ReproHum: #0744-02: Investigating the Reproducibility of Semantic Preservation Human Evaluations

Mohammad Arvan and Natalie Parde
{marvan3, parde}@uic.edu
University of Illinois Chicago

Abstract

Reproducibility remains a fundamental challenge for human evaluation in Natural Language Processing (NLP), particularly due to the inherent subjectivity and variability of human judgments. This paper presents a reproduction study of the human evaluation protocol introduced by Hosking and Lapata (2021), which assesses semantic preservation in paraphrase generation models. By faithfully reproducing the original experiment with careful adaptation and applying the Quantified Reproducibility Assessment framework (Belz and Thomson, 2024a; Belz, 2022), we demonstrate strong agreement with the original findings, confirming the semantic preservation ranking among four paraphrase models. Our analyses reveal moderate inter-annotator agreement and low variability in key results, underscoring a good degree of reproducibility despite practical deviations in participant recruitment and platform. These findings highlight the feasibility and challenges of reproducing human evaluation studies in NLP. We discuss implications for improving methodological rigor, transparent reporting, and standardized protocols to bolster reproducibility in future human evaluations. The data and analysis scripts are publicly available to support ongoing community efforts toward reproducible evaluation in NLP and beyond.

1 Introduction

Reproducibility is a cornerstone of scientific progress, ensuring that research findings are reliable, verifiable, and can form a solid foundation for subsequent studies. In the field of Natural Language Processing (NLP), where models and systems evolve rapidly, reproducibility is especially critical to validate claims and foster cumulative knowledge. Central to this effort are evaluation strategies that assess model performance, broadly categorized into automatic metrics and human judgments. Automatic evaluation offers efficiency and consistency but often fails to capture nuanced language understanding, whereas human evaluation provides richer, more context-sensitive insights, at the expense of scalability, objectivity, and cost (Belz and Reiter, 2006; Reiter and Belz, 2009; Liu et al., 2016). These trade-offs highlight the complementary roles of both approaches in NLP research.

As NLP models increasingly approach or surpass the limits of traditional automatic metrics and benchmarks, the role of human evaluation has become more pronounced. However, the inherent subjectivity in human judgments introduces challenges for reproducibility. Variability in annotator expertise, demographic factors, evaluation environments, and subtle differences in protocol implementation can all introduce variability into human evaluation results (Howcroft et al., 2020). Addressing these challenges requires clear, consistent, and transparent protocols for human evaluation to preserve the integrity and comparability of NLP research.

The ReproHum Project (Belz et al., 2023; Belz and Thomson, 2024a) represents a concerted effort to address these issues by developing systematic approaches to enhance the reproducibility of human evaluations in NLP. By formalizing methodological frameworks and providing practical guidelines, ReproHum seeks to mitigate sources of inconsistency and enable more reliable comparisons across studies. This initiative complements a growing body of meta-analytical work aimed at improving reproducibility and rigor across scientific disciplines (Open Science Collaboration, 2015; Errington et al., 2021a,b). Motivated by these considerations, we undertake a focused reproduction study of the human evaluation protocol introduced by Hosking and Lapata (2021) in their work on "Factorising Meaning and Form for Intent-Preserving Paraphrasing." At the heart of our investigation lies a central research question:

To what extent can human evaluation

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 590–600

results be faithfully reproduced when following the original experimental setup?

The remainder of this paper is organized as follows: Section 2 reviews related work and foundational concepts; Section 3 describes the original protocol and details our reproduction methodology; Section 4 presents results and analyzes observed differences; Section 5 discusses broader implications and proposes guidelines informed by our findings; finally, Section 6 concludes and outlines directions for future research. The input, preference data, and analysis are made available in GitHub (Arvan and Parde, 2025).

