Investigating Crowdsourcing Protocols for Evaluating the Factual Consistency of Summaries

Anonymous submission

Abstract

Current pre-trained models applied for summarization are prone to factual inconsistencies which misrepresent the source text. Thus, evaluating the factual consistency of summaries is necessary to develop better models. However, the optimal human evaluation setup for factual consistency has not been standardized. To address this issue, we crowdsourced evaluations for factual consistency using the rating-based Likert Scale and ranking-based Best-Worst Scaling to determine the factors that affect the reliability of the human evaluation. Our crowdsourced evaluations are conducted on the summaries of CNN-Daily Mail and XSum datasets generated by four state-of-the-art models. Ranking-based Best-Worst Scaling offers a more reliable measure of summary quality across datasets, and the reliability of Likert ratings highly depends on the target dataset and the evaluation design. To improve the reliability, we extend the scale of the Likert rating to make it more flexible and we present a scoring algorithm for Best-Worst Scaling, called value learning. Our crowdsourcing guidelines and evaluation protocols will be publicly available to facilitate future research on factual consistency in summarization.

1 Introduction

Pre-trained language models have achieved promising progress in abstractive text summarization (Edunov et al., 2019; Dong et al., 2019; Song et al., 2019; Zhang et al., 2019, 2020). Despite their strong capability to generate coherent and fluent summaries, a serious limitation of these models is their tendency to produce text that is factually inconsistent with the input. Such inconsistencies render the summary unusable in many applications, including clinical or legal summarization, where factual accuracy is paramount. Thus, evaluating the factual consistency of the generated summaries with respect to the source is an important task (Falke et al., 2019; Cao et al., 2020; Gabriel et al., 2020; Durmus et al., 2020; Huang et al.; Pagnoni et al., 2021).

Recently, metrics have been proposed for evaluating factual consistency, including applying natural language inference (Falke et al., 2019; Mishra et al., 2020; Barrantes et al., 2020) and question answering models (Eyal et al., 2019; Scialom et al., 2019; Durmus et al., 2020; Wang et al., 2020). However, current metrics still have a low correlation with human judgments on factual consistency (Koto et al., 2020; Pagnoni et al., 2021).

To overcome the inherent limitation of automatic metrics, researchers usually adopt crowdsourced human evaluations using platforms such as Amazon's Mechanical Turk (MTurk) (Gillick and Liu, 2010; Sabou et al., 2012; Lloret et al., 2013). Despite the ubiquity of human evaluations, papers often differ in their preferred evaluation protocols (Louis and Nenkova, 2013; Hardy et al., 2019). Furthermore, differences in the evaluation task design affect the consistency and quality of the resulting crowdsourced human judgments (Santhanam and Shaikh, 2019), which ultimately affect system comparisons. Various methodologies have been proposed to measure inter- and intra-annotator consistency in human evaluation (Amidei et al., 2018). Best-Worst Scaling (Louviere and Woodworth, 1991) is a ranking-based method by which the annotator selects the best and worst example out of a set of examples. Prior research has claimed that Best-Worst Scaling produces higher-quality evaluations than widely-used rating scales such as the Likert Scale for tasks such as sentiment polarity analysis (Kiritchenko and Mohammad, 2017). In the context of summarization, Steen and Markert (2021) find that, compared to the Likert Scale, ranking-based protocols are more reliable for measuring the coherence of summaries but less so for measuring repetition. However, previous studies have not analyzed annotation reliability in the context of factual

Models	CNN/DM			XSum		
	R-1	R-2	R-L	R-1	R-2	R-L
PEGASUS	44.19[1]	21.45[1]	41.08[1]	46.84[1]	24.52[1]	39.10[1]
ProphetNet	42.45[3]	19.90[3]	39.31[3]	43.23[3]	19.96[3]	35.16[3]
BART	44.07[2]	21.13[2]	40.89[2]	44.15[2]	21.28[2]	35.94[2]
BERTSUM	41.82[4]	19.39[4]	38.67[4]	38.21[4]	16.11[4]	30.83[4]

Table 1: ROUGE-1/2/L scores for model reproduction on CNN/DM and XSum datasets. We apply models directly when they are already fine-tuned and otherwise re-trained them. Pegasus and BART generally obtain the highest ROUGE scores, with ProphetNet comparable in both cases and BERTSUM notably worse on XSum.

Models	CNN/DM		XSum	
	BWS	LS	BWS	LS
PEGASUS	3.230[2]	3.887[2]	3.247[3]	3.350[1]
ProphetNet	3.100[3]	3.860[4]	3.360[2]	3.293[3]
BART	3.593[1]	4.017[1]	3.570[1]	3.433[2]
BERTSUM	3.087[4]	3.863[3]	2.827[4]	2.790[4]

Table 2: Average model rank across BWS evaluations and average rating score (5-point) across LS evaluations. We average the ranks to obtain the final scores.

consistency for summarization.

Our contributions are the following: 1) We believe to be the first to study the reliability of human evaluation for factual consistency in summarization. 2) We study rating and ranking-based protocols across two summarization datasets with respect to four state-of-the-art abstractive models. We determine the factors that affect the reliability of the human evaluation, and present a novel ranking-based protocol with the highest reliability. 3) We will release our evaluation guidelines and annotations to promote future work on factual consistency evaluation.

