EVADE: LLM-Based Explanation Generation and Validation for Error Detection in NLI

Longfei Zuo[1, 2*] **Barbara Plank**[2, 3] **Siyao Peng**[2, 3]
[1] Technical University of Munich, Heilbronn, Germany
[2] MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[3] Munich Center for Machine Learning (MCML), Munich, Germany
longfei.zuo@tum.de b.plank@lmu.de loganpeng1992@gmail.com

Abstract

High-quality datasets are critical for training and evaluating reliable NLP models. In tasks like natural language inference (NLI), human label variation (HLV) arises when multiple labels are valid for the same instance, making it difficult to separate annotation errors from plausible variation. An earlier framework, VARIERR (Weber-Genzel et al., 2024), asks multiple annotators to explain their label decisions in the first round and flags errors through validity judgments in the second round. However, conducting two rounds of manual annotation is costly and may limit the coverage of plausible labels or explanations. Our study proposes a new framework, EVADE, for generating and validating explanations to detect errors using large language models (LLMs). We perform a comprehensive analysis comparing human- and LLM-detected errors for NLI across distribution comparison, validation overlap, and impact on model fine-tuning. Our experiments demonstrate that LLM validation refines generated explanation distributions to more closely align with human annotations, and that removing LLM-detected errors from training data yields improvements in fine-tuning performance than removing errors identified by human annotators. This highlights the potential to scale error detection, reducing human effort while improving dataset quality under label variation.

1 Introduction

Datasets are fundamental to research, yet annotation errors can undermine model reliability (Goswami et al., 2023), highlighting the need for high-quality data in trustworthy Natural Language Processing (NLP) systems (Klie et al., 2024; Nasution and Onan, 2024; Weber-Genzel et al., 2024). Human Label Variation (HLV, Plank 2022) refers to plausible variation in annotations, in which multiple labels can be assigned to a single instance.

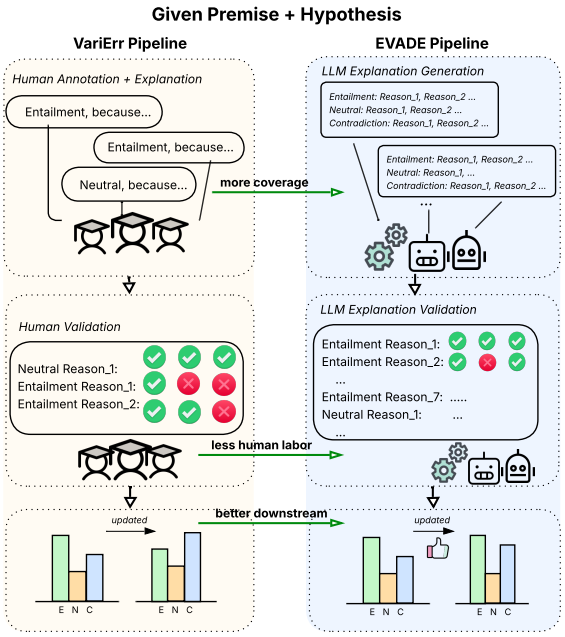

Figure 1: Overview of our LLM-based EVADE framework compared with the human-based VARIERR pipeline (Weber-Genzel et al., 2024). The first two modules, explanation generation and validation, are the core components. Compared with VARIERR, our EVADE framework provides broader explanation coverage, requires less human intervention, and delivers better downstream performance in predicting label distributions.

HLV has gained significant attention, particularly for Natural Language Inference (NLI) (Pavlick and Kwiatkowski, 2019; Nie et al., 2020), where multiple plausible labels (Entailment, Neutral, or Contradiction) can be valid for the same premise-hypothesis pair (Jiang and de Marneffe, 2022; Jiang et al., 2023; Jayaweera and Dorr, 2025). While HLV better reflects real-world ambiguities, it also introduces a challenge: annotation errors may be obscured by the variation.

Many approaches for annotation error detection (AED) have been introduced in previous research (Klie et al., 2023; Weber and Plank, 2023; Weber

*Main work carried out while at LMU Munich.

Findings of the Association for Computational Linguistics: ACL 2026, pages 1286–1300

et al., 2024; Bernier-Colborne and Vajjala, 2024), but to our knowledge, Weber-Genzel et al. (2024) is the only work that explicitly focuses on separating errors from HLV. They propose a two-round error detection procedure in which expert annotators first provide labels with explanations and then reassess their reasoning after reviewing the full set of labels and explanations from all the annotators. For each annotated NLI label, if the annotators do not validate their own previously written explanations in the second round, the label is considered erroneous. Previous studies typically introduce artificial noise by randomly flipping labels (Zheng et al., 2021; Jinadu and Ding, 2024), but these synthetic errors are often easily identifiable (Larson et al., 2019). In contrast, by leveraging self-validation, VARIERR captures naturally occurring errors and enables a more realistic evaluation of error detection.

However, a key limitation of this approach is its reliance on human experts to generate and validate explanations, both resource-intensive and difficult to scale. Fortunately, recent work (Chen et al., 2025) shows that LLM-generated explanations are comparable to humans in approximating label distributions in NLI, suggesting the potential to extend the VariErr framework to an LLM-based pipeline.

This paper presents EVADE, an LLM-driven method for error detection that leverages explanation generation and validation. Our main research questions are: **RQ1:** How effectively can LLM-generated explanations reflect valid HLV, and how do these compare to human-written explanations? **RQ2:** To what extent does LLM-validation enable reliable error detection compared to humans? **RQ3:** What is the impact of removing LLM-detected errors on downstream fine-tuning performance in terms of alignment with HLV? [1]

Figure 1 presents the structure of EVADE. We first prompt LLMs to generate explanations given a premise–hypothesis pair and a candidate label. Subsequently, LLMs evaluate the explanations by assigning validity scores, indicating how effectively each explanation supports the assigned label (§3). To assess the effectiveness of our framework, we compare it with the human-driven pipeline from multiple perspectives, including explanation distribution, validated label overlap, linguistic similarity, and downstream task performance. Our findings demonstrate that EVADE aligns closely with human annotators in terms of label agreement, and can effectively identify errors, yielding annotations that more accurately reflect HLV (§4 & §5).

[1]Our code and data are publicly available at https://github.com/mainlp/LLM_AED.

