Once Upon a Replication: It is Humans' Turn to Evaluate AI's Understanding of Children's Stories for QA Generation

Andra-Maria Florescu[1,*], Marius Micluța-Câmpeanu[1,*], Liviu P. Dinu[2]
[1]Interdisciplinary School of Doctoral Studies
[2]Faculty of Mathematics and Computer Science
University of Bucharest, Romania
{andra-maria.florescu,marius.micluta-campeanu}@s.unibuc.ro
ldinu@fmi.unibuc.ro

Abstract

The following paper presents the outcomes of a collaborative experiment on human evaluation from the ReproNLP 2024 shared task, track B, part of the ReproHum project. For this paper, we evaluated a QAG (question-answer generation) system centered on English children's storybooks that was presented in a previous research, by using human evaluators for the study. The system generated relevant QA (Question-Answer) pairs based on a dataset with storybooks for early education (kindergarten up to middle school) called FairytaleQA. In the framework of the ReproHum project, we first outline the previous paper and the reproduction strategy that has been decided upon. The complete setup of the first human evaluation is then described, along with the modifications required to replicate it. We also add other relevant related works on this subject. In conclusion, we juxtapose the replication outcomes with those documented in the cited publication. Additionally, we explore the general features of this endeavor as well as its shortcomings.

Keywords: ReproNLP, QAG system, FairytaleQA, Reproduction

1. Introduction

In the field of Natural Language Processing (NLP), reproducibility is crucial for democratizing and understanding better the mechanisms of the field (Storks et al., 2023). Nevertheless, there are still issues and no widely recognized, appropriate procedure for carrying out replications of earlier research. A major factor that continues to make reproduction challenging to accomplish is the evaluations conducted by both humans and computers (Belz et al., 2023a; Pineau et al., 2021). A wide range of variables, such as imprecise data, incorrect experiments, and disagreement among the human assessors, make human evaluation one of the major obstacles to accurately replicating previous research (Thomson et al., 2024; Belz et al., 2023b; Popović, 2021).

The present study focuses on human evaluation of prior NLP research and is part of the ReproNLP 2024 shared task on Reproducibility of Evaluations in NLP (Belz and Thomson, 2024), namely on the Track B task associated with the ReproHum project. The plan was to undertake the study again and try to replicate the findings. For this project, we replicated an NLP study in which we evaluated a QAG (Question and Answer Generation) system conducted by Yao et al. (2022) and compared the outcomes of this replication to the original findings. To our knowledge, the present study represents the first attempt of replicating these results.

Section 2 focuses on presenting the original study, QAG systems, the common strategy for evaluating QAG systems, and related studies presented in section 3. Section 4 explains how the NLP evaluation was replicated. It begins by outlining the contents of the selected paper and then goes into depth about every aspect of the evaluation that was replicated. Section 5 presents and discusses the findings from the replicated evaluation concerning the original study. Lastly, Section 7 offers some closing thoughts and future works related to this project.

In alignment with open science principles, we make available all code and data employed in this investigation for the benefit of the scientific community and future research endeavors[1].

2. QAG system

The original study, "It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books" (Yao et al., 2022) examined the question-answer pair generation task (QAG) in the context of early childhood education (kindergarten through middle school). The original study implemented a QA-pair generation pipeline, which, as observed in human and automated evaluation, effectively supported the objective of automatically generating high-quality questions and

*These authors contributed equally to this work.

[1]https://github.com/mcmarius/ReproNLP-2024

Proceedings of the HumEval @LREC-COLING 2024, pages 106–113
Turin, Italy. May 21st, 2024.

answers at scale. This was achieved by leveraging a newly-constructed expert-annotated QA dataset built upon child-oriented fairy tale storybooks (FairytaleQA, Xu et al., 2022).

Five non-native English speakers were selected for the study's human evaluation in order to assess the QAG system's capacity to produce high-quality Question-Answer pairs. Furthermore, in addition to the QAG system, the ground truth and the PAQ system (Lewis et al., 2021) were evaluated by human evaluators who were blind to the system they were assessing. Ground truth QA pairs were written by human annotators of the FairytaleQA dataset, while the PAQ system consists of two components: a passage selection model and an answer extraction model. The PAQ system is supported by the PAQ dataset, a corpus of 65 million QA pairs that were automatically generated.