2 Study Design

Each study consists of 100 input documents randomly sampled from each dataset, and associated four summaries generated using four models.

2.1 Datasets and Models

Datasets: The CNN/DailyMail dataset (Hermann et al., 2015; Nallapati et al., 2016) is a standard benchmark for summarization models (Fabbri et al., 2021) consisting of online articles and bullet-point summaries, typically including three sentences. XSum (Narayan et al., 2018) consists of 227K online articles with single-sentence summaries.

Models: Our study uses the following abstractive summarization models: **BART** (Lewis et al., 2020), a denoising autoencoder for pretraining sequence to sequence and natural language understanding tasks; **ProphetNet** (Qi et al., 2020), a pre-trained encoder-decoder model that performs n-gram language modeling; **PEGASUS** (Zhang et al., 2020), a model pre-trained with a summarization-specific objective function; and **BERTSUM** (Liu and Lapata, 2019), a two-stage fine-tuning approach. The models' ROUGE scores are shown in Table 1.

2.2 Reliability

For computing reliability, **Krippendorff's alpha** (α) is a reliability coefficient developed to measure the agreement among multiple annotators (Krippendorff, 2011). α measures instance-level reliability, especially how reliable judgments are over individual summary instances. For system-level rankings, to measure the reliability of the rankings of summarization models, we compute **Split-Half Reliability (SHR)**. SHR computes Pearson correlations and the annotations are split into two independent groups over which correlations are calculated.

We follow a similar block-design described in Steen and Markert (2021). We divided our corpus into 20 blocks of 5 documents and included all 4 generated summaries for each document in the same block, which results in $5 \times 4 = 20$ summaries per block. We require 3 annotators per block, and each annotator is limited to annotating at most two blocks total across all tasks. Crowdsourcing is done via Amazon Mechanical Turk (MTurk).

2.3 Protocols

The **Likert Scale (LS)** is a common rating-based evaluation protocol (Asghar et al., 2018). Likert Scales usually have 5 points (Steen and Markert, 2021). **Best-Worst Scaling (BWS)** is a type of ranking-oriented evaluation that requires annotators to specify only the best and the worst example in a set of summaries (Hollis and Westbury, 2018; Kiritchenko and Mohammad, 2017). For BWS, the annotator labels the most factually consistent summary and the least factually consistent summary.

Furthermore, we include the article as the context of the summaries as opposed to the coherence and repetition dimensions studied in Steen and Markert (2021), which do not require reading the input article. Including the article allows annotators to better differentiate summaries with similar quality, as the annotators may instinctively rely on additional contextual features to decide rankings.

Scale	CNN/DM		XSum	
	α (%)	SHR (%)	α (%)	SHR (%)
	Protocols			
LS	4.43	45.61	22.02	**92.77**
BWS	**15.82**	**87.65**	**24.77**	90.31
	Ours			
LS_{10}	12.87	51.36	29.51	**94.85**
BWS_{value}	**29.31**	**92.48**	**30.62**	92.98

Table 3: Instance and system-level reliability computed by Krippendorff's alpha (α) and split-half reliability (SHR) on the CNN/DM and XSum datasets.

	CNN/DM			XSum		
	BWS	LS	LS_{10}	BWS	LS	LS_{10}
Change Rate (%)	**74.71**	87.75	96.00	**70.25**	92.25	96.25
Percentage Scale Overlap	-	0.67	0.61	-	**0.88**	**0.82**

Table 4: Results of (a) inconsistencies in annotations by different (lower is better) and (c) egion bias of Likert Scales (higher is better).

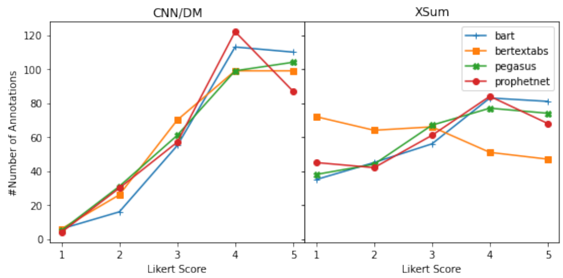


Figure 1: Score distribution of LS with 5-point scale for faithfulness. Each data point shows the number of times a particular score was assigned to each system.

2.4 Research Questions

We examine protocols in the context of factual consistency, while the problem of human evaluation presses for analysis. We organize our study along three main research questions (RQ):

RQ1: Ranking (BWS) vs. LS? We aim to determine the more reliable evaluation protocol.

RQ2: What will affect reliability? We aim to determine the factors that affect the reliability of the human evaluation.

RQ3: What are protocols' limitations and how to improve it? Based on the analysis, we propose two protocols to improve the reliability.

3 Analysis

The primary results for reliability across datasets and scales are found in Table 3. We show the average model ranking and rating across `BWS` and `LS` scales in Table 2. Despite the consistently higher ROUGE scores, Pegasus was not always ranked highest, which aligns with previous work suggesting that ROUGE score does not correlate with factual consistency (Durmus et al., 2020).