2 Related Work

Human label variation (HLV) refers to cases where annotators assign different but plausible labels to the same instance (Plank, 2022). HLV may arise from subjectivity (Cabitza et al., 2023), semantic ambiguity in the target instances (Aroyo and Welty, 2013, 2015) or guideline divergence (Peng et al., 2024), which challenges the assumption of a single ground truth in annotation. Aroyo and Welty (2013) propose the crowd truth, capturing subjective and diverse human interpretations.

HLV has been widely examined in the context of Natural Language Inference (NLI). Pavlick and Kwiatkowski (2019) show that NLI annotation disagreements are often systematic rather than noise. ChaosNLI (Nie et al., 2020) provides a human judgment distribution (HJD, Chen et al. 2024) by crowdsourcing 100 annotations per instance over a subset of MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), and αNLI (Bhagavatula et al., 2020) instances. LiveNLI (Jiang et al., 2023) and VariErrNLI (Weber-Genzel et al., 2024) supply ecologically valid explanations that verbalize different reasoning behind label decisions.

Annotation error detection (AED) is crucial in NLP, as many widely-used benchmarks contain annotation errors (Northcutt et al., 2021; Klie et al., 2023; Bernier-Colborne and Vajjala, 2024; Larson et al., 2020; Weber et al., 2024). Klie et al. (2023) conducted a comprehensive survey of error detection methods focusing on single-label tasks. However, annotation errors also arise when multiple labels are valid at the same time. To separate annotation error from HLV, VariErrNLI (Weber-Genzel et al., 2024) incorporates a second round of validity judgment to assess whether the provided explanations support the annotator's label decisions.

LLM-generated explanations are frequently used across a variety of tasks (Kunz and Kuhlmann, 2024), addressing whether they can approximate human reasoning. Within NLI, recent studies investigate how LLM-generated explanations can be leveraged to understand and model human annotation variations. Jiang et al. (2023) find that GPT-3 can generate fluent explanations, though not always aligned with the target label. Chen et al. (2025) prompt LLMs to generate explanations guided with

human-annotated labels and achieve similar performance in approximating human judgment distribution as when using human-written explanations. Hong et al. (2025) propose a linguistic taxonomy to further guide LLMs in generating more diverse explanations. The studies above highlight the potential of LLMs in generating explanations. Our paper addresses LLMs' ability to validate their explanation generation and detect annotation errors.

3 Generating and Validating Explanations Using LLMs

This section details our EVADE pipeline using LLMs. We introduce our dataset and model setups (§3.1), explanation generation and filtering process (§3.2), as well as validation scenarios (§3.3).

3.1 Setups

Dataset We experiment on the VARIERR dataset (Weber-Genzel et al., 2024), including NLI labels, explanations, and second-round validations from four human expert annotators. All 500 VARIERR instances are also in ChaosNLI (Nie et al., 2020), containing label distributions from 100 crowdworkers. This setup provides references for evaluating label distribution and explanation generation.

Models We select four state-of-the-art LLMs from two families due to their demonstrated capabilities in instruction following, critical in our generation and validation experiments: *Llama-3.1-8B-Instruct*, *Llama-3.3-70B-Instruct* (Grattafiori et al., 2024),[2] *Qwen2.5-7B-Instruct* and *Qwen2.5-72B-Instruct* (Qwen et al., 2025).

3.2 Explanation Generation and Filtering

Generation We follow Chen et al. (2025) to generate all distinct explanations for each label in a given premise-hypothesis pair to obtain a broad coverage. To maintain output validity, we adapted their prompt to allow LLMs to abstain if the label cannot be reasonably justified. Our prompt also explicitly prohibits introductory phrases or semantic repetitions. We kept all qualified explanation generations in our repository; thus, Chen et al. (2025)'s label-guided or label-free approaches in picking the longest or first explanations are irrelevant to this study. See Appendix A.1 Figure 3 for the template.

[2]At submission time, v3.1 is the most recent Llama-8B-Instruct model release and v3.3 for Llama-70B-Instruct.

Model	# Expl.	Avg. Length	# Label/Item	# Expl./Label
Llama-8B	**8845**	17.60	3.00	**5.90**
Llama-70B	4022	**22.35**	3.00	2.68
Qwen-7B	4047	16.19	2.87	2.82
Qwen-72B	3920	18.43	3.00	2.61
VariErr	1933	13.89	1.76	2.20

Table 1: Generation statistics on 500 VariErr examples. # Expl.: total number of explanations; Avg. Length: average number of words per explanation; # Label/Item: number of LLM-explained labels per instance; Expl./Label: average number of explanations per label.

Filtering Despite explicit instructions, we still find two main issues with LLM-generated explanations: *fallback responses* and *formatting errors*, which would not occur in human-written explanations. First, rather than omitting the output when uncertain, Llama-72B model sometimes responds with fallback statements such as *"Note: Since the statement is not supported by the context, there are no explanations for why the statement is true."* These responses signal the model's uncertainty or disagreement with the target label. While they may be interpreted as a form of "non-variation", they fail to provide substantive explanatory content for any specific label. Second, we observe formatting issues in the outputs of certain models, such as truncated outputs from Llama-8B caused by the default generation limit (256 tokens), and occasionally generated Chinese sentences from Qwen-7B, even all prompts are in English. As our framework aims not only to evaluate error detection performance but also to compare the alignment of LLM-generated explanations and relevant label distributions with human annotations, we manually filter out such explanations to ensure a cleaner distribution. [3]

Results We observe noticeable differences across model generation results in Table 1. Llama-8B generates nearly twice as many explanations as the others, suggesting potential redundancy between explanations, and Llama-70B yields the longest explanations (22.35 tokens on average) among all the models. These discrepancies become more evident when compared with human annotations in VARIERR. Human annotators generated much fewer explanations in quantity and, in general, much shorter explanations (13.89 tokens).

We then look at the average number of NLI

[3]In practice, we removed 37 incomplete outputs from Llama-8B, excluded 17 fallback generations from Llama-70B, and filtered out 6 explanations containing Chinese texts from Qwen-7B.

labels that are supported by LLM-generated explanations per instance (#Label/Item), as well as the average number of explanations per label (#Expl./Label). VARIERR receives on average 1.76 explained labels. However, LLMs tend to produce explanations for every label (or almost all in the case of Qwen-7B), which is unlikely to reflect a realistic human label variation. As for the number of generated explanations per label, Llama-8B still produces twice as many as humans under each label, due to its extensive output. These LLM overgenerations necessitate the explanation validation step in the EVADE framework.

