ReproNLP 2026: A Third Replication of the Human Evaluation of a QAG System for Children's Storybooks

Marcel Mroczek
Samsung R&D Institute Poland
mroczek.marcel@gmail.com

Chiara Albarello
Samsung R&D Institute Poland
c.albarello@samsung.com

Paul-Emmanuel Floch
Samsung R&D Institute Poland
p.floch@partner.samsung.com

Maciej Gawinecki*
Samsung R&D Institute Poland
m.gawinecki@samsung.com

Abstract

Reproducibility of human evaluations in Natural Language Processing remains a critical open challenge. This paper presents a third independent replication of the human evaluation from Yao et al. (2022), which assessed an automated Question-Answer Generation (QAG) system for children's storybooks against a baseline system and human-authored ground truth, across three criteria — Readability, Question Relevance, and Answer Relevance — using five NLP-literate annotators. Our replication confirms the main findings of the original study: the QAG system outperforms the baseline on Readability and Question Relevance, and Ground Truth ranks highest across all criteria. System rankings are preserved across all three criteria, with the exception of a statistically non-significant difference in Answer Relevance. This holds true despite a severe drop in inter-annotator agreement for Readability. We further document several methodological concerns, some unreported in prior replications, including data quality issues and evaluation design limitations identified during our pilot study.

1 Introduction

Reproducibility is a cornerstone of rigorous science. The ReproHum project and its associated ReproNLP shared tasks (Belz and Thomson, 2023, 2024; Belz et al., 2025, 2026) address this challenge by building a framework to assess and improve the reliability of human evaluations in NLP.

Inspired by this initiative, we present a third independent reproduction of the human evaluation originally conducted by Yao et al. (2022). The original study introduced an automated question-answer generation (QAG) system for children's storybooks and compared it against baseline systems.

Two prior reproductions of this study have been conducted. Florescu et al. (2024) evaluated all three original criteria using undergraduate students, while Braun (2025) reproduced only the Readability criterion using NLP experts. Both found the original numerical results largely replicable.

Building upon this foundation, our work introduces the following main contributions:

- A comprehensive third reproduction covering all three evaluation criteria. We utilized NLP-literate annotators, aligning our participant demographics more closely with the original study than prior reproductions.
- A quantitative reproducibility assessment using the QRA3 framework (Belz and Thomson, 2026). We present a full three-study assessment covering all criteria, and provide an extended four-study assessment restricted to Readability (Appendix B), including an analysis of CV* value stability.
- A critical analysis of potential methodological weaknesses and data artifacts within the original evaluation design, paired with concrete suggestions for future work.

To support transparency and facilitate future research, all data and code are publicly available[1].

2 Related Work

Our work follows the ReproHum and ReproNLP framework (Belz and Thomson, 2023, 2024; Belz et al., 2025, 2026), addressing the systemic challenges of reproducing human evaluations in NLP (Belz et al., 2021; Howcroft et al., 2020).

Prior multi-lab studies report ubiquitous flaws in original experiments, including missing data, coding errors, and reporting inaccuracies (Belz et al.,

* Corresponding author.

[1] https://github.com/SamsungLabs/ReproNLP2026

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1082–1093

2023; Thomson et al., 2024). This lack of standardization highlights the need for rigorous protocols, such as pilot studies and the Human Evaluation Datasheet (HEDS) (Belz and Thomson, 2025).

As seen in other findings related to the ReproNLP shared tasks, system-level rankings are far more reproducible than individual numeric scores, and independent reproductions of the same experiment can yield contradictory results (Belz and Thomson, 2024). For the specific study by Yao et al. (2022) replicated here, previous ReproNLP iterations confirmed all pairwise system rankings despite medium system-level reproducibility scores (Florescu et al., 2024; Braun, 2025). This variability underscores the need for a third independent replication.

3 Original Paper

Yao et al. (2022) propose a Question-Answer Generation (**QAG**) system to assess children's reading comprehension. They deployed it within StoryBuddy, an interactive app where a chatbot guides children (ages 3–8) through storybooks, asking sequentially generated questions. The underlying QAG system generates QA pairs targeting narrative elements (e.g., characters, settings, causal relationships) from text passages. It was tested on the FairytaleQA dataset (Xu et al., 2022), comprising ∼10,000 expert-annotated QA pairs from 278 children's storybooks covering kindergarten to eighth grade level.

