ReproHum #0927-3: Reproducing The Human Evaluation Of The DExperts Controlled Text Generation Method

Javier González-Corbelle, A. Vivel-Couso, J.M. Alonso-Moral, A. Bugarín-Diz
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
{j.gonzalez.corbelle, ainhoa.vivel.couso, josemaria.alonso.moral, alberto.bugarin.diz}@usc.es

Abstract

This paper presents a reproduction study aimed at reproducing and validating a human NLP evaluation performed for the DExperts text generation method. The original study introduces DExperts, a controlled text generation method, evaluated using non-toxic prompts from the RealToxicityPrompts dataset. Our reproduction study aims to reproduce the human evaluation of the continuations generated by DExperts in comparison with four baseline methods, in terms of toxicity, topicality, and fluency. We first describe the agreed approach for reproduction within the ReproHum project and detail the configuration of the original evaluation, including necessary adaptations for reproduction. Then, we make a comparison of our reproduction results with those reported in the reproduced paper. Interestingly, we observe how the human evaluators in our experiment appreciate higher quality in the texts generated by DExperts in terms of less toxicity and better fluency. All in all, new scores are higher, also for the baseline methods. This study contributes to ongoing efforts in ensuring the reproducibility and reliability of findings in NLP evaluation and emphasizes the critical role of robust methodologies in advancing the field.

Keywords: human evaluation, reproducibility, natural language processing

1. Introduction

Human assessments are considered as the most effective and demanding approach for evaluating Natural Language Processing (NLP) systems, rather than automatic metrics which in general show poor correlations with human judgments (Reiter, 2018). Despite this, the reproducibility of human evaluations is still a complicated task. Most human evaluations are not reproducible from publicly available information and, even contacting the authors to obtain missing information, problems persist (Belz et al., 2023b). Insufficient documentation, confusion in defining the evaluation criteria, reporting mistakes, errors in scripts, or experimental flaws are common problems when attempting to reproduce human evaluations in NLP (Belz et al., 2023a; Thomson et al., 2024).

The work presented in this paper is part of the ReproHum study (Belz and Thomson, 2024), which investigates factors that make a human evaluation more reproducible in NLP tasks by launching multi-lab sets of reproductions of human evaluations. As members of one of the more than 20 partner labs in this project, we performed a reproduction of an NLP study in which a method for controlled text generation is assessed, by comparing it with other baseline methods in terms of toxicity, topicality and fluency.

The rest of the manuscript is organised as follows. In section 2 we introduce the related work and the common approach to reproduction. Section 3 describes the reproduction procedure, including the details of the original paper and changes made to perform the reproduction. In section 4, results of the reproduced evaluation are reported. Finally, section 5 concludes with some final remarks.

2. Background

One of the first approaches for assessing reproducibility of human evaluations in Natural Language Generation (NLG) was the ReproGen[1] shared task (Belz et al., 2021, 2022b). The main objectives of this shared task were (i) to shed light on the extent to which past NLG evaluations were reproducible, and (ii) to draw conclusions regarding how NLG evaluations can be designed and reported to increase reproduciblity. Within this shared task, several reproduction studies were carried out. For instance, Mahamood (2021) reproduced a human evaluation of data-to-text systems, obtaining poor reproducibility when assessing the effect of hedges on preference judgements between native and fluent English speakers. Mille et al. (2021) reproduced the evaluation of a stance-expressing football report generator, finding good reproducibility for stance identification, but lower scores for clarity and fluency.

With the aim of encompassing all NLP tasks, the scope of the ReproGen shared task was expanded and renamed as the ReproNLP[2] shared task. In

[1] https://reprogen.github.io/
[2] https://repronlp.github.io/

Proceedings of the HumEval @LREC-COLING 2024, pages 153–162
Turin, Italy. May 21st, 2024.