2 Background

Human evaluation remains the most trusted method for assessing NLP system outputs, yet reproducibility in these evaluations continues to pose major challenges. These difficulties largely arise from inconsistent methodologies, ambiguous reporting, and incomplete experimental details. The field suffers from considerable heterogeneity in quality criteria and terminology, as demonstrated by Howcroft et al. (2020), who surveyed 165 NLG papers using human evaluation and found over 200 distinct terms describing quality. Crucially, about two-thirds of those papers failed to define what quality aspects they measured, and many omitted essential information such as system inputs, outputs, or participant demographics. This fragmentation hampers comparability and aggregation of results, underscoring the need for consistent evaluation designs and terminology.

Building on this foundation as part of the ReproHum initiative, Belz et al. (2023) examined 177 NLP papers with human assessments, identifying that only around 13% contained sufficiently accessible information to allow confident reproduction. Frequent issues included missing participant details, unclear instructions, and methodological flaws, all exacerbated by incomplete documentation and limited author involvement. They recommended adopting structured recording protocols such as the Human Evaluation Data Schema (HEDS) (Shimorina and Belz, 2022) and called for stronger standardization in experimental design to improve reproducibility.

Efforts to assess reproducibility directly have been made through organized shared tasks like those in the HumEval and ReproNLP initiative (Belz et al., 2023; Belz and Thomson, 2024b; Balloccu et al., 2024; Belz et al., 2025). In 2023 and 2024, ReproHum partners attempted to reproduce existing human evaluation studies, revealing common obstacles such as inconsistent bug fixes, procedural deviations, and variability in evaluator numbers. These factors often led to divergent results and illustrated the persistent lack of uniformity in quality criteria and evaluation protocols. Such challenges emphasize the necessity for multiple reproductions and diverse quantitative reproducibility measures to ensure reliable conclusions.

Recent critical analyses echo and extend these concerns by highlighting fundamental limitations in human evaluation methodologies. For example, Hosking et al. (2024) demonstrate that human preference feedback used in training and evaluating large language models tends to systematically underrepresent important aspects such as factuality, and is biased by factors like assertiveness and output complexity. Similarly, Gehrmann et al. (2023) survey widespread flaws affecting human evaluation, automatic metrics, and datasets, arguing that current protocols are unsustainable for distinguishing advanced models and advocating for causal frameworks in evaluation reporting. Complementing these perspectives, Thomson et al. (2024) document pervasive experimental flaws in repeated human evaluations, ranging from coding errors to deviations from scientific best practices and inaccurate result reporting. Together, these studies caution that human evaluation, while indispensable, is imperfect and vulnerable to methodological weaknesses. They recommend comprehensive reforms including multiple annotators, improved experimental rigor, better annotation aggregation, transparent reporting, and stronger oversight to increase reliability and scientific rigor.

Taken together, these findings—from foundational surveys to reproducibility analyses, shared task experiences, and broad methodological critiques—paint a consistent picture: human evaluation is essential for NLP system assessment but is hindered by inconsistent reporting, incomplete details, ambiguous quality definitions, and practical execution flaws. There is a collective imperative to adopt comprehensive standardization of evaluation methodologies, rigorous documentation protocols, better annotator training and incentivization, and more reliable aggregation and analysis techniques. Addressing these challenges is crucial for advancing scientific rigor and trustworthiness

in NLP evaluation.

3 Experimental Setup and Reproduction

This section details the human evaluation methodology reproduced from Hosking and Lapata (2021). We first describe the original evaluation protocol used to assess paraphrase generation models on semantic preservation, dissimilarity, and fluency. Next, we outline our reproduction setup focusing exclusively on the semantic preservation criterion, including participant recruitment and survey implementation. Then, we summarize key deviations from the original experiment and justify the necessary adaptations, highlighting how these changes were managed to maintain the integrity and validity of the reproduction. Finally, we describe the Quantified Reproducibility Assessment (QRA) framework (Belz and Thomson, 2024a; Belz, 2022) used to evaluate the reproducibility of our results. This framework provides a structured approach to assess reproducibility across various dimensions, including statistical consistency, inter-annotator agreement, and qualitative findings.

