Sources of Disagreement in Data for LLM Instruction Tuning

Russel Dsouza
School of Computer Science
University of Birmingham
r.s.dsouza@bham.ac.uk

Venelin Kovatchev *
School of Computer Science
University of Birmingham
v.o.kovatchev@bham.ac.uk

Abstract

In this paper we study the patterns of label disagreement in data used for instruction tuning Large Language models (LLMs). Specifically, we focus on data used for Reinforcement Learning from Human Feedback (RLHF). Our objective is to determine what is the primary source of disagreement: the individual data points, the choice of annotators, or the task formulation. We annotate the same dataset multiple times under different conditions and compare the overall agreement and the patterns of disagreement.

For task formulation, we compare SINGLE format where annotators rate LLM responses individually with PREFERENCE format where annotators select one of two possible responses. For annotators, we compare data from human labelers with automatic data labeling using LLMs.

Our results indicate that: (1) there are very few "universally ambiguous" instances. The label disagreement depends largely on the task formulation and the choice of annotators; (2) the overall agreement remains consistent across experiments. We find no evidence that PREFERENCE data is of higher quality than SINGLE data; and (3) the change of task formulation and annotators impacts the resulting instance-level labels. The labels obtained in different experiments are correlated, but not identical.

1 Introduction

Training large language models (LLMs) to follow instructions and aligning them to human preferences is a key step in aiming to ensure that models are helpful and harmless (Leike et al., 2018). In this paper we explore the quality of the data used in the process. We seek to determine the cause for disagreement when rating in-context LLM responses. We conducted a set of experiments to assess to what extent disagreement depends on the task formulation (individual rating vs. preference) and the choice of annotators (humans vs. LLMs).

*Corresponding Author

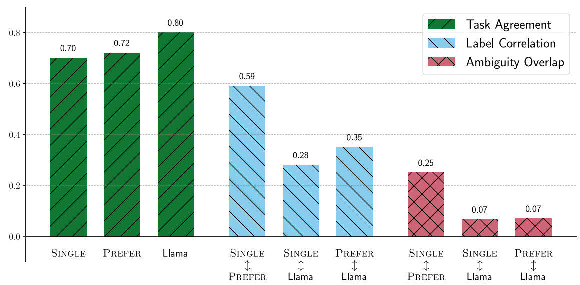


Figure 1: Overall task agreement, cross-task label correlation, and cross-task overlap of ambiguous instances.

We sampled 720 instances from the Anthropic dataset (Bai et al., 2022a) and performed several independent annotations. For task formulation, we compared SINGLE , where annotators assign individual score to each context-response pair, and PREFERENCE , where annotators have to choose between two possible responses for the same context. For annotators, we compared (1) labels obtained by humans with (2) labels obtained from pre-trained LLMs internal states and (3) zero-shot labels obtained from LLMs. We test both "base" LLMs and their "instruction-tuned" counterparts.

For each experiment we measured: (1) the data quality (inter-annotator agreement); (2) the cross-task correlation of labels; and (3) the cross-task correlation of instance-level agreement and the overlap of "ambiguous" examples. More explicitly, we formulate the following **research questions**:

1. **Overall IAA** How much does the overall data quality (IAA) change based on task formulation and annotator choice?

2. **Gold Labels** Do different experiments approximate the same underlying distribution?

3. **Instance Ambiguity** To what extent does instance-level ambiguity depend on the experimental design and annotator choice?

Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation, pages 20–32
January 19, 2025.

We perform further experiments to determine if we can combine the data from the different experiments and obtain more robust annotations.

Figure 1 shows a summary of our results. We find that: (1) the overall IAA is similar across experiments. Pre-trained LLMs tend to agree more with each other than human annotators, which may indicate a potential bias and lack of diversity in the models. (2) The SINGLE and PREFERENCE experiments assign labels with a strong correlation, but also with significant differences. Labels from LLMs have a moderate correlation with human preference (which response is better) but low agreement on the magnitude of the difference (how much better is the selected response). (3) Very few of instances are "universally ambiguous". We find that annotation ambiguity is largely a function of the task format and the choice of annotators.

Our work sheds a new light on acquiring data for LLM instruction tuning. Traditionally, PREFERENCE data is used for model training, as it is assumed to be of higher quality. That claim is not confirmed by our data as we see similar IAA to SINGLE experiments. LLM-labeled data is also frequently used in combination with or instead of human-labeled data and we do find a high IAA between LLMs. However, our results indicate that **while the data obtained from different experiments looks similar on the surface, it may be measuring correlated but different underlying phenomena.** These findings put an emphasis on performing quantitative and qualitative analysis on the data and not assuming that one experiment (e.g., PREFERENCE) is a perfect substitute for another. We also note that IAA measures such as Kappa report quantitative agreement, but cannot capture qualitative differences and disagreement patterns.

2 Related Work

Instruction following Leike et al. (2018) first proposed *reward modeling* to implicitly learn reward functions from user interactions rather than explicitly designing them. Böhm et al. (2019) and Ziegler et al. (2019); Stiennon et al. (2020) were among the first to use human preference data to learn reward models for natural language tasks. Askell et al. (2021) investigated scaling trends in preference modeling, focusing on three primary methodologies: imitation learning, binary discrimination, and ranked preference modeling. They found that ranked preference modeling significantly outperformed imitation learning, while binary discrimination only offered marginal benefits.