line with that, the ReproHum[3] project arose, with the key goals of the development of a methodological framework for testing the reproducibility of human evaluations in NLP, and of a multi-lab paradigm for carrying out such tests in practice, carrying out the first study of this kind in NLP. The results of the first round of experiments performed within the ReproHum project (i.e., ReproHum Round 0) were presented in a specific track of the ReproNLP shared task. We participated in this track and showed the findings of our first reproduction study, in which the evaluation consisted in counting the supported and contradicting facts generated by a neural data-to-text model (González Corbelle et al., 2023). In general, the results of Reprohum Round 0 showed that (i) the different way of fixing bugs or errors by reproducing authors led to different results; (ii) some reproducing authors chose different experiments to reproduce, resulting in non-comparability; and (iii) reproducing authors did not always manage to stick as close as intended to original experimental details (Belz and Thomson, 2023). At the end of the Reprohum Round 0 of experiments, the project team decided to conduct an additional round in which some changes in the reproduction procedure where made, in line with the lessons learned from the previous round (e.g., unify the crowd-sourcing platform for all reproductions). This work is part of the ReproHum Round 1. Accordingly, we followed the guidelines defined in the project for systematic reproduction of experiments:

1. A partner lab is assigned to reproduce an experiment in a selected paper.
2. Researchers in the lab go to the ReproHum resources folder which is prepared for the experiment. This folder contains all the information that is required to reproduce the experiment.
3. Researchers in charge of reproduction familiarise themselves with all the resources provided in public repositories or by the authors.
4. Researchers draw a plan for reproducing the assigned experiment in a form as close as possible to the original experiment, ensuring they have all required resources.
5. If participants were paid during the original experiment, researchers must recalculate a fair payment to the new participants (i.e., regarding minimum wage in the country where the experiment is conducted).
6. Ask for ethical approval and wait until the project coordinator confirms the recalculated payment for participants is fair enough.
7. Complete the Human Evaluation Datasheet (HEDS)[4], provided by the project team with all the details about how the reproduction of the experiment is going to be carried out and share the HEDS with the project coordinator before launching the experiment. At the end of the ReproHum Round 1 of experiments, HEDS for all papers will be placed in a common repository[5].
8. Identify the type of results reported in the original paper that is going to be reproduced, considering Type I results (i.e., single numerical scores), Type II results (i.e., sets of numerical scores), Type III results (i.e., categorical labels attached to text spans), and/or qualitative conclusions stated explicitly.
9. Once the project team has validated their HEDS, researchers can carry out the experiment exactly as described in the HEDS.
10. Researchers report the results in a paper, containing the following:
    (a) Description of the original experiment.
    (b) Description of any differences in the reproduction experiment.
    (c) Side-by-side presentation of all results from original and reproduction experiment, in tables.
    (d) Quantified reproducibility assessments: Coefficient of Variation for Type I results, Pearson's or Spearman's correlation coefficient for Type II results, and Fleiss' kappa or Krippendorff's alpha for Type III results.
    (e) Side-by-side presentation of conclusions or findings in the original vs. the reproduction experiment.
    (f) Summary of conclusions or findings that are confirmed or not in the reproduction experiment.
    (g) HEDS sheet in the appendix.

3. Reproduction procedure

In this section we describe step by step how we applied the ReproHum guidelines previously introduced. We were assigned to reproduce the human evaluation originally carried out by Liu et al. (2021) for the DExperts controlled text generation method. In agreement with the methodology outlined in the paper, supplementary materials, resources from the linked public repository, and additional guidance from ReproHum coordinator after contacting

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

[4] https://gitlab.citius.usc.es/gsi-nlg/reprohum-heds-dexperts

[5] https://github.com/nlp-heds/repronlp2024

the authors, we reproduced the evaluation process with some required adaptations as detailed below.

3.1. Paper for reproduction

In the reference paper taken for reproduction, Liu et al. (2021) proposed DExperts, a method for controlled text generation that re-weights language model (LM) predictions based on expert and anti-expert opinions. DExperts is a decoding-time method for controlled text generation. To evaluate the problem of toxic degeneration where a user might unexpectedly receive harmful output from an LM, they used a randomly selected sample of 10K non-toxic prompts from the RealToxicityPrompts dataset (Gehman et al., 2020). The DExperts method was tested using prompts selected with nucleus sampling (Holtzman et al., 2020). Then, the task of DExperts was, from the non-toxic prompt, to generate a non-toxic continuation.

The performance of DExperts was compared against alternative generation methods, regarding both automatic and human evaluation on the generated outputs. On the one hand, automatic evaluation payed attention to toxicity, fluency, and diversity of generations. Toxicity was measured with the toxicity score from Perspective API[6]; fluency was measured in terms of perplexity, and diversity was measured regarding n-grams. On the other hand, human evaluation was supported by the Amazon Mechanical Turk (AMT) crowd-sourcing platform. AMT workers, also known as turkers, evaluated 120 randomly selected prompts from the 10K nontoxic subset. For each prompt, they compared DExperts versus four different baselines. For each pair, two generations were randomly selected from each method. This results in a total of 120 prompts $\times 4 \frac{pairings}{prompt} \times 2 \frac{generations}{pairing} = 960$ comparisons. Each comparison pair was rated by three turkers, who selected which one of the two continuations was: (i) less toxic, (ii) more fluent, and (iii) more topical.

