

# “Seeing the Big through the Small”: Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations?

Beiduo Chen<sup>▲</sup><sup>✉</sup> Xinpeng Wang<sup>▲</sup><sup>✉</sup> Siyao Peng<sup>▲</sup><sup>✉</sup> Robert Litschko<sup>▲</sup><sup>✉</sup>  
 Anna Korhonen<sup>▲</sup> Barbara Plank<sup>▲</sup><sup>✉</sup>

<sup>▲</sup> MaiNLP, Center for Information and Language Processing, LMU Munich, Germany

<sup>✉</sup> Munich Center for Machine Learning (MCML), Munich, Germany

<sup>▲</sup> Language Technology Lab, University of Cambridge, United Kingdom

{beiduo.chen, xinpeng.wang, siyao.peng, robert.litschko}@lmu.de

alk23@cam.ac.uk, b.plank@lmu.de

## Abstract

Human label variation (HLV) is a valuable source of information that arises when multiple human annotators provide different labels for valid reasons. In Natural Language Inference (NLI) earlier approaches to capturing HLV involve either collecting annotations from many crowd workers to represent human judgment distribution (HJD) or use expert linguists to provide detailed explanations for their chosen labels. While the former method provides denser HJD information, obtaining it is resource-intensive. In contrast, the latter offers richer textual information but it is challenging to scale up to many human judges. Besides, large language models (LLMs) are increasingly used as evaluators (“LLM judges”) but with mixed results, and few works aim to study HJDs. This study proposes to exploit LLMs to approximate HJDs using a small number of expert linguists to provide detailed explanations for their chosen labels and explanations. Our experiments show that a few explanations significantly improve LLMs’ ability to approximate HJDs with and without explicit labels, thereby providing a solution to scale up annotations for HJD. However, fine-tuning smaller soft-label aware models with the LLM-generated model judgment distributions (MJDs) presents partially inconsistent results: while similar in distance, their resulting fine-tuned models and visualized distributions differ substantially. We show the importance of complementing instance-level distance measures with a global-level shape metric and visualization to evaluate MJDs more effectively against human judgment distributions.

## 1 Introduction

In Natural Language Processing (NLP), we are faced with many situations in which more than one plausible label (or reading) exists, a phenomenon referred to as Human Label Variation (HLV, Plank 2022). HLV could be caused by inherent disagreement (Pavlick and Kwiatkowski, 2019), subjectivity (Cabitza et al., 2023), or cases where multi-

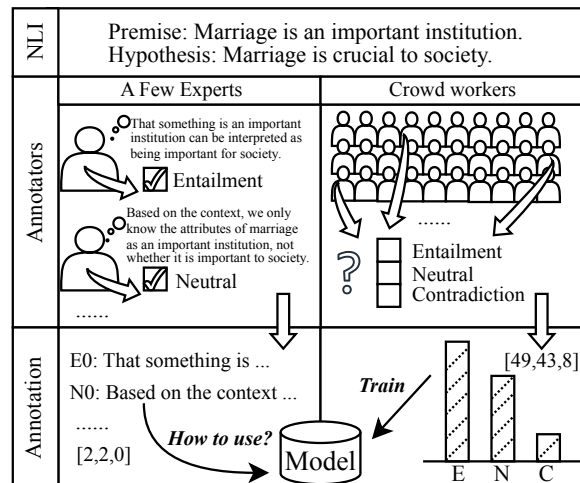


Figure 1: Comparison between approaches to investigate HLV in NLI. Experts first explain the sample individually and then select a label, while crowd workers only record their choices. Explanations provide details for labels to understand HLV. However, it is not clear how to use explanations effectively to model HLV.

answers are plausible. An increasing body of work suggests that HLV provides rich information that should not be discarded as noise (e.g. Aroyo and Welty, 2015; Plank et al., 2014; Uma et al., 2021) as it impacts every step of the learning process (Plank, 2022). As one of the fundamental natural language understanding tasks, Natural Language Inference (NLI) (Dagan et al., 2005; Bowman et al., 2015; Williams et al., 2018; Manning, 2006) has embraced HLV especially (e.g. Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang and de Marneffe, 2022a; Zhou et al., 2022).

There are two common approaches to investigating HLV in NLI, illustrated in Figure 1. One way is to collect annotations from a “big” number of crowd workers (Pavlick and Kwiatkowski, 2019), enabling smoother label probabilities from a statistical standpoint to represent “collective” human opinion (Nie et al., 2020). The obtained hu-

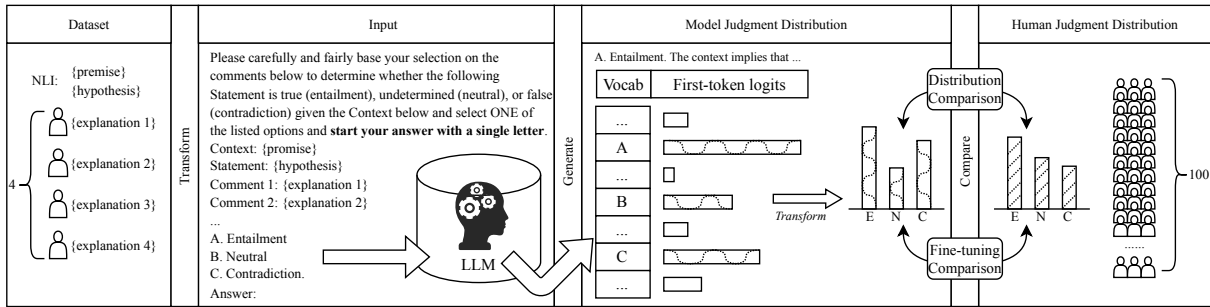


Figure 2: The overall structure of our LLM approximation system. Explanations from 4 annotators in VariErr NLI (Weber-Genzel et al., 2024) are transformed with corresponding NLI samples together into multiple-choice questions, and the generated soft labels (model judgment distributions) are compared with human judgment distributions from 100 crowd workers in Chaos NLI (Nie et al., 2020). Two SOTA open-source LLMs, Mixtral (Jiang et al., 2024), and Llama3 (Meta, 2024), interpret the explanations, and we conduct comparisons on *distribution* and *fine-tuning*.

man judgment distributions (HJDs) are typically soft labels suitable for model training and evaluation (Fornaciari et al., 2021a; Uma et al., 2021; Anand et al., 2024). It provides a rich resource to study and model HLV via soft labels or other methods (Uma et al., 2021; Gruber et al., 2024; Davani et al., 2022; Fornaciari et al., 2021a; Jiang and de Marneffe, 2022a). However, this annotation scheme is labor intensive, and despite offering limited insights on how label variation arises (Jiang and de Marneffe, 2022b). Contrary to the former dense HJD-based annotation, some NLI variation datasets are annotated by a “small” number of expert linguists, delivering annotated labels along with corresponding explanations shown to contain richer linguistic information than explicit labels alone (Jiang et al., 2023b; Weber-Genzel et al., 2024), as each annotator explains the label they selected. These accompanying explanations however, are scarce, and it is not clear how to use them effectively for modeling.

LLMs, benefiting from enormous training data, are capable of generalizing across various tasks (Zhao et al., 2023), from text generation (Lee et al., 2023a), model distillation (Xu et al., 2024), to name a few, to more recently, functioning as “LLM judges,” for example in evaluation (Chiang and Lee, 2023; Verga et al., 2024) or linguistic annotation (e.g. Ettinger et al., 2023). This suggests that LLMs can serve as a good bridge between humans and machines. In this paper, our main research questions are: *Can LLMs provided with a “small” number of detailed explanations better approximate the human judgment distributions collected by a “big” number of annotators?* If this is the case, *are the obtained model judgment distribu-*

*tions (MJDs) suitable as soft labels for fine-tuning smaller models to predict distributions?* To investigate these questions, we provide two sets of main experiments, as illustrated in Figure 2.

Experiments on *distribution comparison* show that a few explanations can improve the capabilities of LLMs to approximate human judgment distributions among various metrics. To demonstrate the practical significance of the generated MJDs from our method, we applied them in a *fine-tuning comparison* to train pre-trained transformer models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). Interestingly, our analysis shows that the effectiveness of the generated MJDs serving as fine-tuning data cannot be predicted well by *distribution comparison*. We hypothesize that instance-level measures, e.g., KL Divergence (Kullback and Leibler, 1951) and Jensen-Shannon Distance (Endres and Schindelin, 2003), overlook the dependencies across sample points, which captures the global-level HLV useful for training models. Therefore, we utilize distance correlation (Székely et al., 2007) to measure the global-level association between the generated MJDs over the whole dataset and the corresponding target HJD. We empirically show that distance correlation can reliably predict the performance of MJDs on *fine-tuning comparison*, as supported by visual investigations. All our code is available at <https://github.com/mainlp/MJD-Estimator> for reproduction.

## 2 Related Work

**Human label variation** Human label variation represents a phenomenon in which inherent disagreement exists in annotation due to genuine disagreement, subjectivity or simply because two (or

more) views are plausible (Plank, 2022). Aroyo and Welty (2015) propose that disagreement is not noise but signal, which is giving us information, as human labels are bound to be scarce yet at the same time critical as they provide human interpretations and values. With substantial datasets, providing many judgments by high-quality coders for each item, training directly with soft labels achieved better results than training from aggregated or even gold labels (Uma et al., 2021). For example, Fornaciari et al. (2021b) find that the soft-label prediction auxiliary task reduces the penalty for errors on ambiguous entities and thereby mitigates overfitting. Thus, Pavlick and Kwiatkowski (2019) argue for a refined evaluation objective that requires models to explicitly capture the full distribution of plausible human judgments.

**Human label variation in NLI** For NLI—the task to determine whether a given premise entails, contradicts, or is neutral towards a target hypothesis—there exist several datasets addressing human label variation (HLV). On the one hand, Variation NLI (Pavlick and Kwiatkowski, 2019) and Chaos NLI (Nie et al., 2020) collect annotations from 50 and 100 crowd workers for human judgment distributions (HJDs). On the other, Live NLI (Jiang et al., 2023b) and VariErr NLI (Weber-Genzel et al., 2024) are annotated by only 5 or 4 linguists, but add textual explanations to their NLI labels. While these works have used NLI with explanations, their goals differ: either to unravel more on reasons for disagreement from the annotators (Jiang et al., 2023b) or to use explanations to facilitate separation of plausible variation from annotation errors (Weber-Genzel et al., 2024). However, little research has been done on estimating HJDs from a few labels and explanations.

**Human explanations & LLM estimators** Human explanations, particularly ecologically valid ones where the same annotator provides both the label and explanation (Jiang et al., 2023b), are effective in improving LLMs’ performance (Wei et al., 2022b; Lampinen et al., 2022). Wadhwa et al. (2023) investigate the levels of missing information in a provided answer for a target question in the context of a given article. They use LLMs to rescale the coarse-grained (4-level) labels and accompanying explanations to a 100-point scale and compare them with manually annotated HJDs. Pavlovic and Poesio (2024) use GPT-3.5 to estimate HJDs directly, but in contrast to us focus on datasets with

inherently subjective NLP tasks provided by the SemEval 2023 shared task 11 (Leonardelli et al., 2023). Their exploration shows that GPT-3.5 tends to produce distributions not well aligned with HJDs for subjective tasks.

LLMs are also employed on the two NLI datasets with explanations. Jiang et al. (2023b) prompt GPT-3 to predict labels and generate explanations on LiveNLI items. They show that the predict-then-explain (post-prediction explanation) strategy significantly outperforms explain-then-predict (chain-of-thought prompting). They also observe through qualitative analysis that over half of the LLM-generated explanations lack informativeness, i.e., only restate the premise/hypothesis. Weber-Genzel et al. (2024) ask GPT-3.5 and GPT-4 to judge the probability of whether individual explanations in VariErr make sense for the corresponding NLI labels. Results show that GPT-4 outperforms traditional error detection methods, but the latter only evaluate the labels without factoring in the explanations, leaving several questions open.

Lee et al. (2023b) is the closest to our work. They propose an LLM Distribution Estimator that reads premise-hypothesis pairs and generates label distributions. However, their MJDs align poorly with HJDs. Moreover, explanation-contained NLI datasets were not yet available when Lee et al. (2023b) was published. Our paper bridges the recently released detailed explanation annotations with LLMs to estimate HJD on NLI efficiently.

### 3 LLMs to Estimate HJDs

As LLM outputs are typically in text form and fluctuate based on inputs and parameters, obtaining model judgment distributions (MJDs) directly from the outputs to approximate human judgment distributions (HJDs) is challenging. We propose LLM prompts with multiple-choice questions (§3.1),<sup>1</sup> to illustrate how we estimate MJD using first-token probability (§3.2) and reduce biases via permutations of choice options and explanations (§3.3).