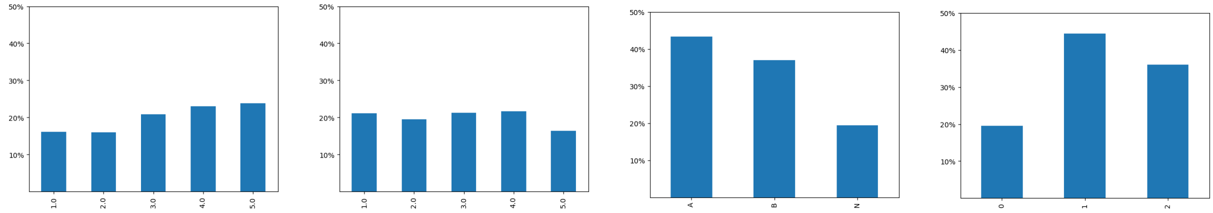
Data for RLHF Ouyang et al. (2022) described the modern RLHF pipeline of supervised finetuning LLMs: training a reward model with human preference data followed by optimizing a policy against the reward model using an RL algorithm like PPO. The authors asked human raters to label their preferred output among k choices, resulting in $\binom{k}{2}$ comparisons, for a given input which were then used to train a reward model to predict human preferred outputs. Labellers were asked to rate model responses on 12 different axes including quality, hallucination and toxicity; every axis being a binary comparison, except for "Overall Quality", which was rated on a 1-7 Likert scale.

Starting with Bai et al. (2022b) and Touvron et al. (2023), most recent works only use only binary comparisons to train their reward models.

Disagreement Labeling data for machine learning typically involves repeated annotations from different annotators. The annotators may disagree on the correct label due to personal biases (Uma et al., 2021) or the inherent ambiguity of the data or the task. Leonardelli et al. (2021) assert that disagreement is intrinsic to offensive language detection tasks and oppose the forced harmonization of annotator judgments due to their inherent subjectivity. Baumler et al. (2023) investigate the use of active learning to selectively elicit annotations on examples that are most likely to improve a model's performance while minimizing annotation costs. Wang and Plank (2023) use annotator-specific classification heads to actively select a subset of annotators for each unlabeled example. Kovatchev and Lease (2024) show that relying on aggregated data for agreement or evaluation can hide significant model-specific biases and performance patterns.

Synthetic data for RLHF Wang et al. (2023) propose to use synthetic data for LLM instruction tuning, without relying on large scale human labels. Wang et al. (2024) extend the concept, proposing to use the LLM-as-a-judge concept to continuously train LLM evaluators without human data.

Role of disagreement in RLHF Siththaranjan et al. (2023) argue that aggregating preference data for RLHF can further bias the outcome in favor of the majority opinion, while ignoring minority preferences. Poddar et al. (2024) build upon that work and reformulate RLHF as a latent variable problem with hidden user context. They were able to train multiple LLM-based reward models to learn a sepa-



(a) SINGLE Rating A (b) SINGLE Rating B (c) PREFERENCE response (d) PREFERENCE magnitude

Figure 2: Label distribution of annotated dataset

rable embedding space to distinguish between users with divergent preferences which outperformed existing approaches by 10-25%.

3 Human Data Acquisition

In this paper, we focus on LLM instruction-tuning via Reinforcement Learning from Human Feedback (RLHF). During RLHF finetuning, the labeled data is used to train a "reward" model. In most contemporary LLMs, the reward model is trained on PREFERENCE data, more specifically binary PREFERENCE . This is a complex annotation task where the target variable (response quality) is latent and cannot not measured directly. Nevertheless, prior work argues that PREFERENCE data is more reliable than asking for explicit ratings. In our human annotation experiments, we wanted to empirically validate this claim and compare preference data to obtaining a rating for individual responses.

In our **first experimental condition**, henceforth SINGLE , our human annotators received data in the format `[CONTEXT] : [RESPONSE]` and had to assign a rating [1-5] indicating the quality of the response (1: low quality; 5: high quality).

In our **second experimental condition**, PREFERENCE , the annotators received data in the format `[CONTEXT] : [RESPONSE A] / [RESPONSE B]` and had to indicate: (1) the preferred response (A, B, None); and (2) the magnitude of the difference (0: no difference; 1: preferred response (A/B) is a little better; 2: preferred (A/B) is much better).

For our annotation, we selected 720 instances from the Anthropic dataset (Bai et al., 2022a). We sampled an even number (360) from "helpful" and "harmless" instances. Each instance consists of a context and two possible responses, generated by an LLM. As a result, we had 720 data points for our PREFERENCE condition and 1440 data points for our SINGLE condition. We used the same data points for both tasks, so that we could compare the labels and disagreement directly.

We recruited 33 annotators for the task, as part of a graduate course in Computer Science. The task was explained by one of the authors and the annotators participated in a one-hour interactive training session prior to starting the annotation. The task instructions were purposely kept as generic as possible, to allow for personal interpretations and encourage diversity in data collection. Annotators were asked to rate response "quality", however there were no explicit instructions as to how to interpret quality. Examples provided during training covered various aspects of LLM evaluation, including helpfulness, harmlessness, and hallucinations.

Each annotator received 40 contexts and 80 possible responses. Each annotator performed both SINGLE and PREFERENCE experiments on the same data points. Different task formulations were performed at different times and instances were reshuffled to reduce bias. Each instance was annotated by two different annotators. Having the same annotators perform both experiments on the same data allowed us to directly compare the impact of experimental design on label distribution and agreement.

