ReproHum 0031–01: Reproducing a Human Readability Evaluation for Question–Answer Generation Systems

Manuela Hürlimann
Centre for Artificial Intelligence,
Zurich University of Applied Sciences,
Winterthur, Switzerland
manuela.huerlimann@zhaw.ch

Mark Cieliebak
Centre for Artificial Intelligence,
Zurich University of Applied Sciences,
Winterthur, Switzerland
mark.cieliebak@zhaw.ch

Abstract

Human evaluations play a central role in assessing natural language processing systems, yet their robustness and reproducibility remain incompletely understood. This paper reports on a reproduction of the human readability evaluation from Yao et al. (2022) for question–answer generation (QAG) systems, conducted within the ReproHum project and the ReproNLP 2026 shared task (Belz et al., 2026). The original evaluation compared three QAG systems with respect to three criteria. We reproduced the evaluation of one of these criteria, readability, using a new group of five evaluators. We report descriptive results, inter-annotator agreement, system-level comparisons, and cross-study robustness metrics compared to the original study and two previous reproductions. Our results support all conclusions of the original evaluation and are largely consistent with two previous reproductions.

1 Introduction

Human evaluation is widely used in Natural Language Processing (NLP) to assess properties of NLP systems that are difficult to capture automatically, such as readability, relevance, or overall quality of generated texts (Howcroft et al., 2020). At the same time, human evaluation outcomes can be sensitive to design choices such as participant background, instructions, and interfaces, as well as to properties of the evaluated items themselves (Belz et al., 2020; Shimorina and Belz, 2022). The ReproHum project addresses these issues via reproduction studies that revisit published human evaluations and systematically document or vary experimental conditions to investigate which factors influence the robustness of human evaluation results (Belz, Agarwal, Shimorina, and Reiter, 2021; Belz, 2025).

We contribute to ReproHum a reproduction of the human readability evaluation reported by Yao et al. (2022) in the context of question–answer generation (QAG). In the original study, human-written (ground-truth) question–answer pairs were compared against two automated QAG systems. The human evaluation was conducted by five evaluators using an Excel-based annotation interface.

We reproduce this evaluation as closely as possible by using the same evaluation interface format and the same data points, but with a new group of five evaluators. Our aim is not to reassess the quality of the systems themselves, but to examine the robustness of the human evaluation outcomes. To this end, we compare our results to the original study and to two prior reproductions (Florescu et al., 2024; Braun, 2025), quantifying agreement and cross-study variability and testing whether system-level differences persist. In doing so, we provide an additional data point for ReproHum's systematic mapping of robustness and reproducibility in human evaluation.

While confirming the original evaluation's claims, our analysis highlights systematic differences in inter-annotator agreement across studies and provides further evidence that variability in human evaluation arises both within and across studies.

We release the data and code on GitHub [1]. The Human Evaluation Data Sheet (HEDS; (Belz and Thomson, 2025) for this reproduction is available through the ReproNLP HEDS repository[2].

2 Background

2.1 Original Evaluation Task

Yao et al. (2022) compare three different QAG approaches using both automated metrics and human evaluations: (i) ground-truth QA pairs written by humans (which we will refer to as HUMAN), (ii)

[1] https://github.com/manhue/ReproNLP2026_0031-01

[2] https://github.com/nlp-heds/repronlp2026

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1111–1116

the PAQ system by Lewis et al. (2021), and (iii) the QAG system proposed by Yao et al. (which we will refer to as YEA, short for "Yao et al."). In their human evaluation, five evaluators rated 722 question–answer (QA) pairs[3] according to three criteria: Readability, Question Relevancy and Answer Relevancy. Rating was done using a five-point Likert scale (1 = worst, 5 = best) and evaluators were blind to system identity.

Each evaluator annotated items for 16 sections from four books and all three systems. They were provided the section context, the question, and the answer. Each item was annotated by two evaluators.

