Common Law Annotations: Investigating the Stability of Dialog System Output Annotations

Seunggun Lee[1] **Alexandra DeLucia**[2] **Nikita Nangia**[1] **Praneeth S. Ganedi**[1]
Ryan Min Guan[3] **Rubing Li**[1] **Britney A. Ngaw**[1] **Aditya Singhal**[1]
Shalaka Vaidya[1] **Zijun Yuan**[1] **Lining Zhang**[1] **João Sedoc**[1]
[1]New York University [2]Johns Hopkins University [3]University of Pennsylvania
jsedoc@stern.nyu.edu

Abstract

Metrics for Inter-Annotator Agreement (IAA), like Cohen's Kappa, are crucial for validating annotated datasets. Although high agreement is often used to show the **reliability** of annotation procedures, it is insufficient to ensure **validity** or **reproducibility**. While researchers are encouraged to increase annotator agreement, this can lead to specific and tailored annotation guidelines. We hypothesize that this may result in diverging annotations from different groups. To study this, we first propose the **Le**e **e**t **a**l. **P**rotocol (**LEAP**), a standardized and codified annotation protocol. **LEAP** strictly enforces transparency in the annotation process, which ensures **reproducibility** of annotation guidelines. Using **LEAP** to annotate a dialog dataset, we empirically show that while research groups may create **reliable** guidelines by raising agreement, this can cause divergent annotations across different research groups, thus questioning the **validity** of the annotations. Therefore, we caution NLP researchers against using **reliability** as a proxy for **reproducibility** and **validity**.

https://github.com/jsedoc/common-law-annotations

1 Introduction

The acquisition of **reliable**, **valid**, and **reproducible** human annotations is an essential component of Natural Language Processing (NLP) research. However, human annotations are inherently subjective (Basile et al., 2021) and each annotator has their own biases (Paun et al., 2022). To overcome this subjectivity, research groups aim to develop annotation guidelines that increase **Inter-Annotator Agreement** (IAA) among annotators, also known as inter-rater reliability. **Reliability**–the level of agreement between the annotators–is a necessary, but not sufficient condition for **reproducibility** (Artstein, 2017). If the precise details of the annotation process–from creating the annotation guidelines to executing the annotations themselves–are not transparent, the annotations may not be reproducible. Furthermore, high **reliability** does not guarantee **validity**–the extent to which annotations accurately capture what is intended to be measured (Paun et al., 2022).

To address these challenges, we first propose **Le**e **e**t **a**l. **P**rotocol (**LEAP**) a codified annotation guideline creation process that standardizes the way research groups create, publicize, and implement annotation guidelines. **LEAP** ensures transparency in the annotation process through its step-by-step procedure, which is crucial to allow for better reproducibility and cross-paper analyses.

Second, we use **LEAP** to investigate agreement by having pairs of researchers simulate the annotation procedure of different research groups on a common dataset, in order to observe the change in agreement within and between these groups. Within the simulation, we observe that each group creates their own unique guidelines, despite working on the same dataset and annotation categories. We leverage the metaphor of a **common law**, in which country/region-specific laws are based on precedent, much like researchers agreeing on common rules for edge cases to increase agreement. Similar to common laws differing between countries, the rules governing annotation guidelines can become increasingly research group-specific and divergent from other groups, as each group strives to raise their IAA. After developing annotation guidelines, we analyze if these observations persist when crowdsourcing the data with each guideline.

While **LEAP** has broader applications, here we apply it to a conversational AI task, where common human annotation metrics include *Appropriateness*, *Information content of output*, and *Humanlikeness* (Howcroft et al., 2020). The popularity and recent advances in dialog agents, such as OpenAI's GPT-

Findings of the Association for Computational Linguistics: ACL 2023, pages 12315–12349
July 9-14, 2023

3 (Brown et al., 2020), ChatGPT,[1] and YouChat[2] motivated us to showcase our method in the dialog domain.

In our investigation, we ask the following research questions:

1. How are the agreement levels different for researchers within and across groups?
2. Do groups converge or diverge in their annotation guidelines?
3. Which groups are able to get the crowdsource workers to agree most? Is it the same as the other groups?
4. Do crowdsource workers converge or diverge within and between groups?

Ultimately, we make the following contributions:

- Empirically show that while groups may create reliable guidelines by artificially raising agreement, this can lead to divergent annotations across different research groups, thus questioning annotation **validity**.
- Propose **LEAP** as a standardized and transparent annotation protocol which ensures **reproducibility** of annotation guidelines, while also allowing for deeper analysis of **validity** influenced by divergent annotation guidelines.

2 Related Work

This paper contributes to the ongoing discourse on the 'science' of annotations (Hovy and Lavid, 2010). Similar to Hober et al. (2023), we call for improved reliability and transparency in annotations.

2.1 Reporting Pitfalls & Errors

The NLP / NLG community generally lacks error reporting (van Miltenburg et al., 2021). Agreement studies and works involving annotations are no exception to this problem. We assert that papers should report the caveats of their work, especially regarding agreement analysis, which we believe makes research more robust. We offer a standardized solution through **LEAP**, where our protocol ensures each published work exposes its entire annotation life-cycle.

2.2 Annotation Protocols

The benefits of crowdsourcing methods are widely recognized and used in fields beyond NLP, including healthcare studies (Hamilton et al., 1994) as well as Psychology (Cuccolo et al., 2021). In particular, the Psychology research community has established notable researcher crowdsourcing initiatives, such as CREP (Grahe et al., 2020), the Pipeline Project (Schweinsberg et al., 2016), and Psi Chi's Network for International Collaborative Exchange: Crowd component (NICE: Crowd) (Cuccolo et al., 2022), which outline standardized practices and methodologies to ensure quality data collection.

Within the NLP field, there are several annotation protocols that outline steps within the annotation development cycle. The MATTER cycle (**Model**, **Annotate**, **Train**, **Test**, **Evaluate**, **Revise**) offers a high-level outline for collecting annotations to train and develop machine learning models (Pustejovsky and Stubbs, 2012). The MAMA (Model-Annotate-Model-Annotate) cycle–a subsection of the MATTER cycle–describes the iterative procedure of refining guidelines and collecting annotations to arrive at an optimal annotation model (Pustejovsky and Stubbs, 2012). The CASCADES model further extends the Model and Revise portion of MATTER, with the steps **Conceptual Analysis**, **Abstract Syntax**, **Semantics**, and **Concrete Syntax** (Bunt et al., 2010). For a deeper analysis of these protocols and their implementations, see Artstein (2017). With GENIE, Khashabi et al. (2022) address reproducibility concerns by providing a platform to run and study annotation results across a variety of text generation tasks.

While such annotation protocols help standardize the annotation procedures, they do not entirely enforce the total transparency of the annotation procedures. To the best of our knowledge, LEAP is the first annotation protocol to strictly require complete transparency in the annotation guideline creation process through recorded discussions and transcripts to ensure full reproducibility and effective cross-paper analysis.

2.3 Divergent Annotation Guidelines

Though divergent annotation guidelines between research groups may seem natural due to each group's unique research purpose, this often occurs among research groups who have similar purposes (e.g., evaluating a new dialogue system). Researchers would benefit by using consistent standardized annotation approaches.

For example, numerous papers created their own definitions for the category of *Appropriateness* (Reiter et al., 2000; van Deemter, 2000), *Information*

[1] https://openai.com/blog/chatgpt/
[2] https://youchat.com/

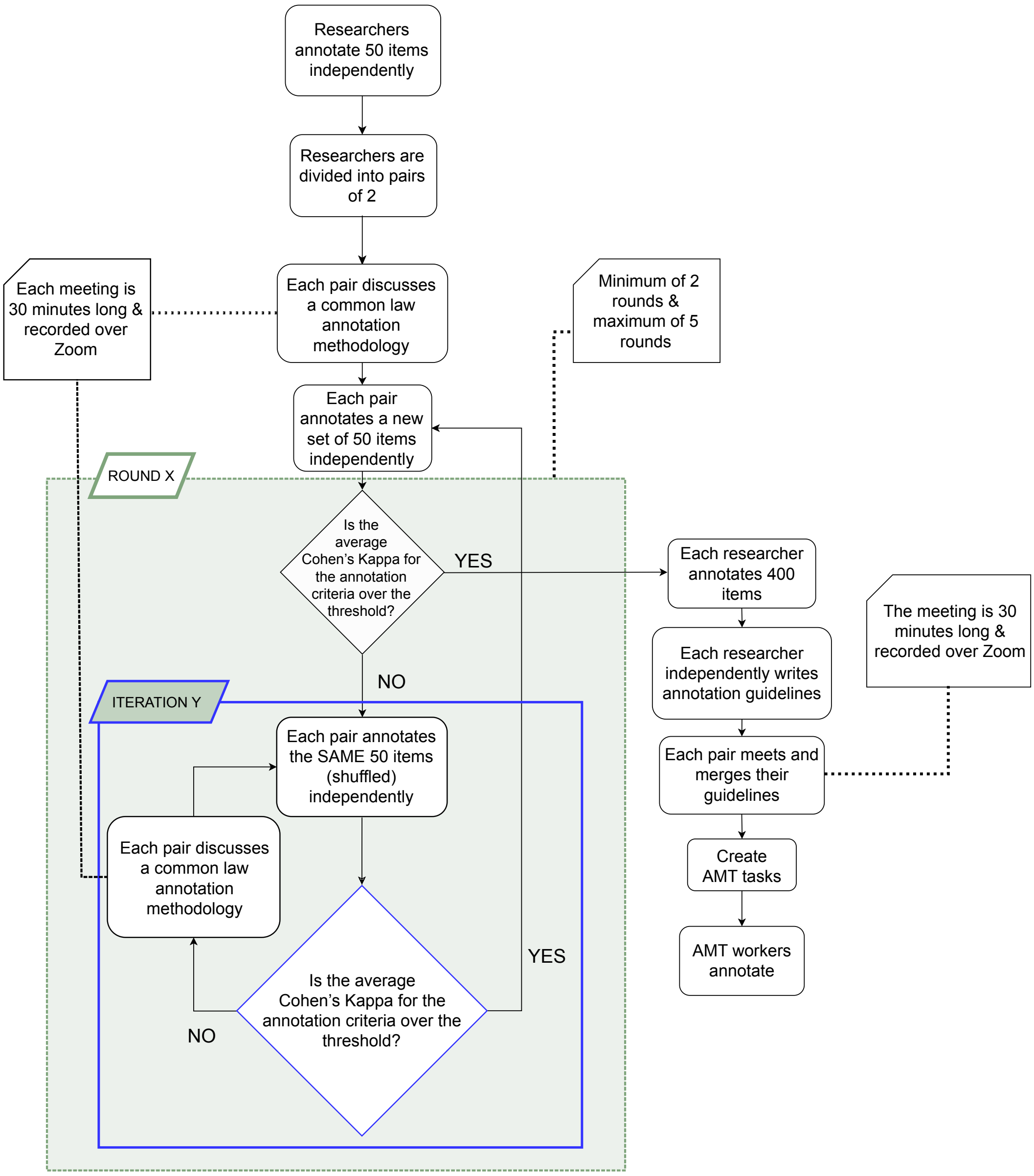


Figure 1: A flowchart of **LEAP** used for our experiments.

content of outputs (Carenini and Cheung, 2008; Filippova and Strube, 2008), and *Humanlikeness* (Agirrezabal et al., 2013; Cercas Curry et al., 2015) (See Appendix A.3 for more examples). Furthermore, though papers may use annotation categories that are different verbatim, the categories often overlap in meaning and purpose (Finch and Choi, 2020).