RQ1: BWS outperforms LS on CNN/DM. in the first two rows of Table 3, we first analyze the performance of `BWS` and `LS` on the CNN/DM dataset, as this dataset has been a benchmark for recent work in text summarization. `BWS` outperforms `LS` by a large margin on both instance-level (α) and system-level (SHR) reliability. As seen in the distribution of the `LS` ratings in Figures 1 and 2, many models are rated as factually consistent with scores of 4 or 5. This coincides with previous investigations on CNN/DM which conclude that recent summarization systems produce fluent texts with relatively few factual errors (Fabbri et al., 2021). However, as seen in Table 2, BART performs the best in factual consistency and ProphetNet performs relatively worse. Forcing the annotator to choose the best may help distinguish these close summaries rather than allowing e.g. the annotator to give both a score of 5. Thus, while many models score similarly, selecting only the best and worst summaries reduces some of the noise associated with rating similar summaries. We think that the `BWS` is preferable in cases where summaries have similar factual consistency, such as CNN/DM.

Though agreement on individual summaries (Krippendorff's alpha) is relatively low for all annotation methods, comparable to those obtained in (Steen and Markert, 2021), studies still arrive at consistent system scores when we average over many annotators as demonstrated by the SHR. This reflects similar observations made by Gillick and Liu (2010). System-level ranks such as SHR, are also more important for evaluation purposes as the goal is generally to rank models to determine the best performing (or most factually consistent) system as opposed to examining individual examples as Krippendorff's alpha measures.

RQ2: Dataset Characteristics Affect Reliability. We extend our experiments to the XSum dataset to see whether the reliability of the protocols changes as the characteristics of the dataset change. XSum models are known to suffer from factual inconsistencies because of the high compression ratio and high level of abstraction of the reference summaries (Maynez et al., 2020). As seen in Table 3, `BWS` and `LS` both perform well, with `LS` slightly outperforming `BWS`. As seen in Figures 1 and 2, the model scores are more spread out along the scale. This coincides with the large range of ROUGE scores and larger differences between models, as seen in Table 1, which likely explains why annotators can differentiate the model outputs better. Thus, the LS is a viable option when the corpus contains a diverse quality of summaries, like XSum.

RQ3: Limitations and Improvements. We first study the four limitations: (a) inconsistencies in annotations by different annotators, (b) inconsistencies in annotations by the same annotator; (c) scale region bias, while different annotators are often biased towards different parts of the rating scale; and (d) fixed granularity, while too narrow of a rating scale range may fail to capture nuanced differences in the text quality, too wide of a rating scale range may overwhelm the annotator. Then we propose two new protocols to improve the reliability. For LS, we extend the scale from 5 to 10, we call it LS-10. Because a finer-grained scale captures more nuanced differences in data points with more choices. And previous work suggests that Best-Worst Scaling fails to yield an unbiased estimate of the true quality value (Hollis, 2018). For BWS, we improve the reliability by incorporating information about the quality of competition, called BWS_{value}. The annotator is asked to give a score (3-point scale) for the difference between the best and the worst summary. The final overall ranking uses a weighted sum. The results in Table 3 prove the effectiveness of our proposed protocols.

To verify aforementioned problems, we conduct the following studies. We first analyze **(a) the inconsistencies in annotations by different annotators**, measured by the percentage of summaries that receive different ratings or rankings from different annotators, which we call **change rate**. As shown in Table 4, annotators are more likely to agree on the same ranking in BWS as opposed to the same rating for LS. We further test **(b) inconsistencies by the same annotator**, especially whether annotations done by the same worker are consistent over time. We ask workers who have previously annotated XSum samples to re-do their annotations one week after their initial annotations. We notified the workers to re-annotate only one week after they finished, instead of at the beginning as we do not want to introduce design bias. In total, 43 workers redid 860 annotations. For LS-10, the change in the rating of the two annotations one week apart by the same worker was 0.92. For BWS, 39% of the time the annotator changed the ranking.

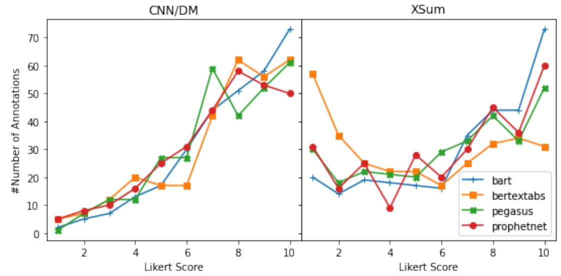


Figure 2: Score distribution of LS_{10} for faithfulness. Each data point shows the number of times a particular score was assigned to each system.

Additionally, we examine whether LS suffers from **(c) region bias**. For a given block and two annotators, we calculate the rating range given by each annotator. We then calculate the overlap length between those two ranges divided by the length of the overall range from both annotators. We call this **the percentage scale overlap** and average over all pairs of annotators and blocks. For LS, the percentage scale overlap is (0.67, **0.88**) for (CNN/DM, XSum), respectively, and (0.61, **0.82**) for LS-10. Thus, greater diversity in summary quality as in XSum may force the annotators to expand their use of the scale and mitigate region bias, which explains why Likert is better than BWS on XSum as opposed to CNN/DM. Finally, we analyze **(d) the effect of scale granularity**. From Table 3, we find that LS-10 is more reliable than LS. Scores tend to move towards the extremes when we use a finer-grained scale (10 vs 5), as seen in the difference in distributions in Figures 1 and 2. Thus, for LS-10, larger range and being less biased towards a specific region promote better reliability. Future work may investigate further what exactly constitutes too wide of a scaling range.