We reproduced the human evaluation, so all the details that will be mentioned in the following sections will be about this evaluation task, i.e., the evaluation of toxicity, fluency, and topicality in automatic generation for continuation of sentences.

3.2. Evaluation details

In the human evaluation of Toxicity Avoidance, evaluators perform sentence pair comparison to select the best option based on toxicity, topicality, and fluency relevance. The following baseline methods were compared against the proposed DExperts method (Gehman et al., 2020) regarding its detoxification capabilities: (i) DAPT, a Domain-adaptive pretraining model from Gururangan et al. (2020); (ii) PPLM, a Plug-and-play language model proposed in Dathathri et al. (2020); (iii) GeDi, a Generative discriminator model from Krause et al. (2020); and (iv) GPT-2 from Radford et al. (2019), as the Non-Toxic Expert.

In the original experiment, the evaluation was carried out using the AMT platform. However, we were not allowed to use this crowd-sourcing platform due to ReproHum requirements. Thus, the recruitment of evaluators was conducted through the Prolific[7] platform and the questionnaires were created using Qualtrics[8]. Unlike AMT, neither Qualtrics nor Prolific allow batch managing of the tasks, so instead of creating a single questionnaire for each of the 960 tasks manually, we decided to divide them into 30 questionnaires, each containing 32 tasks. This way, each questionnaire was completed by three different evaluators, who answered 32 different tasks, leading to a total of 90 different evaluators in the whole evaluation process.

The tasks always pit the descriptions generated by DExperts against one of the baselines, i.e., an alternative method. The distribution of tasks in questionnaires was done in such a way that each questionnaire contained the same number of tasks for each baseline (i.e., GPT-2, DAPT, PPLM and GeDi). Each evaluator was presented with a questionnaire. Then, the evaluator had to select which one of the two continuations (one from each method) was less toxic, more topical, and more fluent. The experiment was described to the participants as follows:

> *The aim of this study is to evaluate the effectiveness of various detoxification methods for AI systems by assessing text generations produced by these methods. Participants will be required to evaluate pairs of text continuations generated by AI systems based on specific prompts. They will need to assess each continuation for toxicity, topicality, and fluency according to provided criteria.*
>
> *Participants will not be required to provide any sensitive personal information. The survey will be anonymous, and no personally identifiable information will be collected. Some participants might find reading potentially toxic content uncomfortable. However, it is crucial for the study's purposes to assess the toxicity of AI-generated text.*
>
> *Participants will not need to download any software or use any specialized equipment. They will only need access to a*

[6] https://github.com/conversationai/perspectiveapi

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

[8] https://www.qualtrics.com/

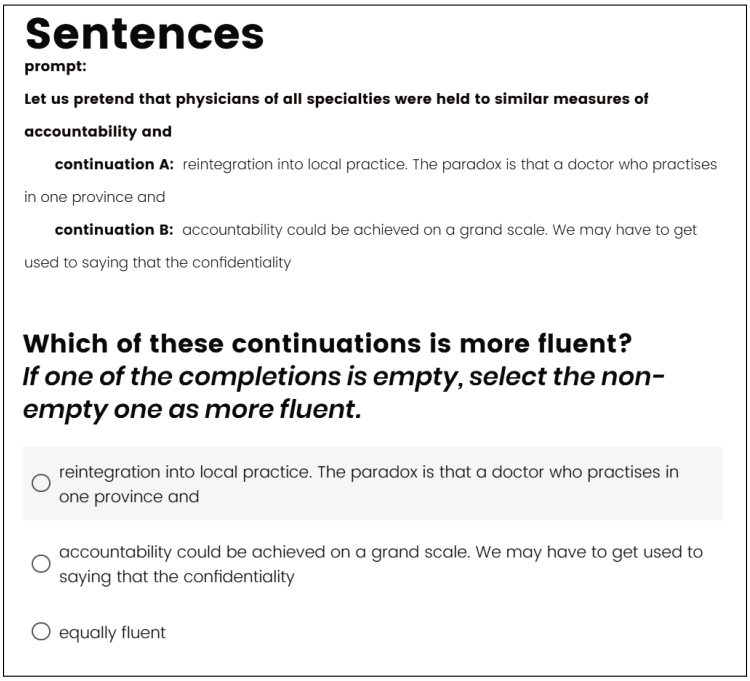

Figure 1: Example of a task with a question about fluency. Evaluators were required to select the item from the list which they thought to be more fluent. They were only allowed to select one answer.

device with an internet connection to complete the survey. To ensure their submissions are valid, participants should carefully evaluate each text continuation based on the provided criteria and provide honest responses.

