ReproHum #0033-05: Human Evaluation of Factuality from A Multidisciplinary Perspective

Andra-Maria Florescu[1,2*] **Marius Micluța-Câmpeanu**[1,2*]
Ștefana-Arina Tăbușcă[1,2*] **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, stefana.tabusca}@s.unibuc.ro
ldinu@fmi.unibuc.ro

Abstract

The following paper is a joint contribution for the 2025 ReproNLP shared task, part of the ReproHum project. We focused on reproducing the human evaluation based on one criterion, namely, factuality of Scientific Automated Generated Systems from August et al. (2022). In accordance to the ReproHum guidelines, we followed the original study as closely as possible, with two human raters who coded 300 ratings each. Moreover, we had an additional study on two subsets of the dataset based on domain (medicine and physics) in which we employed expert annotators. Our reproduction of the factuality assessment found similar overall rates of factual inaccuracies across models. However, variability and weak agreement with the original model rankings suggest challenges in reliably reproducing results, especially in such cases when results are close.

1 Introduction

Although Natural Language Processing (NLP) represents a field that strongly focuses on computational approaches and the use of automatic evaluation, human evaluation remains an important practice for assessing NLP systems. Automated metrics may be scalable and robust, however, they often fail to capture nuances of natural language, such as emotional tone, cohesion, and coherence, in the same manner as humans, as stated by Celikyilmaz et al. (2021) and first noticed by (Papineni et al., 2002).

Since humans are also prone to errors, there still is a need for proper guidelines of human evaluation (Thomson et al., 2024). Belz and Reiter (2006) reviewed several evaluation methods in NLP and showcased the role of human evaluators in assessing aspects that computational approaches struggle to take into account. However, reproducibility of human evaluation and proper guidelines remain a complex and difficult-to-achieve task (Belz et al., 2023).

This paper is a contribution to the ReproNLP 2025 shared task (Belz et al., 2025), which is part of the ReproHum project[1], a multi-lab cooperative project aiming to test the reproducibility of human evaluations through large-scale reproduction.

Our paper focuses on evaluating the factuality of artificially generated scientific definitions from August et al. (2022). The original work was reproduced before by van Miltenburg et al. (2024), and (Li et al., 2024), who concentrated on evaluating the fluency of the generated definitions.

The paper begins with introducing the research and then discussing about related studies. Next up we present our reproduction steps followed by our additional experiment, concluding with results, participant feedback and final remarks.

According to standard scientific procedures, we provide all data and code used to help with further research in this area. We also follow the project's coordination team's guidelines by completing the Human Evaluation Datasheet (HEDS) (Shimorina and Belz, 2022; Belz and Thomson, 2024).[2]

2 Original Study

In the original study (August et al., 2022), human evaluation is employed on a dataset of 300 generated scientific definitions by three models considered to be the best performing: DExperts, GeDI, and an SVM reranker. The authors define two types of generated definitions that are in scope: "high-complexity" (which use more academic and technical language) and "low-complexity" (which use terms more approachable for the general public). In this sense, they compile two separate sets of data: scientific news articles designed for training of lower-complexity definitions and scientific

*Equal contribution.

[1]https://reprohum.github.io/
[2]https://github.com/nlp-heds/repronlp2025

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 583–589

journal abstracts, which are later used for high-complexity definitions training.

The first model uses the DExperts architecture introduced by Liu et al. (2021), which consists of an ensemble of three different language models: an "expert" (trained on text with desired features), an "anti-expert" (trained on text with unwanted qualities), and a base model. The difference between the logits from the first two is merged with the logits of the latter. For their study, the authors opt to continue the training of pretrained BART-large models for the expert and anti-expert models, utilizing the abstracts dataset and the news dataset. For the generation of more complex definitions, the abstracts dataset is used in the training of the expert model, while the news dataset is employed for the anti-expert model training, with the setup being reversed to generating less complex definitions.

The original work also includes the usage of the GeDi (Generative discriminators) method (Krause et al., 2021), which utilizes a class-conditioned language model trained on text with required attributes, more specifically in this context referring to the two sets of data.

Another well-performing approach is the one of reranking, which is introduced in the original paper. This method consists of producing 100 candidate definitions for each term during test time using a BART model, which are then reranked by a discriminator model trained to discern between text from scientific journals and from science news.

