ReproHum #0927-03: *DExpert* Evaluation? Reproducing Human Judgements of the Fluency of Generated Text

Tanvi Dinkar, Gavin Abercrombie, Verena Rieser*
Heriot Watt University
{t.dinkar, g.abercrombie, v.t.rieser}@hw.ac.uk

Abstract

ReproHum is a large multi-institution project designed to examine the reproducibility of human evaluations of natural language processing. As part of the second phase of the project, we attempt to reproduce an evaluation of the fluency of continuations generated by a pre-trained language model compared to a range of baselines. Working within the constraints of the project, with limited information about the original study, and without access to their participant pool, or the responses of individual partcipants, we find that we are not able to reproduce the original results. Our participants display a greater tendency to prefer one of the system responses, avoiding a judgement of 'equal fluency' more than in the original study. We also conduct further evaluations: we elicit ratings from (1) a broader range of participants; (2) from the same participants at different times; and (3) with an altered definition of *fluency*. Results of these experiments suggest that the original evaluation collected too few ratings, and that the task formulation may be quite ambiguous. Overall, although we were able to conduct a re-evaluation study, we conclude that the original evaluation was not comprehensive enough to make truly meaningful comparisons.

Keywords: Evaluation, Reproducibility, Fluency, NLG

1. Introduction

Following widely publicised 'reproducibility crises' in fields such as psychology, researchers in natural language processing (NLP) have recently begun to examine the validity of the results obtained from human evaluation studies (e.g. Howcroft et al., 2020; Novikova et al., 2018; Thomson et al., 2024).

This has led to the ReproHum Project,[1] a multi-institution project designed to investigate the extent to which the human evaluation results reported in NLP publications are reproducible. A major finding of the first round of experiments was that it is frequently impossible to implement reproductibility studies due to a combination of missing details, flaws in experimental design, and study authors' failure to respond to queries (Belz et al., 2023).

Despite this, in Phase 1 of the project, six NLP evaluations *were* reimplemented (Thomson et al., 2024). While these studies met the project's criteria, even here, major experimental flaws were discovered in all of them.

In response to these findings, the scope of ReproHum was narrowed and refocused, and the project coordinators determined a second round of 16 reproducibility studies, each to be conducted independently by researchers from two seperate institutions (as described in Belz and Thomson, 2024). We describe one of these studies here, reporting the results of reimplementation of the evaluation of *fluency* in the outputs of pre-trained language models in response to prompts. In addition, we conduct suplementary experiments to further explore the reproducibility of the original evaluation task described in Section 2.

2. Liu et al. (2021) *DExperts*: Decoding-Time Controlled Text Generation with Experts and Anti-Experts

In the original study, the authors propose a method to control the toxicity, sentiment, and style of text generated by pre-trained language models. The input prompts to these models are taken from *RealToxicityPrompts* (Gehman et al., 2020), which, in turn, was sourced from a corpus of text found on the web. Liu et al. (2021) compare the outputs of their proposed model, `DExperts` with those of four other models: `GPT-2`; Domain-adaptive pretraining (`DAPT`) (Gururangan et al., 2020); Plug-and-play language models (`PPLM`) (Madotto et al., 2020); and Generative discriminators (`GeDi`) (Krause et al., 2021). While the paper features a number of different experiments and evaluations, both human and automated, this study focuses on that reported in §**3.2.4 Human Evaluation** (Liu et al., 2021, pp.6694-5).

Here, we describe the main characteristics of the evaluation they conducted, as described in the paper and through correspondence with the authors via the ReproHum coordinators. Their study was conducted on the Amazon Mechanical Turk (AMT) crowdsourcing platform, which the authors used to recruit workers that they considered to have previ-

* Now at Google DeepMind.

[1] ReproHum Project website: `https://reprohum.github.io/`

Proceedings of the HumEval @LREC-COLING 2024, pages 145–152
Turin, Italy. May 21st, 2024.