3.1 Original Evaluation Protocol

The original evaluation compared several paraphrase generation models to assess their performance in balancing semantic preservation, syntactic variation, and fluency. The primary models evaluated are summarized below:

SEPARATOR. Introduced by Hosking and Lapata (2021), SEPARATOR employs an encoder-decoder architecture featuring a Vector-Quantized Variational Autoencoder (VQ-VAE) bottleneck that explicitly disentangles semantic and syntactic information in the latent space. Specifically, semantic content is encoded as continuous latent variables, while surface form is represented as discrete latent variables. At test time, manipulating the discrete syntactic latent codes while fixing the semantic codes enables generation of paraphrases with substantial syntactic variation that preserve the original meaning. This design allows for a principled trade-off between semantic fidelity and syntactic novelty without the need for access to target exemplars.

DiPS. This method enhances paraphrase diversity by applying submodular optimization over outputs from a standard encoder-decoder paraphrasing model (Kumar et al., 2019) . The approach encourages varied surface realizations, fostering greater lexical variation.

Latent BoW. This method uses a discrete bag-of-words latent representation within an encoder-decoder framework (Bowman et al., 2016). This explicitly models word presence, promoting lexical diversity in generated paraphrases.

VAE Baseline. This baseline shares the overall encoder-decoder architecture with SEPARATOR but encodes semantic and syntactic features jointly as continuous Gaussian latent variables, without disentangling them. This joint encoding limits the model's capacity for controlled syntactic variation.

In the original evaluation, crowdworkers on Amazon Mechanical Turk (MTurk) compared an original question with two paraphrases generated by different models. Annotators selected their preferred paraphrase based on three criteria:

- **Dissimilarity**: How distinct the paraphrase is in surface form from the original question;
- **Semantic preservation**: How well the paraphrase retains the original meaning or intent; and
- **Fluency**: The naturalness and grammaticality of the paraphrase.

The authors reported sampling 200 questions evenly from the Paralex (Fader et al., 2013) and Quora Question Pairs (QQP) (DataCanary et al., 2017) datasets. Each paraphrase pair was evaluated independently by three annotators, resulting in 600 judgments per criterion. Annotators made forced-choice selections, assigning +1 to the preferred paraphrase and -1 to the alternative for each criterion. These scores were averaged over annotators, where negative values indicate less frequent preference.

The evaluation interface presented the original question alongside two paraphrases side-by-side. Annotators were instructed to consider surface differences, semantic equivalence, and fluency carefully, promoting consistent judgments. Compensation was set at $3.50 per Human Intelligence Task (HIT), each containing 32 paraphrase pairs with an expected completion time of 20 minutes.

Additional information from the ReproHum team, via communication with the original authors, included exact evaluated outputs and the user interface used. Notably, the evaluation incorporated *attention checks*: control samples with known labels embedded within each HIT. Two controls were

Aspect	Original Experiment	Reproduction
Evaluation criterion	Semantic preservation, dissimilarity, and fluency	Semantic preservation only
Crowdsourcing platform	`Amazon Mechanical Turk` (MTurk)	`Prolific`
Region restrictions	United States, United Kingdom	United States, United Kingdom, Australia, Canada
Participant approval rate	Minimum 96%	Minimum 99%
Minimum HITs completed	5,000 HITs	200 HITs
Expected time per HIT	20 minutes	8 minutes
Payment per HIT	$3.50 ($10.50/hour)	£1.60 / $2 (£12 / $15.14/hour)

Table 1: Summary of key differences between the original experiment and our reproduction.

deployed. In one, system output was a random paraphrase with a completely different meaning (intended to fail the *meaning* criterion), and in the other, output was identical to the input (intended to fail the *dissimilarity* criterion). Since our reproduction focuses solely on semantic preservation, we excluded the second control. HITs failing attention checks were relisted to ensure data quality.

Key findings reported in the original study indicate that while the VAE baseline best preserves question meaning, it produces the least variation. By contrast, SEPARATOR yields significantly more variation, better preserves original question intent, and generates more fluent paraphrases. These differences were statistically significant (one-way ANOVA with post-hoc Tukey HSD test, $p < 0.05$). We focus exclusively on the semantic preservation criterion; findings relating to dissimilarity and fluency are beyond the scope of this reproduction.

