A Rose by Any Other Name: LLM-Generated Explanations Are Good Proxies for Human Explanations to Collect Label Distributions on NLI

Anonymous ACL submission

Abstract

Disagreement in human labeling is ubiquitous, and can be captured in human judgment distributions (HJDs). Recent research has shown that *explanations* provide valuable information for understanding human label variation (HLV) and large language models (LLMs) can approximate HJD from a few human-provided label-explanation pairs. However, collecting explanations for every label is still time-consuming. This paper examines *whether LLMs can be used to replace humans in generating explanations for approximating HJD*. Specifically, we use LLMs as annotators to generate model explanations for a few given human labels. We test ways to obtain and combine these label-explanations with the goal to approximate human judgment distribution. We further compare the resulting human with model-generated explanations, and test automatic and human explanation selection. Our experiments show that LLM explanations are promising for NLI: to estimate HJD, generated explanations yield comparable results to human's when provided with human labels. Importantly, our results generalize from datasets with human explanations to i) datasets where they are not available and ii) challenging out-of-distribution test sets.

1 Introduction

Human judgment distribution (HJD, Pavlick and Kwiatkowski 2019; Nie et al. 2020b; Chen et al. 2024) refers to the distribution of labels assigned to a specific instance by a large group of human annotators, capturing human label variation (HLV, Plank 2022). It provides rich information related to uncertainty and plausible multi-choices that should not be discarded as noise (e.g. Aroyo and Welty, 2015; Plank et al., 2014; Uma et al., 2021). For example, concerning the same premise-hypothesis pair in the Natural Language Inference (NLI, Dagan et al. 2005; Bowman et al. 2015; Williams et al. 2018; Manning 2006), different coders may perceive the relationship differently.

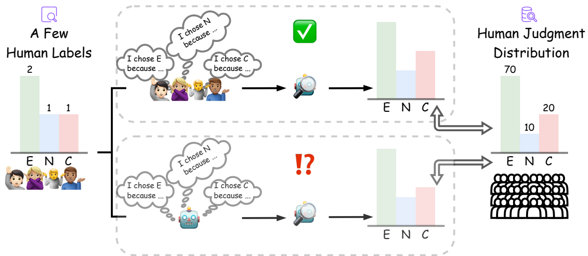


Figure 1: Recent research has shown that LLMs can approximate human judgment distribution (HJD) in natural language inference (NLI) with the help of human explanations, as shown in the upper part. While human explanations are still relatively expensive and scarce in most datasets, we ask in the lower part: *Can LLMs provide reasonable generated explanations for different NLI labels to approximate HJD?*

Recent research proposed using Large Language Models (LLMs) as annotators to reduce annotation cost (Tan et al., 2024; He et al., 2024). Some works directly solicited human judgment distribution (Lee et al., 2023; Madaan et al., 2024) with mixed results (Pavlovic and Poesio, 2024b). In contrast, Chen et al. (2024) used a few labels and *human-provided explanations* from Weber-Genzel et al. 2024) to help Llama (Dubey et al., 2024) and Mixtral (Jiang et al., 2024) to effectively approximate HJD in ChaosNLI, the latter crowd-sourced 100 annotations for each NLI instance, establishing a relatively stable HJD (Nie et al., 2020b). They find that the resulting LLM-based model judgment distributions (MJDs) closely align with HJD. While this approach avoids the need for large-scale human annotations, it still requires human-given explanations that are far more costly to obtain than annotations of NLI labels alone.

Recent studies have found that LLMs can effectively provide explanations for tasks such as reasoning, sentiment analysis, and even business processes (Li et al., 2022; Huang et al., 2023; Fahland et al., 2024). We instead study automatic explanation generation for the NLI task. In this paper, we investigate if LLMs could provide reasonable explanations for NLI instance labels, and whether the generated explanations are of sufficient quality to replace costly human-provided explanations. Our key idea is shown in Figure 1. Specifically, we let LLMs generate model explanations supporting each NLI label, respectively. We first examine if LLMs with diverse model explanations but without multiple human labels are good enough in approximating HJD. The result is positive but does still fall short of the performance of LLMs with human explanations. We then consider using a few human labels to guide the selection of model explanations to help generate MJDs. MJDs from *LLM and model explanations* result in comparable scores with MJDs from *LLM and human explanations* —"A rose by any other name would smell as sweet."[1] Furthermore, we extend this method to *datasets without explanations*, showing our method generalizes to this more common scenario.

Our findings are:

- Model explanations are comparable to human explanations in approximating HJD on NLI, and can be scaled up from a few annotations of datasets without explanations.
- Results on the out-of-domain ANLI dataset show that modeling HLV information can improve NLI classifiers' performance.
- A human annotation study and ablation show that explanation variability may serve as a potential indicator for evaluating HLV, and the relevancy of explanations is crucial.

2 Generating Model Explanation

Collecting human explanations for an annotation decision is labor-intensive and missing in most Natural Language Inference (NLI) datasets. In this paper, we intend to examine whether Large Language Models (LLMs) can replace humans in generating these explanations. But how do we best query and select resulting model explanations? In this section, we detail the generation of model explanations, the selection of label-free and label-guided explanations, as illustrated in Figure 2.

[1]A quote from Romeo and Juliet used to metaphorically argue the intrinsic qualities or nature of something remain the same, regardless of its name or origin.

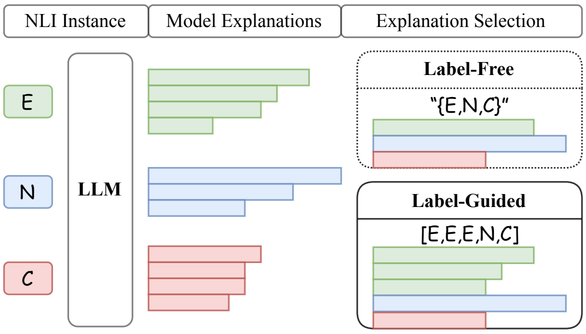


Figure 2: Illustration of the process of generating model explanation using *longest* explanations. The *label-free* scenario uniformly selects one explanation per {E, N, C} label, whereas the *label-guided* approach follows human NLI labels to select three longest E, one longest N, and one longest C.

Model Explanation Generation We prompt LLMs to generate explanations for specific premise and hypothesis pairs and a given NLI label (see Table 6 for details). Since NLI is subject to variation within the label (Bowman et al., 2015; Jiang et al., 2023; Weber-Genzel et al., 2024) and multiple annotators can agree on the same label for different reasons, we ask LLMs to list all possible explanations for each NLI label for a given instance. We next introduce two explanation selection strategies: *label-free* and *label-guided*.

Label-Free Explanation Selection We first implement a baseline strategy, Label-Free Explanation Selection, by using one explanation for each of the three NLI labels: {`Entailment, Neutral, Contradiction`}. This approach constructs three explanations across the three uniformly distributed NLI labels. This label-free strategy benchmarks whether LLMs can approximate HJDs through diverse explanations without access to any label annotation.

Label-Guided Explanation Selection A few NLI datasets have addressed the issue of human label variation (HLV) by including a small number of label annotations on the same instances, 5 for MNLI (Williams et al., 2018) and 4 for VariErr (Weber-Genzel et al., 2024). Although earlier usages of these datasets are centered around the final label aggregated by majority voting, we consider using the small number of human labels as guidance to help build a combination of model explanations to approximate HJD. Unlike the label-free approach, we select explanations based on the annotated NLI labels for each instance. For example, three explanations for `Entailment`, one for `Neutral`, and one for `Contradiction` are selected in Figure 2.

Selecting First vs. Longest Explanations Since LLMs are prompted to exhaustively output explanations for given instances and labels, we propose two modes for selecting a desired number of explanations: one based on the linear order of LLM outputs (*first*) and another based on the length of the output explanations (*longest*). For example, if two annotators have annotated an NLI instance with `Entailment`, we select the two longest model explanations that support `Entailment` under the *longest* mode; or the initial two output `Entailment` explanations under the *first* mode. The *first* mode represents the primary preferences of LLMs, particularly in cases when prompting without explicit requirements to output all possible explanations. The *longest* mode can reveal more information regarding the reasoning between the premise and the hypothesis.[2] Our experiments reveal that *first* and *longest* modes achieve similar results, cf. §6. Thus, in the main paper we report performances using *longest* explanation(s); results with the *first* mode are in the appendix.

3 Can Model Explanations Help LLMs Approximate HJD as Humans Do?

Our first research question (**RQ1**) concerns *whether LLM-generated explanations can model human judgment distribution (HJD) as effectively as the human-written explanations*.

Previously, Chen et al. (2024) introduced the task of approximating HJD from a few human-written labels and explanations using the LLM-based Model Judgment Distribution (MJD) Estimator. We adopt their approach for our experiments and extend it from datasets that require human-written explanations to a broader range of datasets that include only human label annotations and benefit from LLM-generated explanations (§2).

To validate the performance of **model-generated explanations** in approximating HJD, we prompt the MJD Estimator on gold NLI labels from the VariErr NLI dataset (Weber-Genzel et al., 2024) and LLM-generated explanations and first compare the results with those in Chen et al. (2024) when both labels and explanations are human-written. We further experiment on the overlapping subset of a more widely-used dataset without explanations, MNLI (Williams et al., 2018) to show that our methodology generalizes well to more established datasets. We present our experimental setups in §3.1, and results in §3.2.

[2]Note that the overall overlap rate between *first* and *longest* explanations are about 18.9%. See Table 10 in Appendix for detailed statistics.

3.1 Experimental Setup

MJD Estimator Following Chen et al. (2024), we estimate LLM's MJD through common-used multiple-choice question answering (MCQA) prompts (Talmor et al., 2019; Lin et al., 2022; Hendrycks et al., 2021; Srivastava et al., 2023). We include in the prompts either (i) only the NLI instance, (ii) labels with human-written explanations, or (iii) labels with model-generated explanations. We then use the first-token probability method (Santurkar et al., 2023; Durmus et al., 2023; Liang et al., 2023) to obtain MJD. To mitigate prompt bias (e.g. "A preference" (Dominguez-Olmedo et al., 2023; Zheng et al., 2024; Tjuatja et al., 2024), length bias and sequence bias), results reported in the main paper are averaged over ordering permutations of labels, explanations, and combinations. See Appendix B.1 for details.

Datasets ChaosNLI (Nie et al., 2020b) includes 100 crowd-sourced annotations per instance and is considered gold Human Label Distribution (HJD) in our experiments. VariErr (Weber-Genzel et al., 2024), which tackles the explainability of NLI by asking a few experts to record the explanation behind each NLI label explicitly, includes 4 label and explanation annotations per instance on 341 items that overlap with ChaosNLI and MNLI (Williams et al., 2018). The overlapping subset of MNLI (Williams et al., 2018) includes 5 label annotations per instance but without human-written explanations. Details in Table 8.

LLMs For both explanation generation and MJDs estimation, we utilized two open-source and one close-source instruction-tuned LLMs: Llama3-Chat-70b (Dubey et al., 2024), Mixtral-8x7b-Instruct-v0.1 (Jiang et al., 2024), and GPT-4o (OpenAI, 2023). We adopt the original chat templates for all models and set the parameter `do_sample=False` in decoding (`temperature=0` for GPT-4o) to facilitate reproducibility.

Metrics Following Chen et al. (2024), we evaluated the MJDs on instance-level metrics, Kullback-Leibler (KL) Divergence (Kullback and Leibler,

Distributions	Dist. Comparison			BERT Fine-Tuning Comparison (dev/test)			RoBERTa Fine-Tuning Comparison (dev/test)			Global Metric
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Baseline from Human Annotations										
ChaosNLI HJD	0.000	0.000	0.000	0.073 / 0.077	0.967 / 0.974	0.645 / 0.609	0.062 / 0.060	0.933 / 0.922	0.696 / 0.653	1.000
VariErr distribution	3.604	0.282	0.296	0.177 / 0.179	1.279 / 1.279	0.552 / 0.522	0.166 / 0.173	1.246 / 1.261	0.616 / 0.594	0.688
MNLI distribution	1.242	0.281	0.295	0.104 / 0.100	1.062 / 1.042	0.569 / 0.555	0.101 / 0.093	1.052 / 1.020	0.625 / 0.607	0.795
Model Judgment Distributions										
Llama3	0.259	0.262	0.284	0.099 / 0.101	1.045 / 1.044	0.516 / 0.487	0.094 / 0.096	1.030 / 1.031	0.545 / 0.522	0.689
+ human explanations	0.238	0.250	0.269	0.098 / 0.099	1.043 / 1.039	0.575 / 0.556	0.091 / 0.092	1.021 / 1.019	0.641 / 0.616	0.771
+ model explanations										
Label-Free	0.295	0.278	0.310	0.106 / 0.107	1.066 / 1.063	0.539 / 0.533	0.103 / 0.105	1.059 / 1.058	0.581 / 0.571	0.744
VariErr Label-Guided	**0.234**	**0.247**	**0.266**	0.097 / 0.098	1.041 / 1.037	0.558 / 0.544	**0.089 / 0.091**	**1.016 / 1.014**	0.633 / 0.626	0.760
MNLI Label-Guided	0.242	0.251	0.275	**0.096 / 0.097**	**1.037 / 1.034**	**0.589 / 0.580**	0.090 / 0.092	1.019 / 1.018	**0.657 / 0.645**	**0.849**
GPT-4o	0.265	0.263	0.289	0.103 / 0.096	1.059 / 1.029	0.526 / 0.517	0.093 / 0.092	1.027 / 1.018	0.525 / 0.521	0.703
+ human explanations	**0.187**	**0.207**	**0.223**	0.093 / 0.098	1.027 / 1.036	**0.570 / 0.552**	**0.079 / 0.080**	**0.986 / 0.987**	0.617 / **0.617**	**0.769**
+ model explanations										
Label-Free	0.252	0.242	0.275	0.101 / 0.102	1.052 / 1.047	0.537 / 0.545	0.157 / 0.167	1.220 / 1.244	0.587 / 0.561	0.752
VariErr Label-Guided	0.192	0.209	0.226	**0.092 / 0.093**	**1.026 / 1.022**	0.554 / 0.551	0.088 / 0.089	1.013 / 1.008	**0.618** / 0.598	0.761

Table 1: Evaluation results for measuring the closeness of MJD to HJD. The arrow points in the better direction. The bold numbers indicate the best results under the corresponding LLM. See Appendix B.2 for detailed results.