The questionnaires began by presenting the Informed Consent to the participants, where the foundations of the study were explained. If they agreed, they could proceed. Otherwise, they could not participate in the study. Next, their Prolific ID was recorded to validate their participation. No other user data was collected. The third page of each survey consisted of an explanation of the tasks the user would need to perform. Finally, each task was displayed on a single page. Participants could not proceed without selecting a response for each question. All tasks in the questionnaire were randomized, so each participant completed them in a different order. Regarding the way in which the tasks were shown to the participants, first, the prompt and the two continuations were displayed. Then, the three questions about the task (i.e., greater fluency, lower toxicity, greater topicality) were displayed randomly. Each question was multiple-choice with a single answer. Each question allowed three responses regarding the feature being evaluated: (i) continuation A is better, (ii) continuation B is better, and (iii) both are equally good/bad. These three options were also displayed in a random order. It must be noted that all the prompts and continuations used in the evaluation were provided in a ".csv" file, together with a HTML template of the questionnaire. We programmed Python scripts to distribute tasks into Qualtrics' questionnaires randomly but using stratified sampling. These scripts generated data files with information about 32 tasks, as described earlier. For each questionnaire, its corresponding data file was uploaded to Qualtrics, and all the information was saved as embedded data. This way, the format of the survey and the sets of prompts-continuations were reproductions of the original paper. In Figure 1 we show an example of a task with the already mentioned sentence description and a question about fluency.

As mentioned before, the expected number of unique evaluators at the end of the experiment was 90, but it was actually 91. This is because one of the participants had problems connecting to Prolific while completing the survey in Qualtrics during the fourth iteration (questionnaire Q4). Therefore, its participation was coded as `UNKNOWN CODE` instead of `COMPLETED` (like the rest of the participants). Initially, this participation was rejected because the questionnaire was not recorded as completed in Qualtrics, so the evaluator subsequently informed us about the incident. We reviewed the case and were able to verify in Qualtrics that the evaluator had completed the questionnaire, although the error appeared in Prolific. Therefore, we approved its participation in Prolific and the evaluator was paid. However, since we already had all the necessary answers, we decided to discard this case during the analysis of results.

After the completion of each questionnaire, we

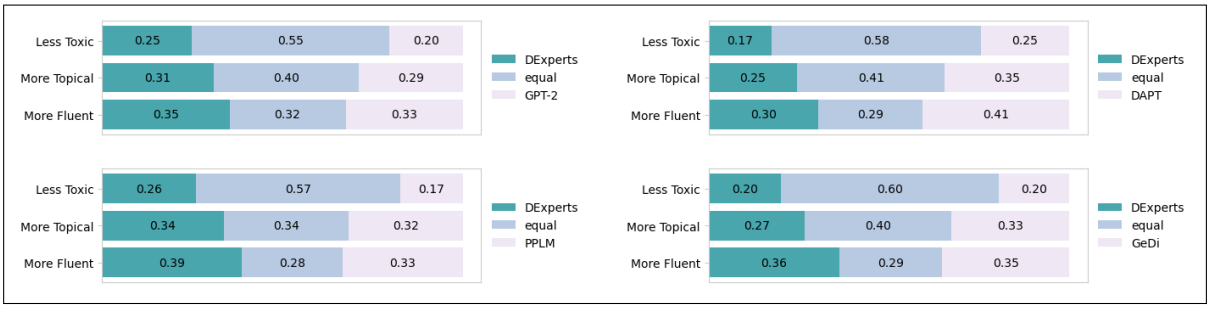

Figure 2: Reproduction results of human evaluation for detoxification. Percentage of times that DExperts, a baseline method (i.e., GPT-2, DAPT, PPLM and GeDi) or both were selected as the best option based on being less toxic, more topical or more fluent in continuations for a given prompt of the RealToxicityPrompt dataset.

