ReproHum #0124-03: Reproducing Human Scores on Neural REG Models

Maurice Langner
Linguistic Data Science Lab
Ruhr-University Bochum
Maurice.Langner@rub.de

Abstract

In the context of the ReproNLP'26 shared task, I report on a single-criterion reproduction study of a human evaluation experiment for neural referring expression generation models (Castro Ferreira et al., 2018a), which has already been reproduced once by Mahamood (2024) for the ReproHum 2024 shared task. The experiments reported on in this paper therefore seek to second the findings from both previous experiments.

1 Introduction

The reproduced empirical study evaluates the performance of different generative models for referring expressions. Referring expression generation, or short, REG, is a subtask of NLG that deals with deictic descriptions of objects in contrast sets or discourse (van Deemter, 2016). Castro Ferreira et al. (2018a) test neural REG models against two baselines on a delexicalised subset of the WebNLG corpus (Gardent et al., 2017; Castro Ferreira et al., 2018b). Input to the models are 2 to 7 RDF triples, which are transformed into a surface expression. Castro Ferreira et al. (2018a) performed a human evaluation for assessing the quality of the generated output from three neural models and two baselines, which will be reproduced in this study. In the context of the 2024 shared task on reproducibility (Belz and Thomson, 2024), this experiment was already successfully reproduced once (Mahamood, 2024) and attested excellent reproducibility. The experiment is reproduced on the basis of a predefined evaluation datasheet (Shimorina and Belz, 2022). This includes all design and documentation flaws (Thomson et al., 2024) that were present in the original experiment. These flaws and resulting issues in reproducibility, which have already been identified in Mahamood (2024), are reported below. Since conclusions must be drawn with caution when only one reproduction is available (Belz et al., 2021b; Belz and Thomson, 2023), this reproduction experiment for the ReproNLP 2026 shared task (Belz et al., 2026) is supposed to solidify the findings in the original study and the first reproduction.

2 Study Details

The study under discussion shows model-generated referring expressions, as well as the tabular data from which the expression was generated, to participants and lets them rate the referring expressions on three different criteria, namely, **fluency**, **grammaticality** and **clarity**, on three separate, vertically stacked 7-point Likert scales. In agreement with Mahamood (2024) and the ReproHum multilab study design, only the criterion **clarity** will be reproduced, which is supposed to encode *how clearly the text expresses the data*. The study description does not specify whether clarity is quantitative (*how many factoids are correctly expressed?*) or qualitative (*is the data expressed in an easily comprehendible way?*) or a combination of both. Neither Mahamood (2024) nor this reproduction resolves this ambiguity, since the original study should be reproduced as faithfully as possible. This means that any bias and interpretative variability also influence the result of reproductions.

The referring expressions are generated by five different models, three of which are neural models with RNN cells (Rumelhart and McClelland, 1987; Hochreiter and Schmidhuber, 1997) with different decoder components and attention mechanisms, namely `Seq2Seq`, `CAtt` (Bahdanau et al., 2015) and `HierAtt` (Libovický and Helcl, 2017). Furthermore, there are two non-neural baseline models, namely the `Ferreira` model, which uses Naive Bayes (Castro Ferreira et al., 2016), and the `OnlyNames` model, the latter producing reformatted Wikipedia IDs without complex coreference. Finally, the set of items to be rated by

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 1094–1103

participants also comprises the gold label string from the corpus, or a randomly chosen gold label string if there are several in the data. This allows for a comparison between the gold label and model outputs with respect to human assessment.

As described in the original paper, the empirical study encompasses six different lists with 24 different items (or trials) each. Each item consists of an input - output pair. The output is a generated referring expression. The input is a set of two to seven RDF triples. Lists are constructed such that the proportion of models and RDF triple number are balanced within each list and across lists. Each participant is assigned one of six lists. The input triples are identical across lists, but the outputs differ. For some specific trial $n \in$ [1,...,24], each list provides the output of a different model for n, such that system output quality is comparable across lists. An example trial and the distribution of model outputs across lists is exemplified in Table 1. Each system output is rated by 10 participants.

