A Needle in a Haystack:
An Analysis of High-Agreement Workers on MTurk for Summarization

Lining Zhang,[1]* Simon Mille,[2] Yufang Hou,[3] Daniel Deutsch,[4] Elizabeth Clark,[5] Yixin Liu,[6] Saad Mahamood,[7] Sebastian Gehrmann,[5] Miruna Clinciu,[8] Khyathi Chandu,[9] João Sedoc[1]

[1]New York University, [2]ADAPT Centre, DCU, [3]IBM Research, [4]Google, [5]Google Research, [6]Yale University, [7]trivago N.V.,
[8]University of Edinburgh, [9]Allen Institute for AI

Abstract

To prevent the costly and inefficient use of resources on low-quality annotations, we want a method for creating a pool of dependable annotators who can effectively complete difficult tasks, such as evaluating automatic summarization. Thus, we investigate the recruitment of high-quality Amazon Mechanical Turk workers via a two-step pipeline. We show that we can successfully filter out subpar workers before they carry out the evaluations and obtain high-agreement annotations with similar constraints on resources. Although our workers demonstrate a strong consensus among themselves and CloudResearch workers, their alignment with expert judgments on a subset of the data is not as expected and needs further training in correctness. This paper still serves as a best practice for the recruitment of qualified annotators in other challenging annotation tasks.

1 Introduction

Natural language generation (NLG) tasks like text summarization are challenging to evaluate both in terms of automatic metrics and human evaluations (Gehrmann et al., 2022). Although automatic metrics are inexpensive proxies for human annotations for tasks like dialog evaluation (Mehri et al., 2022), they may have problems dealing with paraphrases, capturing distant dependencies, or identifying nuances in human languages (Banerjee and Lavie, 2005; Isozaki et al., 2010; Manning et al., 2020). Thus, it is still crucial to obtain high-quality human annotations as gold labels for evaluation. Amazon Mechanical Turk (MTurk)[1] is a commonly used crowdsourcing platform for collecting human annotations on designed tasks, known as Human Intelligence Tasks (HITs). However, finding qualified workers for high-quality annotations with a better inter-annotator agreement (IAA) is challenging, especially for difficult tasks such as text summarization. Best practices for recruiting high-quality workers are also poorly understood, and the relationship between high quality and high agreement needs further investigation.

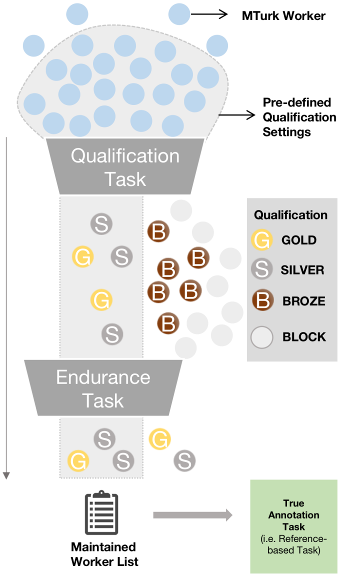


Figure 1: Two-step pipeline for finding high-agreement MTurk workers: participants who satisfy basic qualification settings and answer designed questions correctly (Qualification) are subsequently filtered in a longer task (Endurance). The maintained worker list is tested for the true annotation task later (Reference-based).

To tackle the above issues, we design a recruitment pipeline to identify workers who are able to produce high-agreement annotations for the evaluation of text summarization on MTurk. It comprises a qualification task and an endurance task, followed by a reference-based task (see Figure 1). In the qualification task, workers who meet predefined qualification settings receive instructions and qualification questions, including an attention check (Oppenheimer et al., 2009). The qualification questions are designed to assess the annotator's ability to evaluate multiple dimensions of a

* Correspondence to lz2332@nyu.edu

[1] https://www.mturk.com/

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 14944–14982
July 9-14, 2023 ©2023 Association for Computational Linguistics

summary correctly. Performance on this task determines whether they are categorized into GOLD, SILVER, BRONZE, or BLOCK. Only the best workers (GOLD and SILVER) move on to the endurance task, which consists of 10 HITs with 4 summaries in each to evaluate. This task only tests the summary's saliency, which is the most subjective dimension (Howcroft et al., 2020), but it challenges the annotator's capacity for handling a heavy annotation workload. GOLD and SILVER workers who complete all HITs are added to a maintained worker list as high-agreement annotators for future tasks. To ensure their general performance for the true annotation task, a reference-based task to evaluate information coverage between summaries is conducted with these workers later.

While serving as a best practice beyond its scope, our study has the following contributions:

- establish a cost-effective recruitment pipeline on MTurk to consistently build a pool of annotators for high-agreement annotations.
- successfully recruit 12 out of 200 (6%) superior annotators for text summarization evaluation, while reducing costs and guaranteeing high agreement.
- rigorously demonstrate that the annotators identified through our pipeline can match or surpass the IAA of expert annotators and standard statistical techniques, though further calibration may be required for correctness.

2 Related Work

Challenges of Human Evaluation Compared to automatic evaluation metrics for NLG tasks like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), human annotations from non-expert annotators on MTurk can reach an agreement with gold standards or expert judgments (Callison-Burch, 2009). Although recent works leverage language models like BERT (Devlin et al., 2019) to get better automatic evaluations (Zhang et al., 2020), human judgments are still indispensable in identifying nuances in specific language tasks (Manning et al., 2020). Finding qualified workers to carry out the evaluations is crucial. This is especially true for tasks like text summarization, which lacks consensus on evaluation protocols (Fabbri et al., 2021) and is often inconsistent with previous human evaluations (Hardy et al., 2019). However, human evaluation from non-expert crowdsourcing platforms have low quality (Gillick and Liu, 2010) and a simple qualification filter is not sufficient to identify qualified workers (Berinsky et al., 2012; Robinson et al., 2019). Some studies applied quality control mechanisms to filter out poor quality annotations, resulting in a relatively low pass rate for a variety of tasks (Graham et al., 2017, 2018; Mille et al., 2019). The fact that up to 70% of the HITs are eventually discarded indicates a huge resource waste.

Even with qualified workers, human annotations might still be adversely affected by factors like incomplete instructions or unfair wages paid to annotators (Huynh et al., 2021), and workers need clear references, schemes, or standards to follow (Howcroft et al., 2020; Karpinska et al., 2021). Thus, our study serves as a detailed reference for finding qualified MTurk workers for a summarization evaluation task and further identifying those who can assist in a large number of annotations.

Inter-Annotator Agreement For annotations without true labels or those evaluated with a qualitative scale such as Likert scale (Likert, 1932), the inter-annotator agreement (IAA) among MTurk workers measures the reliability of the annotations. For example, Cohen's Kappa (Cohen, 1960) measures IAA between a pair of results of the same length from two annotators, while Krippendorff's Alpha (Hayes and Krippendorff, 2007) measures the agreement of a set of results from any number of annotators, even with unequal sample sizes. Both range from -1 to 1, with 1 indicating complete agreement. Further studies also continue to mitigate annotator bias through complementary methods to IAA (Amidei et al., 2020), aimed at high-quality annotations. In our study, we utilize both Cohen's Kappa and Krippendorff's Alpha as the measurement of annotation reliability.

3 Methods

In this section, we detail how the workers were recruited and which tasks were carried out.[2]

3.1 MTurk Qualification Settings

To narrow down the pool of our target workers, we set a few pre-defined qualifications for workers on MTurk before publishing the qualification task: (i) the **Location** is set to "UNITED STATES (US)"; (ii) the **Number of HITs Approved** is set to be "greater than 1000" to target workers who are already experienced on MTurk; (iii) the **HIT Approval Rate (%)** is set to be "greater than or

[2] Appendix A.9 shows instructions given during the tasks.

equal to 99" to target workers who are able to finish tasks with high quality and have stable performance. We also set the task visibility as "Private", which means our tasks are visible to any worker, but only workers who meet all qualification requirements can preview and accept.

Paolacci et al. (2010) show that the annotations collected with the "Location" setting on MTurk are representative of the population of our target country in terms of demographic data. This helps mitigate biases introduced by samples from traditional recruitment methods like college undergraduate samples (Buhrmester et al., 2011). We set qualification settings (ii) and (iii) based on previous work (Whiting et al., 2019; Oppenlaender et al., 2020; Kummerfeld, 2021) and our own experience on MTurk. Workers who meet all qualification requirements are eligible to participate in the qualification task.

3.2 Qualification Task

Summarization task In summarization, the input is the text of a document and the output is a short summary. We evaluate a summary S according to 6 dimensions based on the criteria taxonomy presented in Howcroft et al. (2020), and workers are asked for a binary answer as to whether a dimension is satisfied in a summary or not:

- **Understandability**: can the worker understand S and is S worth being annotated.
- **Compactness**: S does not contain duplicated information.
- **Grammaticality**: S is free from grammatical & spelling errors.
- **Coherence**: S is presented in a clear, well-structured, logical, and meaningful way.
- **Faithfulness**: all of the information in S can be found in the article; S accurately reflects the contents of the article.
- **Saliency**: S captures the most important information of the article and does not include parts of the article that are less important.

Training and qualification There are two main parts of the qualification task. The ***training part*** guides the workers through the above evaluation dimensions and instructs them on how to annotate. The definition of each dimension is illustrated with positive and negative examples, and full annotation examples are shown (summary and binary rating for each dimension). Then, workers are required to write an instruction summary in their own words to make sure they have understood the task and are ready to annotate. The ***qualification part*** tests the worker's understanding of the task. Three documents are provided, each with one summary. The worker reads the document and annotates the corresponding summary according to each dimension. The ratings are then compared to expert ratings provided by the authors of this paper. The last document comes with an attention check to test whether a worker is just randomly assigning scores without reading: a highlighted instruction asks the worker to ignore the task and select specific answers. Finally, an optional field is provided to collect feedback.

Worker categorization Upon finishing their task, workers are categorized into four types:

- **GOLD**. The GOLD workers pass the attention check and annotate every dimension of every document in the qualification part correctly.
- **SILVER**. The SILVER workers pass the attention check and make only one mistake when annotating each dimension of the documents in the qualification part.
- **BRONZE**. The BRONZE workers pass the attention check and make more than one mistake when annotating each dimension of the documents in the qualification part.
- **BLOCK**. The BLOCK workers fail to pass the attention check.

The GOLD and SILVER workers are assigned a qualification score and proceed with the endurance task. Besides, we conducted multiple rounds of the qualification task to avoid influence from the time or day when the task was conducted and randomly sampled workers (Arechar et al., 2017; Berinsky et al., 2012).

3.3 Endurance Task

The endurance task is designed to test whether a worker can reliably perform a large number of annotations. The workers who finish all HITs of this task are assigned the highest qualification score and are added to a maintained worker list.

The endurance task comprises 10 HITs. For each HIT, a document and 4 corresponding summaries generated by different models are provided; each HIT takes around 5 minutes to finish (approximately an hour for all HITs). To keep the task simple we only evaluate each summary on one dimension, but to ensure that the task is challenging enough we (i) use the most subjective of the 6 di-

Round Number	1	2	3	4	Total
Total participants at the beginning	50	50	50	50	200
# GOLD workers passed qualification task	1	3	2	2	8
# SILVER workers passed qualification task	4	5	3	6	18
# workers entered endurance task	5	8	5	8	26
# GOLD workers passed endurance task	1	1	1	1	4
# SILVER workers passed endurance task	0	3	2	3	8
# workers passed both tasks	1	4	3	4	12

Table 1: Number of MTurk workers qualified after each task.

mensions, Saliency, and (ii) use a more fine-grained 10-point Likert scale (from 1 to 10).

Rationale for choosing 10 HITs Our motivation is two-fold: to find workers who were able to complete many tasks and whose annotations are better than random. As the number of HITs increases, the number of remaining workers drops from 26 to 12. The survival rate defined by the Kaplan–Meier estimator (Kaplan and Meier, 1958) is 38.59% when the number of HITs is set to 10 which is an estimate of a worker's capacity to be able to complete many tasks. We empirically found that we need a minimum of 8 HITs completed by a worker in order to validate that their annotations are statistically significantly different from random noise (see Table 2).

Num. of HITs finished	Num. of workers remaining	Survival rate % (Kaplan–Meier estimator)	Confidence interval of Cohen's Kappa	
			Lower bound	Upper bound
-	**26**[1]	100	-	
1	19	63.16	-	
2	18	59.65	-	
3	17	56.14	-	
4	16	52.63	-	
5	15	49.12	-0.18	0.44
6	15	49.12	-0.18	0.44
7	15	49.12	-0.18	0.44
8	14	45.61	0.06	0.44
9	13	42.10	0.08	0.42
10	12	**38.59**	**0.09**	**0.42**

[1] This (26) is the number of workers who entered the endurance task (GOLD and SILVER workers passed the qualification task).

Table 2: Statistical results as number of HITs grows.

3.4 Reference-based Task

Finally, to test whether the selected MTurk workers actually perform better at annotating summaries in general, we conduct a reference-based task that comprises 30 HITs. In each HIT, a reference summary and 4 candidate summaries are provided. The worker is asked to assign each candidate summary two scores ("can2ref" score and "ref2can" score) on a scale from 1 to 5. The "can2ref" score indicates whether all of the information in the candidate summary can also be found in the reference summary, while the "ref2can" score checks the converse coverage direction. A score of 1 means that almost no information in one summary can be found in the other, while a score of 5 indicates complete information coverage. The worker is provided with instructions and examples of the rating at the beginning of the task.

4 Results

4.1 Annotation Data and Cost

The collected experimental data not only contained annotation results but also metadata reflecting annotator behaviors.[3] The cost of annotation on MTurk included both the wages paid to MTurk Workers and the fees paid to MTurk (which may vary according to the task). A worker who participated in the qualification and the endurance tasks earned $8.5 ($1 for the qualification task plus $7.5 for the endurance task) on average, while a worker who participated only in the qualification task (i.e. who did not qualify) earned $1 on average. Given the total cost of $514 for the entire pipeline which yielded 12 workers, the cost of identifying a qualified worker is $42.8. For details, the breakdown of the cost is shown in Table 3.

4.2 Qualification Task Results

We conducted four rounds of the qualification task, each round included 50 MTurk workers (see Table 1). This choice of multiple rounds aimed to guarantee the stability of the annotation results (Berinsky et al., 2012; Arechar et al., 2017). The overall pass rate of the attention check was 0.69; thus, 62 workers in total did not pass the attention check and

[3] The data and code used for the analysis of all tasks are available at `https://github.com/GEM-benchmark/MTurkRequirementPipeline`.

Annotation Task		Reward per Assignment	Num. of Assignment per Task	Total Reward	Fees to MTurk	Total Cost	Hourly Wage
Qualification Task (Each of 4 rounds)		$1.00	50	$50	$20	$70	$2
Endurance Task	Round 1	$0.75	5	$37.5	$7.5	$45	$7.5
	Round 2	$0.75	8	$60	$12	$72	$7.5
	Round 3	$0.75	5	$37.5	$7.5	$45	$7.5
	Round 4	$0.75	8	$60	$12	$72	$7.5

Table 3: Wage Paid to MTurk Workers and total amount spent on annotation. The number of assignment per task indicates the number of workers who entered the task, which is not equal to the number of workers who passed the task. The hourly wage is calculated for one MTurk worker given a task.

were categorized as BLOCK. Out of 200 MTurk workers, there were only 8 GOLD workers and 18 SILVER after the qualification task. Thus, only 26 MTurk workers (13% of all participants) qualified for the endurance task.

