Some lessons learned reproducing human evaluation of a data-to-text system

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

Abstract

This paper presents a human evaluation reproduction study regarding the data-to-text generation task. The evaluation focuses in counting the supported and contradicting facts generated by a neural data-to-text model with a macro planning stage. The model is tested generating sport summaries for the ROTOWIRE dataset. We first describe the approach to reproduction that is agreed in the context of the ReproHum project. Then, we detail the entire configuration of the original human evaluation and the adaptations that had to be made to reproduce such an evaluation. Finally, we compare the reproduction results with those reported in the paper that was taken as reference.

1 Introduction

An experiment or study is reproducible when independent researchers can replicate it by following the documentation shared in the original report and draw the same conclusions, which is also a clear synonym of reliability. In Natural Language Processing (NLP), reproducibility is not limited to specifying the parameters chosen to train a model, but it goes beyond that and requires the specification of all the details of the evaluation process by which the reported results are obtained. In NLP, until recently, not too much attention has been paid to the reproducibility of neither automatic nor human evaluations. In the case of automatic metrics, there is a reproducibility checklist (Pineau, 2020), but in the case of human evaluations not so much progress has been made.

In addition, some papers have been published about reproducibility in NLP, regarding reproducibility tests based on the fulfillment of certain properties in human evaluations (Belz et al., 2020) but also proposing a template for recording the details of human evaluations in NLP experiments, with the aim of improving the replicability of these processes (Shimorina and Belz, 2022).

The work presented in this paper is part of the ReproHum[1] project, that investigates the factors that make a human evaluation more reproducible in NLP by launching multi-lab sets of reproductions of human evaluations. As members of one of the 21 partner labs in this project, we performed a reproduction of an NLP study in which a data-to-text system is assessed and compare the results obtained in the reproduction with the original ones.

The rest of the manuscript is organised as follows. In section 2 we introduce related work and the common approach defined as a global requirement for all the reproducibility experiments within ReproHum project. Section 3 describes the reproduction of the NLP evaluation, first, explaining the content of the paper chosen for reproduction and then, explaining all the details of the evaluation that is going to be reproduced. In section 4, the results of the reproduced evaluation compared to the original paper are reported and discussed. Finally, section 5 concludes with final remarks and future work.

2 Background

In the context of the shared task REPROLANG (Branco et al., 2020) a replication of a human evaluation of a neural text simplification system by Nisioi et al. (2017) was performed (Cooper and Shardlow, 2020), obtaining worse results in the reproduction study, in terms of Grammaticality and Meaning Preservation.

With the aim of developing theory and practice of reproducibility assessment, the ReproGen shared task arose and in its two editions (Belz et al., 2021, 2022) several studies involving the reproduction of different experiments were carried out. Popović

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

Proceedings of the Third Workshop on Human Evaluation of NLP Systems,
pages 49–68, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-088-5_005

and Belz (2021) replicated an evaluation of Machine Translation outputs where errors related to comprehensibility and meaning correctness were annotated in texts by marking up word involved in an error (Popović, 2020). They found that 4 out of 6 system rankings were the same in both studies, but error rates for minor error types have lower reproducibility than those classified as major error types.

Mahamood (2021) reproduced human evaluations of data-to-text systems. Despite differences in the number and type of raters, authors found poor reproducibility when assessing the effect of hedges on preference judgments between native and fluent English speakers. Mille et al. (2021) faced the evaluation reproduction of a stance-expressing football report generator (van der Lee et al., 2017), finding good reproducibility for stance identification accuracy, but lower reproducibility for Clarity and Fluency.

In addition, it is worth noting that in the context of the ReproHum project, adhering to the following guidelines is mandatory when reproducing experiments:

1. You are allocated an experiment in a paper.
2. Go to the resources folder which is prepared adhoc for the experiment. This folder contains all the information you will need to reproduce the experiment.
3. Familiarise yourself with the experiment that was assigned for reproduction and all the resources provided in the public repositories or by the authors.
4. Plan for repeating the allocated experiment in a form that is as far as possible identical to the original experiment, ensuring you have all required resources, and apply to your research ethics committee for approval.
5. If participants were paid during the original experiment, follow the project procedure to recalculate a fair pay to the workers (regarding minimum wage, original study wage, and so on).
6. Ask for ethical approval and wait until the project team confirms the payment to the workers.
7. Complete the Human Evaluation Datasheet (HEDS, see appendix A) provided by the project team with all the details about how the repetition of the experiment is going to be carried out and share the HEDS with the project before launching the experiment.
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 have validated your HEDS, carry out the experiment exactly as described in the HEDS.
10. Report the results in a paper describing the original experiment, any differences in your reproduction experiment, presentation of the results and conclusions in the original vs. reproduction experiment, and finally draw overall conclusions and share the HEDS in the appendix.

It must be noted that during all the reproduction process described above is not allowed to contact the authors of the original paper or communicate with other project labs carrying out this or any other reproduction experiment to avoid affecting the reported outcomes. Thus, all the information and resources provided should be in the common resources folder provided by the project team and in case of any question we were asked to only contact the ReproHum project managers who act as a proxy with the authors of the work to be reproduced.

3 Reproduction of an NLP evaluation

In this section we describe how we applied the ReproHum guidelines previously introduced. For the purpose of human evaluation reproduction, we were assigned the paper published by Puduppully and Lapata (2021). Based on the evaluation details described in the paper, the appendices, the resources available in the associated public repository, and the resources provided by the ReproHum managers after contacting the authors, we reproduced the evaluation as close as possible to the original one.

3.1 Paper for reproduction

As described above, our experiment consisted in performing a reproduction as accurate as possible

of a human NLP evaluation. In the reference paper taken for reproduction, Puduppully and Lapata (2021) propose a neural model with a macro-planning stage followed by a generation stage reminiscent of traditional methods comprising separate modules for planning and surface realization. The proposed model (Macro) is tested with two datasets for data-to-text from the sports field: ROTOWIRE (Wiseman et al., 2017) and MLB (Puduppully et al., 2019). The former consists of a dataset composed of tables with NBA basketball game statistics, aligned with summaries describing such data; while the later maintains the same format, but the data are about MLB baseball games. Therefore, the task of the generation model is, from the data tables, to generate sports summaries describing the game statistics.

To demonstrate that Macro improves the results of other architectures for data-to-text generation, they make a comparison against different systems, applying both automatic and human evaluation on the system outputs. On the one hand, the metrics used to automatically evaluate the texts generated by the different models are BLEU, and the set of Information Extraction (IE) metrics proposed in (Wiseman et al., 2017) to evaluate the relation generation (RG), content selection (CS) and content ordering (CO) stages of the systems. On the other hand, in terms of human evaluation, two experiments using the Amazon Mechanical Turk (AMT) crowd-sourcing platform were performed. First, the quality of the generated texts was evaluated in terms of grammar, coherence and conciseness. Second, quantifying how many of the facts mentioned in the generated texts supported or contradicted the data in the box score, i.e., the table provided as input to the system.

We reproduced the first experiment for the ROTOWIRE dataset, so all the details that will be mentioned in the following sections will be about this evaluation task, i.e., the count of supported/contradicting facts in automatic generation of NBA summaries.

