The 2025 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results

Anya Belz[1], Craig Thomson[1], Javier González Corbelle[1,2], Malo Ruelle[1,3]
[1]ADAPT, Dublin City University
[2]CiTIUS, Universidade de Santiago de Compostela, Spain
[3]École Centrale de Lille, France
Corresponding author: anya.belz@dcu.ie

Abstract

This paper presents an overview of, and the results from, the 2025 Shared Task on Reproducibility of Evaluations in NLP (ReproNLP'25) which followed on from four previous shared tasks on reproducibility of evaluations, 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'25 shared task, summarise results from the reproduction studies submitted, and provide additional comparative analysis of their results, including for the first time additional, 'sanity-check' evaluations by LLMs.

1 Introduction

Natural language processing (NLP) and machine learning (ML) are still far from solving the reproducibility crisis that has been well documented over recent years (Belz et al., 2021a; Thomson et al., 2024). Authors still don't make enough resources and information available about published work to enable repetitions of it despite reproducibility checklists being introduced by conferences.[1] When reproducibility is tested, results often fail to confirm original findings (Wieling et al., 2018; Belz et al., 2021a; Belz and Thomson, 2024a).

The core aim of this sixth 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), ReproNLP'23 (Belz and Thomson, 2023), and ReproNLP'24 (Belz and Thomson, 2024a), is to continue to add to the body of reproduction studies in NLP and ML, but also to produce and analyse multiple reproductions of shared original evaluations, to shed more light on how best to assess reproducibility in NLP/ML and ultimately how to improve the degree to which our findings in the field are reproducible.

The eight new reproduction studies (for an overview see Table 1) reported in ReproNLP this year add new data points to the body of directly comparable evaluations available for investigations of reproducibility. Our new analyses point towards further reasons for low reproducibility of evaluations, and ways to improve experimental design likely to improve 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 (Section 3). We report results from LLM sanity checks carried out in those cases where at least one reproduction disagreed with the original study (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. We conclude with some discussion (Section 6) and a look to future work (Section 7).

2 ReproNLP 2025

Like its predecessor, ReproNLP 2025[2] consisted of two tracks, one an 'unshared task' in which teams repeat 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 experiments for which the shared-task organisers make available all necessary

[1]For an example see the AAAI'26 one at https://aaai.org/conference/aaai/aaai-26/reproducibility-checklist/.

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

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 1002–1016

Original Study	Qual. Criterion	#evors	#sys	items-per-sys	Labs reproducing study for ReproNLP 2025
Yao et al. (2022)	Readability	5	3	120.33‡	a) University of Twente
August et al. (2022a) B†	Factual Truth	2	3	300	a) University of Bucharest
Bai et al. (2021)	Informativeness	7	4	60	a) Tianjin University
Reif et al. (2022)	Semantic Similarity	6	6	50	a) Charles University
Gu et al. (2022)	Overall	2	4	31.50‡	a) Dublin City University b) Bielefeld University
Hosking and Lapata (2021)	Meaning preservation	varies	4	300	a) Heidelberg University b) University of Illinois Chicago

Table 1: ReproNLP 2025 experiments performed by ReproHum partner labs. All experiments were in the English language. For Hosking and Lapata (2021) the number of evaluators varies because only the number of participants per item is controlled, not the number of items per participant. An item is defined as one system output evaluated absolutely, or a set of system outputs evaluated relatively. † = marked B because another experiment by the same authors was included in ReproNLP 2024. ‡ = values varied for the different studies, showing the mean.

information and resources (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[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., 2023).

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[4]).

The information provided for each study below includes (i) whether the assessment of systems was *relative* to other systems or *absolute* without comparators; (ii) what the language*(s)* of the systems were; (iii) how many *datasets* were used; (iv) how many *systems* were evaluated and (v) by how many *evaluators*; and (vi) whether the evaluation was run on a *crowd-sourcing* platform.

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

[4] https://repronlp.github.io/

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; between 4 and 6 systems and between 200 and 300 evaluation items per dataset-criterion combination; crowdsourced.

2. **Bai et al. (2021)**: *Cross-Lingual Abstractive Summarization with Limited Parallel Resources*: https://aclanthology.org/2021.acl-long.538

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

3. **Hosking & Lapata (2021)**: *Factorising Meaning and Form for Intent-Preserving Paraphrasing*: https://aclanthology.org/2021.acl-long.112

   Relative evaluation study; English; 3 quality criteria; 1 dataset; 4 systems and 1200 evaluation items per criterion; crowdsourced.