2.4 Disagreement in Annotations

Basile et al. (2021) emphasizes the importance of observing and embracing inherent disagreement in annotation tasks, arguing that focusing on a single 'ground truth' reference obscures the complexity and subjectivity of human-annotated data (Pavlick and Kwiatkowski, 2019; Uma et al., 2022).

In fact, in SummEval (Fabbri et al., 2021) they found that crowd worker annotations had reasonable IAA but were uncorrelated to expert annotators who also had high IAA. This suggests a flaw in the current annotation paradigm. Instead, we propose in our work that a pair of researchers should first converge with high IAA on a subset of the dataset. Then the pair should create the instructions and design for the crowd annotation task and validate the agreement.

In our work, we extend this study of disagreement by empirically illustrating how artificially eradicating irreconcilable disagreement can harm accuracy (and thus potentially harm **validity**).

3 Experiment Design

3.1 LEAP

Figure 1 illustrates the codified steps of **LEAP** for our experiment. In the following paragraphs, we explain the core components of **LEAP**. While the procedure outlined below is tailored for dialog annotations, the overall method can be adapted to other tasks (see Appendix subsection A.1).

Parameters To customize LEAP for a specific annotation task, several parameters need to be selected. These parameters include:

- Minimum and maximum number of rounds
- Agreement criteria and threshold
- Common law discussion time limit
- Number of researchers involved (minimum of 2)
- Number of items per small iteration
- Number of items for the larger annotation

Our advice to practitioners is that while each of these can be modified during an annotation process, the best practice is to set them *a priori* based on a smaller pilot, prior studies, or budget limitations.

Annotations Annotations are done independently, on the same subset of data. During the annotation, annotators are *not* allowed to communicate with each other. After each iteration of annotations, the agreement score is calculated for each annotation category. The agreement scores are shared with the annotators.

Annotation Discussions Each pair of annotators in a research group use discussions to walk through and compare their annotations. During discussions, annotators are asked to resolve edge cases that are causing disagreement, ultimately working towards a shared understanding of each category's annotation guidelines.

All discussions are conducted using a *recorded* video-conference platform, such as Zoom,[3] to ensure full transparency of the annotation process. Discussions are limited to a pre-specified amount of time. As researchers compare individual annotation examples, screen-share is enabled to make the process transparent, while transcript tools are enabled to allow for efficient analysis post-experiment.[4]

The quintessential idea for the records is to ensure that the decisions made during the meeting are documented as they may provide insights into construct validity and also help in understanding survey design. Since recording may not be available for all situations (e.g., automatic transcription does not support all languages), an alternative is to maintain detailed notes during the discussions.

Rounds & Iterations Prior to developing the final annotation guidelines, **LEAP** requires researchers to annotate multiple subsets of data.

Each *round* consists of a subset of a given dataset. Each annotation session is termed as an *iteration*. After a pair of researchers complete an *iteration* of annotations, the agreement score for each annotation category is calculated. The *average* of the agreement scores across the annotation categories is used to compare against a pre-designated threshold level of agreement.

If the category average agreement score meets the threshold, the researchers move on to the next *round* of annotations. This next round uses a new subset of the dataset. However, if the category average agreement score does *not* meet the threshold, the researchers are unable to move to the next *round* of annotations. Rather, the researchers discuss the most recent iteration of annotations to fine-tune their shared understanding of the annotation categories. Then the researchers conduct the next *iteration* of annotations. In the new *iteration*, researchers annotate the *same* subset of data, however, the presentation order is shuffled. Iterations allow researchers to test their level of convergence. This step is repeated until the researchers are able to meet their desired threshold, upon which they move on to the next *round* of annotations.

Post Convergence Round Once the researchers complete their rounds of annotations, they annotate a larger set of new items.[5] This larger round of items is used to compare the crowd worker ratings with the researchers and evaluate consistency over a large set of annotations.

Creating the Annotation Guidelines The final component of the protocol is creating the annotation guidelines. Similar to the discussions, this process is made transparent through recorded screen share and live transcripts.

There are several benefits to such an iterative annotation procedure. First, researchers are able to find and fix pitfalls and mistakes in the anno-

[3]https://zoom.us/

[4]The recordings will not be shared publicly.

[5]After conducting a pilot round of annotations, we chose 400 items to be the appropriate amount of annotations which would guarantee statistical significance.

tation process by experiencing them directly. Furthermore, through the iterative process, researchers are able to systemically fine-tune their annotations to construct a shared understanding of the annotation categories. Finally, the iterative process allows the researchers to retroactively analyze the discussions conducted after each annotation session in a structured manner.

3.2 Experimental Design

Data For this task, we generated model responses using prompts from the English as a Second Language (ESL) (Sedoc et al., 2019) and Daily Dialog (Li et al., 2017) evaluation sets (1,323 prompts). For each prompt, we generated model responses using eight state-of-the-art conversational models, including DialoGPT (Zhang et al., 2020), GPT-3 (Brown et al., 2020), Plato2 (Bao et al., 2021), and BlenderBot 2 (Weston and Shuster, 2021; Komeili et al., 2022; Xu et al., 2022). In total, we created 11,907 prompt-response pairs. The prompts and model responses have been detokenized to avoid revealing the model origins to the annotator. We used the dialog prompts and the language generation systems within their intended usage. For more information on the model parameters, see Appendix A.4.

Instructions The experiment followed the **LEAP** architecture. The goal of each group was to create annotation guidelines that would help other annotators annotate conversational text data as similarly as possible. The annotations consisted of *static* evaluations, as they are one of the most used forms of human evaluations in NLP (Finch and Choi, 2020). Following Howcroft et al. (2020), we provided the following base definitions for three annotation categories:

1. *Appropriateness*: The degree to which the output is appropriate in the given context.
2. *Information content of outputs*: The amount of information conveyed by an output.
3. *Humanlikeness*: The degree to which the output could have been produced by a human.

We intentionally kept the category definitions simple to give each group freedom in devising their own annotation guidelines. See Appendix Figure 6 for an example of the prompt and response annotated by the researchers.

See Appendix Figure 8 for the tabular step-by-step instructions–created using **LEAP**–shared with all researcher annotators. For specific instructions on creating annotation guidelines, shared with all researcher annotators, see Appendix Figure 7.

LEAP parameters In order to maintain inter-group consistency, each group was instructed to use a five-point ordinal scale. For our agreement criteria, we chose linear Cohen's κ as it is commonly used. We ran a small pilot and estimated that the Cohen's κ 95% confidence interval was ± 0.1 with 50 annotations and ± 0.05 with 400 annotations.[6] Cohen's κ of 0.6 to 0.8 is commonly regarded as a threshold for sufficient inter-annotator agreement in NLP research (Landis and Koch, 1977) thus we chose a category average Cohen's κ of 0.7 as the threshold. To time-bound the process, we chose a minimum of 2 rounds and a maximum of 5 rounds. In our pilot, we also found that 30 minutes was sufficient to discuss annotation differences.

Groups We simulated the process of six individual research groups (Group 1-6) defining guidelines for human annotation of conversational data. Each group consisted of two researchers. The group pairings had diverse members in terms of gender identity and annotation experience.[7]

3.3 Crowdsourced Annotation Parameters

Once all the annotation guidelines have been created, we used Amazon Mechanical Turk (MTurk)[8] to collect crowdsourced data. For our experiment, groups did not iterate over the annotation guidelines with crowd workers.

Instructions Each crowd worker was given the following instructions:

> The annotation task is to label responses to a given prompt. The prompt consists of two people (A and B) talking to each other. The response is the next utterance after the final utterance in the prompt.

Then the annotators were given the annotation guideline based on the group task chosen (see Figures 9 to 14 in the Appendix).

All MTurk tasks were deployed using the same portion of the dataset as the round of 400 prompts and responses that were annotated by the researchers. This choice was made because the round

[6] We used `https://search.r-project.org/CRAN/refmans/DescTools/html/CohenKappa.html`.

[7] All data was collected without any information that names or uniquely identify individuals.

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

of 400 annotations was the latest set of annotations done by the researchers, meaning the researchers' annotations were most calibrated with the annotation guidelines.[9] Groups 1 and 2 used the full LEAP protocol with iterations.

3.4 Testing Iteration-Free LEAP

While the iterations in **LEAP** give researchers the opportunity to converge on their common law annotation guidelines, we acknowledge that this may require additional time and resources. Furthermore, it reduces the independent relationship between annotations. Thus, we tested an iteration-free version of **LEAP** (see Figure 5 for a flowchart).

The iteration-free version of **LEAP** excludes the iteration component. If a group is unable to reach the pre-designated agreement threshold, they move on to the next *round* of annotations. This allows researchers to annotate more data while converging; however, they cannot discuss a subset of data multiple times. Iteration-free **LEAP** favors coverage over convergence. A new round of annotations consists of a new subset of data. Groups 3, 4, 5, and 6 used the iteration-free **LEAP**.

4 Results & Discussion

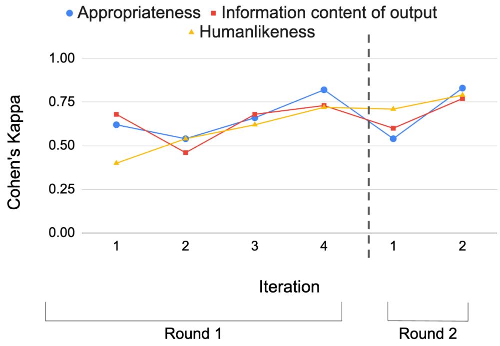
4.1 Agreement Analysis - LEAP

***Within* Group** By using the iterative annotation procedure of **LEAP**, Group 1 and Group 2 were able to achieve a high level of agreement on the second iteration of the second round of annotations. Figure 2 illustrates the change in agreement for Groups 1 and 2.

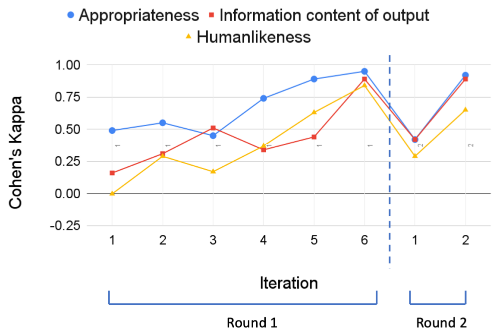
We also observed a drop in agreement for both groups when moving from round 1 to round 2. This is expected, as the change in annotated data introduces new edge cases, causing divergence between annotators. However, as both groups were able to calibrate their annotations via the iterations in round 1, round 2 required substantially fewer amounts of iterations to achieve the threshold of 0.7.

***Between* Groups** Taking advantage of the standardized annotation protocol codified through **LEAP**, we analyzed the changes in the agreement between annotators of different groups. Figure 3 illustrates the changes in agreement for annotators within the same group and between different groups.

[9]For additional information regarding crowd worker metadata, compensation, qualifications, and quality checks, see Appendix A.7.



(a)



(b)

Figure 2: Agreement scores for Groups 1 (above) and 2 (below) with using **LEAP**.

In *round 1* and *round 2*, for all three categories, *within*-group agreement–that is the level of agreement between annotators of the *same* group–was relatively higher than *between*-group agreement, or the level of agreement between annotators of *different* groups. Such observation suggests that raising agreement levels through fine-tuned annotation guidelines can cause divergence across different research groups.

Interestingly, we observed a relatively higher level of *between*-group agreement for *Appropriateness*, despite the fact that researchers in Group 1 and Group 2 never communicated with one another. This suggests that certain annotation categories, such as *Appropriatness*, have a stronger shared construct than others.

