ReproHum #0033-3: Comparable Relative Results with Lower Absolute Values in a Reproduction Study

Yiru Li, Huiyuan Lai, Antonio Toral, Malvina Nissim
CLCG, University of Groningen
Groningen, the Netherlands
y.li.170@student.rug.nl
{h.lai, a.toral.ruiz, m.nissim}@rug.nl

Abstract

In the context of the ReproHum project aimed at assessing the reliability of human evaluation, we replicated the human evaluation conducted in "*Generating Scientific Definitions with Controllable Complexity*" by August et al. (2022). Specifically, humans were asked to assess the fluency of automatically generated scientific definitions by three different models, with output complexity varying according to target audience. Evaluation conditions were kept as close as possible to the original study, except of necessary and minor adjustments. Our results, despite yielding lower absolute performance, show that relative performance across the three tested systems remains comparable to what was observed in the original paper. On the basis of lower inter-annotator agreement and feedback received from annotators in our experiment, we also observe that the ambiguity of the concept being evaluated may play a substantial role in human assessment.*

Keywords: human evaluation, reproducibility, ReproHum

1. Introduction

In spite of substantial advances in the automatic evaluation of Natural Language Processing (NLP), especially with the development of trained metrics highly correlating with human judgements, such as COMET (Rei et al., 2020), eventually it is the actual human evaluations that are still widely considered the most significant and reliable performance assessment. This is particularly true in language generation tasks, where the availability of a human gold standard produced in advance, as it is common practice in classification tasks, is not an option due to the large variability of valid outputs.

And yet, human evaluation, both in classification and generation tasks, is surely not free of problems. First, humans might not be great judges on a given task as they cannot tell one category from another; this has been shown for example in profiling (Flekova et al., 2016), in the detection of political leaning (De Mattei et al., 2020), and in discerning AI-generated from human-written texts (Clark et al., 2021; Freitag et al., 2021). Second, even when people might be able to yield judgements in a given task, how to perform human evaluations which are dependable, for example on conversations (Smith et al., 2022), is an open problem. Third, and most importantly, human judgements are tainted by a somewhat natural variability, which might yield idiosyncratic results that are not reproducible in subsequent studies and thus eventually not that indicative of system performance beyond a specific and single experiment. This is especially true if evaluation settings are not systematically and clearly defined and reported. Recent research has shown that due to these and related factors, reproducing human evaluation in NLP studies proves an almost impossible task (Belz et al., 2023).

This paper situates itself in this last line of research, in the context of the larger ReproHum[1] project (Belz and Thomson, 2024), which is a multi-lab cooperative project aiming to test the reproducibility of human evaluations through large-scale reproductions.

Our reproduction work follows the schedule provided by the project coordination team, and this paper reports our results accordingly. The experiment was pre-registered through the Human Evaluation Data Sheet (HEDS[2]) proposed in Shimorina and Belz (2022)'s work, providing records for possible future usage. In this report, we first summarize the original study and provide a detailed explanation of the human evaluation task we are reproducing (Section 2). Next, we introduce the adjustments we had to make to successfully replicate the experiment (Section 3). Lastly, we report our results and bring forward our observations and comments on the feasibility and meaning of this reproduction (Section 4).

*In the ReproHum project this reproduction study has code #0033-3.

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

[2] Details in Appendix A, also on https://github.com/nlp-heds/repronlp2024

Proceedings of the HumEval @LREC-COLING 2024, pages 238–249
Turin, Italy. May 21st, 2024.

2. Original Study

The original study we have reproduced is one of the human evaluation tasks described in the paper "*Generating Scientific Definitions with Controllable Complexity*" by August et al. (2022). This research proposes a new method for generating scientific definitions with controllable complexity, varying according to target audience. Several systems are trained using a newly collected dataset of scientific definitions and both automatic and human evaluations are performed on the generated outputs.

2.1. Task and Model

The core task in the research is to generate scientific definitions with controllable complexity that are appropriate answers to a "term question" in the form of "What is (are) X," where X is a scientific term or concept (August et al., 2022, Section 3). In the first part of their paper, the authors explore the performance of different models in generating scientific definitions without complexity control. Pairs of the "term questions" and corresponding definitions are then used as training/finetuning data for multiple language models. The authors have also collected additional data from scientific abstracts serving as supporting documents. Through the use of automatic metrics, they conclude that the `BART` model (Lewis et al., 2020) trained with term question concatenated with the supporting document (`BART SD`) outperforms the rest of the models they tested. Therefore, `BART SD` is used as the base generation model for all subsequent experiments.

After the selection of the base generation model, the authors explore four complexity control methods, including their proposed new method called *reranking*. A *Reranker* is composed of two parts: a `BART SD` generator that provides 100 definitions of the same scientific question, and a discriminator that was trained to distinguish scientific journals from science news. The logits of the discriminator are then used to determine the complexity of the definitions. In their work, the original authors have trained one model for each method other than *reranking*, and two models using *reranking* - one of which uses a Linear SVM Classifier as the discriminator and the other one uses the SciBERT uncased pretrained model (Beltagy et al., 2019).

