ReproHum #0031-01: Reproducing the Human Evaluation of Readability from “It is AI’s Turn to Ask Humans a Question”

Daniel Braun
Marburg University
Department of Mathematics and Computer Science
daniel.braun@uni-marburg.de

Abstract

The reproducibility of results is the foundation on which scientific credibility is built. In Natural Language Processing (NLP) research, human evaluation is often seen as the gold standard of evaluation. This paper presents the reproduction of a human evaluation of a Natural Language Generation (NLG) system that generates pairs of questions and answers based on children’s stories that was originally conducted by Yao et al. (2022). Specifically, it reproduces the evaluation of readability, one of the most commonly evaluated criteria for NLG systems. The results of the reproduction are aligned with the original findings and all major claims of the original paper are confirmed.

1 Introduction

Reproducibility is one of the main measures for good science. By reproducing studies from other researchers and confirming their results, scientific findings can be independently verified. In recent years, surveys about reproducibility revealed widespread problems across disciplines (Baker, 2016). Natural Language Processing (NLP) is no exception and also suffers from a variety of problems with regard to the reproducibility of scientific results (Cohen et al., 2018; Belz et al., 2023).

Evaluation metrics, like Precision, Recall, and Accuracy, but also more sophisticated metrics, like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or the BERT-Score (Zhang et al., 2019), are widely used in the evaluation of NLP systems. They are not only cheap and fast to calculate but also promise a high degree of reproducibility: Most metrics guarantee that for the same input, they will always produce the same score.[1] However, they do not always correlate well with human judgments (Reiter and Belz, 2009; Reiter, 2018) and, particularly for Natural Language Generation (NLG), often fail to take into account the many different aspects that influence the overall assessment made by humans.

Human evaluation, therefore, in the field of NLP plays an important role and is often seen as the best available option for evaluation (Howcroft et al., 2020). However, designing good, reproducible human evaluations is much more difficult than the use of automated metrics and many existing human evaluations suffer from problems that limit their reproducibility (Schuff et al., 2023; Thomson et al., 2024). The ReproHum project and the associated ReproNLP shared taks (Belz and Thomson, 2023, 2024; Belz et al., 2025) aim to address this problem by analysing the reproducibility of human evaluations in NLP and developing a methodological framework to assess the reproducibility of human evaluations.[2]

As part of the project, multiple partner labs attempt to reproduce results from selected papers that report on human evaluations in NLP. This paper presents the results of such a reproduction study for the paper “It’s AI’s turn to ask Humans a Question” by (Yao et al., 2022). The original paper introduces a new approach for the generation of question-answer pairs and compares that approach against two baselines in a human evaluation.

A previous reproduction of the study was conducted by Florescu et al. (2024) who concluded that “All in all, we managed to replicate the original study”. We believe that the design of the reproduction presented in this paper, which comes to a similar conclusion, is closer to the original design, particularly, because we recruited participants with the same background as the original participants (namely NLP experts; see also Section 3.2), while the participants recruited by Florescu et al. (2024) have a different background (namely undergraduate students).

[1] Yet, reproduction of metric-based results can still be difficult as pointed out by Chen et al. (2022).

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

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 576–582
July 31 - August 1, 2025

While the original paper and the reproduction by Florescu et al. (2024) investigate three quality criteria in the human evaluation, namely readability, question relevancy, and answer relevancy, we only reproduced the evaluation of one quality criterion, namely readability. As shown by Howcroft et al. (2020), readability is one of the most frequently evaluated quality criteria in human evaluation of NLG systems, although it is not always evaluated under this specific term and the definitions of it vary.[3]

Based on the information provided in the original paper and additional information obtained from the original authors by the ReproHum project team, we were able to reproduce the main findings of the original paper: While the scores obtained in the reproduction of the human evaluation slightly differ from the original scores, the ranking of the compared systems and all major claims made in the original paper with regard to the readability evaluation could be verified.

The results of our reproduction, the code used to calculate the reported metrics, and a Human Evaluation Datasheet (HEDS, Shimorina and Belz (2022)) for the reproduction experiment are available on GitHub.[4] The HEDS file is also available in the central ReproNLP 2025 HEDS repository.[5]

2 Original Study

Yao et al. (2022) introduce a question-answer pair generator that is designed for educational purposes: based on story books for readers from kindergarten to eighth-grade, the system automatically generates question-answer pairs that are designed to test different dimensions of comprehension skills.

The architecture of the system consists of three main components:

1. A heuristic-based **answer generation module** that generates candidate answers from story passages.
2. A BART-based (Lewis et al., 2020) **question generation module** that, based on the candidate answers and the story passage, generates corresponding questions.
3. And finally, a DistilBERT-based (Sanh et al., 2019) **ranking module** that selects the final question-answer pairs from the generated candidate pairs.