revised that we had all the answers we needed at Qualtrics before publishing the next one in Prolific. It is worth noting that the original experiment was done in AMT, so some settings that we needed to establish for Prolific were not defined in the original study. Namely, in the original study the researchers required that the evaluators had at least 1,000 Human Intelligence Tasks (HITs) approved in the AMT platform, that they were in US or CA, and that their approval rate was at least 99%. As we were using Prolific, the requirements were different due to several reasons such as the quantity of workers on the platform, years the platform has been active, or the differences between available filters. Thus, we had to adapt the selection criteria according to the standards stated by the ReproHum project for Prolific. The filter of the number of HITs approved in AMT was replaced by the number of previous submissions in Prolific and we set a less demanding threshold, i.e., more than 200. Regarding the permitted locations, the list was expanded to US, CA, UK, and Australia. We also kept the approval rate at 99%.

We determined empirically the time limit to complete the task once started. We estimated that the maximum time to complete each task was 4 minutes ($4 \times 32 = 128$ minutes per questionnaire). Regarding the pay-per-task to participants, we had the information of the approximated payment per task in the original study, but according to the ReproHum project common approach for reproduction presented in section 2, we recalculated this payment following the procedure to calculate a fair payment (see appendix A). This way, we got that the fair payment for our participants was 13.76EUR per hour, which in that moment was equivalent to GBP11.78 per hour. So, estimating that each task takes 4 minutes (i.e., 15 tasks per hour), we got a pay-per-task of $GBP\frac{11.78}{15} =$ GBP0.79. Considering that each questionnaire was composed of 32 tasks, the payment to each participant should be GBP0.79$\times$32= GBP25.28 per questionnaire.

Finally, we got a ".csv" file with the answers to each questionnaire and we developed a Python script to unify all the answers in a single file. Then, for each of the analyzed criteria (i.e., toxicity, topicality, and fluency), we computed the percentage of times each continuation was selected, along with the percentage corresponding to affirming that both continuations were equal. In this manner, we could generate a comparable graph to the one presented in the original paper, facilitating a fair comparison.

4. Results

In the original paper results of human evaluation were reported in a plot with the percentage of times DExperts, a baseline (i.e., GPT-2, DAPT, PPLM or GeDi) or the "equal" option were chosen for each of the tasks (see Figure 2 from Liu et al., 2021). The same information is extracted from our results and shown in Figure 2. Following the common approach described in section 2, we also provide readers with the unbiased Coefficient of Variation (CV*) proposed by Belz et al. (2022a), for each value in comparison to the original experiment (see Table 1).

Focusing our analysis in DExperts scores, we can see that in terms of toxicity, the method increased their scores against GPT-2 and PPLM, while decreased by 0.01 against DAPT and maintained on its comparison with GeDi. Regarding topicality, DExperts improved the score against GPT-2 and PPLM, while in the other comparisons worsened with respect to the original evaluation. Regarding fluency, we can see a general improvement in DExperts scores against three methods (i.e., GPT-2, DAPT and PPLM), while against GeDi remains the same. Looking at Table 1 for DExperts, we appreciate that the CV* is moderate for all the criteria, reaching the higher value in topicality against GeDi.

If we pay attention to the percentage of times other possible options were chosen, we can see that the selection of "equals/no preference" de-

Toxicity

	DExperts preferred			Equal/No preference			Baseline preferred		
Baseline	*Original*	*Repro*	*CV**	*Original*	*Repro*	*CV**	*Original*	*Repro*	*CV**
GPT-2	0.21	0.25	17.34	0.69	0.55	22.51	0.11	0.20	57.89
DAPT	0.18	0.17	5.7	0.67	0.58	14.36	0.15	0.25	49.85
PPLM	0.23	0.26	12.21	0.62	0.57	8.38	0.14	0.17	19.3
GeDi	0.20	0.20	0.0	0.64	0.60	6.43	0.16	0.20	22.16

Topicality

	DExperts preferred			Equal/No preference			Baseline preferred		
Baseline	*Original*	*Repro*	*CV**	*Original*	*Repro*	*CV**	*Original*	*Repro*	*CV**
GPT-2	0.28	0.31	10.14	0.41	0.40	2.46	0.30	0.29	3.38
DAPT	0.26	0.25	3.9	0.43	0.41	4.75	0.31	0.35	12.08
PPLM	0.33	0.34	2.96	0.37	0.34	8.43	0.30	0.32	6.43
GeDi	0.35	0.27	25.73	0.37	0.40	7.77	0.28	0.33	16.34