#### 3.1 Prompt Types

To facilitate asking LLMs to estimate MJDs, we transform the NLI problem into a multiple-choice question answering (MCQA) prompt, selecting one answer from the three options [A, B, C]. The prompts are shown in detail in Table 5 in Appendix

<sup>1</sup>In preliminary experiments, we found that directly asking LLMs for MJDs was challenging. We also found another interesting prompt format in Appendix J.

A. We design three prompt types: a base prompt “without explanations”, a prompt “with explanations”, and one “with explicit explanations” which contains both label and explanation.

**Without explanation** Our base prompt informs LLMs about the NLI task, provides the premise and hypothesis of NLI instances and instructs the LLMs to choose an NLI label via a MCQA format. We constrain the label selection space by asking LLMs to “select ONE of the listed options” and restrict the initial letter of the output to one of [A, B, C] by instructing “start your answer with a single letter.”

**With explanations** This prompt incorporates human explanations of label choices as “comments” without disclosing the annotators’ chosen labels directly. Specifically, we place these comments in the prompt after the hypothesis and premise but before the MCQA part.

**With explicit explanations** This prompt type reveals the NLI labels in addition to the corresponding explanations to the LLMs. Our preliminary experiments found that LLMs sometimes misinterpret phrases that support a Neutral label (e.g., “do not mention” or “cannot explain”) as supporting a Contradiction label. Therefore, we include a prompt type where the corresponding explanation by the annotator is appended with the phrase “so I choose X” to clarify the intended NLI label.

### 3.2 First-token Probability

Conditioned by the prompts above, we next map LLMs’ output from [A, B, C] to probabilities as MJDs. In particular, we set up a one-to-one mapping  $f: O \rightarrow L$  from the option set  $O$  to the label space  $L$ , where  $O = \{A, B, C\}$  and  $L = \{Entailment, Neutral, Contradiction\}$  with permutations on both  $O$  and  $L$  (cf. §3.3).

Denote the text output of LLMs as a list of words  $w = [w_1, w_2, \dots, w_k]$ ,  $w_i \in V$  where  $k$  is the length of the text output and  $V$  is the vocabulary used for LLMs. We extract the logits of the first-token  $w_1$  before the decoding process as  $s_{w_1} = [s_{w_1}, s_{w_2}, \dots, s_{w_k}, s_{w_{k+1}}, \dots, s_{w_n}]$ ,  $w_i \in V$  where  $n$  is the vocabulary size. As shown in Figure 2, we only use part of the first-token logits  $s_{w_1}^O = [s_A, s_B, s_C]$  which present the distribution scores of the option set  $O$ . To transform  $s_{w_1}^O$  into a probability distributions  $p^O$ , we utilize normalization and softmax (with temperature  $\tau$ ) functions:<sup>2</sup>

<sup>2</sup>The default  $\tau$  is set to be 20; more studies in Appendix H.

$$p_{\text{norm}}^O(j) = \frac{s_j}{\sum_j^{|O|} s_j}, \quad (1)$$

$$p_{\text{softmax}}^O(j) = \frac{\exp(s_j/\tau)}{\sum_j^{|O|} \exp(s_j/\tau)}, \quad (2)$$

Finally, we obtain the model judgement distribution  $p^L$  through the mapping function  $f$ :

$$p^L = f(p^O). \quad (3)$$

### 3.3 Bias Consideration

Previous studies (e.g., Dominguez-Olmedo et al., 2023; Zheng et al., 2024; Tjuatja et al., 2023) reveal that LLMs are biased when processing multiple-choice questions, such as preferring the first option A. To address this, we shuffle the option set  $O$  of the mapping relationship  $f$ , totaling  $A\binom{3}{3} = 6$  permutations. We ask LLMs to process all permutations, thus averaging all MJDs obtained to mitigate bias. Namely, each of the three NLI labels has been mapped to option A, reducing biases caused by the initial letters of the options. Moreover, the order of multiple explanations may also exert unequal influence on LLMs, as shown in a case study in Appendix K. To address this bias, we adopt two ways to feed explanations (i.e., “comments”) to the LLMs: “serial” and “parallel” modes.

**Serial mode** We input all explanations together to LLMs as in the prompt shown in Table 5, asking LLMs to process them at once. To mitigate the bias caused by the position of the explanations in the prompt, we shuffle the order of all  $m$  explanations with full permutations  $A\binom{m}{m}$  and use the average output as the model’s final answer.

**Parallel mode** We feed one explanation (i.e., “comment”) at a time under the “parallel” mode to the prompt. Namely, to process  $m$  explanation annotations on an NLI item, we create  $m$  prompts and ask LLMs to handle them separately. We then average the  $m$  outputs to obtain the final MJD. Although this approach prevents LLMs from obtaining an overall impression of all explanations, it allows them to focus more on interpreting each explanation and significantly reduces the computational cost from  $A\binom{m}{m}$  (“serial”) to  $m$  (“parallel”).



Datasets	Samples	Valid overlap	Annotators	Explanations	Description
MNLI (Williams et al., 2018)	433k	341	1 or 5	✗	Majority with single label, subset with 5.
VariErr NLI (Weber-Genzel et al., 2024)	500	341	4	✓	Ecologically valid explanations.
MNLI subset of Chaos NLI (Nie et al., 2020)	1,599	341	100	✗	Human judgment distributions.

Table 1: Datasets statistics. Numbers represent either the sample count or the annotator count.

## 4 Experimental Setup

### 4.1 Datasets

We experiment with two NLI datasets containing HLV, Chaos NLI (Nie et al., 2020) and VariErr NLI (Weber-Genzel et al., 2024). The former contains 1,599 MNLI (Williams et al., 2018) instances with HJDs collected from 100 crowd workers. In contrast, the latter contains 500 MNLI instances randomly sampled from Chaos NLI re-annotated with explanations and labels by 4 expert linguists. Note that the source MNLI corpus is much larger than Chaos NLI and VariErr NLI,<sup>3</sup> and is used only for pre-training of the fine-tuning experiments. To ensure fairness and to facilitate follow-up analyses, we focus our comparisons on a subset of 341 VariErr items that receive exactly 4 explanation annotations.<sup>4</sup> Table 1 presents detailed data statistics.

### 4.2 Models

We utilized two open-source instruction-tuned LLMs: Mixtral-8x7b-Instruct-v0.1 (Jiang et al., 2024) and Llama3-Chat-70b (Meta, 2024). We adopt the original chat templates for both models and set the parameter `do_sample=False` in decoding to ensure consistent outputs for the same input. Discussion regarding the risk of data leakage is elaborated in Appendix G.

### 4.3 Distribution Comparison

Our first experiment compares the LLM-derived MJDs to the HJDs. To derive MJDs, we fed LLMs with the three prompt types exemplified in Table 5 in the Appendix using the 341 VariErr instances. We then compare the resulting MJD to the HJDs of the corresponding Chaos NLI instances. We investigate these distribution differences between humans and LLMs at the instance level following prior work (Nie et al., 2020; Chiang and Lee,

<sup>3</sup>MNLI contains 433k instances, but most of them with a single label, and only a small subset of the dev and test data were re-annotated by 4 additional annotators.

<sup>4</sup>In VariErr, each annotator can give one or more label-explanations pairs to an NLI item, and “I don’t know” explanations are dropped after validation. Among 500 VariErr NLI items, 4 items received 1 validated explanation, 30 items 2, 62 items 3, 341 items 4, 59 items 5, and 4 items 6.

2023; Lee et al., 2023b): Kullback-Leibler Divergence (KL, Kullback and Leibler 1951) and Jensen-Shannon Distance (JSD, Endres and Schindelin 2003). In addition, we follow Baan et al. (2022) to measure the human Distribution Calibration Error, namely measured as Total Variation Distance (TVD, Devroye and Lugosi 2001) between MJDs and HJDs. MNLI single labels were transformed into one-hot vectors to compute the metrics.

### 4.4 Fine-tuning Comparison

Our second experiment investigates how well the resulting MJDs approximate human labels for model training. To do so, we compare the generated MJDs and original HJDs to annotated labels of the parallel instances in MNLI, VariErr NLI and Chaos NLI, 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 HJDs or MJDs above; see Appendix B for fine-tuning details.

To evaluate the resulting classifiers, we split the remaining 1,258 MNLI instances from Chaos NLI that do not overlap with VariErr NLI into the development and test sets. We use KL and weighted F1 scores as evaluation metrics between the outputs of the fine-tuned models and HJDs from Chaos NLI. All metrics are detailed in Appendix C.

## 5 Results

### 5.1 Distribution Comparison

Table 2 presents the distribution comparison results. Firstly, we analyze baseline HJDs and observe that the MNLI single-label data (i.e., with no HLV) is the farthest from Chaos NLI’s HJD, followed by MNLI and VariErr distributions. We add a distribution comparison to the uniform distribution as a sanity check to understand the obtained MJDs.

When comparing MJDs, we observe that for the “without explanation” prompt, Llama3 is closer than Mixtral to Chaos NLI HJD. However, both models benefit from adding explanations, i.e., their MJDs gradually get closer to the HJD. The two

Distributions\Metrics	KL ↓	JSD ↓	TVD ↓
<i>Baseline</i>			
Chaos NLI	0	0	0
MNLI single label	9.288	0.422	0.435
MNLI distributions	1.242	0.281	0.295
VariErr distributions	3.604	0.282	0.296
Uniform distribution	0.364	0.307	0.350
<i>MJDs from Mixtral</i>			
$p_{\text{norm}}$ of Mixtral	0.433	0.291	0.340
+ “serial” explanations	0.407	0.265	0.306
+ “serial” explicit explanations	0.382	0.246	0.286
+ “parallel” explanations	0.339	0.258	0.295
+ “parallel” explicit explanations	<b>0.245</b>	<b>0.211</b>	<b>0.239</b>
$p_{\text{smax}}$ of Mixtral	0.434	0.292	0.342
+ “serial” explanations	0.349	0.258	0.296
+ “serial” explicit explanations	0.305	0.235	0.269
+ “parallel” explanations	0.310	0.255	0.290
+ “parallel” explicit explanations	<b>0.217</b>	<b>0.208</b>	<b>0.232</b>
<i>MJDs from Llama3</i>			
$p_{\text{norm}}$ of Llama3	0.259	0.262	0.284
+ “serial” explanations	0.255	0.259	0.281
+ “serial” explicit explanations	<b>0.235</b>	<b>0.247</b>	<b>0.266</b>
+ “parallel” explanations	0.257	0.261	0.283
+ “parallel” explicit explanations	0.243	0.253	0.273
$p_{\text{smax}}$ of Llama3	0.231	0.245	0.260
+ “serial” explanations	0.226	0.243	0.258
+ “serial” explicit explanations	<b>0.212</b>	<b>0.232</b>	<b>0.245</b>
+ “parallel” explanations	0.226	0.245	0.260
+ “parallel” explicit explanations	0.214	0.237	0.254

Table 2: Distribution comparison results. MJDs and HJDs are compared on 341 overlapping instances.

LLMs reach similar scores overall, exceeding those of MNLI and VariErr distributions regardless of whether the transformation method is  $p_{\text{norm}}$  or  $p_{\text{smax}}$ . Mixtral benefits more from the additional information (KL/JSD/TVD drop more from its base), and the best setup uses explicit explanations.

**Parallel vs serial** Regarding bias considerations, the two LLMs exhibit slightly different patterns. “Serial” mode is better for Llama, while “parallel” suits Mixtral. “Serial” mode intuitively feels better because it provides multiple explanations at the same time, allowing models to estimate each label relative to all other labels and explanations. For example, LLMs may discriminate which explanation is more convincing and thus favor the label indicated by that explanation. We conduct an ablation study by adding 4, 3, 2, or 1 explanations at a time to LLM prompts. However, Figure 3 shows that adding more explanations to a prompt makes the Mixtral MJDs less similar to HJD but more for Llama3 MJDs. The distribution similarity be-

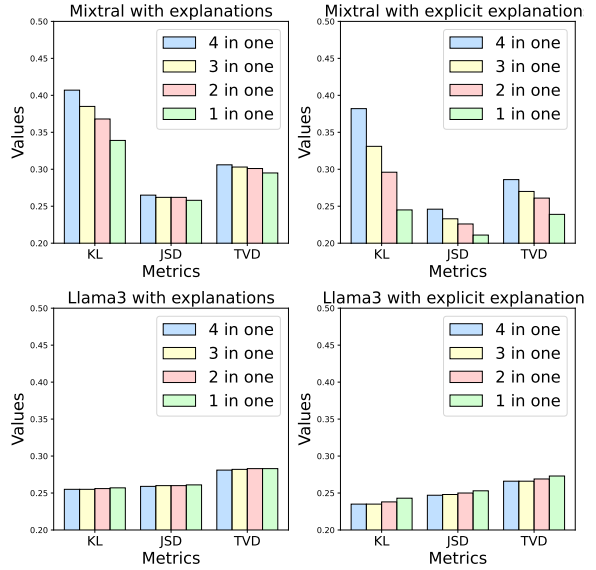


Figure 3: Distribution comparison results. “n in one” denotes the way LLMs process  $n$  explanations at a time.

tween Mixtral MJDs and HJDs seems to gradually decrease as the number of input explanations increases, whereas Llama3 shows almost no fluctuation and even performs better with longer texts. We hypothesize that Llama3 is better at longer prompts. More detailed scores are listed in Table 7 in Appendix D.