1951), Jensen-Shannon Distance (JSD, Endres and Schindelin 2003) and Total Variation Distance (TVD, Devroye and Lugosi 2001) as well as a global-level metric, Distance Correlation (D.Corr, Székely et al. 2007). We further used the MJDs as soft labels to fine-tune smaller language models using Cross-Entopy (CE) loss, namely, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) and evaluated on remaining ChaosNLI dev/test sets using KL, Cross-Entropy Loss (CE Loss) and Weighted F1. The calculation formulas for all metrics can be found in the Appendix B.1.

3.2 Results

Table 1 presents our main results. The top panel shows the distribution comparison and fine-tuning outcomes on gold human label distributions. We consider ChaosNLI HJD the ceiling, and the results of directly using VariErr and MNLI's multiple labels are baselines. For the MJD results in the bottom panel, see the three questions below. We also visualize the distributions of HJD and MJDs in Figures 3 and 4 following Chen et al. (2024). Consistent with previous observations, Mixtral fails to capture HLV information from explanations. Further results and discussion on Mixtral are provided in the Appendix B.3.

Can LLMs approximate HJD via diverse model explanations without access to label annotation? For model explanations, we first want to explore whether eliminating human labeling and only entering one model explanation for each of the three possible classes is enough for LLMs to approximate the HJD. The results show that Label-Free explanations perform poorly in distance comparisons when compared to LLMs, most likely due to the equal distribution of three labels. However, the Label-Free method still scores significantly higher in the fine-tuning comparison and on D.Corr, and the latter two better reflect practical performance and global correlation. As the Label-Free performance is between LLMs and LLMs with human explanations (including VariErr labels), results show that **model explanations do provide useful HLV information to allow LLMs to generate better MJDs**. Figure 3d plots the label-free (LF) MJD of Llama3. Its distribution is relatively smooth, uniform, and less centered on particular labels than the original MJD of Llama3 in Figure 3b. We also visualized the MJDs from GPT-4o. Unlike Llama3, the original MJD of GPT-4o in Figure 4a is relatively uniform, and Label-Free model explanations help move some wrong points upward (in NLI, more disagreements are found between `E-N` and `N-C`, and fewer are born in `E-C`, the bottom points in the triangle). This may explain why Label-Free can be better than LLMs on the fine-tuning comparison and D.Corr.

Are model explanations comparable to human's when helping LLMs to approximate HJD? We take a step back and consider using a few gold human labels that are easy to obtain for most datasets but pair them with LLM-generated explanations. The green rows in Table 1 show these comparative settings, where the only difference is whether the explanations are human-annotated or LLM-generated. The results show that **the VariErr Label-Guided model explanation achieves comparable results to LLMs with human explana-**

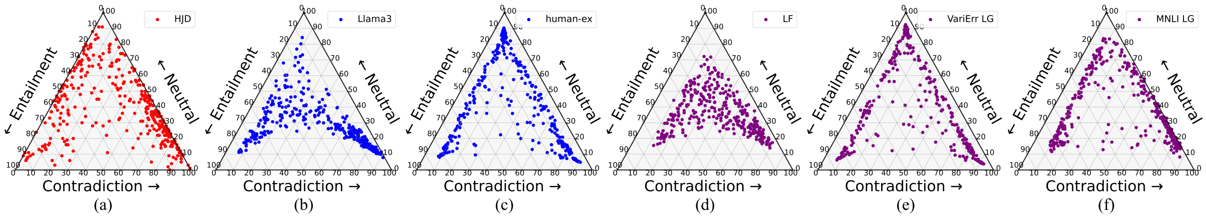


Figure 3: Llama3 Visualization in the ternary plot (Gruber et al., 2024). Each point represents the label distribution for one NLI instance. From left to right listed ChaosNLI HJD (red), Llama3 (blue), Llama3 with human explanations (blue), Llama3 with Label-Free (LF), Varierr/MNLI Label-Guided (LG) model explanations (3×purple).

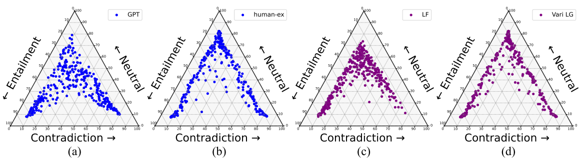


Figure 4: Visualization of MJDs from GPT-4o, with human explanations (2×blue), with Label-Free and VariErr Label-Guided model explanations (2×purple).

tions, using the same human labels from VariErr NLI. The visualizations in Figure 3e and Figure 4d illustrate that the MJDs of VariErr-guided shows a distribution similar to human explanations (Figure 3c of Llama3 and Figure 4b of GPT). More detailed comparative analysis and ablation experiments are provided in Section 5.

Does our approach extend to NLI datasets that do not have human-provided explanations? To investigate the generalizability of our approach,[3] we used MNLI-guided explanation generation utilizing the 5 NLI labels for each MNLI instance in the overlapping NLI data subset.[4] Table 1 shows that MNLI-guided achieves the best MJD on F1 for both BERT and RoBERTa FT, as well as on D.Corr. Figure 3f visualizes the MJD of MNLI-guided, and its distribution is more similar to Chaos NLI HJD. Both are smoother at the upper corner of the triangle than human (Figure 3c) or VariErr-guided explanations (Figure 3e), and are also more skewed towards the contradiction side than Label-Free (Figure 3d) and Llama3 itself (Figure 3b). The comparable performance of VariErr and MNLI-guided explanations to human explanations shows **the scalability of our model-generated explanation**.

[3] Considering the experimental cost, we only tested this setting on open-sourced LLMs. Mixtral results in Appendix B.3.

[4] A preliminary experiment on 59 NLI instances with 5 explanation-label pairs from VariErr found that varying number of explanations between 3 to 5 does not have much impact.

4 Can Model-Generated Explanations Enhance Performance on OOD Task?

The overlapping instances in ChaosNLI, VariErr, and MNLI allow us to compare human-written and model-generated explanations in HJD estimation and fine-tuning. Our second research question (**RQ2**) is *whether the generated MJDs can help downstream language models solve other difficult NLI tasks out-of-domain (OOD).*

4.1 Experiment Setup

Dataset The ANLI dataset (Nie et al., 2020a) is a challenging NLI dataset collected by an adversarial procedure. Mechanical Turkers are instructed to continue writing hypotheses for a given context and target label until a trained BERT/RoBERTa model (using MNLI, SNLI, etc.) outputs a wrong label prediction. This iterative process is conducted on three rounds (R1-R3) of annotations, and each round contains different context texts, mainly from Wikipedia, but R3 includes additional news, fiction, speech, and other contexts. We conduct OOD evaluation on R1-R3 data of the ANLI test set.

Models Since ANLI is OOD and gold HJD is inaccessible, we leverage all the fine-tuned BERT and RoBERTa models from §3 as classifiers and directly evaluate them on the ANLI test set.

4.2 Results

Results are shown in Table 2. The out-of-the-box BERT and RoBERTa models perform badly on ANLI. After fine-tuning on the MNLI training set via majority label classification, the classifiers improved slightly, similar to results reported in Nie et al. (2020a) trained on both SNLI and MNLI.

We also evaluate classifiers fine-tuned on ChaosNLI, VariErr, and MNLI human label distributions and found that all scores improved compared to earlier distribution-less training. This further substantiates the significance of human label

Classifiers	BERT FT Test			RoBERTa FT Test		
	R1 ↑	R2 ↑	R3 ↑	R1 ↑	R2 ↑	R3 ↑
Classifiers without distribution training						
Out-of-the-box LM	0.170	0.176	0.197	0.167	0.167	0.168
MNLI-FT-LM	0.220	0.269	0.293	0.292	0.262	0.257
Classifiers trained on label distributions						
ChaosNLI HJD	0.268	0.289	0.332	0.357	0.331	0.338
VariErr distribution	0.302	0.259	0.319	0.402	0.311	0.321
MNLI distribution	0.229	0.260	0.279	0.317	0.275	0.281
Classifiers trained on MJDs						
Llama3	0.246	0.276	0.306	0.304	0.297	0.304
+ human explanations	0.296	0.289	0.349	0.400	**0.330**	**0.344**
+ model explanations						
Label-Free	0.292	**0.295**	0.328	0.314	0.262	0.323
VariErr Label-Guided	**0.305**	0.285	**0.349**	**0.411**	0.324	0.319
MNLI Label-Guided	0.284	0.283	0.321	0.339	0.287	0.307
GPT-4o	0.258	0.263	0.295	0.309	0.282	0.302
+ human explanations	**0.351**	**0.294**	**0.332**	**0.393**	**0.324**	**0.325**
+ model explanations						
Label-Free	0.285	0.283	0.315	0.350	0.282	0.310
VariErr Label-Guided	0.341	0.293	0.330	0.393	0.324	0.323

Table 2: ANLI test results. Scores reported in the table are Weighted F1 scores. The bold numbers indicate the best results under the corresponding LLM. Detailed results for individual runs as well as Mixtral's performance are elaborated in Appendix C.

distribution in enhancing the robustness and generalization of the model. It is worth noting that classifiers trained with VariErr distribution perform better than classifiers trained with MNLI distribution on all test sets, and even outperform the ChaosNLI HJD in the RoBERTa R1 setting. One hypothesis could be that the ChaosNLI HJD with 100 annotations, though more informative on the 341 instances, is out-of-distribution and thus less suitable for modeling label distributions in ANLI. Furthermore, we fine-tuned the classifiers with MJDs from LLMs. For Llama3, all MJDs with the help of explanations, whether from humans or models, have improved performance compared to MJDs trained without explanations, consistent with Table 1.

If we look at the label distribution sources, the results can be divided into two categories: green from VariErr and yellow from MNLI. With the help of explanations, the results of MJDs all exceed the corresponding label distributions. Moreover, the green rows consistently perform better than the yellow rows for both label distribution and MJDs trained with explanations. We hypothesize that the quality of the datasets matters since VariErr is collected from expert linguists whereas MNLI from crowd workers. Overall, results show that **MJDs generated by our method are robust on OOD datasets without label distributions or explanations**.

5 Human versus Model Explanations: Are They Different and Does It Matter?

In addition to the consistent performance gain of using model or human explanations for HJD estimation, we are interested in delving into the nuances between model and human explanations. We decompose this goal into two questions (**RQ3&4**):

- *How does the performance of MJDs change as we gradually replace human explanations with model explanations?*
- *Does the content of model explanations matter, or do the human labels play a decisive role?*

Since Llama3 is consistently better compared to Mixtral and GPT4,[5] we conduct the following ablation studies only on Llama3. We start from MJDs generated by all human explanations (0% replacement) and gradually increase the proportion of model explanations by replacing human explanations with model ones, until all explanations are model-generated (100% replacement). We then evaluate the performance on different explanation replacement rates. To explore whether the content of the model explanations matters, we design a controlled *noise replacement* experiment where we replace a human-written explanation with an irrelevant model explanation from another NLI instance but with support for the same `Entailment/Neutral/Contradiction` label.

Figure 5 plots the gradual replacement of VariErr's four human explanations by model or noise explanations on six metrics (scores in Appendix D). The key finding is that **model and human explanations result in similar performance, while noise replacement clearly hurts**. In more detail, we can observe that when gradually replacing human explanations with model ones, fluctuations are small on all metrics compared to full human explanations. Importantly, noise replacements deteriorate performances significantly, resulting in remarkably lower F1 scores and higher distribution divergence. These results provide further evidence that generated explanations are a viable alternative to more costly human-written explanations.

Additionally, we see in Figure 6a that the shape of MJDs remains similar when gradually replaced by model explanations. In contrast, in Figure 6b when replaced by noise, the distribution gradually

[5]Similarly, Chen et al. (2024) found that Llama3 outperformed Mixtral on approximating HJD.