3.2 Reproduction Setup and Deviations

Our reproduction aimed to reproduce the original experiment as closely as possible with a narrowed focus on semantic preservation. We used all available information from the original paper (Hosking et al., 2022) and follow-up communications coordinated by the ReproHum team. We also completed the Human Evaluation Datasheet (HEDS) (Shimorina and Belz, 2022; Belz and Thomson, 2024c) documenting the evaluation details.[1]

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

Certain deviations were necessary due to differences in scope, platform, and participant recruitment. These deviations are summarized in Table 1 alongside the original experiment for clarity. These adaptations reflect practical constraints and the ReproHum project's standards for participant compensation and consistency. We implemented our own data analysis scripts and conducted additional analyses to quantify the degree of reproducibility.

3.3 Statistical Analysis

Power Analysis. A priori power analyses were conducted to confirm that the sample size ($n = 200$ per group across $k = 4$ models) would be adequate to detect differences at a conventional significance level ($\alpha = 0.05$). Effect sizes were based on established conventions (Cohen, 2013), ensuring sufficient sensitivity to detect effects of practical significance in our group comparisons.

Group Comparisons. Differences in semantic preference scores among the four paraphrase models were assessed using a one-way Analysis of Variance (ANOVA). ANOVA is a standard parametric test for evaluating mean differences across multiple independent groups, ideally under assumptions of normality and homogeneity of variances. In this study, we did not formally test these assumptions. However, with balanced and relatively large sample sizes per group, ANOVA is generally robust to moderate violations of normality and heteroscedasticity (Lix et al., 1996). Consequently, ANOVA was deemed appropriate for identifying differences

System	Wins	Losses	Win %	Best-Worst Score	Best-Worst Scale
VAE	1413	387	78.5%	1026	57.00
SEPARATOR	913	887	50.7%	26	1.44
Latent BoW	779	1021	43.3%	-242	-13.44
DiPS	495	1305	27.5%	-810	-45.00

Table 2: Summary of human preferences for semantic preservation across paraphrase models, including best-worst scores and normalized scales.

in mean semantic preference scores. Upon a statistically significant ANOVA result, Tukey's Honest Significant Difference (HSD) test was chosen for post-hoc pairwise comparisons, as it controls the family-wise error rate when performing multiple group comparisons.

Inter-rater Agreement. To assess the reliability of categorical ratings provided by multiple annotators, Fleiss's κ statistic was employed. Fleiss's κ is specifically designed for measuring agreement among more than two raters on nominal scales, and assumes all data are fully rated with no missing labels. Unlike Krippendorff's α, which accommodates missing data and a variety of measurement levels, Fleiss's κ is directly applicable and interpretable given our annotation design: nominal data, complete ratings, and uniform measurement scale across annotators. This makes Fleiss's κ the most relevant and suitable choice for evaluating inter-rater agreement in our study.

3.4 Quantified Reproducibility Assessment

We adopt the Quantified Reproducibility Framework (QRA++) as described by Belz (2025), which categorizes results commonly reported in NLP and machine learning into four types and associates each with appropriate reproducibility metrics. The small-sample coefficient of variation (CV*) is used as a key indicator of reproducibility for numerical results, with the following interpretation: CV* values from 0 up to approximately 10 indicate a good degree of reproducibility; values between 10 and approximately 30 indicate medium reproducibility; and values above 30 indicate poor reproducibility.

The four result types and their associated reproducibility measures are:

1. **Type I results:** Single numerical scores, such as mean quality ratings or error counts. Reproducibility is assessed using the small-sample coefficient of variation (CV*) (Belz, 2022).

2. **Type II results:** Sets of related numerical scores (e.g., multiple Type I results). These are evaluated using correlation coefficients such as Pearson's r and Spearman's ρ.

3. **Type III results:** Categorical labels are attached to text spans of variable length. In the context of reproducibility, inter-rater agreement metrics such as Fleiss's κ or Krippendorff's α are commonly reported to assess consistency among annotators on the same dataset. However, since responses from the original experiment are not available, we cannot report inter-rater agreement as a measure of reproducibility for the original study.

