The Shared Task on Reproducibility of Evaluations in NLP (ReproNLP) 2026: Overview and Results

Anya Belz[1], Craig Thomson[1], Javier González Corbelle[1,2]
[1]ADAPT, Dublin City University
[2]CiTIUS, Universidade de Santiago de Compostela, Spain
Corresponding author: anya.belz@dcu.ie

Abstract

We present the 2026 Shared Task on Reproducibility of Evaluations in NLP (ReproNLP'26) which followed on from five predecessor shared tasks on reproducibility of evaluations, ReproNLP'25, ReproNLP'24, 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 the topic across the two fields. We describe the ReproNLP'26 shared task, summarise results from the reproduction studies submitted, and provide additional comparative analysis of their results.

1 Introduction

Natural language processing (NLP) and machine learning (ML) still have some way to go in fully addressing how to design, implement and document experiments in order to achieve high degrees of reproducibility. Authors generally still do not fully share information and resources when reporting experiments (Belz et al., 2023a; Schmidtova et al., 2025; Thomson et al., 2026), and there is little standardisation and resource sharing in experiment design and implementation, in particular human evaluation experiments (Thomson et al., 2026). The core aim of the ReproNLP[1] shared tasks focusing on reproducibility in NLP has been to encourage, coordinate, report and analyse reproduction studies, but also to bring the results together at a higher level and identify characteristics of experiments with good reproducibility and to obtain insights into how the degree of reproducibility of experiments can be improved.

The sixth shared task event in the ReproGen/ReproNLP series is something of a milestone. The ReproNLP shared tasks (and their predecessors REPROLANG and ReproGen) have resulted in several dozen individual reproducibility studies (Caines and Buttery, 2020; Belz et al., 2021, 2022b; Belz and Thomson, 2023, 2024; Belz et al., 2025), to which we add a further seven this year, together providing a rich source of repeated experiments for the study of reproducibility in NLP.

Since the 2021 ReproGen shared task, we have quantitatively measured reproducibility using the Quantified Reproducibility Assessment (QRA) formalism (Belz et al., 2022a; Belz and Thomson, 2026). Under the umbrella of the ReproHum[2] multi-lab multi-test (MLMT) study (Belz et al., 2023a), we have been reporting (in the ReproHum Track) on intermediate results from a long-term coordinated, controlled study of how different factors affect reproducibility. To facilitate collaboration on this study, we have developed the ReproHum common approach (Appendix B) to reproduction studies which ReproNLP participants are encouraged to adopt. Taken together, researchers now wishing to conduct reproduction studies can avail themselves of a toolbox of resources to reduce knowledge and effort thresholds in reproducibility.

The seven new reproduction studies reported in ReproNLP'26 add new data points to the body of directly comparable evaluations available for investigations of reproducibility. Our new analyses point towards further reasons for worse/better reproducibility of evaluations. The six studies conducted in the ReproHum track are shown in overview in Table 1; additionally we present one study in the Open Track.

We start in Section 2 with a description of the organisation and structure of ReproNLP'26. We then report results by track, starting with the ReproHum Track (Sections 3 and 4) and followed by the Open Track (Section 5). We examine connections

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

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

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1055–1070
July 4, 2026 ©2026 Association for Computational Linguistics

between experiment properties and reproducibility (Section 6), and conclude with a discussion (Section 7) and directions for future work (Section 8).

2 ReproNLP 2026

ReproNLP 2026 followed the same two-track format as recent editions: an unshared open track (Track A) and a shared ReproHum track (Track B).

- **A Open Track:** Reproduce previously reported evaluation results from any paper, 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, repeat one or more of the studies and compare results, using the information provided by the ReproNLP organisers only, and following the ReproHum common reproduction approach.

Track B forms part of the ReproHum project[3] and the original 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., 2023b).

This year, Track A had one submission, the study by Ardestani et al. (2026). In Track B, submissions were received for each of the following experiments from the ReproHum set:

1. **Reif et al. (2022)**: *A Recipe for Arbitrary Text Style Transfer with Large Language Models*.

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

2. **August et al. (2022)**: *Generating Scientific Definitions with Controllable Complexity*.

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

3. **Yao et al. (2022)**: *It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books*.

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

4. **Gabriel et al. (2022)**: *Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines*.

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

5. **Castro Ferreira et al. (2018)**: *NeuralREG: An end-to-end approach to referring expression generation*.

   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 repeat assessments of individual quality criteria on individual datasets as indicated above (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 (size, evaluator type, cognitive complexity). For each experiment, we obtained agreement from the original authors to use their experiment in the ReproHum study and ReproNLP shared task. They provided very detailed information about the experiments which were shared with all participants.

Each of these experiments is being repeated in two separate reproduction studies in ReproHum. Those that have completed in the current batch (and were not previously reported as part of ReproNLP'24 or ReproNLP'25) are included here in the ReproNLP'26 report. All experiments have also been open to other ReproNLP participants for reproduction since 2023. In this way, we have accumulated some additional, albeit less tightly controlled, results for some of the ReproHum experiments, enabling us to present degree of reproducibility results for up to five aligned experiments.