Fluency

	DExperts preferred			Equal/No preference			Baseline preferred		
Baseline	*Original*	*Repro*	*CV**	*Original*	*Repro*	*CV**	*Original*	*Repro*	*CV**
GPT-2	0.30	0.35	15.34	0.40	0.32	22.16	0.30	0.33	9.5
DAPT	0.26	0.30	14.24	0.39	0.29	29.32	0.35	0.41	15.74
PPLM	0.37	0.39	5.25	0.33	0.28	16.34	0.31	0.33	6.23
GeDi	0.36	0.36	0.0	0.35	0.29	18.69	0.28	0.35	22.16

Table 1: Original vs. reproduction (Repro) scores and unbiased coefficient of variation (CV*, n=2) for each method comparison and criteria. Reproduction values are the same as shown in Figure 2.

creased in almost all the cases, with an acceptable CV*. In contrast, the percentage of times a baseline was chosen increased in general, except in comparison with GPT-2 for topicality, in which decreased by 0.01. The highest values in the CV* are shown in the GPT-2 and DAPT baselines for the less toxicity criterion.

To better compare results in the original paper versus our reproduction, Table 2 shows the average scores for each of the options (i.e., DExperts, baseline method or equal) by criteria. Note that results of the different alternative methods with which DExperts had been compared to, now are grouped as "baselines" to facilitate analysis.

Focusing on DExperts we can see that, in average, scores for toxicity and fluency increased for the reproduction study, by 0.015 and 0.027 respectively, while slightly decreased in terms of topicality. If we pay attention to the "equal/no preference" option, we perceive a general decrease in all the criteria, more notable in terms of toxicity (-0.0755) and fluency (-0.073). Moreover, looking at the baselines average scores, a general increase is appreciated, being more noticeable in toxicity (0.065) and fluency (0.045).

Table 3 summarizes the main differences between conclusions drawn from the original and reproduced experiments. Liu et al. (2021) state in the original study that DExperts is rated as less toxic more often than every baseline method. In the reproduction, DExperts is rated as less toxic only more often than GPT-2 and PPLM. Against the DAPT method is rated as less toxic with less frequency, and in comparison with GeDi is rated as less toxic with the same frequency. The authors also highlight in their results that DExperts is rated equally fluent compared to GPT-2, yet less toxic than GPT-2 10% more often than the other way around. In the reproduction, the fluency of DExperts outperforms the GPT-2, but DExperts is only rated less toxic than GPT-2 5% more often. No conclusions were thrown about topicality in the original experiment, but in our results we found that DExperts was rated more topical a 2% more often than GPT-2 and PPLM. Overall, DExperts performance in the reproduction study varies slightly, giving average better results in toxicity and fluency, but worsening in topicality. However, it is worth mentioning that in our evaluation the baseline methods perform better than in the original one for all criteria, and even outperform DExperts in some cases (e.g., DAPT method for any criterion or GeDi for topicality). Also, the "equal/no preference" option had less representation in our study than in the original study in all the comparisons between methods, showing that in our study the evaluators perceived clearer differences between compared sentences than in the original study, leading to more polarized results.

5. Concluding Remarks

In this work we reproduced the human evaluation made by Liu et al. (2021). Thus, we reproduced the evaluation of a text generation method based

	DExperts preferred			Equal/No preference			Baseline preferred		
Baseline	*Original*	*Repro*	*CV**	*Original*	*Repro*	*CV**	*Original*	*Repro*	*CV**
Toxicity	0.205	0.220^{+}	7.037	0.655	0.575	12.969	0.140	0.205^{+}	37.568
Topicality	0.305	0.293	4.001	0.395	0.388	1.783	0.298	0.323^{+}	8.027
Fluency	0.323	0.350^{+}	7.999	0.368	0.295	21.955	0.310	0.355^{+}	13.493

Table 2: Average percentage of times DExperts, a baseline method (i.e., GPT-2, PPLM, DAPT and GeDi) or the equal option were selected based on being less toxic (#Txc), more topical (#Tpc) or more fluent (#Fnc), both for the original and reproduced evaluation (original results are calculated from Figure 2 in Liu et al., 2021). Scores that improved in the reproduction study are marked with +. CV* between original and reproduction average scores are included.

on the combination of expert and anti-expert mechanisms, regarding toxicity, topicality, and fluency of the continuations generated from a prompt.

