Same Trends, Different Answers: Insights from a Replication Study of Human Plausibility Judgments on Narrative Continuations

Yiru Li, Huiyuan Lai, Antonio Toral, Malvina Nissim
CLCG, University of Groningen / The Netherlands
y.li.170@student.rug.nl
{h.lai, a.toral.ruiz, m.nissim}@rug.nl

Abstract

We reproduced the human-based evaluation of the *continuation of narratives* task presented by Chakrabarty et al. (2022). This experiment is performed as part of the ReproNLP Shared Task on Reproducibility of Evaluations in NLP. Our main goal is to reproduce the original study under conditions as similar as possible. Specifically, we follow the original experimental design and perform human evaluations of the data from the original study, while describing the differences between the two studies. We then present the results of these two studies together with an analysis of similarities between them. Inter-annotator agreement (Krippendorff's alpha) in the reproduction study is lower than in the original study, while the human evaluation results of both studies have the same trends, that is, our results support the findings in the original study.

1 Introduction

Reproduction studies of human evaluations in the field of Natural Language Processing (NLP) are attracting increasing attention (Belz et al., 2021b, 2022b). Due to the inherent limitations of automatic evaluation, especially in Natural Language Generation tasks which often imply high variability in the output, human evaluation is often considered to provide more reliable assessments (van der Lee et al., 2019). However, initial results observed in the context of ReproHum[1], a coordinated, multi-lab reproducibility project which the present work is also part of, suggest that the majority of human evaluations in NLP face the challenge of being unreproducible due to various reasons (Belz et al., 2023). This clashes with the importance of ensuring high levels of experimental reproducibility, which has been gaining increasing recognition in the NLP community (Fokkens et al., 2013; Belz et al., 2021a, 2022a).

In the context of our participation in the ReproNLP Shared Task on Reproducibility of Evaluations in NLP (Track C – ReproHum Project)[2], this paper reports on our experience when trying to reproduce as closely as possible a previously run human evaluation. Specifically, we aimed to reproduce human evaluations conducted by Chakrabarty et al. (2022) on the continuations of narratives generated with various systems or written by humans. In order to harmonise and coordinate all replication efforts within ReproHum, the project leaders have created a spreadsheet that each lab in charge of a reporduction experiment was asked to fill in and submit, acting as pre-registration for the replication. This Human Evaluation Datasheet (HEDS) is included in Appendix B. Following the shared reproduction approach provided by the ReproHum's coordinators, we first summarize the original study explaining the task addressed, and the human evaluation setting (Section 2), followed by our replicated experiment (Section 3). Although we did try to perform our new experiments under conditions as similar as possible to those of the original study, we still ended up with some differences between our setup and the original paper (e.g. we raise the payment to give the annotator a fairer reward); we discuss these in detail. Finally, we report and analyze the results obtained in our reproduction study by comparing them to the original experiments (Section 4), and draw some conclusions on the feasibility of a full experimental reproduction (Section 5).

2 Original Study

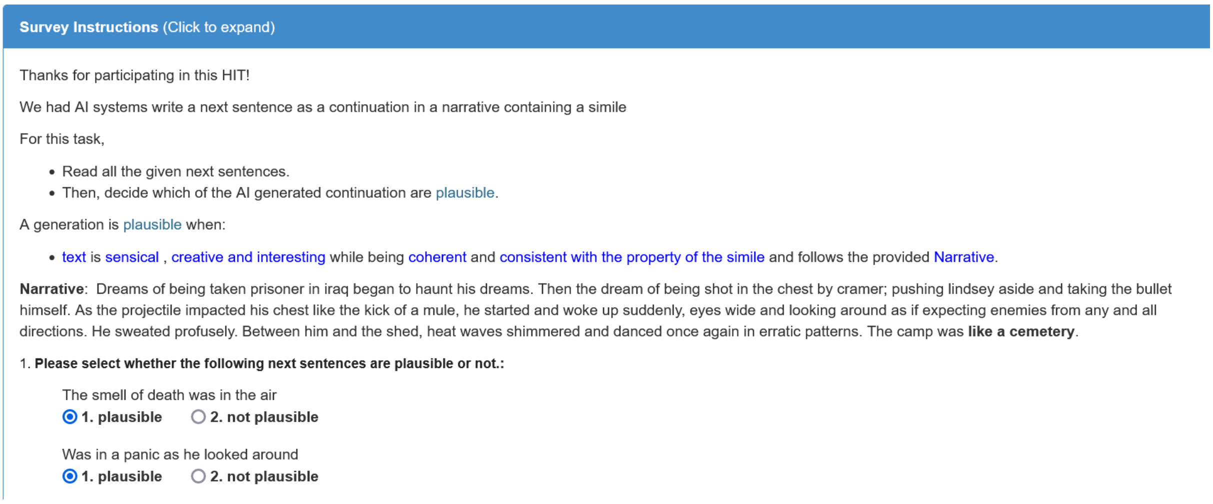
We aim to repeat the experiment conducted in "*It's not Rocket Science: Interpreting Figurative Language in Narratives*" by Chakrabarty et al. (2022).

