ReproHum #0866-04: Variability in Human Judgments of Sociopolitical Acceptability Across Studies

Rui Fan[1,2,3] and **Guanyi Chen**[1,2,4*]
[1]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning,
[2]National Language Resources Monitoring and Research Center for Network Media,
[3]School of Chinese Language and Literature, Central China Normal University
[4]School of Computer Science, Central China Normal University
ruifan@ccnu.edu.cn, g.chen@ccnu.edu.cn

Abstract

Human evaluations are essential for assessing NLP systems, but their reproducibility can be limited when judgments involve socially sensitive constructs. This paper reproduces the perceived sociopolitical acceptability evaluation in Gabriel et al. (2022), where annotators judged whether model-generated writer-intent implications reflected mainstream or fringe viewpoints. Using the same 600 headline–belief pairs, we collected new annotations on Prolific and compared our results with both the original study and a prior reproduction. Our scores are lower than the original results. Under a 70% threshold, these findings do not support the original conclusion that most generations were socially acceptable. Overall, our results align more closely with the prior reproduction, while also showing substantial variability, especially for GPT2-large. We argue that this variability may arise from a combination of platform differences, task framing, topic effects, and changes in social context over time. These findings highlight the importance of reporting not only annotation results, but also the evaluation setting in which subjective social judgments are collected.

1 Introduction

Human evaluation plays an important role in Natural Language Processing (NLP), especially when system outputs must be assessed in terms of qualities that are difficult to capture with automatic metrics. At the same time, human evaluation results can be difficult to reproduce because they depend on task design, annotator interpretation, platform settings, and participant populations. This issue becomes particularly important when the evaluated dimension involves social or political judgments rather than purely linguistic properties.

This paper reports a reproduction study conducted in the context of the ReproHum shared task on the reproducibility of NLP evaluations (Belz and Thomson, 2024; Belz et al., 2026). We reproduce one human evaluation dimension from Gabriel et al. (2022), who introduced Misinfo Reaction Frames (MRF), a framework for reasoning about how readers react to news headlines. Their work goes beyond binary misinformation detection by modeling the implications that headlines may evoke, including perceived writer intent, reader perception, reader action, likelihood of spread, and perceived veracity.

Our reproduction focuses on perceived sociopolitical acceptability. In the original study, annotators judged whether the belief or viewpoint invoked by a model-generated writer-intent implication represented a majority or mainstream perspective, as opposed to a minority or fringe perspective. The original study reported acceptability scores above 74% for T5-base, T5-large, and GPT2-large, and concluded that most model generations were socially acceptable. A prior reproduction by Mahlaza et al. (2024) reported lower scores for the same evaluation dimension and argued that the original conclusion depends on how "most" is interpreted. If a 70% threshold is used, the original finding is not reproduced. This motivates our study: we ask whether the perceived sociopolitical acceptability results can be reproduced in a new human evaluation setting, and whether our results align more closely with the original study or with the prior reproduction.

Using the same set of 600 headline–belief pairs, we collected new judgments on Prolific and computed the percentage of generated implications judged to be socially acceptable. To quantify reproducibility, we report CV*, Pearson's r, and Spearman's ρ, and we compare conclusions across the original study, the 2024 reproduction, and our reproduction. Our results show that all three model scores are lower than those in the original study. Under a 70% threshold, our results do not support

*Corresponding Author

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1104–1110

the original conclusion that most generations are socially acceptable. At the same time, our results are closer to the 2024 reproduction, especially for the two T5 models. These findings suggest that perceived sociopolitical acceptability is sensitive to evaluation conditions and that subjective human evaluation dimensions require more detailed reporting and clearer interpretation criteria.

2 Original Study

2.1 Misinfo Reaction Frames

Gabriel et al. (2022) introduce Misinfo Reaction Frames (MRF), a pragmatic framework for reasoning about readers' reactions to news headlines. Instead of only classifying a headline as real or misinformation, MRF aims to capture how readers interpret the headline, what they infer about the writer's intent, how they feel, what actions they might take, and whether they perceive the headline as real or misleading. To support this framework, the authors construct a corpus of news headlines from domains such as COVID-19, climate change, and cancer, and train language models, including T5 and GPT2 variants, to generate reaction-frame implications from headlines.