4. **Type IV results:** Qualitative findings stated explicitly or implied by quantitative results in the original paper. Reproducibility is quantified by the proportion of original findings confirmed in the reproduction experiment.

4 Results

This section presents the outcomes of our reproduction study evaluating semantic preservation in paraphrase generation models. We closely followed the original human evaluation protocol (Hosking et al., 2022), comparing four systems: the VAE baseline, SEPARATOR, Latent BoW, and DiPS.

4.1 Semantic Preference Outcomes

Table 2 summarizes crowdworker preferences aggregated over 600 pairwise comparisons per system pair. Columns indicate the number of times a model's paraphrase was preferred (*wins*), disfavored (*losses*), the net preference score (*wins - losses*), and the overall win percentage.

The VAE baseline clearly dominates, winning nearly 79% of comparisons and achieving the highest net preference score, indicating the strongest semantic fidelity. SEPARATOR's paraphrases were moderately preferred, reflecting its design trade-off

between preserving meaning and encouraging syntactic variation. Latent BoW and DiPS trailed, with DiPS showing the lowest semantic preservation according to annotator judgments.

4.2 Statistical Analysis

We conducted a power analysis to determine whether our sample size ($n = 200$ per group across $k = 4$ models) was sufficient to detect the expected effects at the conventional significance level $\alpha = 0.05$. Specifically, power calculations for small, medium, and large effect sizes, based on established conventions (Cohen, 2013), yielded approximate powers of 0.65, 1.00, and 1.00, respectively, for effect sizes $f = 0.10, 0.25$, and 0.40.

Although the power to detect small effects (approximately 0.65) is slightly below the commonly accepted threshold of 0.80, the study is well-powered to identify medium and large effects. This indicates strong sensitivity to differences between models that are of practical significance.

Following this, a one-way ANOVA revealed significant differences in mean semantic preference scores across the four paraphrase models (F = 140.08, $p < 0.001$). Tukey's HSD post-hoc tests confirmed all pairwise comparisons were statistically significant (family-wise error rate 0.05). Key contrasts include:

- VAE significantly outperformed SEPARATOR (mean difference = 5.0, $p < 0.001$), Latent BoW (6.34, $p < 0.001$), and DiPS (9.18, $p < 0.001$).
- SEPARATOR significantly outperformed Latent BoW (1.34, $p = 0.019$) and DiPS (4.18, $p < 0.001$).
- Latent BoW significantly outperformed DiPS (2.84, $p < 0.001$).

These findings statistically support the observed semantic preservation ranking among models.

Inter-rater agreement for the semantic preservation ratings was quantified using Fleiss's κ statistic, yielding a value of 0.539. According to the interpretation ranges of Landis and Koch (1977), this represents "moderate agreement" (0.41–0.60), and also qualifies as "fair to good agreement" (0.40–0.75) as per Fleiss's original guidelines (Fleiss et al., 2013). While these thresholds aid interpretability, we note κ values are context-dependent and may vary according to task and domain.

System	O	R	CV*
VAE	58	57.00	0.63
SEPARATOR	-6	1.44	7.60
Latent BoW	-12	-13.44	1.65
DiPS	-39	-45.00	10.31

Table 3: Original (O) and reproduced (R) semantic preservation scores after subtracting 100 from all values (the original scores were shifted by +100 for CV* calculations). CV* denotes the coefficient of variation, with lower values indicating higher reproducibility.

4.3 Quantified Reproducibility Assessments

To quantify reproducibility of semantic preservation results between the original and reproduced experiments, we applied the four types of reproducibility assessments outlined in the Quantified Reproducibility Assessment (QRA) framework.

4.3.1 Type I: Coefficient of Variation (CV*)

The adjusted coefficient of variation (CV*) was computed for each system's paired original and reproduction mean semantic scores to measure relative variability, accounting for small sample sizes (Belz, 2022). Table 3 presents the CV* values and descriptive statistics.