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

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

Given Narrative	Continuations	Produced by	Plausible
Dreams of being taken prisoner in iraq began to haunt his dreams. Then the dream of being shot in the chest by cramer; pushing lindsey aside and taking the bullet himself. As the projectile impacted his chest like the kick of a mule, he started and woke up suddenly, eyes wide and looking around as if expecting enemies from any and all directions. He sweated profusely. Between him and the shed, heat waves shimmered and danced once again in erratic patterns. The camp was **like a cemetery**.	The smell of death was in the air	Model (baseline)	yes
	Was in a panic as he looked around	Model (+Context)	yes
	The usual welcomed silence is not welcomed here...it makes for crazy dreams.	Human	no
	You could hear a pin drop with the lack of sounds.	Human	yes

Table 1: An example of narrative ending in simile with corresponding continuations either generated by NLP systems or written by humans.



Survey Instructions (Click to expand)

Thanks for participating in this HIT!

We had AI systems write a next sentence as a continuation in a narrative containing a simile

For this task,

- Read all the given next sentences.
- Then, decide which of the AI generated continuation are plausible.

A generation is plausible when:

- text is sensical , creative and interesting while being coherent and consistent with the property of the simile and follows the provided Narrative.

Narrative: Dreams of being taken prisoner in iraq began to haunt his dreams. Then the dream of being shot in the chest by cramer; pushing lindsey aside and taking the bullet himself. As the projectile impacted his chest like the kick of a mule, he started and woke up suddenly, eyes wide and looking around as if expecting enemies from any and all directions. He sweated profusely. Between him and the shed, heat waves shimmered and danced once again in erratic patterns. The camp was **like a cemetery**.

1. **Please select whether the following next sentences are plausible or not.:**

The smell of death was in the air
● **1. plausible** ○ **2. not plausible**

Was in a panic as he looked around
● **1. plausible** ○ **2. not plausible**

Figure 1: A screenshot of the annotation interface.

This paper studies the interpretation of two figures of speech in narratives, namely *idioms* and *similes*, by means of a generation task.

2.1 Task and Models

The task consists in producing a plausible continuation of a given paragraph ending with a figure of speech, ensuring such continuation is coherent with the narrative and complies with the meaning of the figurative expression. A plausible continuation would serve as an indication that the given figure of speech is interpreted correctly. Table 1 shows an example of a provided narrative with human- or machine-generated continuations, some of which are deemed plausible, and some implausible.

To perform the task, the large pre-trained model GPT-2 XL (Radford et al., 2019) is fine-tuned on narrative-continuation pairs. Also, the authors propose two knowledge-enhanced models ("context-enhanced model" and "literal-enhanced model") that add, respectively, some context or a literal explanation of the figurative expression at the beginning of the narrative.

The continuations generated by different systems are assessed by means of human evaluation and also used for comparison with human-written ones.

2.2 Human Evaluation Settings

The original paper includes two types of human evaluations for both the simile task and the idiom task: absolute evaluation and comparative evaluation. The absolute evaluation asks the worker to evaluate whether the single continuation is plausible, independent of other continuations. The comparative evaluation asks the workers to compare multiple continuations and then choose the most plausible (neither or all are plausible are also possible options). We reproduce only the absolute evaluation as described in the original paper.

