1.82	12.53
NTS	2.34	2.46	8.55
ORIG	2.79	2.76	1.69
PTB	2.94	2.96	1.02
Mean CV*	–	–	5.95

There were also strong correlations between the two studies and the pairwise ranks were confirmed in all cases:

Study A	Study B	r	ρ	Type IV
O	R1	0.98	1.00	6/6

With all three QRA measures across both reproductions strongly confirming the original results, Shardlow & Narwaz has one of the three highest overall degree of reproducibility of any of the human evaluations in ReproNLP'24 (the other two being for Hosking et al. for meaning preservation, above, and Yao et al. for readability, below).

3.1.9. Gabriel et al. (2022)

For this experiment, participants were shown output texts from three systems that generate statements of the writer's intents given news headlines as input. Their task was to decide whether the text was socially acceptable or not. The below table shows the percentage of times a system was deemed socially acceptable alongside the corresponding CV* (n=2) values for *O* (the original study) and *R1* (Mahlaza et al., 2024), finding a good to medium degree of reproducibility.

System	O	R1	CV*
T5-base	75.30	68.67	9.18
T5-large	74.66	68.31	8.86
GPT-2 (large)	74.66	65.30	13.34
Mean CV*	–	–	10.46

Pearson's r was only moderate, with a stronger Spearman's ρ, and 67% of findings confirmed:

Study A	Study B	r	ρ	Type IV
O	R1	0.58	0.87	2/3

3.2. Track B: Other teams

Track B of ReproNLP was also open to non-ReproHum partner labs. Participants in this track reproduce experiments of their choice from the same set of Track B papers, but do not necessarily follow the exact common approach (Appendix A) as ReproHum partner labs do.

3.2.1. Yao et al. (2022)

Florescu et al. (2024) repeated this evaluation of generated questions and answers for children's stories, performing the evaluation for all three quality criteria in the original study.

Readability: Participants are asked to rate what was named the "readability" of the question-answer pair. The exact prompt used, however, was "readability(grammarly [sic] correct and clear language. worst 1 to 5)", which references three different quality criteria (readability, grammaticality and clarity), making it a clear example of the confusion in quality criteria found by Howcroft et al. (2020).

The below table shows the mean readability ratings for each system alongside CV* (n=2) values for the original study *(O)* and *R1* (Florescu et al., 2024), finding a good degree of reproducibility.

System	O	R1	CV*
Ours	4.71	4.52	5.24
PAQ Baseline	4.08	4.17	2.87
Groundtruth	4.95	4.71	6.25
Mean CV*	–	–	4.79

Correlations were near perfect, and the pairwise ranks of systems were confirmed in all cases:

Study A	Study B	r	ρ	Type IV
O	R1	0.99	1.00	3/3

With all three QRA measures strongly confirming the original results, Yao et al. has one of the three highest overall degree of reproducibility of any of the human evaluations in ReproNLP'24 (the other two being for Hosking et al. for meaning preservation, and Shardlow & Narwaz for ease of understanding, above).

Relevancy (Question): The following table shows the mean question relevancy ratings for systems alongside the corresponding CV* (n=2) values for the original study *(O)* and *R1* (Florescu et al., 2024), finding only a medium degree of reproducibility for the system outputs, with a good degree of reproducibility for the human-authored ground truth.

System	O	R1	CV*
Ours	4.39	3.83	17.95
PAQ Baseline	4.18	3.61	19.63
Groundtruth	4.92	4.71	5.49
Mean CV*	–	–	14.36

However, correlations are still perfect, with the pairwise ranks of systems confirmed in all cases:

Study A	Study B	r	ρ	Type IV
O	R1	0.99	1.00	3/3

Relevancy (Answer): Finally, for the relevancy of the answer, system scores are again less reproducible than the ground truth, with a medium to poor degree of reproducibility for the systems, and good degree of reproducibility for the ground truth:

System	O	R1	CV*
Ours	3.99	3.20	30.35
PAQ Baseline	3.90	3.20	27.37
Groundtruth	4.83	4.46	10.12
Mean CV*	–	–	22.61

Correlations are strong with Spearman's lower, and the two systems being scored identically in the reproduction experiment, as opposed to only being similar in the original study; this also affects the pairwise rankings confirmed score (Type IV) for the two systems:

Study A	Study B	r	ρ	Type IV
O	R1	0.99	0.87	2/3

4. Track A

We accepted two submissions in the open track, where participants could carry out reproduction experiments for any paper, focusing on human and/or metric-based evaluations.