## 5.2 Fine-tuning Comparison

We present results on fine-tuning smaller models and comparing their MJDs to the HJDs on the held-out Chaos NLI dev and test sets. As comparison to LLM-predicted HJDs, we also train models on the existing datasets; see Table 3 for the results. All detailed scores of fine-tuning comparison are in Table 8 and Table 9 in Appendix E.

Models trained on the Chaos NLI train set perform best, noticeably higher than MNLI and VariErr models. Regarding the KL and CE Loss metrics, both LLMs demonstrate strong approximation performance, approaching Chaos NLI more closely than MNLI or VariErr NLI. Mixtral gets slightly better KL and CE Loss on “parallel”, while Llama3 wins on “serial”, mirroring results in §5.1.

However, fine-tuned LLM models show divergent results on F1. Overall, adding explicit explanations contributes to the best models. Llama3 improves fine-tuning results by yielding closer MJD to HJD and achieving better F1 scores than MNLI/VariErr HJDs. In contrast, while Mixtral only achieves slightly worse results in distribution comparison, it is much inferior in F1 score, even be-

Distributions	BERT FT (dev / test)			RoBERTa FT (dev / test)		
	Weighted F1 ↑	KL ↓	CE Loss ↓	Weighted F1 ↑	KL ↓	CE Loss ↓
<i>Baseline</i>						
Chaos NLI train set	<b>0.626 / 0.646</b>	<b>0.074 / 0.077</b>	<b>0.972 / 0.974</b>	<b>0.699 / 0.650</b>	<b>0.061 / 0.067</b>	<b>0.932 / 0.943</b>
MNLI single label	0.561 / 0.589	0.665 / 0.704	2.743 / 2.855	0.635 / 0.603	0.844 / 0.867	3.281 / 3.344
MNLI distributions	0.546 / 0.543	0.099 / 0.102	1.046 / 1.048	0.613 / 0.604	0.100 / 0.096	1.047 / 1.029
VariErr distributions	0.557 / 0.559	0.179 / 0.186	1.286 / 1.299	0.617 / 0.589	0.174 / 0.197	1.269 / 1.333
<i>MJDs from Mixtral</i>						
$p_{\text{norm}}$ of Mixtral	0.416 / 0.422	0.134 / 0.133	1.152 / 1.142	0.486 / 0.466	0.123 / 0.127	1.118 / 1.123
+ “serial” explanations	0.443 / 0.454	0.145 / 0.141	1.183 / 1.166	0.509 / 0.514	0.128 / 0.128	1.132 / 1.126
+ “serial” explicit explanations	0.506 / 0.511	0.130 / 0.130	1.139 / 1.132	<b>0.569 / 0.572</b>	0.114 / 0.122	1.091 / 1.107
+ “parallel” explanations	0.404 / 0.428	0.134 / 0.131	1.150 / 1.136	0.483 / 0.502	0.123 / 0.122	1.118 / 1.109
+ “parallel” explicit explanations	<b>0.507 / 0.514</b>	<b>0.108 / 0.108</b>	<b>1.074 / 1.065</b>	0.558 / 0.565	<b>0.092 / 0.098</b>	<b>1.025 / 1.037</b>
$p_{\text{sfnorm}}$ of Mixtral	0.427 / 0.432	0.131 / 0.129	1.140 / 1.130	0.497 / 0.472	0.121 / 0.125	1.112 / 1.118
+ “serial” explanations	0.452 / 0.462	0.121 / 0.118	1.113 / 1.096	0.506 / 0.525	0.110 / 0.109	1.078 / 1.069
+ “serial” explicit explanations	0.509 / <b>0.520</b>	0.105 / 0.105	1.064 / 1.057	<b>0.568 / 0.573</b>	0.093 / 0.098	1.026 / 1.036
+ “parallel” explanations	0.397 / 0.429	0.121 / 0.119	1.112 / 1.098	0.497 / 0.505	0.110 / 0.111	1.079 / 1.074
+ “parallel” explicit explanations	<b>0.522 / 0.517</b>	<b>0.095 / 0.095</b>	<b>1.035 / 1.026</b>	0.567 / <b>0.576</b>	<b>0.082 / 0.087</b>	<b>0.994 / 1.003</b>
<i>MJDs from Llama3</i>						
$p_{\text{norm}}$ of Llama3	0.514 / 0.526	0.097 / 0.098	1.038 / 1.036	0.541 / 0.528	0.091 / 0.094	1.023 / 1.025
+ “serial” explanations	0.574 / 0.574	0.096 / 0.097	1.037 / 1.033	0.618 / 0.601	0.091 / 0.093	1.020 / 1.022
+ “serial” explicit explanations	0.578 / 0.574	<b>0.091 / 0.092</b>	<b>1.022 / 1.018</b>	0.634 / 0.598	<b>0.085 / 0.088</b>	<b>1.003 / 1.006</b>
+ “parallel” explanations	0.573 / 0.582	0.098 / 0.098	1.041 / 1.038	0.636 / 0.598	0.093 / 0.095	1.026 / 1.028
+ “parallel” explicit explanations	<b>0.582 / 0.586</b>	0.094 / 0.095	1.030 / 1.026	<b>0.639 / 0.620</b>	0.089 / 0.091	1.014 / 1.016
$p_{\text{sfnorm}}$ of Llama3	0.528 / 0.524	0.091 / 0.093	1.023 / 1.021	0.546 / 0.535	0.085 / 0.089	1.005 / 1.009
+ “serial” explanations	0.567 / 0.576	0.091 / 0.091	1.021 / 1.016	0.626 / 0.608	0.082 / 0.086	0.996 / 1.000
+ “serial” explicit explanations	<b>0.585 / 0.568</b>	<b>0.086 / 0.087</b>	<b>1.008 / 1.004</b>	<b>0.646 / 0.610</b>	<b>0.077 / 0.081</b>	<b>0.981 / 0.987</b>
+ “parallel” explanations	0.584 / <b>0.583</b>	0.092 / 0.093	1.024 / 1.020	0.643 / 0.611	0.085 / 0.089	1.004 / 1.008
+ “parallel” explicit explanations	0.581 / 0.578	0.088 / 0.089	1.014 / 1.010	0.645 / <b>0.621</b>	0.081 / 0.085	0.993 / 0.996

Table 3: Results of fine-tuning comparison on Chaos NLI dev/test set. The KL and Cross-Entropy (CE) Loss reflected the distance between distributions, whereas Weighted F1 reflected the capability in handling NLI problems.

low that of MNLI/VariErr HJDs. The next section investigates this discrepancy in LLM performances between distribution and fine-tuning comparisons.

## 6 Analysis and Discussion

We observe from §5 that even though Llama3 and Mixtral are equally similar to Chaos NLI HJDs in distribution comparisons (see Table 2), their fine-tuned F1 scores differ, with Llama3 MJDs achieving a higher F1 than MNLI/VariErr HJDs while Mixtral being lower (see Table 3). We further inspect these distributions visually (§6.1) and using distance correlation (§6.2). We then explore the potential causes for the observed differences (§6.3). Finally, we highlight the contributions of this paper and suggest future directions for reference (§6.4).

### 6.1 Visualization

To inspect MJDs against HJD, we use the visualization tool by Gruber et al. (2024) to plot each instance’s distribution in the ternary plot. For consistency, we focus on a single setting with “explicit

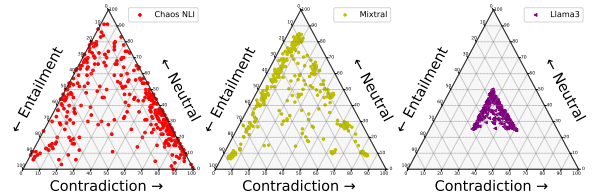


Figure 4: Visualization of distributions in ternary plot. Each point represents one of the 341 samples.

explanations” prompt, “parallel” mode, and  $p_{\text{norm}}$  transformation. Figure 4 compares the Chaos NLI HJD to Mixtral and Llama3’s MJDs. Interestingly, Llama3 and Mixtral exhibit rather different clusters: Llama3 has an overall higher entropy with instances closer to the center, whereas Mixtral is seemingly closer to the Chaos NLI HJD. We also see that the original Chaos NLI HJD is slightly skewed towards Contradiction, i.e., the right side of the triangle, while Mixtral MJD is slightly skewed towards Entailment, i.e., the opposite left side. Figure 5 further zooms in on Llama3 MJD and shows that Llama3 is slightly skewed towards the right

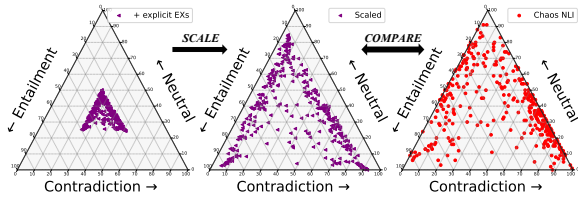


Figure 5: Zooming in ( $scale=3.3$ ) on Llama3 MJD.

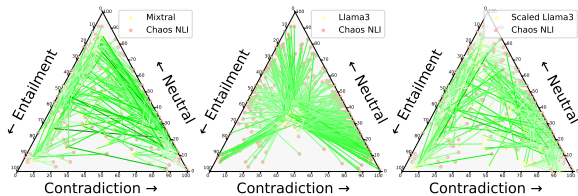


Figure 6: Error visualization. Corresponding sample points from the MJD and Chaos NLI HJD connected by a green line. Darker green means more distant.

side (Contradiction), more in line with Chaos NLI.

**Pairwise distance plots** While triangle plots provide insights on the overall distributions, they lack information regarding *how far* a particular instance (sample point) is in MJD, e.g.,  $[E=0.4, N=0.5, C=0.1]$  from its correspondence in the HJD. Therefore, we calculate and visualize pair-wise distances (or errors) between the corresponding sample points in MJDs and Chaos NLI HJDs. Namely, if many samples “moved” dramatically from an HJD to an MJD, this is a worse MJD estimation.

Figure 6 illustrates that Mixtral MJD has more substantial movements, i.e., erroneous estimates, than Llama3 MJD to Chaos NLI HJD. Figure 7 in Appendix F further provides pairwise error comparisons and shows that the absolute errors of Mixtral are scattered while Llama3 errors are more concentrated. These observations motivate us to look for a more suitable error measure in §6.2.

## 6.2 Quantifying the Visual Observations: Distance Correlation

Visualizations clearly show that, compared to Mixtral, the MJDs produced by Llama3 exhibit shapes more similar to HJDs, which corroborates Llama’s superior performance in fine-tuning comparisons. Given that the previous metrics used in *distribution comparison* are focusing on instance-level, while visualizations represent the distribution of all data points, we propose to further evaluate MJDs against HJDs using a global-level measure, distance correlation (D.Corr, Székely et al. 2007), to capture the differences between general distributions.

Distributions\Metrics	D.Corr $\uparrow$
Uniform distribution	0
MNLI single label	0.612
MNLI distributions	0.795
VariErr distributions	0.688
<i>MJDs from Mixtral</i>	
$p_{norm}$ of Mixtral	0.609
+ “parallel” explicit explanations	<b>0.719</b>
<i>MJDs from Llama3</i>	
$p_{smax}$ of Mixtral	0.593
+ “parallel” explicit explanations	<b>0.709</b>
<i>MJDs from Llama3</i>	
$p_{norm}$ of Llama3	0.689
+ “parallel” explicit explanations	<b>0.809</b>
$p_{smax}$ of Llama3	0.677
+ “parallel” explicit explanations	<b>0.802</b>

Table 4: Distance Correlation (D.Corr) on distribution comparisons between MJDs and Chaos NLI HJD. A higher correlation indicates better performance.