The notably low CV* for the VAE baseline (0.63) indicates good reproducibility. SEPARATOR and Latent BoW show slightly more variability. DiPS demonstrates the highest variability (CV* = 10.31); nonetheless, this value is just past the threshold for medium degree of reproducibility. The median CV* across all systems is 4.625,[2] indicating a good level of reproducibility overall.

4.3.2 Type II: Correlation Analysis

Pearson's correlation coefficient ($r = 0.99$, $p = 0.01$) and Spearman's rank correlation coefficient ($\rho = 1.00$, $p = 0.00$) between the original and reproduced semantic preservation scores both indicate extremely strong, statistically significant agreement in relative system rankings and absolute scores. This affirms that the reproduction closely matches the original behavioral patterns.

4.3.3 Type III: Agreement Metrics

As the responses from the original experiment are not available, we cannot report inter-annotator agreement metrics (such as Fleiss's κ or Krippendorff's α) as a measure of reproducibility. There-

[2] (1.65 + 7.6) / 2 = 4.625

fore, Type III results regarding inter-rater consistency in semantic preservation judgments are not available for the reproduction. Inter-rater agreement statistics for our own collected data are reported separately in Section 4.2.

4.3.4 Type IV: Side-by-Side Comparison of Findings

The reproduction confirms the original study's conclusions that the VAE baseline outperforms other paraphrase models in semantic preservation, with SEPARATOR occupying a middle ground and Latent BoW and DiPS exhibiting lower semantic fidelity. All primary findings reproduce, reinforcing the robustness of the original experimental conclusions.

5 Discussion

Our quantitative reproducibility analysis results demonstrate successful reproduction of the original human evaluation protocol for assessing semantic preservation in paraphrase generation models. This conclusion is supported by low coefficient of variation (CV*) values, strong correlation coefficients, moderate inter-annotator agreement (as measured by Fleiss' κ), and confirmation of all original findings, collectively indicating a high degree of reproducibility.

Nonetheless, several aspects warrant further exploration. First, although the power analysis was conducted using analysis of variance (ANOVA), linear mixed-effects models (McLean et al., 1991) may be more appropriate for this type of data since the requirements for using analyses of variances are often not met (Boisgontier and Cheval, 2016). Second, the interpretation and acceptable ranges of Fleiss' κ values are context-dependent; different tasks and domains often yield varying κ distributions (Artstein and Poesio, 2008). Currently, clear guidelines for interpreting Fleiss' κ within the context of human evaluation across diverse tasks and settings are lacking.

The automated evaluation of NLG systems, particularly for open-ended and creative tasks, remains an open challenge. Reference-based automated metrics and emerging LLM-based evaluation methods are the two primary approaches in this domain. Consequently, developing automated evaluation frameworks that are both more reliable than existing metrics and more cost-effective than human evaluations would represent a significant advancement (Gilardi et al., 2023). Such frameworks could enable researchers to conduct broader and more systematic evaluations, facilitate robust model comparisons across diverse tasks, and assist practitioners in selecting models best suited to specific applications. Moreover, these frameworks hold promise for supporting continuous model evaluation and monitoring systems that adapt dynamically to evolving requirements and user needs.

Although automated evaluation techniques have advanced, traditional reference-based metrics for open-ended text generation continue to exhibit significant shortcomings. Metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) primarily measure n-gram overlap between generated and human-written texts. While widely used, these metrics frequently fail to capture semantic equivalence and correlate poorly with human judgments (Gaizauskas, 1998; Belz and Reiter, 2006; Reiter and Belz, 2009; Liu et al., 2016; Schluter, 2017; Novikova et al., 2017; Lowe et al., 2017; Post, 2018; van der Lee et al., 2019; Xu et al., 2023; Fabbri et al., 2021; Ernst et al., 2023).