Models representing the four complexity control methods are trained to provide definitions of either high complexity or low complexity and the resulting definitions are then evaluated by means of automatic metrics. See Table 1 for an example of generated definitions, directly taken from August et al. (2022).

2.2. Human Evaluation Task

The original paper includes several human evaluation tasks on the generated definitions to test the robustness of their proposed *reranking* approach. 50 terms were randomly selected from the test split as target terms. The corresponding definitions generated for these 50 terms, both with high and low complexity, by the three models that showed the best performance in the automatic evaluation task are then put through human evaluation. These three models are *Reranker* utilizing an SVM classifier as the discriminator, the Generative discriminators (GeDi) proposed by Krause et al. (2021), and the Ensemble of language models (DExperts) proposed by Liu et al. (2021). These $50 \times 2 \times 3 = 300$ definitions were then rated by human annotators.

Besides the main evaluation task that targets the complexity of definitions generated by different systems, three additional side human evaluation tasks are conducted to ensure the generations are all fluent, relevant to the questions, and factual. In our reproduction study we only focused on one of the side evaluation tasks: fluency.

Two trained annotators performed the fluency task, but the specific training they underwent was left unspecified in the original paper. One annotator was one of the authors of the original paper, and the other annotator was a research assistant.

For the evaluation, the annotators were first shown an instruction page, received instructions on the nature of this task. They were informed that they will be evaluating definitions based on their fluency, and the answers to the question "How fluent is this definition?", will need to be given on a scale from "Not at all" to "Very". They were provided with two "Not at all" fluent definitions and two "Very" fluent definitions. Then, on the annotation interface, more specific instructions were given.

In the actual annotation part, the annotators are asked to evaluate to what extent a definition is fluent. They are expected to rate the definitions using 1–4 Likert scales (1 = "Not at all" to 4 = "Very"). The two annotators rate the same 300 definitions. The definitions are presented to them in different, random orders. It is not disclosed to them whether a definition is of high complexity or low complexity, nor which system has generated this definition. In the case of "nan" (empty definition), they should rate it as "Not at all" fluent.

Figure 1 and Figure 2 show the original instruction page and an example of the original annotation interface, respectively. Note that these screenshots were taken from the interface utilized by the main evaluation task on complexity performed in the original experiment, as we do not possess the actual interface used in the fluency evaluation task. Nevertheless, according to the information provided by the original authors, the interfaces used by these

Control Method	Direction	
	Low (News)	High (Journal)
SVM-Rerank	A type of computing in which there are many computers running at the same time in different parts of the world.	In computer science, distributed computing is the process of computing on a large scale **without a single centralized data center**.
BERT-Rerank	A type of computer system in which there are more than a few computers working together.	In computer science, distributed computing is the process of computing on a large scale **without a single centralized data center**.
GeDi	Is the implementation of computer programs across multiple computers on similar hardware and/or software resources.	In computer science, **a concept that states that data must be shared across computing resources**.
DExpert	An **Internet-driven by-computing** that portion of different computers from start to finish.	In computer science and communication between-Consequently-integrates.
PPLM	**Easeless, self-organized, and often self-organizing** networked computer systems intended for the purposes of optimization.	Multi-purpose, distributed **system software** with or without a single datum storage system.

Table 1: Generated definitions from each complexity control method for the question: What is (are) distributed computing? Factually incorrect information is labeled in **bolded red**.
Note: From "*Generating scientific definitions with controllable complexity*" by August et al. (2022).

two tasks are identical except for the task-specific instructions and questions.