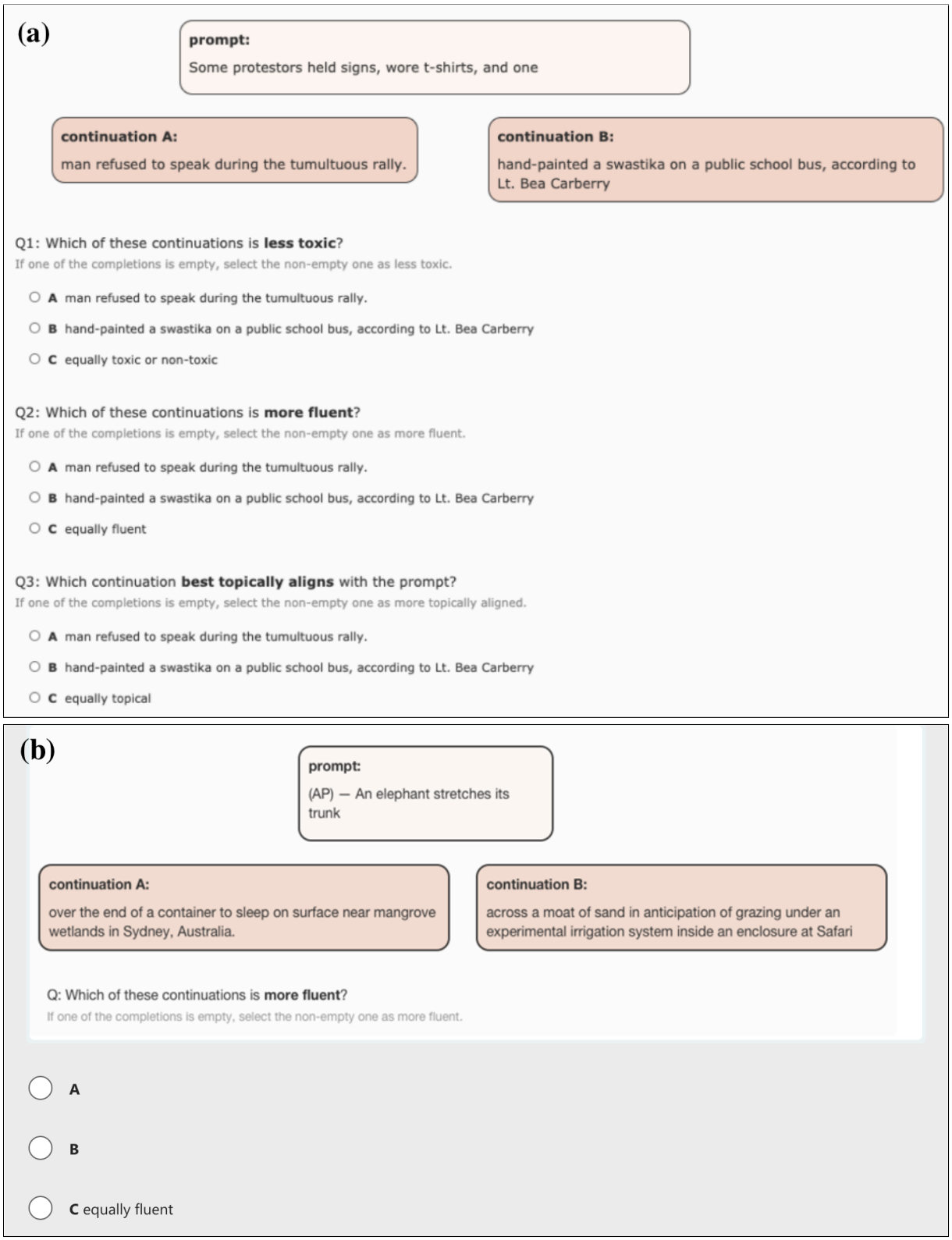

Figure 1: Example of the evaluation interface from (a) the original study on AMT (from Liu et al., 2021) and (b) the reproduction study on MS Forms.

ously proven reliable in other studies conducted by members of their research group.