3 Issues in Reproduction

While the trials on each list are documented and therefore reproducible, the repository does not state in which order the trials in each list were shown to the participants. A script in the repository (humaneval/stats.py) randomizes the order once and writes pairs of links for each list into a database of the form (LIST_ID, URL, NEXT_URL), which hence encodes trial order, but the corresponding database is lost. Executing the code does not reproduce the order due to the absence of a fixed random seed. Apart from that, it is known that trials were not randomized per participant, which means that every candidate of a list rates the trials in the same order. From the HTML files, it was possible to reconstruct the first item on each list. Hence, the first element of each list was fixed according to the original one and the order of the remaining items was randomly selected. During the study, the order of items is not changed. Furthermore, personal data, such as age, gender and country, that is collected on the introductory page, and their granularity were adapted to the requirements of the ReproHum project. Finally, a few sentences of the introductory text were modified grammatically in order to match the limitation to a single evaluation criterion. No other study-related changes were made, in order to keep the induced effect on the study outcome as small as possible (Belz et al., 2021a, 2022).

With regard to the original study data, it is noteworthy that the file containing the participants' evaluations, which is the base of the computed results in the paper, is missing in the repository. Consequently, the raw data and recycled code from (humaneval/stats.py) were used to filter out surplus data points and reconstruct the originally evaluated dataset [1].

4 Technical Details of the Reproduction

The original web application[2] is implemented in PHP; the authors use a separate database server. Database structure must be inferred from the PHP code and the existing SQL queries in the code in order to make the original app functional, because no database schema is present in the GitHub repository (Mahamood, 2024). The design of the User Interface, in other words the website design, is determined by externally sourced CSS. These CSS files determine size, look and layout of HTML forms and pages the participants interact with, and are crucial for faithfully reproducing the human evaluation study. A javascript implementation of a timer is responsible for guaranteeing a minimum time of 30 seconds each participant has to spend on each of the 24 trials of the presented list.

In order to adapt the web application to the blueprint flask application for the ReproHum shared task (Watson and Gkatzia, 2023), the web-technological foundation is swapped, while preserving user interface, design and functionality.

The same CSS files for the website design are used, but instead of pre-generating 144 different html files for all trials across all lists, a templating library creates a generic interface. This is used by the web app to generate HTML markup for each of the trials on the fly. The markup used for the model output of each trial is extracted from the original HTML files and re-used in the templates. Therefore, the highlighting and formatting of the tabular input data as well as the target sentence are identical to the original study. In this way, a

[1]The raw study data contained 61 participants, and 4 of the participants provided more than 24 data points, most probably due to submitting data for a trial twice (user error). The original script just sequentially iterates data entries from top to bottom and stops as soon as data limits (24 trials of the 6 pre-defined lists for 10 participants each) are reached. Neither the paper nor code contains any information on filter criteria, which is why the method is assumably sound.

[2]The original web application source code can be found here: https://github.com/ThiagoCF05/NeuralREG

List	Model	model Output
L1	Ferreira	alan shepard is united states who was born in new hampshire> on 1923-11-18 . he graduated from nwc, m.a. 1957 . shepard was chosen by nasa in 1959 and he crewed apollo.
L2	Seq2Seq	alan shepard is an american who was born in new hampshire on 1923-11-18 . he graduated from nwc, m.a. 1957 . he was chosen by nasa in 1959 and he crewed apollo 14 .
L3	CAtt	alan shepard is an american citizen who was born in new hampshire on 1923-11-18 . he graduated from nwc, m.a. 1957 . he was chosen by nasa in 1959 and he crewed apollo 14 .
L4	HierAtt	alan shepard is an american who was born in new hampshire on 1923-11-18 . he graduated from nwc, m.a. 1957 . shepard was chosen by nasa in 1959 and he crewed apollo 14 .
L5	Gold	alan shephard is a us citizen who was born in new hampshire on november 18th , 1923 . he graduated from nwc with an ma in 1957 . he was chosen by nasa in 1959 and he crewed apollo 14 .
L6	Only	alan shepard is united states who was born in new hampshire on 1923-11-18 . alan shepard graduated from nwc, m.a. 1957 . alan shepard was chosen by nasa in 1959 and alan shepard crewed apollo 14 .