3.2 Evaluation details & Changes

In the human evaluation of supported/contradicting facts, the following baseline systems were compared against the proposed Macro model (Puduppully and Lapata, 2021): (1) Templ, a template-based generator from (Wiseman et al., 2017) for ROTOWIRE; (2) ED+CC, a vanilla encoder-decoder model with an attention and copy mechanism (Wiseman et al., 2017); (3) RBF-2020 (Rebuffel et al., 2020), a Transformer encoder model, with a hierarchical attention mechanism over entities and records within entities, which represents the state of the art on ROTOWIRE dataset. In addition, the gold summaries were also included for comparison, i.e., summaries from the dataset.

Twenty summaries from the tested dataset (i.e., ROTOWIRE) were selected, which gave us a total of 100 summaries generated by the 5 different systems (including the gold summaries). For each summary, using the AMT platform, 3 different evaluators performed the task of counting the supported/contradicting facts on the texts, which yielded a total of 300 HITs (Human Intelligence Tasks). Each evaluator was presented a questionnaire with sentences randomly selected from one of the summaries under consideration along with their corresponding box scores. Then, he/she was asked to count the facts that support and contradict the data (ignoring hallucinations, i.e., unsupported facts).

To carry out the evaluation, the AMT crowd-sourcing tool was used. In order to ensure a minimum quality of the results, only crowd-workers with a minimum of 1,000 previously completed HITs were allowed to take part in the experiment. Furthermore, quality of work requirements were stated, such as only workers with an approval rate greater than 98% in the platform and from English-speaking countries (i.e., US, UK, Canada, Ireland, Australia, or NZ) were admitted.

All the details mentioned so far would allow us to perform an approximate reproduction of the evaluation, yet not as detailed as we aim in this work. We are aware that the general trend in the NLP field when writing a paper is to focus more on the analysis of the results than on exhaustively detailing the evaluation process. This is normal due to the strict length limit of papers. However, we wanted to make a faithful reproduction of the evaluation, so we asked the ReproHum project managers to contact the authors of the paper to obtain extra details on how to carry out the evaluation. They kindly replied to all the questions with full transparency and accordingly we received extra resources to carry out an evaluation as close as possible to the original one.

Regarding the way in which the questions or HITs were shown to the workers, a box score along

with 4 sentences extracted from a longer system generated summary were shown in each of the HITs. These sentences could belong to any of the 5 systems that were compared in the evaluation. Thus, for each of the 4 sentences, the worker had to count the number of contradicting and supported facts with respect to the box score and indicate it by means of a dropdown menu in a range from 0 to 20. It must be noted that all the sentences used in the evaluation were provided in a .csv file, together with the corresponding HTML template of the questionnaire for each of the HITs. This way, the format of the survey and also the sentences evaluated were exactly the same as in the original paper. In figure 1 we show an example of a HIT with the already mentioned dropdowns to fill the count of supported/contradicting facts.

In AMT the tasks must be published in batches, so we followed the same strategy as the original study to publish the different batches in which the tasks were splitted. Each dataset was divided into 4 mini-batches, i.e., taking into account the ROTOWIRE dataset, we had 100 different HITs to evaluate, so there were 4 mini-batches of 25 HITs size. The order in which the mini-batches, HITs and sentences inside each HIT were presented was the same as in the original experiment. Each posted HIT had to be completed by 3 different evaluators and there were no restrictions on the maximum number of different HITs that an evaluator could perform. Therefore, the number of unique evaluators at the end of the experiment was variable depending on how many HITs each worker had decided to complete.

After the completion of each mini-batch and before publishing the next one, certain conditions had to be checked. Answers in which the sum of contradicting and supported facts was equal to or greater than 20 must be excluded. This is because none of the sentences under evaluation had so many contradicting + supported facts.

At the end of each mini-batch the following procedure was applied:

1. Compute *FC* as the total number of facts (contradicting + supported), given by the crowd-worker for each sentence (see figure 1).

2. If $FC \geq 20$:

   2.1 The response should be excluded from the final results and a replacement HIT posted on AMT. To do this, use custom qualifications to ensure a crowd-worker who has already done this HIT, is not assigned it again.

   2.2 This crowd-worker should be prevented from doing any future task (using custom qualifications).

   2.3 Keep records of both the original response and the repeated response, but mark the final one that passed the check, so that it can be included in the final results (it is possible for the HIT to be repeated multiple times before one crowd-worker finally passes the check and that response is marked as valid for inclusion in the final results).

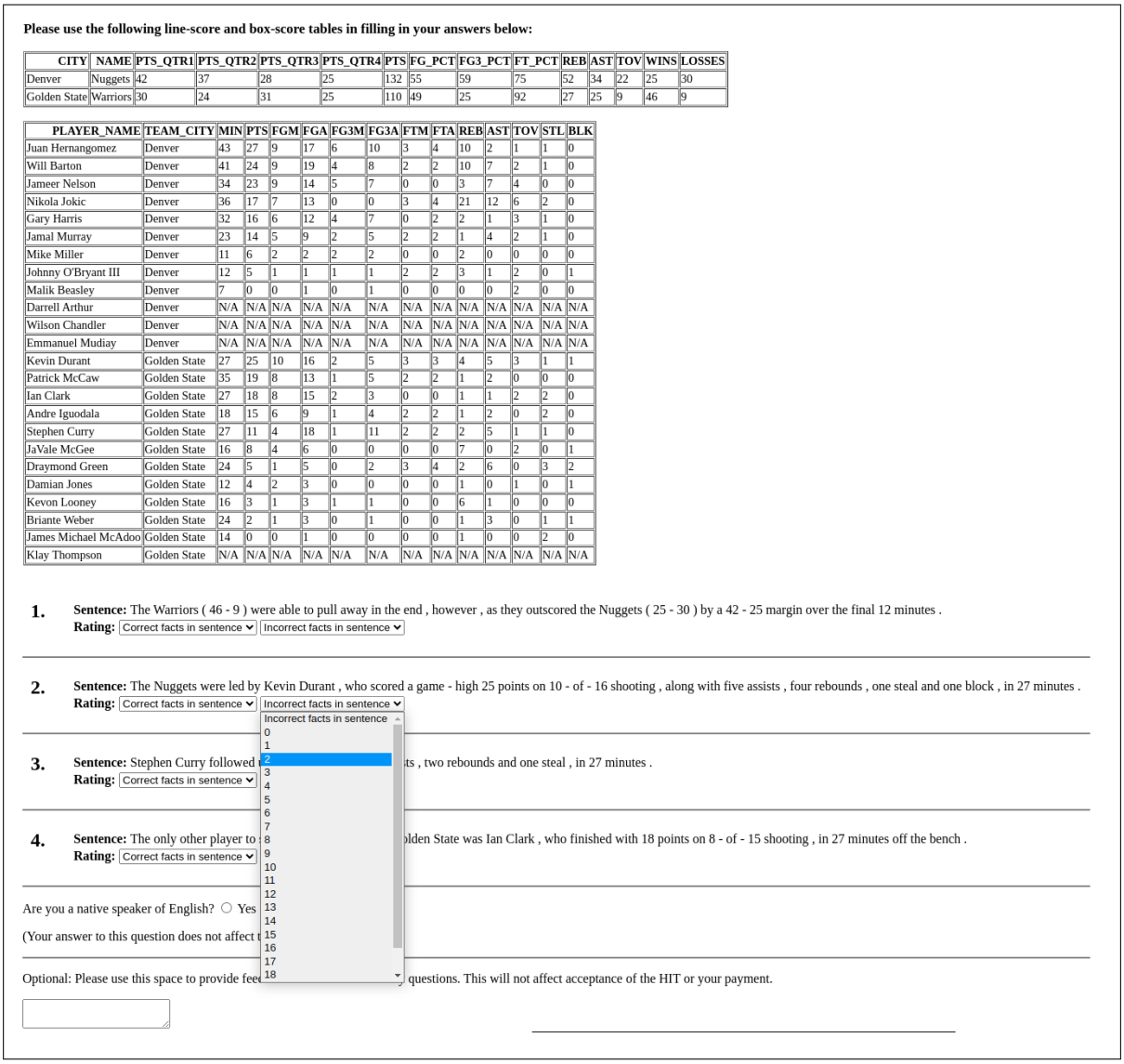
   2.4 Still pay the crowd-workers even if $FC \geq 20$, accept their work but exclude them from future tasks. This way their reputation in the platform is not affected.