3.3 Explanation Validation

To address the excessive generation of explanations, we instruct LLMs to validate their own explanations. Precisely, we ask an LLM to give validity judgments on whether each of its generated explanations supports the associated label, mirroring VARIERR's second-round self-validation process. We aim to align LLMs' explanations and label distribution more closely with human annotations and to demonstrate that LLMs can likewise be used for explanation validation and error detection.

Following Weber-Genzel et al. (2024), we prompt LLMs to assign a **validity score** (between 0.0 and 1.0) assessing how much each of its generated explanations justifies the corresponding label for the given premise–hypothesis pair. We consider an explanation as **validated** if its validity score p exceeds a predefined threshold τ; otherwise, it is **not-validated**. To determine whether a label is erroneous, we follow the criterion from VARIERR: a label is regarded as an "error" if the scores of all associated explanations fall below the threshold τ.

Prompts To systematically examine the role of context in LLM validation, we experiment with three prompting scenarios: (1) ***one-expl***: an LLM scores its generated explanations one at a time, without reference to other explanations; (2) ***one-llm***: an LLM scores each of its generated explanations in the context of all its generations; and (3) ***all-llm***: an LLM scores each of its generated explanations in the context of explanations generated by all four LLMs. When multiple explanations are in context, the LLM receives all explanations simultaneously and outputs a json file with set of scores in a single pass. Our prompts are adapted from Weber-Genzel et al. (2024); see Appendix A.2 Figures 4-5 for details.

Model	*one-expl*	*one-llm*	*all-llm*
Llama-8B	0.8003 $_{0.1290}$	0.5009 $_{0.3134}$	0.5154 $_{0.3616}$
Llama-70B	**0.8456** $_{0.1126}$	0.6483 $_{0.2114}$	0.6439 $_{0.2333}$
Qwen-7B	0.7107 $_{0.1442}$	0.5675 $_{0.2628}$	0.3825 $_{0.3894}$
Qwen-72B	0.8299 $_{0.0780}$	**0.7279** $_{0.1352}$	**0.7006** $_{0.1835}$

Table 2: Average validation scores and standard deviations across three prompting scenarios for four LLMs.

Results Table 2 presents the average validation scores of different LLMs across the three validation scenarios. The *one-expl* setting consistently yields the highest validity, with most LLMs producing average validity scores above 0.8. This is the only prompt setup that does not provide contextual information from other candidate explanations, which most likely leads to models' over-confidence in the explanations' validity. When additional context is introduced in the *one-llm* and *all-llm* settings, models can review alternative explanations for the same instance. This broader perspective may result in more comprehensive evaluations. However, providing contextual explanations generated by other LLMs (*all-llm* versus *one-llm*) only subtly changes the average validity score for all models except the Qwen-7B model. We then turn to the standard deviation (std), obtained by scoring each label for each instance and averaging across labels. The *one-expl* scenario achieves the lowest std, indicating more consistent scoring behavior for each label. However, this stability may reduce sensitivity to differences between valid and invalid explanations within the same label. In contrast, *one-llm* and *all-llm* appear to balance multiple explanations with higher std, and can thereby filter relatively more invalid explanations and yield refined distributions.

Moreover, smaller models tend to assign lower scores than large ones across all settings. The gap between differently sized models becomes particularly large in the *one-llm* and *all-llm* settings. One possible explanation is that longer contextual inputs in these settings may overwhelm smaller models, reducing their confidence when confronted with multiple competing explanations. For example, *all-llm* provides an average of approximately 41.7 explanations per instance, showing an extreme difference in average validity score between Qwen-7B (0.3825) and Qwen-72B (0.7006).

However, one should be cautious in assuming that differences in average validity scores directly indicate that one model is better or worse. To assess

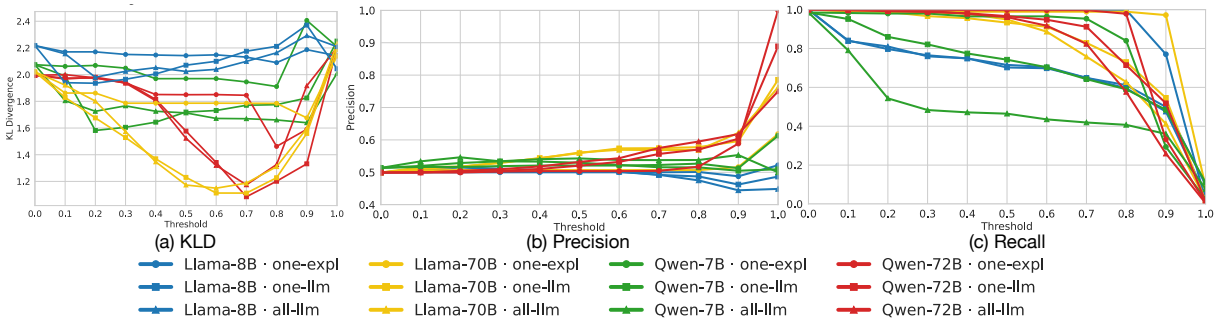

Figure 2: (a) shows the KL divergence curves between model distributions and ChaosNLI annotations across three prompting scenarios with validation threshold from 0.1 to 0.9. (b) and (c) present the precision and recall of the LLM-validated labels, computed against the VariErr-validated labels as ground truth, with validation thresholds from 0.1 to 0.9.

how well EVADE aligns with humans, we next examine whether the low validity scores correspond to genuine "errors" and analyze the overlap between LLM-validated labels and those approved by human annotators in VARIERR.

4 LLM versus Human Validation

After generation and validation, we compare the LLM-based (EVADE) and human-based error detection (VARIERR in Weber-Genzel et al. 2024) pipelines from different angles. We analyze the label distribution associated with LLM-generated explanations before and after LLM validation (§4.1) and examine the validated label overlap with those from human annotators (§4.2). Since VARIERR's validity judgments are binary and our LLM outputs range from 0.0 to 1.0, we also show in these sections how much the validity threshold τ affects our comparison. We last assess the semantic similarity between LLM-generated and human-written explanations, both before and after validation (§4.3).