Table 1: Example model outputs for trial 20185. Each list contains the output of a different model for this trial and its corresponding RDF triples. Each item is rated by 10 different participants that are associated with the respective list. The highlighted **gold** sentence is taken from the corpus and not generated by a model.

faithful reproduction of the look and feel of the original user interface is guaranteed.

While many human evaluation studies present all items participants have to rate on a single page, such that they need to scroll down to the next element, Castro Ferreira et al. (2018a) only present one trial per page and submit. This also means that participants cannot go back to former items and adjust ratings on the basis of later trials and decisions. Since this design has a strong influence on the participants' decisions, javascript is used for hiding previous trials and showing the next, while also permitting to send the whole data as one json object on the final submit, as implemented in the blueprint application for the ReproGen project.

As middleware, the flask web framework is used, data storage is realized as a PostgreSQL database. Gunicorn serves as WSGI server and Nginx as reverse proxy. For easy deployment, docker images are created for the web application, the database as well as for Nginx. Study data is stored in persistent docker volumes in order to minimize risk of data loss. For easier access to the data, a URL is added for downloading the database tables as CSV files directly through the web app[3]. The application was deployed on an AWS EC2 instance. Code and data of this reproduction experiment are made available online[4].

5 Reproduction Results (QRA++)

The reproduction study was conducted twice with disjoint sets of participants. Due to organisational miscommunication, the first experiment was run with demographic filters (ReproA), which did not agree with the original experiment setup. Therefore, a second experiment with no demographic filter (ReproB) was conducted. This gives the opportunity to measure whether demographic constraints and language proficiency have a strong influence on quality estimates with regard to the involved criteria.

The two experiments ReproA and ReproB are identical in every detail (also the order of trials), except for the recruited participants. Study configurations of all studies are shown in Table 2. For the reproduction study, participants were recruited via Prolific instead of Mturk. In contrast with Castro Ferreira et al. (2018a) and Mahamood (2024), ReproA defines three participant controls. Nationality is limited to US, UK, Australia and Canada.

[3]This functionality is not visible to participants and does not alter by any means the study or user interface.

[4]https://github.com/MMLangner/Reprohum-2026-0124-03

Aspect	Original	ReproB (ReproA)	Mahamood (2024)
Item count	144	144	144
Systems count	6	6	6
Participant count	60	60	60
Participants per Item	10	10	10
Items per Participant	24	24	24
Platform	Amazon MTurk	Prolific	Prolific
Compensation	unknown	£13.50 / hr	£12.00 / hr
Age	36 Years avg.	55% $<$ 35 yrs. (77% $>$ 35 yrs.)	majority 18-24 43%
Gender Split	27F, 33M	33M, 27F (30F, 29M, 1D)	35F, 25M
Quality Criterion	fluency grammaticality clarity	clarity	clarity
Participant controls	unknown	(1. nationality US/UK/CAN/AUS) 2. min. accept rate 99% 3. completed studies 200	none
English Proficiency	Native:44 Fluent:14 Basic:2	Native:41 (59) Fluent:17 (1) Basic:2 (0)	Native:37 Fluent: 21 Basic:2

Table 2: Comparison of study aspects and demographic information on participants for all four studies. Where ReproB differs from ReproA, values for ReproA are added in round brackets.

Participants were required to have an acceptance rate of 99% and a minimum number of completed studies of 200. No other selection controls were employed. For ReproB, the demographic filter was removed and participants from ReproA were excluded.

With regard to demographic features, ReproB matches the original participant data well. For ReproB, 55 % of participants were younger than 35 years, which means that the age distribution is complementary to ReproA, and much closer to the original study. Only 41 are native speakers, 17 are fluent and 2 basic speakers. This distribution agrees well with the original study. 29 participants are from US/UK, 14 preferred not to answer, 14 are from European countries and one each from Australia, Canada and Mexico.

