ReproHum: #0033-05: Human Evaluation Report on "Generating Scientific Definitions with Controllable Complexity"

Ines Arous
York University
inesar@yorku.ca

Jackie Chi Kit Cheung
McGill University
Canada CIFAR AI Chair, Mila
jackie.cheung@mcgill.ca

Abstract

Human evaluation is a central component in assessing natural language generation (NLG) systems, especially for open-ended and creative tasks. Yet, the field still lacks standardized practices for designing and reporting such evaluations. In this paper, we present a reproduction study of the human evaluation conducted by (August et al., 2022). They evaluate systems that generate scientific definitions with controllable complexity and leverage human evaluation to compare these approaches. By reproducing the experimental setup under comparable conditions, we find that the pairwise ranking of systems is fully consistent with the original study, while the absolute scores diverge substantially, suggesting a moderate level of reproducibility.

1 Introduction

Human evaluation remains a cornerstone of NLP research. It provides nuanced, context-aware judgments, particularly for generation tasks where outputs are open-ended, creative, or domain-sensitive (Howcroft et al., 2020; Hämäläinen and Alnajjar, 2021). Unlike automatic metrics, which optimize for surface-level signals such as lexical overlap or embedding similarity (Mathur et al., 2020; Sai et al., 2021), human evaluation captures insights about users' perceptions of a system's output. It also reveals whether a system is useful in practice, including its appropriateness and reliability, and enables the analysis of systematic errors (van Miltenburg et al., 2021).

Despite its importance, human evaluation receives relatively little attention in NLP research, and when it does appear, details are often under-reported (Schmidtova et al., 2025). Subtle differences in instructions, interfaces, annotator backgrounds, or aggregation methods can meaningfully influence results, yet these details are rarely documented in sufficient depth (Thomson et al., 2024). This lack of transparency, combined with the often overlooked variation in human judgment (Plank, 2022), undermines the reliability of findings from human evaluation and diminishes their value as a basis for subsequent work. For human evaluation to serve as a solid scientific foundation, its findings must be reproducible. Reproducibility, in turn, requires clear, consistently applied, and fully specified protocols that make the evaluation process traceable from design to conclusion. Without such standards, the field risks building on results that are neither verifiable nor comparable, eroding the integrity of human-evaluation practice in NLP. These issues are exacerbated by recent trends in LLM-as-a-Judge approaches, which are often justified by their ability to reproduce human judgments, thereby placing additional pressure on the reliability and consistency of human evaluation itself (Chehbouni et al., 2025).

The ReproHum Project[1] addresses these concerns by developing a methodological framework for systematically testing the reproducibility of human evaluations in NLP. Throughout the project's phases, the ReproHum team identifies limitations and inconsistencies in current practices, translating these findings into actionable recommendations and working toward a community consensus on how to confront the reproducibility crisis. This initiative complements a growing wave of meta-research aimed at strengthening reproducibility across disciplines. The knowledge generated through ReproHum has the potential to reshape human evaluation methodology, making it more dependable and comparable. The insights emerging from ReproHum are driving a shift in the NLP domain and informing future evaluation practices through workshops, reproducibility studies (Arvan and Parde, 2024, 2025), and surveys of current practices (Thomson et al., 2024). Most concretely, the

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

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1117–1126

insights emerging from the project are informing evaluation practices by introducing human evaluation datasheets, which aim to standardize how evaluation decisions are recorded and reported across the NLP community (Belz and Thomson, 2025; Shimorina and Belz, 2022).

Building on this framework, and as part of the broader ReproHum initiative, we reproduce the human evaluation protocol introduced by (August et al., 2022) in their work "Generating Scientific Definitions with Controllable Complexity." We investigate the extent to which their human evaluation experiment can be reproduced. We leverage the original paper and the authors' communication with the ReproHum team to reflect the realistic conditions under which reproducibility is most commonly attempted. We find a gap between the numerical scores reported by the original study and ours, suggesting that the task instructions may be ambiguous and interpreted differently across annotator pools. Nonetheless, we obtain a system ranking similar to that of the original study, indicating a moderate level of reproducibility.