2.2 Previous Reproductions

There are two previous reproduction reports of this evaluation. Florescu et al. (2024) reproduced the human evaluation of all three criteria with undergraduate student participants. They reported challenges around inter-annotator agreement but in general confirmed the claims of the original study. Braun (2025) reproduced the readability criterion with NLP-expert participants and reported results aligned with the original study's main readability claims. In the present report, we compare our reproduction to the original study and to both previous reproductions.

3 Reproduction Study

3.1 Participants and Compensation

The new evaluators were recruited via relevant mailing lists and provided informed consent prior to starting the evaluation. With respect to language background, potential evaluators were asked to self-assess whether they possess a good working knowledge of English, which we considered a prerequisite for participating in the evaluation.

As intended by the reproduction instructions, five evaluators participated in our reproduction. They were four PhD students and one non-student researcher, all with an NLP/AI background. They completed the task in approximately 90 minutes to two hours and each received CHF 40 in vouchers.

3.2 Materials, Interface, and Instructions

An Excel spreadsheet provided by the original authors (Yao et al., 2022) served as the interface for evaluation. This spreadsheet contained the results of the original study and included the following columns:

- `labeller_id`
- `qa_id`
- `story_name`
- `section_id`
- `section` (text that the question and answer refer to)
- `source` (the system: HUMAN (labelled as "groundtruth"), PAQ (labelled as "PAQ"), or YEA (labelled as "Ours"))
- `split` (test or validation)
- `question`
- `answer`
- `readability` (grammatically correct and clear language; 1 = worst, 5 = best)
- `relevancy_Q` (question is relevant to the section; 1–5)
- `relevancy_A` (answer can correctly answer the question; 1–5)

We preserved these column names from the original study in accordance with ReproHum's instructions. Using the `labeller_id`, we split the original spreadsheet into five separate sheets to be provided to our evaluators, preserving the original assignment of items to evaluators.

In these sheets, we only showed the five columns relevant to a blind assessment of the readability criterion: `qa_id` (to map answers during analysis), `section`, `question`, `answer`, and `readability`. The numeric contents of the `readability` column were cleared before providing the sheets to evaluators.

3.3 Differences Between Original Study and Our Reproduction

We are aware of the following differences to the original study:

- **Evaluators**: in the original study, there were three NLP experts and two education experts. Our evaluators were all NLP/AI experts.

[3] 242 from HUMAN and 240 each from PAQ and YEA.

System	Mean	SD
HUMAN	4.18	0.99
YEA	3.60	1.54
PAQ	3.25	1.62

Table 1: Mean and standard deviation of readability ratings in our reproduction.

System 1	System 2	t	p
HUMAN	PAQ	7.6051	2.10e-13
HUMAN	YEA	4.9684	9.96e-07
PAQ	YEA	-2.3962	0.0170

Table 2: Pairwise Welch t-tests between system ratings in our reproduction.

- **Compensation**: there was no compensation in the original study, whereas our evaluators received 40 CHF in vouchers.
- **Instructions**: Because the original detailed instruction text was not recoverable, we used the standardised instructions provided by the ReproHum project team.

4 Evaluation Results

In this section, we present the results of our human evaluation as if it were a stand-alone study. In section 5, we compare our results to those of the original study and the two previous reproductions.

4.1 Descriptive Statistics

Table 1 shows the means and standard deviations of readability ratings per system in our evaluation. HUMAN answers receive the highest mean readability score, followed by YEA and then PAQ.

Pairwise Welch t-tests (Table 2) indicate that readability ratings differ between all system pairs. HUMAN is rated significantly higher than YEA ($p < 10^{-6}$) and PAQ ($p < 10^{-12}$). YEA outperforms PAQ ($p = 0.017$), although this difference would be considered not significant if a conservative multiple-comparison correction was applied.