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. **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.

6. **Gu et al. (2022)**: *MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes*: https://aclanthology.org/2022.acl-long.450

   Relative evaluation study; English; 3 quality criteria; 1 dataset; 2 systems; between 63 and 67 evaluation items per criterion.

7. **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; 100 evaluation items; 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 criterion they assess.

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) are included here in the ReproNLP'25 report. All 21 experiments from the current batch were open to all other ReproNLP'25 participants.

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

2.1 Participation

There were no submissions for Track A this year, and eight for Track B. The ReproHum partners reporting in Track B are listed in Table 1. There were no non-ReproHum participants this year.

2.2 Approach to reproduction and reproducibility assessment

We encouraged all participants to complete a HEDS datasheet (Belz and Thomson, 2024b) in the ReproHum version,[5] and to follow the ReproHum Common Approach to reproduction laid out in Appendix A which includes QRA++ (Belz, 2025), a set of quantitative reproducibility assessment measures for four common types of results in NLP/ML 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 facilitate comparison across submissions. In brief summary (for full details see Belz, 2025), 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.

In QRA++, 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, 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.[6]

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

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

[6] To obtain comparable results we restrict ourselves to pairwise system ranks as findings.

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 is a difference between the studies reported in 2023 on the one hand, and 2024 and 2025 on the other. For the earlier batch, our aim was to achieve maximum similarity between design and implementation of 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 insufficient 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 ReproNLP'25 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.

2.3 LLM sanity checks

In past editions of ReproNLP, a recurring theme was two reproductions giving contradictory results regarding the reproducibility of the original evaluation experiment, e.g. one agreeing strongly with the original, the other disagreeing equally strongly. Previously, we had no way of deciding if one of the reproductions was more likely to give a true picture of the reproducibility of the original experiment than the other. Since then, results have been reported that indicate that in such situations, LLM-based evaluations (commonly known as 'LLM-as-judge' methods) tend to agree very strongly with one reproduction while disagreeing with the other (Huidrom and Belz, 2025a,b). In fact, this was found to be the case across five different sets of experiments tested by Huidrom and Belz (2025b), across a wide variety of different types and sizes of LLMs and LLM ensembles. So, for the first time this year, we apply such sanity checks to situations where there is disagreement among the two (or in one case three) reproductions carried out (Section C).

Note that the results from these sanity checks should not be interpreted as implying that there's something wrong with the reproduction that the LLMs disagree with. The reason may simply be that the sample of evaluators used represented a population outlier. In this report, we don't offer potential explanations; we simply report the correlation results and state which evaluation the LLMs agree with. Overall, based on Huidrom and Belz (2025b), we assume that if the LLMs all strongly agree with the original evaluation and one of the reproductions, as well as strongly agreeing with each other, then it is more likely that these agreeing evaluations give the true picture, than the one single disagreeing evaluation.

To reiterate, this does not however mean that the latter is lacking in quality or rigour, as the evaluator cohort may simply be a statistical outlier.

3 Track B Results

In this section, we report results for the eight submissions (listed in Table 1) received in Track B, where related submissions area grouped together into subsections headed by the paper reference for the original study. 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* (Type I QRA) computed on all either two or three scores. We then report the pairwise Pearson's r and Spearman's ρ correlation coefficients (Type II QRA) and the proportion of pairwise system ranks upheld (Type IV QRA). (For details see Section 2.2.) 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 Yao et al. (2022)

For this experiment, participants were shown a spreadsheet where each row contains a section for

a children's story, a generated question, and a generated answer for that question. They were 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 below table shows the mean system scores, alongside the corresponding CV* (n=2) values (Type I results) for O (the original study) and R1 (Braun, 2025). CV* scores here indicate a medium degree of reproducibility (in terms of the categorisation introduced at the end of Section 2.2).

System	O	R1	CV*
Yao et al.	4.71	3.85	26.14
PAQ Baseline	4.08	3.14	35.91
Ground truth	4.95	4.38	15.51
Mean CV*	–	–	25.85

The table below shows Type II (Pearson's r and Spearman's ρ correlations) and Type IV (P, the proportion of identical pairwise system ranks) QRA scores. On both Type II and IV measures, the alignment is perfect or near perfect, indicating that the Yao et al. study has excellent reproducibility.

Study A	Study B	r	ρ	P
O	R1	0.99	1.00	1.00 (3/3)

3.2 August et al. (2022a) B

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 were reported in terms of percentage of definitions with errors.