4.1. Chakravarthi et al. (2020)

In the original study, a code-mixed Malayalam language dataset was annotated for sentiment (5 labels) by human participants and then used to train classifiers which were in turn evaluated by automatic metrics. Sasidharan Nair et al. (2024) recreate this complete pipeline.

4.1.1. Label counts (human evaluation)

The count of labels recorded in the reproduction varied greatly from the original study, resulting in a moderate degree of reproducibility for some labels, and a very poor degree of reproducibility for others, as shown in this table:

System	O	R1	CV*
Positive	565	626	10.21
Negative	138	162	15.95
Mixed Feelings	70	144	68.95
Neutral	398	327	19.53
Non-Malayalam	177	89	65.97
Mean CV*	–	–	36.12

However, the correlations were strong, with pairwise ranks also confirmed for most labels. Note that rather than comparing systems we are comparing label counts of an annotated corpus.

Study A	Study B	r	ρ	Type IV
O	R1	0.94	0.70	8/10

4.1.2. Automated metrics

After completing their re-annotation of the corpus, Sasidharan Nair et al. (2024) then evaluated LR and BERT sentiment classifiers on both the original corpus and their newly created one, using F1 score. The below table shows Mean CV* for *O* (results from the original paper) and *Re-Imp* (a re-implemented classifier by Sasidharan Nair et al. (2024) but trained on the original corpus). The table also shows the Mean CV* for *Re-Imp* and *Re-Ann*, where *Re-Ann* refers to the re-implemented classifier trained on the re-annotated corpus.

This reproduction study clearly shows the effect that the reproducibility of human evaluation can have on the reproducibility of downstream tasks.

Classifier	Study A	Study B	Mean CV*
LR	O	Re-imp	7.65
BERT	O	Re-imp	22.73
LR	Re-imp	Re-ann	47.70
BERT	Re-imp	Re-ann	24.10

Note that we calculate Mean CV* in this report differs from how Sasidharan Nair et al. (2024) calculate it, where the macro and weighted averages of F1 score are calculated first, with the mean CV* then calculated at that level.

In terms of Type IV results, the reproducing team find that by both macro-average and weighted-average, for both setups (*O* vs *Re-imp* and *Re-Imp* vs *Re-ann*), the BERT classifier is always better than LR. This corresponds to 8/8 findings upheld. Even at the per-label level, 9/10 findings are upheld for *O* vs *Re-imp*, and 8/10 for *Re-Imp* vs *Re-ann*.

4.2. Gu et al. (2022, 2023)

Lorandi and Belz (2024) reproduce original studies found in Gu et al. (2022) and Gu et al. (2023). They calculate the CV* between original and reproduc-

ReproNLP 2024							mean CV*		
Orig Study // *Repro a / Repro b* **measurands**	**3.2.1**	**4.3.4**	**4.3.8**	**4.1.1**	**4.1.2**	**4.1.3**	**a(n=2)**	**b(n=2)**	**n=3**
Liu et al. (2021) // *Dinkar et al. (2024) / González Corbelle et al. (2024)* **Fluency**	UNK / 96 / 90	A,B,Tie	RQE	Good	Both	iiOR	34.55	14.58	21.99
Hosking et al. (2022) // *Arvan and Parde (2024) / Watson and Gkatzia (2024)* **Preservation of meaning**	UNK / 180 / 180	A,B	RQE	Good	Cont	RtI	3.37	6.62	6.15
Feng et al. (2021) // *Fresen et al. (2024) / Lango et al. (2024)* **Informativeness**	4 / 4 / 4	1-5	DQR	Good	Cont	RtI	52.07	70.53	55.52
Atanasova et al. (2020) // *Loakman and Lin (2024) / Gao et al. (2024)* **Coverage**	3 / 3 / 3	1-3	RQE	Good	Cont	RtI	18.49	32.56	28.16
August et al. (2022) // *van Miltenburg et al. (2024) / Li et al. (2024)* **Fluency**	2 / 2 / 2	1-4	DQE	Good	Both	iiOR	20.50	40.62	26.87
Castro Ferreira et al. (2018) // *Mahamood (2024)* **Clarity**	60 / 60	1-7	DQE	Good	Both	iiOR	3.44	-	-
Kasner and Dusek (2022) // *Klubička and Kelleher (2024)* **Number of redundancies per system**	2 / 2	count	Count	Good	Cont	iiOR	149.72	-	-
Shardlow and Nawaz (2019) // *Mondella et al. (2024)* **Ease of understanding**	98 / 40	1-4	RQE	Good	Both	iiOR	5.95	-	-
Gabriel et al. (2022) // *Mahlaza et al. (2024)* **Social acceptability**	UNK / 42	Yes,No	DQE	Feature	Both	EFoR	10.46	-	-