4.2 Agreement Analysis - Iteration-Free LEAP

Groups 3, 4, 5, and 6 tested the iteration-free **LEAP**. None of the groups were able to reach the designated threshold of an average Cohen's $\kappa > 0.7$. In addition, we found supporting evidence of divergence across annotators of different groups.

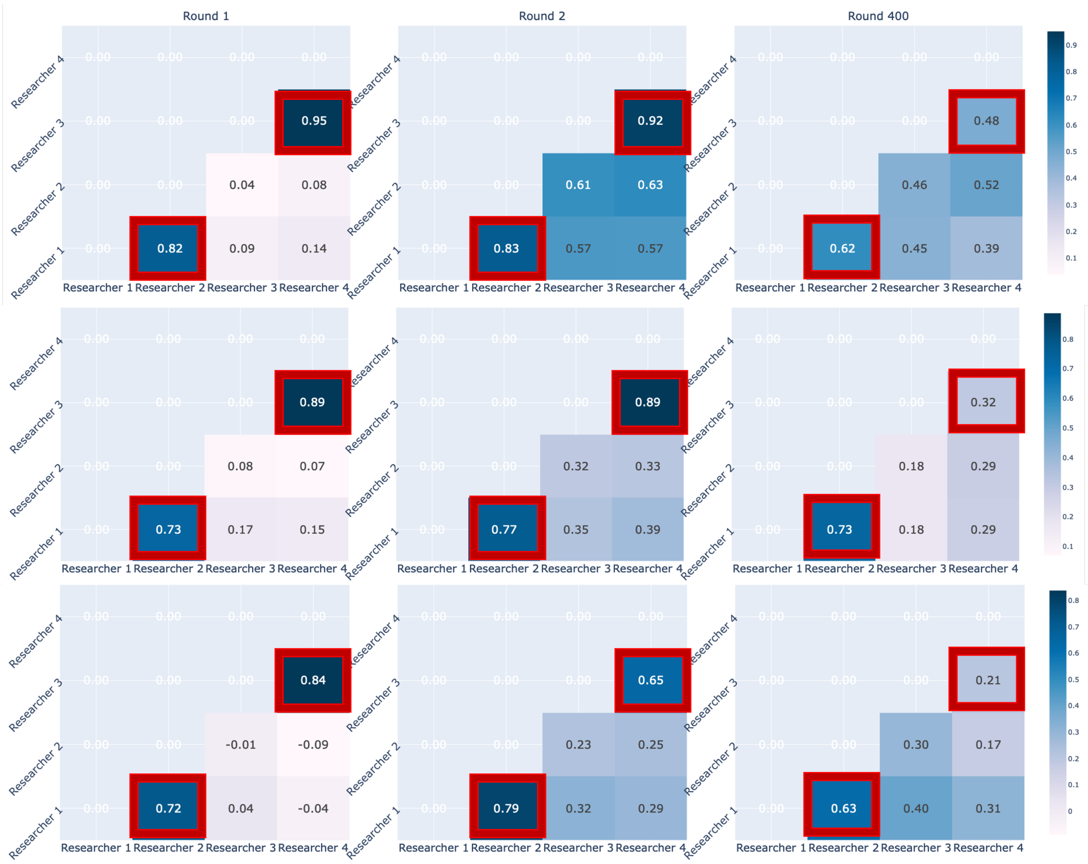


Figure 3: Contingency table of annotations for Group 1 (researchers 1 and 2) and Group 2 (researchers 3 and 4) - From top to bottom: *Appropriateness*, *Information content of outputs*, and *Humanlikeness*. The graphs for Round 1 and Round 2 show the figures for the *final* iteration of each round. Round 400 indicates the final round of annotations in **LEAP** with 400 items. Red borders indicate within-group agreement. Darker blue indicates higher agreement (Cohen's κ).

We present the detailed results in Appendix A.9. Remarkably, the iterations were important to solidify common law rules, since moving to new samples (i.e., new rounds) caused more confusion and the rules did not ground well.

4.3 Annotation Guidelines

We analyzed each group's annotation guideline and its creation process by examining the Zoom recordings of discussions. For the final version of the guidelines for all groups see Figures 9 to 14 in the Appendix.

Appropriateness The group discussion transcripts and written guidelines showed that the different groups took a similar approach when annotating *Appropriateness*. Primarily, all groups based their *Appropriateness* score on whether the model response "made sense" in relation to the prompt itself. Also, all groups considered the contextual relevance of the response in relation to the prompt. This reinforces our observation that annotators overall had a strong shared construct of *Appropriateness*, which resulted in high levels of agreement for the category.

Information content of output Unlike *Appropriateness*, agreement levels between groups for *Information content of output* were relatively low. While Group 1 gauged the category based on the specificity of the information provided by the response, Group 2 based the category score on the length of the response (ie. the number of sentences), as well as the correctness of the response (ie. if the information provided is factually correct). Such divergences in annotation guidelines explain the low level of agreement between annotators of different groups.

We conducted a similar analysis on Groups 3, 4, 5, and 6. As discussed in Appendix A.9, we observed two distinct silos of convergence in agreement. The annotation discussion transcripts revealed that Group 3 and Group 6 quantified the amount of new information stated in the response to score *Information of content*, while Group 4 and Group 5 did not. For example, if a response did not reveal any new information, but was relevant to the prompt, Group 4 and Group 5 would give at least a 3 for *Information content of output*. However, as Group 3 and Group 6 focused on the quantity of new information when annotating *Information content of output*, they would give it a low score.

Furthermore, Groups 3 and 6 solely looked at the response field to judge *Information content of output*, meaning a short, generic response would receive a low score for this category. In comparison, Groups 4 and 5 created guidelines that looked at both the prompt and response to judge the level of information given, meaning a short, generic response could still receive a higher score depending on the broader context.

The divergence in annotation guidelines not only explains the low average agreement between groups for *Information content of output* but also uncovers why different clusters of agreement occur between certain groups.

Humanlikeness While both Groups 1 and 2 based *Humanlikeness* on whether a real human would have said the response, both groups had diverging approaches for the annotation category. Group 1 emphasized that the annotator should *not* consider the appropriateness of the response when judging *Humanlikeness*. On the other hand, Group 2 simply evaluated whether a real human could have said the response, while also taking into consideration grammatical errors.

For Groups 3, 4, 5, and 6 two separate clusters of agreement occurred between the groups–one between Group 3 and Group 6, another between Group 4 and Group 5. The clusters of agreement can be attributed to the differing annotation procedures that emerged between these silos. Group 3 and Group 6 annotated by ignoring the prompt and judging solely the *Humanlikeness* of the response. On the other hand, Group 4 and Group 5 took into consideration the response's context. For example, following Group 3 and Group 6's guidelines, even if the response was a complete replica of an utterance in the prompt, the response could receive a high score for *Humanlikeness*. In contrast, if the response repeated content from the prompt, Group 4 and Group 5 gave the response a low *Humanlikeness* score.

The two different interpretations of a category reinforce the notion that a "ground truth" annotation value is difficult to reach, especially for categories that have less of a shared construct - like *Information content of output* and *Humanlikeness*.

4.4 Crowdsourced Data

In order to examine how diverging annotation guidelines impact agreement levels for crowdsourced annotations, we employed batches of Human Intelligence Tasks (HITs) on Amazon Mechanical Turk (MTurk). We recruited and filtered MTurk workers who were able to achieve a category average $\kappa > 0.7$ agreement with the researchers on a pilot HIT. These workers were then given a larger MTurk task of annotating the same set of 400 prompt-response questions from the guideline creations, with 55 prompt-response questions per HIT (for details see subsection A.7).

Agreement Between Researchers & Crowdsource Workers The average agreement between the crowdsource workers and the researcher for each Group is illustrated in Figure 15 in the Appendix. For all Groups except Group 1, *Appropriateness* was the category with the highest agreement between the researchers and the HIT workers. Overall, HIT workers who used Group 4's guideline had the highest average agreement scores with the Group's researchers. Furthermore, the variable levels of agreement for **LEAP** indicate that annotations are relatively noisy even with a well-defined protocol.

Group 1 & Group 2 We calculated the agreement between MTurk annotators of the *same* group's annotation guidelines, as well as the agreement between annotators of Groups 1 and 2 (see Table 1.

Groups	App.	Info.	Human.
Group 1	0.37	0.09	0.19
Group 2	0.58	0.20	0.30
Between Groups 1 & 2	0.37	0.13	0.09

Table 1: IAA *within* and *between* crowd workers using Group 1's and Group 2's guidelines

Of the three categories, again, *Appropriateness* had the strongest shared construct with the highest level of agreement. Group 1 and Group 2 had higher agreement *within* groups for *Humanlikeness* compared to the IAA from *between* Groups 1 and 2. As with the researcher annotators, crowd workers who followed different annotation guidelines were unable to achieve high agreement.

Groups 3, 4, 5, & 6 Similarly, we analyzed the differences in agreement levels for crowd workers using guidelines created by Groups 3, 4, 5, and 6 (see Table 2 and Table 3).

Groups	App.	Info.	Human.
Group 3	0.38	0.16	0.25
Group 4	0.46	0.54	0.56
Group 5	0.38	0.30	0.23
Group 6	0.47	0.22	0.43
Average	0.42	0.31	0.37

Table 2: IAA *within* group for crowdsource workers using guidelines created by Groups 3, 4, 5, and 6.

Groups	App.	Info.	Human.
Groups 3 & 4	0.38	0.20	0.20
Groups 3 & 5	0.32	0.23	0.18
Groups 3 & 6	0.37	0.27	0.23
Groups 4 & 5	0.32	0.17	0.22
Groups 4 & 6	0.3	0.15	0.24
Groups 5 & 6	0.57	0.27	0.27
Average	0.38	0.22	0.22

Table 3: IAA *between* groups for crowdsource workers using guidelines created by Groups 3, 4, 5, and 6.

Similar to Groups 1 and 2, crowd workers for Groups 3, 4, 5, and 6 had relatively higher agreement *within* group compared to *between* different groups.

5 Conclusion

In this paper, we caution NLP researchers against using **reliability** as a proxy for **reproducability** and **validity**. While LEAP does not strictly enforce validity, it creates transparency in the "common law" annotation rules. This transparency can enable others to assess the validity of the choices. We propose and encourage researchers to use **LEAP** as a solution to ensure **reproducibility** by rendering the annotation protocol completely transparent while allowing for deeper cross-paper analysis on **validity** through the standardized annotation procedure.

Using **LEAP**, we simulated a parallel series of independent annotation procedures, illustrating how even if a research group achieves agreement, their agreement with annotators from different groups can be low for certain categories due to diverging annotation guidelines.

Overall, research groups should use agreement metrics with care. While a high agreement score is often a community-recognized threshold required for research groups to publish their annotated datasets, research groups should be aware of the pitfalls in raising agreement metrics. Furthermore, research groups should follow a standardized annotation guideline creation process, such as **LEAP**, and make the entire procedure transparent. With such standardization and transparency, we will be able to better understand the issues associated with simply using agreement metrics as the main threshold to cross for publications.

6 Limitations

LEAP requires access to a telecommunication platform, such as Zoom, which can record, screen-share, and save live transcripts of the discussions. The dialogue data used in the annotations, as well as the annotation categories and their respective guidelines, were all in English. Furthermore, the researcher participants of the study were all co-authors of the paper and did not include professional annotators. We tested **LEAP** using only conversational dialogue.We only used three annotation categories. Though there are other protocols that could have helped in the analysis, we only experimented with LEAP and an ablation of LEAP. Some model responses may have contained bias.

Acknowledgements

We thank the reviewers for their comments and suggestions. We thank Mark Liberman for suggesting the common law metaphor and for his helpful feedback. We also thank Nicholas Lourie for his help with the project. We would like to thank Claire Daniele for her help with figures and editorial support.