For the human evaluation, the authors select 50 terms from the test data at random, generating for each both low- and high-complexity definitions with all three best-performing models described earlier, which results in a set of 300 texts to be manually annotated.

They use two annotators to rate the generated definitions based on three criteria: fluency, relevancy, and factuality. Specifically for factuality, the annotators assign a binary label to each definition, indicating whether or not it contains any factually wrong information. In case of errors, they use a 1–4 Likert scale to score the extent of the inaccuracies. The paper reports a Krippendorff's alpha of 0.59 when assessing whether a definition contained factually incorrect information, showcasing a modest agreement between the human evaluators, as per the official interpretation guidelines (Krippendorff, 2019). The inter-rater agreement is maintained to almost the same extent in the case of evaluating the severity of errors (Krippendorff's alpha of 0.55).

3 Previous Work

In the reproduction by van Miltenburg et al. (2024), they closely followed the original paper (August et al., 2022) with minor changes due to missing details, looking specifically at the fluency ratings. Their results showed similar patterns, with fluency rating being significantly different among the SVM model and both GeDi and DExperts. However, inter-annotator agreement was lower (Krippendorff's alpha decreased by 0.11). Additionally, they conducted a second study where they gathered evaluations from eight additional annotators and analyzed the variability of the ratings. Also focused on fluency, Li et al. (2024) reproduced and found out that even if overall performance was lower, the relative performance of the three systems matched the original findings. The authors stated that lower agreement among annotators and their feedback suggests that ambiguity significantly affects human judgment.

4 Our reproduction

Our reproduction concentrates on the factuality criterion. Following the standard procedures for human evaluation reproducibility, by using Quantified Reproducibility Assessment (Belz, 2025), we try to track the original study as closely as possible. This entails the setup of two annotators that rate 300 definitions, first with a binary label of Yes/No as an answer to the question "Does this definition contain factually incorrect information?" and, in case of factual inaccuracies, to further provide a score on a set scale from 1 to 4 for the extent of the error, with the specification that 1 represents the lowest severity of error, while 4 is the highest.

Moreover, an additional experiment is carried out on domain-specific questions. In this sense, we define separate question sets depending on their domain, focusing on medical and physics-related questions. These term sets are then each assessed by a pair of participants with expert knowledge in the respective question set domain. The category of each term is established automatically, with subsequent human validation. This pipeline utilizes a Llama 3.2 LLM with the setup shown in Box 1.

All terms are assessed in this manner, after which a manual validation by one of the authors is performed. For the additional experiment, only the medicine (174 questions) and physics (42 questions) categories are considered in order to align with the participants' areas of expertise.

Box 1: Llama 3.2 Prompt

SYSTEM Prompt: You are a helpful assistant.

USER Prompt: What exact science is this term from? examples: medicine, geography, physics, chemistry, computer science. Respond with the name for each term, no explanations

Terms: [
{"id":<term_id>,
"term_text":"<term_text>"},
...]

After the annotation process is performed, the raters receive a feedback form regarding their participation.

4.1 Platform

As the platform used in the original experiment is unavailable for new experiments, we recreated the survey form from scratch as a hosted web application. We implemented the interface using Next.js, backed by a Redis-like database (Upstash).[3,4]

We strove to mimic the original look and feel of the interface as closely as possible with the aid of screenshots provided by the ReproHum team. Furthermore, we fixed several issues found in the initial interface, adding client-side and server-side validations and allowing participants to resume their progress in an intuitive manner. We note that adding validations and fixing critical issues is allowed under the ReproHum protocol.

The experiment instructions explicitly stated that going back or refreshing the page was not allowed, presumably due to software defects in the previous implementation, where previous answers were not retrieved in the UI, and the survey could only be completed in one iteration. While we kept the same instructions and did not document these enhancements, we noticed that, based on the server logs, one participant used a second pass to calibrate their answers.

To promote an open research environment, we make the source code of this interface publicly available with demo access to the hosted version.[5,6]

[3] https://nextjs.org/
[4] https://upstash.com/
[5] https://github.com/mcmarius/repronlp-2025-app
[6] URL: https://repronlp-2025-app.vercel.app/ Demo credentials: username: demo-user, Password: demo-password. Accounts can only be created by an admin user, no sign up is possible.