For ReproA, the resulting demographic configuration of the participant group differs strongly from both the original and Mahamood (2024)'s reproduction. Participants of ReproA were significantly older, with 77% older than 35 Years and hence the average of the original study. In Mahamood (2024), the majority of participants were much younger (18-24). The ratio of native speakers of English is at 98%, due to the nationality criterion. Therefore, the participants of ReproA are more proficient in the target language than those in Castro Ferreira et al. (2018a) and Mahamood (2024), which provided ratios for non-native speakers of 26.6% and 38.3% respectively. This difference in age and language proficiency may have an influence on the study results.

5.1 Type I: Scores and CV*

Table 3 lists the average Likert scale rating per system for the original study, ReproA, ReproB and the previous reproduction by Mahamood (2024). First, all average ratings of ReproB and ReproA are higher than 5, which is not the case for the other two studies. Further, except for NeuralREG+`HierAtt` in ReproA, all models in ReproA and ReproB have a higher average rating than in the other studies. In ReproB and ReproA, NeuralREG+`CAtt` is the best performing model with an average rating of 5.48 and 5.59 and comes closest to the gold sentences with 5.64 and 5.62 respectively. NeuralREG+`HierAtt` provides the second best average score in ReproB with 5.31, but the worst average with 5.11 in ReproA. Except NeuralREG+`Seq2Seq` in ReproB and NeuralREG+`HierAtt` in ReproA, neural models outperform the baselines, but not always by a large margin, as seen in the difference between *Ferreira* (5.27) and NeuralREG+`Seq2Seq` with a distance of 0.09. In ReproB, *OnlyNames* scores unusually

high at 5.27. Summing up, all four studies only agree on that the gold sentences score highest.

For ReproB, *OnlyNames* is the furthest away from the original study with a CV of 9.03, while the closest being *Ferreira* with 2.26. For the other models, CV values range between 4.25 and 5.85, which indicate a very good reproducibility. In general, the CV values for the comparison between ReproB and Mahamood (2024) are larger than those between ReproB and the original study, with 5 CV scores between 6.45 and 12.03. This means that ReproB is far closer to the original study than it is to Mahamood (2024), while Mahamood (2024) is closer to the original than ReproB is to the original. This roots in the fact that average ratings in ReproB and Mahamood (2024) deviate from the original study in opposite directions: ReproB provides higher ratings, while Mahamood (2024) provides lower ratings. When comparing ReproB, original data and Mahamood (2024) (N=3), CV scores are good between 4.04 and 7.38.

The coefficients between ReproB and ReproA are very small, which indicates a fit as good as between the original data and Mahamood (2024)'s data, underlining once more the closeness of results between all studies. For the comparison between Mahamood (2024) and ReproB, *Ferreira* is much closer than in the comparison between ReproA and Mahamood (2024), but coefficients for NeuralREG+`HierAtt` are much higher at 6.45.

Coefficients of Variation (CV) are also low ($<$ 9.4) for the comparisons between ReproA and the original paper, while for the model NeuralREG+`HierAtt`, it's even below 1, where the scores are near identical. Some of these CV values are close to those between the original study and Mahamood (2024), while for NeuralREG+`Seq2Seq` and *OnlyNames*, they are further apart. For coefficients of variation between ReproA and Mahamood (2024)'s reproduction, values for *Ferreira* and NeuralREG+`CAtt` are larger ($>$ 14.4). When computing the coefficients across all three studies (N=3), CV values range between 1.41 and 8.99. Smallest coefficients and largest consensus can be found for model NeuralREG+`HierAtt`, which is the worst model in ReproA, while being positioned in the middle field of ReproB and the original study and being the best model in Mahamood (2024).

In general, coefficients indicate a good fit between all studies, with only 6 percent of all CV values between 12 and 14 that indicate a medium fit (Belz et al., 2025). The CV values, calculated across all four experiments according to the measurand definition[5] in Belz and Thomson (2026) and shown in Table 7, emphasize the excellent reproducibility of results.