August et al. reported 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.

(i) Evaluators agree there is an error

The below table shows the system-level scores based on the stricter criterion that both evaluators had to agree there was an error, alongside the corresponding CV* (n=2) values for the original study (O) and reproduction R1 (Florescu et al., 2025). The degree of reproducibility in terms of CV* is poor, with the best system-level CV* being slightly better at about 20, but the mean being nearly 60.

System	O	R1	CV*
SVM	16.00	57.00	111.99
GeDi	33.00	51.00	42.73
DExperts	67.00	54.00	21.42
Mean CV*	–	–	58.71

Type II QRA (correlations) indicated a medium *negative* correlation, while we can see from the Type IV QRA that only one of the three pairwise ranks was the same between the two studies.

Study A	Study B	r	ρ	P
O	R1	-0.33	-0.50	0.33 (1/3)

(ii) At least one evaluator finds an error

The next table below shows the mean system scores based on the less strict criterion that just one evaluator has to indicate a definition has an error for it to count towards the evaluation score. CV* scores improve when aggregating responses by this method, now being closer to the medium good range for human evaluations.

System	O	R1	CV*
SVM	38.00	78.00	68.76
GeDi	52.00	78.00	39.88
DExperts	86.00	78.00	9.73
Mean CV*	–	–	39.46

As we can see from the above table (see also discussion by Florescu et al. (2025)), all system-level percentages ended up being the same (78%) with this method of aggregation; we are therefore unable to report correlations. The Type IV results below show that none of the three system ranks were the same in O and R1.

Study A	Study B	r	ρ	P
O	R1	nan	nan	0 (0/3)

3.3 Bai et al. (2021)

For the Informativeness evaluation of cross-lingual summarisation systems reported by Bai et al. (2021), participants were asked to select the best of 4 system outputs (marking it with a 1). They then marked the worst system as -1, and the other two as 0. Reported scores are the percentage of times each system is selected as best minus the times it is selected as worst. Bai et al. (2021) reported results

for three resource settings, minimum, medium, and maximum, each indicating a proportion of the test set used (maximum referring to the whole of the test set). The reproductions however were conducted only for the maximum setting so this is the setting we report results for.

The below table shows the aggregated system scores, alongside the corresponding CV* (n=2) values for O (the original study) and R1 (Supryadi et al., 2025). The degree of reproducibility is extremely high, with the lowest (best) CV* values seen for any human evaluation experiment to date in ReproNLP.

System	O	R1	CV*
MCLAS	0.06	0.08	2.05
NCLS	-0.13	-0.13	0.00
NCLS+MS	-0.18	-0.19	1.71
GOLD	0.26	0.25	0.56
Mean CV*	–	–	1.08

For Type II results, we see (near) perfect correlations. Pearson's is only 0.99 because we do not round up to 1.0 unless the two series are identical (see our rounding policy in appendix B). Pairwise system ranks are the same in both studies.

Study A	Study B	r	ρ	P
O	R1	0.99	1.00	1.00 (6/6)

3.4 Reif et al. (2022)

Participants are asked to rate, on a 0–100 slider scale, the **Meaning Preservation** of an output sentence, given the input sentence. The below table shows the mean scores, alongside the corresponding CV* (n=2) for O (the original study) and R1 (Onderková et al., 2025). CV* values are mostly in the medium range; the Paraphrase system stands out for having poor CV*, in fact O considers it to be the best system and R1 the worst.

System	O	R1	CV*
Paraphrase	90.29	45.81	65.17
Zero-shot	69.71	49.44	33.92
Unsup. MT	86.76	73.32	16.74
Dual RL	85.29	68.24	22.14
Aug. zero-shot	86.47	65.10	28.11
Human	85.29	74.81	13.05
Mean CV*	–	–	29.86

The correlations show a mixed picture with Pearson's indicating a mild to medium correlation, but Spearman's a mild *negative* correlation. The Type IV QRA score shows that only 6 of 15 of pairwise ranks are the same between the two studies.

Study A	Study B	r	ρ	P
O	R1	0.32	-0.20	0.4 (6/15)

Reif et al. (2022) did not report scores in their paper, but did show them in a bar chart. Onderková et al. (2025) were able to estimate the scores by counting pixels in the chart (with an accuracy of $\pm 0.3\%$). Given the large differences in per-system scores (over 10 in all cases) the effect on QRA++ results is negligible.