References

Manex Agirrezabal, Bertol Arrieta, Aitzol Astigarraga, and Mans Hulden. 2013. POS-tag based poetry generation with WordNet. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 162–166, Sofia, Bulgaria. Association for Computational Linguistics.

Ron Artstein. 2017. Inter-annotator Agreement. In Nancy Ide and James Pustejovsky, editors, *Hand-*

book of Linguistic Annotation, pages 297–313. Springer Netherlands, Dordrecht.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Harry Bunt, Alex Chengyu Fang, Nancy M. Ide, and Jonathan J. Webster. 2010. A methodology for designing semantic annotation languages exploiting syntactic-semantic iso-morphisms. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong*, pages 29–46.

Joan Byamugisha, C. Maria Keet, and Brian DeRenzi. 2017. Evaluation of a Runyankore grammar engine for healthcare messages. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 105–113, Santiago de Compostela, Spain. Association for Computational Linguistics.

Giuseppe Carenini and Jackie C. K. Cheung. 2008. Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 33–41, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Amanda Cercas Curry, Dimitra Gkatzia, and Verena Rieser. 2015. Generating and Evaluating Landmark-Based Navigation Instructions in Virtual Environments. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 90–94, Brighton, UK. Association for Computational Linguistics.

Hyungtak Choi, Siddarth K.M., Haehun Yang, Heesik Jeon, Inchul Hwang, and Jihie Kim. 2018. Self-Learning Architecture for Natural Language Generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 165–170, Tilburg University, The Netherlands. Association for Computational Linguistics.

Philipp Cimiano, Janna Lüker, David Nagel, and Christina Unger. 2013. Exploiting ontology lexica for generating natural language texts from RDF data. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 10–19, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Kelly Cuccolo, Jon E Grahe, Martha S Zlokovich, John Edlund, rick miller, Susana Gallor, Jordan R Wagge, Kaitlyn M Werner, Albert L Ly, Fanli Jia, and et al. 2022. NICE: CROWD.

Kelly Cuccolo, Megan S. Irgens, Martha S. Zlokovich, Jon Grahe, and John E. Edlund. 2021. What Crowdsourcing Can Offer to Cross-Cultural Psychological Science. *Cross-Cultural Research*, 55(1):3–28. Publisher: SAGE Publications Inc.

Seniz Demir, Sandra Carberry, and Kathleen F. McCoy. 2008. Generating textual summaries of bar charts. In *Proceedings of the Fifth International Natural Language Generation Conference on - INLG '08*, page 7, Salt Fork, Ohio. Association for Computational Linguistics.

Jan Milan Deriu and Mark Cieliebak. 2018. Syntactic Manipulation for Generating more Diverse and Interesting Texts. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 22–34, Tilburg University, The Netherlands. Association for Computational Linguistics.

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

Abdurrisyad Fikri, Hiroya Takamura, and Manabu Okumura. 2018. Stylistically User-Specific Generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 89–98, Tilburg University, The Netherlands. Association for Computational Linguistics.

Katja Filippova and Michael Strube. 2008. Dependency tree based sentence compression. In *Proceedings of the Fifth International Natural Language Generation Conference on - INLG '08*, page 25, Salt Fork, Ohio. Association for Computational Linguistics.

Sarah E. Finch and Jinho D. Choi. 2020. Towards Unified Dialogue System Evaluation: A Comprehensive Analysis of Current Evaluation Protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Dimitra Gkatzia, Helen Hastie, Srinivasan Janarthanam, and Oliver Lemon. 2013. Generating student feedback from time-series data using reinforcement learning. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 115–124, Sofia, Bulgaria. Association for Computational Linguistics.

Jon E. Grahe, Kelly Cuccolo, Dana C. Leighton, and Leslie D. Cramblet Alvarez. 2020. Open Science Promotes Diverse, Just, and Sustainable Research and Educational Outcomes. *Psychology Learning & Teaching*, 19(1):5–20. _eprint: https://doi.org/10.1177/1475725719869164.

Byron B. Hamilton, Judith A. Laughlin, Roger C. Fiedler, and Carl V. Granger. 1994. Interrater reliability of the 7-level functional independence measure (FIM). *Scandinavian journal of rehabilitation medicine*, 26(3):115–119.

Vrindavan Harrison and Marilyn Walker. 2018. Neural Generation of Diverse Questions using Answer Focus, Contextual and Linguistic Features. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 296–306, Tilburg University, The Netherlands. Association for Computational Linguistics.

Nicole Hober, Tülay Dixon, and Tove Larsson. 2023. Toward increased reliability and transparency in projects with manual linguistic coding. *Corpora*, 18.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Eduard Hovy and Julia Lavid. 2010. Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International journal of translation*, 22(1):13–36.

David M. Howcroft, Anya Belz, Miruna 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.

Stephanie Inglis. 2015. Summarising Unreliable Data. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 95–99, Brighton, UK. Association for Computational Linguistics.

Stephanie Inglis, Ehud Reiter, and Somayajulu Sripada. 2017. Textually Summarising Incomplete Data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 228–232, Santiago de Compostela, Spain. Association for Computational Linguistics.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel Weld. 2022. GENIE: Toward reproducible and standardized human evaluation for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11444–11458, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-Augmented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Kittipitch Kuptavanich, Ehud Reiter, Kees Van Deemter, and Advaith Siddharthan. 2018. Generating Summaries of Sets of Consumer Products: Learning from Experiments. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 403–407, Tilburg University, The Netherlands. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174. Publisher: [Wiley, International Biometric Society].

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140:55–55.

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21, Nancy, France. Association for Computational Linguistics.

Saad Mahamood and Ehud Reiter. 2012. Working with clinicians to improve a patient-information NLG system. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 100–104, Utica, IL. Association for Computational Linguistics.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Priscilla Moraes, Kathy McCoy, and Sandra Carberry. 2014. Adapting Graph Summaries to the Users' Reading Levels. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 64–73, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, and Tatsuya Harada. 2019. Toward a Better Story End: Collecting Human Evaluation with Reasons. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 383–390, Tokyo, Japan. Association for Computational Linguistics.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Alice Oh and Howard Shrobe. 2008. Generating baseball summaries from multiple perspectives by reordering content. In *Proceedings of the Fifth International Natural Language Generation Conference on - INLG '08*, page 173, Salt Fork, Ohio. Association for Computational Linguistics.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. Learning from Multi-Annotated Corpora. In *Statistical Methods for Annotation Analysis*, pages 147–165. Springer International Publishing, Cham.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. Place: Cambridge, MA Publisher: MIT Press.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

Raheel Qader, Khoder Jneid, François Portet, and Cyril Labbé. 2018. Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 254–263, Tilburg University, The Netherlands. Association for Computational Linguistics.

Ehud Reiter, Albert Gatt, François Portet, and Marian van der Meulen. 2008. The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In *Proceedings of the Fifth International Natural Language Generation Conference on - INLG '08*, page 147, Salt Fork, Ohio. Association for Computational Linguistics.

Ehud Reiter, Roma Robertson, and Liesl Osman. 2000. Knowledge acquisition for natural language generation. In *Proceedings of the first international conference on Natural language generation - INLG '00*, volume 14, page 217, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Sashank Santhanam and Samira Shaikh. 2019. Towards Best Experiment Design for Evaluating Dialogue System Output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Björn Schlünder and Ralf Klabunde. 2013. Greetings generation in video role playing games. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 167–171, Sofia, Bulgaria. Association for Computational Linguistics.

Martin Schweinsberg, Nikhil Madan, Michelangelo Vianello, S. Amy Sommer, Jennifer Jordan, Warren Tierney, Eli Awtrey, Luke Lei Zhu, Daniel Diermeier, Justin E. Heinze, Malavika Srinivasan, David Tannenbaum, Eliza Bivolaru, Jason Dana, Clintin P. Davis-Stober, Christilene du Plessis, Quentin F. Gronau, Andrew C. Hafenbrack, Eko Yi Liao, Alexander Ly, Maarten Marsman, Toshio Murase, Israr Qureshi, Michael Schaerer, Nico Thornley, Christina M. Tworek, Eric-Jan Wagenmakers, Lynn Wong, Tabitha Anderson, Christopher W. Bauman, Wendy L. Bedwell, Victoria Brescoll, Andrew Canavan, Jesse J. Chandler, Erik Cheries, Sapna Cheryan, Felix Cheung, Andrei Cimpian, Mark A. Clark, Diana Cordon, Fiery Cushman, Peter H. Ditto, Thomas Donahue, Sarah E. Frick, Monica Gamez-Djokic, Rebecca Hofstein Grady, Jesse Graham, Jun Gu, Adam Hahn, Brittany E. Hanson, Nicole J. Hartwich, Kristie Hein, Yoel Inbar, Lily Jiang, Tehlyr Kellogg, Deanna M. Kennedy, Nicole Legate, Timo P. Luoma, Heidi Maibuecher, Peter Meindl, Jennifer Miles, Alexandra Mislin, Daniel C. Molden, Matt Motyl, George Newman, Hoai Huong Ngo, Harvey Packham, Philip S. Ramsay, Jennifer L. Ray, Aaron M. Sackett, Anne-Laure Sellier, Tatiana Sokolova, Walter Sowden, Daniel Storage, Xiaomin Sun, Jay J. Van Bavel, Anthony N. Washburn, Cong Wei, Erik Wetter, Carlos T. Wilson, Sophie-Charlotte Darroux, and Eric Luis Uhlmann. 2016. The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, 66:55–67.

João Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Advaith Siddharthan, Matthew Green, Kees van Deemter, Chris Mellish, and René van der Wal. 2012. Blogging birds: Generating narratives about reintroduced species to promote public engagement. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 120–124, Utica, IL. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from Disagreement: A Survey. *J. Artif. Int. Res.*, 72:1385–1470. Place: El Segundo, CA, USA Publisher: AI Access Foundation.

Kees van Deemter. 2000. Generating vague descriptions. In *Proceedings of the first international conference on Natural language generation - INLG '00*, volume 14, page 179, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Sebastian Varges. 2006. Overgeneration and ranking for spoken dialogue systems. In *Proceedings of the Fourth International Natural Language Generation Conference on - INLG '06*, page 20, Sydney, Australia. Association for Computational Linguistics.