For each round, we calculated Krippendorff's Alpha[4] to measure the agreement among annotators. The highest Krippendorff's Alpha was 0.33 reached by the first round, and the average Krippendorff's Alpha of all four rounds was 0.25. In addition, the exclusion of BLOCK workers led to an increase in Krippendorff's Alpha, compared to the value calculated on all workers. The highest Krippendorff's Alpha without BLOCK workers was 0.44 (second round), and the average Krippendorff's Alpha of all four rounds increased to 0.41. These results showed that, as expected, BLOCK workers seemed to lack good-faith effort in the task and likely yielded low quality annotations.

4.3 Endurance Task Results

We published the same endurance task for GOLD and SILVER workers separately, and reported IAA using Cohen's Kappa and Krippendorff's Alpha among each type of worker; we also reported similar IAA results from combined GOLD and SILVER workers. We additionally collected endurance task results from volunteer researchers unrelated to this paper for a comparison between MTurk workers and NLG "experts".

SILVER Workers There were 18 SILVER workers after the qualification task, 13 of whom accepted the endurance task. However, only 8 SILVER workers finished all 10 HITs–a yield rate of around 44% given the number of SILVER workers entering this task. To calculate the IAA, we considered the annotation scores of all summaries (40 ratings) for each of the 8 workers and calculated Cohen's Kappa for each worker pair; the highest Cohen's Kappa was 0.451 between workers S_{22} and S_{43}. To avoid influence from a possible unstable performance at the beginning of the task, we also tried to omit the first two HITs, that is, we only used 32 ratings when calculating Cohen's Kappa; the resulting improvement for Cohen's Kappa was very low. In addition, we calculated Krippendorff's Alpha on the entire annotation results for all summaries and workers, and it reached 0.358.

GOLD Workers There were 8 GOLD workers after the qualification task and 6 of them accepted the endurance task. However, only 4 GOLD workers finished all 10 HITs, for a yield rate of around 67% given the number of GOLD workers entering this task. This rate was higher than that of SILVER workers. We calculated pairwise Cohen's Kappa using all the scores, and the highest IAA score increased to 0.48, compared to 0.45 for SILVER workers. There was no significant improvement after omitting the first two HITs. Krippendorff's Alpha for the GOLD workers reached 0.443, which is higher than with SILVER workers (0.358).

GOLD and SILVER Workers To investigate IAA of worker pairs across GOLD and SILVER workers, we combined the results of these two categories of workers and calculated pairwise Cohen's Kappa. The highest pairwise Cohen's Kappa on the 40 ratings per worker was 0.55; see the matrix in Figure 2. Again, omitting the first two HITs also did not change the scores much. For Krippendorff's Alpha, the value was 0.396, which fell in the range between the SILVER worker's (0.358) and GOLD worker's (0.443) values.[5]

In Appendix A.2, we show a breakdown of the results per text position in each HIT (correlations for all first texts, for all second texts, etc.) for each of the three subgroups (SILVER, GOLD, GOLD AND SILVER); the possibly sightly darker heat maps

[4] https://pypi.org/project/krippendorff/

[5] Note that the relatively low Krippendorff's Alpha scores may in part be due to the large size of the scale (10 points).

could indicate higher correlations for the second text of each HIT.

Comparison to Expert Ratings To get an idea of the quality of qualified MTurk workers according to our approach, we compared their IAA with the IAA obtained by conducting the same endurance task with three researchers as NLG "experts". The pairwise Cohen's Kappa for all 40 ratings only reached 0.268 (see Table 10 in Appendix A.3). The IAA among the experts was comparatively lower than the GOLD and SILVER workers, indicating that qualified workers identified by our tasks reached a better agreement at least for the endurance task. Thus, it seems possible to recruit high-quality workers using our pipeline.

Detection of Abnormal Workers From Cohen's Kappa scores shown in Figure 2, the worker S_{42}[6] had much lower agreement scores (heatmap in the yellow colors on the row and column corresponding to the worker). Recent studies have uncovered the presence of bots on MTurk (Webb and Tangney, 2022). To understand the reason for this worker's lower agreement with other workers, we analyzed their online behavior using the metadata extracted from their annotation results.

Figure 3 shows the timeline of each of the 10 HITs as a horizontal gray line. The timelines are plotted from top to bottom, corresponding to the first to the last HIT in the endurance task. The X-axis represents the duration between the time of acceptance and submission, which is normalized by the duration for each HIT (ranging from 0 to 1). Different marks present each annotator behavior, as shown in the legend. Among these behaviors, blue points represent the time when the MTurk worker assigned a score for one of the four summaries, and the corresponding number on top represents the summary index (valued from 0 to 3). Orange crosses denote the suggested reading time of the article in each HIT, given the average human reading speed of 130 words per minute.[7] If the suggested reading time after normalization was longer than the duration, we marked the orange cross as 1 at the time of submission which is at the end of the gray line.

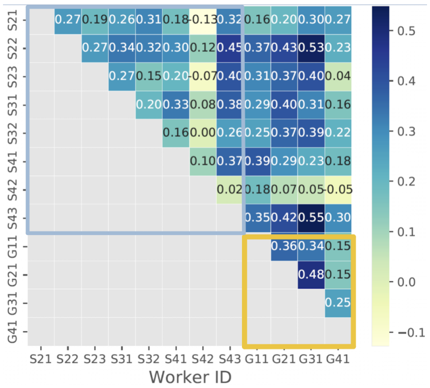


Figure 2: Cohen's Kappa for endurance task (grey frame: SILVER workers; yellow: GOLD workers).

Most of the orange crosses were marked at the end of the timelines in Figure 3 (right), indicating this worker assigned scores and submitted the HIT in less time than it usually takes for a human to even finish reading the article. This result demonstrates that this worker may not have put in good faith in the endurance task, which possibly explains the low IAA with other workers. By removing this worker and calculating Krippendorff's Alpha again within GOLD and SILVER workers, the IAA increased to 0.454 (compared to 0.396 when including the worker).

4.4 Reference-based Task Results

To test the reliability of our qualified workers and compare them to workers who do not undergo our selection process, we launched the reference-based task (see Section 3.4), which is open to our qualified workers as well as to any other workers satisfying basic qualification settings.

Qualified Workers after Pipeline We published the reference-based task to the 12 MTurk workers from four rounds who have passed both the qualification and the endurance task. All 12 workers accepted this task but only 8 workers finished 30 HITs within a week.

There are two scores to evaluate the information coverage between each candidate summary and the reference summary. We use the "can2ref" score to represent whether all information in the candidate summary can be found in the reference summary, and the "ref2can" score to represent the converse coverage. For both types of scores, we calculated Cohen's Kappa for every worker pair (given 4 candidate summaries per HIT, 30 HITS per worker). Cohen's Kappa for "can2ref" score ranges from 0.15 to 0.71, with a relatively high IAA between the first GOLD workers from the first two rounds (G_{11} and G_{21}). Similarly, Cohen's Kappa for "ref2can" score ranges from 0.14 to 0.66.

[6] S_{42} stands for the second SILVER worker from Round 4

[7] https://wordstotime.com/

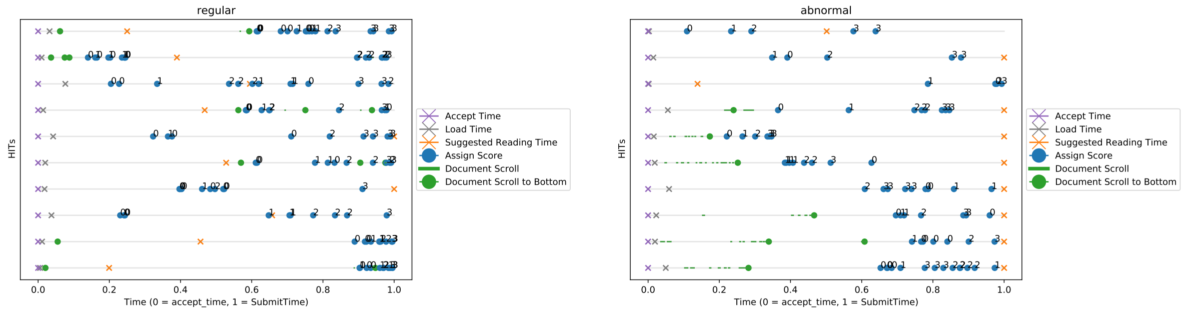


Figure 3: Comparison of online behaviors between the abnormal worker (S_{42}, right) and the regular worker (left).

Finally, Cohen's Kappa for the combined scores ranges from 0.15 to 0.68 (see Figure 4), demonstrating that the agreement numbers are stable across multiple measures. Krippendorff's Alpha for the above scenarios ("can2ref" score, "ref2can" score, and combined) are 0.558, 0.508, and 0.534.

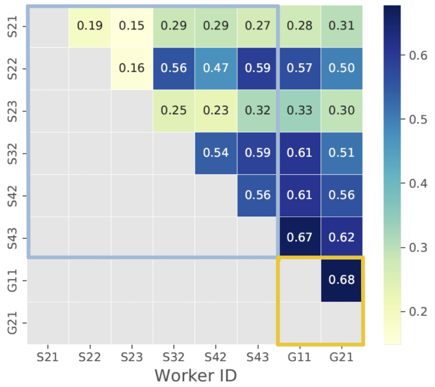


Figure 4: Cohen's Kappa for reference-based task (grey frame: SILVER workers; yellow: GOLD workers).

Baseline MTurk Workers For comparison, we published the same reference-based task to MTurk workers who did not participate in our previous experiments. 276 MTurk workers participated and each worker finished on average 2 HITs (In total 30 HITs × 20 Assignments/HIT). Krippendorff's Alpha for "can2ref", "ref2can", and the two combined were extremely low, at 0.087, 0.077, and 0.080 respectively, demonstrating the necessity of a high-quality recruitment pipeline. We experimented with the following approaches to investigate whether we could increase the agreement between random MTurk workers to a level comparable to qualified workers from our pipeline.

IAA with Median Among the 20 assignments of each HIT, we randomly divided the workers into 4 groups of 5 workers and took the median of each group representing a "new worker" (Lau et al., 2014). Then, we concatenated the results of 20 HITs for the 4 "new workers" to calculate IAA. Krippendorff's Alpha scores increased to 0.191, 0.185, and 0.188 respectively.

Filter on Timing and Number of Finished HITs To exclude unqualified workers whose annotations may decrease IAA, only workers who (i) spent more than the suggested reading time[8] and (ii) finished 3 or more HITs were selected for calculation of IAA. This resulted in 25 workers remaining, but Krippendorff's Alpha remained almost the same as calculated without the filter.

Statistical Filter (MACE) We applied the Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013; Paun et al., 2018) to identify reliable workers based on competence scores calculated on annotations. The workers with competence scores above a threshold were kept. We additionally calculated Spearman's coefficient (Spearman, 1904) within the groups of our pipeline and MACE (see Table 4). We report the results of additional failed attempts to improve Spearman's coefficient across these two groups, in Table 12 in the Appendix.

In summary, the most effective methods to improve agreement numbers among random workers were median grouping and MACE. IAA on median scores can raise Krippendorff's Alpha to almost 0.2. MACE increases Krippendorff's Alpha as the threshold increases, but at the cost of an incomplete HIT coverage (27/30 and 18/30 respectively for the threshold of 0.6 and 0.7 in Table 4) and fewer workers per HIT (1.9 and 1.2, respectively, for the threshold of 0.6 and 0.7 in Table 4). Similarly, Spearman's coefficient of MACE workers

[8] We performed the same timing analysis as in Section 4.3.

Threshold	0.5	0.6	0.7
% of workers kept	19.2%	15.9%	7.6%
HIT coverage	30/30	27/30	18/30
Avg. num. workers per HIT	2.4	1.9	1.2
Krippendorff's Alpha (all scores)	0.380	0.472	0.754
Spearman's coefficient (MACE workers)	0.351	0.414	0.770
Spearman's coefficient (pipeline workers)	0.558	0.565	0.577

Table 4: IAA for different thresholds of MACE.

can be increased above our pipeline workers' only at the same expense as above.

CloudResearch MTurk Workers To further test our pipeline, we conducted the same reference-based task on the CloudResearch platform (cloudresearch.com), which helps researchers recruit high-quality annotators. We recruited the same number (eight) of CloudResearch workers as our pipeline. The Krippendorff's Alpha and Cohen's Kappa[9] for CloudResearch workers is slightly lower than our pipeline workers (see Table 5 and Figure 9). Additionally, we found that our pipeline workers have a higher task acceptance rate. This results in a shorter experimental period compared to the task conducted on CloudResearch.

Worker Source	IAA Metric	can2ref	ref2can	combined score
Pipeline	CK	0.15-0.71	0.14-0.66	0.15-0.68
	KA	0.558	0.508	0.534
Cloud Research	CK	0.18-0.60	0.19-0.61	0.18-0.60
	KA	0.527	0.498	0.513

Table 5: The range of Cohen's Kappa (CK) and Krippendorff's Alpha (KA) of pipeline and CloudResearch workers for reference-based task.

Analysis of Correctness Across Annotation Sources We randomly sampled 50 annotation questions from the reference-based task to test correctness, which is defined as the alignment with expert judgments.[10] In addition, we also compared the expert judgment with scores generated by GPT models: GPT-3.5 ("text-davinci-003") and ChatGPT which are built on InstructGPT (Ouyang et al., 2022), and GPT-4 (OpenAI, 2023). Scores are aggregated by taking the median within groups of pipeline, MACE, and CloudResearch workers, as well as experts.[11] For ChatGPT we ran inference 5 times with default parameters (temperature=1, top_p=1) and took the median. To obtain GPT-3.5 and GPT-4 scores temperature was set to 0 with a single run.

Class	Group Type	Spearman's Coefficient	95% Confidence Interval
Crowd Annotators	Pipeline	0.03	(-0.61, 0.65)
	MACE	0.10	(-0.56, 0.69)
	CloudResearch	0.08	(-0.58, 0.67)
GPT models	GPT-3.5	0.73	(0.18, 0.93)
	ChatGPT	0.73	(0.20, 0.93)
	GPT-4	0.83	(0.41, 0.96)

Table 6: Spearman's coefficient of the expert judgment and groups for crowd annotators and GPT models.

We did not find that pipeline workers were superior to MACE workers in terms of correctness. Pipeline and CloudResearch workers had a significant Spearman's correlation with each other (see Figure 5), which indicates a reproduction of the recruitment procedure on CloudResearch at a lower cost. However, the confidence intervals are too wide to draw any conclusion about the correlation between crowd annotators and expert judgments (see Table 6). This indicates that the pipeline may not guarantee the training of the correctness of annotations. However, we found that GPT models correlated well with expert judgments. Further details can be found in Appendix A.7 and A.8.