2.1 Approach to reproduction and reproducibility assessment

All participants were encouraged to complete a ReproHum HEDS datasheet[4] (Belz and Thomson, 2025), and to follow the ReproHum Common Approach to reproduction laid out in Appendix B which includes QRA (Belz and Thomson, 2026), a set of quantitative reproducibility assessment measures for four common types of results in NLP/ML

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

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

Original Study	Quality Criterion	Num sys-tems	Items per system	Labs reproducing study for ReproNLP 2026
August et al. (2022)	Factual Truth (all)	3	100	a) University of Bucharest* b) **McGill University / York University**
August et al. (2022)	Factual Truth (one)	3	100	a) University of Bucharest* b) **McGill University / York University**
Castro Ferreira et al. (2018)	Clarity	6	24	a) Trivago* b) **Ruhr University Bochum**
Gabriel et al. (2022)	Social acceptability	3	196	a) University of Cape Town* b) **Central China Normal University**
Reif et al. (2022)	Meaning Preservation	4	50	a) Charles University Prague* b) **Shopify**
Yao et al. (2022)	Readability	3	120.33	a) Marburg University* b) **Zurich University of Applied Sciences**

Table 1: ReproNLP 2026 experiments performed by ReproHum partner labs. All experiments were in the English language. An item is defined as one system output evaluated absolutely, or a set of system outputs evaluated relatively. Labs marked with * correspond to papers from previous years; labs submitting to ReproNLP in 2026 are shown in bold.

that accommodates multiple reproduction studies of the same original work and produces results that are comparable across different such sets of reproductions.

In this report we analyse all submissions in terms of QRA measures recomputed by us to ensure comparability across submissions. In brief summary (for full details see Belz and Thomson (2026)), 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. a set of Type I results for comparable systems.
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 and Thomson, 2026).
2. Type II results: Pearson's r, Spearman's ρ, Kendall's τ, Kendall's W.
3. Type III results: Fleiss's κ; Krippendorff's α.
4. Type IV results: Proportion P of identical pairwise system ranks in a set of comparable experiments.[5]

In the submissions presented in this paper we have Type I, II and IV results, so we apply the corresponding quantitative measures above. CV* is a version of the standard coefficient of variation, corrected for small samples.

In ReproNLP'26 we are for the first time using an updated version of CV* (Belz and Thomson, 2026) which avoids the overcorrection for small sample size of the previous version, now providing two formulations of the correction, one for use where normality can be assumed, and one for where it cannot.

As a consequence, CV* values in the present report are not directly comparable to values reported in previous ReproNLP reports.[6] The main results tables (with the blue shading) in this paper are automatically generated using the QRA3 tool available from `https://github.com/DCU-NLG/qra`.

The reproduction studies below that form part of the ReproHum controlled MLMT study are strictly controlled to be comparable to each other and the original work, with the proviso that since 2024, we no longer aim to achieve complete similarity between design and implementation of original and reproduction studies, or try to resolve every last bit of lack of clarity about the original experiment. In the second (current) round of ReproHum reproduction studies, we recognise that evaluation experiments should be robust to minor differences. As

[5]To obtain results that are comparable across different studies, we restrict Type IV assessment to pairwise system ranks as findings.

[6]The QRA3 tool (`https://github.com/DCU-NLG/qra` still supports the (now deprecated) previous method of calculation.)

a result, when there was insufficient clarity about how an aspect of an original experiment was implemented, partner labs drafted solutions which were then moderated by the ReproHum coordinating team to arrive at an agreed solution that both partner labs reproducing the same experiment then used. For more details on such cases, please see the individual ReproNLP participants' reports in this volume and those from 2024 and 2025.

Finally, we use the following verbal descriptors and corresponding CV* ranges in the case of *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 higher CV* scores indicate worse reproducibility.

3 Track B ReproHum MLMT Results

As outlined above, we report degree of reproducibility results below for each ReproNLP'26 submission in terms of QRA measures. For Type I measures, we are using the small sample correction option in the QRA3 tool that assumes normal distribution (Belz and Thomson, 2026) in all cases. While there is discussion regarding where to draw the line, it is generally agreed that when the items in the sample are themselves means computed over large numbers of items, then we are justified in using normality assumptions, by dint of the central limit theorem. In the ReproHum set of experiments used in reproduction studies, all such means are computed over between 150 and 480 items (average 262.13).

3.1 Castro Ferreira et al. (2018)

In this experiment, participants are shown outputs of a data-to-text system (using WebNLG'2017 data, (Gardent et al., 2017)), and rate their **Clarity** on a 1 (very bad) to 7 (very good) scale. The following table shows the mean system ratings, alongside the corresponding CV* (n=2) values, for O (the original study), R1 (Mahamood, 2024), and R2 (Langner, 2026):

System	O	R1	R2	CV*
OnlyNames	4.90	4.92	5.27	5.83
Ferreira	4.93	4.69	5.02	4.96
NREG+Seq2Seq	4.97	4.97	5.21	3.86
NREG+CAtt	5.26	4.97	5.48	6.81
NREG+HierAtt	5.13	5.04	5.31	3.73
Original	5.42	5.22	5.64	5.36
Mean CV*	–	–	–	5.09

CV* scores here indicate an excellent degree of reproducibility for system-level scores (Type I), especially considering this is a human evaluation.

The next table below compares *pairs* of aligned experiments, showing Type II (r, ρ, τ correlations) and Type IV (P, the proportion of identical pairwise system ranks) QRA scores. On both Type II and IV measures, the alignment is high while not close to 1, indicating that the original study has very good reproducibility. Moreover, we can see that O and R2 agree slightly more with each other than the other pairings.

Study A	Study B	CV*	r	ρ	τ	P
O	R1	3.06	0.783	0.841	0.690	0.800 (12/15)
O	R2	4.63	0.908	0.829	0.733	0.867 (13/15)
R1	R2	7.53	0.893	0.812	0.690	0.800 (12/15)

Table 2 shows the full QRA results, as automatically generated by the QRA3 tool. This table provides different perspectives again from the two smaller tables above: QRA is here applied, as intended, to the complete set of aligned experiments, producing degree of reproducibility assessments at the (i) system, (ii) QC, and (iii) study level. This shows us that the experiment originally created by Castro Ferreira et al. (2018), and repeated by (Mahamood, 2024) and (Langner, 2026), has a very good degree of reproducibility: Type I results (CV*, lower is better) are very good. Correlations (Type II) are strong but below 0.9, and the P score (Type IV) shows 82% shared pairwise system rankings across the three runs of the experiment.