The modules have been fine-tuned on the FairytaleQA dataset (Xu et al., 2022). The dataset consists of over 10,000 QA pairs from almost 300 children's books, which have been specifically designed to test reading comprehension.

2.1 Automated Evaluation

While the training split of the FairytaleQA dataset was used to fine-tune the modules, the authors used the validation and test set to evaluate the system. The evaluation consists of both, an automated, metric-based, evaluation and a human evaluation.

For the automated evaluation, the newly introduced system was compared against a state-of-the-art QA pair generation system by Shakeri et al. (2020) that uses a two-step approach and a PAQ baseline system (Lewis et al., 2021). The metric used for the evaluation was ROUGE-L (Lin, 2004). In the metric-based evaluation, the new system introduced by Yao et al. (2022) clearly outperformed the two baseline systems and the PAQ baseline system outperformed the system by Shakeri et al. (2020).

2.2 Human Evaluation

Based on the results of the automated evaluation, the authors of the original paper decided to only use the output of the better performing PAQ as system baseline in the human evaluation. In addition to the output of the newly introduced system and the PAQ baseline, participants in the human evaluation were also shown human-generated QA pairs from the dataset ("groundtruth").

2.2.1 Participants

The original paper only disclosed that "five human participants" participated in the human evaluation. In order to facilitate the replication of the original study, the ReproHum team requested additional information from the authors which they kindly shared: out of the five participants four were faculty (professors or researchers) and one grad student. Two of the five participants were education experts, while the other three were NLP experts.

2.2.2 Quality Criteria

The human evaluation in the original paper investigated three quality criteria and defined them as

[3] Other terms used for readability include fluency, goodness of outputs in their own right, and quality of outputs (Howcroft et al., 2020)

[4] `https://github.com/Responsible-NLP/ReproHum-0031-01`

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

follows:

- **Readability**: "The generated QA pair is in readable English grammar and words."
- **Question Relevancy**: "The generated question is relevant to the storybook section."
- **Answer Relevancy**: "The generated answer is relevant to the question."

2.2.3 Instructions and Interface

Each participant annotated QA pairs for 16 story sections from 4 books. For each of the 16 sections, participants received on average 9 QA pairs, 3 from each model. The exact number of annotated pairs varied between participants. Each QA pair was annotated by 2 annotators. Overall 722 QA pairs have been rated. While the original paper does not provide information about the detailed annotation instructions and interface, additional information have been obtained by the ReproHum team. While the exact instruction that participants received were unfortunately not retrievable anymoe, the Excel that was send to participants to annotate the data itself was provided to the ReproHum team (see Figure 1). The annotation sheet consists of six columns with the headers:

- "section",
- "question",
- "answer",
- "readability (grammarly correct and clear language. worst 1 to 5)",
- "relevancy_Q (Q is relevant to section. 1 to 5)", and
- "relevancy_A (Answer can correctly answer the Q. 1 to 5)".

Notably, the explanations provided in the paper for readability ("The generated QA pair is in readable English grammar and words.") and answer relevancy ("The generated answer is relevant to the question.") differ from the explanations provided in the sheet.

2.2.4 Results and Claims

The main results of the human evaluation are shown in Table 2. For all three criteria, the system proposed by Yao et al. (2022) outperformed the PAQ baseline, but could not beat the "groundtruth". Moreover, the authors point out that their model has "above-average (>3) ratings" in all categories.

3 Reproduction

The paper and its human evaluation have been chosen by the ReproHum team to be reproduced by a partner lab. We conducted the reproduction according to the ReproHum guidelines and instructions.

3.1 Scope

In accordance with ReproHum protocol, our reproduction was restricted to just one of the three quality criteria measured in the original human evaluation, namely readability.

3.2 Participants

Like the original study (see Section 2.2.1), five people participated in the reproduction study. Out of those five, two were non-student researchers and three were grad students (particularly PhD students). All five participants are experts in the field of NLP. Unlike the original study, we compensated the participants. In accordance with ReproHum protocol, the compensation was based on the UK Living Wage (which was higher than the local minimum wage). Based on a pilot annotation conducted by the authors, the maximum completion time was estimated to be 1.5 hours and the compensation was a 25 EUR Amazon voucher.