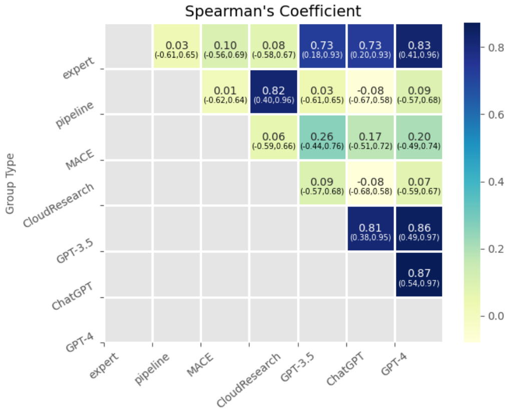


Figure 5: Spearman's coefficient for scores of 50 random samples in reference-based task among groups. 95% confidence interval is shown below the coefficient.

4.5 Discussion

In Section 4.4, we published the same reference-based task as a test to different crowd annotators

[9] The range of Cohen's Kappa is slightly smaller for CloudResearch workers.

[10] Fifty random samples were chosen in order to differentiate between MACE and pipeline assuming 20% superiority in terms of correctness.

[11] We use the median of a group of experts as the expert judgment, which has Krippendorff's Alpha of 0.52.

(pipeline, MACE, and CloudResearch). It showed that filtering workers *before* the actual evaluation task (pipeline) can avoid the waste of time and resources and achieve high agreement at a lower cost and a full coverage of HITs, compared to discarding annotations *after* the task (MACE) (see Table 7). Our pipeline also recruited workers of similar quality to CloudResearch at a lower cost; however, based on further analysis, the correctness of annotations was not guaranteed (see Section 7 for details). Besides, details about the estimated cost of GPT models for the reference-based task can be found in Table 15 in Appendix A.8.2.

	Pipeline	MACE (0.5)	CloudResearch
Num. of initial workers	200	276	45
% of workers kept	4%	19.2%	17.8%
HIT coverage	30/30	30/30	30/30
Avg. num. workers per HIT	8	2.4	8
Krippendorff's Alpha	**0.534**	0.380	0.513
Cost per worker (for Avg. num. workers per HIT)	**$27**	$175	$31

Table 7: Comparison between approaches of crowd annotators (pipeline, MACE, and CloudResearch) for the reference-based task.

5 Statistical Test for Stability of Pipeline

We next examined whether there was a difference in the probability of passing the qualification and endurance task among MTurk workers. Thus, we started by assuming the probability of passing each task for each round came from the same distribution, and we performed a statistical test as follows.

Let $\mathcal{X}$ denote the random variable representing the MTurk worker. For the qualification task, let $q_{x\in\mathcal{X}}(x)$ denote the binary random variable which has the value of 1 if the worker can pass the task, and 0 otherwise. Similarly, let $e_{x\in\mathcal{X}}(x)$ denote the binary random variable indicating whether the worker can pass the endurance task. Given 50 MTurk workers in each round, we use Q to denote the binary random variables in a round as (1). It can also be regarded as examples sampled from $q_{x\in\mathcal{X}}(x)$. Among the samples, the probability of a worker who can pass the qualification task is equal to the expectation of $q_{x\in\mathcal{X}}(x) = 1$ as (2). Since only workers who passed the qualification task are eligible for the endurance task, the probability of a worker passing the endurance task is equal to the expectation of $e_{x\in\mathcal{X},q(x)=1}(x) = 1$ as (3), which is a joint distribution of $q_{x\in\mathcal{X}}(x)$ and $e_{x\in\mathcal{X}}(x)$.

$$Q = \{q_{x_1\in\mathcal{X}}(x_1), ..., q_{x_{50}\in\mathcal{X}}(x_{50})\} \tag{1}$$

$$P(q_{x\in\mathcal{X}}(x) = 1) = \mathbb{E}(q_{x\in\mathcal{X}}(x) = 1) \tag{2}$$

$$\begin{aligned} &P(e_{x\in\mathcal{X},q(x)=1}(x) = 1) \\ =&\mathbb{E}(e_{x\in\mathcal{X},q(x)=1}(x) = 1) \\ =&P(e_{x\in\mathcal{X}}(x) = 1|q(x) = 1)P(q(x) = 1) \end{aligned} \tag{3}$$

Thus, we used the Bootstrap method (Efron, 1992) with 10,000 iterations to estimate the mean and standard deviation of the probability of passing the qualification and endurance task. Table 8 shows the results of all rounds with breakdowns of each round. We can see some variance that might come from MTurk workers given each round. To test whether there is a difference in the probability of passing each task among different rounds, we conducted the permutation test (Fisher, 1936; Pitman, 1937) for every two rounds. The results show that we cannot reject the null hypothesis that the underlying distributions of every two rounds are the same (see Appendix A.4).

Annotation Task	Qual. Task	End. Task
Pass Rate	0.13	0.06
Mean of Pass Rate (Bootstrap)	0.1302	0.0602
Standard Dev. of Pass Rate (Bootstrap)	0.0236	0.0168

Table 8: Statistical test results for stability of pipeline.

6 Conclusion

In this paper, we present a two-step recruitment pipeline that yields 12 qualified workers (4 GOLD and 8 SILVER workers) out of 200 MTurk workers with basic qualification settings in our experiments. We show that workers identified by our pipeline can (i) achieve a higher inter-annotator agreement than expert annotators in the endurance task, (ii) outperform the statistical filter (MACE) that discards annotation *after* the reference-based task, and (iii) replicate a proxy of CloudResearch annotations in the correctness analysis. Though the 6% yield rate is not as expected, our pipeline serves as the **best practice** to deliver high-agreement annotations and addresses the widespread waste of resources on low-quality annotations through filtering out subpar workers *before* they embark on large-scale tasks. In the future, we plan to build up a pool of reliable annotators who can deliver high-quality (both high agreement and correctness) evaluations on a large scale and in multiple tasks, languages, and platforms.

7 Limitations

This research creates a relatively complete pipeline to identify qualified MTurk workers for high-quality human evaluations based on existing techniques, and thoroughly tests the effectiveness of this pipeline both qualitatively and quantitatively compared to other methods. However, there are several limitations of this work:

- **The experiments are only conducted for summarization tasks in English on MTurk platform.** Thus, this pipeline can also be tested on other NLG tasks, in other languages, and on other platforms to see whether our three-step concept generalizes broadly to all human evaluations.
- **The specific questions designed for each task are not "panacea" solutions.** A better HIT design may exist for different experimental purposes, as long as it follows the ideas behind each task. For example, the endurance task aims to ensure the worker's reliable performance on a large number of annotations, so modifications based on this idea might work better in case-by-case scenarios[12].
- **There is no guarantee for the training of correctness in the pipeline though a high agreement is achieved.** An additional correctness check might need to be included along with the endurance task to achieve both high agreement and correctness through the filtering of the pipeline.

8 Ethical Considerations

Considering that crowd workers are often underpaid, experiments in this work all followed fair working wage standards[13] when using MTurk for recruitment purposes (details for each task are in Table 3). In addition, we have not rejected the work from any unqualified workers so far, though we reserve the right to do so when conducting the experiments.

In our experiments, personal data (any information relating to an identifiable natural person) was collected, processed, and stored based on certain data protection regulations,[14] given relevant privacy concerns. Special category information (i.e. personal data revealing racial or ethnic origin, etc.) was not included in this work. More information about the details of human evaluation experiments in this work can be found in the Human Evaluation Datasheet (HEDS) (Shimorina and Belz, 2022) in the Appendix.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful feedback on our paper. We would like to thank Claire Daniele for her editorial support. Mille's contribution was funded by the European Union under the Marie Skłodowska-Curie grant agreement No 101062572 (M-FleNS).

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2020. Identifying annotator bias: A new IRT-based method for bias identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4787–4797, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Antonio A Arechar, Gordon T Kraft-Todd, and David G Rand. 2017. Turking overtime: How participant characteristics and behavior vary over time and day on amazon mechanical turk. *Journal of the Economic Science Association*, 3(1):1–11.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3):351–368.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, 6(1):3–5.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

[12] We encourage starting the design from the reference-based task (which performs as the test of true annotation task) and thinking about what specific training the annotators are expected to have through the qualification and endurance task.

[13] https://livingwage.mit.edu/counties/27053

[14] https://gdpr.eu/article-4-definitions/

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

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.

Ronald Aylmer Fisher. 1936. Design of experiments. *British Medical Journal*, 1(3923):554.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *arXiv preprint arXiv:2202.06935*.

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.

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

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.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. 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.

Jessica Huynh, Jeffrey Bigham, and Maxine Eskenazi. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not. *arXiv preprint arXiv:2111.05241*.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

E. L. Kaplan and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan K. Kummerfeld. 2021. Quantifying and avoiding unfair qualification labour in crowdsourcing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–349, Online. Association for Computational Linguistics.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Emma Manning, Shira Wein, and Nathan Schneider. 2020. A human evaluation of AMR-to-English generation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4773–4786, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shikib Mehri, Jinho Choi, L. F. D'Haro, Jan Deriu, Maxine Eskénazi, Milica Gasic, Kallirroi Georgila, Dilek Z. Hakkani-Tür, Zekang Li, Verena Rieser, Samira Shaikh, David R. Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges. *ArXiv*, abs/2203.10012.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. The second multilingual surface realisation shared task (SR'19): Overview and evaluation results. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 1–17, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872.

Jonas Oppenlaender, Kristy Milland, Aku Visuri, Panos Ipeirotis, and Simo Hosio. 2020. Creativity on paid crowdsourcing platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

E. J. G. Pitman. 1937. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

Jonathan Robinson, Cheskie Rosenzweig, Aaron J. Moss, and Leib Litman. 2019. Tapped out or barely tapped? recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. *PLOS ONE*, 14(12):1–29.

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.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Margaret A Webb and June P Tangney. 2022. Too good to be true: Bots and bad data from mechanical turk. *Perspectives on Psychological Science*, page 17456916221120027.

Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair work: Crowd work minimum wage with one line of code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):197–206.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Proportion of Worker Categories in Qualification Task for Each Round

Annotation Task		Total Number of Workers	GOLD Workers	SILVER Workers	BROZE Workers	BLOCK Workers
Qualification Task	Round 1	50	1 (2%)	4 (8%)	32 (64%)	13 (26%)
	Round 2	50	3 (6%)	5 (10%)	29 (58%)	13 (26%)
	Round 3	50	2 (4%)	3 (6%)	24 (48%)	21 (42%)
	Round 4	50	2 (4%)	6 (12%)	27 (54%)	15 (30%)

Table 9: Proportion of worker categories for each round.

A.2 Cohen's Kappa for Each Summary in Endurance Task

For the figures below, "Answer.score_0" to "Answer.score_3" correspond to the scores aggregated from the 1st to 4th summary separately for each HIT. The dark color indicates a high IAA in terms of Cohen's Kappa score. $S42$ stands for the second SILVER worker from Round 4.

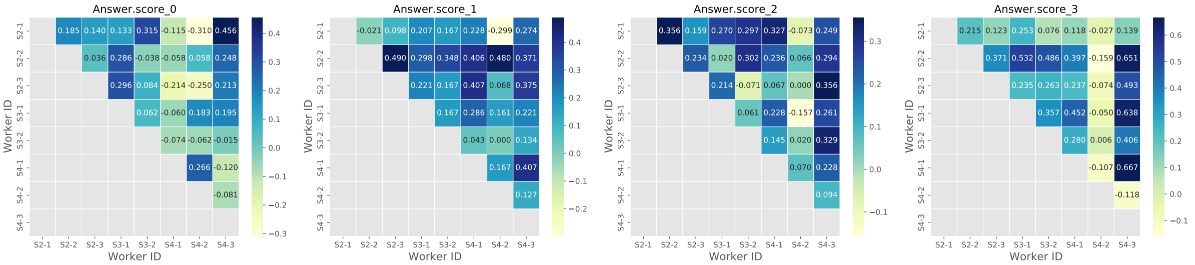


Figure 6: Cohen's Kappa for each summary among SILVER workers (Pairwise).

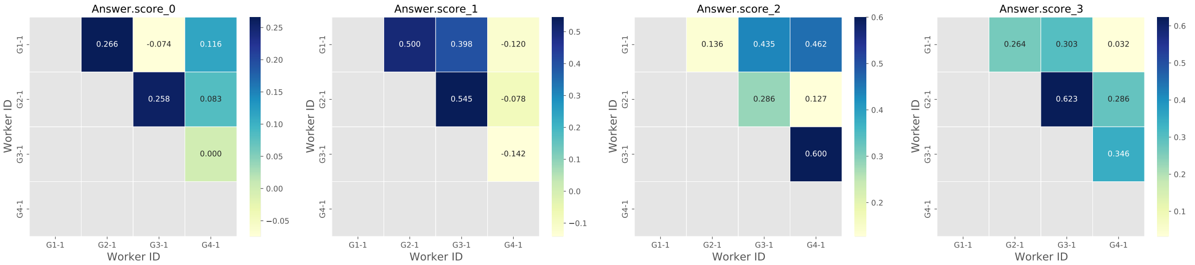


Figure 7: Cohen's Kappa for each summary among GOLD workers (Pairwise).

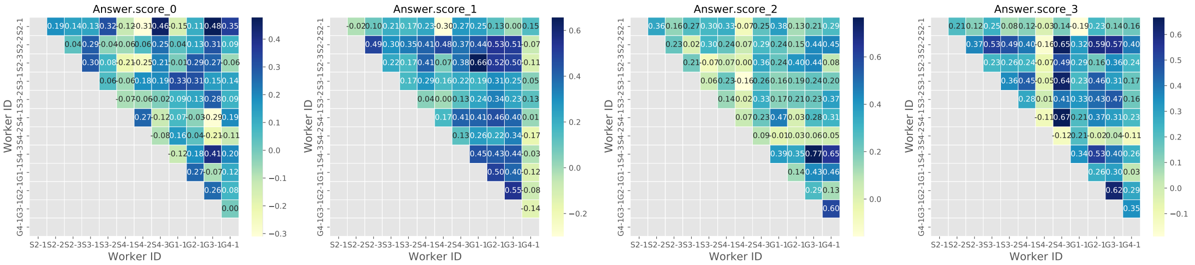


Figure 8: Cohen's Kappa for each summary across SILVER and GOLD workers (Pairwise).

A.3 Endurance Task Result of Lab Members

Worker Combination		A and B	B and C	C and A
Cohen's Kappa (Each Summary)	Answer.score_0	-0.261	-0.083	0.246
	Answer.score_1	0.285	0.13	0.285
	Answer.score_2	0.206	-0.006	-0.049
	Answer.score_3	0.066	0.006	0.387
Cohen's Kappa (Concatenation)		0.1	0.055	0.268
Cohen's Kappa (Omit first 2 HITs)		0.2	0.091	0.196
Krippendorff's Alpha			0.201	

Table 10: Endurance task result of lab members.

A.4 Statistical Test Results of Qualification and Endurance Tasks for Each Round

Annotation Task		Pass Rate	Mean of Pass Rate (Bootstrap)	Standard Dev. of Pass Rate (Bootstrap)
Round 1	Qua. Task	0.1	0.0997	0.0424
	End. Task	0.02	0.0199	0.0198
Round 2	Qua. Task	0.16	0.1611	0.0521
	End. Task	0.08	0.0805	0.0384
Round 3	Qua. Task	0.1	0.1000	0.0482
	End. Task	0.06	0.0599	0.0339
Round 4	Qua. Task	0.16	0.1595	0.0511
	End. Task	0.08	0.0800	0.0380
All Rounds	Qua. Task	0.13	0.1302	0.0236
	End. Task	0.06	0.0602	0.0168

Table 11: Statistical test results of qualification and endurance task.