3. If *FC* for every sentence from the set of 4 within a HIT is < 20, the response is valid. This HIT must be marked for inclusion in the final results.

4. Once there are valid responses for the complete mini-batch, move to the next one.

We set the HIT expiration time the same as in the original study: each crowd-worker had 7 days to perform the task once accepted before sending it without completing it.

It is worth noting that we had to set some AMT settings which were not defined in the original study. Namely, the time limit to complete the task once started was determined empirically by us. Performing several tests with people performing the task for the first time, we estimated that 4 min was the average time to complete the HIT, however we set the maximum time allotted per crowd-worker to 4h, just to ensure that no crowd-worker ran out of time. In addition, regarding the pay-per-task to crowd-workers, we had the information of the approximated payment per task in the original study, but according to the project common approach for reproduction presented in section 2, we recalculated this payment following the procedure to calculate a fair payment (see appendix B). We adjusted the payment to the current minimum wage conditions, taking as reference the UK minimum living wage per hour, that was GBP10.90 in the date of the experiment. Considering that each



Please use the following line-score and box-score tables in filling in your answers below:

CITY	NAME	PTS_QTR1	PTS_QTR2	PTS_QTR3	PTS_QTR4	PTS	FG_PCT	FG3_PCT	FT_PCT	REB	AST	TOV	WINS	LOSSES
Denver	Nuggets	42	37	28	25	132	55	59	75	52	34	22	25	30
Golden State	Warriors	30	24	31	25	110	49	25	92	27	25	9	46	9

PLAYER_NAME	TEAM_CITY	MIN	PTS	FGM	FGA	FG3M	FG3A	FTM	FTA	REB	AST	TOV	STL	BLK
Juan Hernangomez	Denver	43	27	9	17	6	10	3	4	10	2	1	1	0
Will Barton	Denver	41	24	9	19	4	8	2	2	10	7	2	1	0
Jameer Nelson	Denver	34	23	9	14	5	7	0	0	3	7	4	0	0
Nikola Jokic	Denver	36	17	7	13	0	0	3	4	21	12	6	2	0
Gary Harris	Denver	32	16	6	12	4	7	0	2	2	1	3	1	0
Jamal Murray	Denver	23	14	5	9	2	5	2	2	1	4	2	1	0
Mike Miller	Denver	11	6	2	2	2	2	0	0	2	0	0	0	0
Johnny O'Bryant III	Denver	12	5	1	1	1	1	2	2	3	1	2	0	1
Malik Beasley	Denver	7	0	0	1	0	1	0	0	0	0	2	0	0
Darrell Arthur	Denver	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Wilson Chandler	Denver	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Emmanuel Mudiay	Denver	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Kevin Durant	Golden State	27	25	10	16	2	5	3	3	4	5	3	1	1
Patrick McCaw	Golden State	35	19	8	13	1	5	2	2	1	2	0	0	0
Ian Clark	Golden State	27	18	8	15	2	3	0	0	1	1	2	2	0
Andre Iguodala	Golden State	18	15	6	9	1	4	2	2	1	2	0	2	0
Stephen Curry	Golden State	27	11	4	18	1	11	2	2	2	5	1	1	0
JaVale McGee	Golden State	16	8	4	6	0	0	0	0	7	0	2	0	1
Draymond Green	Golden State	24	5	1	5	0	2	3	4	2	6	0	3	2
Damian Jones	Golden State	12	4	2	3	0	0	0	0	1	0	1	0	1
Kevon Looney	Golden State	16	3	1	3	1	1	0	0	6	1	0	0	0
Briante Weber	Golden State	24	2	1	3	0	1	0	0	1	3	0	1	1
James Michael McAdoo	Golden State	14	0	0	1	0	0	0	0	1	0	0	2	0
Klay Thompson	Golden State	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

1. **Sentence:** The Warriors (46 - 9) were able to pull away in the end , however , as they outscored the Nuggets (25 - 30) by a 42 - 25 margin over the final 12 minutes .
Rating: Correct facts in sentence | Incorrect facts in sentence

2. **Sentence:** The Nuggets were led by Kevin Durant , who scored a game - high 25 points on 10 - of - 16 shooting , along with five assists , four rebounds , one steal and one block , in 27 minutes .
Rating: Correct facts in sentence | Incorrect facts in sentence

Incorrect facts in sentence
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18

3. **Sentence:** Stephen Curry followed ... ts , two rebounds and one steal , in 27 minutes .
Rating: Correct facts in sentence

4. **Sentence:** The only other player to ... olden State was Ian Clark , who finished with 18 points on 8 - of - 15 shooting , in 27 minutes off the bench .
Rating: Correct facts in sentence

Are you a native speaker of English? ○ Yes
(Your answer to this question does not affect t...

Optional: Please use this space to provide fee... questions. This will not affect acceptance of the HIT or your payment.

Figure 1: Example of a HIT from the survey. By checking the box score, evaluators must count how many correct/incorrect (i.e., supported/contradicting) facts are mentioned in each of the 4 sentences. Dropdowns allow to choose a value between 0 and 20 for each answer.

task takes about 4 minutes, a crowd-worker can do 15 tasks per hour, so the payment per HIT was set to GBP10.9/15 = GBP0.726 (i.e., 0.88USD = 0.80EUR) per completed task.

Finally, authors shared with the ReproHum project managers the script they used to process the results obtained from the evaluation. Given a file with the responses obtained from AMT, the mean of the scores for each system is automatically calculated and then tested by one-way ANOVA analysis of variance with Tukey posthoc to see if the results obtained for the baseline models and gold summaries show significant differences with respect to the Macro system that is evaluated.

4 Results

The evaluation process was organized in the above-mentioned 4 mini-batches to complete the total 300 tasks. Following the procedure explained in section 3.2, we had to repeat 59 HITs in order to obtain 300 valid responses. At the end, a total of 144 different crowd-workers participated in the evaluation. Notice that, in the original paper it is reported a total of 131 crowd-workers participating in the study, but for 600 tasks instead of 300, i.e., for both the ROTOWIRE and MLB datasets. Since we reproduced the experiment with the ROTOWIRE dataset, the number of unique participants is quite high, considering that we requested half of the HITs.