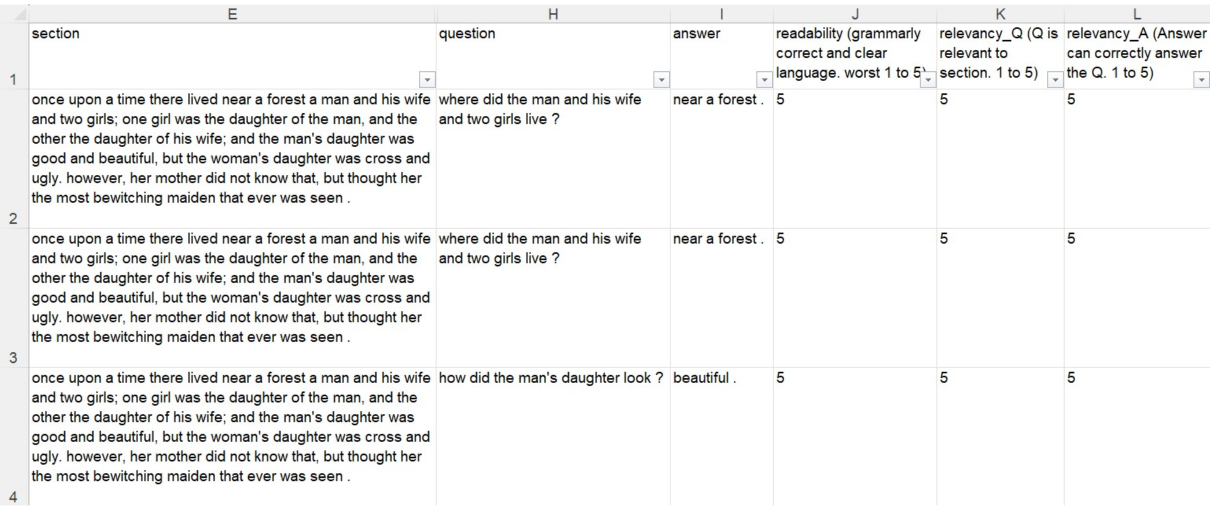
3.3 Instructions and Interface

We used the exact same Excel sheet for the reproduction that was also used during the original study and split the data between participants in accordance with the parameters described in Section 2.2.3. Since the original instructions for participants were not known, and to comply with ethical requirements, we drafted new instructions. In order to minimise any potential influence on the results, we kept the instructions as short as possible (see Appendix A for the full instructions).

3.4 Known Deviations

To summarise, there are three aspects in which we know that our reproduction deviated from the original experiment.

- **Background of Participants**: While the original study recruited three NLP experts and two educational experts, all our participants were NLP experts.
- **Compensation**: While the participants in the original experiment received no compensation, we compensated our participants with 25 EUR vouchers.



	E	H	I	J	K	L
1	section	question	answer	readability (grammarly correct and clear language. worst 1 to 5	relevancy_Q (Q is relevant to section. 1 to 5)	relevancy_A (Answer can correctly answer the Q. 1 to 5)
2	once upon a time there lived near a forest a man and his wife and two girls; one girl was the daughter of the man, and the other the daughter of his wife; and the man's daughter was good and beautiful, but the woman's daughter was cross and ugly. however, her mother did not know that, but thought her the most bewitching maiden that ever was seen .	where did the man and his wife and two girls live ?	near a forest .	5	5	5
3	once upon a time there lived near a forest a man and his wife and two girls; one girl was the daughter of the man, and the other the daughter of his wife; and the man's daughter was good and beautiful, but the woman's daughter was cross and ugly. however, her mother did not know that, but thought her the most bewitching maiden that ever was seen .	where did the man and his wife and two girls live ?	near a forest .	5	5	5
4	once upon a time there lived near a forest a man and his wife and two girls; one girl was the daughter of the man, and the other the daughter of his wife; and the man's daughter was good and beautiful, but the woman's daughter was cross and ugly. however, her mother did not know that, but thought her the most bewitching maiden that ever was seen .	how did the man's daughter look ?	beautiful .	5	5	5

Figure 1: Excel Sheet used by the anntoators in both the original experiment and the reproduction

	Yao et al.		PAQ Baseline		Groundtruth	
	M	SD	M	SD	M	SD
Readability (1 to 5)	4.71	0.70	4.08	1.13	4.95	0.28
Question Relevancy (1 to 5)	4.39	1.15	4.18	1.22	4.92	0.33
Answer Relevancy (1 to 5)	3.99	1.51	3.90	1.62	4.83	0.57

Table 1: Human evaluation results as presented by Yao et al. (2022)

- **Instructions**: Because the original instruction could not be recovered, we wrote new instructions.