Table 2: Summary of some properties of ReproNLP experiments performed by ReproHum partner labs, alongside mean CV* (n=2, or n=3; shown in different columns because different samples sizes are not directly comparable). 3.2.1 = number of evaluators in original/reproduction experiment; 4.3.4 = List/range of possible responses; 4.3.8 = Form of response elicitation (DQE: direct quality estimation, RQE: relative quality estimation, Cl/Lab: classification/labelling, Count: counting occurrences in text); 4.1.1 = Correctness/Goodness/Features; 4.1.2 = Form/Content/Both; 4.1.3 = each output assessed in its own right (iiOR) / relative to inputs (RtI) / relative to external reference (EFoR).

tion scores for each evaluation measure. Based on these, we include per-system mean CV* scores below, along with the maximum and minimum.

System	CV*		
	Mean	Min	Max
Multi-CTG	1.68	0.34	5.56
Prior-CTG	1.25	0.00	4.14
Prior-CTG+optim	1.28	0.00	3.81

Moreover, Lorandi and Belz (2024) report correlations (Type II) in excess of 0.99 for all reproductions and all findings (Type IV) as upheld.

A full breakdown of per-measure CV* results, as well as other analyses, can be found in their paper.

5. Reproducibility by Quality Criterion and other properties

We saw a wide variety of different degrees of reproducibility for the different human evaluations

in previous sections. It seems likely that these differences in degree of reproducibility are explainable by differences between the evaluations. In the QRA++ approach (Belz, 2022; Belz and Thomson, 2023), as in metrology on which it is based, such differences are captured by 'conditions of measurement,' and HEDS was designed to capture these.

Table 2 shows some of the main HEDS properties of the experiments repeated by ReproHum partner labs, along with mean CV* values calculated as follows:

- **a(n=2)**: the mean of two-way CV* values between *Orig Study* and *Repro a*.
- **b(n=2)**: the mean of two-way CV* values between *Orig Study* and *Repro b* (if there was a *Repro b*).
- **(n=3)**: the mean of three-way CV* values between *Orig Study*, *Repro a*, and *Repro b* (if there were 3 sets of results).

What we are looking for in this table is any indication that one of the HEDS properties affects experiment-level mean CV* (last three columns). One such property is number of evaluators (HEDS Question 3.2.1): the pattern is for larger number of evaluators to be associated with better reproducibility, with the exception of Fluency in Liu et al. which bucks the trend somewhat.

Another trend that is observable is that evaluations that are more cognitively complex tend to have poorer reproducibility than cognitively simpler evaluations. An extreme example of this is Kasner and Dusek's count of redundancies per system, which are very hard to match in reproductions. Similar results were obtained in an earlier pair of reproductions of an error analysis experiment, where some of the error counts also reached above 140 CV* (**?**). Another example is Informativeness (fourth from top in table).

Cognitively simpler assessments like Clarity and Fluency have better score-level reproduciblity. This is a trend that we have consistently observed across multiple reproduction experiments. Note however that here too Fluency in Liu et al. bucks the trend which may be explained by other experimental properties we are not examining here.

6. Discussion

As we saw in previous sections, different types of QRA++ assessments (Type I, II, and IV) can show very different degrees of reproducibility for sets of reproductions for the *same* original experiment. For example, for Social Acceptability in Gabriel et al. 2022, CV* levels were reasonable but Pearson's was only 0.58.

Another example is Fluency in August et al. (2022) where the CV* values are quite poor, but Type II and IV reproducibility is excellent.

It can also be the case that one reproduction for the same original experiment indicates excellent reproducibility and another shows very poor reproducibility, as was the case for Content Coverage Atanasova et al. 2020.

The latter observation (observed previously) indicates the importance of conducting more than one reproduction experiment. An alternative may be to increase the number of individual assessments carried out (Simonsohn, 2015), but it is not clear how additional assessments should be created (more evaluators, more system outputs, both?).

The differences between results from different types of QRA assessment highlight that each assesses a different aspect of reproducibility: Type I/CV* looks at how close individual aligned scores are; Type II/correlations look at how similar relative increases and decreases are in aligned sets of scores; and Type IV/findings abstracts away from scores altogether to look at findings which we here interpret as pairwise system ranks, i.e. which of two systems performs better.