For the absolute evaluation, the authors of the original paper randomly sampled 25 narratives for each task, with each narrative containing 5 corresponding human-written continuations for the simile task or 3 for the idiom task. Three

Parameter	Original Setting	Replicated Setting
Reward	0.50 (U.S. Dollar)	**2.21** (U.S. Dollar)
Max Assignments	3	3
Assignment Duration	2 (hour)	2 (hour)
Auto Approval Delay	3 (day)	3 (day)
Expires in	7 (day)	7 (day)
Annotators	7 (simile) + 4 (idiom)	**75** (simile) + **75** (idiom)

Table 2: Parameters of the HIT publication settings, and the changed setting is marked bold.

continuations generated by the baseline GPT2-XL model, context-enhanced model, and literal-enhanced model are assessed by means of human evaluation along with human references.

The evaluation was conducted on *Amazon Mechanical Turk* (MTurk), a crowdsourcing platform, on which requesters may publish so-called Human Intelligence Tasks (HITs) for workers to complete. Each HIT (survey) was designed to have two major parts with instructions: the first part is an example, and the second part is the evaluation questions unique for each HIT. Figure 1 shows a screenshot of the annotation interface.

The example in the HIT is the same for all HITs of the same tasks, and the continuations of one of the selected similes/ idioms are evaluated in the second part. The example has the same layout as the questions: it includes one narrative which uses a given simile/idiom, the meaning of the presented simile/idiom, three model-generated continuations, and several human-written continuations. For the simile task, five human-written continuations are presented; for the idiom task, three are presented.

In all HITs, the positions of continuations were not randomly shuffled, i.e., the first continuation to be evaluated is always generated by the baseline model. Also, the workers are instructed to answer all questions, but it is technically possible for them to submit a HIT with questions unanswered (the script does not include a force-answering mechanism). For each continuation, the workers are instructed to answer a binary question, specifying whether any given continuation is judged as plausible or not.

Each HIT was completed by three unique workers, and each worker was rewarded $0.5 for completing one HIT. Seven and four unique workers were recruited for the simile task and the idiom task, respectively, as we could infer from the provided result file (this information was not included in the original paper, and it is unclear how this was enforced or allowed on the crowdsourcing platform). In the end, 25 HITs were put up for evaluation for each of the two tasks, and 3 responses for each HIT were collected, resulting in 75 responses for each task, and 150 responses in total. We did not observe any rejected or republished HIT in the collected responses, and since no approval time was included, we infer that all HITs were automatically approved.

3 Reproduction study

As mentioned, we only replicated the absolute evaluation from the original paper. Three differences between the reproduced experiment setting and the original one exist.

First of all, we recruited a total of 75 workers for each task with no additional requirements. This was done after thorough consideration: the total number and requirements of workers employed for absolute evaluation are not mentioned in the original paper. Still from the file containing the result data, we could infer, via anonymised ids, that the total number of annotators is much smaller than the number of HITs. However, since all HITs were published in one row, the selecting criteria for workers were unclear and we received no further clarification from the original authors; hence, we chose to also publish the HITs with no additional restrictions on workers in one row for each task, which ultimately led to recruiting one worker for each HIT. No control on whether one worker can work on both the idiom and simile tasks was put in place: in other words, one worker can potentially work on at most two HITs in total, one HIT of each task. Due to invalid results received, we re-published some of the HITs to obtain new assessments so that evaluations from a total of 127 workers were collected. In the original study, no rejecting or re-publishing of HITs was observed, but one invalid result is included in the original outputs for the simile task.

	Idiom		**Simile**	
	original	**reproduced**	**original**	**reproduced**
GPT2-XL	56	76	60	68
+Context	68	**92**	68	72
+Literal	48	68	76 (60)	**80**
Human	**80**	68	**88**	68
Difference Rate (%)	38.7		34	

Table 3: Summarized results of original and replicated experiments. The result we fail to reproduce is marked in red, with our calculated result in parentheses. The best result of each task is marked in bold. We also calculated the difference rate between each evaluated result from the replicated study and the original study to find how well our replicated results agree with the original result. See Table 4 and 5 for more details.