To address these limitations, recent approaches leverage contextual embeddings to better assess semantic similarity. For instance, BERTScore (Zhang et al., 2020) and AlignScore (Zha et al., 2023) use pretrained language model embeddings to evaluate the closeness of generated outputs to references in embedding space. Building on this concept, the LLM-as-a-Judge framework employs large language models directly as evaluators by utilizing their capacity to assess generated texts from multiple perspectives (Zheng et al., 2023; Ashktorab et al., 2024; Hong et al., 2024; Ru et al., 2024; Gilardi et al., 2023). This framework shows promise in aligning well with human judgments; however, challenges remain, including sensitivity to prompts, potential brittleness, and inherent biases within the models (Schroeder and Wood-Doughty, 2024; Thakur et al., 2024).

Given the increasing use of LLMs or foundation models as evaluators, and the growing availability of high-quality human judgment data, including over thirty studies involving human evaluation (Belz and Thomson, 2023, 2024b), it is essential to further investigate the reliability of such models as judgment agents. The availability of this data, combined with detailed annotation of experimental protocols, will facilitate the development of improved recommendations and practical guidelines for evaluation metrics in diverse contexts.

6 Conclusion

This study contributes to advancing reproducibility in NLP human evaluation by successfully reproducing the semantic preservation assessment protocol introduced by Hosking and Lapata (2021). Our reproduction closely matched the original results, confirming the relative semantic fidelity of diverse paraphrase generation models with strong statistical validation and moderate inter-annotator agreement. The application of the Quantified Reproducibility Assessment framework (Belz and Thomson, 2024a; Belz, 2022) provided a multifaceted and quantitative perspective on reproducibility, highlighting areas of strong consistency as well as aspects sensitive to experimental conditions.

Despite this success, challenges remain, including contextual interpretation of agreement metrics, the effects of platform and participant differences, and the necessity for more robust statistical modeling approaches. These results reinforce the need for comprehensive standardization of human evaluation methodologies, detailed and transparent documentation such as the Human Evaluation Datasheet (Shimorina and Belz, 2022), and wider adoption of reproducibility-focused frameworks. Looking forward, integrating improved automated evaluation methods, and particularly those leveraging LLMs (Zheng et al., 2023; Ashktorab et al., 2024), offers promising avenues to complement human judgments and reduce reliance on costly and variable human annotations.

Finally, we encourage the NLP community to embrace collaborative reproducibility initiatives such as the ReproHum Project (Belz and Thomson, 2024a) and to make evaluation data, protocols, and analyses openly accessible. Such collective efforts are crucial to strengthening the scientific rigor and trustworthiness of human evaluation in NLP, thereby accelerating reliable and cumulative progress in the field. Our own data and analysis scripts supporting this reproduction are available in a public repository (Arvan and Parde, 2025).

7 Limitations

While our reproduction study provides valuable insights into the reproducibility of human evaluation for semantic preservation in paraphrasing, some limitations remain. Our study focuses exclusively on a single evaluation criterion: semantic preservation. This reflects a deliberate choice aligned with the ReproHum project's methodology, which emphasizes evaluating one criterion per experiment to keep the experiment manageable for researchers. Although this approach simplifies the evaluation process, it limits the scope of conclusions we can draw about the reproducibility of multi-criteria human evaluations often used in paraphrase assessment, which may behave differently.

Despite this focused scope, our work underscores the importance of meticulous documentation, standardized protocols, and quantitative measures of reproducibility. We hope our findings contribute to the foundation of reproducible human evaluation studies and encourage future research to explore complementary criteria within similarly rigorous frameworks.

8 Ethical Considerations

Our study involves human participants recruited via an online crowdsourcing platform to perform semantic preservation evaluations. We took several measures to ensure ethical standards were upheld throughout the research.

First, all participants were informed about the nature of the task, its purpose, and the approximate time commitment before giving their consent to participate. Participation was entirely voluntary, and workers were free to withdraw at any point without penalty.

Second, we ensured fair and adequate compensation consistent with recommended guidelines for crowdsourcing platforms to respect the participants' time and effort. By providing reasonable payment rates, we aimed to minimize exploitation and support equitable treatment of annotators.