Ultimately, it is the latter, pairwise system ranks, that we care most about in many contexts. What matters is not necessarily maximising the rank correlation, but the proportion of pairwise ranks that are the same (although clearly these are linked).

In the previous section we looked at the effect different experiment properties may have on reproducibility. However, these cannot explain differences between reproductions of the same original experiment where properties are the same. This means that there are other factors affecting reproducibility, e.g. evaluator sampling, and quality of the reproduction experiment. All of this clearly makes it harder to link properties with reproducibility.

Given the finding that score-level QRA can show poor reproducibility even where all findings are upheld, it might be questioned whether inter-annotator agreement (IAA), commonly used as an indicator of experiment quality, is really the right measure. It might be that reproducibility tests assessing multiple different QRA measures are a better pre-experiment test of quality.

7. Conclusion

Shared task result reports tend to be written under considerable pressure of time, and the present paper is no exception. We will conduct additional analyses and more in-depth explorations of our data in due course, as well as reporting the results from the second batch of ReproHum multi-lab multi-test study experiments once all have been completed. The latter will provide more robustly

quantifiable assessment of the impact of selected experiment properties on reproducibility.

This year's edition of the shared task has once again highlighted the considerable extent to which results (i) from different reproduction experiments of the same original experiments, and (ii) from different types of QRA analysis, can differ. This can be interpreted as meaning that we should conduct multiple reproduction experiments, and multiple types of QRA analysis, respectively.

There continues to be little standardisation in evaluation practices, and quality criteria names and definitions in particular, in human evaluation in NLP, despite numerous surveys and studies (Belz et al., 2020; Howcroft et al., 2020; van der Lee et al., 2021; Gehrmann et al., 2023) calling for more standardisation to improve quality and reliability. In the present context, lack of standardisation also has the effect of muddying the waters with respect to conclusions about which quality criteria are associated with better reproducibility: if it is unclear, e.g. due to mere name differences, whether the same quality criterion was in fact assessed, it is hard to draw accurate conclusions beyond the individual experiment.

All in all, it seems clear that human evaluation in NLP would benefit from more standardisation in experimental design and execution, for better comparability, but also so that reproducibility, hence reliability, of standard methods can be established, and once established, benefited from.

Acknowledgments

We thank the authors of the original papers that were offered for reproduction in ReproNLP 2024. And of course the authors of the reproduction papers, without whom there would be no ReproHum project and no ReproNLP shared task.

Our work was carried out as part of the ReproHum project on Investigating Reproducibility of Human Evaluations in Natural Language Processing, funded by EPSRC (UK) under grant number EP/V05645X/1. In particular, we thank our numerous collaborators from NLP labs across the world who carried out the reproductions in Track C as part of the second batch of coordinated reproductions in the ReproHum project.

The ReproNLP work also benefits from the work being carried out in association with the ADAPT SFI Centre for Digital Media Technology which is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

Bibliographical References

Mohammad Arvan and Natalie Parde. 2024. Human evaluation reproduction report for "hierarchical sketch induction for paraphrase generation". In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

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. 2021a. 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, Simon Mille, and David Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021b. The reprogen shared task on reproducibility of human evaluations in nlg: Overview and results. *INLG 2021*, page 249.

Anya Belz, Anastasia Shimorina, Maja Popovic, and Ehud Reiter. 2022. The 2022 reprogen shared task on reproducibility of evaluations in nlg: Overview and results. *INLG 2022*, page 43.

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 Eval-*

uation of NLP Systems, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 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.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Tanvi Dinkar, Gavin Abercrombie, and Verena Rieser. 2024. Dexpert evaluation? reproducing human judgements of the fluency of generated text. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. 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 1479–1491, Online. Association for Computational Linguistics.

Andra-Maria Florescu, Marius Micluta-Campeanu, and Liviu P. Dinu. 2024. Once upon a replication: It is humans' turn to evaluate ai's understanding of children's stories for qa generation. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Vivian Fresen, Mei-Shin Wu-Urbanek, and Steffen Eger. 2024. Humeval 24 reproduction report for paper 0043: Language model as an annotator: Exploring dialogpt for dialogue summarization. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland. Association for Computational Linguistics.

Mingqi Gao, Jie Ruan, and Xiaojun Wan. 2024. A reproduction study of the human evaluation of the coverage of fact checking explanations. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Int. Res.*, 77.

Javier González Corbelle, Ainhoa Vivel Couso, Jose Maria Alonso-Moral, and Alberto Bugarín-Diz. 2024. Reproducing the human evaluation of the dexperts controlled text generation method. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. 2022. A distributional lens for multi-aspect controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1043, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. 2023. Controllable text generation via probability density estimation in the latent space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12590–12616, Toronto, Canada. 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.