4 Results

The results of the reproduction are shown in Table 2 as "Reproduction". In the reproduction experiment the inter-annotator agreement (IAA) was low with Krippendorff's $\alpha = 0.41$. While the original paper reports a much higher IAA at α "between 0.73 and 0.79" (Yao et al., 2022), we have been unable to reproduce those values based on the original data. Florescu et al. (2024) were also unable to reproduce the α values and calculated Krippendorff's $\alpha = 0.43$ on the original data, which is much closer to our results.

In the reproduction the best readability scores were achieved by the "groundtruth" QA pairs (mean 4.38 on a scale from 1 (worst) to 5 (best)), followed by the pairs generated by the system introduced by (Yao et al., 2022) (mean 3.85), and the PAQ baseline (mean 3.14).

4.1 Comparison

With regard to the readability of the generated QA pairs, Yao et al. (2022) conducted t-tests and summarised that their model "performed significantly better than the PAQ model (avg = 4.08, s.d.=1.13, t(477) = 7.33, p $<$ .01), but was not as good as the groundtruth (avg = 4.95, s.d. = 0.28, t(479) = -4.85, p $<$ .01)." (Yao et al., 2022, p. 738).[6]

While, as Table 2 shows, the average scores in the reproduction were lower across all three models and the standard deviation was higher, the reproduction too found that the model introduced by Yao et al. (2022) significantly outperformed the PAQ model (t(477) = 5.56, p $<$ .01) but was not as good as the groundtruth (t(479) = -5.05, p $<$ .01). Similarly, despite the drop in the average score, the observation that the new model "has above-average ($>$3) ratings" (Yao et al., 2022, p. 738) also still holds in the reproduction. In summary, all claims made in the original paper with regard to the readability could be verified in the reproduction (see Table 3).

In comparison to the previous reproduction by Florescu et al. (2024), Table 2 shows that our evaluation resulted in even lower scores than the scores obtained by Florescu et al. (2024), which were already lower than the original results. While, relatively speaking, both reproductions are in line with the original results, the absolute numbers reported

[6] We were able to reproduce and thereby verify the results of the performed t-tests based on the provided data.

	Yao et al.		PAQ Baseline		Groundtruth	
	M	SD	M	SD	M	SD
Yao et al. (2022)	4.71	0.70	4.08	1.13	4.95	0.28
Reproduction	3.85	1.35	3.14	1.43	4.38	0.96
CV*	26.14		35.01		15.51	
Δ (Reproduction - Yao et al.)	-0.86	+0.65	-0.94	+0.3	-0.57	+0.68
Florescu et al. (2024)	4.52	0.75	4.17	1.22	4.71	0.52
CV* (Florescu et al.)	4.10		2.18		4.95	
Δ (Florescu et al. - Yao et al.)	-0.19	+0.05	+0.09	+0.09	-0.24	+ 0.24

Table 2: Results of the human evaluation (original, reproduction, and difference, as well as the results obtained in the previous reproduction by Florescu et al. (2024)) of readability on a scale from 1 (worst) to 5 (best); mean (M) and standard deviation (SD)

Claim	Verified
The model by Yao et al. outperforms the PAQ model with regard to readability	yes
The groundtruth outperformed the model by Yao et al. with regard to readability	yes
The model by Yao et al. achieves an average rating > 3 for readability	yes

Table 3: Claims based on readability in the original paper and their verifications.

by Florescu et al. (2024) are much closer to the original results than ours.

4.1.1 Quantification of Reproducibility

In accordance with ReproHum protocol, we also calculated different measures to quantify the reproducibility of the experiment. First, we calculated the coefficient of variation (CV), which is the ratio of the standard deviation of the results to the means. Particularly, we used the adapted version CV* introduced by Belz (2022, 2025), which is adjusted for small sample sizes.[7] As shown in Table 2, CV* values vary between 15.51 and 35.01, which seems high in comparison to other ReproHum reproductions (see e.g. Van Miltenburg et al. (2023) and Arvan and Parde (2024)).

Additionally, we calculated correlations between the original results and the reproduction. The Pearson correlation indicates a very strong positive correlation ($r = 0.9856$), however, at $p = 0.108$, the correlation is not statistically significant. Spearman's Rho at $\rho = 1.0$ indicates a perfect positive monotonic relationship between the results of the original study and the replication. However, the small sample size should be kept in mind when interpreting both metrics.

Lastly, we find that the ranking of the systems is equal in the original study and the reproduction and that all major claims (see Table 3) can be verified.

5 Conclusion

The results of our reproduction confirm all major claims and results of the original experiment with regard to the readability of the generated question-answer pairs.