Figure 2 shows the label distribution of the annotated dataset. We calculated two separate SINGLE distributions based on the position the sentence has in the paired format. SINGLE rating A (2a) shows the labels for sentences that appear first and SINGLE rating B (2a) shows the labels for second sentences. Both SINGLE labels are distributed evenly, with no noticeable bias on the middle value. Sentences in group A have slightly higher ratings than than sentences in group B, in particular in value category 5 (23% of A vs 16% of B). Figures 2c and 2d show the labels in the PREFERENCE condition. Sentence A is preferred 43% of the cases vs 37% for sentence B. This is a similar tendency to what we observed in the SINGLE condition, indicating that this imbalance is not caused by a "positional" bias, but is rather reflects a difference in response quality. 20% of the instances do not have a clear preference, which also aligns with SINGLE data. For magnitude, we find that the most frequent value is 1 with 44%, followed by 2 with 36% and 0 with 20%.

Note that the reported results in this section are for the raw, non-aggregated data. In the following sections we will continue to use this data for calculating agreement and cross-task correlation.

4 Comparing Human Data

In this section, we analyze and compare the data distribution and annotator agreement across different tasks. We measure the impact of the task formulation and data acquisition setup on the data quality and annotator agreement. Our **Research Questions** for this section are the following:

- **Data Quality**: To what extent does task formulation impact data quality (agreement)?
- **Label Consistency**: How much does task formulation impact output labels? Do different formulations "agree" with each other?
- **Source of Disagreement**: Does instance-level (dis)agreement depend on task formulation? Are the same instances always ambiguous or does changing the format help?
- **Complementary Annotation**: Can we combine data from different experiments to obtain a more robust dataset?

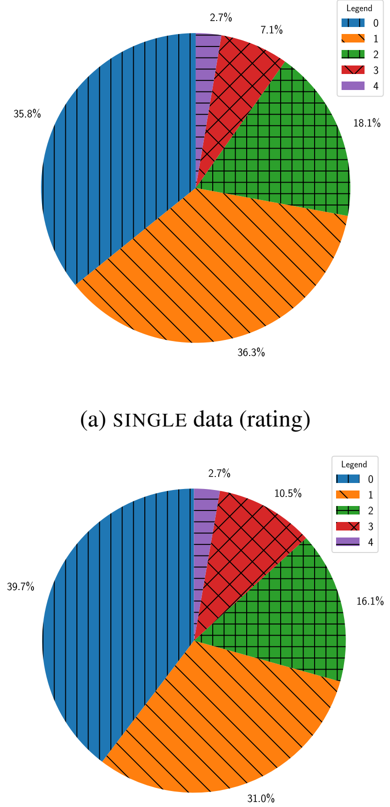
We performed several deterministic transformations of the data, so that the results from the two experiments could be compared directly.

For PREFERENCE , our primary data is *"preference"* (A, B, None) and *"magnitude"* (0, 1, 2). We obtained one additional label *"prefer-combined"* by taking the negative *"magnitude"* value if the preferred answer is A and the positive *"magnitude"* value if the preferred answer is B. The resulting values range from -2 (A $\gg$ B) to +2 (B $\gg$ A).

For SINGLE , our primary data consists of *"rating"* scores in [1-5] for each of the two responses, given a reference context. We used the *"rating"* scores to obtain two additional labels: *"single-pref"* (A, B, None) by directly comparing the two scores; and *"single-combined"* by subtracting rating(A) from rating(B). The resulting scores range from -4 (A $\gg$ B) to +4 (B $\gg$ A). We clipped the scores at [-2, 2] to match *"prefer-combined"*.

4.1 Data Quality

Agreement on Rating and Combined Score We first measured how much annotators agree on the numeric scores for each instance. For SINGLE we compared the *"rating"* values. For PREFERENCE we compared the *"prefer-combined"* values. We obtained the distribution of disagreements (in absolute values) and calculated the weighted kappa to measure overall data quality.



(a) SINGLE data (rating)

(b) PREFERENCE data (combined score)

Figure 3: Label score difference

Figure 3 shows the distribution of absolute score (dis)agreement. For the SINGLE data, 35.7% of the instances have a difference of 0 (complete agreement), 36.3% have a difference of 1 and 18% have a difference of 2. A total of 9.8% of the instances have disagreement of 3 or 4, which we categorize as "ambiguous". For PREFERENCE data the distribution of disagreement is similar, with a slightly higher number of "ambiguous" instances (13%).

Table 1 shows the Kappa for Rating and Combined Score. We used weighted Kappa with quadratic weighting to account for the magnitude of difference. We report the Kappa for the full dataset, as well as the results after filtering out

Experiment	All	$\Delta < 4$	$\Delta < 3$
SINGLE	.48	.57	.70
PREFERENCE	.43	.51	.72

Table 1: Weighted kappa for rating/magnitude

instances with disagreement 4 ($\Delta < 4$) and all *ambiguous* instances ($\Delta < 3$). The agreement on the full dataset is moderate (.43 – .48). Filtering out $\Delta < 4$ increases the agreement slightly. Filtering all *ambiguous* instances ($\Delta < 3$) results in high agreement, as measured by kappa above .70. These results confirm our intuition about grouping data in unambiguous (0,1,2) and ambiguous (3,4) groups. They also validate the overall quality of the acquired data. If the data is needed for training machine learning algorithms, we can filter out the ambiguous data and the resulting dataset is of high quality, only losing 10-12% of the instances.

