# Common Law Annotations: Investigating the Stability of Dialog Annotations

**Seunggun Lee,**[1] **Alexandra DeLucia,**[2] **Ryan Guan,**[3] **Rubing Li,**[1] **Nikita Nangia,**[1] **Shalaka Vaidya,**[1] **Lining Zhang,**[1] **Zijun Yuan,**[1] **Praneeth Ganedi,**[1] **Britney Ngaw,**[1] **Aditya Singhal,**[1] **João Sedoc**[1]
[1]New York University, [2]Johns Hopkins University, [3]University of Pennsylvania

## Abstract

Although metrics for inter-annotator agreement (IAA), like Cohen's Kappa, are often used to measure the reliability of annotation procedures, it is insufficient to ensure survey validity. We conduct an experiment that simulates multiple research groups creating their own annotation guidelines for three main categories. We find that though each pair of researchers raises their agreement by converging on category definitions, the agreement between researchers of different groups falls. We argue that agreement scores should not be blindly raised without considering its implications on the guideline's validity.

## Introduction

Reliable and valid human annotations are an essential component of NLP research, especially for testing and validating datasets or assessing models. While automatic evaluation metrics for text data offer a partial analysis of natural language generation (NLG) model performance, they have yet to become sufficient replacements for human annotations (Liu et al. 2016; Deriu et al. 2021). Nonetheless, it is important to acknowledge that human annotations are inherently subjective (Basile et al. 2021). Each annotator has their own biases (Paun, Artstein, and Poesio 2022) and may have different preconceptions regarding annotation categories.

In order to overcome such ambiguity, research groups strive to develop annotation guidelines that help raise **agreement** among the annotators. Inter-annotator agreement (IAA) metrics, such as Cohen's Kappa (Cohen 1960), are commonly used to measure the agreement between annotators based on a common annotated dataset.

To better understand agreement metrics, we emphasize two distinct qualities: **reliability**—the level of agreement between the annotators—and **validity**— the extent to which annotated data are correct (Paun, Artstein, and Poesio 2022). Reliability implies reproducibility. If annotations have high agreement and reliability, then we expect the annotations to be reproducible when using the same procedure and guidelines (Artstein 2017).

However, having high agreement scores does not necessarily mean that an annotation procedure is valid (Artstein 2017; Paun, Artstein, and Poesio 2022; Basile et al. 2021). This is because *correctness* is a difficult characteristic to judge for inherently subjective tasks. Thus, validity is harder to gauge than reliability for annotation guidelines.

Consider the following trivial annotation guideline: the *Appropriateness* category is scored by counting the number of words in the response. For example, suppose we ask annotators to give an *Appropriateness* score of 1 to a one-word response, an *Appropriateness* score of 2 to a two-word response, and so on. This annotation guideline would yield high IAA, and thus have high reliability, as annotators would be able to produce consistent annotations. However, the guideline has low validity, as it employs an extremely specific set of rules that is not applicable to the broader use of the category of *Appropriateness*.

Similarly, when researchers create annotation guidelines with the intent of raising IAA, it can result in instructions that are highly specific to the paper, but which diverge in definition across different research groups. This is a problem, as the guidelines may reinforce biased annotations (Craggs and Wood 2005). To analyze these issues, we simulate the annotation procedure of multiple research groups on two dialog datasets, observing the change in agreement when these groups create independent guidelines.

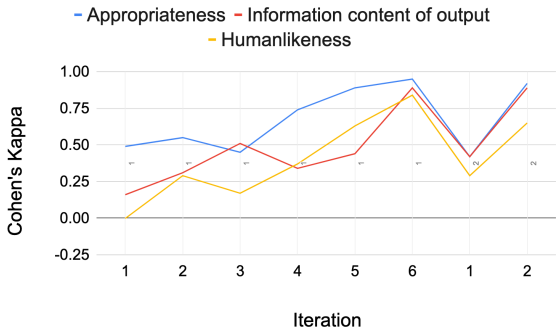


Figure 1: Agreement scores for Group 6 - using **LEAP**

Within the simulation, we observe that each of the groups create different guidelines, despite working on the same dataset and annotation categories. We label each of these guidelines using the metaphor of a **common law** to describe a set of shared understandings regarding an annotation pro-


Figure 2: Agreement scores between annotators across all groups - *Information content of output*.

cedure. We show that though each group is able to raise agreement within itself, the agreement falls between annotators of different groups, especially for the categories *Information content of output* and *Humanlikeness*. We extend the scope of the analysis by using the guidelines to collect crowdsourced annotations on the same dataset.

## Data

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 different models - DialogGPT (Zhang et al. 2019), GPT3 (Brown et al. 2020), Plato2 (24L and 32L) (Bao et al. 2020), BlenderBot (2.7B and 9B) (Miller et al. 2017), BlenderBot 2 (400M and 3B) (Weston and Shuster 2021; Komeili, Shuster, and Weston 2021; Xu, Szlam, and Weston 2021), and the original human response.