Secondly, we raised the monetary compensation from \$0.5 per HIT to \$2.21 per HIT, following the general recommendation of the ReproHum project to meet the minimum hourly salary in the UK, assuming it takes 10 minutes to finish one HIT properly. Besides setting differences, we received some invalid responses due to the original survey layout and thus had to re-publish some HITs. Table 2 presents key HIT parameters in the original and the present study.

Thirdly, we changed the examples given in the idiom tasks. Only the specific simile examples were made available by the authors, therefore we chose a narrative including an idiom and its corresponding continuations from the development set and then used them as examples for the crowdworkers.

4 Results

In this section, we report our results compared to the original experiment, and the reproducibility assessment. For each narrative-continuations pair we collect three responses, and whether a continuation is plausible is determined by majority voting, following what was described in the original paper.

As a first check, we assess inter-annotator agreement (IAA) using Krippendorff's α, which is appropriate for categorical labels attached to text spans, and which was used in the original experiment. The original experiment reports Krippendorff's $\alpha = 0.68$, while our replicated experiment shows Krippendorff's $\alpha = 0.11$ and the Krippendorff's α of the original experiment is 0.33 using our calculation method. This discrepancy might have to do with the fact that in our replication there are many more annotators, and with the way the score was calculated (accounting or not for the same annotator possibly doing more HITs in the original study).

The original paper reports quantitative results of each model for each task, and describes the reported data as the "percent of times that the generation from each of the models and human-written references was chosen as plausible by the majority of workers."

Since in each task we only collect assessments from each worker for one single continuation per model, there is no confusion on how to calculate the quantitative results. However, responses of multiple human-written continuations from each worker are collected, and the original paper did not detail how they came to the reported results for the human-produced continuations. After several attempts, 7 out of 8 the original results of the plausibility of human-written continuations were successfully reproduced using the following procedure:

1. determine whether a human-written continuation is plausible using majority voting;
2. count the total number of plausible continuations for each task;
3. divide the total by the number of human-written continuations in each HIT (3 and 5 for HITs in the idiom and simile task, respectively);
4. round up the calculated mean to an integer;
5. divide the rounded mean by the number of HITs (25 for both tasks) to calculate the percentage.

The calculated results are shown in Table 3.[3] The best result in each column is marked in bold, while the only result that our recalculation procedure described above could not reproduce as reported in the original study is marked in red (we include our recalculated result in brackets).

Surprisingly, our replicated experiments show that knowledge-enhanced models outperform humans, which was not the case in the original study. One plausible assumption is that the workers recruited in the original study have all been given additional training on evaluating continuations. Another possible reason is that the workers in the original study might unconsciously think that the second half of continuations is more plausible. Each of them works on multiple HITs, and in each HIT the first three continuations are always model-generated continuations and the rest are human-written continuations. The second problem is avoided in our replicated study as we avoided letting one worker evaluate several HITs. Nevertheless, both the performance difference and low IAA suggest that normal workers cannot fully understand, or reach an agreement on determining whether a continuation is plausible, with only the example and instructions given on the HIT page.

Overall, the original paper concludes from the results that "a knowledge-enhanced model outperformed the baseline GPT-2 model...the context model was favored for idioms while the literal model was favored for similes," and the general trend of our results, albeit at times largely different in scores, also supports this conclusion.

5 Conclusions

Although the results of our replicated experiment support the general findings of the original paper, the human evaluation process of the original paper could not be fully reproduced properly.

Two aspects need particular attention. First, the reproducing process is intrinsically difficult. Even though we tried our best, and we gained substantial help from the original authors, several questions still emerged during the replication stage which could not be answered. The detail of the worker recruitment process for example was not available and might be not fully known to the original authors either, due to platform specifications that can be not in full control of the researcher. We did stick to the original crowdsourcing platform used although it would not have been our primary choice, also due to logistic issues related to payment and, as said, full control over workers' recruitment.

Secondly, as shown in Table 3, the two rounds of experiments disagree with each other on more than one-third of continuations. Comparing the replicated results to the original results, we see that the high difference rates indicate disagreement between the reproduction results on the original output. We draw the conclusion from these mentioned problems that human evaluation of the plausibility of continuations, no matter the generated ones or the human written ones, is precarious.