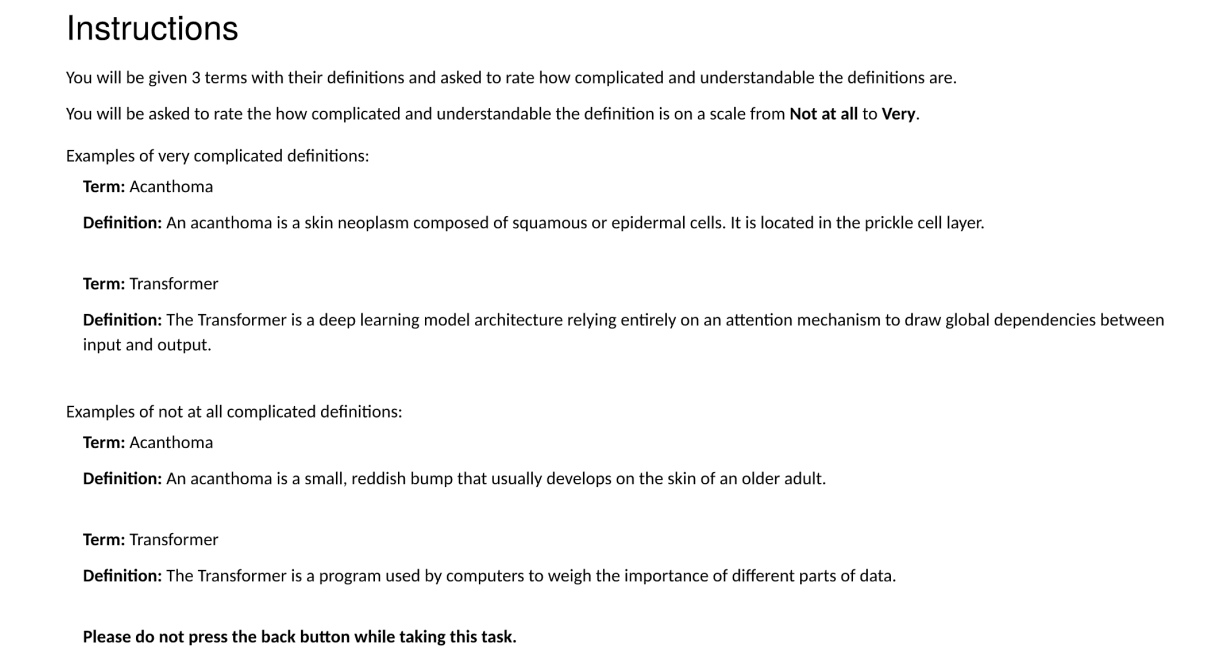
Based on the results, the authors conclude that their SVM-reranked methods can provide definitions that were rated close to "Very" fluent and are significantly more fluent compared to definitions generated by the other two systems. Further discussion of their results, also in comparison with ours, is included later in Section 5.

3. Reproduction Study

In our reproduction study, several adjustments had to be made for various reasons. None of these adjustments are related to the nature of the assessment questions - they remained identical to what was given in the original experiment.

The first adjustment we made was changing the evaluation platform from *LabintheWild* to *Qualtrics*, essentially leading to the re-writing of the evaluating interface. By the time we started reproducing the experiment, *LabintheWild* was inaccessible through its website, forcing us to use another evaluation platform instead; we chose *Qualtrics* since it could replicate the functionality and look-and-feel of the original interface, and we are familiar with it. We tried our best to keep the new interface as similar as possible to the original interface, keeping important features identical. Figure 3 shows our instructions, and Figure 4 shows an example of our new annotation interface. It is important to note that the instructions for the fluency evaluation task were not reported in the paper nor in the additional information kindly provided through email by the paper's author. As the instruction screenshots provided to us only included examples for the complexity evaluation task, we could not replicate what was included in the original instructions and had to include new examples in our guidelines.

The second adjustment we made was removing other unrelated questions from the interface, now giving our annotators one question per page instead of two questions per page. This change is due to the fact that we are only replicating the fluency evaluation task but not the relevance evaluation task which is included in the original paper alongside the fluency one. The annotators in our replication study are now answering only 300 questions in total (one question – fluency – per instance) instead of 600 in the original paper (two questions – fluency and relevance – per instance). Even though the other 300 questions/answers are irrelevant to the fluency evaluation task, the annotators' overall performance may still be affected by this difference,



Instructions

You will be given 3 terms with their definitions and asked to rate how complicated and understandable the definitions are.

You will be asked to rate the how complicated and understandable the definition is on a scale from **Not at all** to **Very**.

Examples of very complicated definitions:

Term: Acanthoma

Definition: An acanthoma is a skin neoplasm composed of squamous or epidermal cells. It is located in the prickle cell layer.

Term: Transformer

Definition: The Transformer is a deep learning model architecture relying entirely on an attention mechanism to draw global dependencies between input and output.

Examples of not at all complicated definitions:

Term: Acanthoma

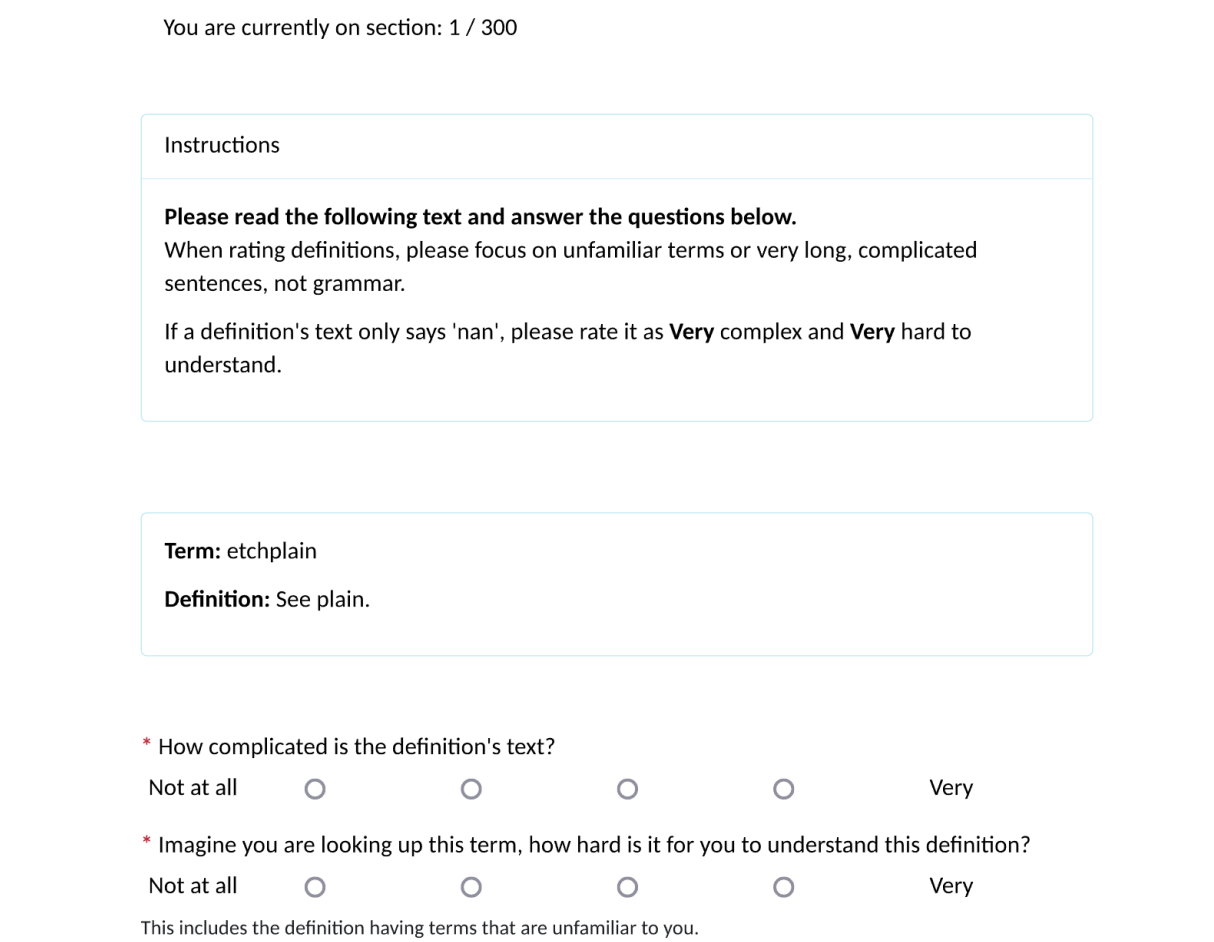
Definition: An acanthoma is a small, reddish bump that usually develops on the skin of an older adult.

Term: Transformer

Definition: The Transformer is a program used by computers to weigh the importance of different parts of data.

Please do not press the back button while taking this task.

Figure 1: A screenshot of the original instruction page.



You are currently on section: 1 / 300

Instructions

Please read the following text and answer the questions below.
When rating definitions, please focus on unfamiliar terms or very long, complicated sentences, not grammar.

If a definition's text only says 'nan', please rate it as **Very** complex and **Very** hard to understand.

Term: etchplain

Definition: See plain.

* How complicated is the definition's text?

Not at all ○ ○ ○ ○ Very

* Imagine you are looking up this term, how hard is it for you to understand this definition?

Not at all ○ ○ ○ ○ Very

This includes the definition having terms that are unfamiliar to you.

Figure 2: A screenshot of the original annotation interface.