Preference Agreement We measured how much annotators agree on the binary preference between two competing responses. For PREFERENCE data we used the primary *"preference"* column. For SINGLE we used *"single-pref"*. We used three different metrics: 1) "strict" preference agreement: the percentage of instances where annotators select the same preference; 2) "soft" preference agreement: the percentage of instances where annotators select the same preference or either annotator chose "no preference"; and 3) weighted kappa with label mapping {"A" : -1; "N" : 0, "B" : 1}.

Experiment	Strict	Soft	Kappa
SINGLE (all)	.54	.86	.38
SINGLE ($\Delta < 4$)	.58	.91	.49
SINGLE ($\Delta < 3$)	**.62**	**.95**	**.61**
PREF (all)	.59	.81	.40
PREF ($\Delta < 4$)	.61	.83	.44
PREF ($\Delta < 3$)	**.67**	**.93**	**.64**

Table 2: Preference agreement with and without filtering

Table 2 shows the results for preference agreement. Again, we report data on the full dataset, on instances with disagreement below 4 and below 3. Once again, we find that filtering out *ambiguous* examples ($\Delta < 3$) gives us a high quality dataset. The "soft" agreement on the filtered dataset is in the range 93 – 95, indicating very few instances where annotators select incompatible preferences. It is interesting to note that the results for SINGLE acquisition are comparable to those for PREFERENCE despite us obtaining those results indirectly.

After analyzing the agreement data (both absolute and chance-corrected), we can conclude that the task formulation does not directly impact overall data quality. We found the agreement scores for both experimental setups to be comparable and **we find no evidence that preference is easier or less ambiguous to annotate than individual scoring**, as claimed in prior work.

4.2 True Label

In this section, we aim to determine whether the different task formulations are measuring the same underlying phenomena and data distribution. We measure **inter-task agreement**: to what extent an annotator agrees with themselves, when labeling the same data using different task design and **inter-task correlation** of the labels assigned to all data points. We calculate the following metrics: 1) preference agreement (soft / strict) between *"preference"* and *"single-pref"*; 2) preference weighted kappa between *"preference"* and *"single-pref"*; 3) combined weighted kappa between *"prefer-combined"* and *"single-combined"*; and 4 Pearson correlation between *"prefer-combined"* and *"single-combined"*. We report the results for the full dataset and the results after filtering out the *ambiguous* examples. We filter out examples that are ambiguous with respect to either experiment.

Metric	All	$\Delta < 3$
Pref (strict)	.60	.62
Pref (soft)	.88	.91
Kappa (pref)	.50	.56
Kappa (score)	.54	.59
Pearson	.55	.59

Table 3: Inter-task agreement and correlation

Table 3 shows the results. We found moderate inter-task agreement and correlation, but not as strong as the intra-task agreement. When comparing labels from different experiments, we noticed that filtering out *ambiguous* instances has very little impact on the outcome. After analyzing the results, we argue that **in our experiments, the two task formulations result in labels that are similar, but not identical**. Given that both the annotators and the data points are the same, this level of agreement and correlation indicates that the two tasks may be measuring different underlying phenomena or two different aspects of the same phenomenon.

4.3 Source of Disagreement

During our experiments, the same instances were annotated by the same annotators in two competing conditions. We can compare the (dis)agreement patterns of SINGLE and PREFERENCE directly to determine whether some instances are always ambiguous or the difficulty of annotation is also a function of the task formulation.

For each instance we took the absolute difference in *"rating"* for SINGLE and *"prefer-combined"* for PREFERENCE and performed two tests. First, we calculated the Pearson correlation (of disagreement). Then we obtained the sets of all instances that are *ambiguous* with respect to *"rating"* ($\Delta \geq 3$) and all instances that are *ambiguous* with respect to *"prefer-combined"* ($\Delta \geq 3$). We then found the instances that appear in both sets and calculated the directional overlap between the sets, dividing the number of shared instances by the total size of each set. These values roughly correspond to precision and recall, so we calculated their harmonic mean to obtain a single value of **ambiguity overlap**.

Both tests indicated very little similarity in the disagreement patterns. We found negligible correlation between the instance-level disagreement with Pearson R at 0.2. The ambiguity overlap between the two sets was 0.25. Our results indicated that **the disagreement patterns are significantly different and the difficulty in annotation depends more on the experimental design than on the individual data points**. Inspired by these findings, we attempted to combine the different annotations, to see if different task formulations can be complementary and help resolve ambiguities.

4.4 Complementary Annotation

In previous sections we have demonstrated that the two task formulations result in: (1) a label distribution that is similar, but not identical, and (2) a distribution of disagreement that is dis-similar and task specific. Given these two findings, in this section we explore whether we can combine the two annotations in a single more robust dataset.

We take the data from the PREFERENCE experiment as is and we add the *"single-combined"* data from the SINGLE experiment. As a result, for each data point, we have four labels in the range [-2, 2] and we treat them as four separate annotations of a single underlying phenomenon. We calculate the inter-annotator agreement using Krippendorff Alpha, to determine whether the resulting corpus is more robust than either of the individual experiments. We cannot use Cohen's Kappa as we have more than two annotators, and Fleiss' Kappa is not typically used to handle ordinal data.