We consider all the samples’ soft labels from a dataset as a 3-D array. The D.Corr between the source dataset  $X$  and the target dataset  $Y$  is calculated as:

$$dCor^2(X, Y) = \frac{dCov^2(X, Y)}{\sqrt{dVar^2(X) dVar^2(Y)}}. \quad (4)$$

where  $dCov^2(X, Y)$  means the distance covariance of the two arrays, and  $dVar$  means the distance standard deviation of the array.

This measure accounts for all soft labels across the dataset and thus should be considered a global measure of the overall datasets. Results from Table 4 show that Llama3 MJDs have a substantially higher distance correlation with Chaos NLI’s HJD than Mixtral. This further proves Llama3 is globally better aligned with the HJD than Mixtral and supports its better fine-tuning performances. Appendix I shows more results with different temperatures  $\tau$  and numbers of explanations.

## 6.3 Why did the Llama3 MJD Work Better than Mixtral’s in Fine-tuning?

To sum up, both the visualization in §6.1 and distance correlations in §6.2 provide additional insights into the differences between Mixtral’s and Llama3’s MJDs. We hypothesize that one advantage of the Llama3-generated soft labels is its **smoothness**. As the benefit of label smoothness for model training has been validated extensively in the past (Müller et al., 2019; Wei et al., 2022a), when we observe that Llama3’s MJD has a smoother



distribution, we then assume that smoother labels could be beneficial based on intuition from previous papers. The overall higher entropy of the Llama3 non-scaled MJD (used for fine-tuning) has a regularizing effect similar to label smoothing. In other words, Llama3-generated MJD looks similar to the temperature-scaled (i.e., squished, within a certain range of  $\tau$ ) version of Chaos NLI. Appendix H further visualizes  $p_{\text{smax}}$ -transformed Llama3 MJDs with  $\tau$  in [5, 10, 20]. From the D.Corr results in Table 11 in Appendix I, we indeed found that higher  $\tau$ -values (smoother label distributions) can lead to better scores with  $\tau$  from 5, 10 to 20. This somehow supports our hypothesis that smoothness might be beneficial.

It is important to note that instance-level distribution distance measures such as KL and JSD cannot reliably predict models' performance when fine-tuned on the generated MJDs. A lower KL divergence cannot guarantee that it contains more information for fine-tuning. In the most extreme case, the KL divergence between a uniform random distribution and Chaos NLI is 0.364, much lower than the one-hot distribution on MNLI (0.665), while providing no helpful information for training. By further inspecting the visualization result, we hypothesize that the second advantage of the Llama3-generated soft labels is the **shape** of the sample distribution. That means the cross-sample dependency of the Llama3-generated soft labels is more similar to Chaos NLI than that of Mixtral. One way to capture this cross-sample dependency is to calculate a *distance matrix* recording the distances between every sample pair in the MJD.

To compare the global level similarity between the generated and the target dataset, we thus proposed to use visual inspections and measure the distance correlation, which measures the distance matrix in the covariance. We empirically show that distance correlation can better reflect the effectiveness of the generated MJDs for fine-tuning. This led us to conclude that metrics like KL, JSD, and TVD, which measure the distance between distributions at the instance level, are better complemented by additional investigations on the shape of the resulting annotations using visualization techniques and global measures.

#### 6.4 Potential benefits beyond NLI

In this study we validate the effectiveness of our approach on NLI. Given that NLI requires different linguistic capabilities (Wang et al., 2019) such as,

e.g., reasoning about word order, or understanding active/passive voice, we strongly believe that our findings generalize to other NLP tasks requiring similar capabilities. We aim to extend our study to other tasks in future work.

Furthermore, for a large number of tasks, constructing HLV datasets requires extensive annotation by crowd workers to obtain HJDs. This paper explores a method that only needs a small number of explanations (reasons why annotators made certain annotations) to approximate labor-intensive HJDs using LLMs. If datasets for more tasks include reasons for annotators' choices, our approach can be applied to approximate HJDs using LLMs for tasks such as sentiment analysis, stance detection and hate speech, etc., where there also exists disagreement and need for HJDs, see for example (Sandri et al., 2023). This would allow researchers to better explore the impact of HLV.

On a broader scale, research into HLV is akin to aligning current machine models with human values and addressing discrepancies in human viewpoints, different interpretations, and aligning them with machine judgment.

## 7 Conclusion

This paper analyzes to what extent LLMs can approximate human judgment distributions from a few explanations. Our results show that a few explanations improve LLM's ability to approximate HJDs. However, measuring the distance of the resulting MJD is insufficient: while similar in distance, their resulting fine-tuned models and visualized distributions differ substantially. We adopt an error visualization tool and a global-level metric, aligning our distribution and fine-tuning results. Our method can also be extended to other tasks beyond NLI, and we encourage an uptake of explanation-informed datasets.

## Limitations

Approximating human judgment distributions from a few explanations is a challenging task. The generalizability of LLMs empowers our approach to transform textual data (prompts with comments) into numeric form (MJDs) and thus approximate the human label distribution. However, we are constrained to existing crowd-annotated NLI datasets as our approximation target, which is not necessarily the best or most representative human label distribution. In future it would be interesting to test

our approach on diverse HJDs. For the same reason, our fine-tuning experiments use the standard soft-label training, which is one of the most prominent but not the only HLV-embracing procedure.

There are further considerations we could take into account in our experimental design. Firstly, the VariErr dataset also includes a second round of validity judgments regarding the explanations to identify erroneous labels. In the current study, this was not in scope. We could further leverage these quality judgments as features to subset better and worse explanations and investigate LLMs' performances when fed with different explanation qualities. Secondly, we could further explore the temperature  $\tau$  or other normalization ways to scale or zoom in to better understand the differences in shapes between HJDs and MJDs. Thirdly, traditional divergence metrics, such as KL, JSD, and TVD, are not particularly designed to measure global-level differences between two distributions, especially concerning the distribution shapes. While we proposed one distance correlation measure to address this, there can be other metrics to measure these macroscopic differences. Similarly, error visualizations partially prove our hypothesis but alone do not provide quantifiable evidence.

## Acknowledgements

We thank the members of the MaiNLP lab for their insightful feedback on earlier drafts of this paper. We specifically appreciate the suggestions of Philipp Mondorf, Bolei Ma, Kassem Sabeh, Verena Blaschke, Diego Frassinelli, Sondre Wold, Jian Lan, and Rob van der Goot. BC acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program. We are also grateful to the anonymous reviewers for their constructive feedback. This research is supported by ERC Consolidator Grant DIALECT 101043235, and UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1.

## References

- Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2024. [MEGAVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks](#). *Preprint*, arXiv:2311.07463.
- Abhishek Anand, Negar Mokherian, Prathyusha Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. [Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations](#). In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 102–113, St Julians, Malta. Association for Computational Linguistics.
- 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.
- Simone Balloccu, Patrícia Schmidová, Mateusz Lango, and Ondrej Dusek. 2024. [Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.
- Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. [LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders](#). *Preprint*, arXiv:2404.05961.
- 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.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- 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.

- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Trans. Assoc. Comput. Linguistics*, 10:92–110.
- 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.
- Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünnler. 2023. [Questioning the Survey Responses of Large Language Models](#). *CoRR*, abs/2306.07951.
- Dominik Maria Endres and Johannes E. Schindelin. 2003. [A new metric for probability distributions](#). *IEEE Trans. Inf. Theory*, 49(7):1858–1860.
- Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. [“you are an expert linguistic annotator”](#): Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore. Association for Computational Linguistics.
- Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021a. [Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2591–2597. Association for Computational Linguistics.
- Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021b. [Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. 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.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. [Mistral 7B](#). *Preprint*, arXiv:2310.06825.
- 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, Léo 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. [Mistral of Experts](#). *CoRR*, abs/2401.04088.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022a. [Investigating reasons for disagreement in natural language inference](#). *Transactions of the Association for Computational Linguistics*, 10:1357–1374.
- Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023b. [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.
- Nanjiang Jiang and Marie-Catherine de Marneffe. 2022b. [Investigating reasons for disagreement in natural language inference](#). *Trans. Assoc. Comput. Linguistics*, 10:1357–1374.
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Andrew K. Lampinen, Ishita Dasgupta, Stephanie C. Y. Chan, Kory W. Mathewson, Michael Henry Tessler, Antonia Creswell, James L. McClelland, Jane Wang, and Felix Hill. 2022. [Can language models learn](#)



- from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 537–563. Association for Computational Linguistics.
- Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen White, and Sujay Kumar Jauhar. 2023a. **Making large language models better data creators**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15349–15360. Association for Computational Linguistics.
- Noah Lee, Na An, and James Thorne. 2023b. **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.
- Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. **SemEval-2023 task 11: Learning with disagreements (LeWiDi)**. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. 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.
- Christopher D. Manning. 2006. **Local textual inference : It’s hard to circumscribe , but you know it when you see it - and nlp needs it**.
- Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. <https://ai.meta.com/blog/meta-llama-3/>.
- Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. **When does label smoothing help?** In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4696–4705.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. **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.
- Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. **Proving test set contamination in black-box language models**. In *The Twelfth International Conference on Learning Representations*.
- Ellie Pavlick and Tom Kwiatkowski. 2019. **Inherent disagreements in human textual inferences**. *Trans. Assoc. Comput. Linguistics*, 7:677–694.
- Maja Pavlovic and Massimo Poesio. 2024. **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.
- 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10671–10682. 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.
- Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. **Why don’t you do it right? analysing annotators’ disagreement in subjective tasks**. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2420–2433. Association for Computational Linguistics.
- 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.
- Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2023. **Do LLMs exhibit human-like response biases? A case study in survey design**. *CoRR*, abs/2311.04076.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu,



- Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- 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.
- Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. [Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models](#). *Preprint*, arXiv:2404.18796.
- Manya Wadhwa, Jifan Chen, Junyi Jessy Li, and Greg Durrett. 2023. [Using Natural Language Explanations to Rescale Human Judgments](#). *CoRR*, abs/2305.14770.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *International Conference on Learning Representations*.
- Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. [VariErr NLI: Separating Annotation Error from Human Label Variation](#). *CoRR*, abs/2403.01931.
- Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. 2022a. [Mitigating neural network overconfidence with logit normalization](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23631–23644. PMLR.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- 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.
- Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. [A Survey on Knowledge Distillation of Large Language Models](#). *CoRR*, abs/2402.13116.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. [A Survey of Large Language Models](#). *CoRR*, abs/2303.18223.
- 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*.
- Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. [Distributed NLI: Learning to predict human opinion distributions for language reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.

## A Prompt

All prompts used in this paper are listed in Table 5. We adopted the original chat templates for both LLMs to input the prompt.

## B Experimental Implementation

We first fine-tuned BERT-base-uncased (Devlin et al., 2019) and Roberta-base (Liu et al., 2019) with the standard NLI training process on MNLI single labels (Williams et al., 2018), and then fine-tuned them on the label distributions of MNLI, VariErr NLI (Weber-Genzel et al., 2024) and Chaos NLI (Nie et al., 2020), as well as the MJDs generated by the LLMs. We used cross-entropy as a loss function for soft-label training. For validation on the dev set, we measured the distribution distance by calculating KL divergence and cross-entropy between the logits of the model and the soft label. We also measured the prediction performance by calculating the F1 score using the largest logits of the model against the majority-voted label. We selected the model with the best macro-F1 score performance on the dev set for final testing. Detailed hyperparameter choices are listed in Table 6. Fine-tuning was conducted with NVIDIA A100 80GB within several hours.

## C Metrics

**Kullback-Leibler Divergence** Kullback-Leibler divergence, often referred to as KL divergence (KL), is a measure of how one probability distribution diverges from a second, reference probability distribution (Kullback and Leibler, 1951). It is

Type	General Instruction Prompt
Without explanations	<p><b>"role": "user", "content":</b>  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:</p>
With explanations	<p><b>"role": "user", "content":</b>  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}  Comment 2: {explanation 2}  ...  A. Entailment  B. Neutral  C. Contradiction.  Answer:</p>
With explicit explanations	<p><b>"role": "user", "content":</b>  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:</p>

Table 5: Instruction prompt of different types to transform NLI into a multi-choice question format.

useful for capturing the relative entropy or information loss when approximating one distribution with another. It is a non-symmetric measure of the difference between two probability distributions  $P$  and  $Q$ .

KL divergence is very sensitive to differences between the two distributions. If there are points where one distribution assigns a high probability and the other assigns a low probability, KL divergence will highlight these differences significantly.