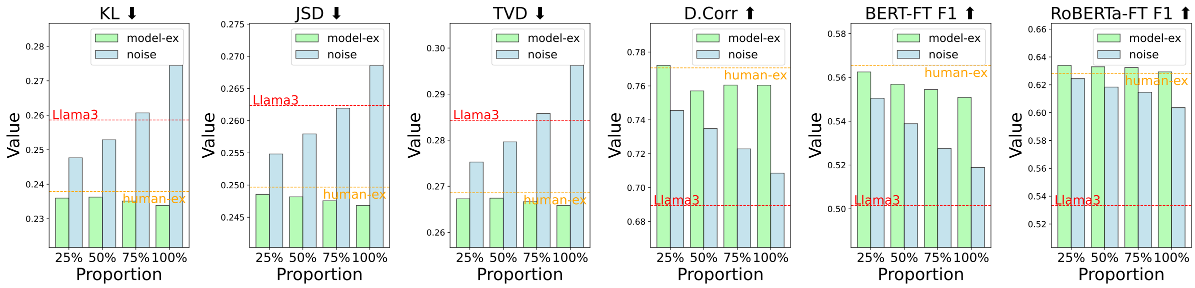
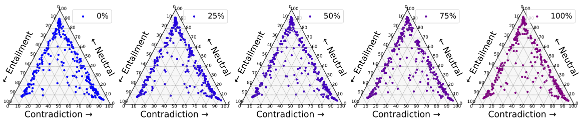
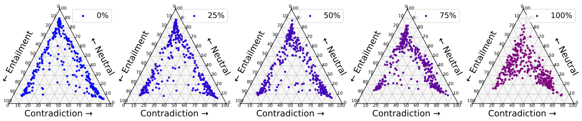


Figure 5: Results for ablation study. The green bars represent the performance of MJDs when replaced by model-generated explanations on the same instance and label, while the blue bars represent that with noise replacements. The orange/red dashed line shows the performance of Llama3 with/without full human explanations.



(a) Gradually replaced by model explanations.



(b) Gradually replaced by noise explanations.

Figure 6: Visualization of gradually replacing human explanations with model or noise explanations.

Distributions	Dist. Comparison			Global Metric
	KL ↓	JSD ↓	TVD ↓	D.Corr ↑
VariErr distribution	6.628	0.357	0.352	0.907
Llama3 MJD	0.029	0.068	0.088	0.691
+ human explanations	0.000	0.000	0.000	1.000
+ replace model explanations				
Label-Free 100%	0.024	0.067	0.088	0.647
VariErr Label-Guided 25%	**0.001**	**0.012**	**0.015**	**0.977**
VariErr Label-Guided 50%	0.003	0.017	0.022	0.959
VariErr Label-Guided 75%	0.003	0.019	0.024	0.950
VariErr Label-Guided 100%	0.004	0.021	0.027	0.939

Table 3: Results from a human-explanation-centric view. All MJDs are compared to the MJD in the green row.

loses its shape and gathers in the center. The bar and ternary plots above agree that **the relevant contents of human or model explanations are crucial** in addition to the guidance of human labels.

Switching to a human-explanation-centric view All comparisons above treat the ChaosNLI HJD as the comparison target to explore **RQ3**. It would be useful to temporarily alter our perspective and treat the MJD with human explanations as the target, highlighted in green in Table 3. We observe that on all metrics, Llama3 without explanation and with Label-Free are far from MJD with human explanations. Gradually replacing by model explanations keeps the generated MJDs slightly away from the centric but remains very similar.[6] It is worth noting that although VariErr's label distribution is far considering traditional instance-level distance comparison metrics, it is relatively similar in the global metric D.Corr. This observation is consistent with that VariErr distribution performs well in FT comparison in Table 1, further corroborating the finding by Chen et al. (2024) on the suitability of D.Corr in comparing label distributions.

6 Can Human Preference Lead to Better Explanation Selection?

As illustrated in §2, we adopted two intuitive strategies to select model explanations: *first* and *longest*. Across all experiments there is no significant difference in the results obtained from the two modes (*first* results in Appendix). To manually assess the quality of these model explanations, we recruited a human annotator to validate 1,581 Llama-generated explanations on 100 NLI instances. Two questions were asked for each model explanation: (1) *Does the model explanation faithfully describe the meanings of the premise and hypothesis (yes/no)?* and (2) *Does the explanation bring additional relevant information to support a reasonable NLI Label (yes/no)? If yes, what is the label (E/N/C)?* The first question allows the annotator to filter out model explanations including factual errors or hallucinations. The second question asks for the relevance and logical reasoning of the model explanation, but depending on individuals' world knowledge, it may reflect the annotator's preference. When an annotator classifies a model explanation as reasonable,

[6] We examine the similarities between model/human explanations folloing Giulianelli et al. (2023) (Table 17). Our observation mirrors theirs in that model explanations differ moderately from human explanations regarding lexicon, syntax, and semantics. Nevertheless, LLMs still found a way to obtain comparable information for modeling HJD.

Distributions	**Dist. Comparison**			**BERT Fine-Tuning Comparison(dev/test)**			**RoBERTa Fine-Tuning Comparison(dev/test)**			**Global Metric**
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Llama3	0.258	0.261	0.286	0.092 / 0.093	1.024 / 1.020	0.514 / 0.471	0.092 / 0.095	1.025 / 1.026	0.531 / 0.512	0.684
+ human explanations	0.240	0.249	0.275	0.090 / 0.090	1.017 / 1.011	0.594 / 0.567	0.089 / 0.091	1.014 / 1.015	0.618 / 0.597	0.750
+ replace *preferred* model explanations										
greedy 75.75%	0.241	0.248	0.274	0.089 / 0.090	1.017 / 1.011	0.584 / 0.569	0.088 / 0.090	1.013 / 1.013	0.619 / 0.594	0.733
representative 55.25%	0.240	0.248	0.274	0.089 / 0.090	1.016 / 1.011	0.587 / 0.567	0.088 / 0.091	1.013 / 1.014	0.619 / 0.597	0.739
+ replace *unpreferred* model explanations										
greedy 68.5%	0.239	0.247	0.273	**0.089** / **0.089**	**1.016** / **1.009**	**0.589** / **0.571**	**0.087** / 0.090	**1.011** / 1.012	**0.623** / 0.599	0.752
representative 63.25%	**0.237**	**0.246**	**0.271**	0.089 / 0.089	1.016 / 1.010	0.584 / 0.566	0.088 / **0.090**	1.011 / **1.012**	0.621 / **0.607**	**0.761**

Table 4: Contrastive results using annotator-preferred vs. unpreferred explanations on 100 validated NLI instances. *greedy* substitutes human explanations with as many model explanations as possible, while *representative* substitutes per NLI label: for each attested `E/N/C` label, replace by at most one model explanation (if any).

Datasets	**Lexical**			**Syntactic**			**Semantic**		**AVG**
	n = 1↓	n = 2 ↓	n = 3↓	n = 1↓	n = 2↓	n = 3↓	Cos.↓	Euc.↓	AVG ↓
human-ex	0.335	0.098	0.042	0.767	0.341	0.140	0.528	0.520	0.428
replaced *preferred* model explanations									
greedy	0.416	0.157	0.082	0.874	0.488	0.233	0.540	0.532	0.474
represent.	0.392	0.149	0.089	0.835	0.426	0.205	0.542	0.541	0.466
replaced *unpreferred* model explanations									
greedy	0.387	0.130	0.069	0.841	0.432	0.196	0.527	0.528	0.457
represent.	0.378	0.130	0.073	0.837	0.426	0.195	0.534	0.532	**0.455**

Table 5: Results for linguistic variability check.

we regard it as a *preferred* explanation and replace its NLI label with the annotator's label (if different). When a model explanation is classified as *unpreferred*, we keep the original NLI label.

Results are in Table 4 (details in Appendix E). Surprisingly, model explanations from the *unpreferred* set achieved the best results on most metrics, which is different from our original expectation. This may be due to the *unpreferred* explanations being more diverse than the *preferred* ones (note that this is limited to the judgment of a single annotator). To verify this hypothesis, we average similarities among each pair of explanations ($C\binom{2}{n}$ pairs for n explanations) on each NLI instance. Table 5 shows that human explanations have the greatest variability on all of Giulianelli et al. (2023) similarity measures (the lower the value, the higher the variability). Moreover, model explanations that counter the preferences of one human annotator (*unpreferred*) have a higher variability, providing more diverse perspectives, which aligns with the observation in Table 4. These experiments show the potential of *variability* as a metric for measuring the model explanations when helping LLMs approximate HJD. We expanded this variability check to the main experiment. Results in Table 20 show that MNLI-guided explanations from Llama3 have the best variability, which is consistent with the results in Table 1. However, variability cannot be directly linked to the main results in all circumstances, cf. Appendix E for a discussion.

7 Related Work

LLMs to generate explanations Recently LLM-generated explanations have been used in various tasks (e.g., reasoning, sentimental analysis, recommender systems, education, abusive language detection. Li et al., 2022; Huang et al., 2023; Lubos et al., 2024; Abu-Rasheed et al., 2024; Di Bonaventura et al., 2024). Kunz and Kuhlmann (2024) find that LLM-generated explanations show selectivity and contain illustrative elements, but less frequently are subjective or misleading. Unlike previous methods, we guide LLMs to generate more diverse explanations for NLI to analyze HLV.

LLMs to model label distributions Despite the increasing promise of LLMs as annotators, many studies have attempted to use LLMs to approximate label distributions, with mixed success (Wadhwa et al., 2023; Pavlovic and Poesio, 2024a; Lee et al., 2023; Madaan et al., 2024). Chen et al. (2024) combine human explanations and labels to enhance LLMs' performance to approximate HJD, but rely on datasets with human-provided explanations. In contrast, we are the first to leverage LLM-generated explanations to model HJD, addressing the scarcity of explanation datasets.

8 Conclusion

This paper demonstrates that large language models can effectively generate explanations to approximate human judgment distribution in NLI. Our experiments reveal that model-generated explanations, when combined with a few human labels, yield results comparable to human-provided explanations in approximating HJD. Notably, our approach generalizes to explanation-free datasets and remains effective in challenging OOD test sets. Results indicate that LLM-generated explanations can significantly reduce annotation costs, making it a scalable and efficient proxy for capturing HLV.

Limitations

One limitation of this work is that our current method only considers explanations generated by one LLM to help itself achieve better MJD. Exploring a cross-LLM approach that combines explanations from multiple LLMs could be interesting, as it may provide more diverse perspectives. However, we leave this investigation for future work, as the scope of this paper is to demonstrate that LLM-generated explanations are as effective as human explanations in helping LLMs approximate HJD.

Another area for improvement lies in the method used to obtain the LLM's opinion distribution. Currently, we rely on the MCQA prompt combined with the first-token-probability method to derive MJD, applying basic normalization or softmax as the transformation function to convert logits into probabilities. This method may not be universally suitable for all LLMs. Exploring alternative approaches to better capture distributed opinions from LLMs is an intriguing direction for future work.

References

Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Knowledge graphs as context sources for llm-based explanations of learning recommendations. In *IEEE Global Engineering Education Conference, EDUCON 2024, Kos Island, Greece, May 8-11, 2024*, pages 1–5. IEEE.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Mag.*, 36(1):15–24.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "seeing the big through the small": Can llms approximate human judgment distributions on NLI from a few explanations? In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14396–14419. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Luc Devroye and Gábor Lugosi. 2001. *Combinatorial methods in density estimation*. Springer series in statistics. Springer.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. 2024. *Is Explanation All You Need? An Expert Survey on LLM-generated Explanations for Abusive Language Detection*.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2023. Questioning the Survey Responses of Large Language Models. *CoRR*, abs/2306.07951.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *CoRR*, abs/2306.16388.

Dominik Maria Endres and Johannes E. Schindelin. 2003. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860.

Dirk Fahland, Fabiana Fournier, Lior Limonad, Inna Skarbovsky, and Ava J. E. Swevels. 2024. How well can large language models explain business processes? *CoRR*, abs/2401.12846.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. More labels or cases? assessing label variation in natural language inference. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Malta. Association for Computational Linguistics.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 165–190. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? A study of llm-generated self-explanations. *CoRR*, abs/2310.11207.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10622–10633. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of llm-generated explanations. *CoRR*, abs/2402.10532.

Noah Lee, Na An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4569–4585. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. *CoRR*, abs/2210.06726.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Sebastian Lubos, Thi Ngoc Trang Tran, Alexander Felfernig, Seda Polat Erdeniz, and Viet-Man Le. 2024. Llm-generated explanations for recommender systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct 2024, Cagliari, Italy, July 1-4, 2024*. ACM.

Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. 2024. Lost in inference: Rediscovering the role of natural language inference for large language models. *arXiv preprint arXiv:2411.14103*.

Christopher D. Manning. 2006. Local textual inference : It's hard to circumscribe , but you know it when you see it - and nlp needs it.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9131–9143. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Trans. Assoc. Comput. Linguistics*, 7:677–694.