2.2 Human Evaluation

The original study includes human evaluation of generated writer-intent implications. The headlines used in the study were drawn from published misinformation-related datasets covering COVID-19 (Cui and Lee, 2020; Nørregaard et al., 2019a; Shapiro et al., 2020; Network, 2024), climate change (Nørregaard et al., 2019a,b), and cancer (Cui et al., 2020). Based on these headlines and associated domain information, the authors trained models to generate writer-intent implications using pretrained language models, including T5 (Raffel et al., 2020) and GPT2 (Radford et al., 2019).

Annotators were asked to judge generated implications along several dimensions, including overall quality, influence on trust, and perceived sociopolitical acceptability. The present reproduction focuses specifically on the perceived sociopolitical acceptability evaluation. In this task, annotators assessed whether the belief or viewpoint invoked by a generated implication represented a majority or mainstream perspective, as opposed to a minority or fringe perspective. The original study clarifies that "minority" refers to less commonly adopted or more extreme social beliefs, rather than viewpoints held by historically marginalized groups.

The original study reports the percentage of generated implications judged to be socially acceptable as 75.30% for T5-base, 74.66% for T5-large, and 74.66% for GPT2-large. Based on these values, the authors conclude that most model generations were rated as socially acceptable.

3 Prior Reproduction Study

Mahlaza et al. (2024) reproduced the human evaluation of Gabriel et al. (2022). Their study was conducted through Prolific[1] and LimeSurvey[2]. They used a dataset of 600 headline–intent pairs and divided the task into batches to reduce participant fatigue.

Their reproduction covered multiple human evaluation dimensions, including quality, influence on trust, sociopolitical acceptability, and potential perpetuation of negative stereotypes. For the sociopolitical acceptability dimension, they reported scores of 68.67% for T5-base, 68.31% for T5-large, and 65.30% for GPT2-large. These values are lower than those reported in the original study.

Based on these results, Mahlaza et al. (2024) argue that the original conclusion that most generations are socially acceptable depends on how "most" is interpreted. If a threshold of 70% is adopted, the finding is not reproduced. If the threshold is relaxed to an above-chance criterion, the finding can be considered partially supported.

This prior reproduction is important for our study because it provides an intermediate comparison point. Rather than comparing only our results with the original study, we compare results across three studies to assess whether the lower acceptability scores observed in the prior reproduction also appear in our reproduction.

4 Reproducibility Study Design

4.1 Scope of the Reproduction

This study focuses on reproducing the perceived sociopolitical acceptability evaluation reported by Gabriel et al. (2022). We selected this dimension because it relies on subjective social judgments and because a prior reproduction found lower acceptability scores than the original study. Our reproduction therefore targets this specific human evaluation component, rather than the full MRF modeling pipeline or all evaluation dimensions. We set up

[1] https://www.prolific.com/

[2] https://survey.cs.uct.ac.za/limesurvey/

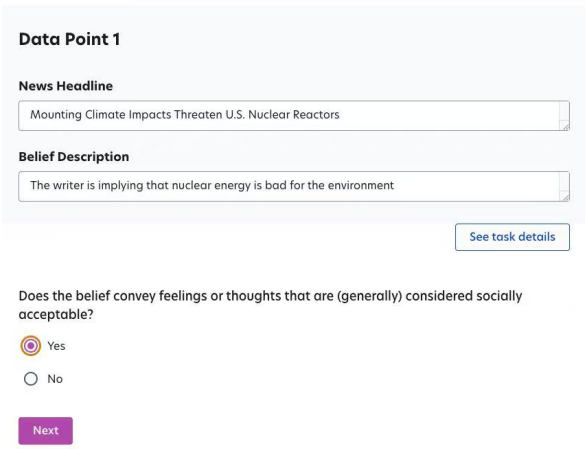

Figure 1: Screenshot of the annotation interface used in our reproduction study.

the experiment using all information available from the original paper (Gabriel et al., 2022) and from follow-up communications with the authors by the ReproHum leadership team.

4.2 Data

The evaluation data consists of 600 headline–belief pairs, the same dataset used in the original study and the prior reproduction. Each item contains a news headline and a generated belief description. The generated implications are associated with three model conditions: T5-base, T5-large, and GPT2-large, with 200 items for each model. Although the original authors reportedly excluded 12 of the 600 items post hoc because they were malformed or otherwise unsuitable, we were unable to identify which specific items were excluded from the available materials. Therefore, our reproduction uses the full set of 600 headline–belief pairs.