Each QA pair's model of origin was unknown to the participants. Using a five-point Likert scale, the participant was asked to rate the QA pairs along three dimensions:

- 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.

According to the original paper, seven novels were chosen at random, and then ten sections from those seven books were chosen randomly (for a total of seventy-QA pairings). To ensure coding consistency, each participant was asked to rate these identical 70 QA pairs. After this step, ten books (five from the test and five from the validation divides) were then chosen at random, and four sections from each book were chosen randomly once again. There are, on average, nine QA-pairs in each section (three for each model). Two coders were assigned at random to each section. Overall, each coder coded four volumes, or sixteen sections and about 140 QA-pairs. A total of 722 QA-pairs were scored. T-tests were also used in the original study to determine if the difference between models is statistically significant.

3. Related works

The goal of automatic question generation, or QG, is to extract meaningful questions and desired responses from text sections. In the past, rule-based or neural models were employed; new developments have made neural models more popular. These models—sequence-to-sequence models in particular—are capable of creating excellent questions by utilizing prior knowledge and anticipated responses. However, their value is limited because they frequently require another system in order to obtain the correct answers. Additionally, there aren't many publicly available data sets for QG systems that may generate both questions and answers. An alternative method concentrates on teaching QG models solely on context, enabling them to produce distinct question kinds for varying text lengths. State-of-the-art (SOTA) systems use pre-trained language models (PLMs) like Google T5 and GPT-3 for instructional neural question generation[2]. These pre-trained models on large-scale text corpora allow for the creation of questions with zero effort and no further training. GPT models have the ability to generate educational questions, as a recent study has shown (Bulathwela et al., 2023). Therefore, since the first study was published, numerous additional studies have been carried out on QA AI systems, some of which have been especially focused on issues related to education. Ushio et al. (2023) released `AutoQG`, a multilingual web-based quality assurance system, and `lmqg`, a Python module for QA generation, fine-tuning, and evaluation. This user-friendly code might be advantageous to both developers and end users who require customized models or fine-grained controls for development.

4. Reproduction of the human evaluation

We were given a document by the task organizers with more details regarding the human evaluation procedure to help with our reproduction experiment, even though we weren't able to interact with the authors directly. The ReproNLP 2024 project team corresponded with the authors to obtain this information prior to initiating the ReproNLP 2024 shared task. The document covers details regarding the task configuration given to the human evaluators, including the methods used. In addition, we fulfilled the experiment's requirements by filling out a Human Evaluation Datasheet (HEDS, Shimorina and Belz, 2022)[3]. This form consists of details on the assignment, the evaluators' characteristics, and the gathered annotation information from them.

For this paper, we attempted to follow the original procedures given by the prior study and the extra information obtained as closely as possible in order to replicate the human evaluation. Five human subjects were employed in the initial study to rate each of the three QA systems. The system outputs were selected randomly by the authors of the

[2]This was the state-of-the-art when the initial study was published in 2022.

[3]The HEDS document is available here: https://github.com/nlp-heds/repronlp2024

original paper, so we used the same set of examples. To ensure our reproducibility, we aimed for the same number of evaluators. We first posted our requirements in an announcement sent to the student representative who shared it on their communication channels. The only inclusion criterion was for the student to be at least in 3rd year. As a result, five undergraduate male BSc and BEng students who are not native English speakers, but speak it fluently answered our request. Gender was not a criterion used for selection, other students could have participated as well. They received no monetary compensation for their involvement, which was instead taken into account as part of their educational curriculum practice hours for which they needed academic credits.

To enable our evaluators to score the narrative sections and QA pairings from the three systems—Ground truth, PAQ, and the original paper's system (called "Ours")—blindly on a scale of 1 to 5 for readability and relevance for questions and answers, the students were each given an Excel file with 7 columns: id (internal), section text, question text, answer text and 3 columns corresponding to each of the 3 ratings they have to provide. The students annotated at their own pace from their place of choosing, but they were instructed that they had a deadline of one week to complete the Excel sheets with their evaluations. Each student read the sections, questions and answers not knowing what QA system they were assessing as well as not being aware of the other annotators in order to have an unbiased evaluation. They reported that on average their annotation took up to 5 hours. We had no pre-coding training or detailed coding guidelines as indicated in the document received from the ReproNLP task organizers regarding the original study.