Acknowledgments

We are very grateful to the anonymous reviewers for their useful comments, which contributed to strengthening this paper. We would also like to thank the annotators for helping us evaluate the data. Finally, we thank the reproHum project leaders for all their help including bridging contacts with the authors of the original study and project coordination.

References

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

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021b. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th*

[3] Table 4 and Table 5 in Appendix A show the detailed results collected from the original and the replicated experiments.

International Conference on Natural Language Generation, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.

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

Anya Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

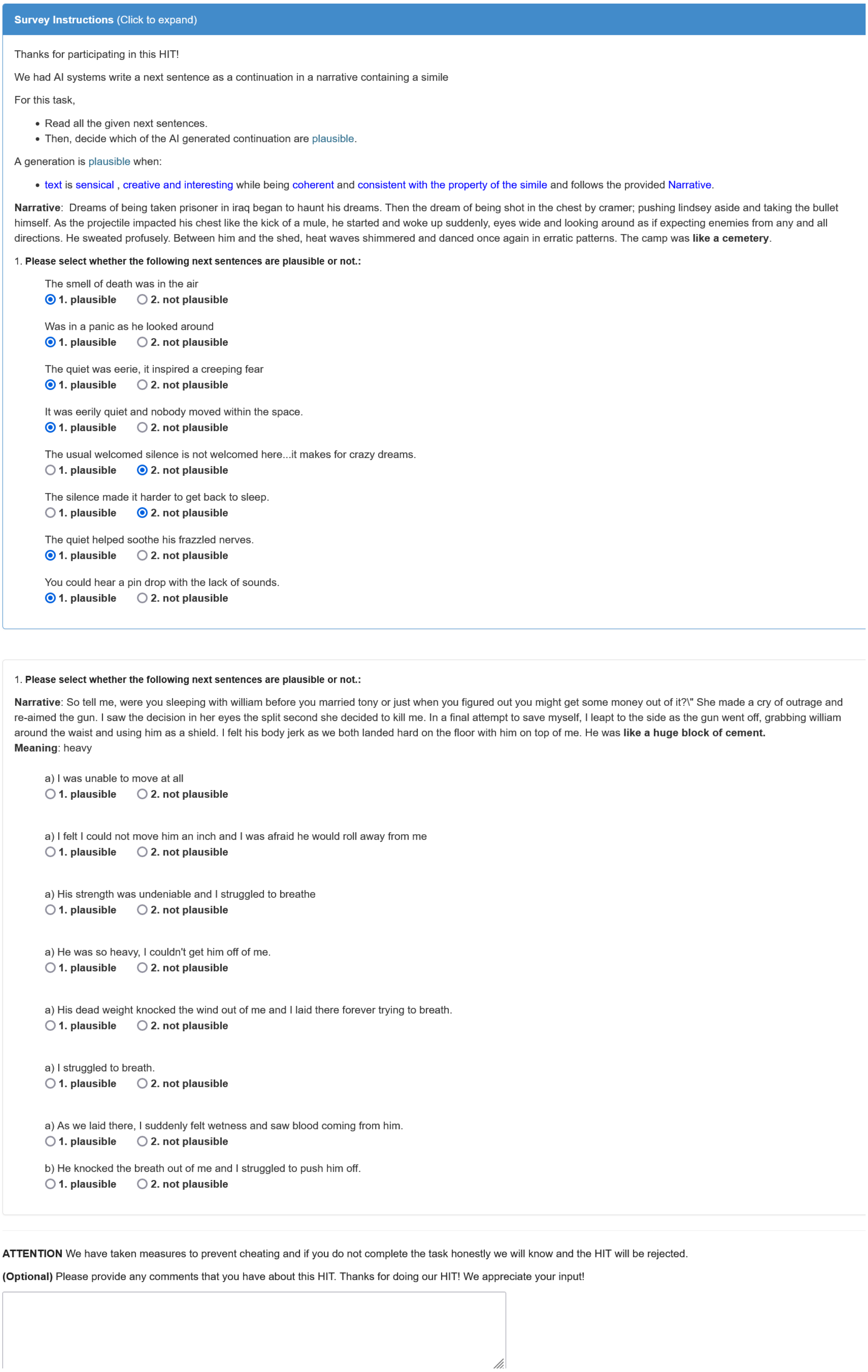
A Examples and Result Tables

	GPT2-XL	+Context	+Literal	H. 1	H. 2	H. 3
Idiom Results	1	1	0	1	1	1
	0	1	1	1	1	1
	0	0	1	1	0	1
	1	1	0	1	1	1
	0	1	0	1	1	1
	0	1	0	0	1	1
	1	1	0	1	1	0
	1	1	1	1	0	1
	1	0	1	1	0	0
	1	0	0	1	1	1
	0	1	1	0	1	1
	1	1	1	1	1	0
	0	1	0	1	1	1
	0	0	1	1	1	1
	0	0	1	1	0	1
	0	1	1	1	1	0
	1	0	0	0	1	1
	1	1	0	1	1	0
	1	1	0	1	1	1
	0	1	0	1	1	1
	1	0	0	1	1	1
	1	1	1	1	1	1
	0	0	1	1	0	1
	1	1	0	0	0	1
	1	1	1	1	0	1
Difference Rate (%) Overall: 38.7	36	40	52	32	44	28

Table 4: Results of the original voted plausibility of continuations generated/ collected for idioms. “H.” is the abbreviation of Human Reference. 1 represents plausible continuation, and 0 represents non-plausible continuation. If the determined plausibility of the continuation is different in replicated study, the value is marked red.

	GPT2-XL	+Context	+Literal	H. 1	H. 2	H. 3	H. 4	H. 5