3.2 Gabriel et al. (2022)

Participants in this experiment are shown output texts from three systems that generate statements of the writer's intents given news headlines as input, and are asked to mark system output texts for **Social acceptability**, using a binary yes/no rating instrument. The following table shows the percentages of socially acceptable outputs, alongside the corresponding CV* (n=2) values for *O* (the original study), *R1* (Mahlaza et al., 2024), and R2 (Fan and Chen, 2026), finding good to medium degrees of Type I reproducibility.

System	O	R1	R2	CV*
T5-base	75.30	68.67	66.17	7.60
T5-large	74.66	68.31	67.33	6.41
GPT2-large	74.66	65.30	58.67	13.69
Mean CV*	–	–	–	9.23

Table 2: **ReproHum 0124-03**: QRA reproducibility assessment for three comparable experiments (n=3), Castro Ferreira et al. (2018), Mahamood (2024), and Langner (2026); n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 3$)		
				System level	QC level	Study level
Type I	Clarity	OnlyNames	(mean) CV*↓	5.83	5.09	5.09
		Ferreira		4.96		
		NeuralREG+Seq2Seq		3.86		
		NeuralREG+CAtt		6.81		
		NeuralREG+HierAtt		3.73		
		Original		5.36		
Type II	Clarity	all	mean r ↑	n/a	0.861	n/a
		all	mean ρ ↑	n/a	0.827	
		all	W↑	n/a	0.876	
Type IV	Clarity	all	P ↑	n/a	0.822	0.822

Table 3: **ReproHum 0866-04**: QRA reproducibility assessment for three comparable experiments (n=3), Gabriel et al. (2022), Mahlaza et al. (2024), and Fan and Chen (2026); n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 3$)		
				System level	QC level	Study level
Type I	Social acceptability	T5-base	(mean) CV*↓	7.60	9.23	9.23
		T5-large		6.41		
		GPT2-large		13.69		
Type II	Social acceptability	all	mean r ↑	n/a	0.649	n/a
		all	mean ρ ↑	n/a	0.455	
		all	W↑	n/a	0.583	
Type IV	Social acceptability	all	P ↑	n/a	0.556	0.556

The next table below again shows the *pairwise* Type I, II and IV QRA results:

Study A	Study B	CV*	r	ρ	τ	P
O	R1	9.30	0.582	0.866	0.816	0.667 (2/3)
O	R2	13.95	0.389	0	0	0.333 (1/3)
R1	R2	4.68	0.976	0.500	0.333	0.667 (2/3)

This reveals that O and R2 produce less similar results than the other two pairings. In fact, in spite of the system-level scores being fairly similar (CV* is at the good end of medium), the correlations are poor, and just one paired rank is the same.

In contrast, O and R1 have good ρ and τ, but lower r; R1 and R2 have excellent r, but less good ρ and τ. These differing results show that all five measures in our tables provide individually interesting perspectives on degree of reproducibility.

Table 3 shows the corresponding full QRA results (for all three experiments jointly), revealing good (bordering on medium) Type I results (CV*), medium to poor correlations (Type II), and 55% shared pairwise ranks (P, Type IV).

3.3 Yao et al. (2022)

For this experiment, participants are shown a spreadsheet where each row contains a children's story, a generated question, and a generated answer for that question. They are then asked to evaluate the **Readability** of the generated question and answer pair (defined as "grammarly [sic] correct and clear language") on a scale of 1 (worst) to 5, which they enter in the adjacent column as an integer.

The table below shows the mean system scores, alongside the corresponding CV* (n=3) values (Type I results) for O (the original study), R1 (Braun, 2025), and R2 (Hürlimann and Cieliebak, 2026). CV* scores here indicate a medium degree

of reproducibility (in terms of the verbal descriptors introduced above).

System	O	R1	R2	CV*
Ground Truth	4.95	4.38	4.18	12.87
QAG	4.71	3.85	3.60	21.52
PAQ	4.08	3.14	3.25	23.29
Mean CV*	–	–	–	19.23

The next table below again shows the *pairwise* Type I, II and IV QRA results. This reveals that even though CV* is only medium overall, the remaining measures are near perfect (r) or in fact perfect (ρ, τ, P)

Study A	Study B	CV*	r	ρ	τ	P
O	R1	22.98	0.986	1	1	1.000 (3/3)
O	R2	25.97	0.923	1	1	1.000 (3/3)
R1	R2	5.99	0.975	1	1	1.000 (3/3)

Table 4 shows the corresponding full QRA results, for the three experiments jointly. These very much accord with what we saw from the pairwise results.

3.4 Reif et al. (2022)

In this experiment, participants are asked to rate, on a 0–100 slider scale, the **Meaning Preservation** of output sentences that have undergone style transfer to make them more positive, given the input sentence. The table below shows the mean system-level scores, alongside the corresponding CV* (n=3) for O (the original study), R1 (Onderková et al., 2025) and R2 (Mahamood, 2026). CV* values are very varied here, depending on the system; the Paraphrase system stands out for having poor CV*, in fact O considers it to be the best system and R1 and R2 the worst. The human-created style-transferred texts in contrast stand out for having particularly *good* reproducibility.