4.1 Alignment with ChaosNLI Distribution

Since HLV treats label distribution rather than a single label as the gold standard, we want to measure the similarity between label distributions derived from model-validated explanations with those from human annotations in ChaosNLI (Nie et al., 2020). The distributions in ChaosNLI are aggregated from 100 annotators, providing a softer representation that better reflects a reference "gold" distribution (Chen et al., 2024). Specifically, we examine model distributions before and after validation against the crowdworker distribution to assess whether the validation step improves the alignment of model outputs with human consensus. We employ Kullback-Leibler divergence (KL, Kullback and Leibler 1951) following previous works (Chen et al., 2024, 2025) to measure the discrepancy between distributions.

An appropriate threshold has a substantial impact on the resulting distribution. To assess this effect, we compute the KL divergence between the model-validated distributions and the ChaosNLI annotations across three validation scenarios and a range of validation thresholds. Results are presented in Figure 2a. A threshold of τ**=0.0** represents the pre-validation condition, as the validation procedure does not permit negative scores. After validation, KL divergence generally decreases across most models and prompting scenarios as the threshold increases up to approximately τ**=0.8**. This reduction is particularly pronounced for the larger models, indicating better agreement with the human label distribution. The point in the plot where the KL divergence is minimized corresponds to the validation threshold that yields the closest alignment with the ChaosNLI distribution. However, as the threshold increases further, the alignment with the human distribution weakens. This suggests that while a high validation threshold can filter out invalid information, an overly strict threshold may also inadvertently discard correct information.

4.2 Comparing with Validated VARIERR

Given that LLMs generate substantially more explanations than humans and tend to provide explanations for every label, it is expected that LLMs identify more errors during validation, as their outputs may initially include a large number of inconsistent arguments. To obtain a more reliable comparison between LLM- and human-validation, we therefore focus on **the overlap of validated**

Models	Lexical n=3 ↑	Syntactic n=3 ↑	Cosine↑	Euclidean↑
within-human	0.052	0.144	0.529	0.522
within-LLM				
Llama-8B	**0.103** / 0.101 / 0.097 / 0.099	**0.265** / 0.264 / 0.263 / 0.263	0.599 / 0.608 / **0.611** / 0.606	0.542 / **0.545** / **0.545** / 0.544
Llama-70B	0.030 / **0.031** / 0.029 / 0.029	0.201 / 0.201 / 0.200 / **0.202**	0.595 / 0.606 / 0.605 / **0.608**	0.535 / 0.539 / 0.538 / **0.540**
Qwen-7B	0.010 / **0.012** / 0.010 / 0.010	0.160 / 0.161 / 0.164 / **0.170**	0.533 / **0.548** / 0.536 / 0.536	0.517 / **0.521** / 0.518 / 0.518
Qwen-72B	**0.023** / **0.023** / 0.021 / 0.022	0.188 / 0.189 / 0.187 / **0.191**	0.578 / 0.583 / 0.583 / **0.584**	0.530 / **0.532** / 0.531 / **0.532**
LLM-vs-human				
Llama-8B	**0.020** / **0.020** / 0.019 / 0.018	0.116 / 0.116 / **0.117** / 0.116	0.450 / 0.452 / **0.453** / **0.453**	0.497 / **0.498** / **0.498** / **0.498**
Llama-70B	0.028 / 0.029 / 0.029 / **0.030**	0.129 / 0.131 / **0.132** / **0.132**	0.515 / 0.521 / 0.521 / **0.523**	0.514 / 0.516 / 0.516 / **0.517**
Qwen-7B	0.015 / **0.016** / 0.015 / 0.014	0.101 / **0.103** / 0.100 / 0.095	0.469 / **0.477** / 0.469 / 0.470	0.502 / **0.504** / 0.502 / 0.502
Qwen-72B	0.023 / **0.024** / 0.023 / **0.024**	0.125 / **0.127** / 0.126 / 0.126	0.502 / 0.506 / **0.507** / 0.504	0.511 / **0.512** / **0.512** / **0.512**

Table 3: Linguistic similarity (lexical, syntactic, semantic) of explanations across three evaluation regimes: *within-human*, *within-LLM*, and *LLM-vs-human*. Each cell lists the scores in order: before validation, after *one-expl* validation, after *one-llm* validation, and after *all-llm* validation. We underline lowest score(s) in each cell and **bold highest**.

labels rather than on the overlap of detected errors.

When comparing the overlaps, we observe strong similarity. The majority of human-validated labels are also confirmed by the LLM validation process. We thereby compute the precision and recall of LLM-validated labels by considering the VariErr labels as the gold standard, varying the validation threshold from 0.1 to 0.9.

Figures 2b-2c present the results. For precision, all models start around P=0.5 when τ=0. Larger models exhibit a generally increasing trend as the threshold rises, whereas smaller models show more fluctuation and a noticeable decline when the threshold exceeds 0.6. This pattern suggests that low thresholds effectively filter out low-confidence predictions and reduce false positives. Still, for smaller models, overly high thresholds begin to remove true positives as well, leading to a drop in precision. Regarding recall, all models across different validation scenarios exhibit a consistent decreasing trend as the threshold increases, with a particularly sharp decline beginning around 0.7. This occurs because higher thresholds exclude more true-positive labels, thereby reducing the overall recall. Among the three validation modes, *all-llm* is the most conservative: it validates substantially fewer labels and yields lower recall than the other two modes.

Overall, excessively low validation thresholds introduce numerous unreliable labels, while overly high thresholds risk discarding valid ones. An optimal threshold should balance precision and recall. Considering these metrics together with the KL divergence, we select the threshold with low KL divergence that also corresponds to relatively high precision and recall values as our final validation threshold (detailed in Appendix C Table 8). Interestingly, the selected thresholds closely align with the average validation scores reported in Table 2 across most models and prompting settings, supporting the calibration of validation scores as meaningful indicators, with low scores generally corresponding to genuine annotation errors. These thresholds are set to facilitate experiments and analyses in the following sections.

4.3 Explanation Similarity

To better understand the underlying generation pattern of LLMs, as well as the similarities and differences between LLM and human reasoning, we analyze the linguistic similarity between human and LLM-generated explanations before and after validation. Following Chen et al. (2025); Hong et al. (2025), we measure lexical, syntactic, and semantic similarity between explanations (Giulianelli et al., 2023), where lower scores indicate greater variation and the higher the more similar. We compare pairwise similarity within-VariErr human explanations (*within-human*), within each LLM (*within-LLM*), and between each LLM and human explanations (*LLM-vs-human*). For each item, we compute pairwise scores among all explanations under the same label and average across labels. Table 3 reports results before validation and after three validation scenarios (*one-expl*, *one-llm*, and *all-llm*); more results are in Appendix D Table 9.