4.2 Participants

All our participants are fluent, non-native English speakers with an English language proficiency level of C1 and above according to the Common European Framework of Reference for Languages (CEFR). For the main reproduction, we employed two PhD students from the Faculty of Psychology, one male and one female. For the additional experiments, we included two medical students and two physics experts (one student and one PhD-level professional), who were tasked with the evaluation of medical and physics-related questions. In total, we had six annotators, with an average age of 25 years.

The participants were compensated with vouchers valued at 433 RON per annotator for the main study, which involved evaluating the 300 questions. Since factuality is more difficult to asses, even with the help of online resources, we estimated a total time of 6 hours by completing 10% of the questions. For the main study, the ReproHum team covered the costs, with a conversion rate of GBP to RON of 6.00928 and an hourly rate of 12 GBP $\approx$ 72 RON, according to the ReproHum procedure for calculating fair pay.[7] For the second experiment, a total of 174 questions were selected for the medicine participants, equal to a pay of 251 RON per individual, and 42 questions were selected for the physics participants, equal to a pay of 61 RON per individual.

4.3 Experiment

For the main reproduction, we followed the original experiment as closely as possible, given the available information on the original setup. This included the evaluation of the same 300 generated definitions, as well as adopting the same structure for the given instructions and examples. As described in an earlier section, the platform was also reproduced, given that access to the original environment is not possible. The definitions were labeled by the two main annotators; to be noted that, as opposed to the original study, none of the authors of this work acted as annotators, as per the standards of the ReproHum project. The instructions and examples given to the participants in the platform are presented in Boxes 2 and 3.

[7] Conversion via Oanda.com, 5 March 2025.

Box 2: Instructions

You will be given <no. of specific experiment terms> terms with their definitions and asked to rate the factual truth of the definitions.

You will first be asked whether the definitions contain any factual inaccuracies (yes or no) and then, if yes, you will be asked to rate the severity of the inaccuracies on a scale from 1 (lowest) to 4 (highest)

When you do not know whether a definition is factually inaccurate, please use an internet search to check.

Box 3: Examples

Term: Acanthoma
Definition: Acanthoma is a type of skin cancer. (inaccuracy marked in red; it is benign, not cancerous).

Term: Transformer
Definition: The Transformer is a type of cheese. (inaccuracy marked in red).
Please do not press the back button while taking this task.

4.4 Additional experiment

For the additional experiments, the instructions and platform remained the same as in the main experiment, the only change being the subsets of definitions and the domain knowledge of the participants. These evaluations targeted definitions related to medicine and physics, selected in alignment with the expertise of the annotators involved. The categorization of definitions was performed as previously described, automatically using a LLama 3.2-based classification prompt and manually validated by one of the authors. The final dataset included 174 medical and 42 physics terms. Each category was independently annotated by a domain-specific pair of raters (with dedicated user roles in the platform), enabling a more informed and reliable assessment of factual correctness in specialized contexts.

5 Results

When looking at the results from the original setup with all 300 generated definitions, the following can be stated: annotators agree poorly on whether a definition was factually incorrect (Krippendorff's α = 0.466), but display even more reduced agreement on how severe those errors are (Krippendorff's α = 0.132). Over half of the definitions (54.0 %) were flagged by both annotators as incorrect, and nearly four-fifths (78.0 %) by at least one of them.

We have computed the same values for the additional experiment, separately for each domain.

Focusing on medicine, the agreement on the yes/no decision rose to a substantial level (Krippendorff's α = 0.682), and consistency around severity improved moderately (Krippendorff's α = 0.507). Still, more than half of the medical definitions (58.6 %) were marked wrong by both annotators, and almost three-quarters (73.0 %) by at least one.

In physics, experts showed a more robust consensus on whether a definition was wrong (Krippendorff's α = 0.790), yet their views on the degree of error remained split (Krippendorff's α = 0.369). Almost all physics definitions were labeled as containing factual inaccuracies (92.9 % by both annotators, 95.2 % by at least one of them).

These patterns reveal the following observations: First, having domain experts makes it easier to agree on the presence of factual mistakes; however, domain expertise is less effective at harmonizing severity ratings—even when evaluators share deep subject knowledge. If we want reliable severity scores in future studies, we may need simpler scales or clearer examples of what each level means. This intuition is supported by the participants' feedback, which will be presented in the next section.