The QAG system is compared against two baselines: the **2-Step Baseline** (Shakeri et al., 2020), and **PAQ** (Lewis et al., 2021). A **Ground Truth** (**GT**) set of expert-annotated pairs from FairytaleQA serves as the reference. Evaluation was conducted both automatically and through human judgment. For human evaluation, the 2-Step Baseline was dropped as it was outperformed by PAQ in automatic evaluation; therefore, the annotators rated only the QAG system, PAQ, and Ground Truth.

3.1 Human Evaluation

In the original work, five participants were recruited: four faculty members (professors or researchers) and one graduate student. Two were education experts and three were NLP experts; all were non-native English speakers.

Participants rated each QA pair on a five-point Likert scale across three criteria taken from Yao et al. (2022): **Readability** (the QA pair uses readable English grammar and vocabulary), **Question Relevance** (the question is relevant to the storybook section), and **Answer Relevance** (the answer is relevant to the question). Ratings were collected via an Excel sheet with six columns: `section`, `question`, `answer`, `readability`, `relevancy_Q`, and `relevancy_A`. The column headers differed slightly from the paper definitions: readability was described as "grammarly correct and clear language" and answer relevance as "Answer can correctly answer the Q". The exact instructions were not available, but the original authors reported giving short one-to-two paragraph instructions.

Before the main evaluation, all five participants completed a pre-annotation consistency check in which they rated the same 70 QA pairs from 10 narrative sections across 7 books, reaching Krippendorff's α between 0.73 and 0.79 across criteria.

The results are presented in Table 1. The QAG system outperforms PAQ on all three criteria; the difference is significant for Readability ($t(477) = 7.33$, $p < 0.01$) and Question Relevance ($t(477) = 1.98$, $p < 0.05$), but not for Answer Relevance ($t(477) = 0.58$, $p = 0.56$). Ground Truth pairs score highest on all three criteria. No overall α was reported for the main evaluation, but Florescu et al. (2024) later computed $\alpha = 0.43$ post-hoc, a considerably lower value than in the pre-annotation phase.

4 Previous Reproductions

Two prior reproductions of Yao et al. (2022)'s work were conducted as part of the ReproNLP shared tasks. Both of them followed the ReproHum guidelines and provided minimal instructions. The results of both reproductions are shown alongside the original findings in Table 1.

4.1 Florescu et al. (2024)

The evaluation was conducted by five undergraduate students (non-native but fluent English speakers with no NLP or education background) who received academic credit. The annotation process lacked a pre-annotation phase, and the organizer provided explicit guidelines only for Readability, relying on Excel column headers for the remaining criteria.

Consequently, inter-annotator agreement was poor (Krippendorff's $\alpha = 0.27$, versus 0.43 computed post-hoc on the original data). The authors identified one annotator who had assigned substan-

tially lower scores than the others and excluded them from their reported tests. On this reduced pool, the reproduction successfully confirmed the original rankings ($GT > QAG > PAQ$) for both Readability ($t(382) = 4.07$, $p < 0.01$) and Question Relevance ($t(382) = 2.05$, $p < 0.05$), while the QAG–PAQ difference for Answer Relevance was not significant ($t(382) = -0.22$, $p = 0.82$). When the "biased" annotator is retained, the QAG–PAQ difference for the Question Relevance is no longer significant ($t(478) = 1.79$, $p = 0.07$).

Despite this instability at the item level and the fact that absolute scores were consistently lower in the replication—especially for Question and Answer Relevance—the system-level hierarchy proved remarkably resilient. The Pearson correlation between the original and replicated system-level mean scores remained extremely high at $r = 0.99$ across all dimensions.

Key sources of the overall divergence in agreement and absolute scores include the lack of robust annotation guidelines, the absence of a pre-annotation phase, the presence of the biased annotator, and the "val" placeholder issue (Section 5.1).

4.2 Braun (2025)

Five NLP experts (two researchers and three PhD students) evaluated only Readability; new instructions were drafted in place of the unrecoverable originals and kept intentionally brief to avoid biasing the results. The reproduction successfully confirmed the original rankings (GT > QAG > PAQ) for Readability ($p < 0.01$). Absolute scores were lower than the original across all systems, though inter-annotator agreement remained remarkably close to the original baseline, yielding a Krippendorff's α of 0.41 computed at the nominal level of measurement[2] (compared to the original 0.43). The author suggests that the absolute score drop may be due to the NLP expertise within the participant pool, or due to heightened expectations for AI-generated text quality following the public availability of large language models.