3.5 Gu et al. (2022)

Here, participants had to rate the quality of the outputs of pairs of extractive summarisation systems, ranking the one which was best **Overall** as 1, the other as 2 (in case of identical output both were ranked 1). Aggregated system-level results are reported as the average rank they are assigned. The below table shows the aggregated system scores, alongside the corresponding CV* (n=3) values for O (the original study), R1 (Mille and Lorandi, 2025), and R2 (Junker, 2025). CV* is medium for both systems.

System	O	R1	R2	CV*
MemSum	1.38	1.27	1.49	35.39
NeuSum	1.57	1.33	1.46	32.40
Mean CV*	–	–	–	33.89

There are only two systems so correlations are either 1 or -1. In these simple terms, O and R1 are in agreement, R2 disagreeing with them. The Type IV results below also show that O and R1 agreed on the one pairwise system ranking while R2 disagreed.

Study A	Study B	r	ρ	P
O	R1	1	1	1 (1/1)
O	R2	-1	-1	0 (0/1)
R1	R2	-1	-1	0 (0/1)

3.6 Hosking and Lapata (2021)

For this experiment, participants are asked to select which of two system-generated output summaries are "Closest in meaning" to the input

(**Preservation of meaning**); the selected system is assigned a 1, the other a -1. There are five systems, and for each input, all pairwise combinations (10) of systems are evaluated. Best-worst scaling is then applied, resulting in system scores between -100 and $+100$.

The below table shows the aggregated system scores, alongside the corresponding CV* (n=3) values for O (the original study), R1 (Steen and Markert, 2025), and R2 (Arvan and Parde, 2025). We see an excellent degree of reproducibility in terms of CV*, with only the DiPS system having a CV* value above 5.

System	O	R1	R2	CV*
VAE	58	57	57.00	0.45
Separator	-6	-3	1.44	4.69
Latent BoW	-12	-9	-13.44	3.13
DiPS	-39	-46	-45	8.17
Mean CV*	–	–	–	4.11

Correlations, as shown in below, are as good as they can be. In terms of Type IV QRA, all 6 pairwise system ranks are the same between all studies.

Study A	Study B	r	ρ	P
O	R1	0.99	1.00	1.00 (6/6)
O	R2	0.99	1.00	1.00 (6/6)
R1	R2	0.99	1.00	1.00 (6/6)

The design of this experiment is very similar to, and by the same authors as Hosking et al. (2022a), which was also found to be highly reproducible by Arvan and Parde (2024) and Arvan and Parde (2024) in ReproNLP 2024 (Belz and Thomson, 2024a).

4 LLM Sanity Check Results

In Section 3 we saw three sets of evaluations where at least two evaluations produced contradicting results: August et al., Reif et al., and Gu et al. For these three we report additional LLM evaluations following the general approach outlined in Section 2.3, and using the specific method described in Appendix C.

August et al. (2022a) B

Recall from Section 3 that the August et al. experiment reports results in two ways, where an error is counted if (i) both evaluators agree, and (ii) at least one evaluator identifies an error.

(i) Evaluators agree there is an error

The first table below shows *mean*[7] CV* (n=2), Pearson's r, Spearman's ρ, and proportion of same pairwise ranks P for O (the original study), R1 (Florescu et al., 2025), and the LLM sanity check.

Study A	Study B	CV*	r	ρ	P
O	R1	58.71	-0.33	-0.5	0.33 (1/3)
O	LLM	24.42	0.98	1.0	1.00 (3/3)
R1	LLM	53.13	-0.16	-0.5	0.33 (1/3)

O and the LLM check have medium mean CV* and perfect or near perfect agreement on the other measures. In contrast, R1 has poor QRA++ scores on all measures with both O and the LLM check. This means it is more likely that the original study is closer to the true picture than the reproduction. If we look at the system-level results in the next table below, we see that R1 produced scores for the three systems that were very close together, in the range 51–57. O and the LLM check place the systems much further apart.

System	O	R1	LLM	CV*
SVM	16.00	57.00	25.00	80.64
GeDi	33.00	51.00	35.00	30.40
DExperts	67.00	54.00	85.00	27.71
Mean CV*	–	–	–	46.25

From this table we can also see that the addition of the LLM check has improved CV* (n=3) values except for DExperts where it has increased slightly.

(ii) At least one evaluator finds an error

For the second aggregation method, the picture is similar: medium Type I reproducibility with (near) perfect Type II and IV reproducibility for O and the LLM check, and very poor reproducibility between R1 and each of the other two evaluations. (Recall that all R1 system scores were the same under this aggregation, so we can't report Pearson's and Spearman's.)