System	O	R1	R2	CV*
paraphrase	90.29	45.55	21.74	76.22
zero	60.71	49.66	24.24	48.10
aug_zero	86.47	64.99	67.31	18.50
human	85.29	74.76	79.33	7.56
Mean CV*	–	–	–	37.59

The next table below again shows the *pairwise* Type I, II and IV QRA results. This reveals that R1 and R2 have excellent Type II and IV reproducibility with respect to each other, despite having poor Type I reproducibility. Arguably, if Type II correlations are close and pairwise ranks are reproduced, how close the actual scores is secondary. Results from O are very dissimilar to those from R1 and R2, in fact in terms of both ρ and τ, the correlation is *inverse*, with only 2/6 ranks shared.

Study A	Study B	CV*	r	ρ	τ	P
O	R1	28.65	0.300	-0.400	-0.333	0.333 (2/6)
O	R2	54.31	0.401	-0.400	-0.333	0.333 (2/6)
R1	R2	33.95	0.987	1	1	1.000 (6/6)

Table 5 shows the corresponding full QRA results, for the three experiments jointly. These show poor reproducibility across all measures.

Because there was a substantial difference between the two reproductions on the one hand, and the original results on the other, we performed an LLM sanity check using the same method as in ReproNLP 2025 (Belz et al., 2025). However, it was inconclusive.

3.5 August et al. (2022)

Participants in this experiment were shown definitions of scientific terms and asked whether they contained any errors (yes or no). They were able to use the internet to check the definitions. Results are reported in terms of percentage of definitions with errors. This QC is called **Factual Truth** below.

August et al. report separate results for counting a definition to contain errors if (i) both evaluators indicated there was an error; and if (ii) at least one of the evaluators indicated there was an error. In this report, we first present two repeat runs first for case *ii*, then for case *i*.

(ii) At least one evaluator finds an error

The corresponding percentages for case *ii* are shown in the following table, O by August et al. (2022); R1 by Florescu et al. (2025); R2 by Arous and Cheung (2026):

System	O	R1	R2	CV*
DEXPERT	86.00	78.00	80.00	5.78
GEDI	52.00	78.00	65.00	22.57
SVM	38.00	78.00	39.00	49.82
Mean CV*	–	–	–	26.05

The 3-way CV* values above range from good for the DEXPERT system to poor for SVM. The next table below shows the *pairwise* Type I, II and IV measures. These reveal something previously unseen: O and R2 are in fact very close, whereas R1 is not aligned at all with the other two experiments. And in fact, it turns out in R1 all systems have the

Table 4: **Reproductions of Yao et al., (2022) by reprohum partners**: QRA reproducibility assessment for three comparable experiments (n=3), Yao et al. (2022), Braun (2025), and Hürlimann and Cieliebak (2026); n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 3$)		
				System level	QC level	Study level
Type I	Readability	Ground Truth	(mean) CV*↓	12.87	19.23	19.23
		QAG		21.52		
		PAQ		23.29		
Type II	Readability	all	mean r ↑	n/a	0.961	n/a
		all	mean ρ ↑	n/a	1.000	
		all	W↑	n/a	1.000	
Type IV	Readability	all	P ↑	n/a	1.000	1.000

Table 5: **Reif_etal_2022**: QRA reproducibility assessment for three comparable experiments (n=3), Reif et al. (2022), Onderková et al. (2025), and Mahamood (2026); n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 3$)		
				System level	QC level	Study level
Type I	Meaning Preservation	paraphrase	(mean) CV*↓	76.22	37.59	37.59
		zero		48.10		
		aug_zero		18.50		
		human		7.56		
Type II	Meaning Preservation	all	mean r ↑	n/a	0.563	n/a
		all	mean ρ ↑	n/a	0.067	
		all	W↑	n/a	0.378	
Type IV	Meaning Preservation	all	P ↑	n/a	0.556	0.556

same percentage (78) assigned to them, so the Type II measures cannot be computed.

Study A	Study B	CV*	r	ρ	τ	P
O	R1	35.07	n/a	n/a	n/a	0.000 (0/3)
O	R2	9.47	0.925	1	1	1.000 (3/3)
R1	R2	25.81	n/a	n/a	n/a	0.000 (0/3)

(i) Evaluators agree there is an error

For case *i* (where both evaluators must agree there is an error), the situation is worse in some respects, and improved in others. The 3-way CV* table shows CV* to be much worse:

System	O	R1	R2	CV*
DEXPERT	67.00	54.00	38.00	30.93
GEDI	33.00	51.00	32.00	31.20
SVM	16.00	57.00	16.00	90.03
Mean CV*	–	–	–	50.72

The *pairwise* reproducibility measures again facilitate interesting additional insights: O and R2 are in near perfect agreement again, while R1 is inversely correlated with both O and R2:

Study A	Study B	CV*	r	ρ	τ	P
O	R1	52.19	-0.327	-0.500	-0.333	0.333 (1/3)
O	R2	17.23	0.902	1	1	1.000 (3/3)
R1	R2	56.98	-0.703	-0.500	-0.333	0.333 (1/3)

3.5.1 Discussion of August et al. reproductions

Taken in combination, the results from O, R1 and R2 for the two cases (*i* and *ii)* strongly suggest that something has gone wrong in R1. That all three systems have the same error rate (as happened in case *ii*) is generally unlikely; pronounced negative correlations for all measures and both O and R2 are another warning flag. As a result, we've decided not to present the full QRA tables for all three aligned experiments jointly, because the results would be based on an experiment that's likely to be faulty, and therefore the presentation would

be likely to be misleading.