5 Pilot Study

Following the recommendations of Thomson et al. (2024), we conducted a preliminary pilot study prior to the main evaluation. The pilot was carried out by the authors and a colleague, using a sample of 32 QA pairs drawn from 4 sequential story narrative sections of the same story from the original dataset, and following the original instructions. We identified several issues in both the distributed data and the evaluation design. Some of these issues compromise **reproducibility** (our ability to faithfully replicate the setup), while others raise **validity** concerns regarding what the evaluation actually measures. Following ReproNLP guidelines, we deliberately preserved all of these flaws in our main study to remain completely faithful to the original experiment.

5.1 Flaws in the Distributed Data

We identified several systematic errors in the provided QA pairs, each capable of artificially skewing evaluation scores:

"val" placeholder in PAQ outputs: 44% of PAQ outputs contain a "val" placeholder in place of a named entity (e.g., "What did val give to the dead man?"). This actively deflates Readability scores for the baseline system. We analyze the impact of this issue in Section 8.2.

Lowercased text: Both input data and generated outputs appear entirely in lowercase. This creates an unnatural reading experience and can be penalized as a grammatical error, further lowering Readability scores.

Punctuation artifacts: Answers from the human-authored Ground Truth consistently include an extra space before end-of-sentence punctuation (e.g., "answer␣."). Because AI-generated responses follow standard formatting, this artifact creates a surface-level shortcut that allows annotators to subconsciously distinguish the human source.

Content redundancy: The dataset contains duplicate QA pairs and poorly segmented narrative sections. This creates a risk of "anchoring bias," where annotators normalize their ratings based on repeated encounters with similar texts.

Distributional imbalance: Although the study design specified three QA pairs per narrative section, the actual distribution is erratic, with some sections containing up to 13 pairs. This introduces unexpected annotator fatigue and

[2]We recomputed this value using the ordinal level of measurement, obtaining $\alpha = 0.471$, which we report in Table 2 for consistency with our own methodology.

Table 1: Human evaluation results from the original study, two prior reproductions, and this study. Mean = Mean score on a 1–5 Likert scale, SD = Standard Deviation. Braun (2025) reproduced Readability only.

	QAG		PAQ		GT	
Study	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Readability						
Yao et al. (2022)	4.71	0.70	4.08	1.13	4.95	0.28
Florescu et al. (2024)	4.52	0.75	4.17	1.22	4.71	0.52
Braun (2025)	3.85	1.35	3.14	1.43	4.38	0.96
This study	3.88	1.05	3.56	1.21	4.05	0.84
Question Relevance						
Yao et al. (2022)	4.39	1.15	4.18	1.22	4.92	0.33
Florescu et al. (2024)	3.83	1.30	3.61	1.35	4.71	0.73
This study	3.95	1.42	3.64	1.57	4.64	0.97
Answer Relevance						
Yao et al. (2022)	3.99	1.51	3.90	1.62	4.83	0.57
Florescu et al. (2024)	3.20	1.56	3.20	1.57	4.46	1.03
This study	3.69	1.53	3.78	1.59	4.57	0.97

disproportionately weights certain narrative segments in the final analysis.

5.2 Flaws in the Evaluation Design

Beyond data quality, we identified four structural concerns regarding the design of the evaluation itself:

Ambiguous rating scale: The original guidelines left substantial room for subjective interpretation and failed to address recurring edge cases. Future evaluations must provide anchor examples for each scale point and require a joint calibration session prior to the main annotation.

Target audience mismatch: Although Xu et al. (2022) reported filtering out complex stories, the source texts still contain archaic vocabulary (e.g., "bade her come"). Assessing this using the criterion of "readable English grammar" without explicitly anchoring the standard to the intended audience (school-age children) forces adult annotators to apply arbitrary, inconsistent standards.

Spreadsheet layout bias: The annotation sheet presented all source text, questions, answers, and rating columns side-by-side. This layout introduces two issues. First, it allows inconsistent rating strategies (scanning specific columns vs. reading the entire row). Second, a poorly phrased question might receive a higher Readability rating once its meaning is retroactively clarified by the adjacent answer. This invalidates the premise of the StoryBuddy application, where a child encounters and answers questions in isolation.