Study A	Study B	CV*	r	ρ	P
O	R1	39.46	nan	nan	0 (0/3)
O	LLM	35.18	0.99	1.0	1 (3/3)
R1	LLM	13.41	nan	nan	0 (0/3)

[7] Averaged over the system-level CV* scores.

The below table shows the system-level scores for all three evaluations, and the overall CV* (n=3). Here too the addition of the LLM check has improved CV* except for DExperts.

System	O	R1	LLM	CV*
SVM	38.00	78.00	68.00	41.49
GeDi	52.00	78.00	75.00	25.45
DExperts	86.00	78.00	98.00	14.09
Mean CV*	–	–	–	27.01

4.1 Reif et al. (2022)

The below table shows mean CV* (n=2), r, ρ and P for O (the original study), R1 (Onderková et al., 2025), and the LLM sanity check.

Study A	Study B	CV*	r	ρ	P
O	R1	29.86	0.32	-0.2	0.4 (6/15)
O	LLM	34.17	0.7	0.49	0.66 (10/15)
R1	LLM	16.13	0.33	0.26	0.6 (9/15)

This presents a very mixed picture: in terms of two-way CV*, R1 and LLM are somewhat closer than the other pairs, but on the other measures, O and LLM are closest. The LLM check evaluation appears to be somewhere between the other two. This could indicate that neither O nor R1 reflect the true picture (which would be revealed with more evaluators, and/or more evaluation) well.

For completeness, below we also show the system-level scores for the three evaluations alongside three-way CV* (n=3). Here too CV* has improved through the addition of the LLM results in all cases except the Dual RL system.

System	O	R1	LLM	CV*
Paraphrase	90.29	45.81	65.75	40.48
Zero-shot	69.71	49.44	44.95	29.48
Unsup. MT	86.76	73.32	55.92	26.25
Dual RL	85.29	68.24	53.98	27.70
Aug. zero-shot	86.47	65.10	64.75	21.09
Human	85.29	74.81	74.06	9.83
Mean CV*	–	–	–	25.81

4.2 Gu et al. (2022)

The below table shows mean CV* (n=2), r, ρ and P values for O (the original study), R1 (Mille and Lorandi, 2025), R2 (Junker, 2025), and the LLM sanity check.

Study A	Study B	CV*	r	ρ	P
O	R1	43.46	1	1	1/1
O	R2	23.25	-1	-1	0/1
O	LLM	30.86	1	1	1/1
R1	R2	45.27	-1	-1	0/1
R1	LLM	12.9	1	1	1/1
R2	LLM	32.99	-1	-1	0/1

Since there are only two systems in this experiment, correlations can only be either -1 or 1. What r, ρ and P tell us is that O and R1, O and LLM, and R1 and LLM are all in agreement, and that R2 is in disagreement with all of them (bearing in mind that with only two systems, hence one pairwise rank, these measures are less meaningful than with more systems). CV* nevertheless tells us that the system-level scores of O and R1 (in agreement on the other measures) are as different from each other as those of R1 and R2 (in disagreement on the other measures).

The below table shows the system-level scores for all four studies, alongside the four-way CV* (n=4). Here again, the latter has improved through the addition of the LLM evaluation.

System	O	R1	R2	LLM	CV*
MemSum	1.38	1.27	1.49	1.33	29.26
NeuSum	1.57	1.33	1.46	1.35	29.91
Mean CV*	–	–	–	–	29.58

5 Reproducibility by Quality Criterion and other properties

In this section, we look at some additional properties of our five sets of studies, to see if any pattern emerges as to which properties may be associated with better, and which with worse, reproducibility.

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 O and R1.
- **b(n=2)**: the mean of two-way CV* values between O and R2 (if there was an R2).

ReproNLP 2025							mean CV*		
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**	**a(n=2)**	**b(n=2)**	**n=3**
Yao et al. (2022) // *Braun (2025)* **Readability**	5 / 5	1–5	DQE	Goodness	Both	iiOR	25.85	-	-
August et al. (2022a) B // *Florescu et al. (2025)* **Factual Truth**	2 / 2	yes, no	DQE	Correctness	Content	EFoR	58.71	-	-
Bai et al. (2021) // *Supryadi et al. (2025)* **Informativeness**	7 / 7	-1, 0, 1	RQE	Goodness	Content	RtI	1.08	-	-
Reif et al. (2022) // *Onderková et al. (2025)* **Meaning Preservation**	6 / 6	0–100	DQE	Goodness	Form	RtI	29.86	-	-
Gu et al. (2022) // *Mille and Lorandi (2025) / Junker (2025)* **Overall quality**	4 / 4 / 4	1, 2	RQE	Goodness	Both	RtI	43.46	23.25	33.89
Hosking and Lapata (2021) // *Steen and Markert (2025) / Arvan and Parde (2025)* **Preservation of meaning**	UNK / 120 / 120	+1, -1	RQE	Goodness	Content	RtI	4.81	5.05	4.11

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 sample sizes are not directly comparable). The following columns map to experiment properties as recorded in HEDS 3.0 (Belz and Thomson, 2024b): 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).