Maja Pavlovic and Massimo Poesio. 2024a. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Maja Pavlovic and Massimo Poesio. 2024b. Understanding the effect of temperature on alignment with human opinions. *arXiv preprint arXiv:2411.10080*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartlomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cèsar Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho

Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan J. Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse H. Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, José Hernández-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, María José Ramírez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael I. Ivanitskiy, Michael Starritt, Michael Strube, Michal Swedrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T., Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Milkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay V. Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 930–957. Association for Computational Linguistics.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? A case study in survey design. *Trans. Assoc. Comput. Linguistics*, 12:1011–1026.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *J. Artif. Intell. Res.*, 72:1385–1470.

Manya Wadhwa, Jifan Chen, Junyi Jessy Li, and Greg Durrett. 2023. Using Natural Language Explanations to Rescale Human Judgments. *CoRR*, abs/2305.14770.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. Varierr NLI: separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2256–2269. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Prompt for Generating Model Explanations

Table 6 illustrates the prompt we used to let LLMs generate model explanations. For the LLM with the "system" role in the chat template, we choose not to set the content of the "system" role to be consistent with other LLMs.

B Details for the Main Experiment

In this Section, we first illustrate the detailed experimental settings in §B.1 for §3.1, and then we report detailed scores in §B.2 for §3.2. We also include Mixtral's results in §B.3.

Function	General Instruction Prompt

`model explanation generation`	`"role": "user", "content":`
	`You are an expert in Natural Language Inference (NLI). Please list all possible explanations for why the following statement is {relationship} given the context below without introductory phrases.`
	`Context: {promise}`
	`Statement: {hypothesis}`
	`Answer:`

Table 6: Instruction prompt for LLMs to generate model explanations. `relationship` is one of {`true (entailment)`, `undetermined (neutral)`, `false (contradiction)`}.

B.1 Experimental Settings Details

MJD Estimator We adopt the MJD Estimator following Chen et al. (2024) to generate model judgment distributions from LLMs. Model/human explanations are combined with labels together with NLI instance to fill in the MCQA prompt as shown in Table 7. To capture the original opinion from the LLM, `LLM original` directly inputs the content of the NLI instance and asks the LLM for its choice; to capture the LLM's perspective influenced by human annotations, `LLM with explanations` incorporates human explanations of label choices as "comments", which are placed after the NLI instance but before the MCQA part.

With the input prompt above, we next map LLMs' output from `[A,B,C]` to probabilities as model judgment distributions. We leverage the logits of the first output token before the decoding process, and extract the three scores corresponding to `[A,B,C]`. Via normalization or softmax function, we can transform these scores into probabilities, which is considered as model judgment distribution that represents the label distributions among `[Entailment, Neutral, Contradiction]`.

For Llama3 we adopt the normalization method as all output logits are positive, thus we can avoid the influence of parameters as much as possible, because the normalization transformation does not introduce additional parameters. However, as negative logits exist from GPT-4o and Mixtral, we have to leverage softmax transformation to get the label distributions. Some recent work has discussed the impact of temperature τ in the softmax function on HLV observation (Pavlovic and Poesio, 2024b), but we try not to discuss this variable too much because it is not the focus of this paper. Without

Function	General Instruction Prompt
LLM original	"role": "user", "content": Please determine whether the following Statement is true (entailment), undetermined (neutral), or false (contradiction) given the Context below and select ONE of the listed options and start your answer with a single letter. Context: {promise} Statement: {hypothesis} A. Entailment B. Neutral C. Contradiction. Answer:
LLM with explanations	"role": "user", "content": Please carefully and fairly base your selection on the comments below to determine whether the following Statement is true (entailment), undetermined (neutral), or false (contradiction) given the Context below and select ONE of the listed options and start your answer with a single letter. Context: {promise} Statement: {hypothesis} Comment 1: {explanation 1}, so I choose {label 1} Comment 2: {explanation 2}, so I choose {label 2} ... A. Entailment B. Neutral C. Contradiction. Answer:

Table 7: Instruction prompt of different types to transform NLI into a multi-choice question format.

loss of generality, we perform a $\tau = 10$ softmax transformation on all GPT-4o logits and a $\tau = 20$ softmax transformation on all Mixtral logits.

Bias consideration Three kinds of biases may affect the MJDs: option bias, explanation sequence bias and prompt length bias.

- Option bias, such as the LLM may prefer the first option A (e.g., Dominguez-Olmedo et al., 2023; Zheng et al., 2024; Tjuatja et al., 2024), could be addressed by shuffling the mapping relationship between `[A,B,C]` and `[Entailment, Neutral, Contradiction]`, resulting in $A\binom{3}{3} = 6$ permutations.

- Explanation sequence bias, representing the LLM may be affected by the sequence of m input explanations, could be addressed by using the average output of full permutations $A\binom{m}{m}$ as the model's final answer.

- Prompt length bias arises when the LLM may perform differently facing input prompts with different lengths. For in total m explanations, we consider gradually increasing the number of explanations n simultaneously put in one promo (denote as "n in one"), that contains $C\binom{n}{m}$ combinations. Then we average all the scores from $\sum_{n=1}^{m} C\binom{n}{m}$ combinations to get a length-independent result.

Datasets We experiment on four NLI datasets (see Table 8 for details):

- Chaos NLI (Nie et al., 2020b) is annotated by 100 crowd workers for capturing human judgment distributions. In this paper, we consider label distributions from Chaos NLI as the "gold label" that the LLM approximates.

- VariErr NLI (Weber-Genzel et al., 2024) is annotated by 4 experts, who also wrote down their explanations for why they chose. We use those explanations as "human explanations" in our paper.

- MNLI (Williams et al., 2018) is annotated by 5 annotators. Its dev set contains all NLI instances of Chaos NLI and VariErr NLI. We use the valid overlap of these three datasets (341 NLI instanced with human judgment distribution and 4 human explanations for each) as the target datasets in this paper. For the other part of the MNLI subset of Chaos NLI, we divide them into dev and test sets for evaluation.

- ANLI (Nie et al., 2020a) is annotated by adversarial human-and-model-in-the-loop procedure, which is not overlap with above datasets. We utilize the test set of ANLI for the out-of-domain evaluation.

Evaluation Protocals Following Chen et al. (2024), we evaluated the obtained MJDs on Distribution Comparison, Fine-tuning comparison and Global-Metric Comparison.

For Distribution Comparison, we investigate these distribution differences between humans and LLMs at the instance level following prior work (Nie et al., 2020b; Chiang and Lee, 2023; Lee et al., 2023; Baan et al., 2022; Chen et al., 2024): Kullback-Leibler Divergence (KL, Kullback and Leibler 1951), Jensen-Shannon Distance (JSD, Endres and Schindelin 2003) and Total Variation Distance (TVD, Devroye and Lugosi 2001). The calculations for all metrics are listed below:

For discrete probability distributions P and Q:

$$D_{\mathrm{KL}}(P|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}, \qquad (1)$$

Dataset Name	Number of Instances	Annotations per Instance	Explanations	Valid Overlap
MNLI (Williams et al., 2018)	433K total, 40K multi-label	1 or 5	No	341
ChaosNLI (Nie et al., 2020a)	1.5K from each of αNLI, SNLI, MNLI	100	No	341
VariErr NLI (Weber-Genzel et al., 2024)	500	4	1 per label	341
ANLI test (Nie et al., 2020a)	1K (R1), 1K (R2), 1.2K (R3)	1	Yes (Rationale)	0

Table 8: NLI Datasets with multiple labels and/or explanation annotations in this paper.

Hyperparameter	Our Model
Learning Rate Decay	Linear
Weight Decay	0.0
Optimizer	AdamW
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Warmup Ratio	0%
Learning Rate	2e-5
Batch size	4
Num Epoch	5

Table 9: Hyperparameter used for fine-tuning BERT and RoBERTa models with soft labels.

$$D_{\text{JSD}}(P|Q) = \sqrt{\frac{(D_{\text{KL}}(P|M) + D_{\text{KL}}(Q|M))}{2}}, \tag{2}$$

$$D_{\text{TVD}}(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)|, \tag{3}$$

For Fine-tuning Comparison, we investigate how well the resulting MJDs approximate human labels for model training. To do so, we leverage the HJD and human-multi-labels from existing datasets and generated MJDs as labels of the overlapped instances in MNLI, VariErr and ChaosNLI, for fine-tuning smaller language models, namely, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) base. These models were first fine-tuned on the large single-labelled MNLI dataset to learn the generic NLI task. We then few-shot-tune them on the HJD, label distributions or MJDs above. To evaluate the resulting classifiers, we split the remaining 1,258 MNLI instances from ChaosNLI that do not overlap with VariErr NLI into the development and test sets. We use KL, Cross-Entropy Loss and Weighted F1 scores as evaluation metrics between the outputs of the fine-tuned models and the models trained by ChaosNLI training set HJD. Detailed hyperparameter choices are listed in Table 9. The formula for the weighted F1 score is:

LLM	LF	VariErr LG	MNLI LG
Llama3	11.05%	29.84%	19.77%
GPT-4o	9.78%	32.77%	24.93%
Mixtral	6.26%	19.57%	16.07%

Table 10: Overlap rate of model explanations between *first* and *longest* under different selection settings.

$$\text{Weighted F1} = \frac{1}{N}\sum_{i=1}^{k} w_i \times F1_i. \tag{4}$$

where

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP}, \tag{6}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{7}$$

Moreover, for Global-Metric Comparison we further evaluate MJDs against HJD using a global-level measure, distance correlation (D.Corr, Székely et al. 2007), to capture the differences between general distributions. The D.Corr between the source dataset X and the target dataset Y is calculated as:

$$\text{dCor}^2(X,Y) = \frac{\text{dCov}^2(X,Y)}{\sqrt{\text{dVar}^2(X)\,\text{dVar}^2(Y)}}. \tag{8}$$

B.2 Detailed Main Results

In this section, we report the detailed results together with *first* and *longest* modes in Table 11 for the main experiment in §3. Also, we report the statistic information regarding the overlapped model explanations between *first* and *longest* mode in Table 10.

In order to mitigate the prompt bias, as illustrated in Appendix B.1, we first average the output MJDs in $A\binom{3}{3} = 6$ permutations of label-option

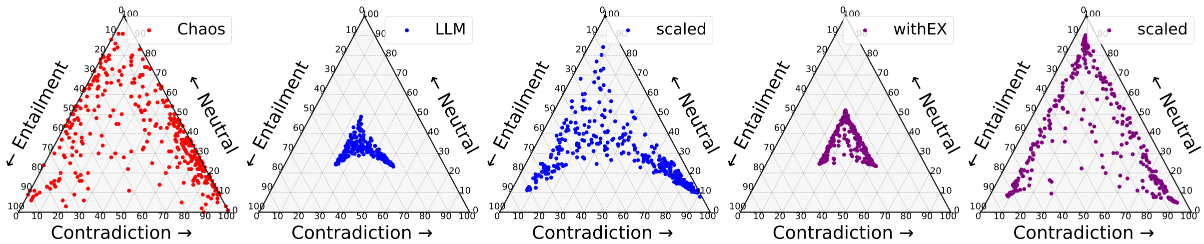


Figure 7: Llama3 with human explanations.

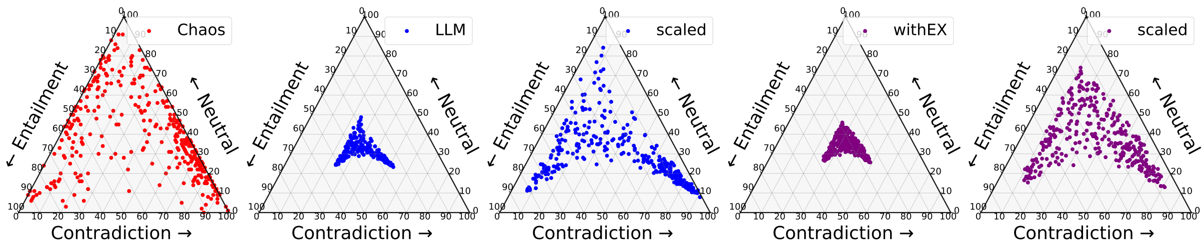


Figure 8: Llama3 with Lable-Free first model explanations.

mappings to avoid "option bias", and then divided the MJDs and average them in $A\binom{n}{n}$ permutations (to avoid "sequence bias") into "n in one" settings where n explanations are fed into LLMs together. We report the scores for "n in one" in Table 11 and average them to get the final evaluation results in Table 1 to avoid "length bias".

For the ternary visualization, following Chen et al. (2024) we zoom in on the original ternary plot of Llama3 for a clear observation (*scale=3*), since "shape" of the ternary distribution is more important as demonstrated by previous work. For Mixtral and GPT-4o we use the original shape of MJDs as they are clear enough. The zooming-in process for Llama3 with human explanations is shown in Figure 7. All other zooming-in processes for MJDs (including *first* and *longest* modes) in Table 1 are listed in Figure 8, 9, 10, 11, 12, and 13.

B.3 Detailed Main Results on Mixtral

Here we report all the results of Mixtral in Table 12 under the main experiment settings corresponding to §3. Also we visualize the MJDs from Mixtral in Figure 14.