Third, to preserve participant privacy, no personally identifiable information was collected or disclosed. Data collected pertained exclusively to the evaluation task and responses relevant for analysis. Furthermore, we anonymized all data to protect participant identities and maintain confidentiality.

Finally, our reproduction effort emphasizes transparency and reproducibility, which are essential ethical principles in scientific research. By openly sharing data, annotation protocols, and analysis scripts, we promote accountability and facilitate community trust. This study, identified as STUDY2023-1217, was reviewed and deemed exempt by the Institutional Review Board at the University of Illinois Chicago, which ensured that all ethical guidelines were adhered to throughout the research process.

Acknowledgments

We would like to thank the ReproHum project (with special thanks to Craig Thomson) for their support and guidance throughout this reproduction. We would also like to thank the original authors for providing additional information and clarifications. This work was supported by the EPSRC grant EP/V05645X/1.

References

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

Mohammad Arvan and Natalie Parde. 2025. reprohum-0744-02.

Zahra Ashktorab, Michael Desmond, Qian Pan, James M. Johnson, Martin Santillan Cooper, Elizabeth M. Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. 2024. Aligning human and LLM judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. *CoRR*, abs/2410.00873.

Simone Balloccu, Anya Belz, Rudali Huidrom, Ehud Reiter, Joao Sedoc, and Craig Thomson, editors. 2024. *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics.

Anya Belz. 2022. A metrological perspective on reproducibility in NLP. *Comput. 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. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

Anya Belz and Craig Thomson. 2024c. HEDS 3.0: The human evaluation data sheet version 3.0. *CoRR*, abs/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, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, and 20 others. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Matthieu P Boisgontier and Boris Cheval. 2016. The anova to mixed model transition. *Neuroscience & Biobehavioral Reviews*, 68:1004–1005.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Routledge.

DataCanary, hilfialkaff, Lili Jiang, Meg Risdal, Nikhil Dandekar, and tomtung. 2017. Quora question pairs. `https://kaggle.com/competitions/quora-question-pairs`. Kaggle.

Ori Ernst, Ori Shapira, Ido Dagan, and Ran Levy. 2023. Re-examining summarization evaluation across multiple quality criteria. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13829–13838, Singapore. Association for Computational Linguistics.

Timothy M Errington, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021a. Challenges for assessing replicability in preclinical cancer biology. *elife*, 10:e67995.

Timothy M Errington, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021b. Investigating the replicability of preclinical cancer biology. *Elife*, 10:e71601.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1608–1618. The Association for Computer Linguistics.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Robert J. Gaizauskas. 1998. Karen sparck jones and julia galliers, *Evaluating Natural Language Processing Systems: An Analysis and Review*. berlin: Springer-verlag, 1996. ISBN 3 540 61309 9, price DM54.00 (paperback), 228 pages. *Nat. Lang. Eng.*, 4(2):175–190.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *J. Artif. Intell. Res.*, 77:103–166.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. 2024. The hallucinations leaderboard - an open effort to measure hallucinations in large language models. *CoRR*, abs/2404.05904.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1405–1418. Association for Computational Linguistics.

Tom Hosking, Hao Tang, and Mirella Lapata. 2022. Hierarchical sketch induction for paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2489–2501. Association for Computational Linguistics.

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, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 169–182. Association for Computational Linguistics.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Lisa M. Lix, Joanne C. Keselman, and H. J. Keselman. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "f" test. *Review of educational research*, 66(4):579–619.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Robert A McLean, William L Sanders, and Walter W Stroup. 1991. A unified approach to mixed linear models. *The American Statistician*, 45(1):54–64.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2241–2252. Association for Computational Linguistics.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Comput. Linguistics*, 35(4):529–558.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 41–45. Association for Computational Linguistics.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust LLM judgments? reliability of llm-as-a-judge. *CoRR*, abs/2412.12509.

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.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *CoRR*, abs/2406.12624.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in NLP. *Comput. Linguistics*, 50(2):795–805.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 355–368. Association for Computational Linguistics.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3225–3245. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with A unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11328–11348. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.