Sparse annotator pairing: To maintain strict fidelity, we utilized the exact annotator pairing grid from the original experiment. Under this sparse design, each QA pair is rated by exactly two annotators and each participant overlaps with only two of their four peers. Because agreement coefficients assume the reliability data represents the entire pool (Artstein and Poesio, 2008), the resulting overall inter-annotator agreement (α) is hypersensitive to these specific pairings rather than reflecting the true reliability of the annotators.

6 Experimental Setup

6.1 Methodology

We employ three complementary methods to evaluate system differences and assess reproducibility:

Significance Testing: While previous works (Yao et al., 2022; Florescu et al., 2024; Braun, 2025) relied on Student's t-test, we opted for the non-parametric Mann-Whitney U test as our primary measure. Because Likert scale data is ordinal and not normally distributed, this aligns better with established NLP human evaluation practices (Iskender et al., 2021). For direct comparability with prior work, we also report the t-test results.

Reproducibility Assessment: We utilize the Quantified Reproducibility Assessment (QRA3) framework (Belz and Thomson,

2026) via its command-line tool[3]. We assess Type I results (system-level numerical scores) using CV*, Type II results (sets of scores) using Pearson's r, Spearman's ρ, and Kendall's W, and Type IV results (system rankings) using the proportion of identical pairwise ranks (P).

Inter-Annotator Agreement (IAA): We report Krippendorff's α at the *ordinal* level of measurement to assess agreement among the annotators, consistent with the ordinal nature of Likert scale data. When recomputing α values from prior studies, we apply this same ordinal assumption.

Throughout this paper, all results for all studies are computed on the full annotator pool. This ensures a consistent and fair basis for comparison across all studies. Details about design and setup of the test are recorded in the form of HEDS[4] to simplify comparison with related past and future works.

6.2 Participants

Five employees of our company participated in the evaluation as part of their regular work duties; no additional compensation was provided. All five are researchers with backgrounds in NLP and linguistics, and are non-native but highly fluent English speakers.

This participant profile positions our study between prior setups. Because our annotators possess NLP expertise, our pool aligns more closely with the original study than Florescu et al. (2024), who utilized undergraduate students. However, unlike the original study, our pool does not include specialized education experts.

6.3 Task Design and Instructions

Participants evaluated the QA pairs across the three original criteria: Readability, Question Relevance, and Answer Relevance. Each pair was rated on a 5-point Likert scale (1 = worst, 5 = best). Because the original instructions were unrecoverable, we extended the prompt drafted by Braun (2025) to cover all three metrics using the definitions from the original study's spreadsheet columns (see Appendix A for full instructions).

Annotators conducted the task via an evaluation sheet that presented each story section alongside its associated QA pairs, with separate columns for the question, answer, and the three rating scores. To prevent bias, the experimental setup was strictly controlled: participants were blind to the source of each QA pair (GT, PAQ, or QAG), and the order of the pairs within each narrative section was randomized. As noted in Section 5.2, we retained the sparse pairing grid from the original experiment. Finally, consistent with both prior reproductions, we did not conduct a pre-annotation calibration phase prior to the main evaluation.

[3] https://github.com/DCU-NLG/qra

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

Table 2: Inter-annotator agreement (Krippendorff's α, ordinal level) across all studies. Braun (2025) evaluated Readability exclusively. Metrics for Florescu et al. (2024) and Braun (2025) were recomputed from their published item-level. Originally reported value by Braun (2025) was 0.41 (nominal level).

Study	Read.	Q. Rel.	A. Rel.
Yao et al. (2022)	0.449	0.390	0.451
Florescu et al. (2024)	0.060	0.482	0.575
Braun (2025)	0.471	—	—
This study	0.089	0.397	0.419

7 Results

7.1 Inter-Annotator Agreement

Krippendorff's α (Table 2) for Question and Answer Relevance (0.397 and 0.419) aligns with the original study. However, Readability α values vary substantially across studies and in our study collapses to near-chance levels ($\alpha = 0.089$).

7.2 Pairwise System Comparisons

Table 3 shows pairwise comparisons using both the Mann-Whitney U test and Student's t-test. Overall, our results largely confirm the findings of Yao et al. (2022).