Firstly, when comparing similarity scores between LLM sizes, larger models more frequently exhibit higher similarity in both *within-LLM* and *LLM-vs-human* regimes. *Within-LLM* scores on

Model	Validation	# Expl.	# L/I	# E/L	AP	P@100	R@100
Llama-8B	*one-expl*	**7474**	**2.99**	**4.98**	**15.3**	**21.0**	**16.3**
	one-llm	5951	2.32	3.97	13.9	12.0	9.3
	all-llm	6283	2.35	4.19	13.9	14.0	10.9
Llama-70B	*one-expl*	**3354**	**2.82**	**2.24**	18.4	24.0	18.6
	one-llm	2931	2.38	1.95	22.6	26.0	20.2
	all-llm	2799	2.33	1.87	**26.8**	**31.0**	**24.0**
Qwen-7B	*one-expl*	**3282**	**2.73**	**2.19**	**16.4**	**19.0**	**15.3**
	one-llm	3157	2.44	2.10	13.6	15.0	12.1
	all-llm	1883	1.49	1.26	14.5	17.0	13.7
Qwen-72B	*one-expl*	**3286**	**2.83**	**2.19**	**18.8**	**26.0**	**20.2**
	one-llm	2878	2.45	1.92	16.7	21.0	16.3
	all-llm	2695	2.15	1.80	**18.8**	22.0	17.1
VariErr	self	1712	1.50	2.29	22.8	23.7	18.3
	random	1712	1.50	2.29	14.7	14.7	11.4

Table 4: Statistics after validation and performance of EVADE on the AED task. The AED values for VARIERR with mean Datamaps and random baseline are copied from Weber-Genzel et al. (2024) for reference.

Llama-7B explanations are an exception, scoring noticeably higher than Llama-70B on lexical and syntactic similarities. Moreover, Llama-8B attains the highest scores across lexical, syntactic, and semantic dimensions among all LLMs in the *within-LLM* regime, likely reflecting the greater volume and thus redundancy of its generated explanations.

Secondly, comparing *within-human* and *within-LLM* similarity scores, LLM generations exhibit greater lexical divergence (except for Llama-8B) while achieving higher syntactic and semantic similarity to each other than between human-written explanations, indicating that LLMs can generate lexically different outputs, while keeping syntactical structure and semantic meaning quite similar.

Thirdly, the similarity between LLM- and human-generated explanations (*LLM-vs-human*) is consistently lower than that observed *within-human/LLM*, indicating that LLM generations still considerably diverge from human-written explanations, particularly in syntax and semantics.

Lastly, validation has minimal effects on similarity scores, typically showing a difference of less than 0.010 before versus after validation. Even so, most similarity scores (except for Llama-8B) increase after validation, reflecting the objective of removing low-quality explanations.

4.4 Post-Validation Analyses

Statistics To assess the effectiveness of EVADE for automatic error detection (AED), Table 4 presents the statistics after each validation scenario. Compared with the pre-validation scores in Table 1, LLM-based validation filters out a substantial proportion of generated explanations. Specifically, averaged across the three prompting strategies, Llama-8B excludes 25.7% of explanations, Llama-70B 24.7%, Qwen-7B 31.5%, and Qwen-72B 24.7%. Across all prompting strategies, *all-llm* excludes the most generated explanations. This further indicates that a validation strategy with a higher standard deviation effectively removes a larger proportion of invalid explanations. LLMs exhibit a more radical filtering behavior than humans. Even under the mildest validation prompt (*one-expl*), LLMs discard on average 16.8% of the generated explanations, compared to 11.4% for human validators. Interestingly, the number of explanations per label in VARIERR slightly increases after validation. The removal of explanations more directly translates into the rejection of labels, likely due to the VARIERR's sparsity that many labels, especially erroneous ones, are supported by only a single explanation.

Automatic Error Detection (AED) Following Weber-Genzel et al. (2024), we also compute the average validation score per label for each instance and then use these scores to rank labels' likelihood of being erroneous. We evaluate this ranking against the self-flagged errors identified by human annotators. We report standard ranking-based metrics: average precision (AP), as well as precision and recall at the top 100 predictions (P@100 and R@100) (Klie et al., 2023). We adopt the Datamaps model (Swayamdipta et al., 2020) and the random baseline from VARIERR as our baselines.

Overall, larger models achieve stronger AED performance, while smaller models perform closely to the random baseline. Among the validation strategies, *one-expl* consistently yields the highest scores, while *all-llm* with Llama-70B is the only configuration that surpasses the DM baseline. Nevertheless, this evaluation setup has limitations. Since these metrics rely solely on the ranking of numerical scores, they may underrepresent labels that receive low scores but are not explicitly ranked lower. For example, two instances with average scores of 0.75 and 0.05 would be classified as errors under a fixed threshold. We thus turn to fine-tuning experiments in §5 using the analyzed thresholds (cf. Table 8) that better preserve the numerical granularity.

5 EVADE for Fine-tuning

To assess whether EVADE benefits downstream fine-tuning experiments, we approximate human judgment distribution (HJD) in ChaosNLI (Nie et al., 2020), following Chen et al. (2024).