4.2 Inter-Annotator Agreement

To assess the consistency of judgements, we report Krippendorff's α per system (ordinal) in Table 3. Overall Krippendorff's α is 0.32, indicating modest agreement. Agreement is lowest for HUMAN, followed by PAQ, and highest for YEA.

System	α
HUMAN	0.22
YEA	0.37
PAQ	0.24

Table 3: Krippendorff's α (ordinal) per system for readability in our reproduction.

E1	E2	Shared items	α
0	1	72	0.3070
0	4	62	0.1858
1	2	63	0.3376
2	3	92	0.4923
3	4	72	0.2075

Table 4: Pairwise Krippendorff's α between evaluator pairs (E1, E2) in our reproduction.

Pairwise Krippendorff's α values between evaluator pairs (Table 4) with overlapping items range from 0.19 to 0.49 (median $\approx$ 0.31), indicating modest agreement overall with substantial variability across rater pairs.

5 Comparison with Previous Studies

We now compare our reproduction to the original study by Yao et al. (2022) and to the two previous reproductions (Florescu et al., 2024; Braun, 2025).

5.1 System means

Table 5 shows the readability results across all four studies. HUMAN QA pairs consistently receive the highest mean readability score, followed by YEA and then PAQ. Absolute scores differ substantially across studies, with our reproduction yielding lower mean scores for HUMAN and YEA than the original and the Florescu et al. (2024) reproduction, and values comparable to those reported by Braun (2025).

5.2 Inter-annotator agreement

Table 6 provides an overview of inter-annotator agreement for readability judgements. Note that the values for Yao et al. (2022) were calculated by Florescu et al. (2024) and that the studies report values at different levels (overall, per system, or both). Previous reproductions also reported modest agreement overall and emphasised that agreement estimates are sensitive under sparse annotation of two evaluators per item. Our overall α of 0.32 is

Table 5: Means and standard deviations of readability ratings in four evaluations.

System	Study							
	Yao et al.		Florescu et al.		Braun		This work (2026)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
HUMAN	4.95	0.28	4.71	0.52	4.38	0.96	4.18	0.99
YEA	4.71	0.70	4.52	0.75	3.85	1.35	3.60	1.54
PAQ	4.08	1.13	4.17	1.22	3.14	1.43	3.25	1.62

in the same general range as the values reported by Florescu et al. (2024) for their reproduction (α=0.27; three evaluation criteria) and by Braun (2025) for readability (α=0.41).

Regarding patterns per-system: agreement in our reproduction is lowest for HUMAN (α=0.22), despite blinding, which is also what Florescu et al. (2024) observe (their negative value even indicates systematic differences between evaluators). A possible interpretation is that high-quality (human-written) QA pairs afford more nuanced readability judgements, leading to greater subjectivity and thus increased evaluator variability. In contrast, Yao et al. (2022) report very high agreement for HUMAN and substantially lower agreement for the other systems. Braun (2025) report overall agreement only.

5.3 QRA++

To quantify cross-study robustness of aggregate results, we compute Quantified Reproducibility metrics using the QRA++ framework (Belz and Thomson, 2026). The results are shown in Table **??**.

The coefficient of variation (CV*) measures the relative dispersion of repeated measurements as a percentage of their mean. Lower CV* values indicate more robust and stable results across studies, whereas higher values indicate greater variability.

On the system level, the four-way CV* values are lowest for HUMAN and highest for PAQ, indicating comparatively stable reproducibility for ground-truth and only moderate reproducibility for automated systems. The aggregate CV for readability across systems (16.88) is also in the moderate range.

Pearson's and Spearman's correlations (Type II) as well as pairwise rank precision (Type IV) show very high agreement on system ordering across the four evaluations.

5.4 Support for Conclusions from Original Study

We now summarise the conclusions drawn by Yao et al. (2022) from their human readability evaluation and evaluate whether they are supported by our reproduction results.

Conclusion 1: YEA outperforms PAQ with regard to readability.