## Experimental Design

We initially created three base annotation criteria: *Appropriateness*, *Information content of output*, and *Humanlikeness*. These definitions inspired by Howcroft et al. (2020):

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. *Humanlikeness*: The degree to which an output could have been produced by a human.

Eight of the co-authors (a.k.a. annotators) were divided into four pairs. Each pair met to discuss their "common law" annotation methodology to maximize annotation agreement. All discussions were recorded through Zoom with the built-in transcripts tool enabled.

However, after the four pairs participated in the aforementioned procedure, we observed that none of the groups were able to reach the intended average IAA level of 0.7. To address this issue, we propose an alternate protocol of annotation: if a group is unable to reach an average agreement of 0.7 across the three categories, they re-annotate a shuffled set of the same 50 prompt and response pairs after their discussion, rather than moving onto the next round of annotations. This re-do of discussions and annotations on the same set of data, which we term *iterations*, is repeated until the group is able to converge on their definitions and reach an average $\kappa > 0.7$ across the three categories. We term this protocol the Lee et al. Protocol (**LEAP**).

Four additional co-authors were grouped into two pairs and annotated data following the **LEAP** protocol. Both pairs of annotators were able to reach an average $\kappa > 0.7$ by the second round 1. Finally, we used the annotation guidelines to collect crowdsourced data on Amazon Mechanical Turk (AMT).

## Results

Figure 1 shows the calculated IAA of Group 6 (for all group agreement numbers see https://common-law-dash.herokuapp.com/). The results show that $\kappa > 0.7$ is difficult to achieve and information content is the most difficult quality to achieve agreement on.

We found that IAA between different groups diverged throughout each round, though at differing levels depending on the annotation category. Groups had the highest agreement between each other for the *Appropriateness* category, compared to *Information content of output* and *Humanlikeness*. Surprisingly, annotators of Groups 1 and 4 showed relatively high agreement with each other for *Information content of output* and *Humanlikeness*, while Groups 2 and 3 showed high agreement with each other for the two categories (e.g., Figure 2).

The recorded transcripts of discussions reveal that the two sets of convergence occur because of similar annotation guidelines were created between the groups with high agreement. Further visualizations of the analyzed data are made available at https://common-law-dash.herokuapp.com/.

## Future Work

Currently, we are using the annotation guidelines to collect crowdsourced annotations through AMT, where six different mutually exclusive groups of workers will annotate the same dataset using the six different annotation guidelines. We hypothesize that although AMT workers will have a high agreement with the researchers that created their respective guidelines, the workers will have low agreement across groups, just as it was for the researchers in different groups.

## References


Artstein, R. 2017. *Inter-annotator Agreement*, 297–313. Dordrecht: Springer Netherlands. ISBN 978-94-024-0881-2.

Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; and Xu, X. 2020. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. *arXiv preprint arXiv:2006.16779*.

Basile, V.; Fell, M.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; Poesio, M.; and Uma, A. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, 15–21. Online: Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Craggs, R.; and Wood, M. M. 2005. Squibs and Discussions: Evaluating Discourse and Dialogue Coding Schemes. *Computational Linguistics*, 31(3): 289–296.

Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; and Cieliebak, M. 2021. Survey on Evaluation Methods for Dialogue Systems. *Artif. Intell. Rev.*, 54(1): 755–810.

Howcroft, D. M.; Belz, A.; Clinciu, M.-A.; Gkatzia, D.; Hasan, S. A.; Mahamood, S.; Mille, S.; van Miltenburg, E.; Santhanam, S.; and Rieser, V. 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*, 169–182. Dublin, Ireland: Association for Computational Linguistics.

Komeili, M.; Shuster, K.; and Weston, J. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 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)*, 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132. Austin, Texas: Association for Computational Linguistics.

Miller, A. H.; Feng, W.; Fisch, A.; Lu, J.; Batra, D.; Bordes, A.; Parikh, D.; and Weston, J. 2017. ParlAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476*.

Paun, S.; Artstein, R.; and Poesio, M. 2022. *Learning from Multi-Annotated Corpora*, 147–165. Cham: Springer International Publishing. ISBN 978-3-031-03763-4.

Sedoc, J.; Ippolito, D.; Kirubarajan, A.; Thirani, J.; Ungar, L.; and Callison-Burch, C. 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)*, 60–65. Minneapolis, Minnesota: Association for Computational Linguistics.

Weston, J.; and Shuster, K. 2021. Blender Bot 2.0: An open source chatbot that builds long-term memory and searches the internet.

Xu, J.; Szlam, A.; and Weston, J. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *CoRR*, abs/1911.00536.