We additionally conducted an LLM sanity check with the method from ReproNLP 2025 which strongly agreed with O and R2, and strongly disagreed with R1, in both settings i (where both evaluators must agree), and ii (where just one evaluator needs to agree). For the full tables, including the LLM results, see Appendix A.

The above illustrates why multiple reproductions are needed in reproducibility assessments: any single one can go wrong, and more evidence means a more more complete a picture emerges.

4 Track B Non-MLMT Results

The results presented in this section were produced for one of the ReproHum MLMT study original papers (see list above), but not under the strict conditions of the MLMT study. Participants were given the same resources as the ReproHum partners that were shared by the original authors, and asked to reproduce the experiment under as similar conditions as possible. However, ultimately it was up to them which quality criteria they evaluate, and how they conduct their reproduction study. Both participants presented in this section, Gawinecki et al. (2026) and Florescu et al. (2024) (from ReproNLP 2024) reproduced the full set of quality criteria from the original study by Yao et al. (2022).

4.1 Full set of QCs from Yao et al. (2022)

To recap briefly from above, in this experiment, participants were asked to evaluate the **Readability** of generated question/answer pairs on a scale of 1–5. However, they also similarly evaluated **Answer Relevance** and **Question Relevance** which were not included in the ReproHum reproductions.

The following table presents the system-level scores and corresponding CV* values for the two reproductions of Yao et al. which included all of the three evaluation criteria, R1 (Florescu et al., 2024), and R2 (Gawinecki et al., 2026):

Criterion	System	O	R1	R2	CV*
Answer Relevance	QAG	3.99	3.20	3.69	17.13
Answer Relevance	PAQ	3.90	3.20	3.78	16.08
Answer Relevance	GND	4.83	4.46	4.57	5.92
Question Relevance	QAG	4.39	3.83	3.95	10.88
Question Relevance	PAQ	4.18	3.61	3.64	12.88
Question Relevance	GND	4.92	4.71	4.64	4.38
Readability	QAG	4.71	4.52	3.88	14.56
Readability	PAQ	4.08	4.17	3.56	12.65
Readability	GND	4.95	4.71	4.05	14.73
Mean CV*	–	–	–	–	12.14

The 3-way CV* scores above are medium good mostly, and are broadly comparable across the three quality criteria (QCs). The scores for the ground truth outputs (GND) have better reproducibility for Answer Relevance and Question Relevance, not however for Readability. The latter is in contrast to the other results for Yao et al., presented above in Section 3.3 where the ground truth outputs had better reproducibility for Readability by some margin.

The three tables below show *pairwise* Type I, II and IV results by QC. We can see that for Answer Relevance, r is near perfect, but ρ and τ are less strong, especially between O and R2, with 2 out of 3 pairwise ranks shared. In contrast, Question Relevance and Readability have perfect Type II and IV reproducibility. CV* values range from good to medium, but are a bit all over the place, each pair having lowest CV* for one of the QCs.

Answer Relevance:

Study A	Study B	CV*	r	ρ	τ	P
O	R1	20.10	0.996	0.866	0.816	0.667 (2/3)
O	R2	6.44	0.984	0.500	0.333	0.667 (2/3)
R1	R2	13.73	0.996	0.866	0.816	0.667 (2/3)

Question Relevance:

Study A	Study B	CV*	r	ρ	τ	P
O	R1	12.76	0.996	1	1	1.000 (3/3)
O	R2	11.77	1.000	1	1	1.000 (3/3)
R1	R2	2.13	0.993	1	1	1.000 (3/3)

Readability:

Study A	Study B	CV*	r	ρ	τ	P
O	R1	4.25	0.996	1	1	1.000 (3/3)
O	R2	20.48	0.997	1	1	1.000 (3/3)
R1	R2	17.97	1.000	1	1	1.000 (3/3)

The corresponding full 3-way QRA results are shown in Table 6. Type I reproducibility is good to medium for all quality criteria, correlations are very strong except for ρ and W for Answer Relevance. P scores are perfect for Question Relevance and Readability, less so for Answer Relevance.

4.2 Reproductions of Readability in Yao et al.

Taking the reproduction studies from the last section together with those from Section 3.3, we have a rare, if not unique in NLP, total of four reproductions of a Readability evaluation.

The table below shows the system-level Readability scores, alongside CV* values (n=5), for O by Yao et al. (2022); R1 by Florescu et al. (2024); R2 by Braun (2025); R3 by Hürlimann and Cieliebak (2026); R4 by Gawinecki et al. (2026).

Table 6: **Yao et al. (2022), non-MLMT**: QRA reproducibility assessment for three comparable experiments (n=3), Yao et al. (2022), Florescu et al. (2024), and Gawinecki et al. (2026); n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 3$)		
				System level	QC level	Study level
Type I	Answer Relevance	QAG	(mean) CV*↓	17.13	13.05	12.14
		PAQ		16.08		
		Ground Truth		5.92		
	Question Relevance	QAG	(mean) CV*↓	10.88	9.38	
		PAQ		12.88		
		Ground Truth		4.38		
	Readability	QAG	(mean) CV*↓	14.56	13.98	
		PAQ		12.65		
		Ground Truth		14.73		
Type II	Answer Relevance	all	mean r ↑	n/a	0.992	n/a
		all	mean ρ ↑	n/a	0.744	
		all	W↑	n/a	0.750	
	Question Relevance	all	mean r ↑	n/a	0.996	
		all	mean ρ ↑	n/a	1.000	
		all	W↑	n/a	1.000	
	Readability	all	mean r ↑	n/a	0.998	
		all	mean ρ ↑	n/a	1.000	
		all	W↑	n/a	1.000	
Type IV	Answer Relevance	all	P ↑	n/a	0.667	0.889
	Question Relevance	all	P ↑	n/a	1.000	
	Readability	all	P ↑	n/a	1.000	

The CV* values are all at the good end of medium.