Jason Weston and Kurt Shuster. 2021. Blender Bot 2.0: An open source chatbot that builds long-term memory and searches the internet.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing Precise and Robust Dialogue Response Evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

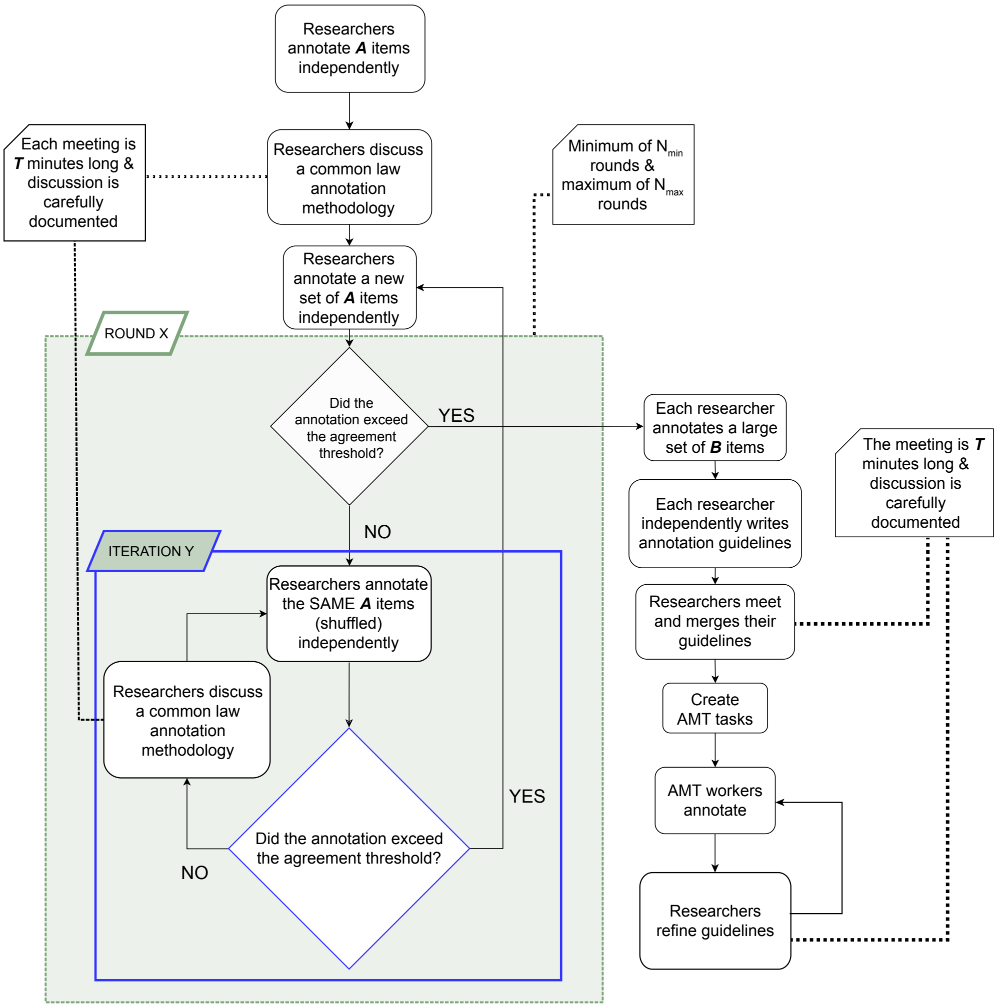


Figure 4: A flowchart of **LEAP** - a standardized and transparent annotation protocol. It has the following parameters: the minimum and maximum number of rounds (N_{min} and N_{max}), agreement criteria and threshold, common law discussion time limit (T), number of researchers involved (minimum of 2), the number of items per small iteration (**A**), the number of items for the larger annotation (**B**).

A Appendix

A.1 Generic LEAP

As stated in subsection 3.1 has task-specific properties that need to be chosen by researchers. We used LEAP with a particular set of parameters for our experiment. Figure 4 is a generic flowchart for LEAP. We acknowledge that LEAP is extremely rigid. For our experiments, this was important as we desired to minimize variance between groups. In practice, researchers may desire more flexibility (e.g., discussion time cap or the number of examples per iteration). If this is the case then we encourage the documentation of deviations from the protocol.

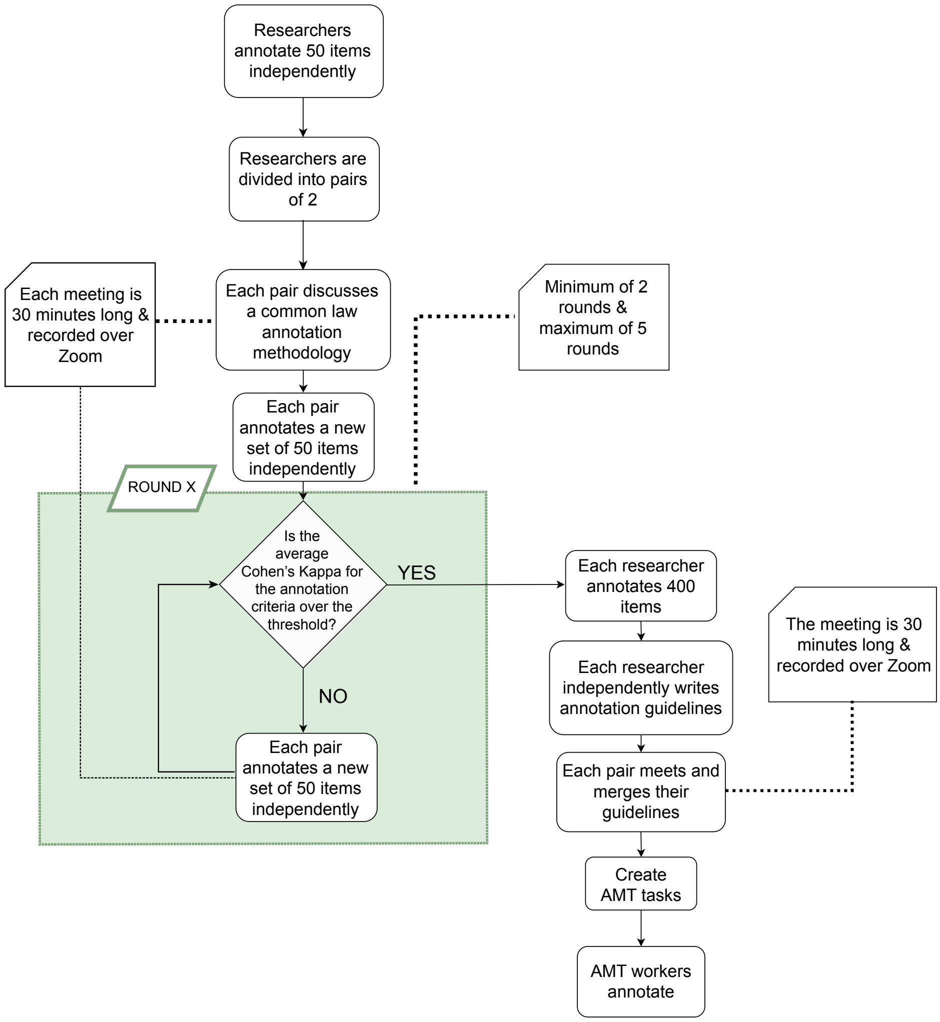


Figure 5: A flowchart of iteration-free **LEAP**.

A.2 Iteration-Free LEAP

Figure 5 shows iteration-free LEAP. Iteration-free LEAP might be more attractive for researchers who are concerned about the independent nature of annotations of each round. We suggest that iterations within a round are important since researchers may not have explicitly agreed on the common law rule.

A.3 NLP Work Using *Appropriateness*, *Information content of output*, and *Humanlikeness*

Various papers created their own definitions for the category of *Appropriateness* (Varges, 2006; Reiter et al., 2008; Oh and Shrobe, 2008; Murray et al., 2010; Mahamood and Reiter, 2011; Schlünder and Klabunde, 2013; Gkatzia et al., 2013; Cimiano et al., 2013; Inglis et al., 2017; Harrison and Walker, 2018; Mori et al., 2019; Santhanam and Shaikh, 2019), *Information content of outputs* (Demir et al., 2008; Siddharthan et al., 2012; Mahamood and Reiter, 2012; Moraes et al., 2014; Inglis, 2015; Kuptavanich et al., 2018; Qader et al., 2018; Choi et al., 2018), and for *Humanlikeness* (Byamugisha et al., 2017; Deriu and Cieliebak, 2018; Fikri et al., 2018).

A.4 Dialog Model Parameters

For DialoGPT, which was trained on 147M dialogue instances created from Reddit threads (Zhang et al., 2020), we used the pre-trained model with the medium (345M) model checkpoint, using the top-K sorting algorithm. For GPT-3 (Brown et al., 2020), we used a temperature of 0.9 and a top-p decoding strategy (Holtzman et al., 2019) with $p = 0.92$. We used the following format for the prompt for GPT-3:

> The following is a conversation between A and B.
>
> A: Oh, I am so tired.
> B: I know what you mean.
> A: I don't know if I can continue working like this.
> B:

For Plato2, we used two model sizes, 24L (with 310M parameters), and 32L (with 1.6B parameters) (Bao et al., 2021). For BlenderBot, two model sizes were used: 2.7B and 9B (Miller et al., 2017). For BlenderBot 2, two model sizes were used as well: 400M and 3B (Weston and Shuster, 2021; Komeili et al., 2022; Xu et al., 2022). Finally, we used the original human responses that are a part of the ESL (Sedoc et al., 2019) and Daily Dialog (Li et al., 2017) evaluation sets.

prompt	response
A: Oh, I am so tired. B: I know what you mean. A: I don't know if I can continue working like this.	Why don't you take a break?

Figure 6: An example prompt and response annotated by the researchers and crowdsource workers.

A.5 Instructions

Common Law Annotations

Creating Annotation Guidelines

The goal is to create guidelines that help people annotate conversational text data as similarly as possible. In order to increase agreement with your annotation partner, you will meet with them to discuss a common annotation methodology.

The annotation task is to label **chatbot responses** to **prompts**, using three annotation criteria:

- **Appropriateness**: The degree to which the output is appropriate in the given context/situation.
- **Information content of outputs**: The amount of information conveyed by an output.
- **Human-likeness**: The degree to which an output could have been produced by a human.

Each criteria is annotated on a **5-point scale** where 1 is worst and 5 is best**.**

Annotating Model Responses

For each round of annotations, you will be provided with a Google Sheets document containing **50** prompts and responses. During these annotations, you ***may not communicate with your partner annotator.***

The **prompt column** contains a single utterance *or* multiple-utterance conversation. The **response** column contains the chatbot's response to the last utterance in the **prompt**.

Utterances are separated by **A:** and **B:**, which indicate two speakers. There are *at most* **two speakers** per prompt, though there *may* be prompts with **only one speaker**.

For example,

prompt	response	Appropriateness	Information content of outputs	Humanlikeness
A: Oh, I am so tired. B: I know what you mean. A: I don't know if I can continue working like this.	Why don't you take a break?	enter annotation here	enter annotation here	enter annotation here

Remember that the annotation values should be a number between **1** and **5**. You will annotate **50 prompt-response pairs** each round. Please time yourself at the start and end of each annotation session.

Figure 7: The annotation and discussion instructions shared to all groups.

Annotation & Discussion Plan [Annotator]

Objective: Annotators repeat annotation and discussion in order to increase their inter-annotator agreement.

Step	Title	Time Needed (Approx.)	Instructions	Notes
1	**Discuss initial annotation methodology**	30 min.	Schedule a common time using the Doodle poll Join this public Zoom call on your scheduled time and discuss annotation methodologies	
2	**1st Annotation Session**	-	Annotate 50 model responses	
3	**Discuss annotation methodology**	30 min.	Schedule a common time using the Doodle poll Join this public Zoom call on your scheduled time and discuss annotation methodologies	
4	**2nd Annotation Session**	-	Annotate 50 model responses	
5	**Discuss annotation methodology**	30 min.	Schedule a common time using the Doodle poll Join your designated Zoom call on your scheduled time and discuss annotation methodologies	
If Inter-Annotator Agreement is below 0.7: proceed to STEP 6 and STEP 7 **If Inter-Annotator Agreement is above 0.7: proceed to STEP 8**				
6	**Annotation Session**	-	Annotate 50 model responses	
7	**Discuss annotation methodology**	30 min.	Schedule a common time using the Doodle poll Join your designated Zoom call on your scheduled time and discuss annotation methodologies	
If Inter-Annotator Agreement is below 0.7: *repeat* STEP 6 and STEP 7 **If Inter-Annotator Agreement is above 0.7: proceed to STEP 8** **→ Maximum of 5 annotation- discussion repetitions**				
8	**Annotate 400 responses**	-	Annotate 400 model responses	
9	**Create Individual Annotation Guideline**	-	Each annotator creates their own annotation guideline	
10	**Merge Annotation Guideline**	-	The annotator pair merges their annotation guideline	
11	**150 AMT Annotations**	-	Annotate 150 items through the Amazon Mechanical Turk (AMT) platform - 50 using your own guideline, 50 using a different group's guideline, and 50 using another group's guideline	

Figure 8: The step-by-step LEAP instructions shared among researcher annotators.