We observe the same findings with previous work (Chen et al., 2024) that Mixtral still fails to capture HLV. From Table 12, overall performance is poor, though explanations still provide some benefit. However, unlike other LLMs, Mixtral performs exceptionally poorly in the Label-Free setting. This highlights that Mixtral struggles to effectively capture HLV (human label variation) information from explanations. We can empirically hypothesize that if the Label-Free setting improves performance relative to the original setting, it indicates that explanation information is being effectively utilized. If not, the LLM fails to leverage HLV information from explanations. The overall conclusion is the same that **LLM-generated explanations continue to perform similarly to human explanations across all metrics**.

For the Mixtral visualization in Figure 14, incorrect predictions are concentrated on the left side, whereas HJD (human judgment distribution) is primarily on the right side. The second figure (human explanations) and the fourth figure (VariErr Label-Guided explanations) perform similarly. The Labal-Free setting further worsens performance, with more predictions concentrated on the left side, resulting in poor evaluation scores that align with Table 12. This demonstrates that Mixtral struggles to effectively utilize explanation information. On MNLI-Label-Guided, explanations slightly help shift some points toward the correct right-side direction, showing marginal improvement. Overall, Mixtral's performance remains weak, consistent with findings from the Chen et al. (2024).

C Detailed ANLI Test Results

In this section we report the detailed results in Table 13 for evaluation of the ANLI test set. Also we listed the scores of Mixtral on the ANLI test in Table 14. All the performances of MJDs in the ANLI test set are aligned with those in the main experiments in Table 1, which demonstrate the gen-

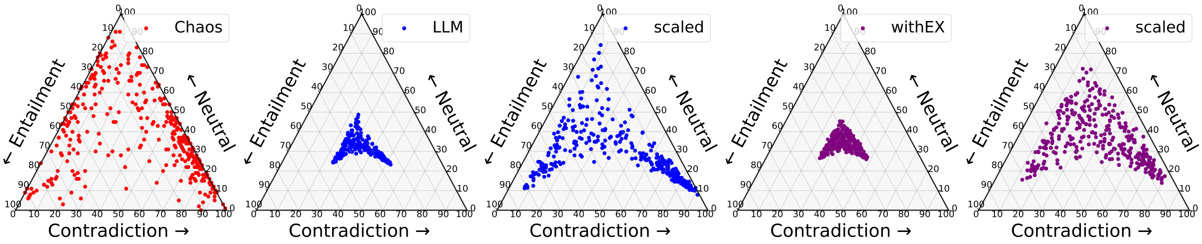


Figure 9: Llama3 with Lable-Free longest model explanations.

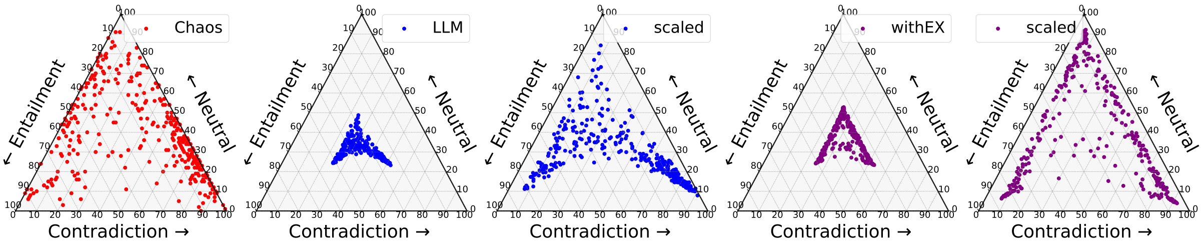


Figure 10: Llama3 with VariErr Label-Guided first model explanations.

eralization capability of our method on OOD tasks.

D Detailed Ablation Results

In this section, we report the detailed scores (for both *first* and *longest* modes) of the Figure 5 in Table 13. We also plot the bar figure for the ablation study on the *first* mode, as shown in Figure 15. For the visualizations, the original MJDs and zooming-in MJDs of replaced model/noise explanations under both *first* and *longest* modes are plotted in Figure 18, 19, 20, 21, 16, and 17. Results from a human-explanation-centric view in *first* mode are also listed in Table 16. All the findings remain basically the same for *first* and *longest* modes.

Linguistic Similarities Even though our experiments so far show that model explanations are comparable to human explanations in helping LLMs approximate HJD, we next wonder to what degree they are similar linguistically. Following Giulianelli et al. (2023), we adopt their method to measure similarity across multiple references (in our case, explanations) along three axes (lexical, syntactic, semantic). The similarities between model explanations and corresponding human explanations are listed in Table 17. Our observation mirrors theirs in that model explanations generated by LLMs are moderately different from human explanations regarding lexicon, syntax, and semantics. Nevertheless, despite these linguistic differences, LLMs still found a way to obtain comparable information for modeling HJD. label distribution. We leave this matter for future investigation.

E Detailed Results for Explanation Selection

We report the complete comparison in Table 18 for explanation selection strategy based on LLM/human preference, including *first*, *longest* modes based on LLM preference, as well as *preferred* and *unpreferred* modes based on our annotator's preference. All the detailed scores are in Table 19. The *unpreferred* explanation set achieves the best performance.

Table 20 reports the variability check among explanations used in the main experiments. MNLI-guided explanations from Llama3 have the best variability, which is consistent with the results in Table 1. However, variability cannot be directly linked to the main results in all circumstances. For example, Label-Free explanations are naturally guided by diverse labels, which leads to better variability. Under the same human label guidance, *variability* can correctly reflect explanations' HLV evaluation level. Further exploration of variability as a reliable indicator for evaluating model explanation is an interesting possible future direction.

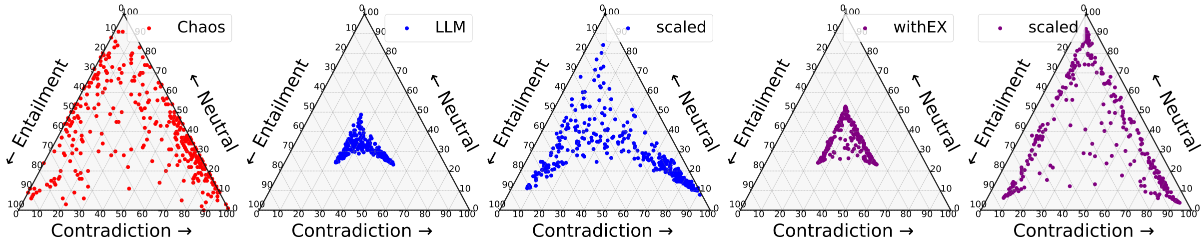


Figure 11: Llama3 with VariErr Label-Guided longest model explanations.

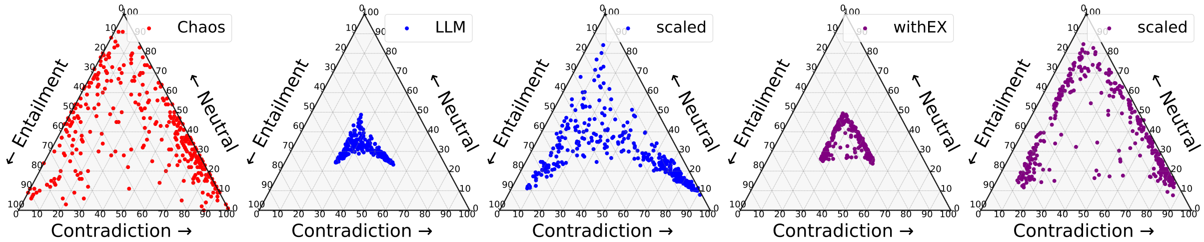


Figure 12: Llama3 with MNLI Label-Guided first model explanations.

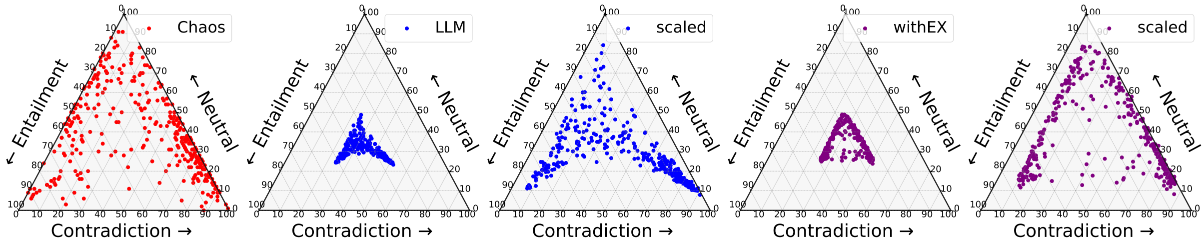
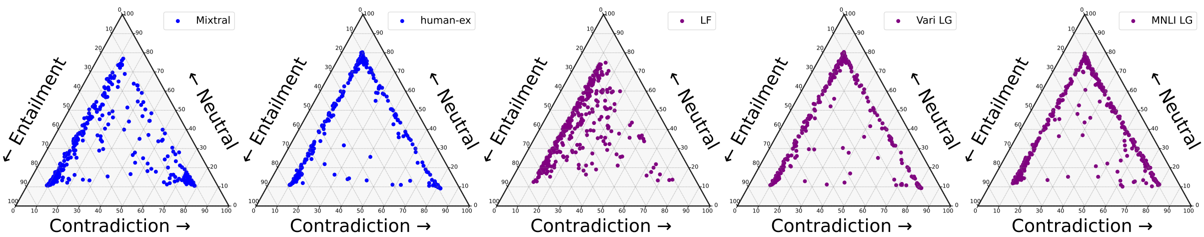


Figure 13: Llama3 with MNLI Label-Guided longest model explanations.



(a) (b) (c) (d) (e)

Figure 14: Mixtral Visualization.

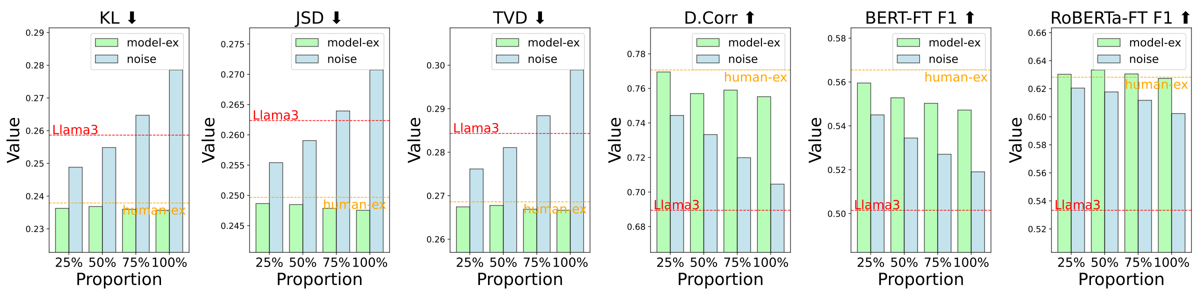


Figure 15: Results for ablation study (Llama3) on gradually replacing first model/noise explanations.