5. Findings

In this section, we present the main outcomes of our study, along with qualitative and quantitative analyses that strive to explain the disparities from the preceding research. We show our approach for determining the inter-annotator agreement, along with the differences in statistical significance of the results between the two experiments. We also include a quantified reproducibility assessment (QRA) (Belz et al., 2022).

5.1. Inter-annotator agreement

First, we attempt to compute the inter-coder reliability score (Krippendorff's alpha, Krippendorff, 2011) for both experiments. If we assume that the authors limited the scope of their pre-coding stage to ground truth examples, we are able to partially confirm the claim from the original paper that shows a high level of agreement between all annotators. We determined this agreement based on the available data that we received.

Given that each sample is coded only by two raters, we compute the overall agreement by averaging the individual agreements between each pair of raters with common examples[4]. There are 3 systems to be evaluated, 5 annotator pairs and 3 evaluation dimensions, leading to a total of 45 individual pairs. Out of these 45 pairs, there are 12 instances with acceptable alpha values over 0.67, resulting in an overall Krippendorff's alpha score of 0.43 for the initial experiment. We note that 9 out of those 12 instances are for ground truth examples. Only 3 out of 30 pairs show an alpha value over 0.67 for their system and the PAQ system, all of them for answer relevancy. A breakdown of these values by system and evaluation dimension is shown in Table 1.

We rely on the only Krippendorff's alpha Python package that provides support for ordinal levels of measurement (Castro, 2017), since we need to distinguish between low and high score differences. Upon some investigations, we find that this implementation does not take into account situations with perfect or almost perfect agreement and only one conforming value (e.g. most ratings have a common score of 5, but there is no example where both labelers give a score of 4), leading to spurious values that erroneously entail no agreement when analyzing subsets of the original data. These issues could be (partially) mitigated by determining the inter-coder reliability score on a larger sample where other identical values are more likely to appear. If this is not feasible, researchers should at least properly specify the software packages used[5].

Next, we calculate Krippendorff's alpha for our evaluators, acknowledging that no pre-coding practice took place due to missing coding guidelines. From the total of 45 pairs, there are 8 instances with an alpha value over 0.67. The automated systems obtain alpha values over 0.67 in 5 out of the same 30 instances, with two additional alpha scores over 0.65. These results show a slightly higher agreement for the automated systems compared to the original paper. Again, this agreement is observed mostly for answer relevancy, with two instances being for question relevancy. The overall

[4] If we compute the overall agreement directly, the score underestimates the real agreement due to the sparsity of data.

[5] Additionally, we computed the agreement scores using the R package `irr`. While the case of identical values is handled correctly by `irr`, it does not allow specifying the domain of possible values. The results are identical, but we note that both implementations show no agreement in other corner cases of almost perfect agreement.

	Ours	PAQ Baseline	Groundtruth	
Readability	0.25	0.24	0.94	
Question relevancy	0.38	0.33	0.35	
Answer relevancy	0.45	0.44	0.45	
Overall agreement				**0.43**

Table 1: Inter-annotator agreement measured by Krippendorff's alpha for the original paper for each system and evaluation dimension. Each cell shows an average of 5 pairs of coders.

	Ours	PAQ Baseline	Groundtruth	
Readability	-0.06	0.05	-0.13	
Question relevancy	0.35	0.46	0.42	
Answer relevancy	0.51	0.50	0.46	
Overall agreement				**0.27**

Table 2: Inter-annotator agreement measured by Krippendorff's alpha for the replication experiment for each system and evaluation dimension. Each cell shows an average of 5 pairs of coders.

Krippendorff's alpha score is 0.27 for the replication study. Table 2 offers a systematic overview of the agreement by system and evaluation dimension, revealing marginally better agreements for relevancy scores than the initial paper.

5.2. Statistical significance of the results

After receiving the annotated files, we perform a sanity check for each evaluator by counting the number of samples for which ratings have any absolute difference in contrast with the original labels. This step reveals that one of our annotators assigned substantially inferior grades, prompting us to omit these biased scores from the statistical tests.