4 Conclusion

In this paper, we conduct studies to understand and improve the reliability of ranking and rating-based human evaluations of factual consistency in summarization models. We find that Best-Worst Scaling is largely reliable in two datasets and the Likert scale also has merits, but the proper scaling and dataset characteristics must be carefully studied to ensure its reliability. We improve these two protocols based on our findings across studies. We believe that our quantitative studies advance the understanding of both models and metrics as we aim to facilitate factually consistent text generation.

5 Ethical Considerations

Intellectual Properties and Privacy Rights All of the datasets (CNN/DM and XSum) used in our study are publicly available. Regarding privacy rights, the authors of the paper completed IRB human subject protection training for conducting this study. We will release the annotations, but rather than releasing the MTurk ID of the worker, we will completely anonymize this ID.

Compensation for Annotators Workers were compensated \$5 per block, calibrated to equal a \$15/hour payrate. We first annotated examples in-house to determine the required annotation speed. Typically, a summary block takes around 20 minutes.

Steps Taken to Avoid Potential Problems Annotations were completed in the form of a survey on a Google Form. We provided space for the Turkers to provide feedback. We manually uploaded the data points (articles and summaries) used in this study to avoid any offensive content.

The Number of Examples We sampled 100 examples from each dataset that did not contain exactly matching summaries. Both Likert and BWS follow the same block design, which includes the same number of examples per block. With the exception that the BWS annotation asks for the most and least factually consistent summary and the Likert asks for ratings for each individual summary. Due to space requirements, we included further details, images of the interface, in the supplementary material. We pay the same amount per block of annotations.

Qualifications of MTurk workers We use the following qualifications to recruit in total 350 MTurk workers with good track records: HIT approval rate greater than or equal to 98%, number of HITs approved greater than or equal to 500, and located in one of the following English native-speaking countries: Australia, Canada, New Zealand, United Kingdom, United States.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, volume 10772 of *Lecture Notes in Computer Science*, pages 154–166. Springer.

Mario Barrantes, Benedikt Herudek, and Richard Wang. 2020. Adversarial nli for factual correctness in text summarisation models. *arXiv preprint arXiv:2005.11739*.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4052–4059, Minneapolis, Minnesota. Association for Computational Linguistics.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019.

Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2020. Go figure! a meta evaluation of factuality in summarization. *arXiv preprint arXiv:2010.12834*.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Geoff Hollis. 2018. Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior research methods*, 50(2):711–729.

Geoff Hollis and Chris Westbury. 2018. When is best-worst best? a comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior research methods*, 50(1):115–133.

Yi-Chong Huang, Xia-Chong Feng, Xiao-Cheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. Ffci: A framework for interpretable automatic evaluation of summarization. *arXiv preprint arXiv:2011.13662*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2013. Analyzing the capabilities of crowdsourcing services for text summarization. *Language resources and evaluation*, 47(2):337–369.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Li, Pavan Kapanipathi, and Kartik Talamadupula. 2020. Looking beyond sentence-level natural language inference for downstream tasks. *arXiv preprint arXiv:2009.09099*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, pages 1–8.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Julius Steen and Katja Markert. 2021. How to evaluate a summarizer: Study design and statistical analysis for manual linguistic quality evaluation. *arXiv preprint arXiv:2101.11298*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

A Appendix

Besides the average model rank and average rating scores across BWS, LS-5, and LS-10 evaluations, we also provide standard deviations in Table 5.

To demonstrate our annotation template and facilitate future research, we show the interface for BWS annotations in Figures 3 and 4 and the interface for Likert annotations in Figures 5 and 6. We made use of the survey feature in Amazon Mechanical Turk (MTurk) to link to these Google Forms 7.

Models	CNN/DM BWS	CNN/DM LS	CNN/DM LS-10	XSum BWS	XSum LS	XSum LS-10
PEGASUS	3.230^{2}/1.150	3.887^{2}/1.051	7.410^{3}/2.160	3.247^{3}/0.936	3.350^{1}/1.334	6.247^{2}/2.978
ProphetNet	3.100^{3}/1.026	3.860^{4}/0.992	7.250^{4}/2.252	3.360^{2}/1.102	3.293^{3}/1.359	6.427^{2}/3.038
BART	3.593^{1}/1.113	4.017^{1}/0.973	7.727^{1}/2.090	3.570^{1}/1.179	3.433^{2}/1.338	6.937^{1}/2.889
BERTSUM	3.087^{4}/0.984	3.863^{3}/1.037	7.453^{2}/2.309	2.827^{4}/0.993	2.790^{4}/1.390	5.163^{4}/3.202

Table 5: Average model rank, rating, and standard deviation across BWS, LS and LS-10 evaluations.