A.6 Annotation Guidelines

Instructions

NOTE: THERE ARE A MAXIMUM OF 5 HITs YOU CAN COMPLETE. COMPLETING ALL 5 HITs WILL GIVE YOU A BONUS! WE ENCOURAGE YOU TO DO ALL 5 HITs

The annotation task is to label responses to a given prompt. The prompt consists of two people (A and B) talking to each other. The response is the next utterance after the final utterance in the prompt. The three base annotation criteria are:

1. **Appropriateness**: The degree to which the output is appropriate in the given context/situation.
2. **Information content of outputs**: The amount of information conveyed by an output.
3. **Human-likeness**: The degree to which an output could have been produced by a human.

Each criteria is annotated on a 5-point scale where **1** is worst and **5** is best.

Specific Definitions

Tips:

- Do **not** consider the humanlikeness of the response when judging its appropriateness. If the response fits the context, but sounds weird, mark it as highly appropriate.
- Do **not** consider the appropriateness of the response when judging its humanlikeness. Humanlikeness should depend only on the response itself, ignoring the context.

#	Appropriateness	Information content of output	Humanlikeness
	Does the response fit the context well?	Is the response specific or generic?	Could a human have said this response?
1	The response is completely irrelevant. It does not match or reference the context at all.	The response could follow almost any statement or context, for example “How so?” or “Yeah”.	The response is garbled text or nonsensical. It has no reasonable interpretation. For example, “He green a”.
2	The response matches or references the context, but is highly unexpected. For example, it might contradict a previous statement, repeat word-for-word something already said, or respond to an earlier part of the conversation.	The response could follow a large category of statements, but not any statement, for example, “Why not?” or “I’m down for that.”	The response has some well-formed parts that you can interpret, but it’s completely unbelievable that the response came from a human. For example, “It’s just north of here UUUU.”
3	The response has at least one interpretation that makes sense in the conversation; however, the interpretation is a stretch or strange given the context.	The response talks about something generic in a vague way, for example: “Yeah, I love that.” or “It really is great.”	The response is well-formed, but doesn’t seem realistic, sounds like a poorly written fictional person, or over-explains. For example, “I am sleepy because it’s night time, so now I will nap in my bed.”
4	The first interpretation of the response makes sense in the conversation; however, it’s a bit strange, awkward, or unexpected.	The response talks about something specific to the conversation in a vague way, for example: “Park street is great.”	The response is mostly believable as coming from a person, but seems a bit strange. For example, the response is unreasonably polite.
5	The response perfectly fits the conversation.	The response is highly specific to the conversation. It discusses something from the conversation in a specific way, for example “I really miss them, now that they’re gone.”	The response is perfectly believable as coming from a human. You would not be surprised at all if a person had actually said it.

Examples:

Prompt	Response	Appropriateness	Information content of output	Humanlikeness
A: This floor is falling apart. B: How can you tell? A: Are you kidding me? Step over here.	That's what I was thinking.	1	2	5
A: The prices on fruit have really gone up this year! B: Yes, they have. It's ridiculous! A: The fruit is not of good quality, either.	I've also noticed the drop in quality.	5	5	5
A: I'm making the food for the party tomorrow. B: I heard you are a fantastic cook! A: Thank you! I do my best.	I'm sure you do !	4	2	4

Figure 9: Annotation Guideline for Group 1.

Instructions

NOTE: THERE ARE A MAXIMUM OF 5 HITs YOU CAN COMPLETE. COMPLETING ALL 5 HITs WILL GIVE YOU A BONUS! WE ENCOURAGE YOU TO DO ALL 5 HITs

The annotation task is to label responses to a given prompt. The prompt consists of two people (A and B) talking to each other. The response is the next utterance after the final utterance in the prompt. The three base annotation criteria are:

1. **Appropriateness**: The degree to which the output is appropriate in the given context/situation.
2. **Information content of outputs**: The amount of information conveyed by an output.
3. **Human-likeness**: The degree to which an output could have been produced by a human.

Each criteria is annotated on a 5-point scale where **1** is worst and **5** is best.

Specific Definitions

Appropriateness

Given the context in the prompt, we will consider the following aspects when assigning the score for appropriateness:

1. Answer the question
2. Talk about the same thing in the prompt
3. The transition is smooth

If the response satisfies all the above requirements, we will assign a score of 5. If the response somehow answers the question but does not satisfy one of the other two requirements, we will assign a score of 4. If the response only answers the question partially, we will assign a score of 3. If the response does not answer the question but satisfies one of the other two requirements, we will assign a score of 2. If the response does not satisfy all the above requirements, we will assign a score of 1.

Information content of outputs

For this part, we will take the information conveyed in the response and the length of the response into consideration. To be detailed, we will consider the following aspects when assigning the score for information content of outputs:

1. The information covered for the question
2. The number of sentence (>= 4 long; >= 3 median; >= 1 short)
3. Information is valid (even if it is not related to the prompt)

If the response satisfies all the above requirements (long sentences), we will assign a score of 5. If the response contains enough and valid information but does not have a reasonable length (median), we will assign a score of 4. If the response contains some information, but the information may not be valid or the response does not have a reasonable length (median), we will assign a score of 3. If the response contains limited information and the length of the response is short, we will assign a score of 2. If the response does not satisfy all the above requirements, we will assign a score of 1.

Humanlikeness

We will evaluate the degree to which the response looks like a human sentence. We will consider the following aspects when assigning the score for human-likeness:

1. First impression of reading as a human sentence
2. Check grammar and syntax error

If the response satisfies all the above requirements, we will assign a score of 5. If the response contains minor grammar or syntax errors but overall looks like a human sentence, we will assign a score of 4. If the response contains a few grammar or syntax errors but still looks like a human sentence, we will assign a score of 3. If the response somehow does not look like a human sentence slightly but there are few grammar or syntax errors in the response, we will assign a score of 2. If the response does not satisfy all the above requirements, we will assign a score of 1.

Figure 10: Annotation Guideline for Group 2.

Instructions

NOTE: THERE ARE A MAXIMUM OF 5 HITs YOU CAN COMPLETE. COMPLETING ALL 5 HITs WILL GIVE YOU A BONUS! WE ENCOURAGE YOU TO DO ALL 5 HITs

The annotation task is to label responses to a given prompt. The prompt consists of two people (A and B) talking to each other. The response is the next utterance after the final utterance in the prompt. The three base annotation criteria are:

1. **Appropriateness**: The degree to which the output is appropriate in the given context/situation.
2. **Information content of outputs**: The amount of information conveyed by an output.
3. **Human-likeness**: The degree to which an output could have been produced by a human.

Each criteria is annotated on a 5-point scale where **1** is worst and **5** is best.

Specific Definitions

Laurel: Ben let out the cats this morning but one of them didn't come back into the house
Dara: oh no, was it Tom?

Appropriateness:

1. Completely irrelevant and non-topical response
   Laurel: Did you get dinner?
2. Response is on topic, but not appropriate
   Laurel: Cats are cute
3. Half of response is appropriate, half is not
   Laurel: Thankfully, it wasn't Tom, but I want some hotdogs
4. Response is mostly appropriate, albeit slightly awkward
   Laurel: Jerry
5. Response is
   Laurel: No, it was Jerry ... He's always crawling under the house.

Information content of output

1. Repetition of something said in the prompt
   Laurel: The cats were let out by Ben
2. A generic answer or a question
   Laurel: Maybe
   Laurel: Do you like cats?
3. One type of information conveyed (information about self or the world)
   Laurel: No, it was Jerry.
   Laurel: Cats often run away.
4. Both types of information conveyed (information about self and the world)
   Laurel: No Jerry:(I am really worried.
5. 3 or more distinct pieces of information conveyed.
   Laurel: No it is Jerry. I'm really worried. 90% of cats that aren't found within 5 hours are roadkill.

Humanlikeness *The extent to which the response BY ITSELF (ignore context) could have been written by a person*

1. Self-contradictory or ungrammatical
   Laurel: Yes it's Jerry. But Jerry is Tom.
2. Incorrect obvious facts about the world
   Laurel: Jerry is mostly a cat name as in Tom and Jerry
3. Too long or unnatural sounding (lack of conversationalist properties)
   Laurel: Jerry got lost then he might have gone under the house just like I drew in my drawing. Maybe this was all a dream
4. Close but not right almost like a non-native English speaker
   Laurel: Jerry ran away ... Sad feelings.
5. Perfectly fluent. You could imagine yourself saying this.
   Laurel: Unfortunately both Jerry and Tom that ran away. I'm not sure what to do.

Figure 11: Annotation Guideline for Group 3.

Instructions

NOTE: THERE ARE A MAXIMUM OF 5 HITs YOU CAN COMPLETE. COMPLETING ALL 5 HITs WILL GIVE YOU A BONUS! WE ENCOURAGE YOU TO DO ALL 5 HITs

The annotation task is to label responses to a given prompt. The prompt consists of two people (A and B) talking to each other. The response is the next utterance after the final utterance in the prompt. The three base annotation criteria are:

1. **Appropriateness**: The degree to which the output is appropriate in the given context/situation.
2. **Information content of outputs**: The amount of information conveyed by an output.
3. **Human-likeness**: The degree to which an output could have been produced by a human.

Each criteria is annotated on a 5-point scale where **1** is worst and **5** is best.

Specific Definitions

Information content of output, Appropriateness:

Appropriateness: 1
Information content of output: 1
If the response doesn't make sense, doesn't relate to the previous conversation, doesn't have some new information

Appropriateness: 1
Information content of output: 2
If the response doesn't make sense, doesn't relate to the previous conversation, has some new information

Appropriateness: 2
Information content of output: 1
If the response doesn't make sense, but still relates to the previous conversation, doesn't have new information

Appropriateness: 4
Information content of output: 3
If the response does make sense, but still relate to the previous conversation, doesn't have new information

Appropriateness: 5
Information content of output: 5
If the response does make sense, but still relate to the previous conversation, has some new information

Appropriateness: 4
Information content of output: 3
If the response does make sense, not quite appropriate, has new information

Appropriateness: 4
Information content of output: 4
If the response does make sense, not as appropriate, has new information

Humanlikeness

Humanlikeness: 1
If the response doesn't make sense, doesn't relate to the previous conversation, repeat previous information,

Humanlikeness: 1
If the response doesn't make sense, doesn't relate to the previous conversation, doesn't repeat previous information,

Humanlikeness: 2
If the response doesn't make sense, but still relates to the previous conversation, repeats previous information,

Humanlikeness: 2
If the response doesn't make sense, but still relates to the previous conversation, doesn't repeat previous information,

Humanlikeness: 1
If the response does make sense, doesn't relate to the previous conversation, repeats previous information,

Humanlikeness: 1
If the response does make sense, doesn't relate to the previous conversation, doesn't repeat previous information,

Humanlikeness: 3
If the response does make sense, but still relates to the previous conversation, repeats previous information,

Humanlikeness: 4
If the response does make sense, relates to the previous conversation, and paraphrases previous information

Humanlikeness: 5
If the response does make sense, still relates to the previous conversation, doesn't repeat previous information,

Figure 12: Annotation Guideline for Group 4.