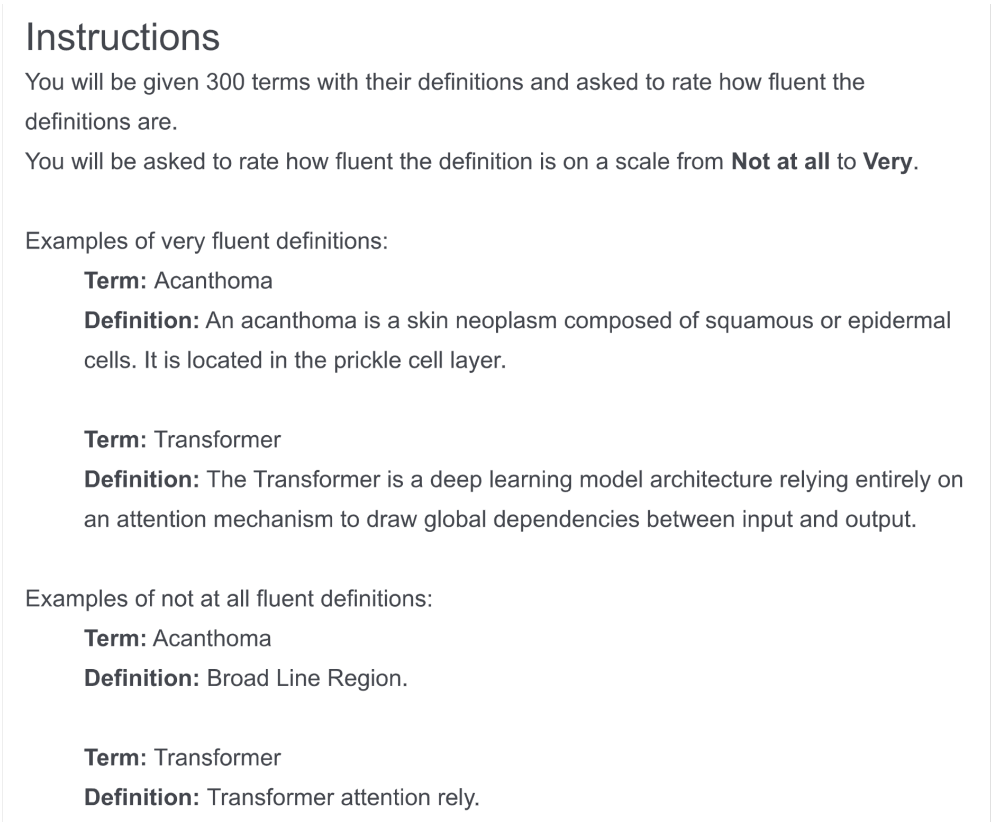


Figure 3: A screenshot of the instruction page in our replication study.

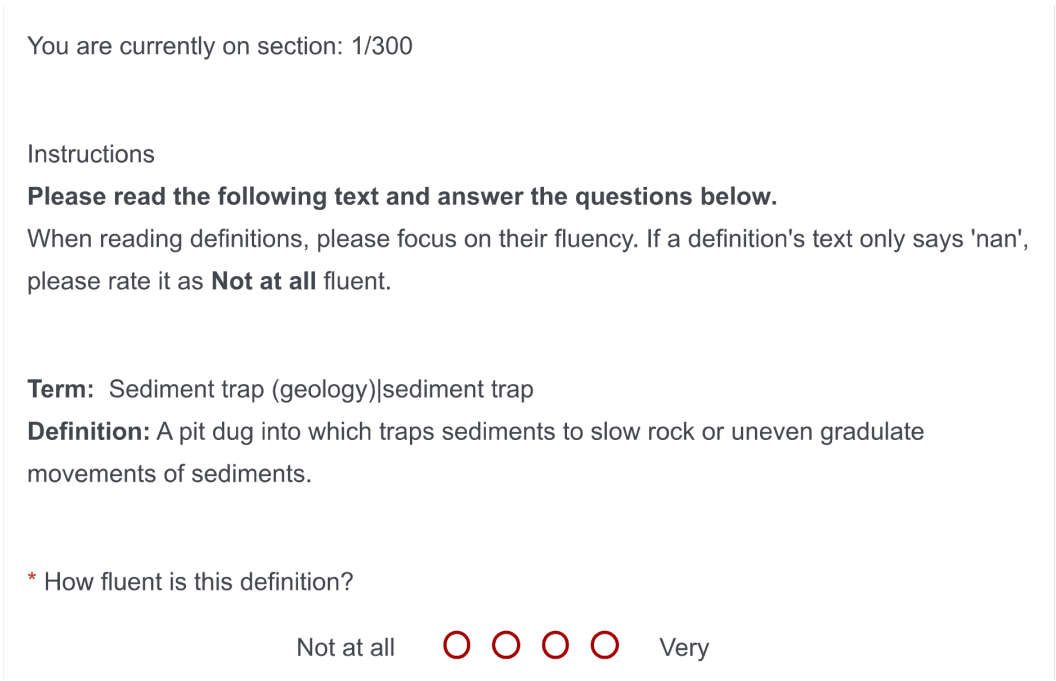


Figure 4: A screenshot of the annotation interface used in our replication study.

however to what extent is unknown.

The third adjustment we made concerned the annotators. We provided the annotators with monetary compensation and they were not related to

this reproduction paper, i.e. none of the annotators is a coauthor. The amount of monetary compensation was determined according to the minimum wage in the U.K. in December 2023. Given the assumption that the annotation task should take approximately 2.5 hours to complete, each annotator was paid 34.6 euros. As said, in the original study, one author of the paper participated in the annotation process; according to the ReproHum reproduction instructions, we have not included one of us in the evaluation task, but instead recruited one NLP PhD student and one linguistics researcher for the task, trying to match as close as possible the background of the original annotators. This adjustment may have had a larger influence on the result than the other modifications we have described: despite the original paper stating that none of their annotators have seen the generations to be evaluated before their evaluation exercise, their familiarity and association with the project could have unintentionally affected the evaluation results.

4. Results

4.1. Side-by-side Presentations

Table 2 shows a side-by-side presentation of our results and the original results.