Furthermore, in the original paper it is reported

Table 1: Average number of Supported (#Supp) and Contradicting (#Contra) facts in game summaries for ROTOWIRE dataset, both for the original evaluation and the reproduction experiment (original results are extracted from Table 5 in Puduppully and Lapata (2021)). CV* column indicates the unbiased coefficient of variation of the reproduction scores for each system, computed following the method explained in Belz (2022). Systems significantly different from Macro are marked with an asterisk * (using one-way ANOVA with posthoc Tukey HSD tests; $p \leq 0.05$)

	Original		**Reproduction**			
	#Supp	#Contra	#Supp	CV*	#Contra	CV*
gold	3.63	0.07	3.36	52.17	0.66	175.7
Templ	7.57*	0.08	6.27*	42.86	0.90	234.26
ED+CC	3.92	0.91*	4.42	39.64	1.95*	119.23
RBF-2020	5.08*	0.67*	4.31	41.37	1.22	161.79
Macro	4.00	0.27	4.08	30.73	0.55	205.31

an inter annotator agreement of 0.44 for supported and 0.42 for contradicting facts, using Krippendorff's α. Of course, these values are calculated for the 600 tasks performed for both datasets, and the total of 131 crowd-workers. We calculated the same agreement measure, by adding the number of supported/contradicitng facts of each task, i.e., by adding the count of each of the four sentences in the HIT to have two scores per HIT: contradicting and supported facts scores. The results gave us an agreement of 0.188 for supported and 0.219 for contradicting facts. Therefore, the agreement in our evaluation is lower than in the original one, although we must take into account that the number of tasks and unique evaluators is different in our reproduction experiment, so a direct comparison is not really fair.

Looking at the original scores shown in Table 1 and as it is stated by Puduppully and Lapata (2021), the number of supported facts for Macro is comparable to gold and ED+CC (not statistically significant differences), but significantly smaller than Templ and RBF-2020. On the other hand, regarding the count of contradicting facts, Macro yields the smallest number among neural models. The number of contradicting facts for Macro is comparable to gold and Templ and significantly smaller than RBF-2020 and ED+CC.

Contrasting the original vs. reproduction results, we can see that for the Macro supported facts the score only differs in 0.08, but in the rest of the compared systems it differs more from the original study, reaching the largest difference in the Templ system (1.3, i.e., a 17% less supported facts in average). Comparing the results of the reproduction for the Macro supported facts with respect to the rest of the systems, Macro obtains significantly smaller values only with respect to the Templ system, while in the original study it is also significantly smaller compared with the RBF-2020 system.

Regarding the contradicting facts, for all the systems the reproduction scores are higher than in the original experiment, being the ED+CC system which yields the largest difference, with a 0.91 of counted contradicting facts in the original paper against a 1.95 in the reproduction experiment (i.e., an increase of the 114%). Surprisingly, the Macro system achieves the lowest score in terms of contradicting facts in our reproduction experiment, while in the original experiment the Templ system had the best performance. Looking at Macro results against the other systems, we can say that Macro achieves only significantly lower scores respect to the ED+CC system, whilst in the original experiment also significantly differences with the RBF-2020 were concluded.

It must be noted that if we pay attention to the unbiased coefficient of variation (CV*) in table 1, there is a big difference between the scores of the supported and contradicting facts. While CV* for supported facts is more stable, ranging from 30.73 to 52.17, the CV* for contradicting facts shows higher values, ranging from 119.23 to 234.26. It denotes a higher level of dispersion around the mean in the scores for contradicting facts.

After analyzing the results shown in Table 1, we can say that the general tendency observed from the reproduction results is similar to that of the results reported in the original study, despite differences in the score values. Table 2 summarizes the main differences between the conclusions obtained in the original experiment vs. those of our reproduction experiment. On the one hand, regarding supported facts, the scores are not very different from those of the original study. Comparing the Macro system with the rest of the systems evaluated, in the original paper the results achieved by the Macro system are comparable with gold and ED+CC system (the difference is not statistically significant), and significantly lower than Templ and RBF-2020 systems. In the reproduction experiment, it is concluded that Macro is comparable to gold, ED+CC, and RBF-2020, while only significantly lower scores are reported with respect to the Templ system. Thus, the tendency observed for supported facts is similar to the original study,

Table 2: Comparison of the conclusions from the original experiment by Puduppully and Lapata (2021) and our reproduction experiment, regarding the Macro system performance. For each type of facts checked, i.e., supported or contradicting, it is indicated with respect to which systems the Macro model is comparable or, on the contrary, obtains significantly lower scores.

Original	**Reproduction**
Supported	*Supported*
Comparable to gold and ED+CC	Comparable to gold, ED+CC, and RBF-2020
Sign. lower than Templ and RBF-2020	Sign. lower than Templ
Contradicting	*Contradicting*
Comparable to gold and Templ	Comparable to Templ, gold, and RBF-2020
Sign. lower than ED+CC and RBF-2020	Sign. lower than ED+CC

except for the RBF-2020, which in the reproduction experiment is comparable to the Macro system, instead of being statistically different.

On the other hand, if we look at the contradicting facts, the original study concluded that Macro results were comparable to gold and Templ, but significant differences were detected only with respect to ED+CC and RBF-2020. In the reproduction experiment, only significantly smaller scores are reported for ED+CC, whereas RBF-2020, Templ and gold yield results which are comparable to Macro. As in the case of the supported facts, the observed tendency is similar in the reproduction and original experiments, but now only significant differences are concluded in one of the two systems for which significant differences were detected in the original study.

This analysis of the scores allows us to say that in terms of supported facts, the reproduction study reports slightly better results for Macro than the original study. Furthermore, compared to the baseline systems, in the reproduction experiment only significantly smaller scores are obtained compared with one of the systems (i.e., Templ), which means that the amount of supported facts generated by Macro is comparable to more systems than in the original study. In terms of contradicting facts, the situation is the opposite. Despite a general increase in the number of contradicting facts for all the systems, only significantly smaller scores are reported with respect to one of the systems (i.e., ED+CC), while in the original study Macro generated significantly less contradicting facts than two other systems. As less generated contradicting facts is better, in this case the results can be considered slightly worse than in the original study.

5 Concluding Remarks and Future Work

In this work we performed a reproduction experiment of a human evaluation in NLP. Following the work by Puduppully and Lapata (2021), in the reproduced evaluation, a data-to-text system with a macro planning stage (Macro) is assessed in terms of contradicting/supported facts generated in the sports domain, i.e., ROTOWIRE dataset.

When counting the supported facts of the different systems, there is not a clear change pattern in the reproduction scores respect to the original ones. All of the scores are slightly different from the original ones, whether higher or lower. But, considering that having more supported facts is better, the Macro system shows a mildly improvement in the reproduction study in terms of score and also obtaining less statistically significant smaller scores with respect to the rest of systems in the comparison. Despite of that, the Templ system (Wiseman et al., 2017) is still the best in terms of supported facts, mainly because, as it is also pointed in the original paper for reproduction, the system essentially parrot facts.

Regarding the count of contradicting facts, there is a clear increase in general for all the systems and, surprisingly the system with the smallest number of contradicting facts is the Macro system, instead of the Templ system which was the best system in the original study. However, the Macro system produces statistically significant smaller scores only with respect to the ED+CC system.

The reproduction results show a similar tendency regarding supported facts, where the Templ system still produces the bigger number of supported facts. However, the tendency changes regarding contradicting facts. In addition to the general increase of contradicting facts, Templ and gold summaries,

which were the baselines with the less contradicting facts, are outperformed by Macro, being the model with the less contradicting facts generated.

There are certain factors in human evaluation that cause the results of a reproduction study, despite replicating all the settings, not to be exactly the same as those reported in the original study. One of the most distinguishing factors are the evaluators. In this case, the same AMT crowd-worker requirements were applied to select a profile of workers in the crowd-sourcing platform equal to that of the original study, but they will never be the same evaluators. Moreover, a different number of evaluators have participated than in the original study, since they could choose how many tasks to perform freely. This is obviously one of the reasons why a reproduction of a human evaluation can lead to a difference in the results.