5.2 Type II: Correlations

The correlation between ReproB and the original data is positive significant (Pearson's r: 0.91, p=0.012; Spearman's ρ: 0.83, p=0.041), the same applies to the correlation between ReproB and Mahamood (2024) (Pearson's r: 0.89, p=0.017; Spearman's ρ: 0.81, p=0.049). While there are significant positive correlations (Spearman's ρ: 0.84, p=0.03; Pearson's r: 0.78, p=0.065) between original results and those reported in Mahamood (2024), Spearman's ρ (0.66, p=0.16) and Pearson's r (0.74, p=0.09) for ReproA show an insignificant, weaker positive correlation with the original data. The Correlations between ReproA and Mahamood (2024) are even weaker (Spearman's ρ: 0.35, p=0.49; Pearson's r: 0.45, p=0.37). At least for the correlations, it seems that ReproB, the experiment without demographic filter, better fits the original data than ReproA (with demographic filter).

5.3 Type IV

The following list repeats the main results determined in the original study (Castro Ferreira et al., 2018a, p. 1966):

- A All three neural models scored higher than the baselines on all metrics
- B NeuralREG+`CAtt` approaches the *gold* label scores
- C There is no clear distinction between RDF triple sizes
- D Baselines (*Ferreira* and *OnlyNames*) are not statistically significantly different
- E Neural models are not significantly different from baselines for clarity
- F There is no significant difference between neural models
- G Gold label texts have significantly higher scores than baselines and NeuralREG+-`Seq2Seq` for the clarity criterion

Claim A is only partially supported by ReproB. NeuralREG+`HierAtt` scores higher and NeuralREG+`Seq2Seq` scores lower than the baselines.

[5]The command line tool can be found at https://github.com/DCU-NLG/qra

System	Original	ReproA	Mahamood (2024)	ReproB
OnlyNames	4.90	5.20	4.92	5.27
Ferreira	4.93	5.27	4.69	5.02
NeuralREG+Seq2Seq	4.97	5.36	4.97	5.21
NeuralREG+CATT	5.26	5.59	4.97	5.48
NeuralREG+HierAtt	5.13	5.11	5.04	5.31
Gold	5.42	5.62	5.22	5.64

Table 3: Mean Avg ratings by human participants for each system

CV* values	Orig. x ReproA	Orig. x Ma24	Ma24 x ReproA	N=3
OnlyNames	7.39	0.51	6.88	5.12
Ferreira	8.27	6.28	14.53	8.99
NeuralREG+Seq2Seq	9.34	0.0	9.34	6.71
NeuralREG+CATT	7.44	7.02	14.44	8.87
NeuralREG+HierAtt	0.48	2.20	1.71	1.41
Gold	4.41	4.62	9.02	5.53
CV* values	Orig. x ReproB	ReproA x ReproB	Ma24 x ReproB	N=3
OnlyNames	9.03	1.65	8.52	6.31
Ferreira	2.26	6.01	8.53	5.37
NeuralREG+Seq2Seq	5.85	3.50	5.85	4.18
NeuralREG+CATT	5.02	2.42	12.03	7.38
NeuralREG+HierAtt	4.25	4.74	6.45	4.04
Gold	4.84	0.43	9.45	5.80

Table 4: CV* values for each system with normalized scales (starting at 0) for cross-study comparability. Numbers for Mahamood (2024) were also re-computed with normalisation for better comparability.