Models	Validation	BERT F1 ↑	BERT KL ↓	RoBERTa F1 ↑	RoBERTa KL ↓
VARIERR R1 *(baseline)*		*0.5842*	*0.1701*	*0.6248*	*0.1667*
VARIERR R2 *(baseline)*		0.5500	0.2510	0.6125	0.2474
(a) Fine-tuning with EVADE *labels*					
Llama-8B	*before*	0.4499	0.1181	0.3312	0.1188
	one-expl	0.4756	0.1172	0.5245	0.1170
	one-llm	**0.5911**	0.0971	0.5975	0.1218
	all-llm	0.5820	**0.0891**	**0.6088**	**0.0984**
Llama-70B	*before*	0.4499	0.1181	0.3312	0.1188
	one-expl	0.5219	0.1011	0.6119	0.0967
	one-llm	**0.6328**	0.0837	0.6151	0.0962
	all-llm	0.6210	**0.0716**	**0.6351**	**0.0818**
Qwen-7B	*before*	0.5074	0.1077	0.5406	0.1076
	one-expl	0.5519	0.1019	0.5624	0.1027
	one-llm	**0.6294**	**0.0783**	**0.5983**	**0.0772**
	all-llm	0.4984	0.2566	0.5172	0.2724
Qwen-72B	*before*	0.4499	0.1181	0.3312	0.1188
	one-expl	0.5310	0.1021	0.5570	**0.1001**
	one-llm	0.5125	0.0994	0.5579	0.1122
	all-llm	**0.6245**	**0.0807**	**0.6545**	0.1109
(b) Using EVADE *to validate* VARIERR *R1 and then finetuning*					
Llama-8B	*one-expl*	**0.5939**	**0.1714**	0.6385	**0.1699**
	one-llm	0.6195	0.1938	**0.6617**	0.1907
	all-llm	0.5929	0.1916	0.6429	0.2042
Llama-70B	*one-expl*	0.5968	**0.1754**	0.6383	**0.1722**
	one-llm	0.5951	0.1924	0.6556	0.1903
	all-llm	**0.6096**	0.1915	**0.6762**	0.1805
Qwen-7B	*one-expl*	0.5892	0.1755	0.6294	0.1787
	one-llm	0.6232	**0.1704**	0.6654	**0.1588**
	all-llm	**0.6415**	0.2456	**0.6731**	0.2793
Qwen-72B	*one-expl*	0.5858	**0.1869**	0.6370	**0.1833**
	one-llm	0.5993	0.1920	0.6376	0.1939
	all-llm	**0.6228**	0.1954	**0.6614**	0.2099

Table 5: Results of directly fine-tuning with EVADE labels or using EVADE to validate VARIERR R1 labels and then finetuning.

Setups We apply EVADE in two setups: (a) directly using EVADE labels before and after each validation scenario, and (b) using EVADE validations to remove LLM-detected errors from VARIERR Round 1 (R1) and then fine-tuning with pruned-VARIERR. This allows us to assess whether EVADE-validation is directly useful to modeling and whether it can help refine human annotations.

Baselines & Ceiling We consider two fine-tuning baselines: directly fine-tuning with VARIERR R1 (Round 1) annotations, and fine-tuning after removing human-identified errors in VARIERR R2 human validation. The latter investigates whether removing self-identified errors from VARIERR leads to improvements in HJD alignment.

Models We adopt two small pre-trained language models, **bert-base-uncased** (Devlin et al., 2019) and **roberta-base** (Liu et al., 2019) as backbones following Chen et al. (2024).

Fine-tuning Models are first fine-tuned on the large single-label MNLI training set (392k examples, Williams et al. 2018) and validated on the matched development set (9.8k examples) to learn the general structure of the NLI task. They are then fine-tuned on the label distributions derived from EVADE validations. For each instance, we construct the label distribution by assigning equal probability to each candidate E/N/C label that appears in the set, and normalizing the values accordingly. To evaluate the performance, we use the 1,099 ChaosNLI instances (Nie et al., 2020) that do not overlap with VARIERR, and split them into development and test sets, containing 549 and 550 instances, respectively. We use the label distributions of 100 crowdworkers in ChaosNLI as the gold standard and evaluate model performance in predicting soft labels using KL divergence and weighted F1. See Appendix B Tables 6-7 for hyperparameters.

Results Table 5 presents the results. When comparing two VARIERR baselines, contrary to expectations, removing human-detected errors (R2) does not improve alignment with HJD. For BERT and RoBERTa finetuning, both EVADE-informed setups improve over the pre-validation results (*before* for setup (a) and R1 for setup (b)) in terms of F1. However, KL divergence is only moderately reduced when directly using EVADE labels in setup (a). While setup (b) does not show improvements over baseline R1 in terms of KL divergence, suggesting that EVADE does not fully optimize the label distribution, it nevertheless demonstrates substantial improvements over R2, highlighting the potential of LLMs in detecting errors.

When evaluating the EVADE validation performance, particularly on the improved weighted F1, we observe that *all-llm* delivers the best performance in both setups. This substantiates our hypothesis that LLM validation with additional labels and explanations as context benefits modeling. In contrast, *one-expl* performs slightly worse, despite achieving the strongest results in ranking-based AED evaluation (as shown in Table 4). We attribute this discrepancy to a mismatch between ranking-based evaluation and distribution-based supervision. While *one-expl* produces well-ordered but overconfident scores that benefit ranking, it retains more low-quality explanations, leading to less reliable and poorly calibrated label distributions for the fine-tuning task.

As for model sizes, the larger Llama-70B has an advantage over Llama-8B, but surprisingly, Qwen-7B beats Qwen-72B in BERT and RoBERTa fine-tuning on both F1 and KL in the EVADE-pruning setup (b). However, Qwen-7B significantly degrades when fine-tuned using *all-llm* validation in setup (a). These outlying results echo Qwen-7B's noticeably lower number of explanations after *all-llm* validation reported in Table 4. Above all, our findings suggest that LLMs can effectively help approximate HJD by validating their annotations using EVADE or by validating human ones.

6 Conclusion

In this paper, we introduce EVADE, an LLM-based explanation generation and validation framework designed to remove annotation errors from human label variation (HLV). Our findings show that LLMs produce more comprehensive explanations than human annotators. By analyzing explanation distributions before and after LLM-based validation, we demonstrate that LLM-generated explanations reliably signal valid instances of HLV, and that the validation process effectively refines these distributions to better capture human annotation variability. Larger context inputs and more capable models further yield more consistent and rational validation outcomes. Moreover, the validation process exhibits a high degree of consensus with human judgments in identifying valid labels, likely attributable to the broader range of candidate explanations considered by LLMs. Finally, we show that removing LLM-detected errors from the dataset yields superior fine-tuning performance compared to removing human-detected errors, underscoring the practical value of LLM-based validation for enhancing dataset quality. Overall, these analyses indicate that LLMs achieve strong performance on error detection tasks, often reaching levels comparable to human annotators, while additionally providing broader explanatory coverage and requiring less human expert involvement.