- **n=3**: the mean of three-way CV* values between O, R1 and R2 (if there was an R2).

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 (Hosking & Lapata, Bai et al.) to be associated with better reproducibility, a pattern also observed in previous ReproNLP shared tasks (see Table 3).

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. An example is the evaluation of Factual Truth in August et al. which had the highest study-level, mean CV* of all studies reported. It also had the smallest number of evaluators. Another example is Meaning Preservation in Reif et al. which had some of the worst QRA++ values, and was also the most inconclusive of our sets of studies.

The two standout studies in terms of reproducibility on all measures were Bai et al. and Hosking & Lapata which share very similar properties as captured in Table 2: both use *relative quality estimation* (RQE) to assess the *goodness* of system outputs in terms of their *content* and *relative to the input* (RtI). Moreover, they both use a form of best-worst scaling.

6 Discussion

As in previous editions of ReproNLP, we saw that degree of reproducibility can look very different depending on 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 25.85).

While we've seen this happen a few times in ReproNLP, the inverse, excellent study-level, mean CV*, and then terrible correlations and P, we have never seen (as one would expect).

In Table 3 we have brought together all studies from ReproNLP 2023–2025 in slighly abbreviated form showing quality criteria, HEDS properties

ReproNLP 2023–2025							mean CV*	
Orig Study: **measurand**	**3.2.1**	**4.3.4**	**4.3.8**	**4.1.1**	**4.1.2**	**4.1.3**	**n=2**	**n=3**
Lin et al.: **Non-Redundancy**	3 / 3 / 3	0, 1, 2	DQE	Good.	Content	iiOR		2.83
Hosking and Lapata: **Pres. of meaning**	UNK / 120 / 120	+1, -1	RQE	Good.	Content	RtI		4.11
Hosking et al.: **Preserv. of meaning**	UNK / 180 / 180	A, B	RQE	Good.	Content	RtI		6.15
Lin et al.: **Informativeness**	3 / 3 / 3	0, 1, 2	DQE	Feature	Content	iiOR		7.18
Lin et al.: **Fluency**	3 / 3 / 3	0, 1, 2	DQE	Good.	Form	iiOR		9.89
Puduppully and Lapata B: **Mean # Supported Facts**	131/167/144	0–20	Count	Corr.	Content	RtI		11.88
Lux and Vu: **Naturalness (speech)**	34/157/37	A, B, Tie	RQE	Good.	Form	iiOR		14.55
Chakrabarty et al.: **Plausibility (simile)**	7/?/45	Yes, No	Cl/Lab	Good.	Both	RtI		15.69
Chakrabarty et al.: **Plausibility (idiom)**	4/?/35	Yes, No	Cl/Lab	Good.	Both	RtI		18.35
Puduppully and Lapata A: **Conciseness**	206/262/?	A, B	RQE	Good.	Both	iiOR		20.48
Puduppully and Lapata A: **Coherence**	206/262/?	A, B	RQE	Good.	Content	iiOR		21.12
Liu et al.: **Fluency**	UNK / 96 / 90	A, B, Tie	RQE	Good.	Both	iiOR		21.99
Puduppully and Lapata A: **Grammaticality**	206/262/?	A, B	RQE	Corr.	Form	iiOR		22.36
August et al. A: **Fluency**	2 / 2 / 2	1–4	DQE	Good.	Both	iiOR		26.87
Atanasova et al.: **Coverage**	3 / 3 / 3	1–3	RQE	Good.	Content	RtI		28.16
Gu et al.: **Overall quality**	4 / 4 / 4	1, 2	RQE	Good.	Both	RtI		33.89
Feng et al.: **Informativeness**	4 / 4 / 4	1–5	DQE	Good.	Content	RtI		55.52
Puduppully and Lapata B: **Mean # Contradicted Facts**	131/167/144	0–20	Count	Corr.	Content	RtI		84.78
Bai et al.: **Informativeness**	7 / 7	-1, 0, 1	RQE	Good.	Content	RtI	1.08	-
Castro Ferreira et al.: **Clarity**	60 / 60	1–7	DQE	Good.	Both	iiOR	3.44	-
Shardlow and Nawaz: **Ease of understanding**	98 / 40	1–4	RQE	Good.	Both	iiOR	5.95	-
Gabriel et al.: **Social acceptability**	UNK / 42	Yes, No	DQE	Feature	Both	EFoR	10.46	-
Yao et al.: **Readability**	5 / 5	1–5	DQE	Good.	Both	iiOR	25.85	-
Reif et al.: **Meaning Preservation**	6 / 6	0–100	DQE	Good.	Content	RtI	29.86	-
August et al. B: **Factual Truth**	2 / 2	yes, no	DQE	Corr.	Content	EFoR	58.71	-
Kasner and Dusek: **# Redundancies**	2 / 2	count	Count	Good.	Content	iiOR	149.72	-