2 Related Work

With the rapid proliferation of NLG tasks, a wide range of evaluation methodologies have been explored. Automatic metrics, while attractive due to their scalability and low cost, have repeatedly been shown to rely on superficial or irrelevant cues. In addition, they often fail to capture the effects of meaning-altering perturbations such as negation. They also tend to overemphasize lexical overlap, rendering them insensitive to semantic errors yet overly sensitive to benign paraphrases (Kryscinski et al., 2019; Sai et al., 2021; Mathur et al., 2020). Recently, LLM-based approaches, such as LLM-as-judge, have gained traction as alternatives. However, these methods have also been shown to suffer from inherent biases and limitations, including inconsistencies in judgment and difficulties in reliably assessing factual correctness (Gao et al., 2025; Bavaresco et al., 2025; Fu et al., 2023; Chehbouni et al., 2025). As a result, despite its cost and variability, human evaluation remains the de facto standard and most trusted approach for assessing the quality of generated text.

Despite its central role, human evaluation poses substantial challenges for reproducibility. These difficulties often stem from ambiguous experiment design, such as underspecified evaluation criteria, poorly defined rating scales, and inconsistent or missing annotator instructions, which collectively undermine the interpretability and comparability of results across studies (Howcroft et al., 2020; Hämäläinen and Alnajjar, 2021; Belz et al., 2023). Compounding this issue, many papers omit critical experimental details, such as experiment guidelines, training procedures, annotation interfaces, and participant demographics, rendering reported evaluations difficult or impossible to reproduce (Ruan et al., 2024). Even when authors are contacted directly, such information is frequently unavailable, leaving many published human evaluations non-reproducible (Thomson et al., 2024; Belz et al., 2023). These problems are further exacerbated by structural weaknesses in annotator selection. Most studies rely on non-expert crowd workers, commonly recruited via platforms such as Amazon Mechanical Turk, even for domain-specific tasks that require specialized knowledge. Prior work has shown that such setups can yield high variance, low inter-annotator agreement, poor calibration, and instability across time or worker pools (Sap et al., 2022; Karpinska et al., 2021; Hämäläinen and Alnajjar, 2021).

In response to these persistent challenges, several community-driven initiatives have been launched to systematically examine and diagnose the state of human evaluation in NLP research. Notable examples include HumEval and the broader ReproNLP ecosystem, which explicitly target issues of reproducibility and methodological transparency in evaluation practices. Within this context, the ReproHum initiative conducted a series of coordinated reproduction efforts from 2023 to 2026, in which participating teams attempted to independently reproduce previously published human evaluation studies. These efforts exposed recurring obstacles, including undocumented bug fixes, deviations from the original experimental procedures, and mismatches in the number or composition of evaluators (Thomson et al., 2024). Such discrepancies frequently led to divergent outcomes, underscoring the lack of standardized quality criteria, evaluation protocols, and reporting practices across the literature. Collectively, these findings highlight the necessity of multiple independent reproductions and the use of complementary quantitative measures of reproducibility to draw robust conclusions about evaluation results. Importantly, these initiatives not only surfaced structural weaknesses in existing practices but also motivated concrete steps

toward standardization, such as proposing human evaluation data sheets to improve transparency, documentation, and reuse of evaluation setups (Belz and Thomson, 2025). The present work is situated within the ReproHum initiative and focuses on reproducing the human evaluation study reported in a prior ACL publication (August et al., 2022).

3 Methods

3.1 Generating Scientific Definitions with Controllable Complexity

August et al. (2022) introduced a new task and dataset for adapting the complexity of scientific term definitions to readers' backgrounds. They leverage two sources for creating their dataset: 1) the medical consumer questions (Ben Abacha and Demner-Fushman, 2019) and 2) Wikipedia science glossaries.[2] They construct their dataset by filtering these sources to identify definitions that answer the question "What is (are) X?" where X is a scientific term or concept. Additionally, they associate each scientific term with a supporting document consisting of 10 related abstracts extracted from S2ORC (Lo et al., 2020), which is a corpus of over 81 million scientific articles.