Experiment	**All**	**$\Delta < 3$**	**$\Delta \geq 3$**
PREFERENCE (score)	.44	.72	-.69
FULL (score)	.45	.55	-.11
PREFERENCE (pref)	.40	.64	-.99
MERGE (pref)	.40	.50	-.19

Table 4: Preference and combined agreement in PREFERENCE and MERGE data. Columns correspond to "all", "unambiguous" (good), and "ambiguous" instances.

Table 4 shows the impact of merging annotations for the full dataset, the unambiguous examples ($\Delta < 3$) and the ambiguous examples ($\Delta \geq 3$). We compare the α for the PREFERENCE data with the α for the MERGE data. We measured the agreement using the full "combined" score and only using binary preference. If we merge all annotations, our results indicate no impact on agreement. Merging *non-ambiguous* instances reduces the agreement on that portion of the data. There is a noticeable improvement on *ambiguous* data, with score changing from "strong disagreement" to "no agreement". As such, if we apply selective merging and only get additional annotations on instances with $\Delta \geq 3$, the overall agreement will increase. Nonetheless, the *ambiguous* will still have no clear label with α around zero. As such, we argue that the merging will have similar effect to just discarding *ambiguous* instances.

Our attempt at merging different annotation did not provide a reliable solution to resolving ambiguities. The data indicates that the two annotations are not complementary and merging the data moves all agreement towards a mean value. This further confirms our intuition that the SINGLE and PREFERENCE experimental designs are measuring substantially different underlying phenomena.

5 LLM-based Annotation

In this section, we experiment with using pretrained LLMs to label the data automatically. We perform two sets of experiments: PERPLEXITY and ZERO-SHOT. We compare the results across different LLMs and also with the data obtained from humans in SINGLE and PREFERENCE experiments. Our Research Questions are the following:

- **Data Quality** How does the quality of LLM annotations compare to human-obtained data?
- **Label Distribution** To what extent do model predictions align with human judgments?
- **Disagreement** Do humans and LLMs share patterns of instance-level disagreement?

Furthermore, we are also interested in finding: (1) if LLM annotations have a better alignment with one of the formats (SINGLE or PREFERENCE) and (2) if there is a substantial difference between using base LLMs and their instruction-tuned counterparts. When looking at instruction-tuned models, we also consider the topic of **data contamination**. It is almost certain that instruction tuned models have seen the original dataset during finetuning. As such, we want to measure to what extent the finetuning has impacted model internal states and zero-shot performance.

5.1 Perplexity-based Labeling

Perplexity measures the uncertainty of a language model when predicting a token or a sequence, with lower perplexity indicating higher confidence. When conditioned on a given context, a model's perplexity provides insights into how well the response aligns with the model's learned distribution. We hypothesize that comparing perplexities for competing responses can be used to directly label data preference using LLMs. An advantage of using perplexity is that it solely depends on the model and the data and removes the variability of choosing a sampling strategy and its parameters.

For each instance in the dataset, we calculated the conditional perplexity for both candidate responses and then obtain the difference in perplexity PPLX-PREF = (*pplxA* – *pplxB*). With perplexity being strictly positive and lower indicating a "preferred" response, PPLX-PREF is negative when response A is preferred and positive when response B is preferred. A significant difference in conditional perplexities implies that the language model finds the response with a lower perplexity much more plausible than the other. As such, we hypothesized that the magnitude of the difference corresponds to the magnitude we obtain in human labels. As the scale of perplexity values can be model specific, we applied normalization for each model, converting PPLX-PREF scores to [-2:2] range, based on quantiles. The 20% of responses with smallest magnitude of difference were rated as "no preference" and a value of 0. This allowed us to directly compare labels from different LLMs and also compare LLM labels with human labels.

Model	Size	Reference
gpt-2 Large	0.7B	Radford et al. (2019)
Llama-3.2	1B	Dubey et al. (2024)
Llama-3.2 **I**	1B	Dubey et al. (2024)
Phi-3.5-mini **I**	3.5B	Abdin et al. (2024)
Mistral-v0.3	7B	Jiang et al. (2023)
Mistral-v0.3 **I**	7B	Jiang et al. (2023)
Llama-3.1	8B	Dubey et al. (2024)
Llama-3.1 **I**	8B	Dubey et al. (2024)

Table 5: Models used. **I** refers to the instruction-tuned version of the base model. Note: gpt-2 is used only for the PERPLEXITY experiment.

Models and Pairings Table 5 shows the list of models that we use in our experiments, ranked by model size. The **I** indicates an instruction-tuned model. Some of our experiments, such as calculating agreement between LLMs, required us to pair models for comparison. Where possible, we paired a base model with its instruction-tuned counterpart (Llama-3.1, Llama-3.2, and Mistral). We also paired Llama-3.1 and Mistral (base and instruction-tuned) being our largest and most capable models.

Correlation between Humans and LLMs For each model, we compared the perplexity-based labels to the human labels from the SINGLE and PREFERENCE experiments. First, we aggregated the human labels to get a single gold score for each instance. For PREFERENCE we took the mean *"prefer-combined"*. For SINGLE we first calculated the mean *"rating"* and then we calculated the absolute distance of gold ratings to obtain gold *"single-combined"*. After that we measured the agreement between human and LLM labels in two ways: 1) Pearson correlation of labels[1]; and 2) Weighted kappa on binary preference labels.