A.5 Cohen's Kappa (combined scores) for CloudResearch Workers in Reference-based Task

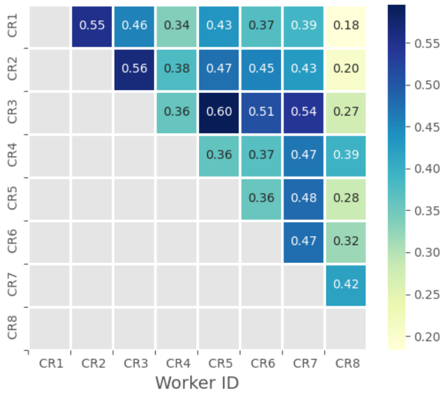


Figure 9: Cohen's Kappa (combined scores) among CloudResearch workers.

A.6 Spearman's Coefficient for Inter-groups (Pipeline & MACE) in Reference-based Task

For the reference-based task, we used 4 methods to calculate Spearman's coefficient:

- **Method 1**: Given the different numbers of remaining MACE workers for each HIT, we calculate Spearman's coefficient between our pipeline and MACE workers in each HIT. Then we take the average of these coefficients as the inter-group Spearman's coefficient shown in Table 12 [15].
- **Method 2**: The only difference between this method and Method 1 is that we take the absolute value when calculating Spearman's coefficient for each HIT.
- **Method 3**: We take the average of each annotation question in each HIT within the group of our pipeline and MACE workers separately, then concatenate these average scores of all HITs together for each group and calculate Spearman's coefficient.
- **Method 4**: The only difference between this method and Method 3 is that we calculate Spearman's coefficient for each HIT and then take the average of all coefficients instead of concatenating first and then calculating the coefficient.

Threshold		0.5	0.6	0.7
% of workers kept		19.2%	15.9%	7.6%
HIT coverage		30/30	27/30	18/30
Avg. num. workers per HIT		2.4	1.9	1.2
Krippendorff's Alpha (all scores)		0.380	0.472	0.754
Method 1	Spearman's coefficient (MACE workers)	0.351	0.414	0.770
	Spearman's coefficient (pipeline workers)	0.558	0.565	0.577
	Spearman's coefficient (inter-group)	-0.081	-0.063	-0.234
Method 2	Spearman's coefficient (MACE workers)	0.396	0.418	0.770
	Spearman's coefficient (pipeline workers)	0.575	0.580	0.591
	Spearman's coefficient (inter-group)	0.307	0.299	0.308
Method 3	Spearman's coefficient (inter-group)	-0.107	-0.067	-0.355
Method 4	Spearman's coefficient (inter-group)	-0.102	-0.113	-0.194

Table 12: Methods for calculation of Spearman's coefficient within and across groups of pipeline and MACE workers in reference-based task.

[15] We also calculate Spearman's coefficient within the group of our pipeline and MACE workers separately for comparison, as shown in Table 12.

A.7 Qualitative Analysis of Correctness Across Annotation Sources in Reference-based Task

For the reference-based task, we first randomly select 50 HITs out of 30 HITs (HIT index ranges from 0 to 29), and then 1 annotation question out of 8 questions (annotation index ranges from 0 to 7) for each of these HITs selected in the above step.

For each randomly selected annotation question, we calculate the median within the groups of our pipeline, MACE, and CloudResearch workers separately, as well as the scores generated by GPT models (GPT-3.5 (“text-davinci-003”), ChatGPT, and GPT-4[16]). The expert judgment (aggregated by the median) and details for 50 randomly selected annotation questions can be found in Table 13 and Table 14.

Figure 10 shows Spearman’s coefficient among different groups aggregated by the median before (left) and after (right) the removal of controversial HITs (HIT with index 15, 16, and 28). We also perform a similar analysis aggregated by the mean shown in Figure 11.

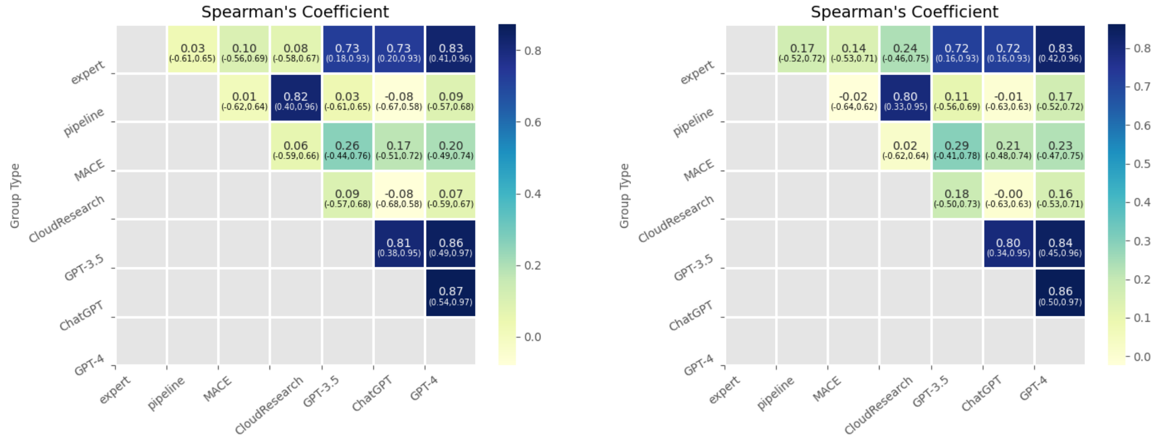


Figure 10: Spearman’s coefficient for scores of 50 random samples aggregated by **median** among groups before (left) and after (right) the removal of controversial HITs (95% confidence interval is shown below the coefficient).

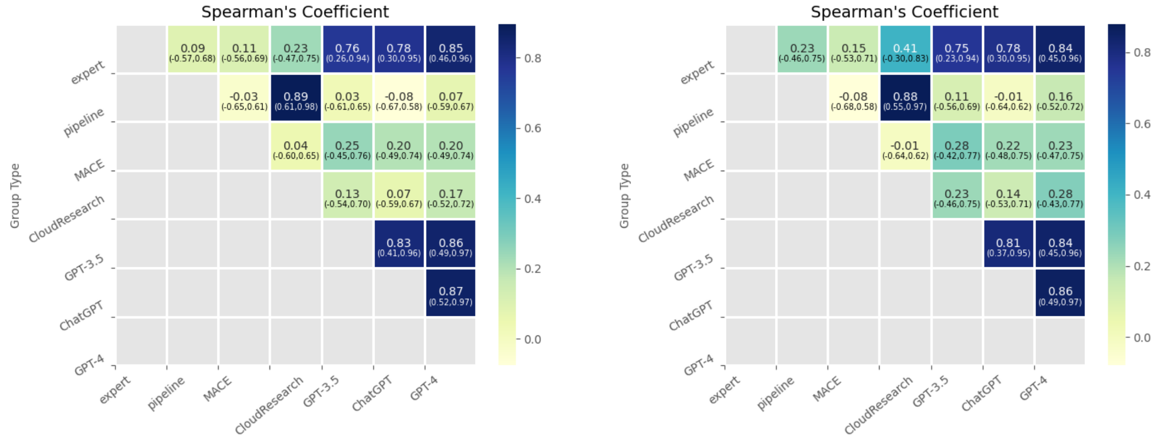


Figure 11: Spearman’s coefficient for scores of 50 random samples aggregated by **mean** among groups before (left) and after (right) the removal of controversial HITs (95% confidence interval is shown below the coefficient).

[16] For the ChatGPT score, we ran 5 times with default parameters (temperature=1, top_p=1) to take the median, but set the temperature as 0 with a single run for GPT-3.5 and GPT-4 scores.

Sample	Two Types of Summaries		Inclusion	Human Annotators (Median)			GPT series scores			Expert
Index			Direction	Pipeline	MACE	CloudResearch	GPT-3.5	ChatGPT	GPT-4	Judgment
1	Reference	The government has given regulators more time to investigate the proposed takeover of broadcaster Sky by 21st Century Fox.	can2ref	5.0	4.0	4.0	4.0	5.0	5.0	5.0
	Candidate	The government has extended the deadline for an inquiry into the takeover of Sky by 21st Century Fox.								
2	Reference	A Chinese woman has been found guilty of trespassing at President Donald Trump's Mar-a-Lago club in Florida and of lying to a federal agent.	can2ref	3.0	5.0	3.5	4	5	4.5	4.0
	Candidate	A Chinese woman who sparked alarm when she walked into US President Donald Trump's Mar-a-Lago resort has been found guilty of trespassing.								
3	Reference	A unique garden is helping Canadians to break a taboo that exists in many societies. It is allowing parents to talk openly about miscarriage.	ref2can	4.0	4.0	4.0	4.0	5.0	4.0	3.0
	Candidate	A Canadian cemetery has created a garden dedicated to the memory of babies lost during pregnancy. It's a place that's especially for those who have had multiple miscarriages.								
4	Reference	Gadgets that track your steps, sleeping and heart rate could help us live longer and cut national healthcare costs by billions - or so we are told.	can2ref	3.0	4.0	2.5	1.0	1.0	1.0	1.0
	Candidate	It is a huge amount of us have a smartphone, a smartphone and a gadget that feeds data from a smartphone.								
5	Reference	A unique garden is helping Canadians to break a taboo that exists in many societies. It is allowing parents to talk openly about miscarriage.	can2ref	2.0	4.0	2.0	4.0	5.0	4.0	3.0
	Candidate	A Canadian garden dedicated to the memory of children lost during pregnancy is helping to heal the pain of grief.								
6	Reference	The 2017 Oscar nominations are out, with La La Land the frontrunner. Here's a round-up of the surprises and talking points from this year's list.	can2ref	4.0	3.0	3.5	3.0	4.0	4.0	4.0
	Candidate	The full list of Oscar nominations has been announced. Here are 10 talking points from the shortlists.								
7	Reference	Welsh victims of the contaminated blood scandal have said it is not fair they get less financial help than people affected in England and Scotland.	ref2can	2.0	4.0	1.5	4.0	4.0	4.0	2.0
	Candidate	A man who contracted hepatitis C from the contaminated blood scandal has said Welsh support payments are not fair.								
8	Reference	An anonymous letter sent to a council outlining an alleged plan to oust head teachers is "defamatory", the leader of Birmingham City Council has said.	can2ref	4.0	4.0	4.0	3.0	1.0	3.0	2.0
	Candidate	A letter written by a council officer calling for schools to be taken over by a council has been defamatory.								
9	Reference	Graduates from ethnic minorities in Britain are less likely to be in work than their white peers, a study says.	ref2can	4.0	4.0	3.5	2.0	2.0	1.0	2.0
	Candidate	The number of ethnic minority graduates in the UK has fallen by almost 5% in the last year, according to a think tank.								
10	Reference	Two endangered red panda cubs have been born at a wildlife park on the Isle of Man.	ref2can	2.0	4.0	4.0	5.0	5.0	5.0	5.0
	Candidate	Two endangered red panda cubs have been born at a wildlife park in the Isle of Man.								
11	Reference	Two endangered red panda cubs have been born at a wildlife park on the Isle of Man.	ref2can	5.0	3.0	5.0	4.0	4.0	5.0	5.0
	Candidate	Two endangered red panda cubs have been born at a wildlife park on the Isle of Man, a year after a giant themed elephant calf escaped from his enclosure.								
12	Reference	Welsh Water has announced pre-tax profits of £7m for the last financial year.	can2ref	5.0	4.0	5.0	5.0	5.0	5.0	4.0
	Candidate	Welsh Water has announced pre-tax profits of £7m for the year to April.								
13	Reference	A "poo-powered" VW Beetle has taken to the streets of Bristol in an attempt to encourage sustainable motoring.	ref2can	4.0	4.0	2.5	4.0	4.0	4.0	3.0
	Candidate	A car powered by biogas has been seen on the streets of Bristol.								
14	Reference	An anonymous letter sent to a council outlining an alleged plan to oust head teachers is "defamatory", the leader of Birmingham City Council has said.	can2ref	5.0	3.0	4.5	4.0	5.0	4.0	3.0
	Candidate	A letter sent to Birmingham City Council by a whistle-blower has been described as "defamatory" by the city council's chief inspector of schools.								
15	Reference	In our media-saturated age, it's rare to have a chief executive who doesn't speak to the press or, indeed, very often publicly.	can2ref	5.0	4.0	5.0	1.0	1.0	1.0	1.0
	Candidate	Chinese entrepreneurs are a familiar sight.								
16	Reference	Parliament has been dissolved and the official election campaign has begun. BBC Reality Check listened in to Prime Minister Boris Johnson's campaign speeches in Downing Street and in Birmingham to check the facts and figures.	ref2can	5.0	4.0	5.0	3.0	3.0	3.0	1.0
	Candidate	Boris Johnson made a series of claims about his government's plans for the next few years. Here are six of the key pledges he made.								
17	Reference	Naturalist Sir David Attenborough and the Queen are the greatest living British man and woman, according to readers of Best of British magazine.	can2ref	3.0	4.0	3.5	4.0	5.0	4.0	4.0
	Candidate	David Attenborough has been voted the best of British by the magazine.								
18	Reference	An Edinburgh adventurer has become the youngest woman to ski solo to the South Pole.	can2ref	4.0	4.0	4.0	5.0	5.0	4.5	4.0
	Candidate	A woman from Edinburgh has become the youngest person to reach the South Pole solo.								
19	Reference	Resurfacing work on a newly-repaired canal towpath that washed away after vandals left a lock gate open has begun.	can2ref	4.0	3.0	3.5	4.0	4.0	4.0	5.0
	Candidate	Work has begun to resurface a canal towpath which was damaged by flooding.								
20	Reference	The Brexit vote is already having a negative impact on business, a survey of bosses from some of the UK's biggest companies has suggested.	ref2can	4.0	5.0	4.0	4.0	5.0	4.0	5.0
	Candidate	The majority of business leaders believe the Brexit vote has already had a negative impact on their company, a survey suggests.								
21	Reference	A campaign has begun to stop the spread of norovirus in Cornwall.	can2ref	5.0	5.0	3.5	4.0	5.0	5.0	5.0
	Candidate	A campaign has been launched to prevent the spread of norovirus in Cornwall.								
22	Reference	Welsh victims of the contaminated blood scandal have said it is not fair they get less financial help than people affected in England and Scotland.	can2ref	3.0	3.5	2.5	3.0	2.0	2.0	3.0
	Candidate	The Welsh Government has said it is not fair to pay for patients who have contaminated blood in the 1970s and 1980s.								
23	Reference	People on Jersey's Ecrehous islands are concerned travellers are arriving from France by boat and not being tested for coronavirus.	ref2can	1.0	4.0	1.0	3.0	3.0	3.0	2.0
	Candidate	People living on Jersey's Ecrehous islands have said they are worried about the number of people arriving ashore.								
24	Reference	The government has given regulators more time to investigate the proposed takeover of broadcaster Sky by 21st Century Fox.	can2ref	2.5	3.0	3.0	4.0	4.0	4.0	3.0
	Candidate	The government has extended its takeover inquiry into Sky's takeover deal with regulator Ofcom.								
25	Reference	Graduates from ethnic minorities in Britain are less likely to be in work than their white peers, a study says.	can2ref	3.0	3.5	3.0	2.0	2.0	1.0	2.0
	Candidate	The number of ethnic minority graduates in the UK has fallen by almost 5% in the last year, according to a think tank.								