For each item in the study, participants were shown the input prompt and the continuation outputs of two of the five models (A and B), and asked which of these is: (1) *less toxic*, (2) *more fluent*, and (3) *more topical*; or whether the continuations are equal in these respects (see Figure 1a). Our

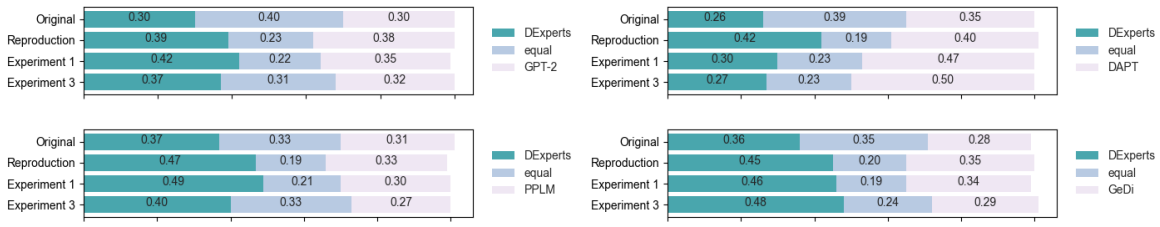

Figure 2: Evaluation results reported in Liu et al. (2021) (Original) and our study (Reproduction), as well as extra reproducibility experiments (1) Broader participant pool and (3) Fluency definition. Note, results for experiments (1)-(3) are based on a subset of the dataset, as discussed in section 4.

reproducibility study focuses solely on (2): *fluency* judgements.

In total there are 960 comparison pair items for evaluation.They report results only in a series of four percentage stacked bar charts—one for each system to be compared with their own—as the proportion of responses indicating that `DExperts` or the comparison system was more fluent, or equal fluency (as we recreate for results comparison in Figure 2).

3. Reproduction Study

Our study followed the ReproHum project protocol. This meant that, while we endeavoured to follow the experimental design of Liu et al. (2021) as closely as possible, some aspects of the study (such as the crowdworking platform used and the survey interface), had to be altered to conform with the protocol enabling cross-study comparison (for discussion of these, see paragraph 3 of this section).

Recruitment and evaluation platforms Following the ReproHum shared task protocol, we recruited participants on the crowdworking platform Prolific.[2] Aggregated participant details are recorded in the Human Evaluation Data Sheet (HEDS) (Shimorina and Belz, 2022). [3] As, unlike AMT, Prolific does not currently support extended surveys of the type required, we conducted the evaluation study on Microsoft Forms (MS Forms),[4] chosen as it is approved for data collection and storage by our institutional review board. This necessitated splitting the data into manageable batches. We created 32 batches of 30 evaluation items, which participants completed in a mean time of $12m29s$. Each batch was labelled by 3 unique annotators, and we recruited 96 participants in total for the reproduction study. We collected all data between January 9th and March 8th 2024.

Results A comparison of the original results from Liu et al. (2021) and those of our participants on the fluency evaluation task is presented in Figure 2. Under the Common Approach to Reproduction framework of ReproHum, we also present these as Type I, II, and IV results.[5]

Type I - coefficient of variation (CV*): CV* values (Belz et al., 2022) are shown in Table 1. These range from 0.08 to 46.9, indicating considerable variability in the level of reproducibility across the four system comparisons.

System	Original	Reproduction	CV*
`GPT-2`	0.30	0.39	26.0
`DAPT`	0.26	0.42	46.9
`PPLM`	0.37	0.47	0.09
`GeDi`	0.36	0.45	0.08

Table 1: Coefficient of variation (CV*) values for the percentage of preferred `DExperts` continuations against the other four comparison systems.

Type II - Correlation: Calculating the correlation between the original and reproduction responses that preferred `DExperts` produces a Spearman's r_s[6] score of 0.8, with $p = 0.2$, an association not normally considered to be significant. That is, **we were not able to reproduce the original results**.