Simile Results	0	1	1	1	1	1	0	1
	1	1	1	1	1	0	1	1
	1	0	0	1	1	1	1	1
	0	1	1	0	1	1	1	1
	1	0	1	1	1	1	1	1
	0	0	1	1	1	1	1	1
	0	1	1	1	1	1	1	1
	1	1	0	1	1	1	1	1
	1	1	1	1	1	1	1	1
	0	1	0	1	1	1	1	0
	1	1	0	1	1	1	1	0
	1	1	1	1	1	1	0	1
	1	1	1	1	1	1	1	1
	1	0	1	1	1	1	1	1
	1	0	1	0	1	1	1	1
	0	0	0	1	1	1	1	0
	1	1	0	1	1	1	1	1
	1	1	1	1	0	1	1	1
	0	1	1	1	1	0	1	0
	1	1	0	1	1	1	1	1
	1	1	1	1	1	1	0	0
	0	1	1	1	1	1	1	1
	0	0	0	1	1	1	1	0
	0	0	0	1	1	1	1	1
	1	1	0	1	0	1	0	0
Difference Rate (%) Overall: 34	40	28	36	40	48	20	12	48

Table 5: Results of the original voted plausibility of continuations generated/ collected for similes. "H." is the abbreviation of Human Reference. 1 represents plausible continuation, and 0 represents non-plausible continuation. If the determined plausibility of the continuation is different in the replicated study, the value is marked red.



Survey Instructions (Click to expand)

Thanks for participating in this HIT!

We had AI systems write a next sentence as a continuation in a narrative containing a simile

For this task,

- Read all the given next sentences.
- Then, decide which of the AI generated continuation are plausible.

A generation is plausible when:

- text is sensical , creative and interesting while being coherent and consistent with the property of the simile and follows the provided Narrative.

Narrative: Dreams of being taken prisoner in iraq began to haunt his dreams. Then the dream of being shot in the chest by cramer; pushing lindsey aside and taking the bullet himself. As the projectile impacted his chest like the kick of a mule, he started and woke up suddenly, eyes wide and looking around as if expecting enemies from any and all directions. He sweated profusely. Between him and the shed, heat waves shimmered and danced once again in erratic patterns. The camp was **like a cemetery**.