Table 13: Qualitative analysis of correctness with 50 random samples (Part 1).

Sample		Two Types of Summaries	Inclusion	Human Annotators (Median)			GPT series scores			Expert
Index			Direction	Pipeline	MACE	CloudResearch	GPT-3.5	ChatGPT	GPT-4	Judgment
26	Reference	Joan Miro's 1927 work Peinture (Etoile Bleue) has sold for more than £23.5 million in London, setting a new auction record for the Spanish painter.	can2ref	4.0	5.0	4.0	3.0	1.0	2.0	3.0
	Candidate	Joan Miro's painting, which inspired the famous Joan Miro, has smashed its auction record for £15m.								
27	Reference	One of Oxford's main routes remains closed because of flooding for the second time in a month.	ref2can	2.5	5.0	2.0	5.0	5.0	5.0	5.0
	Candidate	A major route through Oxford has been closed for the second time in a month due to flooding.								
28	Reference	Holidaymakers say they have been left thousands of pounds out of pocket after a letting company ceased trading without notice.	ref2can	5.0	4.0	5.0	4.0	4.0	4.0	2.0
	Candidate	Brighton Holiday Homes has gone bust with bookings cancelled after a third of its customers claimed their money was lost.								
29	Reference	A £4.4m revamped Denbighshire leisure centre will open on Saturday.	cand2ref	4.0	5.0	3.0	5.0	4.0	4.0	4.0
	Candidate	A Denbighshire leisure centre is reopening on Thursday after a £4.4m revamp.								
30	Reference	Gadgets that track your steps, sleeping and heart rate could help us live longer and cut national healthcare costs by billions - or so we are told.	ref2cand	1.0	3.5	3.0	4.0	1.0	1.0	1.0
	Candidate	Every step we take is going to be tracked by a device that cannot simply put our fingers on our wrists.								
31	Reference	Gadgets that track your steps, sleeping and heart rate could help us live longer and cut national healthcare costs by billions - or so we are told.	cand2ref	2.0	3.0	4.0	4.0	1.0	1.0	1.0
	Candidate	Every step we take is going to be tracked by a device that cannot simply put our fingers on our wrists.								
32	Reference	Joan Miro's 1927 work Peinture (Etoile Bleue) has sold for more than £23.5 million in London, setting a new auction record for the Spanish painter.	ref2cand	2.0	5.0	4.0	4.5	4.0	4.0	4.0
	Candidate	A painting by Joan Miro has sold for £18.8m at auction, breaking the previous record for a work by the artist.								
33	Reference	A unique garden is helping Canadians to break a taboo that exists in many societies. It is allowing parents to talk openly about miscarriage.	cand2ref	3.0	3.0	4.0	3.0	4.0	5.0	5.0
	Candidate	A Canadian memorial garden is helping parents come to terms with the pain of losing a child during pregnancy.								
34	Reference	Holidaymakers say they have been left thousands of pounds out of pocket after a letting company ceased trading without notice.	cand2ref	3.0	4.0	4.0	4.0	4.0	3.0	4.0
	Candidate	A holiday home firm has gone bust after customers were told they had been left "heartbroken" after bookings were cancelled.								
35	Reference	A woman rescued after falling from a North Sea ferry has told how she thought she was going to die.	ref2cand	4.0	4.5	5.0	1.5	5.0	5.0	5.0
	Candidate	A woman who fell from a ferry into the North Sea has described how she thought she was going to die.								
36	Reference	The Brexit vote is already having a negative impact on business, a survey of bosses from some of the UK's biggest companies has suggested.	ref2cand	4.0	3.0	3.0	2.0	4.0	5.0	5.0
	Candidate	The UK's vote to leave the European Union is already having a negative impact on businesses, a survey suggests.								
37	Reference	Welsh Water has announced pre-tax profits of £7m for the last financial year.	ref2cand	4.0	4.5	4.0	4.0	5.0	5.0	5.0
	Candidate	Welsh Water has announced pre-tax profits of £7m for the year to April.								
38	Reference	One of Oxford's main routes remains closed because of flooding for the second time in a month.	ref2cand	5.0	3.0	5.0	3.5	5.0	5.0	5.0
	Candidate	A major route through Oxford has been closed for the second time in a month because of flooding.								
39	Reference	A 10-year-old boy died after he hit his head on a wall while playing football at school, an inquest heard.	ref2cand	4.0	4.0	3.0	3.5	4.0	5.0	5.0
	Candidate	A 10-year-old boy who hit his head while playing football at school died from traumatic brain injury, an inquest heard.								
40	Reference	A video artist who uses YouTube clips, a print-maker and an artist who pairs spoken word with photography are among this year's Turner Prize nominees.	ref2cand	3.0	4.0	3.5	3.5	4.0	4.0	4.0
	Candidate	A YouTube artist who splices together clips of horror films and a print-maker who works with women's groups are among the nominees for this year's Turner Prize.								
41	Reference	Parliament has been dissolved and the official election campaign has begun. BBC Reality Check listened in to Prime Minister Boris Johnson's campaign speeches in Downing Street and in Birmingham to check the facts and figures.	ref2cand	1.0	1.0	3.0	1.0	3.0	4.0	2.0
	Candidate	Boris Johnson has been making his pitch to Conservative voters in the final week of the election campaign. What did he get right and wrong?								
42	Reference	Film director Roman Polanski has been released after being questioned by prosecutors in Poland over sex offences in the US.	cand2ref	3.0	4.0	3.5	4.0	4.0	4.0	4.0
	Candidate	Polish film director Roman Polanski has been freed after prosecutors said they had not made an extradition bid for him.								
43	Reference	A video artist who uses YouTube clips, a print-maker and an artist who pairs spoken word with photography are among this year's Turner Prize nominees.	cand2ref	3.0	3.5	4.0	4.0	3.0	4.0	3.0
	Candidate	A video artist who uses YouTube and a storyteller who uses storytelling techniques are among the nominees for the 2014 Turner Prize.								
44	Reference	DJ Dave Lee Travis has told a court he does not have a "predatory nature".	ref2cand	4.0	3.5	4.0	4.0	4.0	4.0	4.0
	Candidate	Former radio DJ Dave Lee Travis has told a court he is "cuddly" not "predatory".								
45	Reference	Naturalist Sir David Attenborough and the Queen are the greatest living British man and woman, according to readers of Best of British magazine.	cand2ref	3.0	3.0	3.0	3.0	4.0	4.0	3.0
	Candidate	Sir David Attenborough has been named the best living British celebrity in a poll by the Magazine of British History.								
46	Reference	A Chinese woman has been found guilty of trespassing at President Donald Trump's Mar-a-Lago club in Florida and of lying to a federal agent.	ref2cand	2.0	5.0	3.0	4.5	1.0	1.0	1.0
	Candidate	A woman who sparked alarm at Mar-a-Lago has been found guilty of killing herself.								
47	Reference	Graduates from ethnic minorities in Britain are less likely to be in work than their white peers, a study says.	ref2cand	5.0	3.0	3.0	3.0	4.0	5.0	5.0
	Candidate	Black and ethnic minority graduates are less likely to be employed than white British counterparts, a report suggests.								
48	Reference	An anonymous letter sent to a council outlining an alleged plan to oust head teachers is "defamatory", the leader of Birmingham City Council has said.	ref2cand	2.0	4.5	4.0	3.5	3.0	2.0	3.0
	Candidate	A letter written by a council officer calling for schools to be taken over by a council has been defamatory.								
49	Reference	The 2017 Oscar nominations are out, with La La Land the frontrunner . Here's a round-up of the surprises and talking points from this year's list.	ref2cand	3.0	4.0	3.0	4.0	3.0	4.0	4.0
	Candidate	The full list of Oscar nominations has been announced. Here are 10 talking points from the shortlists.								
50	Reference	People on Jersey's Ecrehous islands are concerned travellers are arriving from France by boat and not being tested for coronavirus.	ref2cand	3.0	5.0	4.0	5.0	4.0	3.0	4.0
	Candidate	People living on Jersey's Ecrehous islands have said they fear they are "playing Russian roulette" with coronavirus restrictions after a rise in arrivals.								

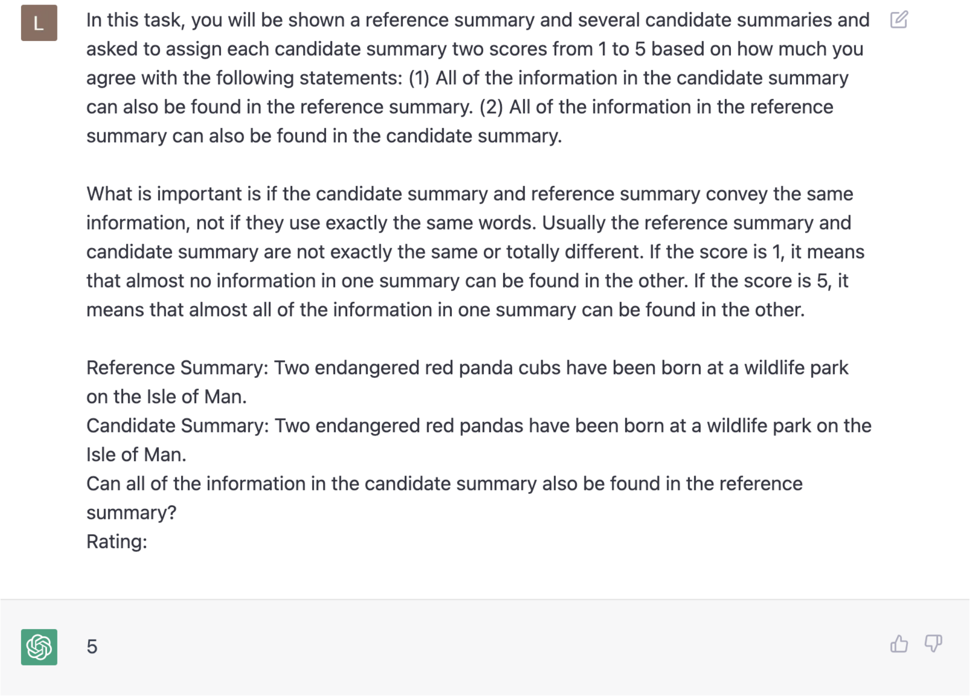
Table 14: Qualitative analysis of correctness with 50 random samples (Part 2).

A.8 Interaction with GPT models in Reference-based Task

A.8.1 Prompt Design

In Figure 12, we show an example of the interaction with ChatGPT and the exact prompt design we use to acquire scores generated by GPT models through API[17] for the analysis of correctness in the reference-based task.

This prompt design follows the instructions we provide to the crowd annotators in the reference-based task (see Figure 16 for details) with minor modifications for the score generation from GPT models. Details about running experiments through API can be found in Section 4.4.



In this task, you will be shown a reference summary and several candidate summaries and asked to assign each candidate summary two scores from 1 to 5 based on how much you agree with the following statements: (1) All of the information in the candidate summary can also be found in the reference summary. (2) All of the information in the reference summary can also be found in the candidate summary.

What is important is if the candidate summary and reference summary convey the same information, not if they use exactly the same words. Usually the reference summary and candidate summary are not exactly the same or totally different. If the score is 1, it means that almost no information in one summary can be found in the other. If the score is 5, it means that almost all of the information in one summary can be found in the other.

Reference Summary: Two endangered red panda cubs have been born at a wildlife park on the Isle of Man.
Candidate Summary: Two endangered red pandas have been born at a wildlife park on the Isle of Man.
Can all of the information in the candidate summary also be found in the reference summary?
Rating:

5

Figure 12: Example of interaction with ChatGPT in the reference-based task.

A.8.2 Estimated Cost of GPT Models

We estimate the cost of running GPT models for the score generation in the reference-based task (240 annotation questions in total) based on the cost of 50 random annotation questions. Details of pricing can be found on OpenAI's website[18]. We assume the GPT model only returns the score without explanations.

GPT Models	Cost per 1K Token	Estimated Cost
GPT-3.5	\$0.02	\$0.21
ChatGPT	\$0.002	\$0.02
GPT-4	\$0.03 (prompt) \$0.06 (completion)	\$0.32

Table 15: Estimated cost of GPT models for the reference-based task.

[17] https://platform.openai.com/docs/api-reference

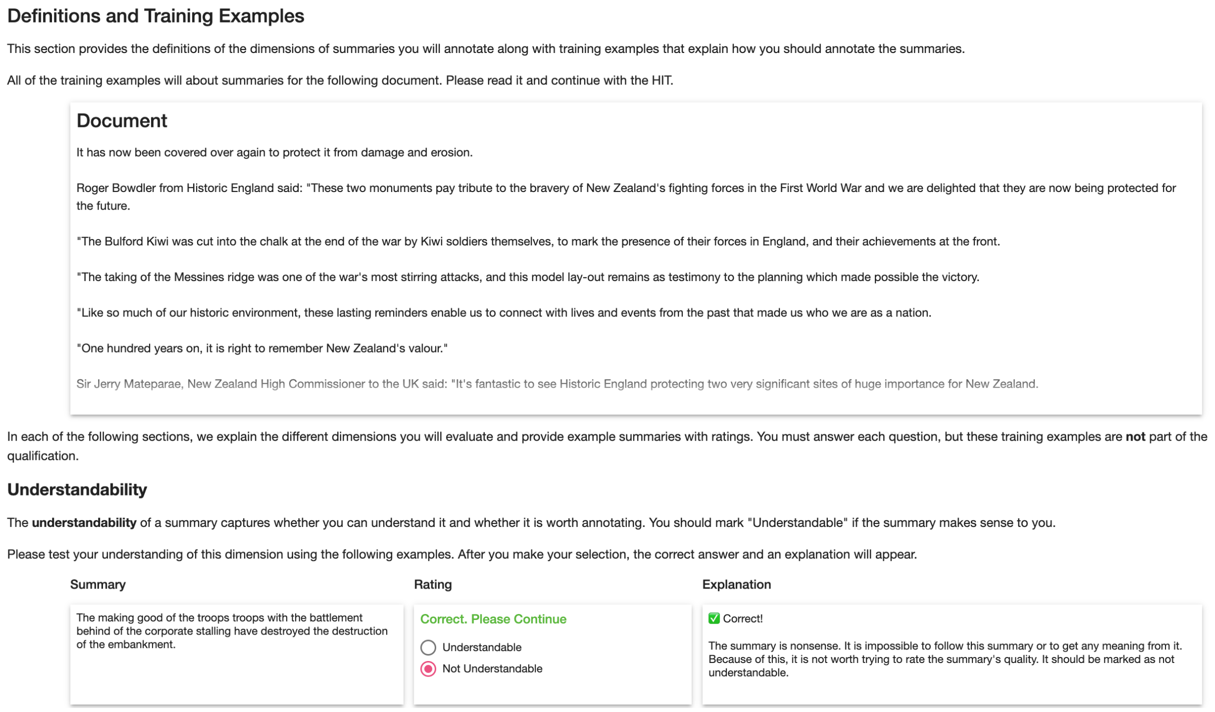
[18] https://openai.com/pricing

A.9 Instruction and Annotation Question Examples of HIT

Here we provide some examples of instructions and annotation questions for all three tasks as screenshots.