Claim B is fully supported by ReproB with 0.16 distance between NeuralREG+CAtt and gold label sentences. Although numerical backing for claim C is scarce, it is a valid interpretation for the diffuse differences between triple sizes in ReproB. Average scores vary between 5.1 and 5.7, with the two highest scores for RDF triple counts 3 and 5 in ReproB. There is no clear proportional or inversely proportional relation between RDF triple count and average score. For further claims, the following conclusions root in the significance test results (Wilcoxon rank test) shown in Tables 5 and 6. Claim D is fully supported by the reproduction data; baselines do not show a significant statistical difference. Claim E is not fully supported by ReproB; NeuralREG+CAtt is not different from *OnyNames*, but it is from *Ferreira*. Still, reproduction provides statistic indication that NeuralREG+CAtt is the closest model to the gold sentences and hence is supportive for the underlying hypothesis. ReproB does not fully support Claim F, since NeuralREG+Seq2Seq is worse than the *OnlyNames* baseline. The reproduction data fully supports Claim G that gold sentences score significantly higher than the two baselines and NeuralREG+Seq2Seq.

Again, ReproA only partially supports Claim A. While models NeuralREG+CAtt and NeuralREG+Seq2Seq score higher than both baselines for the clarity criterion in ReproA, NeuralREG+HierAtt scores lower than both baselines. ReproA fully supports Claim B with 0.03 average rating difference between NeuralREG+CAtt and the gold sentences. With regard to claim C, ReproA does not provide any interpretable relation between RDF triple count and average score either. Claim D is fully supported by ReproA. Claim E is not fully supported by ReproA, because NeuralREG+CAtt is significantly different from both baselines. Since in ReproA, NeuralREG+CAtt is the best and NeuralREG+HierAtt the worst, data also contradicts Claim F that neural models are scoring on a comparable level. ReproA validates Claim G that gold sentences score significantly higher than the two baselines and NeuralREG+Seq2Seq. Claims and their support are summarized in Table 8. The aspect of ReproA that is inconsistent with the findings in the original study is the low performance of the

	Model	only	seq2seq	ferreira	hieratt	catt
0	Gold	5968.5 (0.002)	6215.0 (0.058)	6032.5 (0.012)	5660.0 (0.001)	7342.0 (0.680)
1	only		8581.0 (0.257)	8633.0 (0.932)	8179.5 (0.557)	6532.5 (0.015)
2	seq2seq			8215.0 (0.434)	7325.0 (0.124)	7061.0 (0.145)
3	ferreira				8528.0 (0.469)	7031.5 (0.022)
4	hieratt					6057.5 (0.001)

Table 5: Wilcoxon rank test (statistic (P-value)) between system ratings for ReproA.

	Model	only	seq2seq	ferreira	hieratt	catt
0	Gold	6193.5 (0.009)	6610.5 (0.003)	6217.5 (0.0)	6558.0 (0.029)	6958.5 (0.214)
1	only		7885.5 (0.616)	7853.0 (0.105)	9344.5 (0.695)	7297.0 (0.115)
2	seq2seq			7966.0 (0.215)	7750.0 (0.414)	7280.0 (0.065)
3	ferreira				7646.5 (0.058)	6204.5 (0.001)
4	hieratt					8496.0 (0.312)

Table 6: Wilcoxon rank test (statistic (P-value)) between system ratings for ReproB.

NeuralREG+`HierAtt` model, which simply does not fit previously observed patterns and generalisations. This is of special concern, since NeuralREG+`HierAtt` is the best-performing model in Mahamood (2024) and the second-best in both original data and ReproB.

6 Inter Annotator Agreement

Krippendorff's alpha is computed for the ten participants of each list, and then the average over lists is calculated, in order to assess Inter Annotator Agreement. In the original paper, no IAA was reported. Therefore, the data from the original experiment was used to compute IAA, resulting in an average of 0.14 across lists, with individual agreements per list ranging from 0.07 to 0.28. In ReproB, agreement ranges from 0.03 for list 3 to 0.40 for list 2, the average agreement lies at 0.17. For ReproA, IAA lies at 0.15 and shows the same weak overall agreement. It varies depending on the list of trials, between 0.10 and 0.24. Hence one can conclude that the average degree of inter annotator agreement is equally low, which points at inherent issues with rating the respective criteria **clarity** on a Likert scale. Apart from that, the strong variation between lists may indicate a difference in task complexity between lists despite the original authors' care in balancing RDF Triple count and system exposure.