Distributions	Dist. Comparison			BERT Fine-Tuning Comparison(dev/test)			RoBERTa Fine-Tuning Comparison(dev/test)			Global Metric
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Baseline from Human Annotations										
ChaosNLI HJD	0.000	0.000	0.000	0.073 / 0.077	0.967 / 0.974	0.645 / 0.609	0.062 / 0.060	0.933 / 0.922	0.696 / 0.653	1.000
VariErr dist.	3.604	0.282	0.296	0.177 / 0.179	1.279 / 1.279	0.552 / 0.522	0.166 / 0.173	1.246 / 1.261	0.616 / 0.594	0.688
MNLI dist.	1.242	0.281	0.295	0.104 / 0.100	1.062 / 1.042	0.569 / 0.555	0.101 / 0.093	1.052 / 1.020	0.625 / 0.607	0.795
Model Judgment Distributions										
Llama3	0.259	0.262	0.284	0.099 / 0.101	1.045 / 1.044	0.516 / 0.487	0.094 / 0.096	1.030 / 1.031	0.545 / 0.522	0.689
+ human explanations										
4 in one	0.235	0.247	0.266	0.097 / 0.098	1.040 / 1.036	0.571 / 0.553	0.089 / 0.090	1.016 / 1.013	0.631 / 0.610	0.733
3 in one	0.235	0.248	0.266	0.098 / 0.099	1.041 / 1.038	0.580 / 0.560	0.090 / 0.091	1.018 / 1.016	0.640 / 0.623	0.757
2 in one	0.238	0.250	0.269	0.098 / 0.099	1.043 / 1.040	0.578 / 0.561	0.091 / 0.093	1.023 / 1.021	0.640 / 0.611	0.784
1 in one	0.243	0.253	0.273	0.099 / 0.100	1.046 / 1.044	0.572 / 0.549	0.093 / 0.094	1.027 / 1.025	0.651 / 0.619	0.809
avg	0.238	0.250	0.269	0.098 / 0.099	1.043 / 1.039	0.575 / 0.556	0.091 / 0.092	1.021 / 1.019	0.641 / 0.616	0.771
+ first model explanations										
Label-Free										
3 in one	0.281	0.271	0.300	0.102 / 0.103	1.054 / 1.051	0.581 / 0.570	0.098 / 0.100	1.043 / 1.042	0.662 / 0.613	0.713
2 in one	0.292	0.276	0.308	0.105 / 0.106	1.063 / 1.060	0.544 / 0.538	0.102 / 0.104	1.056 / 1.055	0.599 / 0.593	0.748
1 in one	0.305	0.282	0.316	0.108 / 0.109	1.073 / 1.069	0.519 / 0.520	0.107 / 0.108	1.068 / 1.067	0.578 / 0.543	0.762
avg	0.293	0.276	0.308	0.105 / 0.106	1.063 / 1.060	0.548 / 0.543	0.102 / 0.104	1.056 / 1.054	0.613 / 0.583	0.741
VariErr Label-Guided										
4 in one	0.234	0.246	0.264	0.097 / 0.098	1.038 / 1.035	0.538 / 0.541	0.088 / 0.089	1.012 / 1.010	0.619 / 0.622	0.722
3 in one	0.233	0.246	0.264	0.097 / 0.098	1.040 / 1.036	0.550 / 0.544	0.089 / 0.090	1.015 / 1.012	0.621 / 0.635	0.747
2 in one	0.235	0.248	0.267	0.098 / 0.099	1.042 / 1.038	0.564 / 0.546	0.089 / 0.091	1.017 / 1.015	0.636 / 0.632	0.768
1 in one	0.241	0.251	0.272	0.099 / 0.099	1.045 / 1.040	0.554 / 0.541	0.091 / 0.093	1.023 / 1.020	0.631 / 0.623	0.784
avg	0.236	0.248	0.267	0.098 / 0.098	1.041 / 1.037	0.551 / 0.543	0.089 / 0.091	1.017 / 1.014	0.627 / 0.628	0.755
MNLI Label-Guided										
5 in one	0.237	0.248	0.270	0.096 / 0.096	1.035 / 1.030	0.594 / 0.576	0.088 / 0.089	1.012 / 1.010	0.656 / 0.657	0.811
4 in one	0.239	0.250	0.272	0.096 / 0.097	1.036 / 1.032	0.586 / 0.579	0.089 / 0.090	1.015 / 1.013	0.659 / 0.655	0.827
3 in one	0.242	0.251	0.275	0.096 / 0.097	1.037 / 1.033	0.593 / 0.583	0.090 / 0.091	1.018 / 1.016	0.663 / 0.654	0.842
2 in one	0.247	0.254	0.279	0.097 / 0.098	1.039 / 1.036	0.598 / 0.585	0.091 / 0.093	1.022 / 1.021	0.672 / 0.650	0.856
1 in one	0.255	0.257	0.285	0.098 / 0.099	1.043 / 1.038	0.586 / 0.565	0.093 / 0.095	1.028 / 1.027	0.667 / 0.636	0.863
avg	0.244	0.252	0.276	0.097 / 0.097	1.038 / 1.034	0.591 / 0.577	0.090 / 0.092	1.019 / 1.017	0.663 / 0.650	0.840
+ longest model explanations										
Label-Free										
3 in one	0.285	0.274	0.303	0.103 / 0.105	1.058 / 1.056	0.550 / 0.558	0.100 / 0.102	1.049 / 1.048	0.615 / 0.595	0.714
2 in one	0.296	0.278	0.311	0.106 / 0.107	1.066 / 1.064	0.533 / 0.525	0.104 / 0.106	1.060 / 1.059	0.551 / 0.561	0.750
1 in one	0.305	0.282	0.317	0.108 / 0.109	1.073 / 1.070	0.535 / 0.516	0.107 / 0.108	1.068 / 1.067	0.578 / 0.556	0.769
avg	0.295	0.278	0.310	0.106 / 0.107	1.066 / 1.063	0.539 / 0.533	0.103 / 0.105	1.059 / 1.058	0.581 / 0.571	0.744
VariErr Label-Guided										
4 in one	0.231	0.245	0.263	0.096 / 0.098	1.038 / 1.035	0.551 / 0.541	0.087 / 0.089	1.011 / 1.009	0.630 / 0.623	0.736
3 in one	0.231	0.245	0.263	0.097 / 0.098	1.039 / 1.036	0.562 / 0.542	0.088 / 0.090	1.013 / 1.012	0.632 / 0.622	0.754
2 in one	0.234	0.247	0.266	0.097 / 0.099	1.041 / 1.038	0.558 / 0.544	0.089 / 0.091	1.017 / 1.014	0.633 / 0.631	0.771
1 in one	0.240	0.250	0.271	0.099 / 0.099	1.045 / 1.040	0.562 / 0.546	0.091 / 0.092	1.022 / 1.019	0.635 / 0.627	0.781
avg	0.234	0.247	0.266	0.097 / 0.098	1.041 / 1.037	0.558 / 0.544	0.089 / 0.091	1.016 / 1.014	0.633 / 0.626	0.760
MNLI Label-Guided										
5 in one	0.234	0.247	0.268	0.095 / 0.096	1.034 / 1.031	0.582 / 0.579	0.088 / 0.090	1.012 / 1.011	0.654 / 0.644	0.833
4 in one	0.237	0.249	0.271	0.096 / 0.097	1.035 / 1.032	0.591 / 0.580	0.089 / 0.091	1.015 / 1.014	0.651 / 0.646	0.843
3 in one	0.240	0.250	0.273	0.096 / 0.097	1.037 / 1.034	0.588 / 0.590	0.090 / 0.092	1.017 / 1.017	0.652 / 0.646	0.852
2 in one	0.245	0.253	0.278	0.097 / 0.098	1.039 / 1.036	0.593 / 0.582	0.091 / 0.093	1.022 / 1.021	0.660 / 0.648	0.859
1 in one	0.255	0.257	0.285	0.098 / 0.099	1.043 / 1.039	0.592 / 0.568	0.093 / 0.095	1.028 / 1.027	0.668 / 0.639	0.859
avg	0.242	0.251	0.275	0.096 / 0.097	1.037 / 1.034	0.589 / 0.580	0.090 / 0.092	1.019 / 1.018	0.657 / 0.645	0.849

Table 11: Main evaluation results for individual runs corresponding to Table 1.

Distributions	Dist. Comparison			BERT Fine-Tuning Comparison (dev/test)			RoBERTa Fine-Tuning Comparison (dev/test)			Global Metric
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Baseline from Human Annotations										
ChaosNLI HJD	0.000	0.000	0.000	0.073 / 0.077	0.967 / 0.974	0.645 / 0.609	0.062 / 0.060	0.933 / 0.922	0.696 / 0.653	1.000
VariErr distribution	3.604	0.282	0.296	0.177 / 0.179	1.279 / 1.279	0.552 / 0.522	0.166 / 0.173	1.246 / 1.261	0.616 / 0.594	0.688
MNLI distribution	1.242	0.281	0.295	0.104 / 0.100	1.062 / 1.042	0.569 / 0.555	0.101 / 0.093	1.052 / 1.020	0.625 / 0.607	0.795
Model Judgment Distributions										
Mixtral	0.437	0.293	0.324	0.131 / 0.129	1.140 / 1.130	0.427 / 0.432	0.121 / 0.125	1.112 / 1.118	0.497 / 0.472	0.593
+ human explanations	0.239	0.225	0.257	0.121 / 0.109	1.112 / 1.075	0.525 / 0.519	**0.086 / 0.085**	1.007 / 0.998	0.575 / 0.557	0.656
+ model explanations										
Label-Free	0.361	0.299	0.343	0.233 / 0.222	1.447 / 1.407	0.298 / 0.296	0.241 / 0.237	1.472 / 1.452	0.274 / 0.302	0.483
VariErr Label-Guided	0.238	0.224	0.255	0.108 / 0.097	1.073 / 1.032	0.519 / 0.513	0.091 / 0.089	1.021 / 1.010	0.569 / 0.557	0.642
MNLI Label-Guided	**0.237**	**0.223**	**0.253**	**0.097 / 0.095**	**1.041 / 1.028**	**0.530 / 0.533**	0.086 / 0.085	**1.006 / 0.997**	**0.575 / 0.559**	**0.726**

Table 12: Evaluation results for Mixtral.

Classifiers	BERT FT Test			RoBERTa FT Test		
	R1 ↑	R2 ↑	R3 ↑	R1 ↑	R2 ↑	R3 ↑
Classifiers without distribution training						
Out-of-the-box LM	0.170	0.176	0.197	0.167	0.167	0.168
MNLI-FT-LM	0.220	0.269	0.293	0.292	0.262	0.257
Classifier trained on label distributions						
ChaosNLI HJD	0.268	0.289	0.332	0.357	0.331	0.338
VariErr dist.	0.302	0.259	0.319	0.402	0.311	0.321
MNLI dist.	0.229	0.260	0.279	0.317	0.275	0.281
Classifier trained on MJDs						
Llama origin	0.246	0.276	0.306	0.304	0.297	0.304
+ human explanations						
4 in one	0.294	0.287	0.349	0.403	0.335	0.349
3 in one	0.304	0.288	0.349	0.406	0.325	0.344
2 in one	0.301	0.291	0.351	0.407	0.335	0.344
1 in one	0.298	0.291	0.348	0.397	0.325	0.338
avg	0.296	0.289	0.349	0.400	0.330	0.344
+ first model explanations						
Label-Free						
3 in one	0.300	0.299	0.353	0.356	0.283	0.340
2 in one	0.293	0.293	0.327	0.296	0.244	0.321
1 in one	0.276	0.276	0.297	0.257	0.224	0.284
avg	0.288	0.288	0.325	0.307	0.254	0.312
VariErr Label-Guided						
4 in one	0.294	0.269	0.345	0.412	0.335	0.312
3 in one	0.296	0.271	0.353	0.407	0.341	0.321
2 in one	0.303	0.280	0.358	0.403	0.336	0.312
1 in one	0.318	0.293	0.346	0.391	0.310	0.313
avg	0.306	0.281	0.345	0.402	0.322	0.312
MNLI Label-Guided						
5 in one	0.294	0.281	0.323	0.354	0.300	0.311
4 in one	0.286	0.290	0.324	0.351	0.290	0.314
3 in one	0.280	0.284	0.323	0.346	0.286	0.321
2 in one	0.272	0.285	0.316	0.342	0.280	0.314
1 in one	0.271	0.285	0.304	0.318	0.269	0.286
avg	0.282	0.283	0.314	0.336	0.284	0.298
+ longest model explanations						
Label-Free						
3 in one	0.308	0.302	0.352	0.340	0.288	0.352
2 in one	0.281	0.279	0.312	0.286	0.250	0.321
1 in one	0.277	0.288	0.304	0.288	0.235	0.295
avg	0.292	0.295	0.328	0.314	0.262	0.323
VariErr Label-Guided						
4 in one	0.298	0.284	0.351	0.417	0.335	0.325
3 in one	0.293	0.283	0.350	0.416	0.337	0.314
2 in one	0.295	0.281	0.350	0.419	0.338	0.316
1 in one	0.312	0.287	0.348	0.405	0.313	0.313
avg	0.305	0.285	0.349	0.411	0.324	0.319
MNLI Label-Guided						
5 in one	0.288	0.281	0.330	0.354	0.301	0.327
4 in one	0.283	0.277	0.332	0.351	0.297	0.336
3 in one	0.282	0.283	0.328	0.349	0.289	0.334
2 in one	0.278	0.285	0.323	0.343	0.282	0.319
1 in one	0.279	0.286	0.312	0.325	0.272	0.287
avg	0.284	0.283	0.321	0.339	0.287	0.307

Table 13: ANLI test results for individual runs.

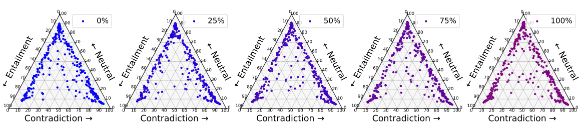


Figure 16: Visualization for gradually replacing first model explanations (Llama3 Scaled by 3).

Classifiers	BERT FT Test			RoBERTa FT Test		
	R1 ↑	R2 ↑	R3 ↑	R1 ↑	R2 ↑	R3 ↑
Classifiers without distribution training						
Out-of-the-box LM	0.170	0.176	0.197	0.167	0.167	0.168
MNLI-FT-LM	0.220	0.269	0.293	0.292	0.262	0.257
Classifiers trained on label distributions						
ChaosNLI HJD	0.268	0.289	0.332	0.357	**0.331**	0.338
VariErr distribution	0.302	0.259	0.319	0.402	0.311	0.321
MNLI distribution	0.229	0.260	0.279	0.317	0.275	0.281
Classifiers trained on MJDs						
Mixtral	0.242	0.252	0.246	0.230	0.240	0.243
+ human explanations	**0.344**	0.280	**0.320**	0.361	**0.292**	**0.300**
+ model explanations						
Label-Free	0.252	0.255	0.255	0.242	0.248	0.243
VariErr Label-Guided	0.340	**0.287**	0.317	**0.362**	0.289	0.296
MNLI Label-Guided	0.275	0.273	0.303	0.329	0.280	0.292

Table 14: ANLI test results for Mixtral.

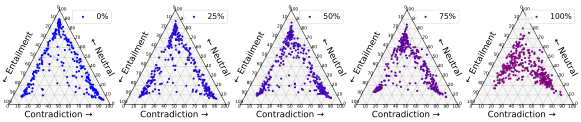


Figure 17: Visualization gradually replacing first noise explanations (Llama3 Scaled by 3).