They use four methods' categories to generate definitions for the scientific terms in their pre-collected dataset: 1) finetuned BART (Lewis et al., 2020) as a sequence-to-sequence model, 2) GPT-2 and GPT-3 as causal language models, 3) GPT-2 finetuned, and 4) bi-directional attention flow (BiDAF) as an information retrieval method. They found that BART, fine-tuned on the supporting documents, outperforms all other methods at generating scientific definitions. Then, they propose controlling the complexity of definitions by re-ranking the generated candidates from BART using two discriminators: SciBERT, an uncased pretrained model (Beltagy et al., 2019), and an SVM trained to distinguish high-complexity from low-complexity definitions. This approach yields two variants of their method: Rerank-BERT and Rerank-SVM. They compare their approach with three baselines: 1) plug-and-play language models (PPLM) (Dathathri et al., 2019), 2) generative discriminators (GeDi) (Krause et al., 2021) and 3) ensemble of language models (DExperts) (Liu et al., 2021). They evaluate both variants of their approach (Rerank-SVM and Rerank-BERT) and the baseline methods using automatic metrics, including sentence length and language model perplexity with GPT. Their automated evaluation suggests that both Rerank-BERT and Rerank-SVM performed best overall.

They select the top three best-performing methods for additional human evaluation: DExperts, GeDi, and Rerank-SVM. They conduct two sets of human evaluation experiments. The first set of experiments involves 233 Amazon Mechanical Turk (MTurk) workers who evaluated, on a 1-4 Likert scale, how complicated a definition was and how difficult it was to understand. They collectively evaluated 50 terms with high- and low-complexity generations for each model, rendering a total of 300 definitions. The second set of experiments involves two trained annotators, one of whom is an author, to evaluate the same 300 definitions on fluency, relevance, and factuality. For fluency and relevance, annotators had to rate definitions on a 1–4 Likert scale (1 = "Not at all" to 4 = "Very"). When evaluating factuality, annotators were asked to indicate whether a definition contained any factually incorrect information (a binary label). If so, they rated the extent of these errors on a 1–4 Likert scale.

The authors provided additional information regarding the human evaluation in the appendix of their paper. In the first set of experiments, participants filled out a short demographics questionnaire and were provided instructions with examples of very complex and not-at-all complex definitions. The authors also provided a screenshot of the interface for this set of experiments. In the second set of experiments, annotators were provided examples, and when evaluating fluency and relevance, the instructions for the task were:

- How fluent is this definition?
- How relevant is this definition for the term?

When evaluating factuality, annotators were first shown examples of both highly extensive and minimal factual errors, and then asked, "Does this definition contain factually incorrect information?" They were encouraged to use the internet if they were unsure whether a definition was correct.

For the first set of experiments, the authors found that DExperts generations differ most between high and low complexity, whereas GeDi showed counter-intuitive results, with low-complexity generations

[2] https://en.wikipedia.org/wiki/Category:Glossaries_of_science

Instructions

You will be given 300 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.

Figure 1: The full instructions provided to annotators

rated as more complicated and difficult to understand than high-complexity generations. They also found that Rerank-SVM showed consistent performance, with high-complexity generations rated as more complex and difficult to understand. In the second set of experiments, they found that Rerank-SVM's definitions were more fluent and relevant than the baselines and had significantly fewer factual errors. The results from their human evaluation were represented in Figure 2 and Table 7 for the first and second sets of experiments, respectively.

3.2 Scope of Reproduction

Our aim was to reproduce the designated experiment as faithfully as possible to the original study. Our reproduction focused on a specific part of the study: the human evaluation of the factuality criterion.