7 Discussion

The reproduction results of the two reproduction experiments ReproB (without demographic filter) and ReproA (with demographic filter) were analyzed and compared with the data of the original paper and the first reproduction by Mahamood (2024). The only claim supported by all three studies (and four experiments) is that the human-written gold-label sentences from the WebNLG corpus receive the highest scores on average. The scores for baseline models and the three neural REG generators vary strongly across the experiments. Noteworthy is the fact that coefficients of variation are lowest on the NeuralREG+`HierAtt` model that takes three different ranking positions in the four experiments, among those best and worst. Since only very few differences in scores are significant, variation may be due to a large baseline of noise. This noise may originate from perceptual differences between participants groups of different age and language proficiency, interpretative room given by the description of the measurand **clarity** or the different trial order, which was not reported in the documentation of the original experiment. Given the fact that in the original study, even for the less subjective criterion **grammaticality**, which is generally well reproducible (Belz et al., 2021b), the distance between worst and best model is only 0.49 score points (4.68 and 5.17), a higher number of participants would have been needed in order to produce reliable and significant results on model ranking.

The difference between ReproB and ReproA

Type of Result	QC	System	Measure applied	Degree of reproducibility ($n = 4$)		
				System level	QC level	Study level
Type I	QC1	OnlyNames Ferreira NeuralREG+Seq2Seq NeuralREG+CATT NeuralREG+HierAtt Gold	(mean) CV*↓	5.06 6.54 5.05 6.87 3.01 4.78	5.22	5.22
Type II	QC1	all all all	mean r ↑ mean ρ ↑ W↑	n/a n/a n/a	0.746 0.662 0.741	n/a
Type IV	QC1	all	P ↑	n/a	0.756	0.756

Table 7: QRA reproducibility assessment for four comparable experiments (n=4), Castro Ferreira et al. (2018a), Mahamood (2024), ReproB, and ReproA. QC1 = clarity; n/a = measure does not apply at this level.

Claim	Mahamood (2024)	ReproB	ReproA
A	True	Partial	Partial
B	False	True	True
C	True	True	True
D	NA	True	True
E	NA	Partial	Partial
F	True	Partial	False
G	True	True	True

Table 8: Overview over claims and their support (True | False | Partial)

with regard to their fit to the original data does not indicate any significant influence of demographic filtering of participants on reproducibility. For some models, the agreement with the original data improved without demographic filtering, while it degraded for a few others. Still, this variation may be subject to the already mentioned noise, and may not be causally related to the differences in language proficiency or average age of participants.

8 Conclusion

I presented the reproduction of a human evaluation study on the output of referring expression generation algorithms on a subset of the WebNLG corpus, which is the second reproduction of the original experiment. With regard to the general reproducibility of the study, the data shows a very good fit to the original reproduction, with only very few coefficients being on only a medium level of agreement. The strong variation in model rank between studies is therefore not a sign of issues in reproducibility, but more likely a symptom of the task complexity, fuzziness of the evaluation criterion *clarity*, and the bias introduced by different trial orders in the experiments. Finally, one can consider this reproduction successful and supportive of the methodological approach taken in the ReproNLP project.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021a. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021b. 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.

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 Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia. ELRA and ICCL.

Anya Belz and Craig Thomson. 2026. Quantified reproducibility assessment for four common types of evaluation results in nlp/ml. *Computational Linguistics*, pages 1–10.

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

Anya Belz, Craig Thomson, Javier González Corbelle, and Malo Ruelle. 2025. The 2025 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 1002–1016, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018a. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018b. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Saad Mahamood. 2024. ReproHum #0124-03: Reproducing human evaluations of end-to-end approaches for referring expression generation. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 250–254, Torino, Italia. ELRA and ICCL.

David E. Rumelhart and James L. McClelland. 1987. *Learning Internal Representations by Error Propagation*, pages 318–362.

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.

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

Kees van Deemter. 2016. *Computational Models of Referring: A Study in Cognitive Science*. The MIT Press.