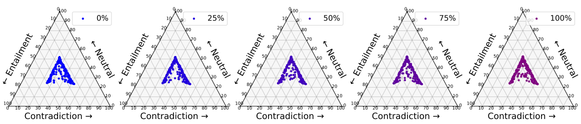


Figure 18: Visualization for gradually replacing first model explanations (Llama3).

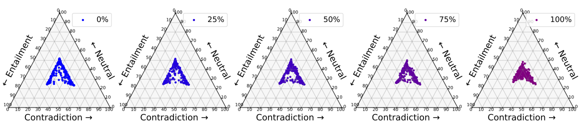


Figure 19: Visualization gradually replacing first noise explanations (Llama3).

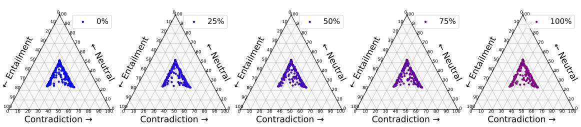


Figure 20: Visualization for gradually replacing longest model explanations (Llama3).

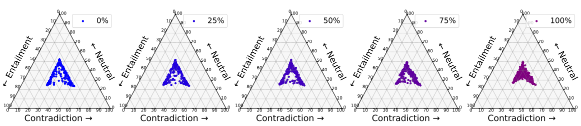


Figure 21: Visualization for gradually replacing longest noise explanations (Llama3).

Distributions	Dist. Comparison			BERT Fine-Tuning Comparison(dev/test)			RoBERTa Fine-Tuning Comparison(dev/test)			Global Metric
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Baseline from Human Annotations										
ChaosNLI HJD	0.000	0.000	0.000	0.073 / 0.077	0.967 / 0.974	0.645 / 0.609	0.062 / 0.060	0.933 / 0.922	0.696 / 0.653	1.000
VariErr distribution	3.604	0.282	0.296	0.177 / 0.179	1.279 / 1.279	0.552 / 0.522	0.166 / 0.173	1.246 / 1.261	0.616 / 0.594	0.688
MNLI distribution	1.242	0.281	0.295	0.104 / 0.100	1.062 / 1.042	0.569 / 0.555	0.101 / 0.093	1.052 / 1.020	0.625 / 0.607	0.795
Model Judgment Distributions										
Llama3	0.259	0.262	0.284	0.099 / 0.101	1.045 / 1.044	0.516 / 0.487	0.094 / 0.096	1.030 / 1.031	0.545 / 0.522	0.689
+ human explanations	0.238	0.250	0.269	0.098 / 0.099	1.043 / 1.039	0.575 / 0.556	0.091 / 0.092	1.021 / 1.019	0.641 / 0.616	0.771
+ replace first model explanations										
Label-Free 100%	0.293	0.276	0.308	0.105 / 0.106	1.063 / 1.060	0.548 / 0.543	0.102 / 0.104	1.056 / 1.054	0.613 / 0.583	0.741
noise	0.292	0.276	0.308	0.105 / 0.106	1.063 / 1.060	0.510 / 0.504	0.102 / 0.104	1.056 / 1.055	0.549 / 0.543	0.702
VariErr Label-Guided 25%	0.236	0.249	0.267	0.098 / 0.099	1.042 / 1.038	0.572 / 0.547	0.090 / 0.091	1.019 / 1.016	0.639 / 0.622	0.769
noise	0.249	0.255	0.276	0.100 / 0.101	1.048 / 1.045	0.551 / 0.539	0.094 / 0.095	1.030 / 1.029	0.628 / 0.613	0.744
VariErr Label-Guided 50%	0.237	0.248	0.268	0.098 / 0.098	1.041 / 1.038	0.559 / 0.546	0.089 / 0.091	1.017 / 1.015	0.639 / 0.628	0.757
noise	0.255	0.259	0.281	0.099 / 0.101	1.047 / 1.044	0.546 / 0.523	0.094 / 0.096	1.032 / 1.031	0.625 / 0.610	0.733
VariErr Label-Guided 75%	0.236	0.248	0.267	0.098 / 0.098	1.041 / 1.037	0.557 / 0.544	0.090 / 0.091	1.018 / 1.015	0.634 / 0.628	0.759
noise	0.265	0.264	0.288	0.101 / 0.102	1.050 / 1.048	0.533 / 0.521	0.096 / 0.098	1.037 / 1.036	0.622 / 0.601	0.720
VariErr Label-Guided 100%	0.236	0.248	0.267	0.098 / 0.098	1.041 / 1.037	0.551 / 0.543	0.089 / 0.091	1.017 / 1.014	0.627 / 0.628	0.755
noise	0.279	0.271	0.299	0.102 / 0.103	1.055 / 1.052	0.525 / 0.513	0.099 / 0.101	1.046 / 1.045	0.612 / 0.592	0.705
+ replace longest model explanations										
Label-Free 100%	0.295	0.278	0.310	0.106 / 0.107	1.066 / 1.063	0.539 / 0.533	0.103 / 0.105	1.059 / 1.058	0.581 / 0.571	0.744
noise	0.288	0.275	0.306	0.104 / 0.105	1.061 / 1.058	0.516 / 0.515	0.101 / 0.103	1.052 / 1.053	0.558 / 0.552	0.709
VariErr Label-Guided 25%	0.236	0.249	0.267	0.098 / 0.099	1.042 / 1.038	0.574 / 0.551	0.090 / 0.091	1.019 / 1.016	0.641 / 0.627	0.772
noise	0.248	0.255	0.275	0.100 / 0.101	1.048 / 1.044	0.561 / 0.540	0.094 / 0.095	1.029 / 1.028	0.631 / 0.618	0.745
VariErr Label-Guided 50%	0.236	0.248	0.267	0.097 / 0.099	1.041 / 1.038	0.571 / 0.543	0.090 / 0.091	1.017 / 1.016	0.639 / 0.627	0.757
noise	0.253	0.258	0.280	0.099 / 0.101	1.046 / 1.044	0.546 / 0.531	0.094 / 0.096	1.031 / 1.030	0.620 / 0.616	0.735
VariErr Label-Guided 75%	0.235	0.248	0.267	0.097 / 0.099	1.041 / 1.038	0.564 / 0.545	0.090 / 0.091	1.018 / 1.016	0.643 / 0.622	0.760
noise	0.261	0.262	0.286	0.100 / 0.101	1.049 / 1.046	0.535 / 0.521	0.095 / 0.097	1.034 / 1.034	0.620 / 0.609	0.723
VariErr Label-Guided 100%	0.234	0.247	0.266	0.097 / 0.098	1.041 / 1.037	0.558 / 0.544	0.089 / 0.091	1.016 / 1.014	0.633 / 0.626	0.760
noise	0.274	0.269	0.296	0.101 / 0.103	1.053 / 1.050	0.526 / 0.511	0.098 / 0.100	1.042 / 1.042	0.608 / 0.599	0.709

Table 15: Detailed scores for the ablation study.

Distributions	Dist. Comparison			Global Metric
	KL ↓	JSD ↓	TVD ↓	D.Corr ↑
VariErr distributions	6.628	0.357	0.352	0.907
Llama3 MJD	0.029	0.068	0.088	0.691
+ human explanations	0.000	0.000	0.000	1.000
+ replace first model explanations				
Label-Free 100%	0.023	0.066	0.087	0.645
VariErr Label-Guided 25%	0.002	0.013	0.017	0.970
VariErr Label-Guided 50%	0.003	0.018	0.023	0.955
VariErr Label-Guided 75%	0.003	0.020	0.026	0.945
VariErr Label-Guided 100%	0.005	0.023	0.029	0.930
+ replace longest model explanations				
Label-Free 100%	0.024	0.067	0.088	0.647
VariErr Label-Guided 25%	**0.001**	**0.012**	**0.015**	**0.977**
VariErr Label-Guided 50%	0.003	0.017	0.022	0.959
VariErr Label-Guided 75%	0.003	0.019	0.024	0.950
VariErr Label-Guided 100%	0.004	0.021	0.027	0.939

Table 16: All results from a human-explanation-centric view. All MJDs are compared to the MJD in the green row.

Datasets	Lexical			Syntactic			Semantic		AVG
	n = 1 ↑	n = 2 ↑	n = 3 ↑	n = 1 ↑	n = 2 ↑	n = 3 ↑	Cos. ↑	Euc. ↑	AVG ↑
human explanations	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
replace first model explanations									
25%	0.773	0.673	0.647	0.929	0.769	0.686	0.841	0.828	0.801
50%	0.603	0.432	0.389	0.873	0.598	0.457	0.712	0.698	0.651
75%	0.459	0.239	0.184	0.832	0.468	0.277	0.608	0.593	0.526
100%	0.358	0.103	0.041	0.805	0.377	0.149	0.529	0.519	0.439
replace longest model explanations									
25%	0.758	0.658	0.635	0.926	0.761	0.674	0.824	0.819	0.789
50%	0.581	0.416	0.378	0.873	0.592	0.447	0.691	0.690	0.635
75%	0.438	0.222	0.173	0.832	0.462	0.267	0.581	0.584	0.511
100%	0.335	0.087	0.033	0.807	0.369	0.141	0.501	0.510	0.422

Table 17: Linguistic similarity results for the ablation study. All sets of explanations are parallelly calculated the similarity with human explanations on Lexical, Syntactic and Semantic levels folloing Giulianelli et al. (2023).

Distributions	Dist. Comparison			BERT Fine-Tuning Comparison(dev/test)			RoBERTa Fine-Tuning Comparison(dev/test)			Global Metric
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Baseline from Human Annotations										
ChaosNLI HJD	0.000	0.000	0.000	0.081 / 0.083	0.993 / 0.992	0.643 / 0.597	0.065 / 0.065	0.944 / 0.936	0.691 / 0.652	1.000
VariErr distribution	4.254	0.313	0.320	0.193 / 0.197	1.329 / 1.333	0.563 / 0.535	0.214 / 0.222	1.391 / 1.407	0.585 / 0.555	0.661
MNLI distribution	1.215	0.281	0.290	0.105 / 0.103	1.064 / 1.051	0.553 / 0.540	0.092 / 0.086	1.024 / 0.999	0.615 / 0.604	0.743
Model Judgment Distributions										
Llama3	0.258	0.261	0.286	0.092 / 0.093	1.024 / 1.020	0.514 / 0.471	0.092 / 0.095	1.025 / 1.026	0.531 / 0.512	0.684
+ human explanations	0.240	0.249	0.275	0.090 / 0.090	1.017 / 1.011	0.594 / 0.567	0.089 / 0.091	1.014 / 1.015	0.618 / 0.597	0.750
+ replace *first* model explanations										
50%	0.238	0.247	0.273	0.089 / 0.089	1.017 / 1.010	0.585 / 0.568	0.089 / 0.091	1.014 / 1.015	0.620 / 0.597	0.758
75%	0.237	0.246	0.272	0.090 / 0.090	1.018 / 1.011	0.577 / 0.565	0.088 / 0.091	1.013 / 1.014	0.620 / 0.586	0.759
100%	0.237	0.246	0.271	0.089 / 0.090	1.017 / 1.011	0.581 / 0.566	0.088 / 0.090	1.013 / 1.014	0.617 / 0.586	0.755
+ replace *longest* model explanations										
50%	0.238	0.247	0.273	0.089 / 0.089	1.016 / 1.009	0.586 / 0.566	0.088 / 0.091	1.013 / 1.014	0.618 / 0.600	0.749
75%	0.239	0.247	0.273	0.090 / 0.090	1.017 / 1.011	0.581 / 0.565	0.088 / 0.091	1.013 / 1.014	0.618 / 0.594	0.744
100%	0.238	0.246	0.272	0.089 / 0.089	1.017 / 1.011	0.581 / 0.565	0.088 / 0.091	1.013 / 1.014	0.616 / 0.587	0.745
+ replace *preferred* model explanations										
greedy 75.75%	0.241	0.248	0.274	0.089 / 0.090	1.017 / 1.011	0.584 / 0.569	0.088 / 0.090	1.013 / 1.013	0.619 / 0.594	0.733
representative 55.25%	0.240	0.248	0.274	0.089 / 0.090	1.016 / 1.011	0.587 / 0.567	0.088 / 0.091	1.013 / 1.014	0.619 / 0.597	0.739
+ replace *unpreferred* model explanations										
greedy 68.5%	0.239	0.247	0.273	**0.089** / **0.089**	**1.016** / **1.009**	**0.589** / **0.571**	**0.087** / 0.090	**1.011** / 1.012	**0.623** / 0.599	0.752
representative 63.25%	**0.237**	**0.246**	**0.271**	0.089 / 0.089	1.016 / 1.010	0.584 / 0.566	0.088 / **0.090**	1.011 / **1.012**	0.621 / **0.607**	**0.761**

Table 18: All the results on 100 validated NLI instances for explanation selection strategy including LLM/human preference.