Instructions

NOTE: THERE ARE A MAXIMUM OF 5 HITs YOU CAN COMPLETE. COMPLETING ALL 5 HITs WILL GIVE YOU A BONUS! WE ENCOURAGE YOU TO DO ALL 5 HITs The annotation task is to label responses to a given prompt. The prompt consists of two people (A and B) talking to each other. The response is the next utterance after the final utterance in the prompt. The three base annotation criteria are:

1. **Appropriateness**: The degree to which the output is appropriate in the given context/situation.
2. **Information content of outputs**: The amount of information conveyed by an output.
3. **Human-likeness**: The degree to which an output could have been produced by a human.

Each criteria is annotated on a 5-point scale where **1** is worst and **5** is best.

Specific Definitions

Appropriateness:

Lower score:

- Confusing response, off-topic
- Offensive, aggressive
- Condtradiction

Higher score:

- Empathetic, compassionate responses
- Apt responses, matching emotional toll of the situation

Information Content:

Lower score:

- Contradiction
- Off-topic
- Repetition
- Standalone response doesn't make sense

Higher score:

- Reasoning, could be indicated by joint statements/multiple clauses

Humanlikeness

Lower score:

- Extensive repetition
- Contradiction
- Off-topic
- Generic responses ("I am sorry to hear that", "How can i help you?"

Higher score:

- Appropriate emojis
- First person pronouns ("I", "We")
- Referring to familial relationships
- Colloquial language ("wanna", etc.)
- Contractions ("I'm", "aren't", etc.)
- Discuss of emotions ("I feel" statements for example)
- Expression of surprise ("oh!", etc.)

Figure 13: Annotation Guideline for Group 5.

Instructions

NOTE: THERE ARE A MAXIMUM OF 5 HITs YOU CAN COMPLETE. COMPLETING ALL 5 HITs WILL GIVE YOU A BONUS! WE ENCOURAGE YOU TO DO ALL 5 HITs

The annotation task is to label responses to a given prompt. The prompt consists of two people (A and B) talking to each other. The response is the next utterance after the final utterance in the prompt. The three base annotation criteria are:

1. **Appropriateness**: The degree to which the output is appropriate in the given context/situation.
2. **Information content of outputs**: The amount of information conveyed by an output.
3. **Human-likeness**: The degree to which an output could have been produced by a human.

Each criteria is annotated on a 5-point scale where **1** is worst and **5** is best.

Specific Definitions

Appropriateness:

1. "this had nothing to do with the conversation whatsoever"
2. "huh, wait, that's very weird"
   - Response has at least a mild relevance to the topic discussed in prompt but otherwise is a very unusual sentence
3. "that's absurd but lets move on"
   - An unusual sentence given topic-prompt or sentence construction but people would just ignore it if someone said that - because its not THAT weird
4. "Hm. I guess that makes sense"
   - Follows logically from the prompt; typical response given the context in the prompt. Sort of what was expected.
5. "That helped the conversation"
   - This adds new and relevant information to the conversation. This is going above and beyond in the right direction for this conversation.

Information Content:

1. Hard to infer anything about the conversation. Very generic.
   - Example: "okay".
2. I know one thing that this conversations is about
   - Example: "Yeah I've been there"
   - You know they are talking about a place which counts as 1 thing
3. I know 2 things. Longer sentence.
   - Example: "Yeah I've been there. I thought it was quite nice actually"
4. Multiple sentences with with 3 or more things.
   - Example:"Yeah I've been there. I thought it was quite nice. My mom liked it too."
5. A long informative sentence (~>10 words) with LOTS to add in terms of specific.
   - Example: "The thing I like about the Taj Mahal is that it is all one block or marble [...]"

Human-likeness:

1. Impossible for a human to say in any context
   - Gibberish: "trhaiotjhoiath ^^ blah_cat"
2. Correct words but very wrong
   - grammar or word order "Go blue stuff very, I said?"
3. some beginner ESL person could say this, i guess
   - ESL: "Very nice that thing is"
4. Someone would only say this in a weirdly specific situation
   - "It would have been nice if they had dunked it"
   - Technical language: "compile down your source code into binary bits via the JVM"
5. literally me or my circle would legitimately say this "Yeah i get you"

Figure 14: Annotation Guideline for Group 6.

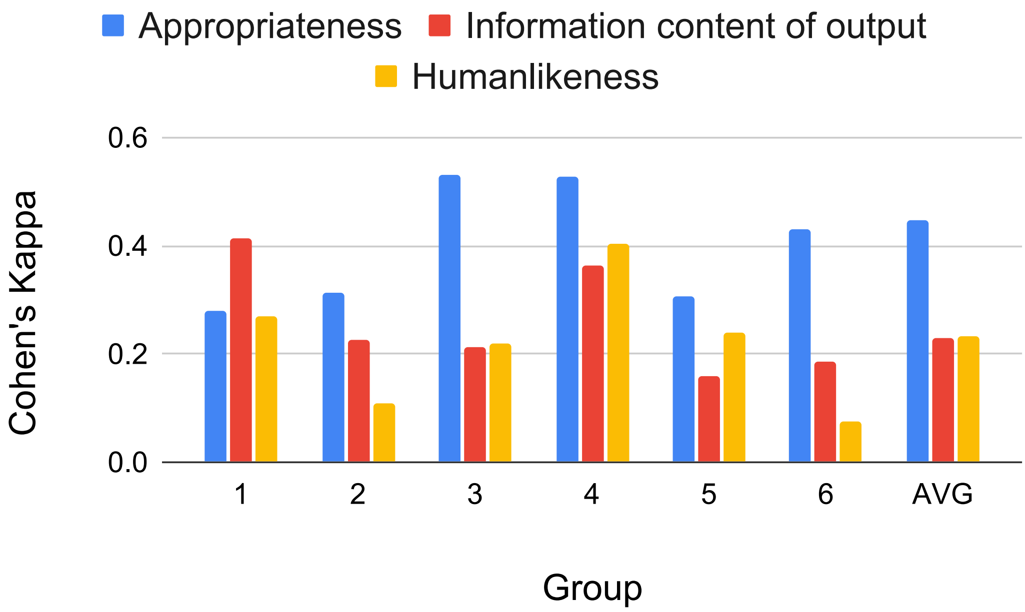


Figure 15: Average agreement between Researchers and Amazon Mechanical Turk Workers, using each Group's guidelines.

To verify that you've read the instructions, please **read the following prompt** and **write an example Response** that satisfies the following: *Appropriateness*: 5, *Information content of Output*: 2, *Humanlikeness*: 5

A: Oh, I am so tired.
B: I know what you mean.
A: I don't know if I can continue working like this.

Figure 16: An example attention check question asked to crowdsource workers.

A.7 Crowdsourced Annotations

Metadata Using guidelines created by Groups 1 and 2, which were created using **LEAP**, we deployed an initial screening round of annotations to distinguish the workers who were able to have high agreement with the researchers of the respective groups. Each screening round consisted of 1 HIT task and 10 unique workers completed the HITs. Workers who were able to achieve a category average $\kappa > 0.7$ agreement with the researchers were noted as quality workers. The qualified workers were then given a larger MTurk task of 400 prompt-response questions, where each HIT asked 55 prompt-response questions.

Three workers qualified for Group 1 and four workers qualified for Group 2. A total of 24 HITs were created for the three workers using Group 1's guidelines and a total of 32 HITs were published for the four workers using Group 2's guidelines. The workers for Group 1 completed a total of 23 HITs and the workers for Group 2 completed a total of 19 HITs.

The workers were notified the annotations will be used for research purposes.

Compensation We conducted an initial pilot run of a HIT and learned the workers took an average of 25 minutes to complete a HIT of 55 items. We paid each worker \$6.25 per HIT.

Qualifications We required a minimum of 500 approved tasks on MTurk. Second, the workers were chosen from a group of workers whose quality was verified for other text-generation evaluation tasks (e.g., summarization evaluation).

Quality Checks In order to ensure the quality of the crowdsource data, we implemented several different quality and attention checks. For each HIT, we asked two quality-check questions to confirm that the worker read and understood the annotation guidelines (Figure 16). We asked an attention-check question to ensure the worker was not randomly participating in the HIT without reading the prompt and responses. Finally, we excluded all workers who did not pass the attention checks or had a category average $\kappa < 0.1$.

A.8 Average Annotation Ratings per Conversational Model

Model	App.	Info.	Human.
BlenderBot 2 - 3b	3.58	3.12	4.78
BlenderBot 2 - 400m	3.10	4.10	4.32
BlenderBot - 3b	2.38	4.75	3.30
BlenderBot - 9b	3.62	4.05	4.25
DialoGPT	3.19	4.13	3.86
GPT-3	4.42	3.90	4.63
Ground truth	4.08	4.48	4.50
Plato 2	3.16	4.07	3.57
Plato 2 - 24L	4.30	4.70	3.80
Plato 2 - 32L	4.40	5.00	4.60

Table 4: Average annotation ratings per conversational model for Group 1.

Model	App.	Info.	Human.
BlenderBot 2 - 3b	3.63	2.53	4.30
BlenderBot 2 - 400m	3.24	3.52	4.58
BlenderBot - 3b	2.84	3.87	4.21
BlenderBot - 9b	3.54	3.46	4.25
DialoGPT	3.54	3.31	4.41
GPT-3	3.95	3.33	4.53
Ground truth	3.89	3.23	4.87
Plato 2	3.47	3.61	4.27
Plato 2 - 24L	3.85	4.30	4.50
Plato 2 - 32L	4.40	4.50	4.80

Table 5: Average annotation ratings per conversational model for Group 2.

Model	App.	Info.	Human.
BlenderBot 2 - 3b	2.74	2.88	4.61
BlenderBot 2 - 400m	2.45	2.94	4.64
BlenderBot - 3b	3.22	3.70	4.71
BlenderBot - 9b	3.21	3.49	4.97
DialoGPT	3.35	2.77	4.60
GPT-3	4.75	2.77	4.86
Ground truth	4.31	3.01	4.95
Plato 2	3.64	3.30	4.28
Plato 2 - 24L	3.02	3.42	3.68
Plato 2 - 32L	3.61	3.40	4.25

Table 6: Average annotation ratings per conversational model for Group 3.

Model	App.	Info.	Human.
BlenderBot 2 - 3b	2.72	2.19	2.42
BlenderBot 2 - 400m	2.24	1.95	2.17
BlenderBot - 3b	3.64	3.85	3.60
BlenderBot - 9b	3.16	3.18	3.18
DialoGPT	3.24	2.94	3.13
GPT-3	4.54	4.31	4.53
Ground truth	4.16	3.94	4.14
Plato 2	3.86	3.78	3.83
Plato 2 - 24L	2.65	2.63	2.65
Plato 2 - 32L	3.92	3.99	3.65

Table 7: Average annotation ratings per conversational model for Group 4.

Model	App.	Info.	Human.
BlenderBot 2 - 3b	2.94	3.02	3.16
BlenderBot 2 - 400m	2.95	3.52	2.92
BlenderBot - 3b	3.99	4.26	3.94
BlenderBot - 9b	3.57	4.10	3.79
DialoGPT	3.58	3.52	3.67
GPT-3	4.49	4.20	4.60
Ground truth	4.48	4.35	4.57
Plato 2	4.08	4.18	4.19
Plato 2 - 24L	3.17	3.90	3.57
Plato 2 - 32L	4.12	4.50	3.57

Table 8: Average annotation ratings per conversational model for Group 5.

Model	App.	Info.	Human.
BlenderBot 2 - 3b	2.85	2.59	4.79
BlenderBot 2 - 400m	2.82	2.96	4.68
BlenderBot - 3b	3.70	3.38	4.74
BlenderBot - 9b	3.42	3.48	4.81
DialoGPT	3.40	2.37	4.66
GPT-3	4.56	2.56	4.95
Ground truth	4.31	2.77	4.92
Plato 2	3.77	3.46	4.27
Plato 2 - 24L	3.12	4.18	3.92
Plato 2 - 32L	3.63	3.90	4.30

Table 9: Average annotation ratings per conversational model for Group 6.