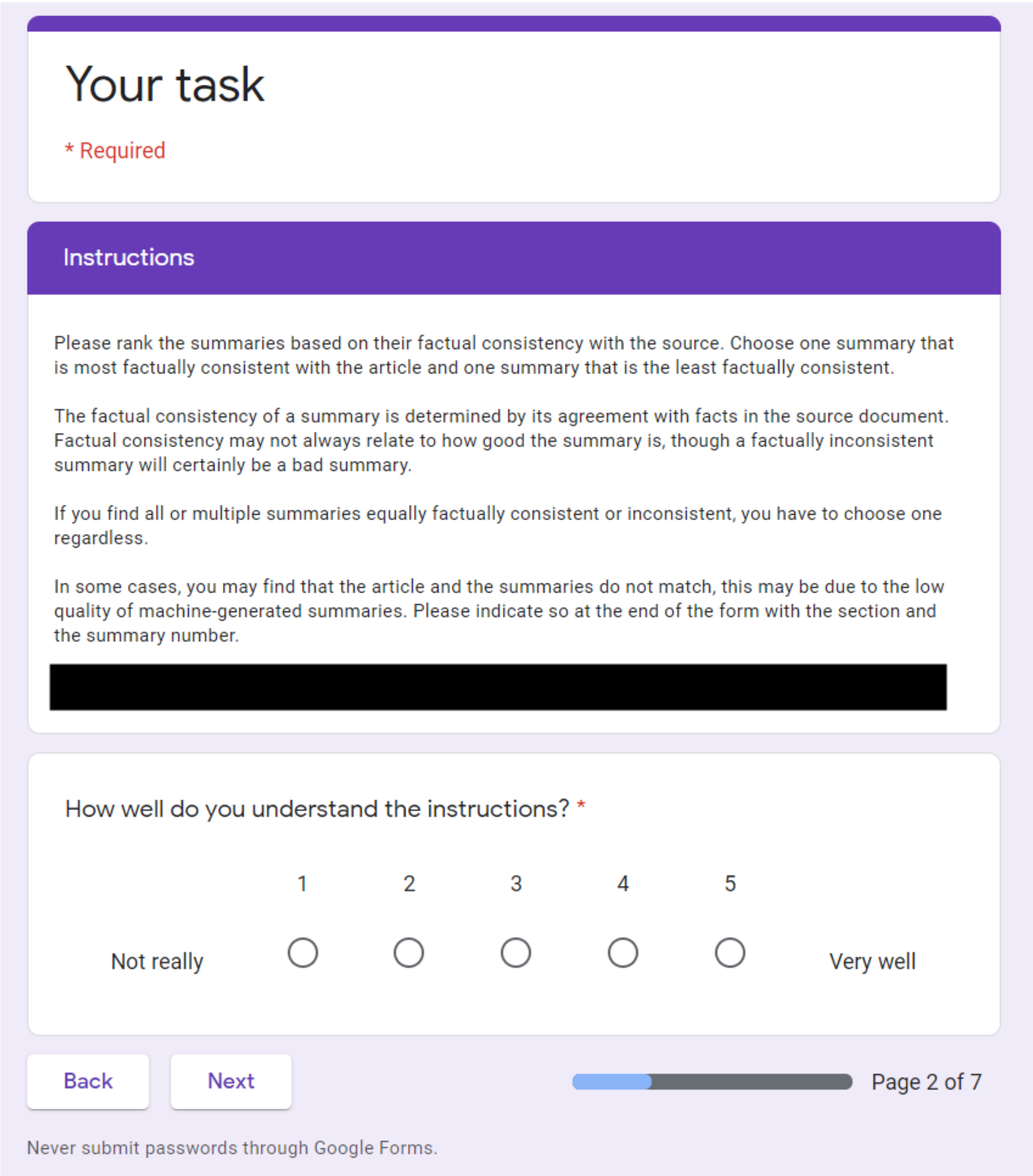


Figure 3: Screenshot of the instruction page for BWS annotation.