For completeness, the initial results are displayed in Table 3, while the same results ignoring the biased labeler are shown in Table 4. We first validate the assumptions of *t-tests* through Shapiro-Wilk tests, confirming that the scores for automated systems ("Ours" and PAQ) are normally distributed. As expected, the ground truth distribution is skewed since most ratings are 4 or above.

The proposed model ("Ours", avg = 4.52, s.d. = 0.79) significantly outperforms PAQ for the *Readability* dimension (avg = 4.13, s.d. = 1.04, $t(382) = 4.07$, $p < 0.01$), albeit not as satisfactory as the ground-truth (avg = 4.67, s.d. = 0.55, $t(392) = -2.64$, $p < 0.01$).

In terms of the *Question relevancy* dimension, ground-truth (avg = 4.77, s.d. = 0.71) surpasses the proposed model (avg = 3.92, s.d. 1.37), which in turn is significantly better than the PAQ baseline (avg = 3.39, s.d. 1.60, $t(382) = 2.05$, $p < 0.05$)[6].

Finally, for the *Answer relevancy* dimension, the ground-truth obtains by far the best ratings (avg = 4.58, s.d. = 0.92). Unlike the original paper, our experiments do not display a considerable distinction between the proposed model (avg = 3.39, s.d. = 1.60) and the PAQ model (avg = 3.42, s.d. 1.62, $t(382) = -0.22$, $p = .82$), though we confirm that this observation is not statistically significant.

We include the human evaluation results of the original paper in Table 5 to aid comparisons with our replication experiments, although we do not repeat the *t-tests* results from the initial paper here since we are able to confirm both the exact numbers and the results of the statistical tests.

5.3. Quantified Reproducibility Assessment

Quantified reproducibility assessment (QRA), introduced by Belz et al. (2022), aims to provide an impartial framework for determining the extent of reproducibility across different tasks and types of evaluation. This is achieved by computing a single score known as precision for each value of interest, which enables comparability between studies.

In accordance with the guidelines provided by the task organizers, we use the unbiased coefficient of variation (CV^*) for small sample sizes (Belz, 2022) as a measure for precision. This score is determined independently for each of the three dimensions, as shown in Tables 6, 7 and 8, along with Pearson's correlation coefficient r and Spearman's correlation coefficient ρ.

As mentioned in the previous section, the results

[6] If we include the biased labeler, the difference between "Ours" and PAQ is no longer significant: $t(478) = 1.79$, $p = 0.07$. This would be the only place where the biased labeler meaningfully affects the statistical tests.

	Ours		PAQ Baseline		Groundtruth	
	M	SD	M	SD	M	SD
Readability	4.52	0.75	4.17	1.22	4.71	0.52
Question Relevancy	3.83	1.30	3.61	1.35	4.71	0.73
Answer Relevancy	3.20	1.56	3.20	1.57	4.46	1.03

Table 3: Human evaluation results of the reproduction study

	Ours		PAQ Baseline		Groundtruth	
	M	SD	M	SD	M	SD
Readability	4.52	0.79	4.13	1.04	4.67	0.55
Question Relevancy	3.92	1.37	3.62	1.45	4.77	0.71
Answer Relevancy	3.39	1.60	3.42	1.62	4.58	0.92

Table 4: Human evaluation results of the reproduction study excluding the biased labeler

	Ours		PAQ Baseline		Groundtruth	
	M	SD	M	SD	M	SD
Readability	4.71	0.70	4.08	1.13	4.95	0.28
Question Relevancy	4.39	1.15	4.18	1.22	4.92	0.33
Answer Relevancy	3.99	1.51	3.90	1.62	4.83	0.57

Table 5: Human evaluation results of the original paper

System	Orig	Repl	CV*	r	ρ
Ours	4.71	4.52	4.10		
PAQ	4.08	4.17	2.18	0.99	1
GT	4.95	4.71	4.95		

Table 6: Precision metrics for the readability dimension showing the degree of reproducibility. CV* is computed using $n = 2$. Pearson's correlation and Spearman's correlation are denoted by r and ρ respectively. *Orig* indicates results from the initial experiment by Yao et al. (2022). *Repl* refers to replicated scores. *GT* represents ground truth scores.