System	O	R1	R2	R3	R4	CV*
QAG	4.71	4.52	3.85	3.60	3.88	16.29
PAQ	4.08	4.17	3.14	3.25	3.56	18.93
Ground Truth	4.95	4.71	4.38	4.18	4.05	11.48
Mean CV*	–	–	–	–	–	15.57

The next table below shows the *pairwise* Type I, II and IV results for the 5 runs of the experiment. CV* ranges from just under 5 to about 26, r is perfect or near perfect in all cases, while the remaining three scores, ρ, τ and P, are all maximum (best) values.

Study A	Study B	CV*	r	ρ	τ	P
O	R1	4.25	0.996	1	1	1.000 (3/3)
O	R2	22.98	0.986	1	1	1.000 (3/3)
O	R3	25.97	0.923	1	1	1.000 (3/3)
O	R4	20.48	0.997	1	1	1.000 (3/3)
R1	R2	20.42	0.996	1	1	1.000 (3/3)
R1	R3	23.45	0.952	1	1	1.000 (3/3)
R1	R4	17.97	1.000	1	1	1.000 (3/3)
R2	R3	5.99	0.975	1	1	1.000 (3/3)
R2	R4	8.62	0.996	1	1	1.000 (3/3)
R3	R4	8.06	0.950	1	1	1.000 (3/3)

Table 7 shows the full 5-way QRA results. This table, and the other two QRA tables for Yao et al., Tables 4 and 6 all paint a similar picture: medium CV* with excellent Type II and IV results.

5 Track A

Ardestani et al. (2026) take a different view of reproduction than the other contributions so far presented. Rather than ask, if we re-run an evaluation experiment under similar conditions, can we expect similar results, Ardestani et al. ask, if we apply this evaluation method to a different dataset and task, can we expect similar correlations with human evaluation.

The evaluation method is AutoNuggetizer which evaluates long-form answers from retrieval-augmented generation (RAG) systems by decomposing evaluation into atomic facts (nuggets) and using LLMs for nugget creation and assignment.

The original study used TREC 2024 RAG Track data, an open-ended query task, comprising 301

Table 7: **Yao et al. (2022), all reproductions**: QRA reproducibility assessment for five comparable experiments (n=5), Yao et al. (2022), Florescu et al. (2024), Braun (2025), Hürlimann and Cieliebak (2026), and Gawinecki et al. (2026); n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 5$)		
				System level	QC level	Study level
Type I	Readability	QAG	(mean) CV*↓	16.29	15.57	15.57
		PAQ		18.93		
		Ground Truth		11.48		
Type II	Readability	all	mean r ↑	n/a	0.977	n/a
		all	mean ρ ↑	n/a	1.000	
		all	W↑	n/a	1.000	
Type IV	Readability	all	P ↑	n/a	1.000	1.000

topics (queries or questions) with community-submitted answers. The main aim for Ardestani et al. was to test if similarly good correlations with human evaluations could be obtained on the QAM-PARI list-QA benchmark, where answers consist of discrete entities, rather than long-form text.

The following table summarises Kendall's τ values observed in the original study O (Pradeep et al., 2025) and in Ardestani et al.'s experiments R1 (these are all the specific τ values that are reported in the paper):

	O (τ)	R1 (τ)
V_{strict} vs E/M	0.887	0.810
A_{strict} vs E/M	0.901	0.810
V_{strict} vs M/M	0.727	0.810
A_{strict} vs M/M	0.758	0.714

Here, E before the forward slash stands for edited automatic nugget identification with AutoNuggetizer, M for manual nugget identification. After the slash, M stands for manual nugget assignment. V_{strict} is the fraction of *vital* nuggets labelled fully supported, A_{strict} the fraction of *all* nuggets fully supported.

The above figures look broadly comparable in the sense that the metrics achieve reasonable correlation with human assessment. However in terms of rank between V_{strict} and A_{strict}, O and R1 do not assign the same relative ranks to the two metrics, neither when compared to E/M, nor when compared to M/M.

Note that it would not be appropriate to apply QRA analysis to the figures in the above table, because task, dataset, and systems were all different (QRA requires at least the object of measurement, i.e. the systems, to be the same).

6 Reproducibility by Properties

As in previous years, we examine associations between experiment properties and QRA measures, to see if any pattern emerges as to which properties may be associated with better, and which with worse, reproducibility.

Table 8 shows some of the main HEDS properties of the experiments repeated by ReproHum partner labs, along with mean CV* r and P values in the last five columns. The three CV* values are calculated as follows:

- **O vs R1**: the mean of two-way CV* values between O and R1.
- **O vs R2**: the mean of two-way CV* values between O and R2.
- **n=3**: the mean of three-way CV* values between O, R1 and R2.

What we are looking for in this table is any indication that one of the HEDS properties is associated with better/worse experiment-level QRA scores.

In previous years, one such property was number of evaluators (HEDS Question 3.2.1): the pattern is for more evaluators to be associated with better reproducibility, a pattern that is also observed in Table 8, where August et al. has the smallest number of evaluators (2) and the worst QC-level CV*, r and P. In contrast, the best set of QRA scores overall is for the experiment with the largest number of evaluators (Castro Ferreira et al.).

Another trend that was previously observed and is also observable here is that evaluations that are more cognitively complex tend to have poorer reproducibility than cognitively simpler evaluations.