Limitations

While our framework demonstrates that LLM-based explanation generation and validation can effectively detect annotation errors from HLV in NLI tasks, several limitations remain. First, the current study is restricted to the NLI task and to the VARIERR dataset. Datasets that simultaneously provide HLV references and identified errors are scarce, which limits us to fully assess the robustness and generalizability of the framework across other tasks and domains. Second, VARIERR incorporates a peer-validation setup, where annotators evaluate not only their own explanations but also those of others. This setting potentially provides richer information for error detection, but it is not explored in the present study. Incorporating peer-validation signals through LLMs remains an important direction for future work. Lastly, our EVADE framework to use LLMs to detect errors is exploratory. LLMs tend to overgenerate explanations across labels, which makes it challenging to use the same evaluation setup for natural or synthetic errors. Although this behavior may preserve ambiguity and variation, it also risks inheriting social or cultural biases from LLMs, potentially marginalizing minority perspectives. Future work could address this by incorporating more diverse LLMs and retaining low-validaty annotations as ambiguous cases for further review.

Use of AI Assistants

The authors acknowledge the use of ChatGPT for grammatical correction, to improve the coherence of the final manuscripts, and to assist with coding-related tasks.

Acknowledgments

We thank the members of the MaiNLP lab and reviewers for their valuable and constructive feedback. We are especially grateful to Beiduo Chen and Maribel Acosta for insightful suggestions on the draft of the paper. BP acknowledges funding by ERC Consolidator Grant DIALECT 101043235. LZ acknowledges support from the TUM-IAS Dieter Schwarz Fellowship.

References

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).

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.

Gabriel Bernier-Colborne and Sowmya Vajjala. 2024. Annotation errors and ner: A study with ontonotes 5.0. *Preprint*, arXiv:2406.19172.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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 *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6860–6868. AAAI Press.

Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. 2025. A rose by any other name: Llm-generated explanations are good proxies for human explanations to collect label distributions on NLI. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 10777–10802. Association for Computational Linguistics.

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

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.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14349–14371. Association for Computational Linguistics.

Mononito Goswami, Vedant Sanil, Arjun Choudhry, Arvind Srinivasan, Chalisa Udompanyawit, and Artur Dubrawski. 2023. Aqua: A benchmarking tool for label quality assessment. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2025. Litex: A linguistic taxonomy of explanations for understanding within-label variation in natural language inference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 34065–34085. Association for Computational Linguistics.

Chathuri Jayaweera and Bonnie J. Dorr. 2025. From disagreement to understanding: The case for ambiguity detection in NLI. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 37–46, Suzhou, China. Association for Computational Linguistics.

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

Nanjiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:1357–1374.

Uthman Jinadu and Yi Ding. 2024. Noise correction on subjective datasets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5385–5395. Association for Computational Linguistics.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Comput. Linguistics*, 50(3):817–866.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation error detection: Analyzing the past and present for a more coherent future. *Comput. Linguistics*, 49(1):157–198.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Jenny Kunz and Marco Kuhlmann. 2024. Properties and challenges of LLM-generated explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Larson, Adrian Cheung, Anish Mahendran, Kevin Leach, and Jonathan K. Kummerfeld. 2020. Inconsistencies in crowdsourced slot-filling annotations: A typology and identification methods. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5035–5046. International Committee on Computational Linguistics.

Stefan Larson, Anish Mahendran, Andrew Lee, Jonathan K. Kummerfeld, Parker Hill, Michael A. Laurenzano, Johann Hauswald, Lingjia Tang, and Jason Mars. 2019. Outlier detection for improved data quality and diversity in dialog systems. 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 517–527. 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. *Preprint*, arXiv:1907.11692.

Arbi Haza Nasution and Aytug Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language NLP tasks. *IEEE Access*, 12:71876–71900.

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.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

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

Siyao Peng, Zihang Sun, Sebastian Loftus, and Barbara Plank. 2024. Different tastes of entities: Investigating human label variation in named entity annotations. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 73–81, Malta. Association for Computational Linguistics.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9275–9293. Association for Computational Linguistics.

Leon Weber, Robert Litschko, Ekaterina Artemova, and Barbara Plank. 2024. Donkii: Characterizing and detecting errors in instruction-tuning datasets. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 197–215, St. Julians, Malta. Association for Computational Linguistics.

Leon Weber and Barbara Plank. 2023. Activeaed: A human in the loop improves annotation error detection. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pages 8834–8845. Association for Computational Linguistics.

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

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

Guoqing Zheng, Ahmed Hassan Awadallah, and Susan T. Dumais. 2021. Meta label correction for

noisy label learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11053–11061. AAAI Press.

A Prompt Templates

This section presents the prompt templates used in this study for explanation generation (§A.1) and validation (§A.2).

A.1 Explanation Generation Prompt

Figure 3 shows the explanation generation prompt adapted from Chen et al. (2025), with additional instructions allowing the LLM to abstain when a label cannot be reasonably justified. The prompt also explicitly discourages introductory phrases and semantic redundancy.

EXPLANATION GENERATION PROMPT

"role": "system", "content":
You are an expert in Natural Language Inference (NLI). List every distinct explanation for why the statement is {relationship} given the context below without introductory phrases.
If you think the relationship is false given the context, you can choose not to provide explanations. Do not repeat or paraphrase the same idea in different words. End your answer after all reasonable distinct explanations are listed.
Format your answer as a numbered list (e.g., 1., 2., 3.)

"role": "user", "content":
Context: {premise}
Statement: {hypothesis}

Figure 3: Explanation generation prompt.

A.2 Explanation Validation Prompts

Following the prompt design of Weber-Genzel et al. (2024) with with several modifications, Figure 4 presents the prompt for *one-expl* setting and Figure 5 illustrates the prompt for *one-llm* and *all-llm* settings. The latter one explicitly instructs the model to utilize the contextual information provided and to generate all validity scores together in a structured JSON format.

ORIGINAL VALIDATION PROMPT

You are an expert linguistic annotator.
We have collected annotations for an NLI instance together with reasons for the labels. Your task is to judge whether the reasons make sense for the label. Provide the probability (0.0–1.0) that the reason makes sense for the label. Give ONLY the probability, no other words or explanation.

For example:
Probability: <the probability between 0.0 and 1.0 that the reason makes sense for the label, without any extra commentary whatsoever; just the probability!>

Context: {premise}
Statement: {hypothesis}
Reason for label {label}: {reason_text()}
Probability:

Figure 4: Validation prompt for *one-expl* scenario.

ALL VALIDATION PROMPT

You are an expert linguistic annotator.
We have collected annotations for an NLI instance together with explanations for the labels. You will first be shown all explanations together so that you understand the overall context, and then your task is to judge whether each reason makes sense for the label. You must output a single JSON object that maps each explanation's index (1,2,3,...) to its probability in one time.
Provide the probability (0.0 - 1.0) that each reason makes sense for the label. Give ONLY the probability, no other words or explanation.