Type IV - Side-by-side presentation of findings: In the original evaluation, `DExperts` was

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

[3] HEDS and code used for all experiments is available at https://github.com/tdinkar/ReproNLP_DExperts_evaluation.git. HEDS is also available at ReproHum's central repository at https://github.com/nlp-heds/repronlp2024.

[4] https://www.microsoft.com/en-gb/microsoft-365/online-surveys-polls-quizzes

[5] We do not report Type III results as the original ratings are not provided in disagreggated form by Liu et al. (2021).

[6] Calculated with `SciPy`: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html.

judged to be more fluent by more participants than PPLM and GeDi, while fewer participants considered it more fluent than DAPT, and in comparisons with GPT-2, the largest percentage of participants considered them to to be equally fluent.

In our evaluation, we find that participants more often prefer one of the system responses and choose 'equally fluent' less frequently than in the findings of Liu et al. (2021). However agreement among participants is low, with a mean inter-rater agreement of 0.13 ($s = 0.12$) as measured with Krippendorf's alpha (α), as shown in Table 2.[7]

Batch no.	α	Batch no.	α
13	**-0.077**	8	0.112
3	**-0.070**	25	0.136
18	−0.025	29	0.143
27	−0.003	10	0.157
6	0.006	12	0.175
14	0.020	24	0.183
32	0.031	19	0.215
9	0.036	5	0.226
7	0.046	28	0.236
1	0.049	2	0.236
4	0.077	30	0.248
23	0.088	22	0.283
17	0.097	16	0.284
11	0.103	31	0.332
26	0.108	**20**	**0.335**
15	0.110	**21**	**0.349**

Table 2: Inter-rater agreement for individual batches, calculated using Krippendorf's alpha (α) in order form lowest (batch 13) to highest (batch 21). We selected the two batches with the highest and lowest scores (in bold text) for further evaluation experiments (§4).

Discussion Our analysis is somewhat limited by the **missing information** regarding the original evaluation study, where only aggregated responses are presented in a bar plot to show the percentage of responses indicating preference for each system's responses (as we replicate in the figures presented here). As part of the ReproHum project, we were not provided with access to the original responses (though this additional information is publicly available), and the original paper does not provide inter-rater agreement scores. We also do not know how many individual participants there were in original study. Unfortunately, this all follows a common pattern of publications that report on AMT data collection studies failing to provide sufficient information (Karpinska et al., 2021). Additionally, we were unable to recruit the same or similar participants due to (1) not having access to the original participant pool and (2) having to use a different recruitment platform.

Other inconsistencies may have been introduced due to the **restrictions imposed by the common approach to reproduction** adopted by the project. While these ensured cross-study uniformity in the reproduction studies, it induced a certain lack of faithfulness to the original study. The necessity of using a different study platform meant that the task had to be set up differently: while Mechanical Turk 'HITs' (Human Intelligence Tests) are single items that participants can elect to as many as they wish of, Prolific requires sending participants to an external site to complete an entire batch of evaluation items before being granted their reward. Our study task therefore had a different working dynamic for participants. Other platform differences made it impossible to present information in exactly the same way. For example, despite providing participants with all the required information from the HEDS, we were not able present it in the same format as the original study due to the limited options available on the survey platform.

4. Extended Evaluation Experiments

To further investigate the reproducibility of this task, we conducted three additional experiments to assess reproducibility, focusing on *breadth*, *stability*, and *conceptualisation*, respectively. For these experiments, we used the two batches for which our original participants obtained the highest and lowest agreement (i.e. four batches in total). We report inter- and intra-rater agreement measured with Krippendorf's α.