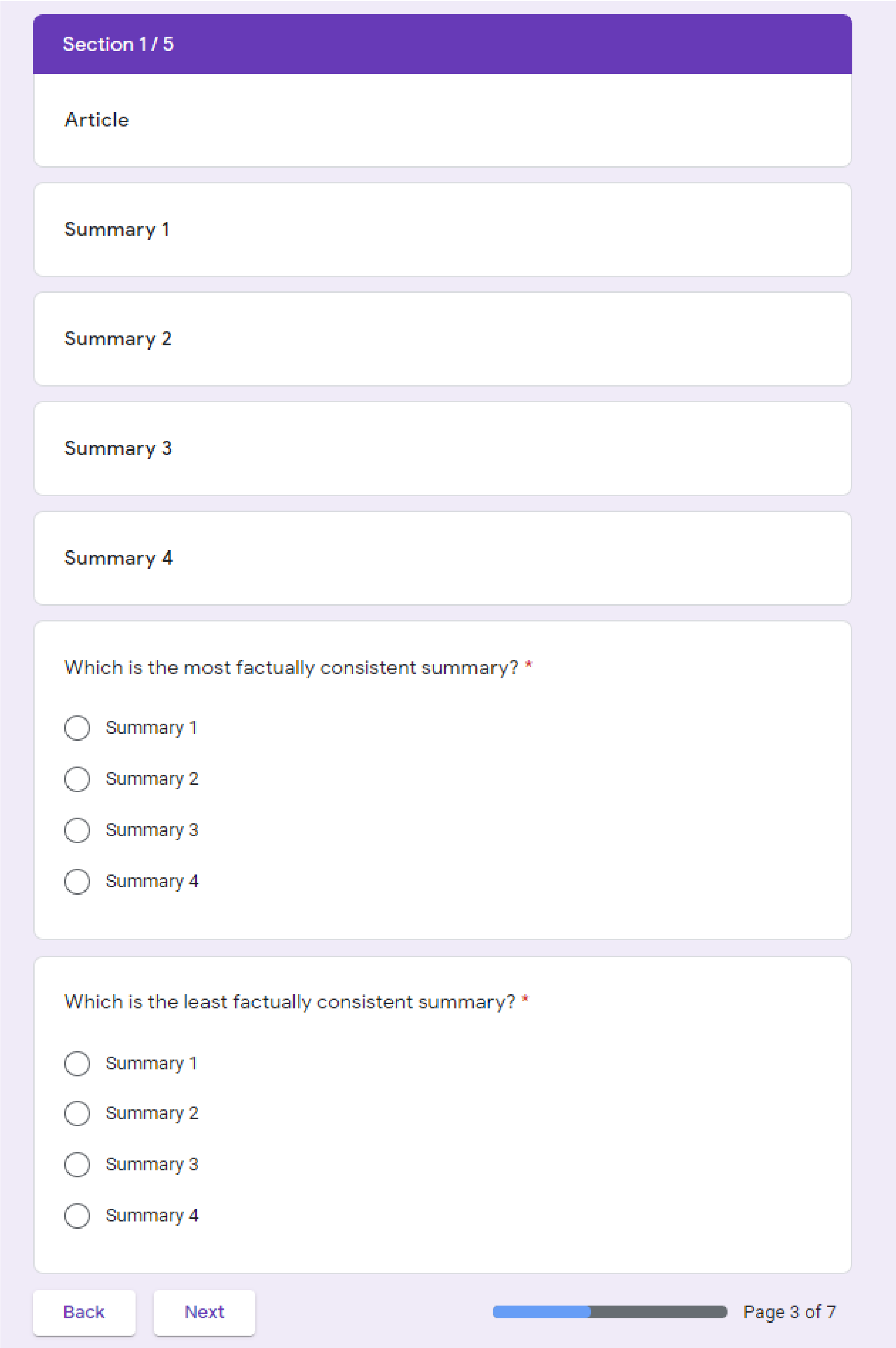
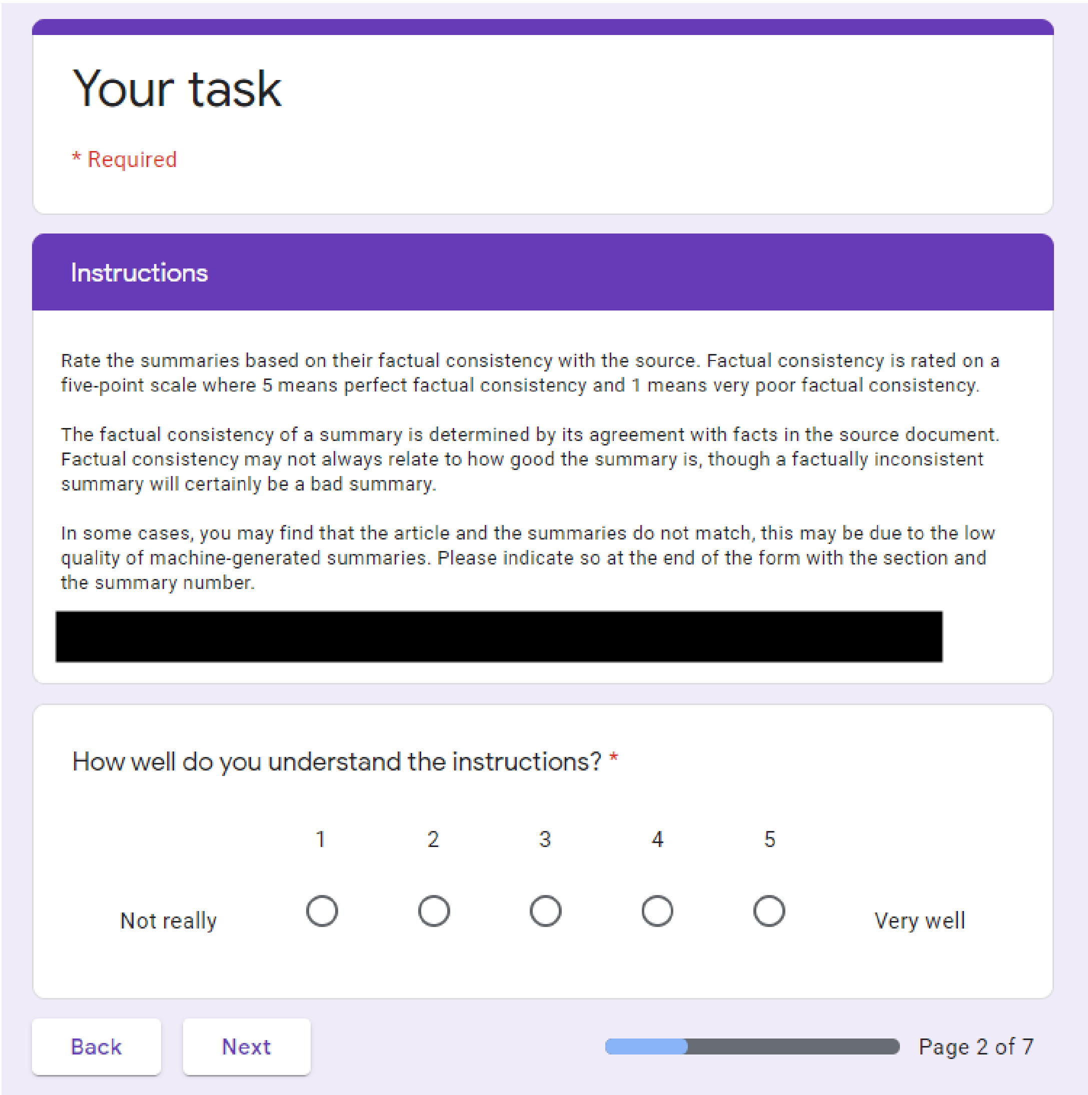


Figure 4: Screenshot of the evaluation page for BWS annotation.