For discrete probability distributions  $P$  and  $Q$ :

$$D_{\text{KL}}(P|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}, \quad (5)$$

For continuous probability distributions, the sum

Hyperparameter	Our Model
Learning Rate Decay	Linear
Weight Decay	0.0
Optimizer	AdamW
Adam $\epsilon$	1e-8
Adam $\beta_1$	0.9
Adam $\beta_2$	0.999
Warmup Ratio	0%
Learning Rate	2e-5
Batch size	4
Num Epoch	5

Table 6: Hyperparameter used for fine-tuning BERT and RoBERTa models.

is replaced by an integral:

$$D_{\text{KL}}(P|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx, \quad (6)$$

In this paper, we adopted the discrete version, as our target distributions are 3-dimensional probabilities. KL divergence is asymmetric ( $D_{\text{KL}}(P|Q) \neq D_{\text{KL}}(Q|P)$ ), which can be beneficial when the distributions have a clear direction of reference or when one distribution is considered the true distribution and the other is an approximation. Thus, we set Chaos NLI HJD as the true distribution  $P$ , and MJDs as the approximation  $Q$ .

**Jensen-Shannon Distance** Jensen-Shannon distance (JSD) is a symmetric and smoothed version of the KL divergence (Endres and Schindelin, 2003). Unlike KL divergence, Jensen-Shannon distance is symmetric ( $D_{\text{JS}}(P|Q) = D_{\text{JS}}(Q|P)$ ). This makes it suitable when there is no inherent reference direction between the two distributions. Jensen-Shannon distance is always finite and bounded between 0 and 1, making it easier to interpret and compare. By averaging the two distributions, it mitigates the impact of any extreme values, providing a more stable measure of similarity.

For discrete probability distributions  $P$  and  $Q$ , let  $M = \frac{1}{2}(P + Q)$ . The Jensen-Shannon divergence is defined as:

$$D_{\text{JS}}(P|Q) = \sqrt{\frac{(D_{\text{KL}}(P|M) + D_{\text{KL}}(Q|M))}{2}}, \quad (7)$$

Same as KL, we also set HJD from Chaos NLI as  $P$ , while MJDs as  $Q$ .

**Total Variation Distance** Total Variation distance (TVD) is a measure of the maximum difference between the probabilities assigned to the same event by two different probability distributions (Devroye and Lugosi, 2001). It provides a way to quantify the difference between two distributions.

TV distance can be interpreted as the maximum proportion of the distribution that needs to be altered to transform one distribution into the other. This makes it an intuitive measure of overall dissimilarity. Unlike KL divergence and JS distance, which involve logarithms and averages, TV distance is based on absolute differences. This can be particularly useful when you need a straightforward measure of discrepancy. TV distance is robust to small changes in probability values, making it a reliable measure when comparing distributions that may have minor variations.

For discrete probability distributions  $P$  and  $Q$ :

$$D_{\text{TV}}(P, Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|, \quad (8)$$

For continuous probability distributions, the sum is replaced by an integral:

$$D_{\text{TV}}(P, Q) = \frac{1}{2} \int_{-\infty}^{\infty} |p(x) - q(x)| dx, \quad (9)$$

In this paper, we still used the discrete version, as same as KL. We set HJD from Chaos NLI as  $P$ , while MJDs as  $Q$ .

**Weighted F1 Score** The weighted F1 score is an extension of the standard F1 score that accounts for class imbalance in multi-class classification problems. In multi-class classification, different classes can have varying frequencies, and the weighted F1 score adjusts for this by giving more importance to classes that have more instances.

The F1 score is the harmonic mean of precision and recall:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (10)$$

where precision (P) is the proportion of true positives among all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (11)$$

and recall (R) is the proportion of true positives among all actual positives:

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (12)$$

In a multi-class setting, we calculate the F1 score for each class, then take a weighted average based on the number of true instances of each class. This alters macro F1 Score to account for label imbalance; it can result in an F-score that is not between precision and recall.

The formula for the weighted F1 score is:

$$\text{Weighted F1} = \frac{1}{N} \sum_{i=1}^k w_i \times F1_i. \quad (13)$$

where  $k$  is the total number of classes,  $F1_i$  is the F1 score for class  $i$ ,  $w_i$  is the weight for class  $i$ , which is proportional to the number of true instances of class  $i$  ( $w_i = \frac{n_i}{N}$ , where  $n_i$  is the number of true instances of class  $i$ , and  $N$  is the total number of instances across all classes.).

The weighted F1 score is particularly useful when dealing with imbalanced datasets because it adjusts the contribution of each class’s F1 score based on how common the class is. This prevents the model from being overly influenced by the performance on the majority class, which can be misleading in an imbalanced dataset. In this paper, the weighted F1 score is implemented with *sklearn*<sup>5</sup>.

## D Detailed Result of Ablation Study

All results of the ablation study (“serial”/“parallel”) were listed in Table 7, which were depicted by Figure 3 in §4.3. To pursue maximizing the evaluation metrics, we can indeed make greater sacrifices in computational resources, especially since Llama3 has a strong capability to process all explanations at once. However, we have considered the possible sequence bias of explanations. To eliminate these biases we consider a shuffling setup, where the computational resource consumption will increase significantly as the number of explanations increases. Therefore, presenting both parallel and serial results is valuable.

## E Details of Fine-tuning Comparison

Here we provided more details regarding the results of *fine-tuning* comparison, as listed in Table 8 and 9, which were the complete versions of Table 3.

<sup>5</sup><https://scikit-learn.org/>

Distributions	$p_{\text{norm}}$			$p_{\text{sfmax}}$		
	KL ↓	JSD ↓	TVD ↓	KL ↓	JSD ↓	TVD ↓
Mixtral original	0.433	0.291	0.340	0.434	0.292	0.342
4 explanations at a time	0.407	0.265	0.306	0.349	0.258	0.296
3 explanations at a time	0.385	0.262	0.303	0.338	0.257	0.295
2 explanations at a time	0.368	0.262	0.301	0.329	0.257	0.293
1 explanations at a time	<b>0.339</b>	<b>0.258</b>	<b>0.295</b>	<b>0.310</b>	<b>0.255</b>	<b>0.290</b>
4 explicit explanations at a time	0.382	0.246	0.286	0.305	0.235	0.269
3 explicit explanations at a time	0.331	0.233	0.270	0.275	0.226	0.257
2 explicit explanations at a time	0.296	0.226	0.261	0.254	0.221	0.251
1 explicit explanations at a time	<b>0.245</b>	<b>0.211</b>	<b>0.239</b>	<b>0.217</b>	<b>0.208</b>	<b>0.232</b>
Llama3 original	0.259	0.262	0.284	0.231	0.245	0.260
4 explanations at a time	<b>0.255</b>	<b>0.259</b>	<b>0.281</b>	0.226	<b>0.243</b>	0.258
3 explanations at a time	0.255	0.260	0.282	0.225	0.243	0.258
2 explanations at a time	0.256	0.260	0.283	<b>0.224</b>	0.243	<b>0.258</b>
1 explanation at a time	0.257	0.261	0.283	0.226	0.245	0.260
4 explicit explanations at a time	<b>0.235</b>	<b>0.247</b>	<b>0.266</b>	0.212	0.232	0.245
3 explicit explanations at a time	0.235	0.248	0.266	0.209	<b>0.232</b>	<b>0.244</b>
2 explicit explanations at a time	0.238	0.250	0.269	<b>0.209</b>	0.233	0.245
1 explicit explanation at a time	0.243	0.253	0.273	0.214	0.237	0.254

Table 7: All results of “serial”/“parallel”. Scores are compared with Chaos NLI HJD. Note that since all the results are averaged scores of  $A^{(m)}$  combinations (described in §3.3), which means LLMs actually obtain all 4 explanations’ information in every setting. “Serial” represents “4 explanations at a time”, while “parallel” represents “1 explanations at a time”. For the settings of 2 and 3, they are the transition form from “serial” to “parallel”.



Distributions	BERT FT					RoBERTa FT				
	ACC. $\uparrow$	KL $\downarrow$	Weighted F1 $\uparrow$	Macro F1 $\uparrow$	CE Loss $\downarrow$	ACC. $\uparrow$	KL $\downarrow$	Weighted F1 $\uparrow$	Macro F1 $\uparrow$	CE Loss $\downarrow$
<i>Baseline</i>										
Chaos NLI train set	<b>0.628</b>	<b>0.074</b>	<b>0.626</b>	<b>0.566</b>	<b>0.972</b>	<b>0.698</b>	<b>0.061</b>	<b>0.699</b>	<b>0.659</b>	<b>0.932</b>
MNLI single label	0.552	0.665	0.561	0.523	2.743	0.628	0.844	0.635	0.616	3.281
MNLI distributions	0.542	0.099	0.546	0.518	1.046	0.607	0.100	0.613	0.598	1.047
VariErr distributions	0.558	0.179	0.557	0.514	1.286	0.617	0.174	0.617	0.594	1.269
<i>MJDs from Mixtral</i>										
$p_{\text{norm}}$ of Mixtral	0.404	0.134	0.416	0.400	1.152	0.464	0.123	0.486	0.451	1.118
+ "serial" explanations	0.459	0.145	0.443	0.452	1.183	0.506	0.128	0.509	0.502	1.132
+ "serial" explicit explanations	<b>0.515</b>	0.130	0.506	0.492	1.139	<b>0.569</b>	0.114	<b>0.569</b>	<b>0.554</b>	1.091
+ "parallel" explanations	0.409	0.134	0.404	0.407	1.150	0.472	0.123	0.483	0.472	1.118
+ "parallel" explicit explanations	0.514	<b>0.108</b>	<b>0.507</b>	<b>0.492</b>	<b>1.074</b>	0.558	<b>0.092</b>	0.558	0.542	<b>1.025</b>
$p_{\text{sfmax}}$ of Mixtral	0.413	0.131	0.427	0.408	1.140	0.477	0.121	0.497	0.460	1.112
+ "serial" explanations	0.467	0.121	0.452	0.459	1.113	0.504	0.110	0.506	0.498	1.078
+ "serial" explicit explanations	0.517	0.105	0.509	0.495	1.064	0.566	0.093	<b>0.568</b>	0.549	1.026
+ "parallel" explanations	0.404	0.121	0.397	0.402	1.112	0.483	0.110	0.497	0.482	1.079
+ "parallel" explicit explanations	<b>0.528</b>	<b>0.095</b>	<b>0.522</b>	<b>0.502</b>	<b>1.035</b>	<b>0.568</b>	<b>0.082</b>	0.567	<b>0.549</b>	<b>0.994</b>
<i>MJDs from Llama3</i>										
$p_{\text{norm}}$ of Llama3	0.556	0.097	0.514	0.473	1.038	0.593	0.091	0.541	0.505	1.023
+ "serial" explanations	0.561	0.096	0.574	0.534	1.037	0.610	0.091	0.618	0.580	1.020
+ "serial" explicit explanations	0.571	<b>0.091</b>	0.578	0.536	<b>1.022</b>	0.630	<b>0.085</b>	0.634	0.589	<b>1.003</b>
+ "parallel" explanations	0.561	0.098	0.573	0.531	1.041	0.634	0.093	0.636	0.594	1.026
+ "parallel" explicit explanations	<b>0.572</b>	0.094	<b>0.582</b>	<b>0.539</b>	1.030	<b>0.634</b>	0.089	<b>0.639</b>	<b>0.596</b>	1.014
$p_{\text{sfmax}}$ of Llama3	0.574	0.091	0.528	0.487	1.023	0.601	0.085	0.546	0.514	1.005
+ "serial" explanations	0.555	0.091	0.567	0.532	1.021	0.617	0.082	0.626	0.588	0.996
+ "serial" explicit explanations	<b>0.577</b>	<b>0.086</b>	<b>0.585</b>	<b>0.544</b>	<b>1.008</b>	<b>0.641</b>	<b>0.077</b>	<b>0.646</b>	0.604	<b>0.981</b>
+ "parallel" explanations	0.572	0.092	0.584	0.541	1.024	0.638	0.085	0.643	0.605	1.004
+ "parallel" explicit explanations	0.572	0.088	0.581	0.540	1.014	0.639	0.081	0.645	<b>0.605</b>	0.993

Table 8: All results for fine-tuning comparison on Chaos NLI dev set.