In connection to the AMT crowd-worker requirements, the following experience with a worker from the platform is worth to be mentioned here. As mentioned in the section 3.2, following the original evaluation settings we indicated as a requirement to perform our task to have a minimum of 1,000 HITs completed. A few hours after launching the first mini-batch of the experiment, we received a message in which an experienced AMT crowd-worker welcomed us to the platform and very kindly told us the following: "If this is your first batch posted here, welcome to Mturk! Just a heads up, posting work with insufficient qualifications tends to yield some terrible results. I'd suggest making the qualifications 10,000 approved HITs". Since this was a reproducibility experiment, we had to stick to the conditions specified in the original paper and kept the minimum number of HITs at 1,000. Anyways, this advice is worth to be mentioned here for consideration in future experiments. Having seen that some of the results of the replicated experiment differed from the original, and that the agreement between the raters was poor, we believe that the minimum number of HITs required may have had an influence. When the original evaluation was launched (in 2021) this requirement was probably enough to achieve good results in the platform, but currently, as the worker recommended, probably we should increase the minimum number of required HITs to 10,000 in order to get equivalent results to those reported by Puduppully and Lapata (2021).

Taking advantage of the fact that this worker had contacted us, we asked him/her about a special qualification granted in AMT that we were curious about, despite not being used in our experiment: the so-called "Masters qualification". On the official AMT website there is no clear information about the requirements that crowd-workers must meet to obtain this qualification and what it exactly means, so we asked the worker what he/she knew about it and the worker told us that AMT is notoriously tight-lipped about the Masters qualification and even the workers of the platform do not know what is the criteria for granting this type of qualification. This fact made us think about the platform's lack of transparency even with workers and why sometimes AMT has bad reputation, despite being a powerful tool that so many people use.

In the light of these findings, this reproduction study emphasizes the critical importance of providing comprehensive details about human evaluations in NLP. The standarization of reporting practices for human evaluations by tools such as the Human Evaluation Datasheet (HEDS) in the framework of a common approach for reproduction, increases the reproducibility and, therefore reliability of any work. Thus, we encourage researchers to further document their NLP evaluations using these standards, with the aim of enhancing the quality of the works in the field.

As future work, we plan to repeat the evaluation for the MLB dataset with the aim of checking if reported results differ in a similar way as observed in the ROTOWIRE dataset. Also, we would like to reproduce the other human evaluation reported in the original paper, i.e., the quality of the generated texts in terms of grammar, conciseness and coherence, by comparing pairs of summaries.

Acknowledgments

This research was funded by MCIN/AEI/10.13039/501100011033 (grants PID2020-112623GB-I00, PID2021-123152OB-C21 and TED2021-130295B-C33), the Galician Ministry of Culture, Education, Professional Training, and University (grants ED431C2022/19 and ED431G2019/04). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program). It was also funded by the Univ. de Santiago de Compostela, Xunta de Galicia, and Spanish Ministry for Economic Affairs and Digital Transformation through the Nós (Ref. 2021-CP081) and ILENIA projects.

References

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

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, 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. 2022. 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.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5588–5594, Marseille, France. European Language Resources Association.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. 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.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Joelle Pineau. 2020. The machine learning reproducibility checklist v2.0.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maja Popović and Anya Belz. 2021. A reproduction study of an annotation-based human evaluation of MT outputs. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 293–300, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035, Florence, Italy. Association for Computational Linguistics.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. A hierarchical model for data-to-text generation. In *Advances in Information Retrieval*, pages 65–80, Cham. Springer International Publishing.

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.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Appendices

A Human Evaluation Datasheet (HEDS)

Here we provide an adapted version of the HEDS that shows all the preregistration details of the experiment. A copy of the original HEDS .json file and all the additional files mentioned below to reproduce the experiment are also shared as supplementary material in the submission of the paper.

Section 1: Paper and supplementary resources

Sections 1.1–1.3 record bibliographic and related information. These are straightforward and don't warrant much in-depth explanation.

Section 1.1: Details of paper reporting the evaluation experiment

Question 1.1.1: Link to paper reporting the evaluation experiment. Enter a link to an online copy of the the main reference (e.g., a paper) for the human evaluation experiment. If the experiment hasn't been run yet, and the form is being completed for the purpose of submitting it for preregistration, simply enter 'for preregistration'.
Answer: For preregistration.

Question 1.1.2: Which experiment within the paper is this form being completed for? Enter details of the experiment within the paper for which this sheet is being completed. For example, the title of the experiment and/or a section number. If there is only one human human evaluation, still enter the same information. If this is form is being completed for pre-registration, enter a note that differetiates this experiment from any others that you are carrying out as part of the same overall work.
Answer: This form is being completed to reproduce the human-based evaluation from the Section 6 of the paper "Data-to-text Generation with Macro Planning" available at https://arxiv.org/abs/2102.02723. Namely, we pay attention here only to the first study, that is "the count of supported and contradicting facts on the generated texts".

Section 1.2: Link to resources

Question 1.2.1: Link(s) to website(s) providing resources used in the evaluation experiment.Enter the link(s). Such resources include system outputs, evaluation tools, etc. If there aren't any publicly shared resources (yet), enter 'N/A'.
Answer: There is a public github repository in the arxiv paper, in which you can find the models and datasets (https://github.com/ratishsp/data2text-macro-plan-py) and the authors also provided by email a repository with the files needed to reproduce the evaluation (https://github.com/ratishsp/data2text-human-evaluation). All the material used to reproduce the evaluation and the details of the procedure will be available at https://drive.google.com/drive/folders/1ZySFzvZh-_2H8iJlBrkemG-9bJ0CFSFH?usp=sharing

Section 1.3: Contact details

This section records the name, affiliation, and email address of person completing this sheet, and of the contact author if different.

Section 1.3.1: Details of the person completing this sheet.

Question 1.3.1.1: Name of the person completing this sheet.
Answer: Javier González Corbelle, Jose María Alonso Moral

Question 1.3.1.2: Affiliation of the person completing this sheet.
Answer: Universidade de Santiago de Compostela, Spain

Question 1.3.1.3: Email address of the person completing this sheet.
Answer: j.gonzalez.corbelle@usc.es, josemaria.alonso.moral@usc.es

Section 1.3.2: Details of the contact author

Question 1.3.2.1: Name of the contact author. Enter the name of the contact author, enter N/A if it is the same person as in Question 1.3.1.1
Answer: N/A

Question 1.3.2.2: Affiliation of the contact author. Enter the affiliation of the contact author, enter N/A if it is the same person as in Question 1.3.1.2
Answer: N/A

Question 1.3.2.3: Email address of the contact author. Enter the email address of the contact author,

enter N/A if it is the same person as in Question 1.3.1.3
Answer: N/A

Section 2: System Questions

Questions 2.1–2.5 record information about the system(s) (or human-authored stand-ins) whose outputs are evaluated in the Evaluation experiment that this sheet is being completed for. The input, output, and task questions in this section are closely interrelated: the value for one partially determines the others,as indicated for some combinations in Question 2.3.

Question 2.1: What type of input do the evaluated system(s) take?

This question is about the type(s) of input, where input refers to the representations and/or data structures shared by all evaluated systems. This question is about input type, regardless of number. E.g. if the input is a set of documents, you would still select text: document below. Select all that apply. If none match, select 'other' and describe.