4.3 Annotation Task

Annotators were shown a news headline and a belief description. They were asked to judge whether the belief represents a mainstream viewpoint or a fringe viewpoint. Following the original study, we interpret mainstream judgments as socially acceptable.

Figure 1 shows a screenshot of the annotation interface used in our reproduction. The interface presented each headline together with the generated belief description and asked participants to judge whether the belief reflected a mainstream or fringe viewpoint.

Model	Original	Ours
T5-base	75.30%	66.17%
T5-large	74.66%	67.33%
GPT2-large	74.66%	58.67%

Table 1: Perceived sociopolitical acceptability scores in the original experiment and our reproduction experiment.

4.4 Participants and Platform

We conducted the evaluation using Prolific. Participant language proficiency was controlled through region-based recruitment. Eligible participants were required to have nationality from Australia, Canada, the United Kingdom, or the United States. In addition, participants were required to have a minimum acceptance rate of 99% and at least 200 previously completed tasks on Prolific.

Each item was judged by 3 annotators. Since multiple annotations were collected for each item, we aggregated judgments by computing the proportion of socially acceptable labels. Model-level scores were then obtained by calculating the percentage of socially acceptable judgments for each model condition.

4.5 Evaluation Strategy

We compare our results with the original study and the 2024 reproduction. The main metric is the percentage of generated implications judged to be socially acceptable. To quantify reproducibility, we compute three forms of Quantified Reproducibility Assessments (Belz, 2025). We calculate CV*, the small-sample-adjusted coefficient of variation proposed by Belz (2022), to measure absolute variability between the original and reproduced per-system scores.

5 Results

We first compare the perceived sociopolitical acceptability scores reported in the original study with those obtained in our reproduction. As shown in Table 1, the original study reports scores of 75.30% for T5-base, 74.66% for T5-large, and 74.66% for GPT2-large. In our reproduction, the corresponding scores are 66.17%, 67.33%, and 58.67%, respectively. Thus, our results are consistently lower than those reported in the original study. If the conclusion that "most generations are socially acceptable" is interpreted using a threshold

Type of Result	QC	System	Measure applied	Degree of reproducibility ($n = 2$)		
				System level	QC level	Study level
Type I	QC1	T5-base T5-large GPT2-large	(mean) CV*↓	11.44 9.15 21.26	13.95	13.95
Type II	QC1	all all all	mean r ↑ mean ρ ↑ W↑	n/a n/a n/a	0.389 0.000 0.000	n/a
Type IV	QC1	all	P ↑	n/a	0.333	0.333

Table 2: QRA reproducibility assessment comparing the original study (Gabriel et al., 2022) with our reproduction (n=2). QC1 = Socially acceptable; n/a = measure does not apply at this level.

of at least 70%, then none of the three models in our reproduction supports the original conclusion. The two T5 models fall slightly below this threshold, while GPT2-large falls substantially below it.

The statistical significance of the differences in acceptance rates across models was further examined. We first applied Cochran's Q test to assess the overall difference among the three models. The result did not reach the conventional significance threshold, although it showed a marginal trend ($Q = 5.4286, p = .0663$). This suggests that while the models may differ in their acceptance rates descriptively, the overall evidence is insufficient to conclude a statistically significant difference among them.

We then conducted pairwise comparisons using McNemar's test, followed by Holm correction to account for multiple comparisons. After correction, none of the pairwise differences remained statistically significant: GPT2-large vs. T5-large ($p = .1704$), GPT2-large vs. T5-base ($p = .2177$), and T5-base vs. T5-large ($p = .7660$). Therefore, although the descriptive results indicate that GPT2-large had the lowest acceptance rate and T5-large the highest, these differences should be interpreted as non-significant under the current sample and annotation setting.

The CV* values in Table 2 show moderate variability for T5-base and T5-large, but substantially higher variability for GPT2-large. This indicates that the reproduced acceptability scores for the T5 models are closer to the original values, whereas the GPT2-large score is much less stable across the two studies. The higher variability observed for GPT2-large may also help explain why the statistical tests reported above did not identify significant differences among models: although GPT2-large has the lowest descriptive acceptance rate, its greater instability increases uncertainty around the estimate, making it harder to establish a reliable model-level difference.

Model	Rep2024	Ours
T5-base	68.67%	66.17%
T5-large	68.31%	67.33%
GPT2-large	65.30%	58.67%

Table 3: Comparison between the 2024 reproduction and our reproduction.