	Original	**Replication**
Fluency (s.d) SVM-Reranker	3.71 (0.59)	3.02 (1.10)
Fluency (s.d) GeDi	3.20 (1.06)	2.40 (1.20)
Fluency (s.d) DExpert	2.33 (0.85)	1.81 (1.04)
t-test between SVM & GeDi	$t_{198} = 5.99$, $p < 0.001*$, Cohen's $d = 0.60$	$t_{198} = 4.42$, $p < 0.001*$, Cohen's $d = 0.62$
t-test between SVM & DExpert	$t_{198} = 18.85$, $p < 0.001*$, Cohen's $d = 1.88$	$t_{198} = 9.65$, $p < 0.001*$, Cohen's $d = 1.36$

Table 2: Comparison of original and reproduction results. * =p-value corrected for multiple hypothesis testing using the Bonferroni-Holm correction.

4.2. Quantified Reproducibility Assessments

According to the Common Approach of Reproduction provided by the ReproHum Team, we report three quantified reproducibility assessments below, including adjusted Coefficient of Variation (CV*), Pearson's r, and Krippendorff's α.

We report an adjusted version of the Coefficient of Variation (CV*) as mentioned in Belz et al. (2022)'s work on quantified reproducibility assessments. CV* was specifically adjusted for small samples. As the experiment utilized a Likert scale from 1-4, we shifted the values from [1,4] to [0,3] to meet the requirement of utilizing CV*. We report the two-way CV* values in Table 3.

System	**CV***
SVM-Reranker	29.09
GeDi	44.31
DExpert	48.45

Table 3: Two-way CV* between original results and replication results

We have calculated Pearson's correlation coefficient between the original results and the reproduction results as instructed. However, it is worth noting that since the sample size in our case is extremely small ($n = 3$), the coefficient (Pearson's $r = .987$) is not reliable. Spearman's ρ is not suitable for such a small sample size either.

To compare the inter-annotator agreement, we also report the Krippendorff's α of our annotations. The original study reports Krippendorff's α = 0.63, while our study reports Krippendorff's α = 0.45.

5. Discussion and Conclusion

Through the analysis of results, we observe that our results support the finding in the original paper, that the definitions produced by the SVM-reranked method are significantly more fluent compared to definitions generated by the other two systems evaluated. However, we observe that in our reproduction experiment the overall fluency is rated lower for all three systems. In one of last year's ReproHum reports on a different reproduction study (Li et al., 2023), the authors noticed the same phenomenon: The reproduction results support the comparative statements made in the original paper (e.g., one system performs better than the others) with the same overall trend, but with lower overall scores. As the fluency score of SVM-Reranker in our evaluation did not surpass 3.5 as it did in the original experiment, we could not confirm the statement suggested in the original paper that the SVM-Reranker can be rated as nearly "Very fluent".

The two-way CV* values suggest medium to low reproducibility, while the reproduced annotations on definitions generated by SVM-Reranker seem

to have a higher agreement with the original annotations, compared to annotations on other models' definitions. We have also noted a decline in Krippendorff's α. The decline of inter-annotator agreement may be attributed to the fact that the original annotators were "trained", while we did not train our annotators since the training process was not specified in the original paper. From the feedback we received from our annotators, the definition of fluency remained ambiguous to some extent, even with the examples and instructions. As a result, the different understanding of the concept of fluency may have caused our two annotators to disagree on a few questions. Lastly, the fact that one of the annotators in the original study was one of the paper's co-author might have influenced the original agreement and thus contributed to the discrepancy observed across the two studies.

Our annotators have provided valuable feedback to us, and both of them have mentioned that in some definitions, unexpected or misplaced punctuation marks or tokens occurred, which affected the overall fluency of the definition, as otherwise the definition would be considered "Very fluent". As we do not possess the original annotations, we do not know how the original authors would rate these definitions. One of the annotators also mentioned that they found the concept of fluency very ambiguous, and this may have led to confusion. From the feedback, we noticed that this annotator has also considered factuality as part of fluency, which would not happen if they were part of the original study, as we know there was an additional, separate factuality evaluation task. Yet this is an unavoidable problem since we do not know exactly what instructions have been given to the annotators, and we can only presume minimum intervention, leading to very few task instructions aside from examples. The confusion in interpreting the concept of fluency may not only lead to a lower overall score but also a lower inter-annotator agreement in the reproduction study, as the two original annotators may have reached some level of agreement on the definition of fluency, while our annotators have not.

Acknowledgments

We wish to thank the anonymous reviewers for their comments on this paper, which we have made our final revisions accordingly. And with gratitude, we thank our two annotators who helped us a lot in annotating the data and provided insightful feedback. We are also grateful to the project coordinators for their consistent help while we carried out the replication experiment.