Your task

* Required

Instructions

Rate the summaries based on their factual consistency with the source. Factual consistency is rated on a five-point scale where 5 means perfect factual consistency and 1 means very poor factual consistency.

The factual consistency of a summary is determined by its agreement with facts in the source document. Factual consistency may not always relate to how good the summary is, though a factually inconsistent summary will certainly be a bad summary.

In some cases, you may find that the article and the summaries do not match, this may be due to the low quality of machine-generated summaries. Please indicate so at the end of the form with the section and the summary number.

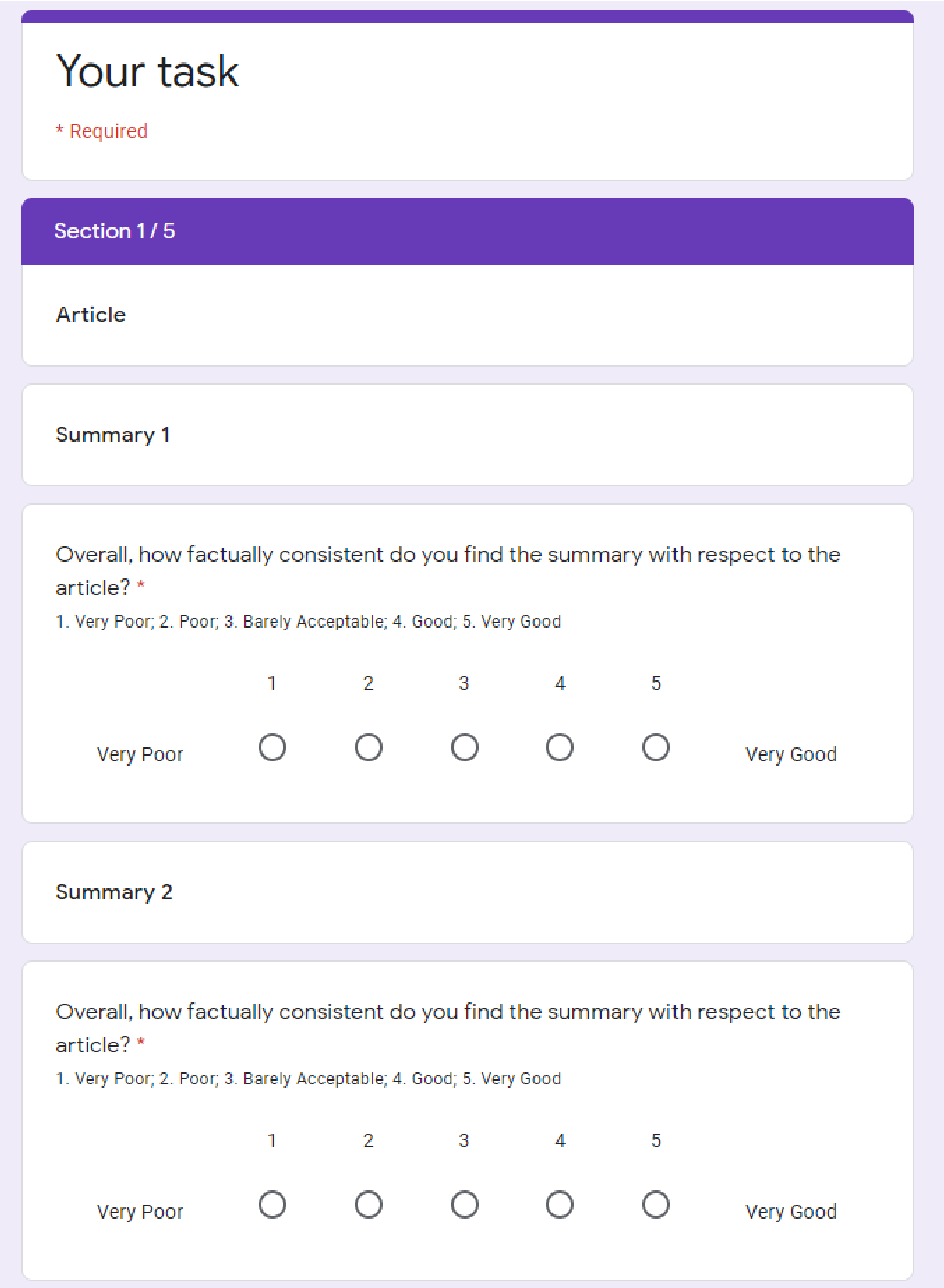
[redacted]

How well do you understand the instructions? *

Not really 1 2 3 4 5 Very well

Back Next Page 2 of 7

Figure 5: Screenshot of the instruction page we used for Likert Scale annotation.