Zdeněk Kasner and Ondrej Dusek. 2022. Neural pipeline for zero-shot data-to-text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3914–3932, Dublin, Ireland. Association for Computational Linguistics.

Filip Klubička and John Kelleher. 2024. Reprohum #1018-09: Reproducing human evaluations of redundancy errors in data-to-text systems. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Mateusz Lango, Patricia Schmidtova, Simone Balloccu, and Ondrej Dusek. 2024. Evaluating summarization models: investigating the impact of education and language proficiency on reproducibility. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Yiru Li, Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2024. Report for reprohum project 0033. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706, Online. Association for Computational Linguistics.

Tyler Loakman and Chenghua Lin. 2024. Human evaluation reproduction report for generating fact checking explanations. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Michela Lorandi and Anya Belz. 2024. Reproducing the metric-based evaluation of a set of controllable text generation techniques. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Saad Mahamood. 2024. Reproducing human evaluations of end-to-end approaches to referring expression generation. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Zola Mahlaza, Toky Raboanary, Kyle Seakgwa, and C. Maria Keet. 2024. Another evaluation of readers' reactions to news headlines. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Irene Mondella, Huiyuan Lai, and Malvina Nissim. 2024. Report for reprohum project 0892-01. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Sachin Sasidharan Nair, Tanvi Dinkar, and Gavin Abercrombie. 2024. Exploring reproducibility of human-labelled data for code-mixed sentiment analysis. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 380–389, Florence, Italy. 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.

Uri Simonsohn. 2015. Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5):559–569. PMID: 25800521.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Martijn Goudbeek, Emiel Krahmer, Chris van der Lee, Steffen Pauws, and Frédéric Tomas. 2024. How reproducible are fluency ratings of generated text? a reproduction of august et al. 2022. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Lewis Watson and Dimitra Gkatzia. 2024. Reprohum #0712-01: Reproducing human evaluation of meaning preservation in paraphrase generation. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. Association for Computational Linguistics.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

A. The ReproHum Common Approach to Reproduction

In order to ensure comparability between studies, we agreed the following common-ground approach to carrying out reproduction studies:

1. Plan for repeating the original experiment in a form that is as far as possible identical to the original experiment, ensuring you have all required resources in place, then apply to research ethics committee for approval. If any aspect of the original experiment is unclear, contact the ReproHum coordinator who will either obtain clarification from the author, or create a sensible design that will then be used by all partner labs reproducing that experiment.
2. If participants were paid during the original experiment, determine pay in accordance with the ReproHum common procedure for calculating fair pay (Belz et al., 2023).
3. Following ethical approval start the reproduction study following the steps below. Contact the ReproHum team with any questions rather than the original authors, as they have already provided us with all the resources and information they have. Don't communicate with other ReproHum teams about their reproduction studies. This is to avoid inadvertently affecting outcomes.
4. Complete HEDS datasheet.
5. Identify the following types of results reported in the original paper for the experiment:
   (a) Type I results: single numerical scores, e.g. mean quality rating, error count, etc.
   (b) Type II results: sets of numerical scores, e.g. set of Type I results .
   (c) Type III results: categorical labels attached to text spans of any length.
   (d) Qualitative conclusions/findings stated explicitly in the original paper.[6]
6. Carry out the allocated experiment exactly as described in the HEDS sheet.
7. Report the results in the following form:
   (a) Description of the original experiment.
   (b) Description of any differences in your repeat experiment.
   (c) Side-by-side presentation of all results (8a-d above) from original and repeat experiments, in tables.
   (d) Report quantified reproducibility assessments as follows:
      i. Type I results: Small-sample oefficient of variation CV* (Belz, 2022).
      ii. Type II results: Pearson's r, Spearman's ρ.
      iii. Type III results: Multi-rater: Fleiss's κ; Multi-rater, multi-label: Krippendorff's α.

[6] We now call these Type IV results.

iv. Conclusions/findings: Side-by-side summary of conclusions/findings that are / are not confirmed in the repeat experiment.

B. Rounding Policy

The python script used to calculate results uses HALF_UP rounding rather than the python default of bankers rounding. Numbers are only ever rounded at the stage of presentation, i.e., the full-precision CV* values are used to calculated the means, rather than the 2 decimal place ones.

For Pearson and Spearman correlations we never round up from 0.99 in order to avoid giving the impression of a perfect correlation where one does not exist.