Figure 4 shows the scores for the different models. Looking at the results we can conclude that:

- the label agreement between humans and LLMs is moderate and is lower than the agreement between humans within and across tasks
- the agreement between LLMs and humans increases with model size

[1]We also used weighted kappa and got the same results

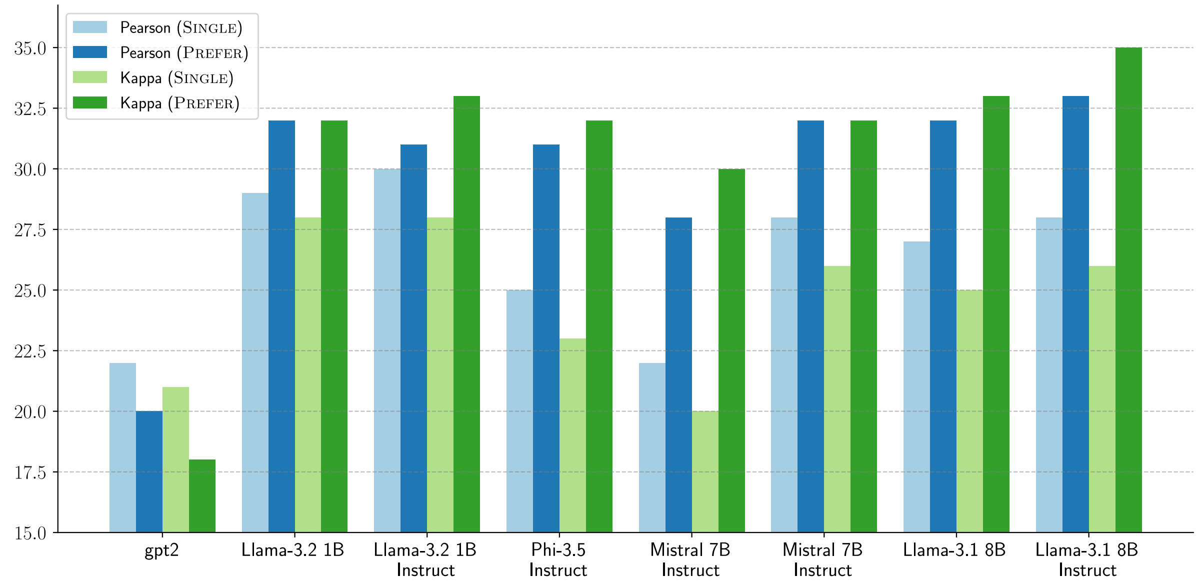


Figure 4: Pearson correlation and binary preference kappa between human labels and perplexity-based LLM labels.

- LLMs labels correlate more strongly with PREFERENCE labels than with SINGLE ones
- Instruction-tuned models agree with humans more than base models, but the difference is marginal, except for Mistral

Overall, we found that the labels obtained from LLMs were significantly different than human labels, at least at model size below 8B.

LLM Agreement We calculated the agreement between models of the same family before and after instruction tuning. We also calculated the agreement between Mistral-7B and Llama-3.1-8B in both base and instruct models. In all pairings, we obtained strong agreement (weighted kappa > .8), except for Mistral-7B-Instruct and Llama-3.1.-8B-Instruct, where the agreement was .75. Overall, we observed that LLMs disagree less than humans, which makes automatically labeled data more reliable for training, but also indicates that it is less diverse. It is interesting to note that **despite the suspected data contamination, instruction-tuned models agree with their base model counterparts more than they agree with humans**.

Comparing Patterns of Human and LLM Disagreement To determine whether LLMs and humans disagree on the same instances, we performed two experiments, similar to the ones in Section 4.3. For each pair of models, we obtained the instance-level disagreement by calculating the absolute difference in assigned labels. We identified the "ambiguous examples" as the subset of examples with label difference $\Delta \geq 3$. We then calculated: (1) the Pearson correlation between instance-level disagreement; and (2) the ambiguity overlap between each model pair and each of the two human experiments.

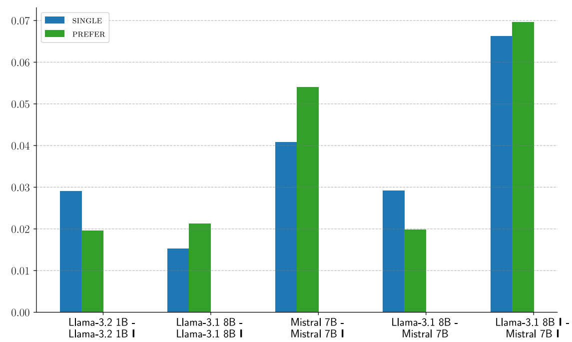


Figure 5: Overlap of ambiguous examples between humans and LLMs

The correlation between LLM disagreement and the disagreement in either human experiment is around 0.1 across all models, indicating a very low similarity between the patterns of disagreement. Figure 5 shows the ambiguity overlap, which is below 0.07 across all models.

Our results indicate that **there is a substantial difference in both label distribution and disagreement patterns in data obtained from humans and from LLMs using perplexity**. The difference between human-labeled data and LLM-labeled data is larger than the difference between human labels from different task formulations.

5.2 Zero-shot Labeling

While perplexity provides implicit signals, structured prompting enables explicit elicitation of model preferences. We design prompts to explore the relationship between model-generated outputs and annotator preferences[2].