Output example: {"1": 0.9, "2": 0.8, ...}

Context: {premise}
Statement: {hypothesis}
Reason {i} for label {label}: {reason_text()}
Reason {i} for label {label}: {reason_text()}
...

Now output the JSON object ONLY.

Figure 5: Validation prompt for *one-llm* and *all-llm* scenarios.

B Fine-Tuning Hyper-parameters

Hyper-parameter configurations for fine-tuning BERT and RoBERTa on the MNLI and VariErr dataset variants are summarized in Tables 6 and 7, respectively.

Hyper-parameter	Value
Learning Rate Decay	Linear
Weight Decay	0.0
Optimizer	AdamW
Max sequence length	128
Learning Rate	2e-5
Batch size	16
Num Epoch	3
Metric for best model	eval_accuracy

Table 6: Hyper-parameters used for fine-tuning BERT and RoBERTa on the MNLI dataset.

Hyper-parameter	Value
Learning Rate Decay	Linear
Weight Decay	0.0
Optimizer	AdamW
Learning Rate	2e-5
Batch size	4
Num Epoch	5
Metric for best model	eval_macro_F1

Table 7: Hyper-parameters used for further fine-tuning BERT and RoBERTa on the VARIERR dataset variants.

C Validation Threshold

Table 8 summarizes the optimal validation thresholds determined for each model and setting. These thresholds correspond to the points that best balance lower KL divergence with higher precision and recall when comparing LLM-validated labels against the human-validated VARIERR references, and are applied in all subsequent analysis.

	one-expl	*one-llm*	*all-llm*
Llama-8B	0.8	0.2	0.2
Llama-70B	0.9	0.6	0.6
Qwen-7B	0.7	0.2	0.2
Qwen-72B	0.8	0.7	0.7

Table 8: Validation threshold.

D Full Similarity Results

Table 9 reports full linguistic similarity results between human and LLM explanations across lexical (n-gram overlap), syntactic (POS n-gram overlap), and semantic (cosine and Euclidean) levels. For lexical and syntactic similarity, we compute n-gram overlaps with $n = 1, 2, 3$. Scores are shown for four stages: before validation, after *one-expl* validation, after *one-llm* validation, and after *all-llm* validation.

Models	Setting	Lexical			Syntactic			Semantic		AVG
		n = 1↓	n = 2↓	n = 3↓	n = 1↓	n = 2↓	n = 3↓	Cos.↓	Euc.↓	AVG ↓
within-human		0.313	0.118	0.052	0.713	0.323	0.144	0.529	0.522	0.339
within-LLM										
Llama-8B	before	0.382	0.191	0.103	0.841	0.483	0.265	0.599	0.542	0.426
	one-expl	0.382	0.189	0.101	0.840	0.483	0.264	0.608	0.545	0.427
	one-llm	0.378	0.183	0.097	0.842	0.484	0.263	0.611	0.545	0.425
	all-llm	0.380	0.186	0.099	0.842	0.483	0.263	0.606	0.544	0.425
Llama-70B	before	0.310	0.100	0.030	0.841	0.444	0.201	0.595	0.535	0.382
	one-expl	0.313	0.100	0.031	0.840	0.445	0.201	0.606	0.539	0.384
	one-llm	0.308	0.097	0.029	0.838	0.446	0.200	0.605	0.538	0.383
	all-llm	0.309	0.097	0.029	0.836	0.447	0.202	0.608	0.540	0.384
Qwen-7B	before	0.211	0.050	0.010	0.804	0.409	0.160	0.533	0.517	0.337
	one-expl	0.219	0.053	0.012	0.806	0.411	0.161	0.548	0.521	0.341
	one-llm	0.207	0.048	0.010	0.804	0.415	0.164	0.536	0.518	0.338
	all-llm	0.200	0.042	0.010	0.805	0.428	0.170	0.536	0.518	0.339
Qwen-72B	before	0.266	0.083	0.023	0.824	0.431	0.188	0.578	0.530	0.365
	one-expl	0.270	0.084	0.023	0.824	0.431	0.189	0.583	0.532	0.367
	one-llm	0.265	0.079	0.021	0.824	0.432	0.187	0.583	0.531	0.365
	all-llm	0.263	0.080	0.022	0.823	0.435	0.191	0.584	0.532	0.366
LLM-vs-human										
Llama-8B	before	0.233	0.062	0.020	0.742	0.308	0.116	0.450	0.497	0.303
	one-expl	0.234	0.062	0.020	0.739	0.307	0.116	0.452	0.498	0.303
	one-llm	0.233	0.061	0.019	0.742	0.310	0.117	0.453	0.498	0.304
	all-llm	0.233.	0.060	0.018	0.745	0.309	0.116	0.453	0.498	0.304
Llama-70B	before	0.257	0.079	0.028	0.754	0.322	0.129	0.515	0.514	0.325
	one-expl	0.262	0.081	0.029	0.755	0.325	0.131	0.521	0.516	0.328
	one-llm	0.262	0.082	0.029	0.756	0.326	0.132	0.521	0.516	0.328
	all-llm	0.263	0.083	0.030	0.753	0.326	0.132	0.523	0.517	0.328
Qwen-7B	before	0.207	0.052	0.015	0.727	0.293	0.101	0.469	0.502	0.296
	one-expl	0.213	0.054	0.016	0.728	0.296	0.103	0.477	0.504	0.299
	one-llm	0.206	0.052	0.015	0.725	0.293	0.100	0.469	0.502	0.295
	all-llm	0.199	0.048	0.014	0.716	0.288	0.095	0.470	0.502	0.291
Qwen-72B	before	0.245	0.072	0.023	0.749	0.323	0.125	0.502	0.511	0.319
	one-expl	0.248	0.074	0.024	0.748	0.324	0.127	0.506	0.512	0.320
	one-llm	0.247	0.073	0.023	0.747	0.324	0.126	0.507	0.512	0.320
	all-llm	0.247	0.073	0.024	0.745	0.324	0.126	0.504	0.512	0.319

Table 9: Full results comparing human and LLM explanations across lexical, syntactic, and semantic similarity. Each cell lists scores in the order: before validation, after *one-expl* validation, after *one-llm* validation, and after *all-llm* validation.