3.3 Additional Information Obtained from Original Authors

Although we did not communicate directly with the original authors, the ReproHum team shared additional information they obtained from them. In particular, the authors provided the exact outputs used in their evaluation. They also provided screenshots of the user interface for the complexity experiment, including the examples and instructions provided to annotators. They also reported the cost of the human-evaluation setup. The complexity study cost $140, while the second set of experiments conducted by two trained annotators incurred no additional expense, as the annotators were either separately funded research assistants or the authors themselves.

Examples

Examples of definitions with no factual inaccuracies: Term: Acanthoma Definition: Acanthoma is a skin lesion that develops from cells in the skin.
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 factually inaccurate definitions: 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).

Figure 2: Examples provided to annotators

3.4 Known Deviations from the Original Experiment

We recognize factors that may have caused our results to deviate from the original experiment. Some are due to unclear or incomplete details in the paper and in the follow-up information we received, while others reflect constraints that made it impossible to recreate the exact experimental conditions. We made sure to communicate these deviations to the ReproHum team to remain as faithful as possible to the original experiment and to other reproduction studies of this work, and we detail them below.

Platform for the Human Evaluation Experiment: The original study was conducted on the LabintheWild platform using a custom, internally developed interface. This platform is no longer available for new experiments. For our reproduction, we used MTurk's sandbox environment to recreate the interface, ensuring that the instructions and rating layout matched the original as closely as possible.

Participant Selection and Payment: In the original study, one annotator was a research assistant and the other was one of the authors, and neither received payment. For our reproduction, we recruited two PhD students in computer science. Based on an estimated hourly rate of 30 CAD and an expected task duration of five hours, we compensated each

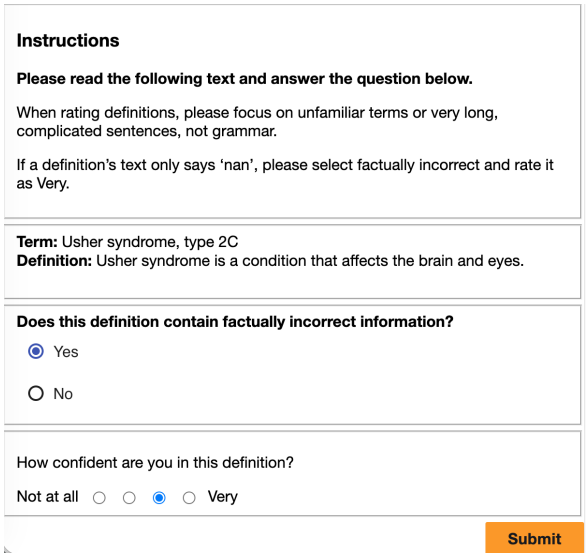

Figure 3: Interface presented to annotators for assessing factuality in the human-evaluation task.

participant with 150 CAD.

Expected Completion Time: It was not clear from the original study whether annotators were expected to complete the task in a single session or over multiple sessions. Given our estimate that the task would take approximately five hours, we allowed participants to complete it over multiple sittings to minimize fatigue.

Instructions and Examples: The instructions and examples provided by the authors covered only the complexity experiment. For our reproduction, the ReproHum team provided instructions and examples tailored to our scope, which focuses on evaluating the factuality criterion. We show the instructions and examples in Figures 1 and 2, and illustrate the interface in Figure 3.

Data Analysis The authors did not release the source code used for their data analysis, so we implemented our own analysis pipeline. We report results for cases where either annotator marked a factual error, and for those where both did.

Extent of Factual Errors Experiment It was not clear from the original paper which aggregation method was used when analyzing the extent of factual errors. One possible interpretation is to exclude cases in which only one annotator identified a factual error, while another is to include these cases by assigning them a value of zero on the Likert scale. We ran the experiment under both interpretations and found that including these cases as zero produces results closest to the original study and avoids the need for the two aggregation conditions, since all instances are accounted for. We therefore report this version.