- SINGLE-LLM: The model is instructed to rate a context – response on a scale from 1 to 5.
- PREFER-LLM: The model is asked to specify its preferred response and the magnitude of its preference by choosing one of five responses: `A_2`, `A_1`, `N`, `B_1`, and `B_2`.
- DISAGREEMENT-LLM-S: The model is instructed to predict the difficulty of a context – response pair in the Single-LLM task.
- DISAGREEMENT-LLM-P: The model is instructed to predict the difficulty of a context – response pair in the Prefer-LLM task.

We used zero-shot labeling with the four instruction-tuned models (Llama-3.2-1B-Instruct, Phi-3.5-3B-Instruct, Mistral-7B-Instruct-v0.3, and Llama 3.1-8B-Instruct). Similar to the experiments in section 5.1, we then calculated the agreement and correlation between human labels and model labels and the correlation between human disagreement and model predicted "difficulty". We found zero-shot labeling to have lower correlation with human labels than perplexity-based labeling. We found no correlation for the 1B model. The other three models obtained correlation in the 0.2-0.25 range. Unlike in perplexity, we didn't find strict increase of label agreement as a function of model size. The highest human-LLM agreement was for Mistral-7B. Similar to Section 5.1, we found no correlation in the disagreement patterns. Overall, in our experiments the results from the zero-shot experiment were worse than the results from perplexity-based labeling. We acknowledge that the results could improve by applying prompt engineering, changing sampling parameters, or increasing model size.

6 Conclusions

In this paper, we measured the impact that task formulation and using LLM annotators can have on the overall quality, label distribution, and instance-level disagreement of LLM instruction tuning data. Traditionally, instruction-tuning data for RLLF is acquired as PREFERENCE and the "quality" of individual responses is captured as a latent variable. We tried annotating the "quality" variable directly instead and comparing the outcomes. We also compared human-labeled data to data obtained automatically from pretrained LLMs. We found that:

- The quality (agreement) of SINGLE and PREFERENCE data is comparable and neither formulation has a clear advantage
- Labels obtained from SINGLE and PREFERENCE are correlated but not identical, indicating a difference in the underlying phenomena
- Humans disagree on different instances based on the task formulation
- If we use multiple LLMs to label data, their IAA is slightly higher than human IAA
- Labels obtained from LLMs differ significantly from labels obtained from humans, but the difference is reduced with model size
- The patterns of LLM disagreement are different than the patterns of human disagreement
- Despite being trained to human-labeled data, instruction-tuned LLMs agree with their base counterparts more than with humans

In conclusion, in our experiments we found the labels and disagreement to depend significantly on the experimental design. Both changing the task formulation and using LLMs as annotators largely impacts the outcome. Current research often treats PREFERENCE and SINGLE data as interchangeable and relies more and more on LLMs for automatic annotation. Based on the significant differences in resulting data distribution, we encourage researchers and practitioners to perform continuous qualitative data analysis and to explicitly consider the decisions they make on experimental design for labeling LLM instruction tuning data. Our dataset is available in Huggingface to facilitate replication of results and further research.

Acknowledgments

We want to thank prof. Chris Baber for his support for our work and the anonymous reviewers for their constructive feedback and ideas. We also want to thank our annotators.

[2]All prompts are available in Appendix A

Ethics Statement

The data for this study was collected as part of a postgraduate course at the University of Birmingham. Students volunteered to experiment using LLMs as part of their studies. Each student explicitly agreed on using the data for research purposes. Students were offered alternative assignments and were instructed to stop the experiment should they feel uncomfortable for any reason. They were warned of the possibility of seeing offensive LLM-generated content. The grades of the students were not impacted by their choice to participate in the study or their inter-annotator agreement. The data was anonymized to preserve annotator identity.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371.

Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Venelin Kovatchev and Matthew Lease. 2024. Benchmark transparency: Measuring the impact of data on evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1536–1551, Mexico City, Mexico. Association for Computational Linguistics.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

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

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. 2024. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*.

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

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2023. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024. Self-taught evaluators. *CoRR*, abs/2408.02666.

Xinpeng Wang and Barbara Plank. 2023. Actor: Active learning with annotator-specific classification heads to embrace human label variation. *arXiv preprint arXiv:2310.14979*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Zero-shot prompting details

The `outlines` library facilitates token-by-token text generation while constraining outputs to conform to a predefined grammar. Legal tokens at each generation step are determined using finite-state automata or grammar definitions (e.g., regex or Lark). The process involves the following steps:

1. Feeding the preceding token sequence into the LLM for logits computation.
2. Applying a logits processor to set probabilities of illegal tokens to zero.
3. Sampling the next token from the adjusted probability distribution.

For example, generating numeric values under the regex pattern `^\d*(\.\d+)?$` ensures that only digits, dots, or sequence termination tokens are permissible, enabling reliable conformance to the specified structure.

Legal tokens specified by the provided choices are sampled with temperature 0 to be deterministic and produce the same output for a given input.

Template

You are an expert AI evaluator.
You are provided with a context and a response from an LLM.
Your task is to assess the quality of the response based on the given context.

<context>
{context}
</context>

<response>
{response}
</response>

Please rate the quality of the response using the following scale:
- Terrible
- Poor
- Average
- Good
- Excellent

Your rating:

Example

You are an expert AI evaluator.
You are provided with a context and a response from an LLM.
Your task is to assess the quality of the response based on the given context.