6. Bibliography

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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

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 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. 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.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, and Malvina Nissim. 2020. Invisible to people but not to machines: Evaluation of style-aware HeadlineGeneration in absence of reliable human judgment. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6709–6717, Marseille, France. European Language Resources Association.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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.

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.

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

Yiru Li, Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Same trends, different answers: Insights from a replication study of human plausibility judgments on narrative continuations. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 190–203, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

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.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

A. HEDS Sheet

A.1. Paper and supplementary resources

Sections 1.1–1.3 record bibliographic and related information. These are straightforward and don't warrant much in-depth explanation.

1.1 Details of paper reporting the evaluation experiment

1.1.1 Link to paper reporting the evaluation experiment.
for preregistration. This is a reproduction experiment, and the original paper is on https://aclanthology.org/2022.acl-long.569/

1.1.2 Which experiment within the paper is this form being completed for?
This form is being completed for preregistration*
Title of experiment: Evaluating Fluency.
Section: 7 & 7.1.
Exact descriptions in Appendix A.7: "Annotators were given examples of very fluent and relevant definitions, and not at all fluent and relevant definitions before starting the task. For fluency, annotators were asked, 'How fluent is this definition?'"

1.2 Link to resources

1.2.1 Link(s) to website(s) providing resources used in the evaluation experiment.
https://drive.google.com/drive/folders/1qqHAl_GvxO14ZoW-XGO3PMvNZnXO9mp-?usp=share_link

1.3 Contact details

This part is hidden for anonymous purposes.

1.3.1 Details of the person completing this sheet

1.3.1.1 Name of the person completing this sheet.
Yiru Li

1.3.1.2 Affiliation of the person completing this sheet.
University of Groningen

1.3.1.3 Email address of the person completing this sheet.
y.li.170@student.rug.nl

1.3.2 Details of the contact author

1.3.2.1 Name of the contact author.
Malvina Nissim

1.3.2.2 Affiliation of the contact author.
University of Groningen

1.3.2.3 Email address of the contact author.
m.nissim@rug.nl

A.2. System Questions

Questions 2.1–2.5 record information about the system(s) (or human-authored stand-ins) whose outputs are evaluated in the Evaluation experiment that this sheet is being completed for. The input, output, and task questions in this section are closely interrelated: the value for one partially determines the others,as indicated for some combinations in Question 2.3.

2.1 What type of input do the evaluated system(s) take?
5. text: sentence

2.2 What type of output do the evaluated system(s) generate?
6. text: multiple sentences

2.3 How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2?
12. question answering

2.4 What are the input languages that are used by the system?
41. English

2.5 What are the output languages that are used by the system?
41. English

A.3. Sample of system outputs, evaluators, and experimental design

3.1 Sample of system outputs
Questions 3.1.1–3.1.3 record information about the size of the sample of outputs (or human-authored stand-ins) evaluated per system, how the sample was selected, and what its statistical power is.

3.1.1 How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment?
100

3.1.2 How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment?
1. by an automatic random process

3.1.3 Statistical power of the sample size.

3.1.3.1 What method was used to determine the statistical power of the sample size?
N/A. Follow the original experiment.

3.1.3.2 What is the statistical power of the sample size?
N/A

3.1.3.3 Where can other researchers find details of the script used?
N/A

3.2 Evaluators
Questions 3.2.1–3.2.5 record information about the evaluators participating in the experiment.

3.2.1 How many evaluators are there in this experiment?
2

3.2.2 Evaluator Type
Questions 3.2.2.1–3.2.2.5 record information about the type of evaluators participating in the experiment.

3.2.2.1 What kind of evaluators are in this experiment?
1. experts

3.2.2.2 Were the participants paid or unpaid?
1. paid (monetary compensation)

3.2.2.3 Were the participants previously known to the authors?
1. previously known to authors

3.2.2.4 Were one or more of the authors among the participants?
2. evaluators do not include any of the authors

3.2.2.5 Further details for participant type.
One participant is a non-student researcher and the other participant is a PhD student.

3.2.3 How are evaluators recruited?
The evaluators are recruited by in-person invitations.

3.2.4 What training and/or practice are evaluators given before starting on the evaluation itself?
Instructions and examples are given on the start pages of the online survey that we use to collect the results.

3.2.5 What other characteristics do the evaluators have?
The evaluators are expected to have high English profiency and have expertise in NLP.

3.3 Experimental Design
Sections 3.3.1–3.3.8 record information about the experimental design of the evaluation experiment.

3.3.1 Has the experimental design been preregistered? If yes, on which registry?
2. no

3.3.2 How are responses collected?
Qualtrics survey.

3.3.3 Quality assurance
Questions 3.3.3.1 and 3.3.3.2 record information about quality assurance.

3.3.3.1 What quality assurance methods are used to ensure evaluators and/or their responses are suitable?
7. None of the above. None quality assurance methods are included in the experiment, following what was in the original paper. We only made sure that the evaluators have expertise in NLP and English fluency.