Table 3: Quality criteria (measurands), HEDS properties and quality-criterion level CV* for all sets of evaluations from ReproNLP 2023–2025. Format is the same as Table 2 (see caption for column headings).

and quality-criterion level mean CV*. The top part of the table contains those studies where we currently have two ReproHum reproductions complete (n=3), while the lower part contains those where we currently have one reproduction (n=2). In each part of the table separately, we have sorted the study sets by CV*.

Among the general tendencies relating to single properties are the following. Larger numbers of evaluators tend to be associated with lower CV*, the one exception to this being Puduppully & Lapata B: Mean # Contradicted Facts. In all 13 other cases where a study has 7 or more evaluators, CV* is under 23, in 8 cases under 16.

Seven of the eight studies with a CV* under 11 have a very small number of possible response values (3 or fewer). Both of the two studies with the worst CV* values by a very large margin asked evaluators to count items directly. Relative quality estimation (RQE) seems to have the edge over direct quality estimation (DQE): the former has an average of 16.35 CV*, the latter 23.06.

In terms of combinations of properties, using a larger number of evaluators together with a small number of response values in RQE of Goodness has in all seven cases resulted in a CV* of under 22, in four cases, under 15. We have three studies assessing Meaning Preservation: two use RQE and achieve excellent CV*; the other one uses DQE and has poor CV*.

We applied LLM sanity checks for the first time in ReproNLP 2025 in order to shed light on which of two disagreeing studies is likely to be closer to the true picture. Of the three cases where there were disagreeing studies, the LLM sanity check was able to answer the question, but in the remaining case (Reif et al.), the LLM results correlated better with the original study than with R1, but

CV* was worse and P was very close for both O and R1. We will return to this analysis once the missing reproduction for Reif et al. is complete.

7 Conclusion

A shared task results report is almost invariably written under pressure of time and to a deadline. There are other aspects than are reported here which we would like to have investigated, but will have to leave for future work.

ReproNLP 2025 is the fifth and likely the last edition of this shared task series. It has contributed new data and insights into reproducibility and the factors that impact it, and we plan to release our resources and results so that further analyses can be conducted and insights gleaned.

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.

Our work was originally 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.

In particular, 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 ReproNLP work has also benefitted 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.

Craig Thomson's work was funded in part by ReproHum, and in part by ADAPT.

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.[8]
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 in terms of QRA++ (Belz, 2025) as follows:
      i. Type I results: Small-sample oefficient of variation CV* (Belz, 2022).

[8] We now call these Type IV results.

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.

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.

C LLM Sanity Check Method

In situations where the two (or in case three) reproductions disagree with each other, we employ a set of LLMs as a sanity check. We report the correlation results and indicate which of the human reproductions the LLM-based evaluations most closely align with, as they tend to show strong agreement with one reproduction while diverging from the other (Huidrom and Belz, 2025b).

The standardised procedure followed for the LLM sanity check is described below:

1. **Determining the number of LLMs.** Use the same number of distinct LLMs as human annotators per item in the original evaluation. That is, if the original evaluation involved 100 items, each annotated by 3 different human evaluators, we use 3 different LLMs to recreate this setup.

2. **Preparing the prompt.** This step involves adapting the original instructions provided to human annotators and clearly specifying the expected response format. The goal is to ensure that the LLMs receive well-structured and unambiguous prompts that reflect the textual and visual information conveyed by the original evaluation interface as closely as possible.