System	Orig	Repl	CV*	r	ρ
Ours	4.39	3.83	13.58		
PAQ	4.18	3.61	14.59	0.99	1
GT	4.92	4.71	4.35		

Table 7: Precision metrics for the question relevancy dimension showing the degree of reproducibility. We use the conventions from Table 6.

are not statistically significant if we include the problematic labeler. We obtain $r = 0.99, p = 0.056$ and $\rho = 1, p = 0.0$ for *Readability*, $r = 0.99, p = 0.056$ and $\rho = 1, p = 0.0$ for *Question relevancy*, and $r = 0.99, p = 0.03$ with $\rho = 0.86, p = 0.33$ for *Answer relevancy*.

System	Orig	Repl	CV*	r	ρ
Ours	3.99	3.20	21.90		
PAQ	3.90	3.20	19.66	0.99	0.87
GT	4.83	4.46	7.94		

Table 8: Precision metrics for the answer relevancy dimension showing the degree of reproducibility. We use the conventions from Table 6.

QRA results display low CV* values for readability, while relevancy scores showcase a substantial gap between QAG systems and ground truth, prompting the need for precise coding instructions.

5.4. Reproduction results

In order to talk about the differences between the original study and ours, we had online meetings with the five human evaluators, focusing on examples with conflicting scores when compared to the original labels. Together, we examined the Excel documents that they had annotated, and we asked them to justify the scores they had given for readability, question relevancy, and response relevancy. It appears that the majority of our annotators based their remarks on their personal interpretations of the

texts, language proficiency, comprehension, and instances where errors resulting from a failure to pay attention to the texts affected the scoring.

We conduct a quantitative analysis stemming from the findings recorded as part of the discussions with our annotators. We synthesize our interpretations for labeling discrepancies in Table 9, noting that we consider examples as belonging exclusively to one error category to better observe systematic mistakes.

One persistent problem with the PAQ system was that it would repeatedly replace the named entities in the questions with "val", for instance: "What did val give to the dead man?". Out of 240 samples (120 unique questions), "val" appears in 106 of them. This caused our annotators to assign readability scores that were lower than those in the original study for 19 occurrences.

The responses were incomplete in 11 cases, like the following:

> Question: What did the man give his son?
>
> Answer: falcon.

The complete answer here would have been "gun, dog and falcon."

Since we made a methodological error by only providing generic scoring instructions without specific restrictions or details, one labeler relied on simple heuristics and primarily assigned low values for single-word responses even if they were otherwise relevant and readable. These account for 19 QA mislabeled pairs. Still, we argue that an educational QA system should seek to include connectives and proper punctuation marks as part of their answers. For example:

> Question: What weapon did val use to cut down a tree?
>
> Answer: axe

Similarly, high ratings were provided for questions or answers that resemble verbatim portions of the story, despite the lack of meaning or importance. The previous article's QA system ("Ours") tends to generate such copy-paste fragments from the story sections, in some cases being illegible: "the son - in - law ate nothing though his wife 's parents , with kind words and friendly gestures , kept urging him to help himself".

It should be noted that the initial labeling is also prone to human errors. These situations are infrequent, but they represent more than 10% of divergent ratings. The following QA pair has received marks of 4 and 5 for readability in the previous experiment, despite the nonsensical nature:

> Question: What kind of garlic would a cow be good for?
>
> Answer: garlic.

Our annotators disregarded the possibility that some questions and answers were pertinent and might have been inferred from the sections, thus focusing only on explicit textual matches. We also noticed that regarding readability, the scoring was influenced by the QA pair, although there were instances in which the question was readable, while the answer was not, thus influencing the rating. We suggest that readability should also be scored independently for question and answer.

6. Discussion

We first reiterate the contributions of the original work and the extent of our replication before discussing the consequences of our findings. In order to enhance the accuracy of automated question-answer generation systems in educational contexts—specifically children's storybooks—Yao et al. (2022) introduced an innovative technique. They demonstrated the superiority of their approach over current state-of-the-art models on two datasets, PAQ and 2-step baseline systems, as well as ground truth (human educational experts), using a combination of automatic and human evaluation approaches. Our replication was restricted to the human evaluation task described in their study, which assessed the produced questions and answers for readability, relevance of the questions, and relevance of the responses in relation to the story's segments.