Lewis Watson and Dimitra Gkatzia. 2023. Unveiling NLG human-evaluation reproducibility: Lessons learned and key insights from participating in the ReproNLP challenge. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 69–74, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Appendix

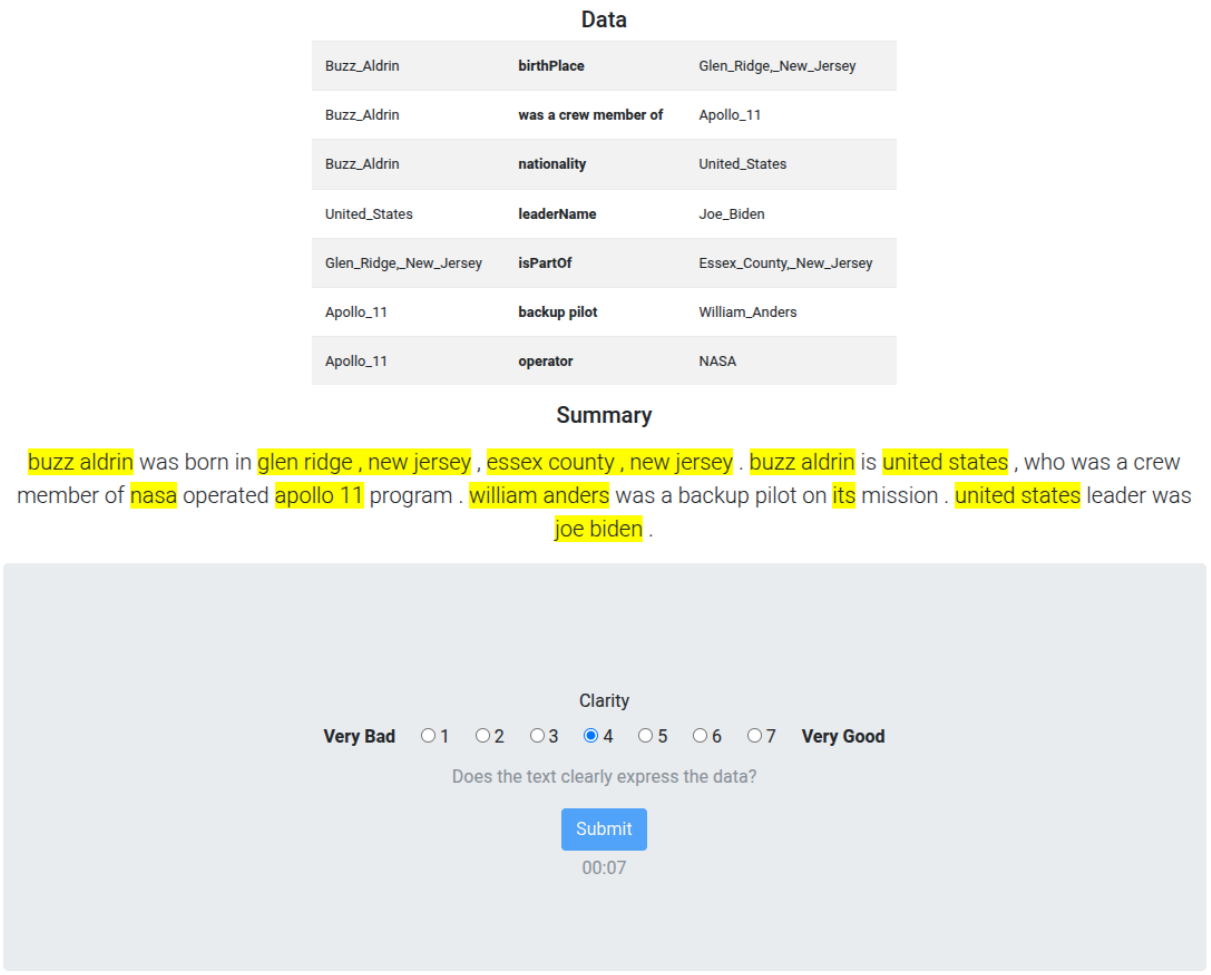

Figure 1: Screenshot of the reproduced user interface