   (a) **Adaption of the instructions.** Use the same instructions provided to human annotators to perform the task, making only minimal modifications (e.g., remove the informed consent or some timing-related instructions, such as the minimum duration required for a valid submission).

   (b) **Verbalisation of the rating instrument.** Describe the rating scale and specify the expected response format (e.g., "Please answer using the following format: <ANSWER>A</ANSWER> in case your answer is A, or <ANSWER>B</ANSWER> in case your answer is B."). Always include a final clarification explicitly instructing the model not to include any information beyond the answer enclosed within the specified tags.

3. **Result extraction process.**

   (a) Apply the predefined extraction patterns, i.e., the response format explicitly indicated to the model in the prompt.

   (b) If it is not possible to extract responses for all items using the predefined patterns, design post-hoc extraction patterns. To do this, randomly sample the 10% of the outputs of each LLM. Use this set of samples as validation set and derive the post-hoc patterns based on the response formats observed in the validation set.

   (c) If there are still items for which responses cannot be extracted in some models, we assign random responses for those specific cases.

4. **Aggregation of the results.** Aggregate the results following the same procedure as in the original experiment with human annotators.

References

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

Mohammad Arvan and Natalie Parde. 2025. Reprohum #0744-02: Investigating the reproducibility of semantic preservation human evaluations. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

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. 2022a. 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.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022b. 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.

Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. 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 6910–6924, Online. Association for Computational Linguistics.

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

Anya Belz. 2025. QRA++: Quantified reproducibility assessment for common types of results in natural language processing. *arXiv e-prints*, pages arXiv–2505.

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

Anya Belz and Craig Thomson. 2024a. 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. 2024b. HEDS 3.0: The human evaluation data sheet version 3.0. *Preprint*, arXiv:2412.07940.

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

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.

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 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

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.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tanvi Dinkar, Gavin Abercrombie, and Verena Rieser. 2024. ReproHum #0927-03: DExpert evaluation? reproducing human judgements of the fluency of generated text. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 145–152, Torino, Italia. ELRA and ICCL.

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 Câmpeanu-Micluța, Stefana Arina Tabusca, and Liviu P Dinu. 2025. Reprohum #0033-05: Human evaluation of factuality from a multidisciplinary perspective. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

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.

Javier González Corbelle, Ainhoa Vivel Couso, Jose Maria Alonso-Moral, and Alberto Bugarín-Diz. 2024. ReproHum #0927-3: Reproducing the human evaluation of the DExperts controlled text generation method. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 153–162, Torino, Italia. ELRA and ICCL.

Nianlong Gu, Elliott Ash, and Richard Hahnloser. 2022. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6507–6522, Dublin, Ireland. Association for Computational Linguistics.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1406–1419, Online. Association for Computational Linguistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2022a. 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.

Tom Hosking, Hao Tang, and Mirella Lapata. 2022b. 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.

Rudali Huidrom and Anya Belz. 2025a. Ask me like i'm human: Llm-based evaluation with for-human instructions correlates better with human evaluations than human judges. In *4th Table Representation Learning Workshop*.

Rudali Huidrom and Anya Belz. 2025b. Using LLM judgements for sanity checking results and reproducibility of human evaluations in NLP. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

Simeon Junker. 2025. Reprohum #0729–04: Human evaluation reproduction report for "memsum: Extractive summarization of long documents using multi-step episodic markov decision processes". In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

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.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021a. 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.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021b. 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.

Florian Lux and Thang Vu. 2022. Language-agnostic meta-learning for low-resource text-to-speech with articulatory features. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6858–6868, Dublin, Ireland. Association for Computational Linguistics.

Simon Mille and Michela Lorandi. 2025. Reprohum #0729–04: Partial reproduction of the human evaluation of the memsum and neusum summarisation

systems. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

Kristýna Onderková, Mateusz Lango, Patrícia Schmidtová, and Ondrej Dusek. 2025. Reprohum #0669-08: Reproducing sentiment transfer evaluation. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

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.

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.

Julius Steen and Katja Markert. 2025. Reprohum #0744-02: A reproduction of the human evaluation of meaning preservation in "factorising meaning and form for intent-preserving paraphrasing". In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

Supryadi, Chuang Liu, and Deyi Xiong. 2025. Reprohum #0067-01: A reproduction of the evaluation of cross-lingual summarization. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

Craig Thomson, Ehud Reiter, and Belz Anya. 2024. Common flaws in running human evaluation experiments in nlp. *Computational Linguistics*.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun 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.