1. **Please select whether the following next sentences are plausible or not.:**

The smell of death was in the air
● **1. plausible** ○ **2. not plausible**

Was in a panic as he looked around
● **1. plausible** ○ **2. not plausible**

The quiet was eerie, it inspired a creeping fear
● **1. plausible** ○ **2. not plausible**

It was eerily quiet and nobody moved within the space.
● **1. plausible** ○ **2. not plausible**

The usual welcomed silence is not welcomed here...it makes for crazy dreams.
○ **1. plausible** ● **2. not plausible**

The silence made it harder to get back to sleep.
○ **1. plausible** ● **2. not plausible**

The quiet helped soothe his frazzled nerves.
● **1. plausible** ○ **2. not plausible**

You could hear a pin drop with the lack of sounds.
● **1. plausible** ○ **2. not plausible**

1. **Please select whether the following next sentences are plausible or not.:**

Narrative: So tell me, were you sleeping with william before you married tony or just when you figured out you might get some money out of it?\" She made a cry of outrage and re-aimed the gun. I saw the decision in her eyes the split second she decided to kill me. In a final attempt to save myself, I leapt to the side as the gun went off, grabbing william around the waist and using him as a shield. I felt his body jerk as we both landed hard on the floor with him on top of me. He was **like a huge block of cement.**
Meaning: heavy

a) I was unable to move at all
○ **1. plausible** ○ **2. not plausible**

a) I felt I could not move him an inch and I was afraid he would roll away from me
○ **1. plausible** ○ **2. not plausible**

a) His strength was undeniable and I struggled to breathe
○ **1. plausible** ○ **2. not plausible**

a) He was so heavy, I couldn't get him off of me.
○ **1. plausible** ○ **2. not plausible**

a) His dead weight knocked the wind out of me and I laid there forever trying to breath.
○ **1. plausible** ○ **2. not plausible**

a) I struggled to breath.
○ **1. plausible** ○ **2. not plausible**

a) As we laid there, I suddenly felt wetness and saw blood coming from him.
○ **1. plausible** ○ **2. not plausible**

b) He knocked the breath out of me and I struggled to push him off.
○ **1. plausible** ○ **2. not plausible**

ATTENTION We have taken measures to prevent cheating and if you do not complete the task honestly we will know and the HIT will be rejected.

(Optional) Please provide any comments that you have about this HIT. Thanks for doing our HIT! We appreciate your input!

Figure 2: Example HIT question page.

B HEDS Sheet

B.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.

1.1 Details of paper reporting the evaluation experiment

1.1.1 Link to paper reporting the evaluation experiment.
ReproHum: pre-experiment record

1.1.2 Which experiment within the paper is this form being completed for?
Absolute evaluation of plausibility (idiom and simile) in Section 5.

1.2 Link to resources

1.2.1 Link(s) to website(s) providing resources used in the evaluation experiment.
https://drive.google.com/drive/folders/
1ruTV4tnkfzTkGuF8VnmxgQr2ToQ3R
gDO?usp=sharing

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.

1.3.1 Details of the person completing this sheet

1.3.1.1 Name of the person completing this sheet.
Huiyuan Lai

1.3.1.2 Affiliation of the person completing this sheet.
University of Groningen

1.3.1.3 Email address of the person completing this sheet.
h.lai@rug.nl

1.3.2 Details of the contact author

1.3.2.1 Name of the contact author.
Malvina Nissim

1.3.2.2 Affiliation of the contact author.
University of Groningen

1.3.2.3 Email address of the contact author.
m.nissim@rug.nl

B.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.

2.1 What type of input do the evaluated system(s) take?
6. text: multiple sentences

2.2 What type of output do the evaluated system(s) generate?
5. text: sentence

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?
17. end-to-end text generation

2.4 What are the input languages that are used by the system?

41. English

2.5 What are the output languages that are used by the system?

41. English

B.3 Sample of system outputs, evaluators, and experimental design

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.

3.1.1 How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment?

25 outputs per system

3.1.2 How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment?

1. by an automatic random process

3.1.3 Statistical power of the sample size.

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

N/A. Follow the original experiment.

3.1.3.2 What is the statistical power of the sample size?

N/A

3.1.3.3 Where can other researchers find details of the script used?

N/A

3.2 Evaluators

Questions 3.2.1–3.2.5 record information about the evaluators participating in the experiment.

3.2.1 How many evaluators are there in this experiment?

N/A

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.