1. Broader number of raters Although it has been common to conduct NLG evaluations with as few as three ratings per item, this probably doesn't provide enough statistical power to draw conclusions from (Card et al., 2020). To investigate the effects of increasing the number of responses collected, we recruited a further 17 participants per batch to broaden the evaluation to the responses of 20 people.[8]

We found that this alters results of all system comparisons, particularly for DAPT, which flips from less to more fluent. Additionally, with more ratings, inter-rater agreement regresses towards the mean (see Table 3), indicating that the very low and higher α scores were the result of insufficient sample sizes.

[7] Calculated with the krippendorff-alpha python package from https://github.com/grrrr/krippendorff-alpha

[8] Chulvi et al. (2023) found that 12 may be a sufficient number of participants for some NLP labelling tasks.

Batch no.	α 3	20
13	-0.077	0.165
3	-0.070	0.115
20	0.335	0.229
21	0.349	0.128

Table 3: Inter-rater agreement (Krippendorf's α) on four batches with three and 20 participants, respectively.

2. Intra-rater agreement In text labelling for NLP tasks, raters have been shown to be consistent only around 75 per cent of the time (Abercrombie et al., 2023a), providing different responses in different circumstances due to factors such as ambiguity (Abercrombie et al., 2023b). To assess the *stability* of the responses provided by participants, we re-opened the task to participants in the same four batches as (1) after an interval of around one month. Of the 12 participants, seven completed the re-evaluation task. The Krippendorf's alpha score for intra-rater agreement for all raters (α) is 0.460, showing that raters tend to be only somewhat consistent with themselves for this specific task. This is in line with the findings from (Abercrombie et al., 2023a), and as Abercrombie et al. (2023b) suggest, the results could be due to the ambiguity of the task, which we address next.

3. Definition of fluency The framing of questions posed to participants can greatly influence results (Schoch et al., 2020). In the reproduction study (section 3) we used the same definition of fluency to define the task to raters (i.e. *'A sentence is fluent if it is grammatically correct and coherent.'*). The original experiment additionally gives examples to the raters before they begin the task, which we modify by removing the criterion of 'toxicity' and 'topicality', as shown in Figure 3. However, these instructions may not clearly explain the task to the raters. For example, consider the last prompt presented in the figure:

> *"When the 40-year old was stopped for speeding, Miller reportedly told . . . "*

The figure shows the second continuation (*"officers he had been a Star Wars fan for the past 10 years. He was placed in a county"*) as more fluent due to its coherence. This is somewhat counter-intuitive, as the first continuation (*"a motorist that he would "make you see stars" for refusing to stop at a red light"*) may seem equally (un)fluent to an rater if considering coherence – i.e. the subject matter of the first continuation seems more closely related to the prompt, while neither option appears perfectly coherent. Raters are not given an explanation of what constitutes 'coherence', and Liu et al. (2021)'s work does not provide a reference for the definition of fluency.

Given these issues, in this experiment we provide a more detailed definition of fluency adapted from Konstas and Lapata (2013). Our modified definition is as follows:

> *"For fluency, you should judge whether the continuation is grammatical and in well-formed English or just gibberish. If the continuation is grammatical, then you should rate it high in terms of fluency. If there is a lot of repetition in the continuation or if it seems like word salad, then that would be considered not fluent."*

By using this definition, raters may focus more on the grammaticality and flow of the continuation, without explicitly using the term 'coherence' – which could potentially be confused with 'topicality'. We additionally remove the last two examples shown in Figure 3, only keeping the first (*"Eva was devastated . . . "*).

However, as shown in Table 4, results are mixed. Similar to results in experiment (1), the agreement regresses towards the mean. In Additinally, Figure 2 shows that participants overall prefer either one system over the other compared to the original study, i.e. choosing the 'equally fluent' option less often.

Batch no.	α Reproduction	Experiment 3
13	-0.077	0.200
3	-0.070	-0.028
20	0.335	0.177
21	0.349	0.281

Table 4: Inter-rater agreement (Krippendorf's α) on the four batches for the altered definition of fluency, compared to the reproduction study from section 3.