A.9.1 Qualification Task

- Figure 13 shows the definition of an evaluation dimension illustrated with examples in the training part.
- Figure 14 shows the example of the qualification question in the qualification part.



Definitions and Training Examples

This section provides the definitions of the dimensions of summaries you will annotate along with training examples that explain how you should annotate the summaries.

All of the training examples will about summaries for the following document. Please read it and continue with the HIT.

Document

It has now been covered over again to protect it from damage and erosion.

Roger Bowdler from Historic England said: "These two monuments pay tribute to the bravery of New Zealand's fighting forces in the First World War and we are delighted that they are now being protected for the future.

"The Bulford Kiwi was cut into the chalk at the end of the war by Kiwi soldiers themselves, to mark the presence of their forces in England, and their achievements at the front.

"The taking of the Messines ridge was one of the war's most stirring attacks, and this model lay-out remains as testimony to the planning which made possible the victory.

"Like so much of our historic environment, these lasting reminders enable us to connect with lives and events from the past that made us who we are as a nation.

"One hundred years on, it is right to remember New Zealand's valour."

Sir Jerry Mateparae, New Zealand High Commissioner to the UK said: "It's fantastic to see Historic England protecting two very significant sites of huge importance for New Zealand.

In each of the following sections, we explain the different dimensions you will evaluate and provide example summaries with ratings. You must answer each question, but these training examples are **not** part of the qualification.

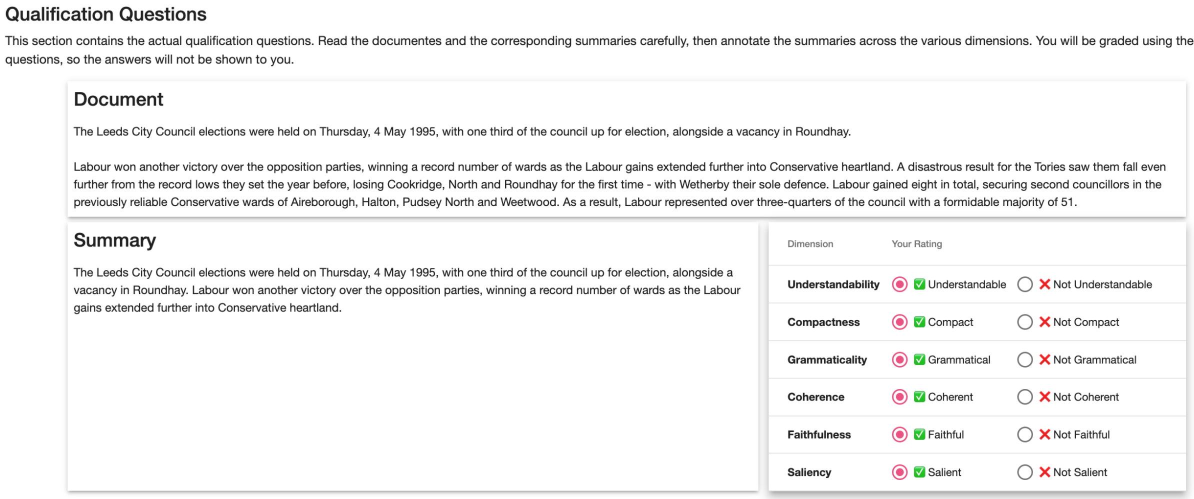
Understandability

The **understandability** of a summary captures whether you can understand it and whether it is worth annotating. You should mark "Understandable" if the summary makes sense to you.

Please test your understanding of this dimension using the following examples. After you make your selection, the correct answer and an explanation will appear.

Summary	Rating	Explanation
The making good of the troops troops with the battlement behind of the corporate stalling have destroyed the destruction of the embankment.	Correct. Please Continue ○ Understandable ● Not Understandable	✓ Correct! The summary is nonsense. It is impossible to follow this summary or to get any meaning from it. Because of this, it is not worth trying to rate the summary's quality. It should be marked as not understandable.

Figure 13: Example from training part of qualification task.



Qualification Questions

This section contains the actual qualification questions. Read the documentes and the corresponding summaries carefully, then annotate the summaries across the various dimensions. You will be graded using these questions, so the answers will not be shown to you.

Document

The Leeds City Council elections were held on Thursday, 4 May 1995, with one third of the council up for election, alongside a vacancy in Roundhay.

Labour won another victory over the opposition parties, winning a record number of wards as the Labour gains extended further into Conservative heartland. A disastrous result for the Tories saw them fall even further from the record lows they set the year before, losing Cookridge, North and Roundhay for the first time - with Wetherby their sole defence. Labour gained eight in total, securing second councillors in the previously reliable Conservative wards of Aireborough, Halton, Pudsey North and Weetwood. As a result, Labour represented over three-quarters of the council with a formidable majority of 51.

Summary

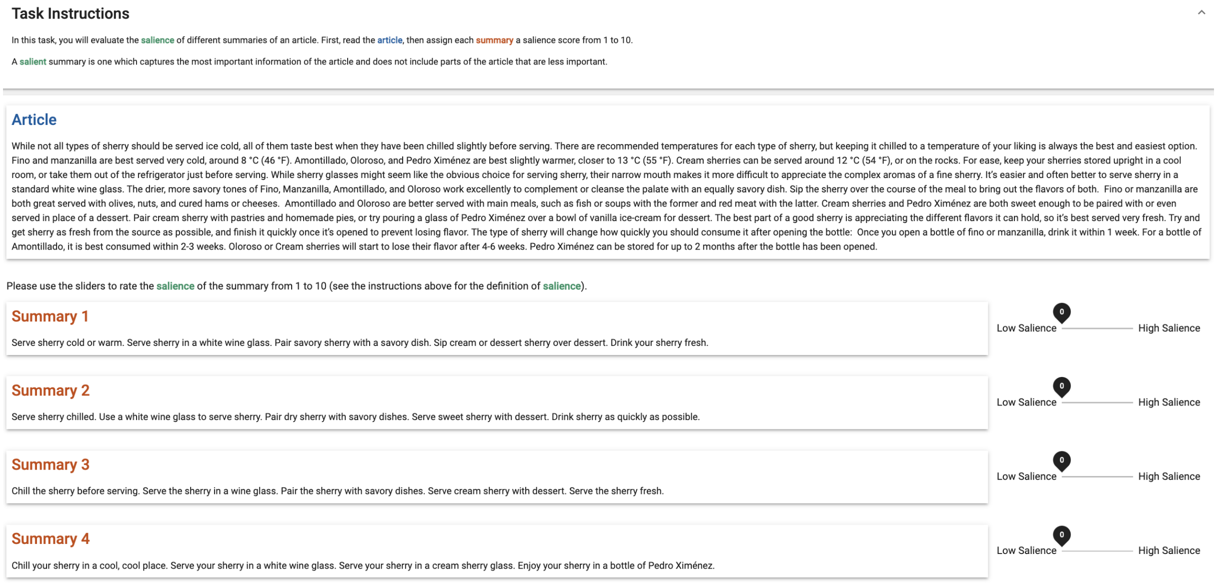
The Leeds City Council elections were held on Thursday, 4 May 1995, with one third of the council up for election, alongside a vacancy in Roundhay. Labour won another victory over the opposition parties, winning a record number of wards as the Labour gains extended further into Conservative heartland.

Dimension	Your Rating	
Understandability	● ✓ Understandable	○ ✗ Not Understandable
Compactness	● ✓ Compact	○ ✗ Not Compact
Grammaticality	● ✓ Grammatical	○ ✗ Not Grammatical
Coherence	● ✓ Coherent	○ ✗ Not Coherent
Faithfulness	● ✓ Faithful	○ ✗ Not Faithful
Saliency	● ✓ Salient	○ ✗ Not Salient

Figure 14: Example from qualification part of qualification task.

A.9.2 Endurance Task

Figure 15 shows the example of the annotation question on a Likert scale of 1 to 10 in the endurance task.



Task Instructions

In this task, you will evaluate the salience of different summaries of an article. First, read the article, then assign each summary a salience score from 1 to 10.

A salient summary is one which captures the most important information of the article and does not include parts of the article that are less important.

Article

While not all types of sherry should be served ice cold, all of them taste best when they have been chilled slightly before serving. There are recommended temperatures for each type of sherry, but keeping it chilled to a temperature of your liking is always the best and easiest option. Fino and manzanilla are best served very cold, around 8 °C (46 °F). Amontillado, Oloroso, and Pedro Ximénez are best slightly warmer, closer to 13 °C (55 °F). Cream sherries can be served around 12 °C (54 °F), or on the rocks. For ease, keep your sherries stored upright in a cool room, or take them out of the refrigerator just before serving. While sherry glasses might seem like the obvious choice for serving sherry, their narrow mouth makes it more difficult to appreciate the complex aromas of a fine sherry. It's easier and often better to serve sherry in a standard white wine glass. The drier, more savory tones of Fino, Manzanilla, Amontillado, and Oloroso work excellently to complement or cleanse the palate with an equally savory dish. Sip the sherry over the course of the meal to bring out the flavors of both. Fino or manzanilla are both great served with olives, nuts, and cured hams or cheeses. Amontillado and Oloroso are better served with main meals, such as fish or soups with the former and red meat with the latter. Cream sherries and Pedro Ximénez are both sweet enough to be paired with or even served in place of a dessert. Pair cream sherry with pastries and homemade pies, or try pouring a glass of Pedro Ximénez over a bowl of vanilla ice-cream for dessert. The best part of a good sherry is appreciating the different flavors it can hold, so it's best served very fresh. Try and get sherry as fresh from the source as possible, and finish it quickly once it's opened to prevent losing flavor. The type of sherry will change how quickly you should consume it after opening the bottle: Once you open a bottle of fino or manzanilla, drink it within 1 week. For a bottle of Amontillado, it is best consumed within 2-3 weeks. Oloroso or Cream sherries will start to lose their flavor after 4-6 weeks. Pedro Ximénez can be stored for up to 2 months after the bottle has been opened.

Please use the sliders to rate the salience of the summary from 1 to 10 (see the instructions above for the definition of salience).

Summary 1

Serve sherry cold or warm. Serve sherry in a white wine glass. Pair savory sherry with a savory dish. Sip cream or dessert sherry over dessert. Drink your sherry fresh.

Summary 2

Serve sherry chilled. Use a white wine glass to serve sherry. Pair dry sherry with savory dishes. Serve sweet sherry with dessert. Drink sherry as quickly as possible.

Summary 3

Chill the sherry before serving. Serve the sherry in a wine glass. Pair the sherry with savory dishes. Serve cream sherry with dessert. Serve the sherry fresh.

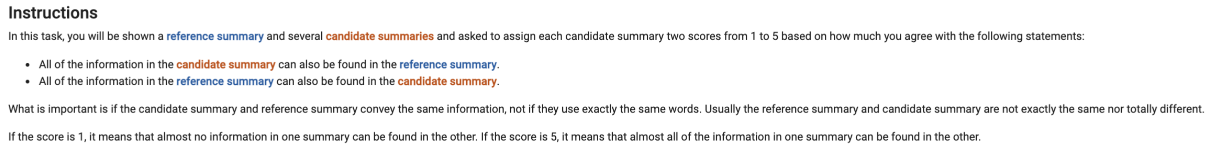
Summary 4

Chill your sherry in a cool, cool place. Serve your sherry in a white wine glass. Serve your sherry in a cream sherry glass. Enjoy your sherry in a bottle of Pedro Ximénez.

Figure 15: Example of the annotation question in endurance task.

A.9.3 Reference-based Task

- Figure 16 shows the instructions for the reference-based task.
- Figure 17 shows the example of the annotation question in the reference-based task.



Instructions

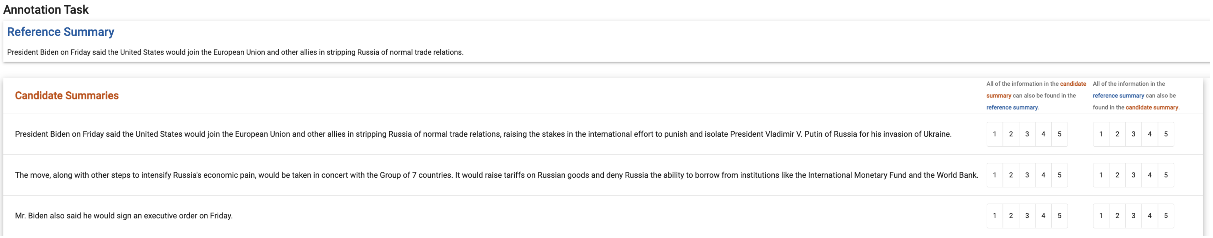
In this task, you will be shown a reference summary and several candidate summaries and asked to assign each candidate summary two scores from 1 to 5 based on how much you agree with the following statements:

- All of the information in the candidate summary can also be found in the reference summary.
- All of the information in the reference summary can also be found in the candidate summary.

What is important is if the candidate summary and reference summary convey the same information, not if they use exactly the same words. Usually the reference summary and candidate summary are not exactly the same nor totally different.

If the score is 1, it means that almost no information in one summary can be found in the other. If the score is 5, it means that almost all of the information in one summary can be found in the other.

Figure 16: Instructions for the reference-based task.



Annotation Task

Reference Summary

President Biden on Friday said the United States would join the European Union and other allies in stripping Russia of normal trade relations.

Candidate Summaries	All of the information in the candidate summary can also be found in the reference summary.	All of the information in the reference summary can also be found in the candidate summary.
President Biden on Friday said the United States would join the European Union and other allies in stripping Russia of normal trade relations, raising the stakes in the international effort to punish and isolate President Vladimir V. Putin of Russia for his invasion of Ukraine.	1 2 3 4 5	1 2 3 4 5
The move, along with other steps to intensify Russia's economic pain, would be taken in concert with the Group of 7 countries. It would raise tariffs on Russian goods and deny Russia the ability to borrow from institutions like the International Monetary Fund and the World Bank.	1 2 3 4 5	1 2 3 4 5
Mr. Biden also said he would sign an executive order on Friday.	1 2 3 4 5	1 2 3 4 5

Figure 17: Example of the annotation question in the reference-based task.

The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP

(described in Shimorina and Belz (2022))

1 Questions about Paper and Supplementary Resources (Questions 1.1–1.3)

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

Question 1.1: Link to paper reporting the evaluation experiment. If the paper reports more than one experiment, state which experiment you're completing this sheet for. Or, if applicable, enter 'for preregistration.'

Paper Link: `https://arxiv.org/abs/2212.10397` This sheet is completed for three experiments in the paper: Qualification Task, Endurance Task, and Reference-based Task.

What to enter in the text box: a link to an online copy of the main reference for the human evaluation experiment, identifying which of the experiments the form is being completed for if there are several. If the experiment hasn't been run yet, and the form is being completed for the purpose of submitting it for preregistration, simply enter 'for preregistration'.

Question 1.2: Link to website providing resources used in the evaluation experiment (e.g. system outputs, evaluation tools, etc.). If there isn't one, enter 'N/A'.