ReproNLP 2026							Type I (mean CV*)			Type II and IV	
Orig Study // *Repro a / Repro b* **measurand**	**3.2.1**	**4.3.4**	**4.3.8**	**4.1.1**	**4.1.2**	**4.1.3**	**O vs R1**	**O vs R2**	**n=3**	mean r	P
August et al. (2022) // *Florescu et al. (2025) / Arous and Cheung (2026)* **Factual Truth (both)**	2 / 2 / 2	Yes, No	DQE	Corr.	Content	EFoR	52.19	17.23	50.72	-0.043	0.556
August et al. (2022) // *Florescu et al. (2025) / Arous and Cheung (2026)* **Factual Truth (one)**	2 / 2 / 2	Yes, No	DQE	Corr.	Content	EFoR	35.07	9.47	26.05	nan	0.333
Castro Ferreira et al. (2018) // *Mahamood (2024) / Langner (2026)* **Clarity**	60 / 60 / 60	1-7	DQE	Good.	Both	iiOR	3.06	4.63	5.09	0.861	0.822
Gabriel et al. (2022) // *Mahlaza et al. (2024) / Fan and Chen (2026)* **Social acceptability**	UNK / 42 / 60	Yes, No	DQE	Feature	Both	EFoR	9.30	13.95	9.23	0.649	0.556
Reif et al. (2022) // *Onderková et al. (2025) / Mahamood (2026)* **Meaning Preservation**	6 / 6 / 6	0–100	DQE	Good	Content	RtI	28.65	54.31	37.59	0.563	0.556
Yao et al. (2022) // *Braun (2025) / Hürlimann and Cieliebak (2026)* **Readability**	5 / 5 / 5	1–5	DQE	Good	Both	iiOR	22.98	25.97	19.23	0.961	1.000

Table 8: 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 sample sizes are not directly comparable). The following columns map to experiment properties as recorded in HEDS 3.0 (Belz and Thomson, 2025): 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).

An example is the evaluation of Factual Truth in August et al. which had the highest study-level, mean CV* of all studies reported. Another example is Meaning Preservation in Reif et al. which has the second worst CV* values. These are both cognitively complex: the former requires evaluators to look things up online and assess the truth value of texts. The latter involves comparing inputs and outputs and determining to what extent they mean the same thing independently of the manner in which they are expressed.

7 Discussion

As in previous editions of ReproNLP, we saw that degree of reproducibility can look very different depending on which single reproduction study we look at. For example, in the August et al. experiment variant where both evaluators must agree on errors (Section 3.5), R1 is inversely correlated with O, while R2 strongly agrees with it.

Degree of reproducibility can also look very different depending which QRA measure is applied.

For example, for Yao et al., the Type II measures applied (Pearson's and Spearman's correlations) showed excellent reproducibility, as did Type IV (P, the proportion of identical pairwise ranks), but CV* was only medium (study-level mean CV* was 19.23). We have observed this pattern a few times in ReproNLP, but the inverse, excellent study-level mean CV*, and then terrible correlations and P, we have never seen (as one would expect).

For the above reasons, it's important to think of the full QRA assessment which takes all aligned experiments in a set into account *jointly* as the truest assessment of the reproducibility of the experiment design (and where shared also the implementation).

A type of experiment that's emerging as having particularly good reproducibility is one that asks 5 or more evaluators to rate system outputs in terms of a cognitively non-complex quality criterion on a small scale, looking only at the system outputs. We saw two examples in this report: Castro Ferreira et al., and Yao et al. With four reproductions to date, the Yao et al. experiment has been thoroughly stress tested: five different teams of researchers have implemented and run this experiment design and all obtained the same system ranks, and near perfect system-level score correlations.

8 Conclusion

As ever, because shared task results reports are written under pressure of time and to a deadline, there are other aspects than those reported here which we would like to have investigated, but will have to leave for future work.

Some of our most important insights from the ReproNLP work have been:

- Multiple reproductions are needed to test the reproducibility of an experiment design: we have seen cases where one reproduction showed terrible reproducibility while the next showed excellent reproducibility.
- Multiple measures of degree of reproducibility are needed (as in QRA): system-level scores can be very different between two runs of an experiment, yet correlations and P (proportion of same pairwise ranks) can still be very strong (and vice versa).
- Degree of reproducibility needs to be assessed at least at the system level and the QC level: large differences can occur between systems in terms of how reproducible their aggregated scores are which is completely obscured at the QC level.

The ReproHum reproductions reported here complete the second ReproHum MLMT study, and we will move on to reporting the results from the whole study in our work next. We hope that the results and methodological, data and computing resources that ReproNLP has accumulated over its six editions will prove useful to the research field. We plan to release our resources and results in full so that further work can be conducted on this basis.

Acknowledgments

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

We thank our numerous collaborators from NLP labs across the world who carried out many of the reproductions reported in this paper as part of the second batch of coordinated reproductions resulting from the ReproHum project.

The first round of ReproHum reproductions 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 which ended in May 2024.

The ReproNLP work has also benefitted from being carried out in the wider context of 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.

Craig Thomson's current work on ReproNLP and ReproHum is funded by ADAPT.

References

MohammadJavad Ardestani, Ehsan Kamalloo, and Davood Rafiei. 2026. Do autonuggetizer's agreement patterns generalize to list-qa? In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. Association for Computational Linguistics.

Ines Arous and Jackie Chi Kit Cheung. 2026. Reprohum: #0033-05: Human evaluation report on "generating scientific definitions with controllable complexity". In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San

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

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

Anya Belz, Anastasia Shimorina, Maja Popović, and Ehud Reiter. 2022b. 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. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anya Belz 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. 2025. HEDS 3.0: The human evaluation data sheet version 3.0. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 60–81, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2026. Quantified reproducibility assessment for four types of evaluation results. *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, 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. 2023b. 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.