Distributions	BERT FT					RoBERTa FT				
	ACC. $\uparrow$	KL $\downarrow$	Weighted F1 $\uparrow$	Macro F1 $\uparrow$	CE Loss $\downarrow$	ACC. $\uparrow$	KL $\downarrow$	Weighted F1 $\uparrow$	Macro F1 $\uparrow$	CE Loss $\downarrow$
<i>Baseline</i>										
Chaos NLI-M train set	<b>0.645</b>	<b>0.077</b>	<b>0.646</b>	<b>0.617</b>	<b>0.974</b>	<b>0.650</b>	<b>0.067</b>	<b>0.650</b>	<b>0.630</b>	<b>0.943</b>
MNLI single label	0.585	0.704	0.589	0.573	2.855	0.599	0.867	0.603	0.594	3.344
MNLI distributions	0.547	0.102	0.543	0.539	1.048	0.599	0.096	0.604	0.594	1.029
VariErr distributions	0.566	0.186	0.559	0.548	1.299	0.590	0.197	0.589	0.569	1.333
<i>MJDs from Mixtral</i>										
$p_{\text{norm}}$ of Mixtral	0.417	0.133	0.422	0.416	1.142	0.459	0.127	0.466	0.453	1.123
+ "serial" explanations	0.479	0.141	0.454	0.466	1.166	0.517	0.128	0.514	0.514	1.126
+ "serial" explicit explanations	0.523	0.130	0.511	0.509	1.132	<b>0.572</b>	0.122	<b>0.572</b>	<b>0.566</b>	1.107
+ "parallel" explanations	0.436	0.131	0.428	0.434	1.136	0.498	0.122	0.502	0.499	1.109
+ "parallel" explicit explanations	<b>0.526</b>	<b>0.108</b>	<b>0.514</b>	<b>0.510</b>	<b>1.065</b>	0.568	<b>0.098</b>	0.565	0.559	<b>1.037</b>
$p_{\text{sfmax}}$ of Mixtral	0.423	0.129	0.432	0.421	1.130	0.464	0.125	0.472	0.457	1.118
+ "serial" explanations	0.485	0.118	0.462	0.473	1.096	0.526	0.109	0.525	0.522	1.069
+ "serial" explicit explanations	<b>0.531</b>	0.105	<b>0.520</b>	<b>0.517</b>	1.057	0.574	0.098	0.573	0.567	1.036
+ "parallel" explanations	0.436	0.119	0.429	0.433	1.098	0.501	0.111	0.505	0.501	1.074
+ "parallel" explicit explanations	0.528	<b>0.095</b>	0.517	0.510	<b>1.026</b>	<b>0.577</b>	<b>0.087</b>	<b>0.576</b>	<b>0.567</b>	<b>1.003</b>
<i>MJDs from Llama3</i>										
$p_{\text{norm}}$ of Llama3	0.561	0.098	0.526	0.506	1.036	0.583	0.094	0.528	0.513	1.025
+ "serial" explanations	0.568	0.097	0.574	0.555	1.033	0.596	0.093	0.601	0.585	1.022
+ "serial" explicit explanations	0.571	<b>0.092</b>	0.574	0.550	<b>1.018</b>	0.595	<b>0.088</b>	0.598	0.581	<b>1.006</b>
+ "parallel" explanations	0.572	0.098	0.582	0.560	1.038	0.595	0.095	0.598	0.579	1.028
+ "parallel" explicit explanations	<b>0.580</b>	0.095	<b>0.586</b>	<b>0.560</b>	1.026	<b>0.615</b>	0.091	<b>0.620</b>	<b>0.598</b>	1.016
$p_{\text{sfmax}}$ of Llama3	0.564	0.093	0.524	0.507	1.021	0.595	0.089	0.535	0.524	1.009
+ "serial" explanations	0.569	0.091	0.576	0.556	1.016	0.603	0.086	0.608	0.591	1.000
+ "serial" explicit explanations	0.566	<b>0.087</b>	0.568	0.548	<b>1.004</b>	0.607	<b>0.081</b>	0.610	0.591	<b>0.987</b>
+ "parallel" explanations	<b>0.574</b>	0.093	<b>0.583</b>	<b>0.560</b>	1.020	0.606	0.089	0.611	0.590	1.008
+ "parallel" explicit explanations	0.574	0.089	0.578	0.551	1.010	<b>0.617</b>	0.085	<b>0.621</b>	<b>0.601</b>	0.996

Table 9: All results for fine-tuning comparison on Chaos NLI test set.

Metrics of KL, CE Loss, Accuracy, Weighted F1 and Macro F1 were included. The same pattern was observed as illustrated in §5.2.

## F Detailed Visualizations

As shown in Figure 7, we plotted the distributions and the absolute errors from Chaos NLI HLV to provide more details. When (explicit) explanations were added, the absolute errors of MJDs became smaller, meaning they got closer to HJD.

## G Data Leakage for LLMs

The lack of complete pre-training details on Large Language Models (LLMs) raises the potential risk of data contamination and the skepticism on LLM performance evaluation (Balloccu et al., 2024; BehnamGhader et al., 2024). To the best of our knowledge, there lacks concrete evidence showing either Mixtral-8x7b-Instruct-v0.1 (Jiang et al., 2024) or Llama3-Chat-70b (Meta, 2024) being contaminated by any of these tested datasets, MNLI (Williams et al., 2018), VariErr NLI (Wei et al., 2022b), and Chaos NLI (Nie et al., 2020). For the VariErr NLI dataset that provides explanations, its release date is later than that of Mixtral and Llama3, so it does not pose a leakage risk. Moreover, closely related to our datasets, Ahuja et al. (2024) conduct the Black Box contamination test (Oren et al., 2024) and show that XNLI (Conneau et al., 2018), a sub-sampled multilingual extension dataset of MNLI, did not leak data to either the Llama-2-7B (Touvron et al., 2023) or the Mistral-7B-Instruct (Jiang et al., 2023a) model. Since Mixtral or Llama have only released their weights, we do not have access to their training corpora. Therefore, we can only offer our perspectives and discussions on data leakage.

For the MNLI training part, to familiarize BERT and RoBERTa with the NLI task, we first fine-tuned the models on MNLI training set and then further fine-tuned them on our samples, which contain both HJDs and MJDs. Please note that in our fine-tuning comparison, the "gold" labels are derived from the distribution of 100 annotations in ChaosNLI, not the single labels from MNLI. Besides, the corpora we evaluated, 341 ChaosNLI training set and 629 dev/test set, were all extracted from MNLI-matched development set (Nie et al., 2020), which did not overlap with MNLI training set used for MNLI fine-tuning. Therefore, there is no risk of data leakage in this process.

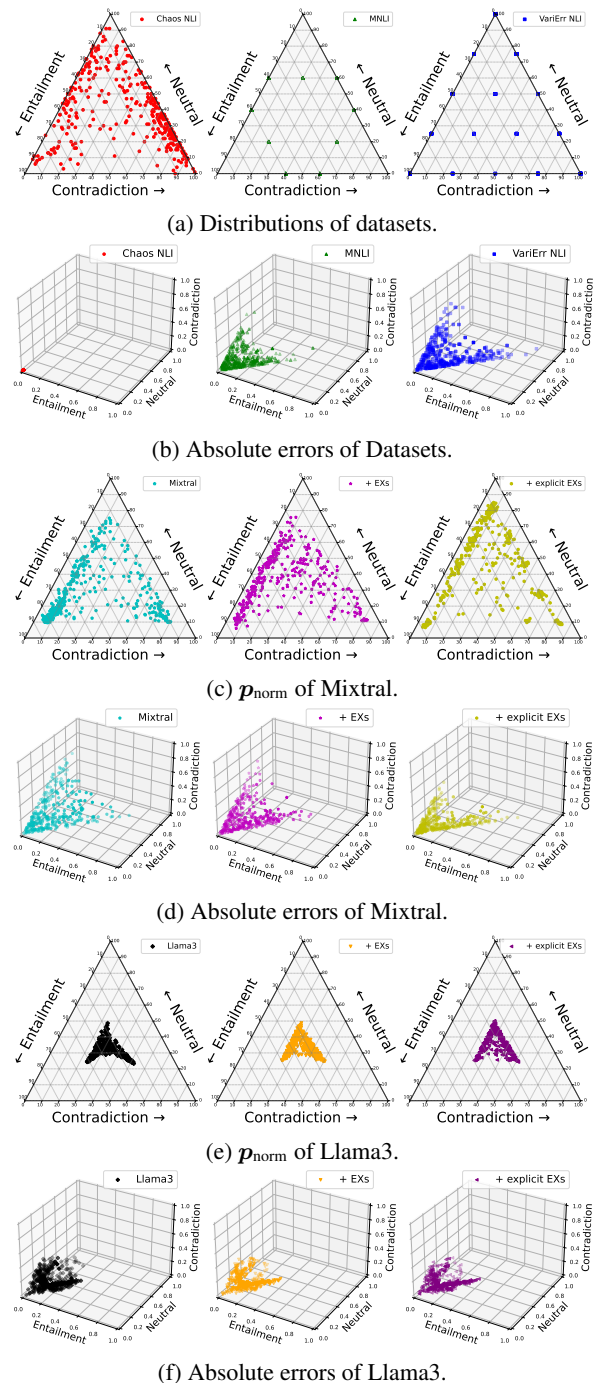


Figure 7: Visualization of distributions and absolute errors in ternary plot and 3D Cartesian coordinate system. Each point represents one sample from the valid overlapped 341 instances.

## H Discussion for Temperature in Softmax

Equation 3.2 is a standard softmax function used to convert the logits obtained from the LLM into a label probability distribution. Zhou et al. (2022) explores various methods for obtaining probability distributions, including Monte Carlo Dropout, among others. Since these methods are not significantly different from the conventional softmax, and exploring which normalization method is better is not the focus of this paper, we have directly adopted the most commonly used non-parametric normalization and parametric softmax approach. In the softmax function, the parameter  $\tau$  is often referred to as the temperature coefficient, which smooths the probability distribution (a smaller  $\tau$  makes the label distribution sharper, while a larger  $\tau$  makes it smoother). Typically, for converting LLM logits into a probability distribution,  $\tau$  should be set to a moderately sized value to prevent excessively small  $\tau$  values from sharpening the probability distribution, thereby potentially losing the HLV information we aim to obtain.

Thus, during the *distribution comparison*, we adopted  $\tau = 20$  to obtain  $p_{\text{softmax}}$  that can be compared with  $p_{\text{norm}}$ . In fact, for the metrics in Table 2, the results of  $p_{\text{softmax}}$  improve initially as  $\tau$  increases from 0, and then decline, as shown in our toy trials. However, we did not specifically adjust the value of  $\tau$  because it is not relevant to the focus of our study.  $p_{\text{norm}}$  does not have any parameters and still yields the desired results.

During our in-depth investigation at §6.1, we discovered that the value of  $\tau$  has an amplifying effect on the distribution shape within a certain range (not a true proportional amplification, but very similar within that range). Given our focus on exploring possible amplification patterns in the Llama3 MJDs, we adjusted  $\tau$  from 20 to 10 to 5. This adjustment allowed us to better observe the intrinsic shape of the Llama3 MJD and investigate its correlation with Chaos NLI HJD. All the “scaled” MJDs are shown in Figure 8, including distributions of Mixtral and Llama3 under different settings. Also, we conducted *fine-tuning comparison*, same as §5.2, and listed all results in Table 10, for further analysis.

The results show that as  $\tau$  decreases, the distribution becomes sharper, but this does not result in a linear change in metrics such as F1, KL divergence, and CE Loss. Since this parameter is highly dependent on the dataset and model, we did

not spend extensive computational resources on a broader exploration of  $\tau$  values.

## I All Results for Distance Correlation

All results for distance correlation (Székely et al., 2007) of MJDs, including  $p_{\text{norm}}$  and  $p_{\text{softmax}}$  with different  $\tau$ , are listed in Table 11 for further exploring. There are many methods to achieve smoothness, and finding the most suitable one for the task at hand goes beyond the current focus of this paper. We will delve deeper into how smoothness can enhance performance in future work.

## J “Assistant” Mode for Mixtral Prompt

Here we’d introduce a really interesting way to prompt. Because this approach lacks a complete theoretical basis and currently proves effective only in Mixtral, not in Llama3, we are presenting the concept here without providing complete experimental evidence. We have proposed a special prompt method for chat templates based on Mixtral, which we refer to as “assistant” mode. We hope that anyone reading this who is interested or knowledgeable about this prompt method will be encouraged to further explore this approach.

We expected that LLM could focus more on the explanations in the “assistant” mode, since it is the text form that contains rich human label variation information. So we put the explanations in the content of assistant role as “comments” as shown in Table 13. In that case, we divide the MCQA into two part, first to ask LLM about their opinion on the original NLI instances, then out human explanations in the position of LLM response (“assistant” role). LLM would assume that the comments are its own previous answers and take more account on them. And finally let LLM give a choice among three NLI labels. It’s a way to extract the potential of LLM’s understanding capability for human label variation. Results are shown in Table 12.