- Evidence:
  - Mean(YEA)=3.60 vs. Mean(PAQ)=3.25
  - Welch $t = -2.40$, $p = 0.017$
- Supported?: Yes.

Conclusion 2: HUMAN outperforms YEA with regard to readability.

- Evidence:
  - Mean(HUMAN)=4.18 vs. Mean(YEA)=3.60
  - Welch $t = 4.97$, $p < 10^{-6}$.
- Supported?: Yes.

Conclusion 3: YEA achieves an above-average (> 3) readability rating.

- Evidence:
  - Mean(YEA)=3.60
- Supported?: Yes.

All three readability-related claims are verified.

6 Conclusion

We repeated the evaluation of Yao et al. (2022) as part of the ReproHum project and the ReproNLP 2026 shared task (Belz and Thomson, 2026). Five human evaluators rated the readability of 722 Question-Answer pairs created by humans (HUMAN) and two automated systems (PAQ, YEA). Our results show significant differences between all system pairs, and we have moderate inter-annotator agreement. When comparing our results

Table 6: Krippendorff's α for readability in four evaluations.

System	Yao et al.*	Florescu et al.	Braun	This work (2026)
all	–	–	*0.41*	*0.32*
HUMAN	0.94	-0.13	–	0.22
YEA	0.25	-0.06	–	0.37
PAQ	0.24	0.05	–	0.24

* Calculation by Florescu et al. (2024).

Table 7: **Yao et al. 2022**: QRA reproducibility assessment for four comparable experiments (n=4), Yao et al. (2022), Florescu et al. (2024), Braun (2025), and This work (2026). QC1 = Readability; n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 4$)		
				System level	QC level	Study level
Type I	QC1	HUMAN	(mean) CV*↓	10.43	16.88	16.88
		YEA		18.19		
		PAQ		22.04		
Type II	QC1	all	mean r ↑	n/a	0.970	n/a
		all	mean ρ ↑	n/a	1.000	
		all	W↑	n/a	1.000	
Type IV	QC1	all	P ↑	n/a	1.000	1.000

with the original study and previous reproductions, we see the same ordering of systems in each evaluation, although absolute readability scores differ. Regarding inter-annotator agreement, our results broadly align with previous reproductions, but an interesting difference concerns the diverging agreement scores for HUMAN items. Our evaluations supports all conclusions of Yao et al. (2022) with respect to the relative perception of the systems. Overall, our results show that system rankings are robust across studies, but absolute scores and agreement levels vary considerably. This indicates that human evaluation results are consistent in relative terms, while still being sensitive to study conditions.

Limitations

As described in Section 3.3, there were differences between the original evaluation experiment and our reproduction.

Acknowledgements

We thank the ReproHum project team for guidance and for providing standardised materials and instructions, and we thank our evaluators for their time and careful work.

References

Anya Belz. 2025. QRA++: Quantified Reproducibility Assessment for Common Types of Results in Natural Language Processing. *arXiv preprint*. ArXiv:2505.17043.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. 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.

Anya Belz, Simon Mille, and Dimitar Shterionov. 2020. Disentangling the Properties of Human Evaluation Methods. In *Proceedings of the 13th International Conference on Natural Language Generation*.

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 Common Types of Evaluation Results in NLP/ML. *Computational Linguistics*, pages 1–10.

Anya Belz, Craig Thomson, and Javier Gonz'alez Corbelle. 2026. "The Shared Task on Reproducibility of Evaluations in NLP (ReproNLP) 2026: Overview and Results". In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San

Diego, USA. Association for Computational Linguistics.

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*.

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 Workshop on Human Evaluation of NLP Systems*.

David Howcroft, Anya Belz, Dimitra Gkatzia, Shailza Hasan, Saad Mahamood, and Simon Mille. 2020. Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Kuttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Anastasia Shimorina and Anya Belz. 2022. The Human Evaluation Datasheet. In *Proceedings of the Workshop on Human Evaluation of NLP Systems*.

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*.