While the order in which the systems have been evaluated is the same, the absolute scores received by each system vary from the original results by up to 20%. Given how vaguely defined *readability* and the scale it was judged on (from 1 to 5) were in the experiment, this does not seem surprising. However, other factors could have also influenced that the results of the reproduction were overall more critical, e.g. the composition of the participants (the reproduction consisted only of NLP researchers) or temporal effects (while the original study was conducted before the public availability of Large Language Models, like ChatGPT, the reproduction was conducted in 2024, when expectations towards the quality of AI-generated texts might already have been higher).

Lastly, it is worth pointing out that our reproduction heavily relied on information that was not available in the original paper, but was kindly provided by its authors to the ReproHum project. If we would have based our reproduction attempt solely on the information provided in the paper, our experimental setup would have, in all likelihood, looked significantly different and might well have yielded different results.

[7] In order to ensure comparability we shifted the values by -1 to ensure that the scale starts at 0.

Limitations

As pointed out in Section 3.4, we know that our reproduction differs in certain aspects from the original experiment.

Acknowledgments

We would like to thank the ReproHum project, and especially Craig Thomson, for facilitating this reproduction study. We would also like to thank the original authors for providing additional information to the ReproHum team.

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.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454.

Anja 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.

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. *Preprint*, arXiv:2505.17043.

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, Craig Thomson, Javier González-Corbelle, and Malo Ruelle. 2025. The 2025 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM^2)*.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Yanran Chen, Jonas Belouadi, and Steffen Eger. 2022. Reproducibility issues for BERT-based evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2965–2989, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, 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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 29(5):1199–1222.

Siamak Shakeri, Cicero Nogueira dos Santos, Henry Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. *arXiv preprint arXiv:2010.06028*.

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in nlp. *Computational Linguistics*, 50(2):795–805.

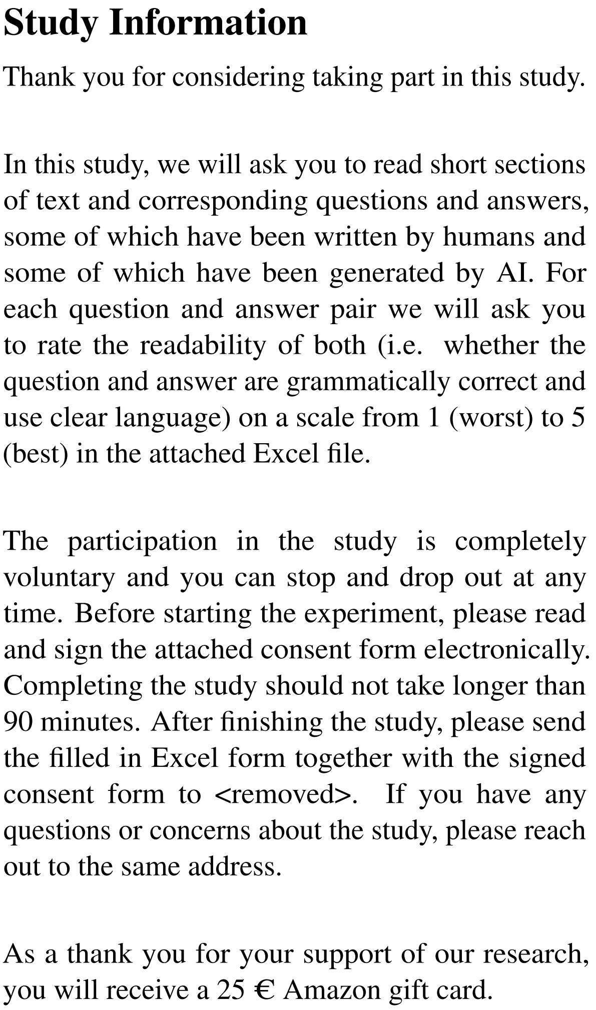
Emiel Van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. 2023. How reproducible is best-worst scaling for human evaluation? a reproduction of ‘data-to-text generation with macro planning’. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 75–88.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Reproduction Instructions



Study Information

Thank you for considering taking part in this study.

In this study, we will ask you to read short sections of text and corresponding questions and answers, some of which have been written by humans and some of which have been generated by AI. For each question and answer pair we will ask you to rate the readability of both (i.e. whether the question and answer are grammatically correct and use clear language) on a scale from 1 (worst) to 5 (best) in the attached Excel file.

The participation in the study is completely voluntary and you can stop and drop out at any time. Before starting the experiment, please read and sign the attached consent form electronically. Completing the study should not take longer than 90 minutes. After finishing the study, please send the filled in Excel form together with the signed consent form to <removed>. If you have any questions or concerns about the study, please reach out to the same address.

As a thank you for your support of our research, you will receive a 25 € Amazon gift card.