Distributions	Dist. Comparison			BERT Fine-Tuning Comparison(dev/test)			RoBERTa Fine-Tuning Comparison(dev/test)			Global Metric
	KL ↓	JSD ↓	TVD ↓	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	D.Corr ↑
Baseline from Human Annotations										
ChaosNLI HJD	0.000	0.000	0.000	0.081 / 0.083	0.993 / 0.992	0.643 / 0.597	0.065 / 0.065	0.944 / 0.936	0.691 / 0.652	1.000
VariErr dist.	4.254	0.313	0.320	0.193 / 0.197	1.329 / 1.333	0.563 / 0.535	0.214 / 0.222	1.391 / 1.407	0.585 / 0.555	0.661
MNLI dist.	1.215	0.281	0.290	0.105 / 0.103	1.064 / 1.051	0.553 / 0.540	0.092 / 0.086	1.024 / 0.999	0.615 / 0.604	0.743
Model Judgment Distributions										
Llama3	0.258	0.261	0.286	0.092 / 0.093	1.024 / 1.020	0.514 / 0.471	0.092 / 0.095	1.025 / 1.026	0.531 / 0.512	0.684
+ human explanations										
4 in one	0.240	0.247	0.273	0.088 / 0.088	1.013 / 1.006	0.589 / 0.566	0.087 / 0.090	1.011 / 1.011	0.607 / 0.591	0.703
3 in one	0.239	0.247	0.273	0.089 / 0.089	1.015 / 1.009	0.599 / 0.566	0.088 / 0.090	1.012 / 1.013	0.613 / 0.598	0.732
2 in one	0.239	0.248	0.274	0.090 / 0.090	1.018 / 1.012	0.596 / 0.569	0.089 / 0.091	1.015 / 1.016	0.629 / 0.604	0.769
1 in one	0.244	0.252	0.279	0.092 / 0.092	1.024 / 1.018	0.593 / 0.567	0.091 / 0.093	1.020 / 1.021	0.622 / 0.596	0.795
avg	0.240	0.249	0.275	0.090 / 0.090	1.017 / 1.011	0.594 / 0.567	0.089 / 0.091	1.014 / 1.015	0.618 / 0.597	0.750
+ replace *first* model explanations										
50%										
4 in one	0.236	0.245	0.271	0.088 / 0.088	1.013 / 1.006	0.587 / 0.571	0.087 / 0.089	1.010 / 1.010	0.616 / 0.598	0.720
3 in one	0.234	0.245	0.270	0.088 / 0.088	1.013 / 1.007	0.590 / 0.567	0.087 / 0.090	1.011 / 1.012	0.619 / 0.595	0.751
2 in one	0.237	0.247	0.272	0.089 / 0.090	1.017 / 1.011	0.586 / 0.570	0.089 / 0.091	1.015 / 1.016	0.621 / 0.605	0.772
1 in one	0.244	0.251	0.279	0.092 / 0.092	1.024 / 1.017	0.578 / 0.566	0.091 / 0.093	1.022 / 1.022	0.624 / 0.592	0.791
avg	0.238	0.247	0.273	0.089 / 0.089	1.017 / 1.010	0.585 / 0.568	0.089 / 0.091	1.014 / 1.015	0.620 / 0.597	0.758
75%										
4 in one	0.236	0.244	0.269	0.088 / 0.089	1.014 / 1.008	0.574 / 0.567	0.086 / 0.089	1.007 / 1.008	0.615 / 0.580	0.721
3 in one	0.234	0.244	0.269	0.089 / 0.089	1.015 / 1.009	0.580 / 0.565	0.087 / 0.089	1.009 / 1.010	0.615 / 0.586	0.752
2 in one	0.236	0.246	0.272	0.090 / 0.090	1.018 / 1.012	0.580 / 0.569	0.088 / 0.091	1.014 / 1.014	0.625 / 0.595	0.773
1 in one	0.243	0.250	0.278	0.092 / 0.092	1.024 / 1.017	0.573 / 0.561	0.091 / 0.093	1.022 / 1.022	0.627 / 0.581	0.788
avg	0.237	0.246	0.272	0.090 / 0.090	1.018 / 1.011	0.577 / 0.565	0.088 / 0.091	1.013 / 1.014	0.620 / 0.586	0.759
100%										
4 in one	0.237	0.243	0.268	0.088 / 0.088	1.013 / 1.007	0.582 / 0.565	0.087 / 0.089	1.009 / 1.009	0.608 / 0.580	0.718
3 in one	0.234	0.244	0.268	0.088 / 0.089	1.014 / 1.008	0.585 / 0.567	0.087 / 0.089	1.009 / 1.010	0.614 / 0.590	0.750
2 in one	0.235	0.246	0.271	0.089 / 0.090	1.017 / 1.011	0.586 / 0.569	0.088 / 0.091	1.013 / 1.014	0.623 / 0.589	0.772
1 in one	0.243	0.250	0.278	0.092 / 0.092	1.023 / 1.017	0.571 / 0.563	0.091 / 0.093	1.021 / 1.021	0.623 / 0.583	0.781
avg	0.237	0.246	0.271	0.089 / 0.090	1.017 / 1.011	0.581 / 0.566	0.088 / 0.090	1.013 / 1.014	0.617 / 0.586	0.755
+ replace *longest* model explanations										
50%										
4 in one	0.237	0.245	0.270	0.087 / 0.087	1.011 / 1.004	0.590 / 0.570	0.086 / 0.089	1.007 / 1.008	0.620 / 0.597	0.707
3 in one	0.236	0.245	0.270	0.088 / 0.088	1.013 / 1.007	0.590 / 0.564	0.087 / 0.089	1.009 / 1.010	0.610 / 0.605	0.737
2 in one	0.237	0.247	0.272	0.089 / 0.089	1.016 / 1.010	0.589 / 0.569	0.088 / 0.091	1.013 / 1.014	0.621 / 0.605	0.765
1 in one	0.244	0.251	0.278	0.092 / 0.092	1.023 / 1.017	0.573 / 0.563	0.091 / 0.093	1.021 / 1.022	0.622 / 0.592	0.786
avg	0.238	0.247	0.273	0.089 / 0.089	1.016 / 1.009	0.586 / 0.566	0.088 / 0.091	1.013 / 1.014	0.618 / 0.600	0.749
75%										
4 in one	0.238	0.245	0.270	0.088 / 0.088	1.013 / 1.007	0.586 / 0.563	0.086 / 0.089	1.007 / 1.008	0.620 / 0.595	0.703
3 in one	0.236	0.245	0.270	0.089 / 0.089	1.015 / 1.008	0.587 / 0.569	0.087 / 0.089	1.009 / 1.010	0.617 / 0.599	0.732
2 in one	0.238	0.247	0.273	0.090 / 0.090	1.018 / 1.012	0.582 / 0.569	0.088 / 0.091	1.014 / 1.014	0.614 / 0.597	0.761
1 in one	0.244	0.251	0.279	0.092 / 0.092	1.024 / 1.017	0.568 / 0.558	0.091 / 0.093	1.022 / 1.022	0.622 / 0.586	0.781
avg	0.239	0.247	0.273	0.090 / 0.090	1.017 / 1.011	0.581 / 0.565	0.088 / 0.091	1.013 / 1.014	0.618 / 0.594	0.744
100%										
4 in one	0.237	0.244	0.269	0.088 / 0.088	1.013 / 1.007	0.586 / 0.568	0.086 / 0.089	1.008 / 1.009	0.613 / 0.589	0.709
3 in one	0.235	0.244	0.269	0.088 / 0.089	1.014 / 1.008	0.587 / 0.566	0.087 / 0.089	1.009 / 1.010	0.615 / 0.589	0.737
2 in one	0.237	0.246	0.272	0.089 / 0.090	1.017 / 1.011	0.587 / 0.566	0.088 / 0.091	1.013 / 1.014	0.614 / 0.590	0.762
1 in one	0.244	0.250	0.278	0.092 / 0.091	1.023 / 1.017	0.566 / 0.559	0.091 / 0.093	1.021 / 1.021	0.622 / 0.579	0.774
avg	0.238	0.246	0.272	0.089 / 0.089	1.017 / 1.011	0.581 / 0.565	0.088 / 0.091	1.013 / 1.014	0.616 / 0.587	0.745
+ replace *aligned* model explanations										
greedy 75.75%										
4 in one	0.240	0.246	0.272	0.088 / 0.088	1.012 / 1.006	0.590 / 0.566	0.087 / 0.089	1.009 / 1.009	0.615 / 0.593	0.692
3 in one	0.239	0.246	0.272	0.088 / 0.089	1.013 / 1.008	0.591 / 0.575	0.087 / 0.090	1.011 / 1.011	0.611 / 0.590	0.719
2 in one	0.239	0.247	0.274	0.089 / 0.090	1.017 / 1.012	0.586 / 0.573	0.088 / 0.091	1.014 / 1.014	0.620 / 0.598	0.747
1 in one	0.244	0.250	0.278	0.092 / 0.092	1.024 / 1.018	0.568 / 0.564	0.090 / 0.092	1.020 / 1.020	0.633 / 0.595	0.774
avg	0.241	0.248	0.274	0.089 / 0.090	1.017 / 1.011	0.584 / 0.569	0.088 / 0.090	1.013 / 1.013	0.619 / 0.594	0.733
representative 55.25%										
4 in one	0.239	0.246	0.272	0.088 / 0.088	1.012 / 1.006	0.599 / 0.570	0.087 / 0.089	1.009 / 1.010	0.616 / 0.591	0.698
3 in one	0.237	0.246	0.272	0.088 / 0.089	1.013 / 1.008	0.595 / 0.568	0.087 / 0.090	1.010 / 1.011	0.609 / 0.603	0.730
2 in one	0.239	0.248	0.274	0.089 / 0.090	1.017 / 1.011	0.587 / 0.567	0.088 / 0.091	1.013 / 1.014	0.617 / 0.605	0.752
1 in one	0.244	0.251	0.278	0.091 / 0.092	1.023 / 1.017	0.567 / 0.561	0.090 / 0.093	1.020 / 1.020	0.635 / 0.589	0.778
avg	0.240	0.248	0.274	0.089 / 0.090	1.016 / 1.011	0.587 / 0.567	0.088 / 0.091	1.013 / 1.014	0.619 / 0.597	0.739
+ replace *aligned* model explanations										
greedy 68.5%										
4 in one	0.237	0.244	0.270	0.088 / 0.088	1.012 / 1.005	0.595 / 0.565	0.086 / 0.088	1.006 / 1.007	0.622 / 0.600	0.712
3 in one	0.235	0.245	0.270	0.088 / 0.088	1.013 / 1.007	0.588 / 0.571	0.086 / 0.089	1.007 / 1.009	0.624 / 0.609	0.742
2 in one	0.238	0.247	0.273	0.089 / 0.089	1.016 / 1.010	0.591 / 0.574	0.088 / 0.090	1.011 / 1.013	0.624 / 0.606	0.768
1 in one	0.246	0.251	0.280	0.091 / 0.091	1.022 / 1.015	0.583 / 0.573	0.091 / 0.093	1.021 / 1.021	0.622 / 0.580	0.787
avg	0.239	0.247	0.273	0.089 / 0.089	1.016 / 1.009	0.589 / 0.571	0.087 / 0.090	1.011 / 1.012	0.623 / 0.599	0.752
representative 63.25%										
4 in one	0.235	0.244	0.268	0.088 / 0.088	1.012 / 1.006	0.587 / 0.560	0.086 / 0.088	1.006 / 1.006	0.622 / 0.605	0.721
3 in one	0.233	0.244	0.268	0.088 / 0.088	1.013 / 1.007	0.586 / 0.567	0.086 / 0.089	1.008 / 1.009	0.625 / 0.613	0.753
2 in one	0.236	0.247	0.272	0.089 / 0.089	1.016 / 1.010	0.588 / 0.573	0.088 / 0.090	1.012 / 1.013	0.624 / 0.615	0.776
1 in one	0.244	0.251	0.279	0.091 / 0.091	1.023 / 1.016	0.574 / 0.563	0.090 / 0.093	1.020 / 1.021	0.612 / 0.593	0.792
avg	0.237	0.246	0.271	0.089 / 0.089	1.016 / 1.010	0.584 / 0.566	0.088 / 0.090	1.011 / 1.012	0.621 / 0.607	0.761

Table 19: Results on 100 validated NLI instances of explanation selection strategy in individual runs.

Datasets	Lexical			Syntactic			Semantic		AVG
	n = 1 ↓	n = 2 ↓	n = 3 ↓	n = 1 ↓	n = 2 ↓	n = 3 ↓	Cos. ↓	Euc. ↓	AVG ↓
human explanations	0.339	0.103	0.045	0.753	0.340	0.140	0.512	0.516	0.343
first model explanations									
Label-Free	0.465	0.188	0.105	0.878	0.482	0.229	0.599	0.543	0.436
VariErr Label-Guided	0.456	0.170	0.083	0.897	0.510	0.241	0.584	0.538	0.435
MNLI Label-Guided	0.431	0.147	0.066	0.890	0.487	0.215	0.567	0.531	0.417
longest model explanations									
Label-Free	0.439	0.139	0.065	0.920	0.520	0.227	0.559	0.527	0.425
VariErr Label-Guided	0.457	0.162	0.079	0.920	0.535	0.252	0.569	0.532	0.438
MNLI Label-Guided	0.437	0.141	0.064	0.917	0.523	0.235	0.549	0.525	0.424

Table 20: Linguistic variability check for the main results in Table 1.