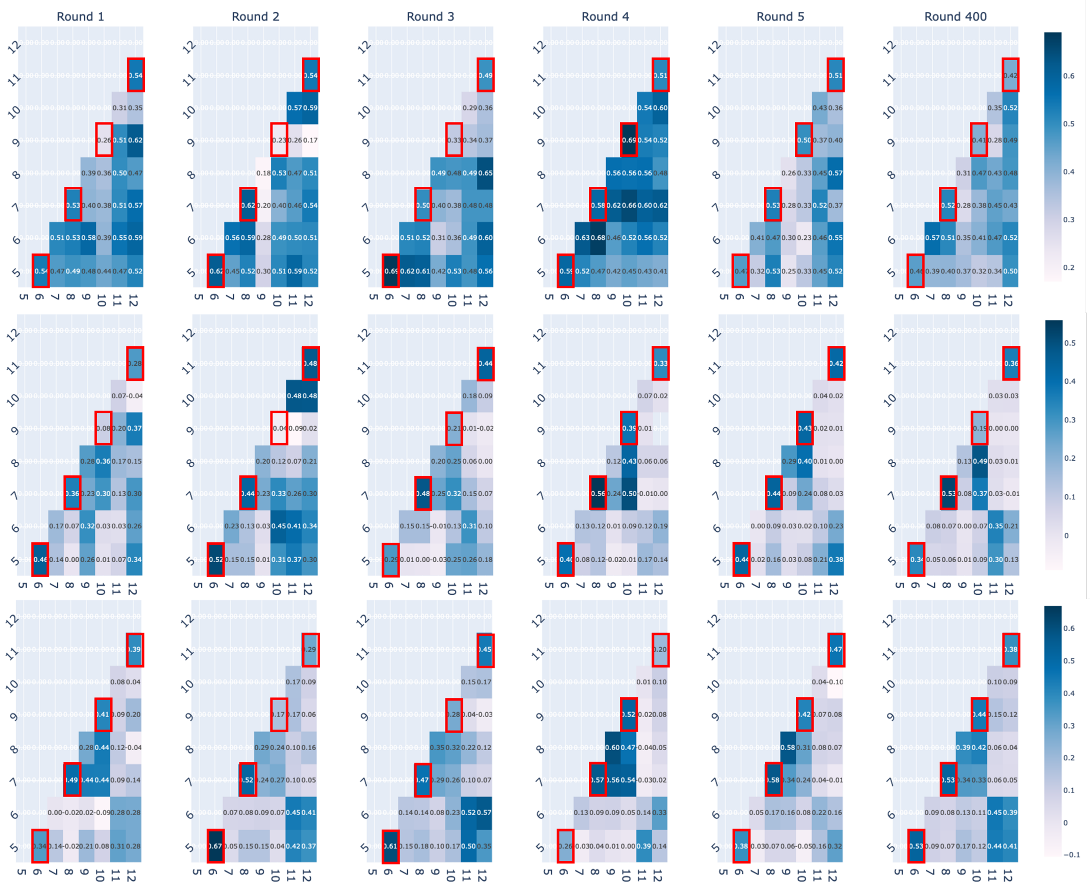


Figure 17: Contingency table of annotations for Groups 3, 4, 5, and 6 - From top to bottom: *Appropriateness*, *Information content of outputs*, and *Humanlikeness*. Round 400 indicates the final round of annotations in **LEAP** with 400 items. Red borders indicate within-group agreement.

A.9 IAA analysis - Iteration-free LEAP

***Within* Group** The red borders in Figure 17 show the change in *within*-group agreement for Groups 3, 4, 5, and 6. We observed that agreement scores for *Appropriateness* were relatively higher than other categories for most rounds across all groups. This coincides with our earlier findings that certain categories, such as *Appropriateness*, may have stronger shared constructs than others.

***Between* Groups** While each group's annotation guideline helped the researchers achieve high agreement within-group, Figure 17 shows that agreement between annotators of different groups remained low throughout the five rounds. Surprisingly, we can observe that agreement between annotators across different groups remained high throughout all five rounds for *Appropriateness*, suggesting that certain annotation categories have a strong shared understanding across annotators of the different groups.

Another interesting observation can be seen in Figure 18, which shows the level of agreement for *Information content of output* during Round 4. The green border shows a distinct silo of agreement between annotators of Groups 4 and 5. We can see that Researcher 10 (Group 5) has *low* agreement scores of 0.09 and 0.01 with Researchers 5 and 6 (Group 3) and 0.07 and 0.02 with Researchers 11 and 12 (Group 6).

However, Researcher 10 has a relatively high agreement of 0.5 and 0.43 with Researchers 5 and 6 (Group 5). With Researcher 7, who also belongs to Group 4, Researcher 10 has an agreement score of 0.39. While the distinction is not as clear, annotators of Group 3 (Researchers 5 and 6) show higher agreement with annotators of Group 6 (Researchers 11 and 12) compared to annotators of Group 4 and Group 5.

Similar distinct silos of agreement can be observed in Figure 17 for *Humanlikeness*, one between Groups 4 and 5 and another between Groups 3 and 6.

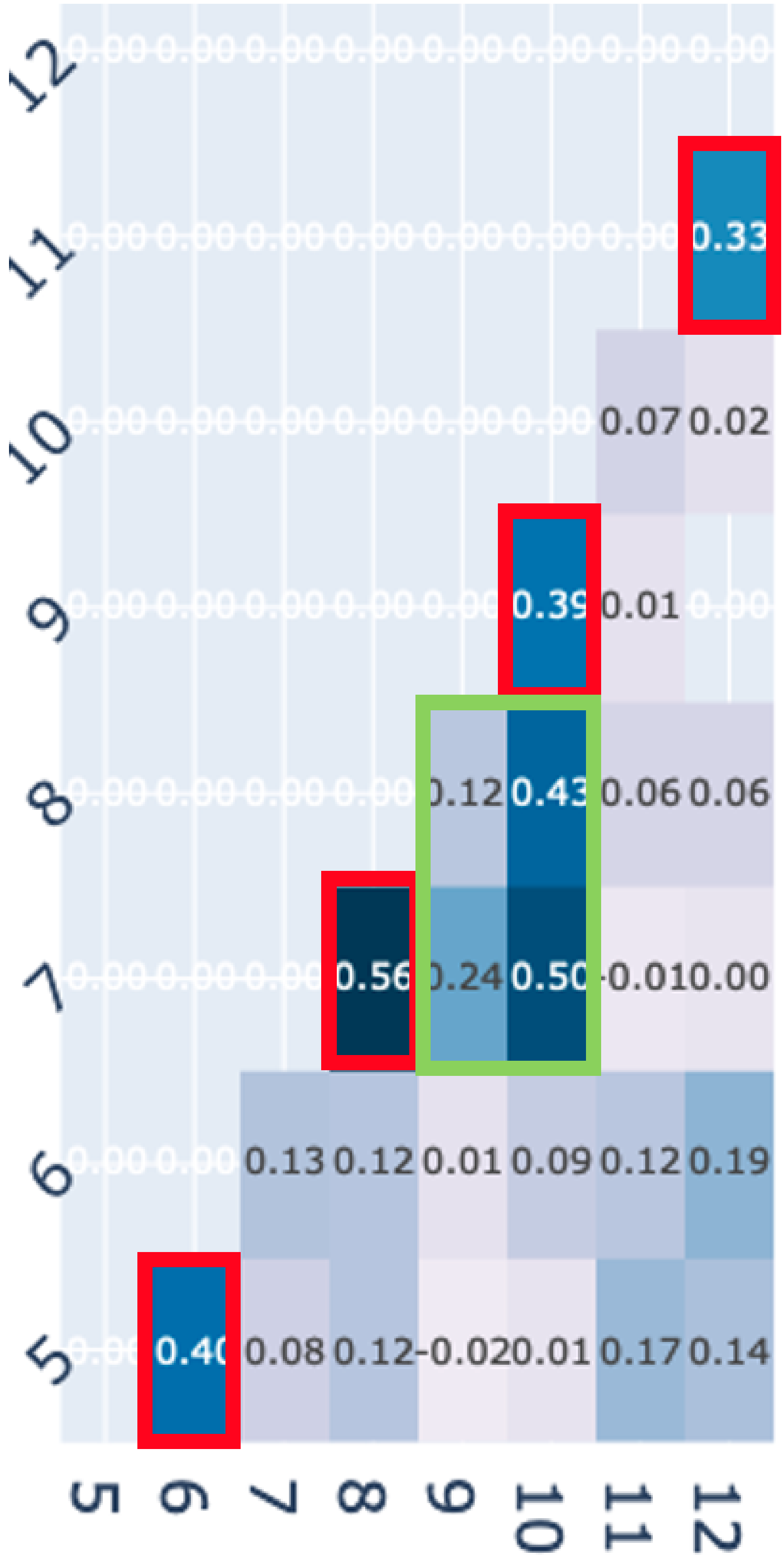


Figure 18: Example of distinct silo of agreement *between* Group 4 and Group 5 for *Information content of output, Round 4*. The Green border show agreement between annotators of Group 4 and Group 5.

A.10 Cohen's Kappa

Counting the raw number of matching annotations is one of the simplest ways to measure agreement. However, the raw agreement fails to account for the possibility of random chance agreement, which becomes problematic when the random chance is very high (Artstein, 2017). To overcome this limitation, **Cohen's Kappa (**κ**)** measures observed agreement above the expected agreement (Cohen, 1968), more formally stated,

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the relative observed agreement among annotators, and p_e is the expected probability of random chance agreement. Cohen's Kappa measures agreement between two annotators, treating any disagreement linearly. If a pair of annotators matches on all annotations (thus $p_o = 1$), then $\kappa = 1$. On the other hand, if the pair has no agreement other than what is expected by chance (thus $p_o = p_e$), then $\kappa = 0$. $\kappa < 0$ is also possible when the pair annotates worse than expected chance agreement ($p_0 < p_e$).

Some annotation studies require different weights to be applied to different levels of agreement between annotators. For example, on a 5-point Likert scale (Likert, 1932), annotation scores 4 and 5 should be regarded as being in higher agreement than annotation scores 1 and 5. To account for this, the **weighted Cohen's Kappa** (Cohen, 1968) is often used to measure IAA in annotation tasks, in order to weigh disagreement differently, thus,

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} x_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} m_{ij}},$$

where w_{ij} is the weight matrix, x_{ij} is the observed matrix, and m_{ij} is the expected matrix.

Cohen's Kappa of 0.6 to 0.8 is commonly regarded as a threshold for sufficient inter-annotator agreement in NLP research (Landis and Koch, 1977). In order to strengthen the reliability of annotation guidelines, various methods have been used to raise the kappa above the threshold, such as taking out outlier anomalous annotations from the dataset (Zhao et al., 2020). However, this is no guarantee that the validity of the dataset is improved by the discarded outliers.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 6 - particularly some models generated responses containing biases.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
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?

Yes, Section 3

☑ B1. Did you cite the creators of artifacts you used?
Yes, Section 3.2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not relevant.

☒ 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)?
Section 1

☑ 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?
Section 3.2

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.2

☑ 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.
Section 3.2

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?
Section 3.3

☑ 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.?
Section 3.3, Appendix A.5

☑ 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)?
Appendix A.5

☑ 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?
Appendix A.5

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
All of the expert annotators are co-authors of the paper. The annotations from AMT workers would be considered IRB exempt.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3.3