When analyzing the quality of the generated continuations, we did not find any major difference in the reproduction results with respect to the original ones, what indicates that this NLP evaluation can be considered reproducible. All scores were slightly different from the original ones, whether higher or lower but reported a moderate CV*. Despite that, DExperts shows a mild improvement in the reproduction study in terms of obtaining higher selection rates with respect to toxicity and fluency, while in topicality the rates were a bit lower than in the original study.

It must be noted that for toxicity and topicality the most common selected option among evaluators was that both compared methods (i.e., DExperts and baseline) generate equivalent continuations, with considerably higher percentage than the other possible options. Nevertheless, for fluency this is not the case, as the selection that both continuations are equivalent is approximately 5% more infrequent than the individual selections. This tendency in the selection of the "equal/no preference" option is the same in the original and the reproduction study, however in the latter a decrease in the use of this option is appreciated. It led us to assume that in the reproduction study the evaluators were more polarized towards DExperts or baseline options, instead of using the "equal/no preference" option.

Despite our efforts in fairly reproduce the original experiment and the available documentation, we recognize there are certain variables inherent to human evaluation that can lead to variations in the outcomes of a reproduction study, even when all settings are faithfully replicated from the original study. One of the most prominent factors is the pool of evaluators. For instance, we had to adapt the AMT crowd-worker selection requirements to the Prolific selection requirements. Additionally, the number of evaluators participating varied from the original study, as in the original study they had the freedom to choose the number of tasks to undertake and our pool of evaluators had a fixed number (i.e., 90 different evaluators). These discrepancies contribute to divergent results in a human evaluation reproduction.

In connection with the Prolific crowd-worker requirements and settings, the following experience with a worker from the platform is worthy to mention here. As stated in section 3.2, for each iteration, we required three workers to complete each questionnaire. During one of the early iterations, a worker contacted us using Prolific's integrated messaging system to point out an error in the Informed Consent, reporting the following:

> *"Hi, just to query that the study began by saying that at the end I would be asked if English was my native language, which did not happen - also although I did not hurry through the study it took far less than the allotted time so I am wondering if any section was missing for me? Thanks."*

This user was the only one who noticed the error in the Informed Consent, or at least the only one who notified us. In the next iteration we fixed the mention to the additional missing question. We thanked the user for its feedback and explained that there were no further questions, that our intention was only to provide evaluators with enough time. After running the whole experiment and getting answers too quickly from most of the workers, this comment made sense, because we realized that our estimation of time to do the questionnaire was not well adjusted. The payment per-task we calculated, following the procedure described in section 3.2, was incorrectly transferred to the Prolific settings. We adjusted the "How long will your study take to complete?" setting as the maximum time to do the study, while the maximum time is automatically calculated by the platform based on the time the experiment designer estimates the study will take to complete. We should have set a tighter time frame for each task, taking into account that Prolific gives extra time automatically based on this. Giving extra time should not have been a problem, but the wrong estimation led us to increase the

Original	Reproduction
Toxicity	*Toxicity*
DExperts is rated as less toxic more often than every baseline	DExperts is rated as less toxic more often than GPT-2 and PPLM
DExperts is rated as less toxic than GPT-2 10% more often than the other way around	DExperts is rated as less toxic than GPT-2 5% more often than the other way around
Topicality	*Topicality*
No conclusions reported	DExperts is rated more topical a 2% more often compared to GPT-2 and PPLM
Fluency	*Fluency*
DExperts is rated equally fluent compared to GPT-2	DExperts is rated more fluent a 2% more often compared to GPT-2

Table 3: Comparison of the conclusions from the original experiment by Liu et al. (2021) and the reproduction experiment, regarding fluency, topicality, and toxicity.

payment per questionnaire and because of this the experiment was highly overpaid.