3.3.3.2 Please describe in detail the quality assurance methods that were used.
Expertise in NLP is expected.

3.3.4 Form/Interface
Questions 3.3.4.1 and 3.4.3.2 record information about the form or user interface that was shown to participants.

3.3.4.1 Please include a link to online copies of the form/interface that was shown to participants.
To be determined.

3.3.4.2 What do evaluators see when carrying out evaluations?
The evaluators see an information letter page which inform them of this experiment and their rights, an introduction page including examples, and then the question pages with some additional instructions.

3.3.5 How free are evaluators regarding when and how quickly to carry out evaluations?
3. neither of the above. We expect the evaluators to complete the whole evaluation within a set time.

3.3.6 Are evaluators told they can ask questions about the evaluation and/or provide feedback?
1. evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation

3.3.7 What are the experimental conditions in which evaluators carry out the evaluations?
1. evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.

3.3.8 Briefly describe the (range of different) conditions in which evaluators carry out the evaluations.
N/A

A.4. Quality Criteria - Definition and Operationalisation

Questions in this section collect information about each quality criterion assessed in the single human evaluation experiment that this sheet is being completed for.

4.1 Quality Criteria
Questions 4.1.1–4.1.3 capture the aspect of quality that is assessed by a given quality criterion in terms of three orthogonal properties. They help determine whether or not the same aspect of quality is being evaluated in different evaluation experiments. The three properties characterise quality criteria in terms of (i) what type of quality is being assessed; (ii) what aspect of the system output is being assessed; and (iii) whether system outputs are assessed in their own right or with reference to some system-internal or system-external frame of reference. For full explanations see Belz et al. (2020).

4.1.1 What type of quality is assessed by the quality criterion?
2. Goodness

4.1.2 Which aspect of system outputs is assessed by the quality criterion?
1. Form of output

4.1.3 Is each output assessed for quality in its own right, or with reference to a system-internal or external frame of reference?
1. Quality of output in its own right

4.2 Evaluation mode properties
Questions 4.2.1–4.2.3 record properties that are orthogonal to quality criteria (covered by questions in the preceding section), i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are more common than others).

4.2.1 Does an individual assessment involve an objective or a subjective judgment
2. Subjective

4.2.2 Are outputs assessed in absolute or relative terms?
1. Absolute

4.2.3 Is the evaluation intrinsic or extrinsic?
1. Intrinsic

4.3 Response elicitation
The questions in this section concern response elicitation, by which we mean how the ratings or other measurements that represent assessments for the quality criterion in question are obtained, covering what is presented to evaluators, how they select response and via what type of tool, etc. The eleven questions (4.3.1–4.3.11) are based on the information annotated in the large scale survey of human evaluation methods in NLG by Howcroft et al. (2020).

4.3.1 What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given.
Fluency

4.3.2 Question 4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given.
N/A. We provide examples though.

4.3.3 Are the rating instrument response values discrete or continuous? If so, please also indicate the size.
1. Discrete
Size of the instrument: 4

4.3.4 List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument.
1-4 Likert Scale

4.3.5 How is the scale or other rating instrument presented to evaluators? If none match, select 'Other' and describe.
1. Multiple-choice options

4.3.6 If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument.
N/A

4.3.7 What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)?
Instructions
Please read the following text and answer the questions below.
When reading definitions, please focus on their fluency. If a definition's text only says 'nan', please rate it as Not at all fluent.
Term:
Definition:
∗ How fluent is this definition?

4.3.8 Form of response elicitation. If none match, select 'Other' and describe.
2. direct quality estimation

4.3.9 How are raw responses from participants aggregated or otherwise processed to obtain reported scores for

this quality criterion?
macro-averages

4.3.10 Method(s) used for determining effect size and significance of findings for this quality criterion.
Pairwise independent t-tests corrected for multiple hypothesis testing using the Bonferroni-Holm correction

4.3.11 Inter-annotator agreement
Questions 4.3.11.1 and 4.3.11.2 record information about inter-annotator agreement.

4.3.11.1 Has the inter-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used?
1. yes, Krippendorff's α

4.3.11.2 What was the inter-annotator agreement score?
0.45

4.3.12 Intra-annotator agreement
Questions 4.3.12.1 and 4.3.12.2 record information about intra-annotator agreement.

4.3.11.1 Has the intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used?
3. N/A. In our experiment, each evaluator only evaluate each item once.

4.3.11.2 What was the intra-annotator agreement score?
N/A

A.5. Ethics

The questions in this section relate to ethical aspects of the evaluation. Information can be entered in the text box provided, and/or by linking to a source where complete information can be found.

5.1 Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?
Yes. The Research Ethics Committee (CETO) of the Faculty of Arts, University of Groningen.

5.2 Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions)? If yes, describe data and state how addressed.
No. The responses are anonymized.

5.3 Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited)? If yes, describe data and state how addressed.
No

5.4 Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes.
No