Table 1: Comparison of the original and reproduced results when either annotator identified a factual error, with R1 being the reproduced experiment conducted by (Florescu et al., 2025).

Method	Original (%)	R1	Ours (%)
DEXPERT	86	78	80
GEDI	52	78	65
SVM	38	78	39

4 Results

This section presents the results of our reproduction study evaluating the factuality of definitions generated for scientific terms. We closely followed the original human evaluation protocol (August et al., 2022), comparing SVM-Rerank, DEXPERT, and GEDI. Factuality is evaluated in two steps: the first counts how many factual errors each system has according to the annotators, and the second assesses the severity of those errors. In what follows, we first explain the main measures used to quantify reproducibility, and then we discuss the results of each of the two steps.

4.1 Quantified Reproducibility Assessment

We applied the standardized procedure for reproducibility assessment proposed by the ReproHum team (Belz and Thomson, 2026). This procedure distinguishes between four types of evaluation results. For our case, where we are interested in reproducing numerical values and comparing different systems, only three of these evaluation types apply to us:

- Type I results: evaluates single numerical scores, such as mean quality ratings or error counts. In our case, we evaluate error counts in terms of factuality. Reproducibility

Table 2: Comparison of the original and reproduced results when both annotators identify a factual error, with R1 being the reproduced experiment conducted by (Florescu et al., 2025).

Method	Original (%)	R1	Ours (%)
DEXPERT	67	54	38
GEDI	33	51	32
SVM	16	57	16

Table 3: QRA reproducibility assessment comparing the human evaluation experiment in August et al. (2022) with ours **when one of the two annotators identifies a factual error**; n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 2$)		
				System level	QC level	Study level
Type I	Factuality	DEXPERT GEDI SVM	(mean) CV*↓	5.64 17.34 2.03	8.34	8.34
Type II	Factuality	all all all	mean r ↑ mean ρ ↑ W↑	n/a n/a n/a	0.925 1.000 1.000	n/a
Type IV	Factuality	all	P ↑	n/a	1.000	1.000

Table 4: QRA reproducibility assessment comparing the human evaluation experiment in August et al. (2022) **when both annotators identify a factual error**; n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 2$)		
				System level	QC level	Study level
Type I	Factuality	DEXPERT GEDI SVM	(mean) CV*↓	43.10 2.40 0	15.16	15.16
Type II	Factuality	all all all	mean r ↑ mean ρ ↑ W↑	n/a n/a n/a	0.902 1.000 1.000	n/a
Type IV	Factuality	all	P ↑	n/a	1.000	1.000

is assessed using the unbiased coefficient of variation (CV^*), where lower values indicate better reproducibility.

- Type II results: evaluates the correlation between related numerical scores (e.g., multiple Type I results). It uses different correlation coefficients, such as Pearson's r, Spearman's ϕ, and Kendall's W. Higher correlation coefficients indicate better reproducibility.

- Type IV results: provides information about which system performs better than another on a given task. It is quantified through the proportion of identical pairwise system ranks in a set of comparable experiments P, with higher P indicating better reproducibility.

These measures are computed at three different degrees of reproducibility: (i) system level, where we compare scores obtained for the same system across different experiments; (ii) quality-criterion (QC) level, where we compare scores for a given QC across multiple systems and experiments; and (iii) study level, where we compare scores for all systems and all QCs across the full set of experiments that make up the study.