Your task

* Required

Section 1 / 5

Article

Summary 1

Overall, how factually consistent do you find the summary with respect to the article? *

1. Very Poor; 2. Poor; 3. Barely Acceptable; 4. Good; 5. Very Good

	1	2	3	4	5	
Very Poor	○	○	○	○	○	Very Good

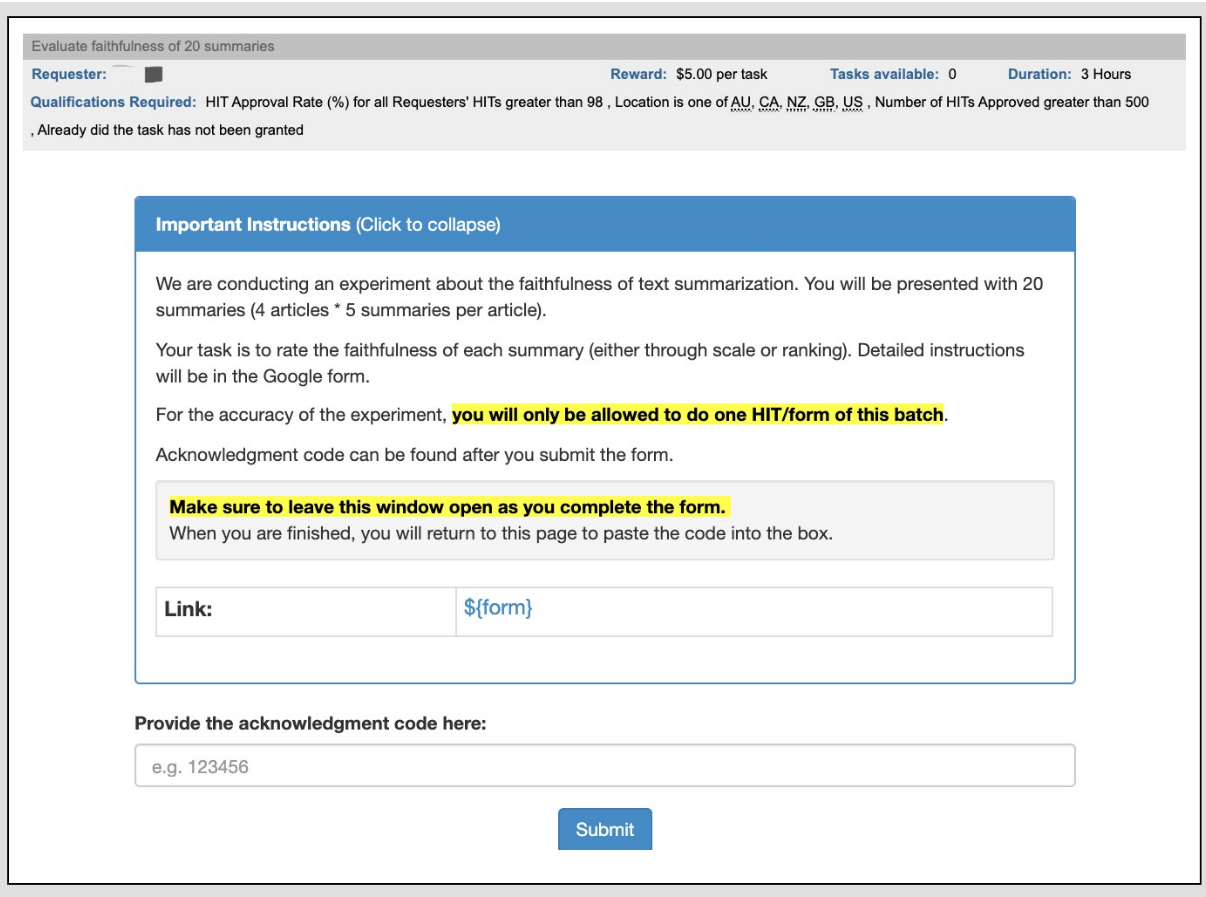
Summary 2

Overall, how factually consistent do you find the summary with respect to the article? *

1. Very Poor; 2. Poor; 3. Barely Acceptable; 4. Good; 5. Very Good

	1	2	3	4	5	
Very Poor	○	○	○	○	○	Very Good

Figure 6: Screenshot of the evaluation page for Likert Scale annotation.



Evaluate faithfulness of 20 summaries

Requester: [redacted] **Reward:** $5.00 per task **Tasks available:** 0 **Duration:** 3 Hours

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than 98 , Location is one of AU, CA, NZ, GB, US , Number of HITs Approved greater than 500 , Already did the task has not been granted

Important Instructions (Click to collapse)

We are conducting an experiment about the faithfulness of text summarization. You will be presented with 20 summaries (4 articles * 5 summaries per article).

Your task is to rate the faithfulness of each summary (either through scale or ranking). Detailed instructions will be in the Google form.

For the accuracy of the experiment, **you will only be allowed to do one HIT/form of this batch**.

Acknowledgment code can be found after you submit the form.

Make sure to leave this window open as you complete the form.
When you are finished, you will return to this page to paste the code into the box.

Link:	${form}

Provide the acknowledgment code here:

e.g. 123456

Submit

Figure 7: This is how our task will look to Mechanical Turk Workers.