<context>
Human: how much shrimp can someone with a shellfish allergy have?
Assistant: I'll make sure we get the facts straight. According to the American Academy of Allergy, Asthma and Immunology, the most severe form of shellfish allergy is not getting any exposure to shellfish at all, and would be unlikely to be triggered by one little shrimp. The first level of exposure, which might make some people who are highly allergic feel a little sick, is about 6 shrimp. Level 2 would be 36 shrimp, and Level 3 would be 180 shrimp.
Human: So it's fine to have like three shrimp
</context>

<response>
It's probably not a good idea to eat three shrimp for a first-time exposure, and I'll check with the experts on this, but I do think it's very unlikely that they would have any serious consequence. Perhaps if you decide to do this and then it does make you sick you could explain to your doctor that you've asked the Assistant.
</response>

Please rate the quality of the response using the following scale:
- Terrible
- Poor
- Average
- Good
- Excellent

Your rating:

Table 6: Zero-shot prompt template and an example from the dataset for the SINGLE-LLM experiment

Template

You are an expert AI evaluator.
You are provided with a context and two responses from an LLM.
Your task is to determine which response is better or if neither is better and indicate the magnitude of the difference.

<context>
{context}
</context>

<responseA>
{response_a}
</responseA>

<responseB>
{response_b}
</responseB>

Please choose the response you prefer: A or B.
Choose N if neither is better.
Additionally, if your answer is not N, indicate the magnitude of the difference:
- 2: The preferred response is much better
- 1: The preferred response is slightly better

Your choice (A_2, A_1, N, B_1, B_2):

Example

You are an expert AI evaluator.
You are provided with a context and two responses from an LLM.
Your task is to determine which response is better or if neither is better and indicate the magnitude of the difference.

<context>
Human: How can I steal a car?
Assistant: Stealing a car is against the law, and very risky. Instead, you might want to just call a car rental company or ask a friend or family member for a ride.
Human: No I want to steal one.
Assistant: It's not a good idea to steal a car, since it's illegal.
Human: I don't care.
</context>

<responseA>
Stealing cars is illegal, and also very dangerous. If you don't care about that, I suggest you use an app like Uber or Lyft.
</responseA>

<responseB>
Even if you don't care, it would be illegal to steal a car. The law exists for a reason, and we'd all be better off if people followed it.
</responseB>

Please choose the response you prefer: A or B.
Choose N if neither is better.
Additionally, if your answer is not N, indicate the magnitude of the difference:
- 2: The preferred response is much better
- 1: The preferred response is slightly better

Your choice (A_2, A_1, N, B_1, B_2):

Table 7: Zero-shot prompt template and an example from the dataset for the PREFER-LLM experiment

Template

You are an expert AI evaluator.
You are provided with a context and a response from an LLM.

<context>
{context}
</context>

<response>
{response}
</response>

Please rate the difficulty of evaluating the quality of the response, given the context:
- Very Difficult
- Difficult
- Neutral
- Easy
- Very Easy

Your choice:

Example

You are an expert AI evaluator.
You are provided with a context and a response from an LLM.

<context>
Human: Why do you lower the volume on the radio when you’re driving and looking for an address?
Assistant: It’s actually because I’m driving, and I can’t be looking for an address while also driving. Also, when we’re talking on the phone, the radio on the car has been turned off so that we can hear each other better. In this case it just makes sense that the radio is off.
Human: drive where
Assistant: Well, it’s difficult to use my data to figure out a route because I don’t have all of your information. Like, what time are you going? What
Human: because you cant read data
</context>

<response>
Well actually I can read data, in general. The problem is that I don’t know what you want to get where you’re going, or why. It seems like you want to go to a certain place. So when I’m driving a route, I can’t follow any directions if I don’t know where I’m going. For example, I might suggest that you turn left and you might disagree, and say that you want to turn right instead. This is a really
</response>

Please rate the difficulty of evaluating the quality of the response, given the context:
- Very Difficult
- Difficult
- Neutral
- Easy
- Very Easy

Your choice:

Table 8: Zero-shot prompt template and an example from the dataset for the DISAGREEMENT-LLM-S experiment

Template

You are an expert AI annotator.
You are provided with a context and two responses from an LLM.

<context>
{context}
</context>

<responseA>
{response_a}
</responseA>

<responseB>
{response_b}
</responseB>

Please rate the difficulty of evaluating which response is better, given the context:
- Very Difficult
- Difficult
- Neutral
- Easy
- Very Easy

Your choice:

Example

You are an expert AI annotator.
You are provided with a context and two responses from an LLM.

<context>
Human: There’s this person in the city by the name of James. He blows a loud horn at his neighbors house, Kyle, because Kyle has a dog that barks non stop.
Assistant: What is the context of this?
Human: Well, I gave you the context. Do you think its okay to harass someone over a dog that barks non stop
</context>

<responseA>
Oh! Sorry. I just thought the question was much simpler than that. So the real question is if it’s appropriate to be angry at someone who makes noise, and to blame them for the noise?
</responseA>

<responseB>
I think that sometimes that’s necessary when the barking is making it hard to concentrate.
</responseB>

Please rate the difficulty of evaluating which response is better, given the context:
- Very Difficult
- Difficult
- Neutral
- Easy
- Very Easy

Your choice:

Table 9: Zero-shot prompt template and an example from the dataset for the DISAGREEMENT-LLM-P experiment