1. raw/structured data
2. deep linguistic representation (DLR)
3. shallow linguistic representation (SLR)
4. text: subsentential unit of text
5. text: sentence
6. text: multiple sentences
7. text: document
8. text: dialogue
9. text: other (please describe)
10. speech
11. visual
12. multi-modal
13. control feature
14. no input (human generation)
15. other (please describe)

Please describe:
Please provide further details for your above selection(s)

Question 2.2: What type of output do the evaluated system(s) generate? This question is about the type(s) of output, where output refers to the and/or data structures shared by all evaluated systems. This question is about output type, regardless of number. E.g. if the output is a set of documents, you would still select *text: document* below. Note that the options for outputs are the same as for inputs except that the *no input (human generation) option* is replaced with *human-generated 'outputs'*, and the *control feature* option is removed. Select all that apply. If none match, select 'other' and describe.

1. raw/structured data
2. deep linguistic representation (DLR)
3. Shallow linguistic representation (SLR)
4. text: subsentential unit of text
5. text: sentence
6. text: multiple sentences
7. text: document
8. text: dialogue
9. text: other (please describe)
10. speech
11. visual
12. multi-modal
13. human generated 'outputs'
14. other (please describe)

Please describe:
Please provide further details for your above selection(s)

Question 2.3: How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2? This question is about the task(s) performed by the system(s) being evaluated. This is independent of the application domain (financial reporting, weather forecasting, etc.), or the specific method (rule-based, neural, etc.) implemented in the system. We indicate mutual constraints between inputs, outputs and task for some of the options below. Occasionally, more than one of the options below may apply. Select all that apply. If none match, select 'other' and describe.

1. content selection/determination
2. content ordering/structuring
3. aggregation
4. referring expression generation
5. lexicalisation
6. deep generation
7. surface realisation (SLR to text)
8. feature-controlled text generation
9. data-to-text generation
10. dialogue turn generation
11. question generation
12. question answering
13. paraphrasing/lossless simplification

14. compression/lossy simplification
15. machine translation
16. summarisation (text-to-text)
17. end-to-end text generation
18. image/video description
19. post-editing/correction
20. other (please describe)

Please describe:
Please provide further details for your above selection(s)

Question 2.4: What are the input languages that are used by the system? This question is about the language(s) of the inputs accepted by the system(s) being evaluated. Select any language name(s) that apply, mapped to standardised full language names in ISO 639-1 (2019). E.g. English, Herero, Hindi. If no language is accepted as (part of) the input, select 'N/A'. Select all that apply. If any languages you are using are not covered by this list, select 'other' and describe.

1. Abkhazian
2. Afar
...
41. English
...
185. N/A (please describe)

Please describe:
Please provide further details for your above selection(s)

Question 2.5: What are the output languages that are used by the system? This field question the language(s) of the outputs generated by the system(s) being evaluated. Select any language name(s) that apply, mapped to standardised full language names in ISO 639-1 (2019). E.g. English, Herero, Hindi. If no language is generated, select 'N/A'.Select all that apply. If any languages you are using are not covered by this list, select 'other' and describe.

1. Abkhazian
2. Afar
...
41. English
...
185. N/A (please describe)

Please describe:
Please provide further details for your above selection(s)

Section 3: Sample of system outputs, evaluators, and experimental design

Section 3.1: Sample of system outputs

Questions 3.1.1–3.1.3 record information about the size of the sample of outputs (or human-authored stand-ins) evaluated per system, how the sample was selected, and what its statistical power is.

Question 3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment? Enter the number of system outputs (or other evaluation items) that are evaluated per system by at least one evaluator in the experiment. For most experiments this should be an integer, although if the number of outputs varies please provide further details here.

Answer: In the experiment, a total of 100 items are evaluated by at least one evaluator. Each of the items is composed of 4 summaries that must be rated. These 100 items are generations from the ROTOWIRE dataset. There are outputs generated by 5 different systems (20 from each). So, a total of 100 items are evaluated, from 5 different systems.

Question 3.1.2: How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment? Select one option. If none match, select 'other' and describe:

1. by an automatic random process
2. by an automatic random process but using stratified sampling over given properties
3. by manual, arbitrary selection
4. by manual selection aimed at achieving balance or variety relative to given properties

5. other (please describe)

Answer: We replicate the evaluation from the original paper, so we manually choose the same system outputs for evaluation. In the original paper these outputs were randomly selected.

Please describe: Please provide further details for your above selection(s)

Section 3.1.3: Statistical power of the sample size.

Question 3.1.3.1: What method was used to determine the the statistical power of the sample size?

Answer: In the paper taken as reference, no method, or criteria to determine the sample size ismentioned. In our reproduction we will evaluate the same number of summaries.

Question 3.1.3.2: What is the statistical power of the sample size? Enter the numerical results of a statistical power calculation on the output sample.
Answer: No method to determine the statistical power of the sample size was used.

Question 3.1.3.3: Where can other researchers find details of the script used? Enter a link to the script used (or another way of identifying the script). See, e.g., Card et al. (2020), Howcroft & Rieser (2021).
Answer: No method to determine the statistical power of the sample size was used.

Section 3.2: Evaluators

Questions 3.2.1–3.2.5 record information about the evaluators participating in the experiment.
Question 3.2.1: How many evaluators are there in this experiment? Enter the total number of evaluators participating in the experiment, as an integer.
Answer: N/A

Section 3.2.2: Evaluator Type

Questions 3.2.2.1–3.2.2.5 record information about the type of evaluators participating in the experiment.
Question 3.2.2.1: What kind of evaluators are in this experiment?

Select one option. These options should be valid for most experiments, but if not, select 'N/A' and describe why:

1. experts
2. non-experts
3. N/A (please describe)

Answer: The raters are crowdworkers required to be from English speaking countries, have a minimum of 1,000 previously completed tasks and have an approval rating in AMT greater than 98%.

Please describe:

Please provide further details for your above selection(s)
Question 3.2.2.2: Were the participants paid or unpaid? Select one option. These options should be valid for most experiments, but if not, select 'N/A' and describe why:

1. paid (monetary compensation)
2. paid (non-monetary compensation such as course credits)
3. not paid
4. N/A (please describe)

Question 3.2.2.3: Were the participants previously known to the authors? Select one option. These options should be valid for most experiments, but if not, select 'N/A' and describe why:

1. previously known to authors
2. not previously known to authors
3. N/A (please describe)

Please describe:
Question 3.2.2.4: Were one or more of the authors among the participants? Select one option. These options should be valid for most experiments, but if not, select 'N/A' and describe why:

1. evaluators include one or more of the authors
2. evaluators do not include any of the authors
3. N/A (please describe)

Please describe:
Question 3.2.2.5: Further details for participant type. Please use this field to elaborate on your selections for questions 3.2.2.1 to 3.2.2.4 above.
Answer: We take as reference the ReproHum global minimum wage per hour (UK living wage), that is GBP 10.90. Considering that each task will take about 4 minutes, an annotator can do 15 tasks per hour, so the payment per HIT will be GBP 10.9/15 = GBP 0.726 (0.88 USD = 0.8 EUR).

Question 3.2.3: How are evaluators recruited? Please explain how your evaluators are recruited. Do you send emails toa given list? Do you post invitations on social media? Posters on university walls? Were there any gatekeepers involved? What are the exclusion/inclusion criteria?
Answer: To recruit evaluators, we use the AMT

platform. The requisites for workers to be selected as valid are as follows: they are from English speaking countries, they have a minimum of 1,000 previously completed tasks and an approval rating greater than 98%.