3.2.2.1 What kind of evaluators are in this experiment?

2. non-experts

3.2.2.2 Were the participants paid or unpaid?

1. paid (monetary compensation)

3.2.2.3 Were the participants previously known to the authors?

2. not previously known to authors

3.2.2.4 Were one or more of the authors among the participants?

2. evaluators do not include any of the authors

3.2.2.5 Further details for participant type.

N/A

3.2.3 How are evaluators recruited?

Post tasks in the crowdsourcing platform (MTurk).

3.2.4 What training and/or practice are evaluators given before starting on the evaluation itself?

We can provide evaluators with detailed guidelines and examples of generated sentences along with plausible assessments. However, it is not known if guidelines and examples were provided in the original paper.

3.2.5 What other characteristics do the evaluators have?

To ensure the quality of annotations, we will require that workers have an acceptance rate of at least 99%. No other demographic constraints are considered, only English as mother tongue (see below). Nothing is known regarding this from the original paper.

3.3 Experimental Design

Sections 3.3.1–3.3.8 record information about the experimental design of the evaluation experiment.

3.3.1 Has the experimental design been preregistered? If yes, on which registry?

2. no

3.3.2 How are responses collected?

Mechanical Turk

3.3.3 Quality assurance

Questions 3.3.3.1 and 3.3.3.2 record information about quality assurance.

3.3.3.1 What quality assurance methods are used to ensure evaluators and/or their responses are suitable?

1. evaluators are required to be native speakers of the language they evaluate.
2. automatic quality checking methods are used during/post evaluation
4. evaluators are excluded if they fail quality checks (often or badly enough)

3.3.3.2 Please describe in detail the quality assurance methods that were used.

2. = pre-selection based on master qualification on MTurk + post-selection based on minimum completion time required

4. = if non masters then excluded; if completion time too short, evaluators excluded.

Unclear, but unlikely, if quality control was done in original experiment and in case what (in paper 100% retention of evaluators)

3.3.4 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.

3.3.4.1 Please include a link to online copies of the form/interface that was shown to participants.

N/A

3.3.4.2 What do evaluators see when carrying out evaluations?

A task instruction, a short narrative and its meaning, and six outputs

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

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

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

5. None of the above

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

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

3.3.8 Briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

N/A

B.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.

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).

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

1. Correctness

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

2. Content of output

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?

2. Quality of output relative to the input

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).

4.2.1 Does an individual assessment involve an objective or a subjective judgment

2. Subjective

4.2.2 Are outputs assessed in absolute or relative terms?

1. Absolute

4.2.3 Is the evaluation intrinsic or extrinsic?

1. Intrinsic

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).

4.3.1 What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given.

Coherence

4.3.2 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.

Coherence: Given a short narrative containing an idiomatic expression, the generated next sentence in the story is plausible.

4.3.3 Are the rating instrument response values discrete or continuous? If so, please also indicate the size.

1. Discrete

Size of the instrument: 0 or 1

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.

0 or 1

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

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.

N/A

4.3.7 What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)?

Title: Choose if generated next sentence in a story containing an idiom is plausible.

Description: Given a short narrative containing an idiomatic expression, annotators need to choose if generated next sentence in the story is plausible.

4.3.8 Form of response elicitation. If none match, select 'Other' and describe.
1. (dis)agreement with quality statement

4.3.9 How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion?
For ground truth we will use majority label. Aggregation strategies are not mentioned in the original paper. We will also keep all assessments for more qualitative and in-depth analysis of single instances.

4.3.10 Method(s) used for determining effect size and significance of findings for this quality criterion.
None

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

4.3.11.1 Has the inter-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used?
1. yes, Krippendorff's alpha

4.3.11.2 What was the inter-annotator agreement score?
N/A

4.3.12 Intra-annotator agreement
Questions 4.3.12.1 and 4.3.12.2 record information about intra-annotator agreement.

4.3.11.1 Has the intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used?
2. no

4.3.11.2 What was the intra-annotator agreement score?
N/A

B.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.

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?
Yes! The Research Ethics Committee (CETO) of the Faculty of Arts, University of Groningen.

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.
No

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: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited)? If yes, describe data and state how addressed.
No

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.
No