4.2 Reproduction Results for Counting Factual Errors

Table 1 and 2 present the raw results of the experiment, using the results obtained by the authors, those reproduced in (Florescu et al., 2025), and ours. Tables 3 and 4 report the QRA reproducibility assessment for the original and our reproduced experiments under the two aggregation conditions, when only one annotator identifies a factual error and when both annotators identify the same factual error, respectively. As shown by the CV^* values in Tables 3 and 4, the three systems differ notably in their reproducibility across the two aggregation conditions. DEXPERT shows good reproducibility when a factual error is counted if at least one annotator identifies it, but its reproducibility drops to poor when errors are counted when both annotators agree. GEDI achieves medium reproducibility under the "either annotator" rule, but improves to good reproducibility when factual errors require agreement between annotators. SVM maintains good reproducibility in both conditions, regardless of whether errors are aggregated using the "either" or "both" annotator criterion. For Type II results, we find that the different correlation coefficients are exceptionally high, indicating (near) perfect correlation. Finally, P is 1, indicating that the repeated experiments obtained exactly the same three pairwise ranks as the original experiment.

In terms of statistical significance, we find that SVM outperforms GEDI, with results in our reproduction (t = 4.58, p = 3.18e-05, Cohen's d = 0.65) closely matching those reported in the original paper (t = 4.71, p < 0.001, d = 0.47). For DEXPERT,

Table 5: QRA reproducibility assessment comparing the human evaluation experiment in August et al. (2022) and ours when annotators indicate **the extent of a factual error**; n/a = measure does not apply at this level.

Type of Result	**QC**	**System**	**Measure applied**	**Degree of reproducibility** ($n = 2$)		
				System level	QC level	Study level
Type I	Factuality	DEXPERT GEDI SVM	(mean) CV*↓	35.05 28.93 59.97	41.32	41.32
Type II	Factuality	all all all	mean r ↑ mean ρ ↑ W↑	n/a n/a n/a	0.967 1.000 1.000	n/a
Type IV	Factuality	all	P ↑	n/a	1.000	1.000

however, the statistical signal is weaker in our reproduction: we obtain (t = 6.73, p = 1.77e-08, d = 0.95), compared with the much stronger effect reported in the original study (t = 12.29, p < 0.001, d = 1.24).

4.3 Reproduction Results for the Extent of Factual Errors

If annotators indicated a factual error, they had to rate its extent on a scale of 1 to 4. The results of this experiment were reported in Table 7 (last column) in the original paper (August et al., 2022). We reproduce the results of this experiment, accounting for cases where annotators indicated no factual errors by assigning a 0 on the Likert scale.

Table 5 presents the QRA-based reproducibility assessment for the original and reproduced experiments. We find, from the CV^*, that the three systems exhibit medium-to-poor reproducibility. Specifically, GEDI shows medium reproducibility, whereas DEXPERT and SVM both fall into the poor reproducibility range. For Type II results, the various correlation coefficients are exceptionally high, indicating (near) perfect correlation between the original and reproduced experiments. Finally, the value of P = 1 shows that the repeated experiment produced exactly the same three pairwise rankings as the original study.

5 Discussion

Our reproduction highlights several important insights into the robustness and interpretability of the original study's findings. First, the QRA reproducibility assessment reveals notable variability across systems. This is accentuated in the second experiment, which evaluates the extent of factual errors. This outcome is likely influenced by how the Likert-scale responses were interpreted: across all annotations, the option "1 — Not at all" was selected only once, and only by a single annotator. This option is arguably confusing, since selecting "Not at all" when rating the extent of a factual error is inconsistent with having answered "Yes" to the preceding question about whether a factual error was present. The resulting lack of variation in the Likert responses likely inflated the CV^* values and contributed to the lower reproducibility scores.

In contrast, the Type II reproducibility results paint a different picture. The correlation coefficients between the original and reproduced experiments are exceptionally high, indicating near-perfect correspondence in the relative ordering of system performance. This is further supported by the finding that P = 1, meaning that the reproduced experiment produced exactly the same three pairwise rankings as the original study. Thus, while the CV^* values might reflect variability in reproducibility, the relative performance of the systems is highly stable across replications.