We observed that in this approach, LLM’s output becomes more sensitive to the input of explanations and seems to be more influenced by their content. However, our understanding is speculative. For example, Mixtral may emphasize “assistant” training during instruct tuning, or the architecture of a Mixture of Experts model might prioritize “assistant” content (as valuable global information to a specific expert). Yet, we lack the means to prove these speculations. Therefore, we present this as an interesting attempt in the appendix, acknowledging

Distributions	BERT FT (dev / test)			RoBERTa FT (dev / test)		
	Weighted F1 $\uparrow$	KL $\downarrow$	CE Loss $\downarrow$	Weighted F1 $\uparrow$	KL $\downarrow$	CE Loss $\downarrow$
<i>Baseline</i>						
Chaos NLI train set	<b>0.626 / 0.646</b>	<b>0.074 / 0.077</b>	<b>0.972 / 0.974</b>	<b>0.699 / 0.650</b>	<b>0.061 / 0.067</b>	<b>0.932 / 0.943</b>
MNLI single label	0.561 / 0.589	0.665 / 0.704	2.743 / 2.855	0.635 / 0.603	0.844 / 0.867	3.281 / 3.344
MNLI distributions	0.546 / 0.543	0.099 / 0.102	1.046 / 1.048	0.613 / 0.604	0.100 / 0.096	1.047 / 1.029
VariErr distributions	0.557 / 0.559	0.179 / 0.186	1.286 / 1.299	0.617 / 0.589	0.174 / 0.197	1.269 / 1.333
<i>MJDs from Mixtral</i>						
$p_{\text{norm}}$ of Mixtral	0.416 / 0.422	0.134 / 0.133	1.152 / 1.142	0.486 / 0.466	0.123 / 0.127	1.118 / 1.123
+ “serial” explanations	0.443 / 0.454	0.145 / 0.141	1.183 / 1.166	0.509 / 0.514	0.128 / 0.128	1.132 / 1.126
+ “serial” explicit explanations	0.506 / 0.511	0.130 / 0.130	1.139 / 1.132	<b>0.569 / 0.572</b>	0.114 / 0.122	1.091 / 1.107
+ “parallel” explanations	0.404 / 0.428	0.134 / 0.131	1.150 / 1.136	0.483 / 0.502	0.123 / 0.122	1.118 / 1.109
+ “parallel” explicit explanations	<b>0.507 / 0.514</b>	<b>0.108 / 0.108</b>	<b>1.074 / 1.065</b>	0.558 / 0.565	<b>0.092 / 0.098</b>	<b>1.025 / 1.037</b>
$p_{\text{sfmax}}$ of Mixtral ( $\tau = 20$ )	0.427 / 0.432	0.131 / 0.129	1.140 / 1.130	0.497 / 0.472	0.121 / 0.125	1.112 / 1.118
+ “serial” explanations	0.452 / 0.462	0.121 / 0.118	1.113 / 1.096	0.506 / 0.525	0.110 / 0.109	1.078 / 1.069
+ “serial” explicit explanations	0.509 / <b>0.520</b>	0.105 / 0.105	1.064 / 1.057	<b>0.568 / 0.573</b>	0.093 / 0.098	1.026 / 1.036
+ “parallel” explanations	0.397 / 0.429	0.121 / 0.119	1.112 / 1.098	0.497 / 0.505	0.110 / 0.111	1.079 / 1.074
+ “parallel” explicit explanations	<b>0.522 / 0.517</b>	<b>0.095 / 0.095</b>	<b>1.035 / 1.026</b>	<b>0.567 / 0.576</b>	<b>0.082 / 0.087</b>	<b>0.994 / 1.003</b>
$p_{\text{sfmax}}$ of Mixtral ( $\tau = 10$ )	0.445 / 0.435	0.210 / 0.214	1.380 / 1.384	0.487 / 0.492	0.207 / 0.209	1.369 / 1.370
+ “serial” explanations	0.457 / 0.461	0.198 / 0.195	1.344 / 1.328	0.522 / 0.533	0.181 / 0.182	1.290 / 1.288
+ “serial” explicit explanations	<b>0.522 / 0.532</b>	0.184 / 0.188	1.302 / 1.306	<b>0.599 / 0.584</b>	0.176 / 0.191	1.275 / 1.314
+ “parallel” explanations	0.426 / 0.432	0.175 / 0.176	1.275 / 1.269	0.518 / 0.514	0.157 / 0.158	1.221 / 1.217
+ “parallel” explicit explanations	0.521 / 0.524	<b>0.143 / 0.145</b>	<b>1.177 / 1.176</b>	0.576 / 0.567	<b>0.126 / 0.140</b>	<b>1.128 / 1.162</b>
$p_{\text{sfmax}}$ of Mixtral ( $\tau = 5$ )	0.453 / 0.439	0.330 / 0.338	1.737 / 1.757	0.559 / 0.542	0.143 / 0.150	1.177 / 1.192
+ “serial” explanations	0.452 / 0.464	0.287 / 0.287	1.610 / 1.604	0.573 / 0.581	0.113 / 0.116	1.088 / 1.091
+ “serial” explicit explanations	<b>0.525 / 0.530</b>	0.295 / 0.308	1.634 / 1.665	<b>0.593 / 0.588</b>	0.131 / 0.138	1.142 / 1.156
+ “parallel” explanations	0.435 / 0.430	0.217 / 0.221	1.398 / 1.406	0.593 / <b>0.599</b>	<b>0.095 / 0.098</b>	<b>1.034 / 1.037</b>
+ “parallel” explicit explanations	0.520 / 0.527	<b>0.180 / 0.184</b>	<b>1.289 / 1.294</b>	0.589 / 0.574	0.098 / 0.101	1.043 / 1.045
<i>MJDs from Llama3</i>						
$p_{\text{norm}}$ of Llama3	0.514 / 0.526	0.097 / 0.098	1.038 / 1.036	0.541 / 0.528	0.091 / 0.094	1.023 / 1.025
+ “serial” explanations	0.574 / 0.574	0.096 / 0.097	1.037 / 1.033	0.618 / 0.601	0.091 / 0.093	1.020 / 1.022
+ “serial” explicit explanations	0.578 / 0.574	<b>0.091 / 0.092</b>	<b>1.022 / 1.018</b>	0.634 / 0.598	<b>0.085 / 0.088</b>	<b>1.003 / 1.006</b>
+ “parallel” explanations	0.573 / 0.582	0.098 / 0.098	1.041 / 1.038	0.636 / 0.598	0.093 / 0.095	1.026 / 1.028
+ “parallel” explicit explanations	<b>0.582 / 0.586</b>	0.094 / 0.095	1.030 / 1.026	<b>0.639 / 0.620</b>	0.089 / 0.091	1.014 / 1.016
$p_{\text{sfmax}}$ of Llama3 ( $\tau = 20$ )	0.528 / 0.524	0.091 / 0.093	1.023 / 1.021	0.546 / 0.535	0.085 / 0.089	1.005 / 1.009
+ “serial” explanations	0.567 / 0.576	0.091 / 0.091	1.021 / 1.016	0.626 / 0.608	0.082 / 0.086	0.996 / 1.000
+ “serial” explicit explanations	<b>0.585 / 0.568</b>	<b>0.086 / 0.087</b>	<b>1.008 / 1.004</b>	<b>0.646 / 0.610</b>	<b>0.077 / 0.081</b>	<b>0.981 / 0.987</b>
+ “parallel” explanations	0.584 / <b>0.583</b>	0.092 / 0.093	1.024 / 1.020	0.643 / 0.611	0.085 / 0.089	1.004 / 1.008
+ “parallel” explicit explanations	0.581 / 0.578	0.088 / 0.089	1.014 / 1.010	0.645 / <b>0.621</b>	0.081 / 0.085	0.993 / 0.996
$p_{\text{sfmax}}$ of Llama3 ( $\tau = 10$ )	0.550 / 0.541	0.090 / 0.093	1.020 / 1.021	0.571 / 0.547	0.085 / 0.089	1.003 / 1.009
+ “serial” explanations	0.571 / 0.582	0.084 / 0.085	1.000 / 0.996	0.644 / 0.619	0.072 / 0.078	0.964 / 0.976
+ “serial” explicit explanations	0.582 / <b>0.602</b>	0.081 / 0.083	0.993 / 0.992	0.656 / <b>0.621</b>	0.070 / 0.079	0.960 / 0.978
+ “parallel” explanations	<b>0.588 / 0.588</b>	0.083 / 0.084	0.998 / 0.995	0.649 / 0.612	0.071 / 0.077	0.962 / 0.973
+ “parallel” explicit explanations	0.576 / 0.587	<b>0.079 / 0.080</b>	<b>0.985 / 0.982</b>	<b>0.661 / 0.616</b>	<b>0.066 / 0.073</b>	<b>0.946 / 0.960</b>
$p_{\text{sfmax}}$ of Llama3 ( $\tau = 5$ )	0.484 / 0.502	0.345 / 0.347	1.785 / 1.783	0.588 / 0.555	0.144 / 0.150	1.182 / 1.193
+ “serial” explanations	0.531 / 0.525	0.271 / 0.276	1.563 / 1.570	0.657 / 0.621	0.104 / 0.118	1.061 / 1.097
+ “serial” explicit explanations	<b>0.599 / 0.590</b>	0.311 / 0.331	1.682 / 1.735	0.663 / 0.622	0.132 / 0.150	1.145 / 1.192
+ “parallel” explanations	0.516 / 0.524	0.207 / 0.210	1.370 / 1.373	0.658 / 0.611	<b>0.082 / 0.093</b>	<b>0.995 / 1.022</b>
+ “parallel” explicit explanations	0.577 / 0.568	<b>0.171 / 0.191</b>	<b>1.262 / 1.315</b>	<b>0.672 / 0.623</b>	0.083 / 0.096	0.996 / 1.030

Table 10: Results for “ $\tau$ ” discussion of fine-tuning comparison on Chaos NLI dev/test set. The KL and Cross-Entropy (CE) Loss reflected the distance between distributions, whereas Weighted F1 reflected the capability to handle NLI problems. When  $\tau$  gets smaller the F1 of Llama3 will improve. However, the change of KL and CE Loss is not linear. It seems we could find a balance point for both performance between single label evaluation and distribution evaluation by searching  $\tau$ , which could be a future study.



<b>Distributions</b>	$p_{\text{norm}}$	$p_{\text{sfmax } \tau = 20}$	$p_{\text{sfmax } \tau = 10}$	$p_{\text{sfmax } \tau = 5}$
	<b>D.Corr</b> $\uparrow$	<b>D.Corr</b> $\uparrow$	<b>D.Corr</b> $\uparrow$	<b>D.Corr</b> $\uparrow$
Chaos NLI	1	-	-	-
MNLI single label	0.612	-	-	-
MNLI distribution	0.795	-	-	-
VariErr NLI	0.688	-	-	-
Mixtral original	0.609	0.593	0.584	0.577
4 explanations at a time	0.642	0.621	0.612	0.605
3 explanations at a time	0.659	0.636	0.625	0.617
2 explanations at a time	0.685	0.662	0.651	0.644
1 explanations at a time	0.731	0.713	0.704	0.697
4 explicit explanations at a time	0.623	0.608	0.601	0.596
3 explicit explanations at a time	0.651	0.637	0.629	0.624
2 explicit explanations at a time	0.678	0.664	0.606	0.652
1 explicit explanation at a time	0.719	0.709	0.704	0.701
Llama3 original	0.689	0.677	0.665	0.656
4 explanations at a time	0.750	0.740	0.730	0.714
3 explanations at a time	0.770	0.763	0.754	0.743
2 explanations at a time	0.795	0.790	0.783	0.774
1 explanations at a time	0.818	0.812	0.807	0.797
4 explicit explanations at a time	0.733	0.725	0.716	0.703
3 explicit explanations at a time	0.757	0.752	0.746	0.739
2 explicit explanations at a time	0.784	0.779	0.774	0.769
1 explicit explanation at a time	0.809	0.802	0.796	0.787

Table 11: Distance Correlation (D.Corr) results. All ablation results for comments numbers. Scores are compared with Chaos NLI’s HJD.