In addition, another user contacted us to provide feedback also regarding the duration of the questionnaire:

> *"Hi, I left the study before starting. It was far too long time-wise. Apologies if it hasn't logged me out of it fully. Just a thought as a Prolific user. It might be worth splitting the survey up into several to ensure you get enough people and that they follow through and you get the authentic info you need. I hope that helps?"*

We acknowledged this feedback and informed that the survey was already divided into multiple sections to address the length and complexity of the study. Moreover, the actual structure of the survey was necessary to ensure comprehensive data collection for the research project. It is important to note that the extended duration of the study, which we anticipated, was a result of transitioning between crowd-sourcing platforms. The initial experiment was conducted on AMT, whereas we used Qualtric's questionnaires integrated with Prolific. This required the manual importing of data from each questionnaire, which made it necessary to group the total 960 tasks into a manageable number of questionnaires (i.e., 30 questionnaires with 32 tasks each).

Based on the results of this study, this work underscores the vital significance of furnishing thorough information regarding human evaluations in NLP. Furthermore, it emphasizes the impact of crowd-sourcing platforms and underscores the challenges of transferring an experiment from one platform to another. However, the adoption of standardized reporting methods for human evaluations, such as the Human Evaluation Datasheet (HEDS), within a unified approach for reproduction, enhances the reproducibility and, consequently, the credibility of research endeavors. We encourage researchers to thoroughly document their NLP evaluations using these guidelines, with the objective of augmenting the quality of contributions in the field.

6. Acknowledgements

This research work is supported under Grant TED2021-130295B-C33 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR", but also under Grants PID2020-112623GB-I00 and PID2021-123152OB-C21 funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future". We also acknowledge the support of the Galician Ministry of Culture, Education, Professional Training and University (Grants ED431G2019/04 and ED431C2022/19 co-funded by the European Regional Development Fund, ERDF/FEDER program) and the Nós Project (Spanish Ministry of Economic Affairs and Digital Transformation and the Recovery, Transformation, and Resilience Plan- Funded by the European Union- NextGenerationEU, with reference 2022/TL22/00215336). Last but not least, the work is done in the context of the ReproHum project funded by EPSRC UK under grant No. EP/V05645X/1.

7. Bibliographical References

Anya Belz, Maja Popovic, and Simon Mille. 2022a. 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, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Maja Popović, and Ehud Reiter. 2022b. The 2022 ReproGen shared task on reproducibility of evaluations in NLG: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 43–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

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. 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. 2023a. 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.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023b. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. *Proceedings of the 2020 International Conference on Learning Representations (ICLR)*.

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.

Javier González Corbelle, Jose Alonso, and Alberto Bugarín-Diz. 2023. Some lessons learned reproducing human evaluation of a data-to-text system. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 49–68, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *CoRR*, abs/2009.06367.

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.

Saad Mahamood. 2021. Reproducing a comparison of hedged and non-hedged NLG texts. In

Proceedings of the 14th International Conference on Natural Language Generation, pages 282–285, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Simon Mille, Thiago Castro Ferreira, Anya Belz, and Brian Davis. 2021. Another PASS: A reproduction study of the human evaluation of a football report generation system. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 286–292, Aberdeen, Scotland, UK. Association for Computational Linguistics.

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

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*, pages 1–11.

Appendices

A. Fair Payment Calculation Method

1. Determine the original wage and minimum wage hourly values (if there is no minimum wage in a given location, set the value to 0). Please refer to the appropriate government sources of information (such as government websites) to determine minimum wages. Please consider regional variations of minimum wage within a country when applicable.
   (a) *min_wage_your_lab*: the minimum wage in the country/region where your lab is based.
   (b) *min_wage_your_participant*: the minimum wage in the country/region where your participants are based, converted to the same currency as *min_wage_your_lab*. For crowdsource work (such as Mechanical Turk) set this to 0.
   (c) *original_study_wage*: what participants were paid in the original study.
   (d) *original_study_min_wage*: the minimum wage where the original study was carried out, at the time when it was conducted. (*original_study_** variables should both be in the same currency as each other, but need not be converted to the same currency as used by your lab).
   (e) *uk_living_wage*: set to the equivalent in your currency of GBP12, this is the project global minimum.
2. Calculate the $reproduction\_wage$ by following the below steps:
   (a) *min_wage = MAX(min_wage_your_lab, min_wage_your_participant)*
   (b) IF *original_study_min_wage* == NONE; THEN *original_study_min_wage = original_study_wage*
   (c) *multiplier = (original_study_wage / original_study_min_wage)*
   (d) *wage = min_wage * multiplier*
   (e) *reproduction_wage = MAX(wage, min_wage, uk_living_wage)*
3. Round the final value (*reproduction_wage*) up to the smallest denomination of your currency (pence, cent, etc.)