The data and code are available at `https://github.com/GEM-benchmark/MTurkRequirementPipeline`.

What to enter in the text box: link(s) to any resources used in the evaluation experiment, such as system outputs, evaluation tools, etc. If there aren't any publicly shared resources (yet), enter 'N/A'.

Question 1.3: Name, affiliation and email address of person completing this sheet, and of contact author if different.

Lining Zhang, New York University
lz2332@nyu.edu

What to enter in the text box: names, affiliations and email addresses as appropriate.

2 System Questions 2.1–2.5

Questions 2.1–2.5 record information about the system(s) (or human-authored stand-ins) whose outputs are evaluated in the Evaluation experiment that this sheet is being completed for.

The input, output, and task questions in this section are closely interrelated: the value for one partially determines the others, as indicated for some combinations in Question 2.3.

Question 2.1: What type of input do the evaluated system(s) take? Select all that apply. If none match, select 'Other' and describe.

Describe the type of input, where input refers to the representations and/or data structures shared by all evaluated systems.

This question is about input type, regardless of number. E.g. if the input is a set of documents, you would still select *text: document* below.

Check-box options (select all that apply):

□ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

□ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; Kamp and Reyle, 2013).

□ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

□ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.

✓ ***text: sentence***: a single sentence (or set of sentences).

✓ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).

✓ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.

□ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

□ ***text: other***: input is text but doesn't match any of the above *text:** categories.

□ ***speech***: a recording of speech.

□ ***visual***: an image or video.

□ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

□ ***control feature***: a feature or parameter specifically present to control a property of the output text, e.g. positive stance, formality, author style.

□ ***no input (human generation)***: human generation[1], therefore no system inputs.

□ ***other (please specify)***: if input is none of the above, choose this option and describe it.

Question 2.2: What type of output do the evaluated system(s) generate? Select all that apply. If none match, select 'Other' and describe.

Describe the type of output, where output refers to the representations and/or data structures shared by all evaluated systems.

This question is about output type, regardless of number. E.g. if the output is a set of documents, you would still select *text: document* below.

Note that the options for outputs are the same as for inputs except that the *no input (human generation) option* is replaced with *human-generated 'outputs'*, and the *control feature* option is removed.

Check-box options (select all that apply):

□ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

□ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; Kamp and Reyle, 2013).

□ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

□ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.

□ ***text: sentence***: a single sentence (or set of sentences).

□ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).

□ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.

□ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

[1] We use the term 'human generation' where the items being evaluated have been created manually, rather than generated by an automatic system.

□ ***text: other***: select if output is text but doesn't match any of the above *text:** categories.

□ ***speech***: a recording of speech.

□ ***visual***: an image or video.

□ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

✓ ***human-generated 'outputs'***: manually created stand-ins exemplifying outputs.[1]

□ ***other (please specify)***: if output is none of the above, choose this option and describe it.

Question 2.3: How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2? Occasionally, more than one of the options below may apply. If none match, select 'Other' and describe.

This field records the task performed by the system(s) being evaluated. This is independent of the application domain (financial reporting, weather forecasting, etc.), or the specific method (rule-based, neural, etc.) implemented in the system. We indicate mutual constraints between inputs, outputs and task for some of the options below.

Check-box options (select all that apply):

□ ***content selection/determination***: selecting the specific content that will be expressed in the generated text from a representation of possible content. This could be attribute selection for REG (without the surface realisation step). Note that the output here is not text.

□ ***content ordering/structuring***: assigning an order and/or structure to content to be included in generated text. Note that the output here is not text.

□ ***aggregation***: converting inputs (typically *deep linguistic representations* or *shallow linguistic representations*) in some way in order to reduce redundancy (e.g. representations for 'they like swimming', 'they like running' → representation for 'they like swimming and running').

□ ***referring expression generation***: generating *text* to refer to a given referent, typically represented in the input as a set of attributes or a linguistic representation.

□ ***lexicalisation***: associating (parts of) an input representation with specific lexical items to be used in their realisation.

□ ***deep generation***: one-step text generation from *raw/structured data* or *deep linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.

□ ***surface realisation (SLR to text)***: one-step text generation from *shallow linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.

□ ***feature-controlled text generation***: generation of text that varies along specific dimensions where the variation is controlled via *control features* specified as part of the input. Input is a non-textual representation (for feature-controlled text-to-text generation select the matching text-to-text task).

□ ***data-to-text generation***: generation from *raw/structured data* which may or may not include some amount of content selection as part of the generation process. Output is likely to be *text:** or *multi-modal*.

□ ***dialogue turn generation***: generating a dialogue turn (can be a greeting or closing) from a representation of dialogue state and/or last turn(s), etc.

□ ***question generation***: generation of questions from given input text and/or knowledge base such that the question can be answered from the input.

□ ***question answering***: input is a question plus optionally a set of reference texts and/or knowledge base, and the output is the answer to the question.

□ ***paraphrasing/lossless simplification***: text-to-text generation where the aim is to preserve the meaning of the input while changing its wording. This can include the aim of changing the text on a given dimension, e.g. making it simpler, changing its stance or sentiment, etc., which may be controllable via input features. Note that this task type includes meaning-preserving text simplification (non-meaning preserving simplification comes under *compression/lossy simplification* below).

□ ***compression/lossy simplification***: text-to-text generation that has the aim to generate a shorter, or shorter and simpler, version of the input text. This will normally affect meaning to some extent, but as a side effect, rather than the primary aim, as is the case in *summarisation*.

□ ***machine translation***: translating text in a source language to text in a target language while maximally preserving the meaning.

✓ ***summarisation (text-to-text)***: output is an extractive or abstractive summary of the important/relevant/salient content of the input document(s).

□ ***end-to-end text generation***: use this option if the single system task corresponds to more than one of tasks above, implemented either as separate modules pipelined together, or as one-step generation, other than *deep generation* and *surface realisation*.

□ ***image/video description***: input includes *visual*, and the output describes it in some way.

□ ***post-editing/correction***: system edits and/or corrects the input text (typically itself the textual output from another system) to yield an improved version of the text.

□ ***other (please specify)***: if task is none of the above, choose this option and describe it.

Question 2.4: Input Language(s), or 'N/A'.

This field records the language(s) of the inputs accepted by the system(s) being evaluated.

English.

What to enter in the text box: any language name(s) that apply, mapped to standardised full language names in ISO 639-1[2]. E.g. English, Herero, Hindi. If no language is accepted as (part of) the input, enter 'N/A'.

Question 2.5: Output Language(s), or 'N/A'.

This field records the language(s) of the outputs generated by the system(s) being evaluated.

English.

What to enter in the text box: any language name(s) that apply, mapped to standardised full language names in ISO 639-1 (2019)[2]. E.g. English, Herero, Hindi. If no language is generated, enter 'N/A'.

3 Questions about Output Sample, Evaluators, Experimental Design

3.1 Sample of system outputs (or human-authored stand-ins) evaluated (Questions 3.1.1–3.1.3)

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

Question 3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment? Answer should be an integer.

Qualification Task: 3*6=18
Endurance Task: 10*4=40
Reference-based Task: 30*8=240

What to enter in the text box: The number of system outputs (or other evaluation items) that are evaluated per system by at least one evaluator in the experiment, as an integer.

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

Multiple-choice options (select one):

○ ***by an automatic random process from a larger set***: outputs were selected for inclusion in the experiment by a script using a pseudo-random number generator; don't use this option if the script selects every nth output (which is not random).

[2] https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

✓ ***by an automatic random process but using stratified sampling over given properties***: use this option if selection was by a random script as above, but with added constraints ensuring that the sample is representative of the set of outputs it was selected from, in terms of given properties, such as sentence length, positive/negative stance, etc.

○ ***by manual, arbitrary selection***: output sample was selected by hand, or automatically from a manually compiled list, without a specific selection criterion.

○ ***by manual selection aimed at achieving balance or variety relative to given properties***: selection by hand as above, but with specific selection criteria, e.g. same number of outputs from each time period.

○ ***Other (please specify)***: if selection method is none of the above, choose this option and describe it.

Question 3.1.3: What is the statistical power of the sample size?

See Section 5 and Appendix A.4 for details.

What to enter in the text box: The results of a statistical power calculation on the output sample: provide numerical results and a link to the script used (or another way of identifying the script). See, e.g., Card et al. (2020); Howcroft and Rieser (2021).

3.2 Evaluators (Questions 3.2.1–3.2.5)

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

Question 3.2.1: How many evaluators are there in this experiment? Answer should be an integer.

Qualification Task: 200
Endurance Task: 26
Reference-based Task: 12

What to enter in the text box: the total number of evaluators participating in the experiment, as an integer.

Question 3.2.2: What kind of evaluators are in this experiment? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

□ ***experts***: participants are considered domain experts, e.g. meteorologists evaluating a weather forecast generator, or nurses evaluating an ICU report generator.

✓ ***non-experts***: participants are not domain experts.

✓ ***paid (including non-monetary compensation such as course credits)***: participants were given some form of compensation for their participation, including vouchers, course credits, and reimbursement for travel unless based on receipts.

□ ***not paid***: participants were not given compensation of any kind.

□ ***previously known to authors***: (one of the) researchers running the experiment knew some or all of the participants before recruiting them for the experiment.

✓ ***not previously known to authors***: none of the researchers running the experiment knew any of the participants before recruiting them for the experiment.

□ ***evaluators include one or more of the authors***: one or more researchers running the experiment was among the participants.

✓ ***evaluators do not include any of the authors***: none of the researchers running the experiment were among the participants.

□ ***Other*** (fewer than 4 of the above apply): we believe you should be able to tick 4 options of the above. If that's not the case, use this box to explain.

Question 3.2.3: How are evaluators recruited?

Qualification Task: On Amazon Mechanical Turk (MTurk) with pre-defined qualification settings (i.e. Location, etc.).
Endurance Task: On MTurk with evaluators who passed Qualification Task.
Reference-based Task: On MTurk with evaluators who passed Endurance Task.

What to enter in the text box: Please explain how your evaluators are recruited. Do you send emails to a given list? Do you post invitations on social media? Posters on university walls? Were there any gatekeepers involved? What are the exclusion/inclusion criteria?

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

Qualification Task: We include a training part to illustrate evaluation dimensions along with examples, and require evaluators to write an instruction summary in their own words.
Endurance Task: Evaluators are provided with task instructions.
Reference-based Task: Evaluators are provided with instructions and examples of the rating at the beginning of the task.

What to enter in the text box: Use this space to describe any training evaluators were given as part of the experiment to prepare them for the evaluation task, including any practice evaluations they did. This includes any introductory explanations they're given, e.g. on the start page of an online evaluation tool.

Question 3.2.5: What other characteristics do the evaluators have, known either because these were qualifying criteria, or from information gathered as part of the evaluation?

Qualification Task: Evaluators are satisfied with: (i) the Location as "UNITED STATES (US)"; (ii) the Number of HITs Approved is "greater than 1000"; (iii) the HIT Approval Rate (%) is "greater than or equal to 99".
Endurance Task: Evaluators pass the attention check and make no (GOLD) or only one mistake (SILVER) when annotating each dimension of the documents in the qualification part.
Reference-based Task: Evaluators (GOLD and SILVER) finish all 10 HITs in Endurance Task.

What to enter in the text box: Use this space to list any characteristics not covered in previous questions that the evaluators are known to have, either because evaluators were selected on the basis of a characteristic, or because information about a characteristic was collected as part of the evaluation. This might include geographic location of IP address, educational level, or demographic information such as gender, age, etc. Where characteristics differ among evaluators (e.g. gender, age, location etc.), also give numbers for each subgroup.

3.3 Experimental Design Questions 3.3.1–3.3.8

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

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

No.

What to enter in the text box: State 'Yes' or 'No'; if 'Yes' also give the name of the registry and a link to the registration page for the experiment.

Question 3.3.2: How are responses collected? E.g. paper forms, online survey tool, etc.

Amazon Mechanical Turk (MTurk).

What to enter in the text box: Use this space to describe how you collected responses, e.g. paper forms, Google forms, SurveyMonkey, Mechanical Turk, CrowdFlower, audio/video recording, etc.

Question 3.3.3: What quality assurance methods are used? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

□ ***evaluators are required to be native speakers of the language they evaluate***: mechanisms are in place to ensure all participants are native speakers of the language they evaluate.

✓ ***automatic quality checking methods are used during/post evaluation***: evaluations are checked for quality by automatic scripts during or after evaluations, e.g. evaluators are given known bad/good outputs to check they're given bad/good scores on MTurk.

✓ ***manual quality checking methods are used during/post evaluation***: evaluations are checked for quality by a manual process during or after evaluations, e.g. scores assigned by evaluators are monitored by researchers conducting the experiment.

✓ ***evaluators are excluded if they fail quality checks (often or badly enough)***: there are conditions under which evaluations produced by participants are not included in the final results due to quality issues.

✓ ***some evaluations are excluded because of failed quality checks***: there are conditions under which some (but not all) of the evaluations produced by some participants are not included in the final results due to quality issues.

□ ***none of the above***: tick this box if none of the above apply.

□ ***Other (please specify)***: use this box to describe any other quality assurance methods used during or after evaluations, and to provide additional details for any of the options selected above.

Question 3.3.4: What do evaluators see when carrying out evaluations? Link to screenshot(s) and/or describe the evaluation interface(s).

See details in Appendix A.9.

What to enter in the text box: Use this space to describe the interface, paper form, etc. that evaluators see when they carry out the evaluation. Link to a screenshot/copy if possible. If there is a separate introductory interface/page, include it under Question 3.2.4.

Question 3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

✓ ***evaluators have to complete each individual assessment within a set time***: evaluators are timed while carrying out each assessment and cannot complete the assessment once time has run out.

□ ***evaluators have to complete the whole evaluation in one sitting***: partial progress cannot be saved and the evaluation returned to on a later occasion.

□ ***neither of the above***: Choose this option if neither of the above are the case in the experiment.

□ ***Other (please specify)***: Use this space to describe any other way in which time taken or number of sessions used by evaluators is controlled in the experiment, and to provide additional details for any of the options selected above.

Question 3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

□ ***evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *before* starting on their assessments, either during or after training.

□ ***evaluators are told they can ask any questions during the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *during* their assessments.

✓ ***evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box***: evaluators are explicitly asked to provide feedback and/or comments about the experiment *after* their assessments, either verbally or in written form.

□ ***None of the above***: Choose this option if none of the above are the case in the experiment.

□ ***Other (please specify)***: use this space to describe any other ways you provide for evaluators to ask questions or provide feedback.

Question 3.3.7: What are the experimental conditions in which evaluators carry out the evaluations? If none match, select 'Other' and describe.

Multiple-choice options (select one):

✓ ***evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.***: evaluators are given access to the tool or form specified in Question 3.3.2, and subsequently choose where to carry out their evaluations.

○ ***evaluation carried out in a lab, and conditions are the same for each evaluator***: evaluations are carried out in a lab, and conditions in which evaluations are carried out *are* controlled to be the same, i.e. the different evaluators all carry out the evaluations in identical conditions of quietness, same type of computer, same room, etc. Note we're not after very fine-grained differences here, such as time of day or temperature, but the line is difficult to draw, so some judgment is involved here.