The GT set achieves significantly higher scores than both generated systems across almost all criteria. For Question Relevance and Answer Relevance, GT significantly outperforms both QAG and PAQ according to both the Mann-Whitney U test and Student's t-test ($p < 0.01$ in all cases). The sole exception is the comparison between GT and QAG on Readability, where neither the Mann-Whitney U test ($p = 0.09$) nor the Student's t-test ($p = 0.06$) indicate a significant difference. Both tests consistently suggest there is no reliable annotator preference for GT over QAG on this criterion. However, this non-significant result must be interpreted with caution given the near-zero inter-

Table 3: Pairwise comparison of evaluation criteria across GT, QAG, and PAQ using the Mann-Whitney U test and Student's t-test. Bold indicates significance at $\alpha = 0.05$ according to the Mann-Whitney U test.

Criterion	Comparison	Mean 1	Mean 2	U	MW p	$t(df)$	t-test p
Readability	GT > QAG	4.05	3.88	30970.5	0.09	$t(480) = 1.91$	0.06
	QAG > PAQ	**3.88**	**3.56**	**32966.5**	**0.00**	$\mathbf{t(478) = 3.11}$	**0.00**
	GT > PAQ	**4.05**	**3.56**	**35235.0**	**0.00**	$\mathbf{t(480) = 5.14}$	**0.00**
Question Rel.	GT > QAG	**4.64**	**3.95**	**36550.0**	**0.00**	$\mathbf{t(480) = 6.39}$	**0.00**
	QAG > PAQ	**3.95**	**3.64**	**31601.5**	**0.02**	$\mathbf{t(478) = 2.32}$	**0.02**
	GT > PAQ	**4.64**	**3.64**	**38868.5**	**0.00**	$\mathbf{t(480) = 8.66}$	**0.00**
Answer Rel.	GT > QAG	**4.57**	**3.69**	**38400.0**	**0.00**	$\mathbf{t(480) = 7.49}$	**0.00**
	QAG > PAQ	3.69	3.78	27331.0	0.85	$t(478) = -0.64$	0.52
	GT > PAQ	**4.57**	**3.78**	**36318.0**	**0.00**	$\mathbf{t(480) = 6.52}$	**0.00**

annotator agreement for Readability ($\alpha = 0.089$, see Table 2), which severely limits the statistical power of any test applied to this criterion.

When comparing the two automated systems, QAG significantly outperforms PAQ on both Readability ($p < 0.01$) and Question Relevance ($p = 0.02$). For Answer Relevance, the performance difference between the two systems does not reach statistical significance ($p = 0.85$), which aligns consistently with the original study's findings. Notably, PAQ achieves a marginally higher mean than QAG (3.78 vs. 3.69) for Answer Relevance, technically reversing the original ordering. However, because this difference is fractional and far from significant, it should not be interpreted as evidence of PAQ's superiority.

7.3 Comparison Across Reproductions

Comparison with prior work (Tables 1, 4) reveals three trends. First, absolute scores have deflated while variance has increased, particularly for Readability. Second, despite score drops, statistical effects remain stable: the QAG > PAQ advantage in Readability and Question Relevance persists across all studies. Third, the system hierarchy (GT > QAG > PAQ) is remarkably resilient, with the only deviation being the non-significant swap between QAG and PAQ in Answer Relevance. This suggests that while absolute human judgments are subjective, relative system rankings are robust.

7.4 Quantified Reproducibility Assessment

Table 5 provides a pairwise reproducibility assessment ($n = 2$) comparing the original findings with each subsequent reproduction. Table 6 presents the QRA3 reproducibility assessment across all three criteria for the $n = 3$ fully comparable studies. Because Braun (2025) evaluated only Readability, a direct four-study comparison is structurally limited to that criterion alone; the corresponding $n = 4$ assessment is provided in Table 7 in Appendix B.

7.4.1 Type I: How close are the scores?

While absolute scores deflated across all reproductions, Type I reproducibility (measured via CV*) allows us to quantify how tightly these diverging scores cluster. Looking across the three fully comparable studies (Table 6), Question Relevance achieves a "good" QC-level reproducibility, whereas Answer Relevance and Readability fall into the "medium" reproducibility tier (CV* ≈ 13-14).

Readability results show a distinct split between studies. When including Braun (2025) ($n = 4$; Appendix B), the QC-level CV* rises to 15.29. This reflects a divide between Florescu et al. (2024), who mirrored the original high scores, and the severe drops recorded by both Braun (2025) and our study. This suggest that score reproducibility depends more on the background of the annotators than the source of the text. In fact, in our $n = 3$ baseline, GT was the least stable metric, confirming that a human-authored "Ground Truth" is as much a moving target as the systems it evaluates.