Question 3.2.4: What training and/or practice are evaluators given before starting on the evaluation itself? Use this space to describe any training evaluators were given as part of the experiment to prepare them for the evaluation task, including any practice evaluations they did. This includes any introductory explanations they're given, e.g. on the start page of an online evaluation tool.
Answer: Before entering each task, evaluators are shown online the informed consent, the instructions of the task, how to read the different tables that will be shown, and an example task.

Question 3.2.5: What other characteristics do the evaluators have? Known either because these were qualifying criteria, or from information gathered as part of the evaluation. Use this space to list any characteristics not covered in previous questions that the evaluators are known to have, either because evaluators were selected on the basis of a characteristic, or because information about a characteristic was collected as part of the evaluation. This might include geographic location of IP address, educational level, or demographic information such as gender, age, etc. Where characteristics differ among evaluators (e.g. gender, age, location etc.), also give numbers for each subgroup.
Answer: The characteristics that evaluators have are the mentioned before: being from English speaking countries, having a minimum of 1,000 previously completed tasks and an approval rating greater than 98%.

Section 3.3: Experimental Design

Sections 3.3.1–3.3.8 record information about the experimental design of the evaluation experiment.
Question 3.3.1: Has the experimental design been preregistered? If yes, on which registry? Select 'Yes' or 'No'; if 'Yes' also give the name of the registry and a link to the registration page for the experiment.

1. yes
2. no

Please provide the name for, and link to the registration page for the experiment:
Please provide further details for your above selection(s)
Question 3.3.2: How are responses collected? Describe here the method used to collect responses, e.g. paper forms, Google forms, SurveyMonkey, Mechanical Turk, CrowdFlower, audio/video recording, etc.
Answer: Responses are collected via AMT.

Section 3.3.3: Quality assurance

Questions 3.3.3.1 and 3.3.3.2 record information about quality assurance.
Question 3.3.3.1: What quality assurance methods are used to ensure evaluators and/or their responses are suitable?

If any methods other than those listed were used, select 'other', and describe why below. If no methods were used, select *none of the above* and enter 'No Method' Select all that apply:

1. evaluators are required to be native speakers of the language they evaluate.
2. automatic quality checking methods are used during/post evaluation
3. manual quality checking methods are used during/post evaluation
4. evaluators are excluded if they fail quality checks (often or badly enough)
5. some evaluations are excluded because of failed quality checks 6. other (please describe)
7. none of the above

Please describe:
Question 3.3.3.2: Please describe in detail the quality assurance methods that were used. If no methods were used, enter 'N/A'
Answer: The task of the evaluators is to count the supported and contradicted facts on the generated texts. They are given two dropdowns to select the number of supported and contradicted facts detected, ranging from 0 to 20. So, when the sum of the supported and contradicted facts in a question is equal or higher than 20, the response is excluded, as there are no more than 20 facts to consider per sentence in none of the tasks

presented to the evaluators. Also, the ID of the evaluator that failed this quality check is saved in order to do not accept more HITs from this worker.

Section 3.3.3: Form/Interface

Questions 3.3.4.1 and 3.4.3.2 record information about the form or user interface that was shown to participants.

Question 3.3.4.1: Please include a link to online copies of the form/interface that was shown to participants. Please record a link to a screenshot or copy of the form if possible. If there are many files, please create a signpost page (e.g., on GitHub that contains links to all applicable resouces). If there is a separate introductory interface/page, include it under Question 3.2.4.

Answer: The HTML that will be used as template in AMT is available in the following link: https://drive.google.com/drive/folders/1ZySFzvZh-_2H8iJlBrkemG-9bJ0CFSFH?usp=share_link.

Question 3.3.4.2: What do evaluators see when carrying out evaluations? Describe what evaluators are shown, in addition to providing the links in 3.3.4.1.

Answer: Evaluators are shown first the informed consent they must fill due to ethic reasons and, then they can read the instructions of the task they must perform, together with an illustrative example, to get them familiarized with the task. Finally, they go into the questionnaire where they can accomplish the required task.

Question 3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations?

Select all that apply:

1. evaluators have to complete each individual assessment within a set time

2. evaluators have to complete the whole evaluation in one sitting

3. neither of the above (please describe)

Please describe:

Please provide further details for your above selection(s)

Question 3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback?

Select all that apply.

1. evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation

2. evaluators are told they can ask any questions during the evaluation

3. evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box

4. other (please describe)

5. None of the above

Please describe:

Question 3.3.7: What are the experimental conditions in which evaluators carry out the evaluations?

Multiple-choice options (select one). If none match, select 'other' and describe.

1. evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.

2. evaluation carried out in a lab, and conditions are the same for each evaluator

3. evaluation carried out in a lab, and conditions vary for different evaluators

4. evaluation carried out in a real-life situation, and conditions are the same for each evaluator

5. evaluation carried out in a real-life situation, and conditions vary for different evaluators

6. evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each evaluator

7. evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators

8. other (please describe)

Please describe:

Question 3.3.8: Briefly describe the (range of different) conditions in which evaluators carry out the evaluations. Use this space to describe the variations in the conditions in which evaluators carry out the evaluation, for both situations where those variations are controlled,and situations where they are not controlled. If the evaluation is carried out at a place of the evaluators' own choosing, enter 'N/A'

Answer: N/A

Section 4: Quality Criteria – Definition and Operationalisation

Questions in this section collect information about each quality criterion assessed in the single human evaluation experiment that this sheet is being completed for.

Many Criteria : Quality Criterion - Definition and Operationalisation

In this section you can create named subsections for each criterion that is being evaluated. The form is then duplicated for each criterion. To create a criterion type its name in the field and press the *New* button, it will then appear on tab that will allow you to toggle the active criterion. To delete the current criterion press the *Delete current* button.

Fact-checking

Section 4.1: Quality Criteria

Questions 4.1.1–4.1.3 capture the aspect of quality that is assessed by a given quality criterion in terms of three orthogonal properties. They help determine whether or not the same aspect of quality is being evaluated in different evaluation experiments. The three properties characterise quality criteria in terms of (i) what type of quality is being assessed; (ii) what aspect of the system output is being assessed; and (iii) whether system outputs are assessed in their own right or with reference to some system-internal or system-external frame of reference. For full explanations see Belz et al. (2020).

Question 4.1.1: What type of quality is assessed by the quality criterion?

1. Correctness
2. Goodness
3. Feature

Please describe:
Please provide further details for your above selection(s)

Question 4.1.2: Which aspect of system outputs is assessed by the quality criterion?

1. Form of output
2. Content of output
3. Both form and content of output

Please describe:
Please provide further details for your above selection(s)

Question 4.1.3: Is each output assessed for quality in its own right, or with reference to a system-internal or external frame of reference?

1. Quality of output in its own right
2. Quality of output relative to the input
3. Quality of output relative to a system-external frame of reference

Please describe:
Please provide further details for your above selection(s)

Section 4.2: Evaluation mode properties

Questions 4.2.1–4.2.3 record properties that are orthogonal to quality criteria (covered by questions in the preceding section), i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are more common than others).

Question 4.2.1: Does an individual assessment involve an objective or a subjective judgment?

1. Objective
2. Subjective

Please describe:

Question 4.2.2: Are outputs assessed in absolute or relative terms?

1. Absolute
2. Relative

Please describe:

Question 4.2.3: Is the evaluation intrinsic or extrinsic?

1. Intrinsic
2. Extrinsic

Please describe:

Section 4.3: Response elicitation

The questions in this section concern response elicitation, by which we mean how the ratings or other measurements that represent assessments for the quality criterion in question are obtained, covering what is presented to evaluators, how they select response and via what type of tool, etc. The

eleven questions (4.3.1–4.3.11) are based on the information annotated in the large scale survey of human evaluation methods in NLG by Howcroft et al. (2020).