○ ***evaluation carried out in a lab, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a lab, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***evaluation carried out in a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out in a real-life situation, i.e. one that would occur whether or not the evaluation was carried out (e.g. evaluating a dialogue system deployed in a live chat function on a website), and conditions in which evaluations are carried out *are* controlled to be the same.

○ ***evaluation carried out in a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation (but not actually a real-life situation), e.g. user-testing a navigation system where the destination is part of the evaluation design, rather than chosen by the user. Conditions in which evaluations are carried out *are* controlled to be the same.

○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.

○ ***Other (please specify)***: Use this space to provide additional, or alternative, information about the conditions in which evaluators carry out assessments, not covered by the options above.

Question 3.3.8: Unless the evaluation is carried out at a place of the evaluators' own choosing, briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

The evaluation is carried out at a place of the evaluators' own choosing.

What to enter in the text box: use this space to describe the variations in the conditions in which evaluators carry out the evaluation, for both situations where those variations are controlled, and situations where they are not controlled.

4 Quality Criterion *n* – Definition and Operationalisation

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

4.1 Quality criterion properties (Questions 4.1.1–4.1.3)

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

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

Multiple-choice options (select one):

○ ***Correctness***: Select this option if it is possible to state, generally for all outputs, the conditions under which outputs are maximally correct (hence of maximal quality). E.g. for *Grammaticality*,[3] outputs are (maximally) correct if they contain no grammatical errors; for *Semantic Completeness*, outputs are correct if they express all the content in the input.

○ ***Goodness***: Select this option if, in contrast to correctness criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for any two outputs which is better and which is worse. E.g. for *Fluency*, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent.

✓ ***Feature***: Choose this option if, in terms of property X captured by the criterion, outputs are not generally better if they are more X, but instead, depending on evaluation context, more X may be either better or worse. E.g. for *Specificity*, outputs can be more specific or less specific, but it's not the case that outputs are, in the general case, better when they are more specific.

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

Multiple-choice options (select one):

○ ***Form of output***: Choose this option if the criterion assesses the form of outputs alone, e.g. *Grammaticality* is only about the form, a sentence can be grammatical yet be wrong or nonsensical in terms of content.

○ ***Content of output***: Select this option if the criterion assesses the content/meaning of the output alone, e.g. *Meaning Preservation* only assesses content; two sentences can be considered to have the same meaning, but differ in form.

✓ ***Both form and content of output***: Choose this option if the criterion assesses outputs as a whole, not just form or just content. E.g. *Coherence* is a property of outputs as a whole, either form or meaning can detract from it. Inherently extrinsic criteria such as *Usefulness* or *Task Completion* also fall in this category.

[3] We take all examples of quality criteria from published reports of evaluations, via the annotated database compiled by Howcroft et al. (2020).

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

Multiple-choice options (select one):

○ ***Quality of output in its own right***: Select this option if output quality is assessed without referring to anything other than the output itself, i.e. no system-internal or external frame of reference. E.g. *Poeticness* is assessed by considering (just) the output and how poetic it is.

✓ ***Quality of output relative to the input***: Choose this option if output quality is assessed relative to the input. E.g. *Answerability* is the degree to which the output question can be answered from information in the input.

○ ***Quality of output relative to a system-external frame of reference***: Choose this option if output quality is assessed with reference to system-external information, such as a knowledge base, a person's individual writing style, or the performance of an embedding system. E.g. *Factual Accuracy* assesses outputs relative to a source of real-world knowledge.

4.2 Evaluation mode properties (Questions 4.2.1–4.2.3)

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

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

Multiple-choice options (select one):

○ ***Objective***: Choose this option if the evaluation uses objective assessment, e.g. any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. Repeated assessments of the same output with an objective-mode evaluation method always yield the same score/result.

✓ ***Subjective***: Choose this option in all other cases. Subjective assessments involve ratings, opinions and preferences by evaluators. Some criteria lend themselves more readily to subjective assessments, e.g. *Friendliness* of a conversational agent, but an objective measure e.g. based on lexical markers is also conceivable.

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

Multiple-choice options (select one):

✓ ***Absolute***: Select this option if evaluators are shown outputs from a single system during each individual assessment.

○ ***Relative***: Choose this option if evaluators are shown outputs from multiple systems at the same time during assessments, typically ranking or preference-judging them.

Question 4.2.3: Is the evaluation intrinsic or extrinsic?

Multiple-choice options (select one):

✓ ***Intrinsic***: Choose this option if quality of outputs is assessed *without* considering their *effect* on something external to the system, e.g. the performance of an embedding system or of a user at a task.

○ ***Extrinsic***: Choose this option if quality of outputs is assessed in terms of their *effect* on something external to the system such as the performance of an embedding system or of a user at a task.

4.3 Response elicitation (Questions 4.3.1–4.3.11)

The questions in this section concern response elicitation, by which we mean how the ratings or other measurements that represent assessments for the quality criterion in question are obtained, covering what is presented to evaluators, how they select response and via what type of tool, etc. The eleven questions (4.3.1–4.3.11) are based on the information annotated in the large scale survey of human evaluation methods in NLG by Howcroft et al. (2020).

Question 4.3.1: What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if criterion not named.

We evaluate a summary according to 6 dimensions: Understandability, Compactness, Grammaticality, Coherence, Faithfulness, and Saliency.

What to enter in the text box: the name you use to refer to the quality criterion in explanations and/or interfaces created for evaluators. Examples of quality criterion names include Fluency, Clarity, Meaning Preservation. If no name is used, state 'N/A'.

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.

For a summary S,

- Understandability: can the worker understand S and is S worth being annotated.
- Compactness: S does not contain duplicated information.
- Grammaticality: S is free from grammatical spelling errors.
- Coherence: S is presented in a clear, wellstructured, logical, and meaningful way.
- Faithfulness: all of the information in S can also be found in the article; S accurately reflects the contents of the article.
- Saliency: S captures the most important information of the article and does not include parts of the article that are less important.

What to enter in the text box: Copy and past the verbatim definition you give to evaluators to explain the quality criterion they're assessing. If you don't explicitly call it a definition, enter the nearest thing to a definition you give them. If you don't give any definition, state 'N/A'.

Question 4.3.3: Size of scale or other rating instrument (i.e. how many different possible values there are). Answer should be an integer or 'continuous' (if it's not possible to state how many possible responses there are). Enter 'N/A' if there is no rating instrument.

Qualification Task: 2 (binary classification)
Endurance Task: 10 (10-point EASL scale)
Reference-based Task: 5 (5-point Likert scale)

What to enter in the text box: The number of different response values for this quality criterion. E.g. for a 5-point Likert scale, the size to enter is 5. For two-way forced-choice preference judgments, it is 2; if there's also a no-preference option, enter 3. For a slider that is mapped to 100 different values for the purpose of recording assessments, the size to enter is 100. If no rating instrument is used (e.g. when evaluation gathers post-edits or qualitative feedback only), enter 'N/A'.

Question 4.3.4: List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument.

Qualification Task: Yes, No
Endurance Task: 1-10
Reference-based Task: 1-5

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

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

Multiple-choice options (select one):

○ ***Multiple-choice options***: choose this option if evaluators select exactly one of multiple options.

○ ***Check-boxes***: choose this option if evaluators select any number of options from multiple given options.

○ ***Slider***: choose this option if evaluators move a pointer on a slider scale to the position corresponding to their assessment.

○ ***N/A (there is no rating instrument)***: choose this option if there is no rating instrument.

✓ ***Other (please specify)***: choose this option if there is a rating instrument, but none of the above adequately describe the way you present it to evaluators. Use the text box to describe the rating instrument and link to a screenshot.

Qualification Task: Multiple-choice options
Endurance Task: Slider
Reference-based Task: Multiple-choice options

Question 4.3.6: If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument.

N/A.

What to enter in the text box: If (and only if) there is no rating instrument, i.e. you entered 'N/A' for Questions 4.3.3–4.3.5, describe the task evaluators perform in this space. Otherwise, here enter 'N/A' if there *is* a rating instrument.

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

Qualification Task:

- In each of the following sections, we explain the different dimensions you will evaluate and provide example summaries with ratings. You must answer each question, but these training examples are not part of the qualification.
- This section contains examples of summaries and ratings for each dimension. These examples show how a summary can be good on one dimension and bad on another. Please read these examples and move on.
- To make sure you understand the instructions, please summarize them briefly in your own words (2-3 sentences). This is required as part of the qualification (min. 100 characters).
- This section contains the actual qualification questions. Read the documents and the corresponding summaries carefully, then annotate the summaries across the various dimensions. You will be graded using these questions, so the answers will not be shown to you.

Endurance Task: In this task, you will evaluate the salience of different summaries of an article. First, read the article, then assign each summary a salience score from 1 to 10. A salient summary is one which captures the most important information of the article and does not include parts of the article that are less important.
Please use the sliders to rate the salience of the summary from 1 to 10 (see the instructions above for the definition of salience).

Reference-based Task: In this task, you will be shown a reference summary and several candidate summaries and asked to assign each candidate summary two scores from 1 to 5 based on how much you agree with the following statements:

- All of the information in the candidate summary can also be found in the reference summary.
- All of the information in the reference summary can also be found in the candidate summary.

What is important is if the candidate summary and reference summary convey the same information, not if they use exactly the same words. Usually the reference summary and candidate summary are not exactly the same nor totally different.
If the score is 1, it means that almost no information in one summary can be found in the other. If the score is 5, it means that almost all of the information in one summary can be found in the other.

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

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

Multiple-choice options (select one):[4]

○ ***(dis)agreement with quality statement***: Participants specify the degree to which they agree with a given quality statement by indicating their agreement on a rating instrument. The rating instrument is labelled with degrees of agreement and can additionally have numerical labels. E.g. *This text is fluent — 1=strongly disagree...5=strongly agree*.

✓ ***direct quality estimation***: Participants are asked to provide a rating using a rating instrument, which typically (but not always) mentions the quality criterion explicitly. E.g. *How fluent is this text? — 1=not at all fluent...5=very fluent*.

○ ***relative quality estimation (including ranking)***: Participants evaluate two or more items in terms of which is better. E.g. *Rank these texts in terms of fluency*; *Which of these texts is more fluent?*; *Which of these items do you prefer?*.

○ ***counting occurrences in text***: Evaluators are asked to count how many times some type of phenomenon occurs, e.g. the number of facts contained in the output that are inconsistent with the input.

○ ***qualitative feedback (e.g. via comments entered in a text box)***: Typically, these are responses to open-ended questions in a survey or interview.

○ ***evaluation through post-editing/annotation***: Choose this option if the evaluators' task consists of editing or inserting annotations in text. E.g. evaluators may perform error correction and edits are then automatically measured to yield a numerical score.

○ ***output classification or labelling***: Choose this option if evaluators assign outputs to categories. E.g. *What is the overall sentiment of this piece of text? — Positive/neutral/negative.*

○ ***user-text interaction measurements***: choose this option if participants in the evaluation experiment interact with a text in some way, and measurements are taken of their interaction. E.g. reading speed, eye movement tracking, comprehension questions, etc. Excludes situations

[4] Explanations adapted from Howcroft et al. (2020).

where participants are given a task to solve and their performance is measured which comes under the next option.

○ ***task performance measurements***: choose this option if participants in the evaluation experiment are given a task to perform, and measurements are taken of their performance at the task. E.g. task is finding information, and task performance measurement is task completion speed and success rate.

○ ***user-system interaction measurements***: choose this option if participants in the evaluation experiment interact with a system in some way, while measurements are taken of their interaction. E.g. duration of interaction, hyperlinks followed, number of likes, or completed sales.

○ ***Other (please specify)***: Use the text box to describe the form of response elicitation used in assessing the quality criterion if it doesn't fall in any of the above categories.

Question 4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion? State if no scores reported.

We use raw responses to calculate Inter-Annotator Agreement (IAA), but sometimes the median of scores is taken to increase IAA.

What to enter in the text box: normally a set of separate assessments is collected from evaluators and is converted to the results as reported. Describe here the method(s) used in the conversion(s). E.g. macro-averages or micro-averages are computed from numerical scores to provide summary, per-system results.

Question 4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion.

See Section 5 and Appendix A.4 for details.

What to enter in the text box: A list of methods used for calculating the effect size and significance of any results, both as reported in the paper given in Question 1.1, for this quality criterion. If none calculated, state 'None'.

Question 4.3.11: Has the inter-annotator and intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used, and what are the agreement scores?

We use Cohen's Kappa and Krippendorff's Alpha for the inter-annotator agreement between evaluators. For the agreement scores, see Section 4 for details.

What to enter in the text box: the methods used to compute, and results obtained from, any measures of inter-annotator and intra-annotator agreement obtained for the quality criterion.

5 Ethics Questions (Questions 5.1-5.4)

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

Question 5.1: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?

This research is conducted by following the equivalent hourly rate listed here: `https://livingwage.mit.edu/counties/27053`

What to enter in the text box: Typically, research organisations, universities and other higher-education institutions require some form ethical approval before experiments involving human participants, however innocuous, are permitted to proceed. Please provide here the name of the body that approved the experiment, or state 'No' if approval has not (yet) been obtained.

Question 5.2: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions/)? If yes, describe data and state how addressed.

In our experiments, personal data (any information relating to an identifiable natural person) was collected, processed, and stored based on certain data protection regulations, given relevant privacy concerns.

What to enter in the text box: State 'No' if no personal data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements such as privacy and security was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.

Question 5.3: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited/)? If yes, describe data and state how addressed.

No.

What to enter in the text box: State 'No' if no special-category data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements relating to special-category data was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.

Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes.

No.

What to enter in the text box: Use this box to describe any *ex ante* or *ex post* impact assessments that have been carried out in relation to the evaluation experiment, such that the assessment plan and process, as well as the outcomes, were captured in written form. Link to documents if possible. Types of impact assessment include data protection impact assessments, e.g. under GDPR.[5] Environmental and social impact assessment frameworks are also available.

Credits

Questions 2.1–2.5 relating to evaluated system, and 4.3.1–4.3.8 relating to response elicitation, are based on Howcroft et al. (2020), with some significant changes. Questions 4.1.1–4.2.3 relating to quality criteria, and some of the questions about system outputs, evaluators, and experimental design (3.1.1–3.2.3, 4.3.5, 4.3.6, 4.3.9–4.3.11) are based on Belz et al. (2020). HEDS was also informed by van der Lee et al. (2019, 2021) and by Gehrmann et al. (2021)'s[6] data card guide.

More generally, the original inspiration for creating a 'datasheet' for describing human evaluation experiments of course comes from seminal papers by Bender and Friedman (2018), Mitchell et al. (2019) and Gebru et al. (2020).

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

[5] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/

[6] https://gem-benchmark.com/data_cards/guide

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.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for datasets.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

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.

David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more underpowered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

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.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

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.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Please see Section 7.

☑ A2. Did you discuss any potential risks of your work?
Please see Section 8.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Please see the abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Please see Section 3.

☑ B1. Did you cite the creators of artifacts you used?
Please see Section 3.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Please see Section 3.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Please see Section 3.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Please see Section 8.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Please see Section 3.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Please see Section 4.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Please see Section 3&4 for details of the implementation and results involving human annotators.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Please see Section 3 and Appendix A.7.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Please see Section 3.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Please see Section 8.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Please see Section 3.1.