Study	Error prevalence		Krippendorff's α	
	Either annotator	Both annotators	Binary (Yes/No)	Severity (1–4)
Original	60.0%	40.0%	0.59	0.55
Reproduction	78.0%	54.0%	0.466	0.132

Table 1: Comparison of original and reproduced factuality results (300 definitions).

When we compare our reproduction with the original evaluation, as seen in Table 1, two patterns stand out:

1. The original study found that 60% of the definitions were flagged as incorrect by at least one annotator (and 40% by both), while our ex-

periment assessed those numbers at 78% and 54%, respectively. This suggests that even small shifts in the annotation instructions or the pool of raters can make annotators more sensitive (or harsher) in spotting factual lapses.

2. While the original experiment achieved substantial consistency on inter-rater agreement for both the "contains an error" and severity judgments, our reproduction shows that the latter in particular can become very noisy if the rubric or calibration is not tight.

These differences can attest that prevalence estimates can shift substantially across studies and that agreement on how bad an error is appears especially fragile. Future work might include more detailed anchor examples or simplified severity scales to boost reproducibility.

Following the original experiment and under the framework of Quantified Reproducibility Assessment (Belz et al., 2025), we have also computed percentages for flagged factual inaccuracies for each generation system (SVM reranker, GeDI, DExperts); to compare the obtained values, we have utilized the coefficient of variation for small samples (CV*), introduced by Belz (2022), with all results available in Table 2 and Table 3.

Model	Original %	Reproduction %	CV*
SVM reranker	16	57	111.9924
GeDI	33	51	42.7288
DExperts	67	54	21.4233

Table 2: Definitions flagged by both annotators as factually inaccurate.

Model	Original %	Reproduction %	CV*
SVM reranker	38	78	68.7590
GeDI	52	78	39.8802
DExperts	86	78	9.7269

Table 3: Definitions flagged by at least one annotator as factually inaccurate.

Across the three generation models, we can observe different patterns of reproducibility when comparing the original study's percentages of definitions flagged for factual inaccuracies to those obtained in our reproduction. For the rate at which both annotators labeled a definition as factually inaccurate, the SVM reranker rose from 16% originally to 57% in our data (mean = 36.5%, CV* $\approx$ 112), indicating extreme divergence relative to its average. GeDI showed a more moderate shift, from 33% to 51% (mean = 42.0%, CV* $\approx$ 43), while DExperts declined a few, from 67% to 54% (mean = 60.5%, CV* $\approx$ 21), suggesting that its error rate is the most stable of the three.

When we consider the rate at which at least one annotator rated a definition as factually inaccurate, the SVM reranker again exhibits high variability, rising from 38% to 78% (mean = 58.0%, CV* $\approx$ 69), while GeDI shifts from 52% to 78% (mean = 65.0%, CV* $\approx$ 40). DExperts shows the smallest proportional change, going from 86% down to 78% (mean = 82.0%, CV* $\approx$ 10), showing its relative reproducibility also in this setup.

The CV* results provide a useful ranking: DExperts' percentages remain within roughly one-fifth and one-tenth of their means, respectively, while the SVM reranker's rates vary by more than half. GeDI consistently falls between these extremes. This suggests that DExperts seems to be the most reproducibly labeled for factual inaccuracy, and the SVM reranker seems to be the least, with GeDI occupying a middle position.

Metric	Value	p	Significance ($\alpha = 0.05$)
Pearson's r	-0.327	0.78	n.s.
Spearman's ρ	-0.500	0.67	n.s.

Table 4: Correlation between the original and reproduced percentages of definitions flagged by both annotators as factually inaccurate.

We have also investigated the correlations between the original study's percentages and our own, using both Pearson's r and Spearman's ρ; the results are visible in Table 4. These values were calculated for the percentages where both annotators labeled a definition as containing factual errors. For the other setup, our reproduction percentages are identical (78%) across all three models, yielding zero variance; as a result, neither Pearson's r nor Spearman's ρ can be computed meaningfully.

Both Pearson's r (–0.327) and Spearman's ρ (–0.500) for the "flagged by both annotators" condition fail to reach statistical significance. This means that we cannot reject the null hypothesis of no linear or monotonic association between the original and reproduction both-flag rates. The apparent inverse relationship could easily arise by chance, given the available observations.