We also observe notable differences between our results and those reported in (Florescu et al., 2025), both in the raw evaluation scores and the resulting system rankings. In their reproduced experiments, all systems obtain the same performance whenever a factual error is identified by either annotator (see Table 1). Similarly, their results are somewhat comparable when both annotators agree on a factual error (see Table 2). This contrasts with our findings and suggests that the interpretation of the evaluation instructions may be ambiguous and vary across annotator pools. Indeed, these discrepancies provide further evidence that annotators may operationalize evaluation criteria differently, even when following the same protocol. Notably, our reproduction yields a system ranking that more closely aligns with that reported by the original authors. We hypothesize that this alignment may be partially explained by an implicit condition related to annotator background: our participants are PhD students in computer science, mirroring the expertise of the original study's annotators, whereas the participants in the ReproHum replication were PhD

students in psychology. Although the annotator's background was not explicitly stated as a controlled condition in the original setup, this difference in domain expertise may have influenced how evaluation criteria were interpreted, ultimately affecting the resulting system rankings.

6 Conclusion

In this work, we attempt to reproduce the human evaluation experiment reported in "Generating Scientific Definitions with Controllable Complexity", focusing specifically on the factuality criterion. We systematically reconstructed the experimental setup described in the original paper and aligned our procedures as closely as possible with the authors' reported methodology. Comparing our reproduced results with those of the original study using the QRA reproducibility assessment, we find that the human evaluation exhibits a medium level of reproducibility. While the pairwise ranking of systems is fully consistent across the two experiments, the numerical scores diverge substantially from those reported by the original authors. This suggests that the relative performance ordering of the systems is stable, but the absolute magnitude of the factuality judgments is sensitive to annotation choices and experimental interpretation.

Acknowledgments

We would like to thank the ReproHum team, especially Craig Thomson, for their support and guidance throughout this reproduction. We would also like to thank our participants for their contribution to our study and the original authors for providing additional information and clarifications. The research was undertaken thanks in part to funding from the Connected Minds Program, supported by the Canada First Research Excellence Fund, Grant #CFREF-2022-00010, and the Canada Research Chairs Program.

References

Mohammad Arvan and Natalie Parde. 2024. ReproHum #0712-01: Human Evaluation Reproduction Report for "Hierarchical Sketch Induction for Paraphrase Generation". In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 210–220, Torino, Italia. ELRA and ICCL.

Mohammad Arvan and Natalie Parde. 2025. ReproHum: #0744-02: Investigating the Reproducibility of Semantic Preservation Human Evaluations. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 590–600, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

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.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. 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 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, 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.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511.

Khaoula Chehbouni, Mohammed Haddou, Jackie CK Cheung, and Golnoosh Farnadi. 2025. Neither Valid nor Reliable? Investigating the Use of LLMs as Judges.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and Play Language Models: A Simple Approach to Controlled Text Generation.

Andra-Maria Florescu, Marius Micluța-Câmpeanu, Stefana Arina Tabusca, and Liviu P Dinu. 2025. ReproHum #0033-05: Human Evaluation of Factuality from A Multidisciplinary Perspective. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 583–589, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan Tn. 2023. Are Large Language Models Reliable Judges? A Study on the Factuality Evaluation Capabilities of LLMs. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 310–316, Singapore. Association for Computational Linguistics.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. LLM-based NLG Evaluation: Current Status and Challenges. *Computational Linguistics*, 51:661–687.

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.

Mika Hämäläinen and Khalid Alnajjar. 2021. Human Evaluation of Creative NLG Systems: An Interdisciplinary Survey on Recent Papers. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 84–95, Online. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. 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.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. 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 540–551, Hong Kong, China. 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.

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.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. *arXiv preprint*. ArXiv:1911.02782.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. Defining and Detecting Vulnerability in Human Evaluation Guidelines: A Preliminary Study Towards Reliable NLG Evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7965–7989, Mexico City, Mexico. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation

CheckLists for Evaluating NLG Evaluation Metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7219–7234, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Patricia Schmidtova, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 2025. Do My Eyes Deceive Me? A Survey of Human Evaluations of Hallucinations in NLG. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79, Hanoi, Vietnam. 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, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.