As stated by Arvan and Parde (2023) in their reproducibility article from ReproNLP 2023, there was insufficient information in the research paper to replicate the original human evaluation in its entirety. This is likely due to the fact that, in the current research climate, NLP research is too focused on novelty and format compliance, rather than providing a clear explanation of the methodologies used.

Given that human evaluation is carried out by humans, personality, culture, expertise, and comprehension can all lead to significant biases (Amidei et al., 2018). This is why, in order to minimize errors made by humans as much as possible, explicit standards for the evaluations are required to obtain less ambiguous interpretations of the annotators. For example, regarding this study and others that focus on QA systems, evaluation dimensions such as readability should be assessed separately for questions and answers.

7. Conclusions

All in all, we managed to replicate the original study. However, our annotators considered answer length, which affected their low scoring because of a methodological error on our part. Furthermore, some questions might have been legitimate even

Error category	Count
Readability	14
Incomplete question	1
Irrelevant question	14
Incomplete answer	11
Right answer in another context	2
Wrong answer	10
Short answer	19
Perception, comprehension	14
“val” mentioned	19
Methodological errors	6
Human error (reproduction study)	13
Human error (original study)	16
Total	**139**

Table 9: Quantitative analysis of divergent answers with an absolute score difference of 3 or 4 in at least one dimension

though the answer was only inferred rather than explicitly stated in the text; nonetheless, our labelers focused solely on information that was explicitly mentioned in the text, which led to another lower score than the original study.

As mentioned in several studies centered on human evaluation (Amidei et al., 2018), one's personality, language knowledge as well as own writing style influence drastically the scoring. This was present in our study as well. After discussing with our annotators, we noticed that in most cases, their personality and English understanding knowledge influenced the scoring. It is clear from comparing the two studies that there are some differences between the replicated and original results. The differences between the first study and ours are more likely to be the result of methodological errors because we were not given access to the entire set of original guidelines, as well as human errors made by the annotators in terms of comprehending and interpreting the assignments.

8. Limitations

Unlike the original research, we only employed BSc and BEng students for this study, and they came from a different field than the original work for the human evaluation. We took the most of the scant information available because we lacked the precise guidelines from the previous research hence having some methodological errors for the evaluations.

9. Acknowledgments

We would like to thank Craig Thomson for providing us with suggestions and materials on how to approach the replication. We also are grateful for our evaluators who answered our call and participated in the study.

10. Bibliographical References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mohammad Arvan and Natalie Parde. 2023. Human evaluation reproduction report for data-to-text generation with macro planning. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 89–96, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

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

Anya Belz, Maja Popović, Ehud Reiter, Craig Thomson, and João Sedoc, editors. 2023a. *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria.

Anya Belz and Craig Thomson. 2024. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes,

Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023b. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable educational question generation with pre-trained language models. In *Artificial Intelligence in Education*, pages 327–339, Cham. Springer Nature Switzerland.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

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

Shane Storks, Keunwoo Yu, Ziqiao Ma, and Joyce Chai. 2023. NLP reproducibility for all: Understanding experiences of beginners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10199–10219, Toronto, Canada. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*, pages 1–11.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. A practical toolkit for multilingual question and answer generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 86–94, Toronto, Canada. Association for Computational Linguistics.

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

11. Language Resource References

Lewis, Patrick and Wu, Yuxiang and Liu, Linqing and Minervini, Pasquale and Küttler, Heinrich and Piktus, Aleksandra and Stenetorp, Pontus and Riedel, Sebastian. 2021. *PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them*. MIT Press.

Xu, Ying and Wang, Dakuo and Yu, Mo and Ritchie, Daniel and Yao, Bingsheng and Wu, Tongshuang and Zhang, Zheng and Li, Toby and Bradford, Nora and Sun, Branda and Hoang, Tran and Sang, Yisi and Hou, Yufang and Ma, Xiaojuan and Yang, Diyi and Peng, Nanyun and Yu, Zhou and Warschauer, Mark. 2022. *Fantastic Questions and Where to Find Them: FairytaleQA – An Authentic Dataset for Narrative Comprehension*. Association for Computational Linguistics.