Distributions	$p_{\text{norm}}$			$p_{\text{sftmax}}$		
	KL ↓	JSD ↓	TVD ↓	KL ↓	JSD ↓	TVD ↓
MNLI ground truth (Williams et al., 2018)	9.288	0.422	0.435	-	-	-
MNLI distribution (Williams et al., 2018)	1.242	0.281	0.295	-	-	-
VariErr NLI (Weber-Genzel et al., 2024)	3.604	0.282	0.296	-	-	-
Mixtral user best	<b>0.245</b>	<b>0.211</b>	<b>0.239</b>	<b>0.217</b>	<b>0.208</b>	<b>0.232</b>
Llama3 user best	<b>0.235</b>	<b>0.247</b>	<b>0.266</b>	<b>0.212</b>	<b>0.232</b>	<b>0.245</b>
Mixtral original	0.433	0.291	0.340	0.434	0.292	0.342
4 explanations at a time	0.266	0.242	0.270	0.264	0.241	0.269
3 explanations at a time	0.260	0.240	0.268	0.257	0.239	0.266
2 explanations at a time	0.258	0.239	0.267	0.252	0.238	0.265
1 explanations at a time	0.250	0.237	0.264	0.245	0.236	0.262
4 explicit explanations at a time	0.224	0.221	0.243	0.230	0.223	0.245
3 explicit explanations at a time	0.220	0.220	0.242	0.224	0.221	0.243
2 explicit explanations at a time	0.216	0.217	0.239	0.217	0.218	0.240
1 explicit explanations at a time	0.203	0.213	0.231	0.198	0.208	0.227

Table 12: “Assistant” mode results. Scores are compared with Chaos NLI’s label probability distributions.

Type	General Instruction Prompt
With explanations	<p><b>"role": "user", "content":</b> Please add some comments for the relationship between Context and Statement. Context: {promise} Statement: {hypothesis}</p> <p><b>"role": "assistant", "content":</b> Comment 1: {explanation 1} Comment 2: {explanation 2} ...</p> <p><b>"role": "user", "content":</b> Please carefully and fairly base your selection on the Comment to determine whether the Statement is true (Entailment), undetermined (Neutral), or false (Contradiction) given the Context and select ONE of the listed options and start your answer with a single letter. A. Entailment B. Neutral C. Contradiction. Answer:</p>
With explicit explanations	<p><b>"role": "user", "content":</b> Please add some comments for the relationship between Context and Statement. Context: {promise} Statement: {hypothesis}</p> <p><b>"role": "assistant", "content":</b> Comment 1: {explanation 1}, so I choose label1 Comment 2: {explanation 2}, so I choose label2 ...</p> <p><b>"role": "user", "content":</b> Please carefully and fairly base your selection on the Comment to determine whether the Statement is true (Entailment), undetermined (Neutral), or false (Contradiction) given the Context and select ONE of the listed options and start your answer with a single letter. A. Entailment B. Neutral C. Contradiction. Answer:</p>

Table 13: Instruction prompt for “assistant” mode.

it as an exploration of prompt methods.

Role	Content
<i>Instance from MNLI</i>	
Premise	Krugman’s column will henceforth be known as The Dismal Science, a phrase too famous to be ownable by anyone, except possibly British essayist Thomas Carlyle (1795-1881), who coined it.
Hypothesis	Krugman writes novels.
<i>Explanations from VariErr NLI</i>	
C0	Krugman is a columnist, he doesn’t write novels.
C1	No, in the context is his column, which appears often in newspaper.
N0	It’s clear that Krugman writes a column, but not clear whether they write novels.
N1	Krugman has a column known as “The Dismal Science”, it can be a novel column, but also can be others like essay column.

Table 14: The sample for the case study. “C0”, “C1” support Contradiction while “N0”, “N1” support Neutral.

## K Case Study Details

We conducted a case study to demonstrate how explanations influence LLM’s output, and empirically showed that the order of explanations does indeed impact the model’s output. This led us to propose the approach in §3.3, which involves reducing the bias introduced by different sequences through permutation and averaging.

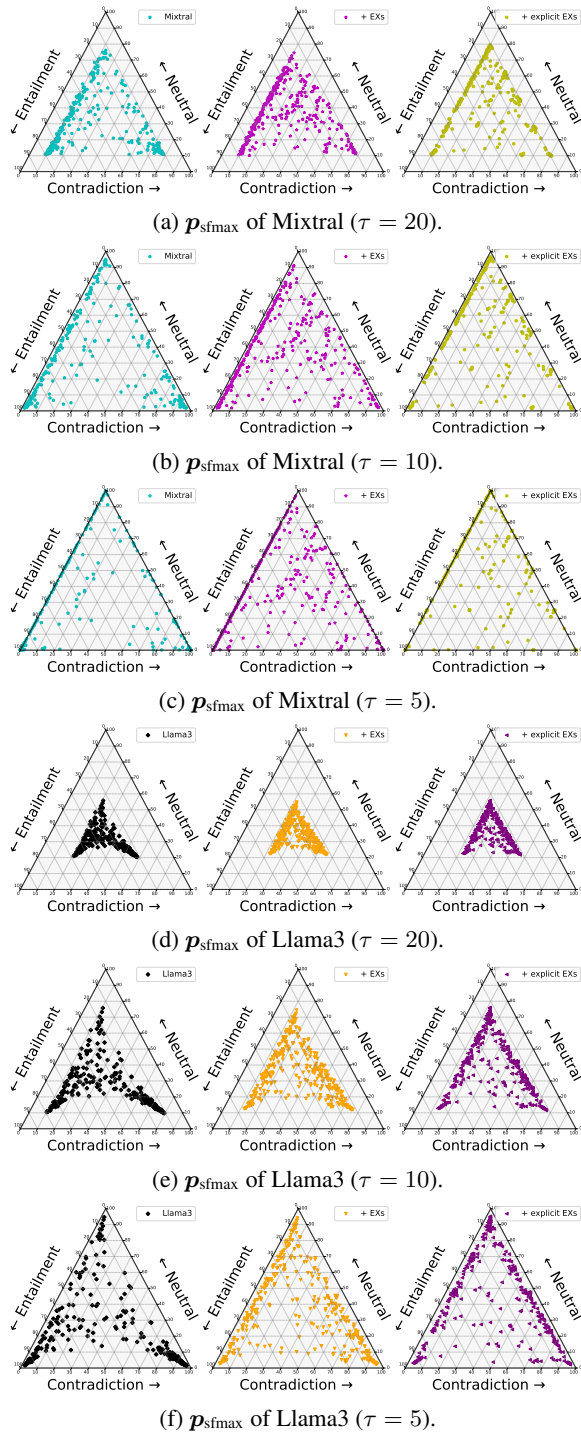


Figure 8: Visualization of distributions in ternary plot for  $\tau$  discussion. Through amplification, the bad becomes worse, and the good becomes better.

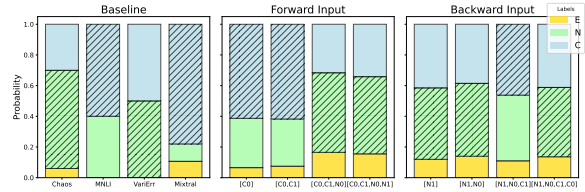


Figure 9: Results of case study. The three colored segments of each bar in the bar chart are labeled with the probabilities of three tags. The ticks below each chart represent the sequence combinations of the input explanations. “C0”, “C1” are explanations supporting Contradiction while “N0”, “N1” support Neutral. Detailed of the used sample are in Table 14.

The sample used is shown in Table 14 and the results are shown in Figure 9. More details of the output logits and texts were recorded in Table 15. We use Mixtral with “assistant” mode to implement this case study.

Through the results, we observed that the output label probability of Mixtral changes, when increasing the number of explanations, whether in a forward direction or backward direction. And, the order does impact. Latter ones usually impact more, but the final results could be similar, which means LLM can capture the HLV information inside, even when facing multiple explanations with diverse opinions.

Setting	Content
no explanations scores/probs output text	- [5.906385898590088, 6.259021282196045, 43.25299835205078] / [0.106578055463, 0.1129412010, 0.78048074346] C. Contradiction. The context provided does not give any information about Krugman writing novels. Instead, it mentions that Krugman's column will be named "The Dismal Science," a phrase coined by Thomas Carlyle. Therefore, the statement that Krugman writes novels is a contradiction given the context.</s>
<i>Forward Input</i>	
input explanations [C0] scores/probs output text	Comment 1: Krugman is a columnist, he doesn't write novels. [4.2198514938, 20.7870941162, 39.63526535] / [0.065280123175, 0.32157152126, 0.61314835556] C. Contradiction. The statement "Krugman writes novels" is a contradiction given the context "Krugman's column will henceforth be known as The Dismal Science" as it implies that Krugman is a columnist and not a novelist.</s>
input explanations [C0,C1] scores/probs output text	Comment 1: Krugman is a columnist, he doesn't write novels. Comment 2: No, in the context is his column, which appears often in newspaper [4.7512874603, 19.5720710754, 39.330707550] / [0.074642324559, 0.30747558292, 0.61788209251] C. Contradiction. Explanation: The context states that Krugman writes a column called "The Dismal Science," and that the phrase was coined by Thomas Carlyle. The statement, "Krugman writes novels," contradicts this context, as it suggests that Krugman writes novels, rather than a column.</s>
input explanations [C0,C1,N0] scores/probs output text	Comment 1: Krugman is a columnist, he doesn't write novels. Comment 2: No, in the context is his column, which appears often in newspaper Comment 3: It's clear that Krugman writes a column, but not clear whether they write novels. [11.1629428863, 35.1051597595, 21.4133796691] / [0.16493348704, 0.51868189878, 0.31638461417] B. Neutral The context only provides information about Krugman's column being known as "The Dismal Science" and that the phrase was coined by Thomas Carlyle. It does not provide any information about Krugman writing novels. Therefore, the statement "Krugman writes novels" is neutral, neither entailed nor contradicted by the context.</s>
input explanations [C0,C1,N0,N1] scores/probs output text	Comment 1: Krugman is a columnist, he doesn't write novels. Comment 2: No, in the context is his column, which appears often in newspaper Comment 3: It's clear that Krugman writes a column, but not clear whether they write novels. Comment 4: Krugman has a column known as "The Dismal Science", it can be a novel column, but also can be others like essay column. [10.3001527786, 33.403881072, 22.699693679] / [0.15511407509, 0.5030422585, 0.34184366636] B. Neutral The statement "Krugman writes novels" is not directly addressed or contradicted by the context provided. The context only mentions that Krugman's column will be known as "The Dismal Science" and that the phrase was coined by Thomas Carlyle. It does not provide any information about Krugman's novel-writing activities. Therefore, the statement is undetermined based on the context.</s>
<i>Backward Input</i>	
input explanations [N1] scores/probs output text	Comment 1: Krugman has a column known as "The Dismal Science", it can be a novel column, but also can be others like essay column. [8.8821163177, 34.580959320, 30.8163871765] / [0.119577013365, 0.4655209219, 0.414870894442] B. Neutral. The context only provides information about the name of Krugman's column, "The Dismal Science", and a brief history of the phrase. It does not provide any information about Krugman writing novels, so the statement "Krugman writes novels" cannot be determined as true or false based on the given context.</s>
input explanations [N1,N0] scores/probs output text	Comment 1: Krugman has a column known as "The Dismal Science", it can be a novel column, but also can be others like essay column. Comment 2: It's clear that Krugman writes a column, but not clear whether they write novels. [9.7944955825, 33.278915405, 26.9777603149] / [0.139819155060, 0.475065795280, 0.385115049658] B. Neutral. The context does not provide enough information to confirm or deny whether Krugman writes novels. The column being referred to as "The Dismal Science" does not have any impact on Krugman's novel-writing abilities or activities.</s>
input explanations [N1,N0,C1] scores/probs output text	Comment 1: Krugman has a column known as "The Dismal Science", it can be a novel column, but also can be others like essay column. Comment 2: It's clear that Krugman writes a column, but not clear whether they write novels. Comment 3: No, in the context is his column, which appears often in newspaper [8.103030204, 31.667663574, 34.176952362] / [0.109577932870, 0.42824437594, 0.462177691185] C. Contradiction. Explanation: The context states that Krugman's column will be known as "The Dismal Science", and it does not provide any information about Krugman writing novels. Therefore, the statement "Krugman writes novels" is not true based on the given context.</s>
input explanations [N1,N0,C1,C0] scores/probs output text	Comment 1: Krugman has a column known as "The Dismal Science", it can be a novel column, but also can be others like essay column. Comment 2: It's clear that Krugman writes a column, but not clear whether they write novels. Comment 3: No, in the context is his column, which appears often in newspaper Comment 4: Krugman is a columnist, he doesn't write novels. [9.7341032028, 32.308975219, 29.41177496] / [0.136227315521, 0.452159266214, 0.411613418264] B. Neutral. The context does not provide enough information to confirm or deny whether Krugman writes novels.</s>

Table 15: Case study records.