Daniel Braun. 2025. ReproHum #0031-01: Reproducing the human evaluation of readability from "it is AI's turn to ask humans a question". In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 576–582, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Andrew Caines and Paula Buttery. 2020. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5614–5623, 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.

Rui Fan and Guanyi Chen. 2026. Reprohum #0866-04: Variability in human judgments of sociopolitical acceptability across studies. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. 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 Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 106–113, Torino, Italia. ELRA and ICCL.

Andra-Maria Florescu, Marius Micluţa-Câmpeanu, Stefana Arina Tabusca, and Liviu P Dinu. 2025. ReproHum #0033-05: Human evaluation of factuality from a multidisciplinary perspective. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 583–589, Vienna, Austria and virtual meeting. 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.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Maciej Gawinecki, Marcel Mroczek, Chiara Albarello, and Paul-Emmanuel Floch. 2026. Repronlp 2026: A third replication of the human evaluation of a qag system for children's storybooks. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. Association for Computational Linguistics.

Rudali Huidrom and Anja 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.

Manuela Hürlimann and Mark Cieliebak. 2026. Reprohum 0031-01: Reproducing a human readability evaluation for question–answer generation systems. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. Association for Computational Linguistics.

Maurice Langner. 2026. Reprohum #0124-03: Reproducing human scores on neural reg models. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. Association for Computational Linguistics.

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.

Saad Mahamood. 2026. ReproHum #0669-08: Reproducing a recipe for arbitrary text style transfer with LLMs. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. Association for Computational Linguistics.

Zola Mahlaza, Toky Hajatiana Raboanary, Kyle Seakgwa, and C. Maria Keet. 2024. ReproHum #0866-04: Another evaluation of readers' reactions to news headlines. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 274–280, 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.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. The great nugget recall: Automating fact extraction and rag evaluation with large language models. *arXiv preprint arXiv:2504.15068*.

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.

Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 2025. Do my eyes deceive me? a survey of human evaluations of hallucinations in NLG. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79, Hanoi, Vietnam. Association for Computational Linguistics.

Craig Thomson, Javier González Corbelle, and Anya Belz. 2026. Process standardisation for human evaluation of nlp system outputs. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. Association for Computational Linguistics.

Bingsheng Yao, Dakuo Chen, Tongshuang Yu, Emily Sheng, Karthik Narasimhan, Tianyi Zhang, Yuan Cao, Kedar Dhamdhere, Dan Ma, Sherry Chang, and Q. Vera Liao Zhou. 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 731–745, Dublin, Ireland. Association for Computational Linguistics.

A LLM Sanity Checks for August et al.

The first table below show the system-level scores for the **Factual Truth** QC, in the setting where *both evaluators must agree there is an error* , augmented by LLM evaluation results as peer the sanity check method from Belz et al. (2025), first proposed by Huidrom and Belz (2025). O by August et al. (2022); R1 by Florescu et al. (2025); R2 by Arous and Cheung (2026); LLM by Belz et al. (2025).

System	O	R1	R2	LLM	CV*
DEXPERT	67.00	54.00	38.00	85.00	35.44
GEDI	33.00	51.00	32.00	35.00	25.65
SVM	16.00	57.00	16.00	25.00	74.14
Mean CV*	–	–	–	–	45.08

The next table below shows the corresponding pairwise QRA results:

Study A	Study B	CV*	r	ρ	τ	P
LLM	O	21.70	0.984	1	1	1.000 (3/3)
LLM	R1	47.22	-0.156	-0.500	-0.333	0.333 (1/3)
LLM	R2	38.19	0.812	1	1	1.000 (3/3)
O	R1	52.19	-0.327	-0.500	-0.333	0.333 (1/3)
O	R2	17.23	0.902	1	1	1.000 (3/3)
R1	R2	56.98	-0.703	-0.500	-0.333	0.333 (1/3)

The first table below show the system-level scores for the **Factual Truth** QC, in the setting where *only one evaluator need agree there is an error*, augmented by LLM evaluation results as peer the sanity check method from Belz et al. (2025). O by August et al. (2022); R1 by Florescu et al. (2025); R2 by Arous and Cheung (2026); LLM by Belz et al. (2025).

System	O	R1	R2	LLM	CV*
DEXPERT	86.00	78.00	80.00	98.00	11.43
GEDI	52.00	78.00	65.00	75.00	18.87
SVM	38.00	78.00	39.00	68.00	39.59
Mean CV*	–	–	–	–	23.30

The next table below shows the corresponding pairwise QRA results:

Study A	Study B	CV*	r	ρ	τ	P
LLM	O	31.27	0.998	1	1	1.000 (3/3)
LLM	R1	11.92	n/a	n/a	n/a	0.000 (0/3)
LLM	R2	26.21	0.899	1	1	1.000 (3/3)
O	R1	35.07	n/a	n/a	n/a	0.000 (0/3)
O	R2	9.47	0.925	1	1	1.000 (3/3)
R1	R2	25.81	n/a	n/a	n/a	0.000 (0/3)

B 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., 2023a).
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. 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.[7]
5. Carry out the allocated experiment exactly as described in the HEDS sheet.
6. 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.

[7] We now call these Type IV results.

(d) Report quantified reproducibility assessments in terms of QRA++ (Belz and Thomson, 2026) as follows:

   i. Type I results: Small-sample coefficient of variation CV*.
   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.
   iv. Type IV results: Proportion of pairwise system ranks maintained.