6 Discussion

6.1 Comparison with Prior Results

The original study reports that most model generations are socially acceptable, with scores above 74% for all three models. Our results do not support this conclusion under a 70% threshold: both T5-base and T5-large fall slightly below 70%, while GPT2-large falls much lower at 58.67%. This pattern suggests that the original study may have overestimated the sociopolitical acceptability of generated implications, or that this judgment is sensitive to evaluation conditions. Since sociopolitical acceptability is a subjective construct, differences in annotator populations, platform settings, or task presentation may affect the results.

Our results are closer to the 2024 reproduction (Mahlaza et al., 2024) than to the original study. As shown in Table 3, the T5-base and T5-large scores in our reproduction are close to those reported in the 2024 reproduction. By contrast, GPT2-large shows a larger descriptive difference between the two reproductions. This suggests that the lower acceptability scores for the T5 models are relatively consistent across reproduction studies, whereas GPT2-large may be more sensitive to evaluation conditions.

Type of Result	QC	System	Measure applied	Degree of reproducibility ($n = 2$)		
				System level	QC level	Study level
Type I	QC1	T5-base T5-large GPT2-large	(mean) CV*↓	3.29 1.28 9.48	4.68	4.68
Type II	QC1	all all all	mean r ↑ mean ρ ↑ W↑	n/a n/a n/a	0.976 0.500 0.333	n/a
Type IV	QC1	all	P ↑	n/a	0.667	0.667

Table 4: QRA reproducibility assessment comparing the 2024 reproduction (Mahlaza et al., 2024) with our reproduction (n=2).

Category	GPT2-large	T5-base	T5-large
COVID	53.42%	62.82%	69.23%
Climate	66.28%	72.80%	68.20%
Cancer	51.43%	57.14%	60.95%

Table 5: Acceptance rates by rough keyword-based topic category.

6.2 Topic Effects and GPT2-large Variability

The above results suggest that GPT-2 Large is an exception in our reproduction. To further explore the reason for this pattern, we examined whether model-level differences were associated with topic domains. Since the original dataset does not provide explicit topic labels, we roughly divided the examples into three categories using keyword matching: COVID, Climate, and Cancer. [3] Table 5 reports the acceptance rates of the three models within each category.

The results suggest that the lower acceptance rate of GPT2-large is not uniform across topics. In the Climate category, GPT2-large performs relatively close to the T5 models. By contrast, the gap is much larger in the COVID category, where GPT2-large obtains an acceptance rate of 53.42%, compared with 62.82% for T5-base and 69.23% for T5-large. This pattern suggests that GPT2-large may be especially unstable on COVID-related examples.

This observation is also consistent with the qualitative finding that GPT2-large more frequently generates vague or overly general terms, such as "bad", rather than producing specific inferences grounded in the input sentence. Such vague wording may be particularly problematic for COVID-related content. Since the annotations in the reproduced study were collected in 2026, several years after the peak of the COVID-19 pandemic, annotators may no longer share the same social context, assumptions, or risk perceptions that were present when COVID-related discourse was more salient. As a result, broad or underspecified inferences about COVID may be interpreted less consistently across annotators, increasing variability in the acceptance judgments. This may help explain why GPT2-large shows both the lowest descriptive acceptance rate and relatively high variability, while the overall statistical tests did not find significant model-level differences.

6.3 Ambiguity of "Most" and Threshold Selection

A central issue is the interpretation of the phrase "most generations were rated as socially acceptable." The original study does not specify a formal threshold for determining "most." The 2024 reproduction uses 70% as a reasonable threshold and notes that conclusions change if the threshold is relaxed to 50%. Our results reinforce this concern. Under a 70% threshold, none of the models supports the original conclusion. Under a weaker threshold of above 50%, all models would still be considered more acceptable than unacceptable. This illustrates how threshold choice can affect reproducibility conclusions.

6.4 Platform and Task Effects

Differences across studies may arise from both annotation platforms and task framing. The original study used MTurk and evaluated generated writer-intent implications using qualified workers from the same pool as the original data annotation, but it does not provide enough detail for us to exactly reproduce the participant selection procedure. In our reproduction, we instead followed the available ReproHum guidance and specified explicit re-

[3] This categorization should therefore be interpreted as an approximate diagnostic analysis rather than a gold-standard domain annotation.

cruitment restrictions for our Prolific study. Since perceived sociopolitical acceptability depends on cultural background, political attitudes, media literacy, and familiarity with misinformation topics, differences in platform, recruitment criteria, and participant composition may reasonably lead to different judgments across studies.