6 Participant feedback

After completing the feedback form, it seems that our main evaluators reported that on average, they spent about 3 hours completing the annotations. They both needed to use the internet on some occasions for their ratings regarding definitions related to biology and chemistry.

One noteworthy aspect that counted in their annotation process was their academic background. This was also seen in the additional experiment participants, with one rater stating that "As I am studying medicine, I know the importance of details, so information should be complete, very clear and precise when it comes to medical terms". They all had difficulties grading incomplete definitions, as well as those with slight inaccuracies.

When they had doubts and rated in the middle of the scale, they justified their ratings for incomplete or vague information, incorrect or imprecise terminology, oversimplification, or definitions that were unclear or failed to fully define the term.

All annotators stated that their capacity to rate factuality would have been enhanced if they had a better understanding of specific terminology and concepts, clearer grading examples for different levels of inaccuracy, and more detailed instructions on completeness and coherence.

7 Conclusion

As numerous studies have shown (Florescu et al., 2024), human evaluation still remains important for properly evaluating technological development. Our findings suggest that employing domain-specific experts and providing proper annotation guidelines represent crucial factors for accurate automated systems. However, both automated and human evaluations in the NLP field have drawbacks, hence the need for hybrid automated-human evaluation systems. Especially when it comes to human evaluation of generated scientific definitions, only experts in such domains should be employed.

While human evaluation of factuality may come off as an objective task, it actually relies heavily on subjective interpretation, human judgment comprising a certain degree of creativity and divergent thinking (a thought process used for generating creative ideas through exploring multiple possible solutions) as it was stated by Guilford (1967), particularly when evaluators draw on multidisciplinary expertise, like in our case, from Psychology, Medicine, and Physics.

While absolute agreement values differed from those originally reported, the general trends regarding which models yield more factual inaccuracies were broadly maintained. However, statistical analysis revealed low and non-significant correlations for the case where both annotators labeled definitions as factually inaccurate. These results outline both the challenges and the value of reproducibility in human evaluation setups.

Limitations

Since this is a reproducibility study, and the original paper had a small sample of only two human evaluators, according to the ReproHum guidelines, we maintained this number. Next, we could not find available annotators who had a demographic background similar to the original experiment, namely an NLP expert that is a trained annotator. Moreover, there was no background information in the original study about the second annotator. It was also stated by the guidelines of this reproduction for the authors not to partake in the annotation process.

Ethics Statement

This study adheres to the ethical guidelines for academic research established by the University of Aberdeen. The experiment design, methodology, and data collection procedures were reviewed and approved by the University of Aberdeen's Physical Sciences & Engineering Ethics Board (Decision from 05.02.2025). Prior to their participation, the annotators gave their written consent after being fully informed about the study's objectives and their role within it, that participation was voluntary and that they could withdraw at any time without facing any repercussions, and the anonymity and confidentiality of their answers, which ensured that no personally identifiable information would be revealed in publications or reports. The study conforms to international ethical norms for research involving human subjects (such as the GDPR for participants residing in the EU) and upholds the values of honesty, openness, and respect for participants' autonomy.

Acknowledgments

We would like to thank Craig Thomson for giving us useful insights for an exhaustive experiment. We are also grateful for the original authors for sharing their resources and our annotators who agreed to take part in this study.

This research is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906 and CNCS/CCCDI UEFISCDI, SiRoLa project, PN-IV-P1-PCE-2023-1701, within PNCDI IV, Romania.

References

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

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

Anya Belz. 2025. QRA++: Quantified reproducibility assessment for common types of results in natural language processing. *Preprint*, arXiv:2505.17043.

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

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.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. *Preprint*, arXiv:2006.14799.

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.

J. P. Guilford. 1967. *The Nature of Human Intelligence*. McGraw-Hill Series in Psychology. McGraw-Hill, New York. [by] J.P. Guilford.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Yiru Li, Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2024. ReproHum #0033-3: Comparable relative results with lower absolute values in a reproduction study. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 238–249, Torino, Italia. ELRA and ICCL.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

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.

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, Martijn Goudbeek, Emiel Krahmer, Chris van der Lee, Steffen Pauws, and Frédéric Tomas. 2024. ReproHum: #0033-03: How reproducible are fluency ratings of generated text? a reproduction of August et al. 2022. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 132–144, Torino, Italia. ELRA and ICCL.