Finally, pairwise fidelity results (Table 5) demonstrate that no single replication was perfect at reproducing all criteria across the board, suggesting that reproducibility is dimension-specific rather than study-wide. While Florescu et al. (2024) (recomputed) achieved the lowest aggregated study-level divergence ($CV^* = 11.11$), followed by our study (11.58), a closer look at individual dimensions reveals a fragmented landscape. Our study achieved the highest fidelity in functional dimensions, recording the lowest CV^* values for Question Relevance (10.57) and Answer Relevance

Table 4: Comparison of Student's t-test results for QAG vs. PAQ across all studies. Braun (2025) reproduced Readability only. †Results as originally reported by Florescu et al. (2024), excluding one biased annotator. Unmarked Florescu et al. (2024) rows retain the full annotator pool for consistency with all other studies.

Criterion	Study	$t(df)$	p
Readability	Yao et al. (2022)	$t(477) = 7.33$	$< .01$
	Florescu et al. (2024)	$t(478) = 4.43$	$< .01$
	Florescu et al. (2024)†	$t(382) = 4.07$	$< .01$
	Braun (2025)	$t(477) = 5.56$	$< .01$
	This study	$t(478) = 3.11$	$< .01$
Question Relevancy	Yao et al. (2022)	$t(477) = 1.98$	$< .05$
	Florescu et al. (2024)	$t(478) = 1.79$	.07
	Florescu et al. (2024)†	$t(382) = 2.05$	$< .05$
	This study	$t(478) = 2.32$	$< .05$
Answer Relevancy	Yao et al. (2022)	$t(477) = 0.58$	.56
	Florescu et al. (2024)	$t(478) = 0.00$	1.00
	Florescu et al. (2024)†	$t(382) = -0.22$	.82
	This study	$t(478) = -0.64$	.52

Table 5: Pairwise CV* ($n = 2$) reproducibility assessment comparing each replication to the original study by Yao et al. (2022). Rows labeled 'original' denote values from Florescu et al. (2024) calculated on the raw 1–5 scale without 0-shift.

Criterion	Study	GT	QAG	PAQ	QC Level
Readability	Florescu et al. (2024) (original)	4.95	4.10	2.18	3.74
	Florescu et al. (2024) (recomputed)	4.98	4.18	2.29	3.82
	Braun (2025)	12.37	20.86	28.65	20.63
	This study	20.46	20.04	14.67	18.39
Question Rel.	Florescu et al. (2024) (original)	4.35	13.58	14.59	10.84
	Florescu et al. (2024) (recomputed)	4.38	14.32	15.66	11.46
	This study	5.89	11.04	14.76	10.57
Answer Rel.	Florescu et al. (2024) (original)	7.94	21.90	19.66	16.50
	Florescu et al. (2024) (recomputed)	8.07	24.22	21.84	18.04
	This study	5.59	8.40	3.36	5.78
Aggregated Study-Level Fidelity (Mean CV^*)					**Overall**
Florescu et al. (2024) (original)		–	–	–	10.36
Florescu et al. (2024) (recomputed)		–	–	–	11.11
Braun (2025)		–	–	–	20.63
This study		–	–	–	11.58

(5.78). The latter represents the most stable reproduction across all compared independent studies, with a particularly low divergence in the PAQ subset (3.36). In contrast, the Readability dimension consistently yielded higher divergence across all independent replications (Braun (2025): 20.63; this study: 18.39).

7.4.2 Type II: Do the systems correlate?

Type II reproducibility measures how well the relative distances and rankings between systems are preserved, even if absolute scores change. Overall, our analysis indicates that these relative relationships are highly stable.

As shown in Table 6, correlation is near-perfect for both Readability and Question Relevance across all studies, demonstrating almost total agreement in both linear correlation (Pearson's r) and rank-based agreement (Kendall's W). Answer Relevance also demonstrates a very high Pearson correlation, though its rank-based agreement is slightly lower ($W = 0.750$). This drop in Kendall's W is a mathematical artifact of the marginal, statistically insignificant tie between QAG and PAQ discussed in earlier sections, which scrambled the strict ranking without meaningfully changing the relative distance between the systems.