Task effects may also have contributed to the differences between studies. In the original study, perceived sociopolitical acceptability was evaluated as part of a broader human evaluation of generated writer-intent implications. Annotators also assessed overall quality, whether the implication made the headline seem more or less trustworthy, and, in an A/B setup, whether revealing the implication changed the perceived trustworthiness of the headline. By contrast, our reproduction focused only on the mainstream-versus-fringe sociopolitical acceptability judgment. This difference may change how annotators interpret the task: when acceptability is evaluated alongside quality and trust-related judgments, annotators may separate whether a generation is coherent, relevant, influential, or socially mainstream, whereas an isolated acceptability task may make the sociopolitical criterion more salient and encourage stricter judgments.

Overall, our findings suggest that subjective human evaluation dimensions require careful reporting and clearer evaluation criteria. The task studied here is not a purely linguistic acceptability judgment; it requires annotators to assess whether a belief is socially mainstream or fringe, and such judgments are deeply context-dependent. Future studies should therefore provide detailed task instructions, examples, platform information, participant eligibility criteria, annotator demographics where appropriate, compensation details, aggregation methods, decision thresholds, and whether the target dimension was evaluated in isolation or together with other evaluation dimensions.

7 Conclusion

We reproduced the perceived sociopolitical acceptability evaluation from Gabriel et al. (2022) and compared our results with both the original study and a prior reproduction (Mahlaza et al., 2024). Our scores are lower than those reported in the original study, with 66.17% for T5-base, 67.33% for T5-large, and 58.67% for GPT2-large. Under a 70% threshold, our results do not support the original conclusion that most generated implications were socially acceptable. Instead, they align more closely with the prior reproduction, suggesting that sociopolitical acceptability judgments are sensitive to evaluation conditions and participant populations.

Limitations

This study reproduces only one human evaluation dimension, perceived sociopolitical acceptability, rather than the full MRF modeling pipeline or all evaluation tasks. We used the full set of 600 headline–belief pairs because the 12 items reportedly excluded in the original study could not be identified from the available materials. In addition, the original study does not provide detailed participant restriction criteria, so we could not exactly reproduce its participant selection procedure. Finally, our correlation analyses are based on only three model-level scores, and annotation-level agreement with the original study could not be computed because the original item-level judgments were unavailable.

Ethical Considerations

This study involves human judgments of sociopolitical acceptability. Such judgments may reflect annotators' personal, cultural, or political perspectives. We therefore avoid interpreting individual annotations as objective truth. Instead, we treat them as subjective evaluations collected under a specific experimental setup.

The evaluation data concerns misinformation-related headlines and generated beliefs. Some items may contain misleading, controversial, or socially sensitive content. Participants should be informed about the nature of the task before participation.

Acknowledgements

We thank the ReproHum organizers for providing the study materials and funding for this reproduction. The authors are supported by the MOE Project of Humanities and Social Sciences, China (Project No. 25YJC740005), the National Language and Character Research Base (Project No. ZDI145-168), and the Fundamental Research Funds for the Central Universities, Academy of Frontier Interdisciplinary Research, Central China Normal University (Project No. JC2026PT-004).

References

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

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

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

Anya Belz, Craig Thomson, and Javier González Corbelle. 2026. The shared task on reproducibility of evaluations in nlp (ReproNLP) 2026: Overview and results. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM²)*, San Diego, USA. Association for Computational Linguistics.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. KDD '20, page 492–502, New York, NY, USA. Association for Computing Machinery.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland. Association for Computational Linguistics.

Zola Mahlaza, Toky Hajatiana Raboanary, Kyle Seakgwa, and C. Maria Keet. 2024. ReproHum #0866-04: Another evaluation of readers' reactions to news headlines. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 274–280, Torino, Italia. ELRA and ICCL.

International Fact-Checking Network Network. 2024. Fighting the infodemic: The coronavirusfacts alliance. [Online; accessed 10-04-2026].

Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. 2019a. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):630–638.

Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. 2019b. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. volume 13, pages 630–638.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Jacob N. Shapiro, Jan Oledan, and Samik-shya Siwakoti. 2020. Fighting the info-demic: The coronavirusfacts alliance. [Online; accessed 10-04-2026].