Question 4.3.1: What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given. The name you use to refer to the quality criterion in explanations and/or interfaces created for evaluators. Examples of quality criterion names include Fluency, Clarity, Meaning Preservation. If no name is used, state 'N/A'.

Answer: Correctness of output relative to input (content)

Question 4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given. Copy and past the verbatim definition you give to evaluators to explain the quality criterion they're assessing. If you don't explicitly call it a definition, enter the nearest thing to a definition you give them. If you don't give any definition, state 'N/A'.

Answer: In the form provided the task to perform is described as "For each sentence, your task is to determine how many of the facts in the sentence are actually supported by the tables, and how many are contradicted by the tables". Also, some examples are provided.

Question 4.3.3: Are the rating instrument response values discrete or continuous? If so, please also indicate the size. Is the rating instrument discrete or continuous? When discrete, also record the number of different response values for this quality criterion. E.g. for a 5-point Likert scale, select *Discrete* and record the size as 5 in the box below. For two-way forced-choice preference judgments, the size would be 2; if there's also a no-preference option, enter 3. For a slider that is mapped to 100 different values for the purpose of recording assessments select discrete and record the size as 100. If no rating instrument is used (e.g. when evaluation gathers post-edits or qualitative feedback only), select *N/A*.

1. Discrete

2. Continuous

3. N/A

Please record the size of the instrument here: 21

Question 4.3.4: List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument. List, or give the range of, the possible values of the rating instrument. The list or range should be of the size specified in Question 4.3.3. If there are too many to list, use a range. E.g. for two-way forced-choice preference judgments, the list entered might be *A better, B better*; if there's also a no-preference option, the list might be *A better, B better, neither*. For a slider that is mapped to 100 different values for the purpose of recording assessments, the range *1–100* might be entered. If no rating instrument is used (e.g. when evaluation gathers post-edits or qualitative feedback only), enter 'N/A'.

Answer: 0-20

Question 4.3.5: How is the scale or other rating instrument presented to evaluators? If none match, select 'Other' and describe.

1. Multiple-choice options

2. Check-boxes

3. Slider

4. N/A (there is no rating instrument)

5. Other (please describe)

Please describe:

Question 4.3.6: If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument. If (and only if) there is no rating instrument, i.e. you entered 'N/A' for Questions 4.3.3–4.3.5, describe the task evaluators perform in this space. Otherwise, here enter 'N/A' if there *is* a rating instrument.

Answer: N/A

Question 4.3.7: What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)? Copy and paste the verbatim text that evaluators see during each assessment, that is intended to convey the evaluation task to them. E.g. *Which of these texts do you prefer?* Or *Make any corrections to this text that you think are necessary in order to improve it to the point where you would be happy to provide it to a client.*

Answer: Correct facts in sentence: dropdown.

Incorrect facts in sentence: dropdown.

Question 4.3.8: Form of response elicitation. If none match, select 'Other' and describe.

Explanations adapted from Howcroft et al. (2020).
1. (dis)agreement with quality statement
2. direct quality estimation
3. relative quality estimation (including ranking)
4. counting occurrences in text
5. qualitative feedback (e.g. via comments entered in a text box)
6. evaluation through post-editing/annotation
7. output classification or labelling
8. user-text interaction measurements
9. task performance measurements
10. user-system interaction measurements
11. Other (please describe)

Please describe:
Question 4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion? Normally a set of separate assessments is collected from evaluators and is converted to the results as reported. Describe here the method(s) used in the conversion(s). E.g. macro-averages or micro-averages are computed from numerical scores to provide summary, per-system results. If no such method was used, enter 'N/A'.
Answer: An average of the correct and incorrect facts is calculated for each system evaluated.

Question 4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion. Enter a list of methods used for calculating the effect size and significance of any results, both as reported in the paper given in Question 1.1, for this quality criterion. If none calculated, state 'None'.
Answer: The results of the different systems wil be compared using a one-way ANOVA with posthoc Tukey HSD tests to determine the significance of the results.

Section 4.3.11: Inter-annotator agreement

Questions 4.3.11.1 and 4.3.11.2 record information about inter-annotator agreement.

Question 4.3.11.1: Has the inter-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used? Select one option. If Yes, enter the methods used to compute any measures of inter-annotator agreement obtained for the quality criterion. If N/A, explain why.

1. yes
2. no
3. N/A

Please describe: Once the experiment finishes, the Krippendorff's agreement will be calculated.
Question 4.3.11.2: What was the inter-annotator agreement score? Enter N/A if there was none.
Answer: We expect an inter-annotator agreement score similar to the one reported in the paper that we took as reference: 0.44 for supported facts and 0.42 for contradicting facts.

Section 4.3.12: Intra-annotator agreement

Questions 4.3.12.1 and 4.3.12.2 record information about intra-annotator agreement.
Question 4.3.12.1: Has the intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used? Select one option. If Yes, enter the methods used to compute any measures of intra-annotator agreement obtained for the quality criterion. If N/A, explain why.

1. yes
2. no
3. N/A

Please describe: We only run the experiment once. To calculate the intra-annotator agreement the same evaluators must evaluate twice the same sentences.

Question 4.3.12.2: What was the intra-annotator agreement score? Enter N/A if there was none.
Answer: N/A

Section 5: Ethics

The questions in this section relate to ethical aspects of the evaluation. Information can be entered in the text box provided, and/or by linking to a source where complete information can be found.

Question 5.1: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee? Typically, research organisations, universities and other higher-education institutions require some form ethical approval before experiments involving human participants, however innocuous, are permitted to proceed. Please provide here the name of the body that approved the experiment, or state 'No' if approval has not (yet) been obtained.
Answer: This experimental evaluation is approved by the ethics committee of the University of Santiago de Compostela (Approval Date: December 22, 2022; Approval Ref.: USC 56/2022). The approval certificate was issued by D. José Manuel Cifuentes Martínez, the Head of the USC Ethics Committee.

Question 5.2: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions)? If yes, describe data and state how addressed. State 'No' if no personal data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements such as privacy and security was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.
Answer: No.

Question 5.3: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1 [2]? If yes, describe data and state how addressed. State 'No' if no special-category data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements relating to special-category data was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.
Answer: No.

Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes. Use this box to describe any *ex ante* or *ex post* impact assessments that have been carried out in relation to the evaluation experiment, such that the assessment plan and process, well as the outcomes, were captured in written form. Link to documents if possible. Types of impact assessment include data protection impact assessments, e.g. under GDPR. Environmental and social impact assessment frameworks are also available.
Answer: No.

[2] urlhttps://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited

B 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.
   - 1.1 *min_wage_your_lab*: the minimum wage in the country/region where your lab is based.
   - 1.2 *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.
   - 1.3 *original_study_wage*: what participants were paid in the original study.
   - 1.4 *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).
   - 1.5 *uk_living_wage*: set to the equivalent in your currency of £10.90 GBP, this is the project global minimum.
2. Calculate the $reproduction\_wage$ by following the below steps:
   - 2.1 *min_wage = MAX(min_wage_your_lab, min_wage_your_participant)*
   - 2.2 IF *original_study_min_wage* == NONE; THEN *original_study_min_wage = original_study_wage*

2.3 *multiplier = (original_study_wage / original_study_min_wage)*

2.4 *wage = min_wage * multiplier*

2.5 *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.)