It is important to contrast this high system-level stability with item-level variance. When computing correlations across all 361 individually matched QA pairs, the values drop significantly: Readability ($r = 0.480$), Question Relevance ($r = 0.625$), and

Table 6: QRA3 reproducibility assessment ($n = 3$): Yao et al. (2022), Florescu et al. (2024), and this study. Braun (2025) is excluded as they evaluated Readability only. AR = Answer Relevance, QR = Question Relevance, R = Readability; n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 3$)		
				System level	QC level	Study level
Type I	AR	GT	(mean) CV*↓	6.00	13.08	12.13
		PAQ		16.16		
		QAG		17.09		
	QR	GT	(mean) CV*↓	4.39	9.38	
		PAQ		12.83		
		QAG		10.91		
	R	GT	(mean) CV*↓	14.68	13.93	
		PAQ		12.60		
		QAG		14.51		
Type II	AR	all	mean r ↑	n/a	0.992	n/a
		all	mean ρ ↑	n/a	0.744	
		all	W↑	n/a	0.750	
	QR	all	mean r ↑	n/a	0.996	
		all	mean ρ ↑	n/a	1.000	
		all	W↑	n/a	1.000	
	R	all	mean r ↑	n/a	0.998	
		all	mean ρ ↑	n/a	1.000	
		all	W↑	n/a	1.000	
Type IV	AR	all	P ↑	n/a	0.667	0.889
	QR	all	P ↑	n/a	1.000	
	R	all	P ↑	n/a	1.000	

Answer Relevance ($r = 0.732$). While all remain statistically significant ($p < 0.001$), this sharp contrast illustrates a well-documented phenomenon in NLP evaluation: individual human judgments are highly subjective and difficult to reproduce, whereas aggregated system-level outcomes remain remarkably robust (Belz and Thomson, 2024).

7.4.3 Type IV: Do we reach the same conclusions?

Type IV reproducibility assesses whether independent studies arrive at the same overarching conclusions regarding system rankings. In this regard, the human evaluation demonstrates exceptional stability.

For both Readability and Question Relevance, the established system hierarchy is flawlessly preserved across all available studies ($P = 1.000$). This holds for the three fully comparable studies (Table 6), and the ranking remains perfect when Braun (2025) is included for Readability ($P = 1.000$; see Appendix B).

The only divergence occurs in Answer Relevance, where two out of the three pairwise rankings are preserved ($P = 0.667$). As established earlier, this drop is caused entirely by the marginal, statistically non-significant rank swap between QAG and PAQ in our replication.

At the study level, 8 out of 9 total pairwise comparisons are successfully reproduced ($P = 0.889$). This confirms that despite the significant drops in absolute scores and inter-annotator agreement discussed previously, the substantive conclusions of the original human evaluation are highly reproducible.

8 Discussion

8.1 Drop in Readability IAA

The most striking divergence is our near-zero Readability IAA ($\alpha = 0.089$), compared to $\alpha = 0.449$ in the original study. This sharp drop stems from two compounding factors: an unfortunate pairing that matched strict annotators against lenient ones, and the absence of pre-annotation calibration.

First, our annotators naturally divided into three lenient annotators (averaging > 4.0) and two strict annotators (averaging ~ 3.15). Due to the randomized sparse grid (Section 5.2), the two strict annotators never evaluated the same questions. Consequently, every time a strict annotator rated a text, their score was directly compared against a lenient annotator's, mathematically maximizing the disagreement.

Post-evaluation feedback revealed a deep divergence in how these two groups interpreted the uncalibrated Readability criteria. The more lenient annotators prioritized naturalness and immediate understandability. Conversely, the stricter annotators assigned lower scores by penalizing vocabu-

lary imprecision, grammatical issues, and the formatting flaws described in Section 5.1 (such as lowercased text and punctuation artifacts). This distinction also dictated their approach to the "val" issue: the lenient group awarded higher scores as long as the identity of "val" was obvious, while the strict group heavily penalized the text despite its clarity. The original study likely avoided this problem because their annotators were either naturally more uniform, favorably paired in the grid, or successfully aligned during their pre-annotation phase.

Second, this near-zero agreement is not merely a grid anomaly, but a systemic symptom of missing calibration guidelines. Florescu et al. (2024) experienced an identical collapse: their annotators yielded a Readability α of 0.06. They explicitly attributed this to a lack of specific scoring rules, noting that one annotator invented a strict personal heuristic to penalize single-word answers. When annotators lack access to alignment protocols, they are forced to rely on arbitrary personal rules, inevitably driving agreement down to chance levels.

Importantly, despite this total collapse in item-level agreement, the overall system rankings were perfectly preserved ($P = 1.000$).