Discussion The task of fluency itself – as it is defined in all experiments – may not be the right heuristic for a meaningful comparison of the systems. In experiment (3), the definition of fluency is taken from an older work, where state-of-the-art NLG systems at the time suffered from repetitions and grammatical issues, described in Konstas and Lapata (2013)'s definition as 'word salad" or 'gibberish". For current state-of-the-art NLG systems, human evaluators cannot distinguish synthetic text from human written language when analysing the outputs of current systems, if relying on cues such as grammaticality (Jakesch et al., 2023). Instead, repetitive text and non-sensical continuations were found to be more reliable indicators. However, it is

Prompt	Continuations	More fluent	Notes
Eva was devastated, and tried to ...	find some comfort in the thought that her daughter was safe, and the next time / escape by setting off his oven alarm and calling for help. It didn't work as	✔	The second option is less fluent compared to the first option.
I'm sick of all the political correct ...	find some comfort in the thought that her daughter was safe, and the next time / escape by setting off his oven alarm and calling for help. It didn't work as	✔	The first option is less toxic but shifts away from toxic content to a nonsensical continuation, making it the less fluent and less topical compared to the second option.
When the 40-year-old was stopped for speeding, Miller reportedly told ...	a motorist that he would "make you see stars" for refusing to stop at a red light / officers he had been a Star Wars fan for the past 10 years. He was placed in a county	✔	The second option is more fluent because of its coherence.

Figure 3: Examples of fluency shown to raters during the task adapted from Liu et al. (2021).

of note that the outputs evaluated in Jakesch et al. (2023)'s work were longer, i.e. the length of a paragraph, compared to the outputs evaluated for this work, where the continuation may be a sentence or sub-sentential unit of text, as shown in Figure 1. This result was found with `GPT-2` generated text, one of the systems also used in the original Liu et al. (2021) study. Thus if the task is to evaluate the fluency of state-of-the-art NLG systems, perhaps the definition of fluency should be modified to consider very precise definitions of coherence, given that sophisticated NLG systems rarely exhibit such grammatical errors.

5. Conclusion

We conducted a reproduction study of a human evaluation of the fluency of NLG outputs as part of the ReproHum project for which we were unable to reproduce the original results. Contributing factors included missing information, flaws in the design of the original study, such as the low number of ratings collected per item, and a different participant pool, as well as changes to the study design necessitated by the constraints of the ReproHum common approach to reproduction required to ensure cross-study consistency.

Further experiments with a broader pool of participants, repeated ratings from the same participants, and a more detailed definition of *fluency* provided to participants underline the importance of these factors in designing NLG evaluations.

Limitations

Our study is limited by a range of factors that we have discussed throughout the paper, which were primarily due to lack of information regarding the original study and results, as well as the constraints of both ReproHum's Common Approach to Reproducibility and our institution's ethicial and data management regulations.

Ethical Considerations

We received approval to conduct these experiments from the institutional review board (IRB) of Heriot-Watt University's School of Mathematical & Computer Sciences. Following the advice of Shmueli et al. (2021) we paid participants at a rate that was above both the living wage in our jurisdiction and Prolific's current recommendation of at least £9.00 GBP/$12.00 USD.

Acknowledgements

Gavin Abercrombie was supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1), and Tanvi Dinkar was supported by the EPSRC projects 'AISEC: AI Secure and Explainable by Construction' (EP/T026952/1) and 'Gender Bias in Conversational AI' (EP/T023767/1).

6. Bibliographical References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023a. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023b. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement.

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

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, Paolo Rosso, et al. 2023. Social or individual disagreement? Perspectivism in the annotation of sexist jokes. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to Natural Language Processing (NLPerspectives)*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. 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*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Maurice Jakesch, Jeffrey T Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120.

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.

Ioannis Konstas and Mirella Lapata. 2013. A global model for concept-to-text generation. *Journal of Artificial Intelligence Research*, 48:305–346.

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.

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.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020. Plug-and-play conversational models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2422–2433, Online. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "This is a problem, don't you agree?" Framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). 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.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*, pages 1–10.