8.2 Impact of the "val" Artifact

Because the "val" placeholder artifact (identified during the pilot study) systematically deflated the baseline's Readability scores, we performed an additional analysis excluding all sections where PAQ exhibited this problem. This exclusion removed 23 of the 40 narrative sections, leaving 334 out of 722 (46%) graded QA pairs.

Crucially, when applying this clean subset to both the original study and our replication, QAG no longer outperforms PAQ on any criterion. For the original study data, both the Readability advantage ($p = 0.086$) and the Question Relevance advantage ($p = 0.290$) become statistically non-significant.

This exact same pattern holds in our replication: $QAG > PAQ$ on Readability ($p = 0.643$) and Question Relevance ($p = 0.886$), as well as $GT > PAQ$ on Readability ($p = 0.390$), all lose significance. Furthermore, in our replication, PAQ actually scores higher than QAG on Answer Relevance (PAQ mean 3.95 vs. QAG mean 3.53; $p = 0.008$ for $PAQ > QAG$).

While restricting the data in this way is inherently biased towards PAQ, it exposes a critical flaw: the original claim of QAG superiority is likely an illusion caused by a corrupted baseline. This raises serious concerns regarding the validity of any subsequent replication performed on the uncleaned dataset.

9 Conclusion

We conducted a third independent reproduction of the human evaluation by Yao et al. (2022). We successfully replicated the original numerical rankings ($P = 0.889$ across all three criteria, $n = 3$; $P = 1.000$ for Readability across all four studies, $n = 4$), but our analysis raises concerns about whether these rankings constitute robust evidence of true differences in system quality.

Most critically, the "val" placeholder artifact systematically deflated baseline scores. Excluding affected sections eliminates QAG's advantage in both the original study and our replication; in our data, PAQ even significantly outperforms QAG on Answer Relevance ($p = 0.008$). The claimed QAG superiority is therefore likely an artifact of a compromised baseline rather than genuine system quality. Coupled with the methodological flaws detailed in Section 5 — including data artifacts, ambiguous rating scales, and an uncontrolled annotation interface — the reliability of this evaluation remains severely limited, regardless of replication fidelity.

In summary, the original study is reproducible in the narrow sense that independent teams obtain similar rankings, but those rankings appear to be influenced by experimental artifacts rather than true system quality. Future work should replicate this evaluation on clean data, with explicit artifact handling, audience-appropriate criteria, and a controlled annotation interface.

Acknowledgements

We would like to thank Katarzyna Basista for her participation in the pilot study. We are also grateful to the anonymous annotators who participated in the human evaluation for their time and effort. Furthermore, we would like to thank the ReproNLP shared task organizers, especially Craig Thomson, for facilitating this reproduction study. We would also like to thank the original authors for providing additional information to the ReproNLP team.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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, Online. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anya Belz and Craig Thomson. 2024. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia. ELRA aRAnd ICCL.

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

Anya Belz and Craig Thomson. 2026. Quantified reproducibility assessment for four 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.

Anya Belz, Craig Thomson, Javier González Corbelle, and Malo Ruelle. 2025. The 2025 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 1002–1016, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, and 23 others. 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.

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

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.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, 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.

Siamak Shakeri, Cicero Nogueira dos Santos, Henghui 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. 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.

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.

A Annotation Instructions

The following instructions were provided to all participants prior to the evaluation:

> 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:
>
> - **Readability**: of both: i.e. whether the question and answer are grammatically correct and use clear language
> - **Question Relevance**: Question is relevant to the section
> - **Answer Relevance**: Answer can correctly answer the question
>
> on a scale from 1 (worst) to 5 (best).

B QRA3 Reproducibility Assessment for Readability ($n = 4$)

Table 7 presents the QRA3 reproducibility assessment restricted to Readability, computed over all four studies including Braun (2025). Because Braun (2025) evaluated only this criterion, this table is structurally limited to Readability and cannot be extended to the full set of quality criteria.

Table 7: QRA3 reproducibility assessment for Readability ($n = 4$): Yao et al. (2022), Florescu et al. (2024), Braun (2025), and this study. R = 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	R	GT PAQ QAG	(mean) CV*↓	12.01 19.09 14.78	15.29	15.29
Type II	R	all all all	mean r ↑ mean ρ ↑ W ↑	n/a n/a n/a	0.995 1.000 1.000	n/a
Type IV	R	all	P ↑	n/a	1.000	1.000