SHALA-LLM: Smartly Handling Ambiguous Labels in Aligning LLMs

Anonymous ACL submission

Abstract

Many human-centered tasks, including natural language inference (NLI) and emotion recognition (ER), have multiple plausible interpretations, leading to label ambiguity and challenging disagreements across human annotators. As LLMs are increasingly deployed in real-world settings, faithfully modeling such ambiguity is essential to identify contested inputs, preserve variability in ambiguous cases, and capture the full distribution of human judgments. Yet, existing LLM alignment approaches have predominantly assumed a single correct label, excluding annotator disagreement during optimization. Instead of treating this ambiguity as noise, we show how to treat it as information that improves model behavior through a new algorithm called SMARTLY HANDLING AMBIGUOUS LABELS IN ALIGNING LLMS (SHALA-LLM). This reinforcement learning framework provides a new way for LLMs to learn directly from annotator distributions while dynamically prioritizing highly ambiguous samples during optimization. Experiments on ambiguity-sensitive NLI and ER benchmarks, including ChaosNLI, GoEmotions, and MSP-Podcast, demonstrate that SHALA-LLM improves agreement with annotator label distributions, e.g. on ChaosNLI, it reduces Jensen–Shannon Distance by up to 62.1%. At the same time, SHALA-LLM improves F1 by up to 16.7%, showing that modeling annotator disagreement can also strengthen classification performance[1].

1 Introduction

In many human-centered tasks, different individuals may interpret the same text, speech, or interaction differently depending on contextual understanding, personal experience, cultural background, or emotional perception (Uma et al., 2021). This phenomenon is especially common in natural language inference (NLI) (Nie et al., 2020; Chen et al., 2025) and emotion recognition (ER) (Sethu et al., 2019; Wu et al., 2026) leading to disagreement across human annotators and complicating the alignment of large language models (LLMs). Most existing LLM alignment paradigms treat disagreement as annotation noise, collapsing ambiguous annotations into single target labels through majority voting, label averaging, or calibration methods that assume a single gold standard (Radharapu et al., 2025). In doing so, they discard the disagreement embedded within human annotations, particularly in highly ambiguous settings where annotator uncertainty and conflicting judgments are prominent (Baan et al., 2022). Although recent studies have explored distributional and pluralistic alignment in LLMs, including uncertainty-aware distribution elicitation and alignment with diverse human viewpoints or preferences (Sorensen et al., 2024; Meister et al., 2025), they primarily operate at inference time through prompting or distribution estimation, rather than directly optimizing LLMs with annotator disagreement during training, limiting learning dynamics and model adaptation under highly ambiguous settings (Baan et al., 2022).

In this paper, we present SHALA-LLM: SMARTLY HANDLING AMBIGUOUS LABELS IN ALIGNING LLMS, a new ambiguity-aware distributional alignment framework that directly learns from human disagreement under ambiguous supervision. SHALA-LLM dynamically reweights rollout rewards according to the degree of annotator disagreement, enabling highly ambiguous samples containing richer disagreement structures to exert greater influence during learning while remaining aligned with the underlying distributions of human judgments under ambiguous supervision.

Evaluations on ambiguity-sensitive NLI and ER benchmarks, including ChaosNLI (Nie et al., 2020), GoEmotions (Demszky et al., 2020), and MSP-Podcast (Busso et al., 2025), demon-

[1] Code will be available upon publication

strate that SHALA-LLM consistently improves distributional alignment and classification results compared with majority-label supervision. On ChaosNLI, SHALA-LLM reduces Jensen-Shannon Distance (JSD) by 62.1% while improving F1 by 16.6%; and on MSP-Podcast, SHALA-LLM reduces JSD by 6.2% and improving F1 by 29.2%. Systematic analyses reveal SHALA-LLM exhibits stronger robustness under highly ambiguous conditions, provides particularly strong benefits for inherently ambiguous semantic categories, and encourages reasoning behaviors that better reflect uncertainty and multiple plausible interpretations under ambiguous supervision. This demonstrates that preserving disagreement structures leads to more robust and human-aligned LLM behavior under ambiguity.

2 Related Work

Modeling ambiguity in NLP. Annotations for human-centered tasks from multiple human annotators may disagree, reflecting important information (Aroyo and Welty, 2015; Davani et al., 2022; Chen et al., 2024b; Plank, 2022). This phenomenon has been studied across multiple domains, including semantic reasoning tasks such as natural language inference (NLI) (Jayaweera and Dorr, 2025; Pavlick and Kwiatkowski, 2019) and subjective perception tasks such as emotion recognition and affective computing (Niu et al., 2025; Chou et al., 2025; Wu et al., 2024b; Sethu et al., 2019). Researchers increasingly recognize the importance of addressing annotation ambiguity with prior work in categorical representation settings such as NLI and discrete ER investigating soft-label supervision (Wu et al., 2026; Fard et al., 2025), multi-label formulations (Ando et al., 2019; Ju et al., 2020), and disagreement-aware learning approaches (Chou and Lee, 2019). In continuous affect prediction settings, others have explored Gaussian Distribution (Han et al., 2017), Beta Distribution (Bose et al., 2024; Wu et al., 2024a), and non-parametric label distributions (Wu et al., 2022). However, these approaches have been primarily developed for conventional supervised learning, while LLM alignment needs advancing.

LLM alignment for ambiguous labels. As LLMs expand use in human-centered applications, they increasingly encounter ambiguous labels. To this end, recent studies have begun analyzing how LLMs behave under annotator disagreement and subjective supervision settings (Lu et al., 2025; Jia et al., 2026). Efforts have explored ambiguity-aware instruction fine-tuning (Hong et al., 2025), soft-label supervision (Mostafazadeh Davani et al., 2022; Chen et al., 2025, 2024a), distributional alignment (Meister et al., 2025) and disagreement-aware alignment strategies (Baan et al., 2022) to better align LLM outputs with distributions of human judgments. Smart handling of ambiguous labels for alignment remains to be extended to RL reasoning methods such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024).

3 Proposed SHALA-LLM Framework

3.1 Problem Overview

Fig. 1 overviews our problem. We consider human subjective tasks (Fig. 1a) where annotations are derived from human perceptions, interpretations, and opinions. Labels are collected from multiple annotators, and we treat ambiguity that arises between them as information, not noise.

Assume the labels collected from N human annotators for sample q are denoted as:

$$\{l_{(q,1)}, l_{(q,2)}, \ldots, l_{(q,N)}\}, \tag{1}$$

where each $l_{(q,i)} \in \{1, \ldots, C\}$ corresponds to one of C candidate classes.

Instead of collapsing annotations into a single majority label, we derive an *Empirical Label Distribution* over annotator judgment, which serves as the supervision signal:

$$p_{q,c} = \frac{n_{q,c}}{N}, \quad c = 1, \ldots, C, \quad \sum_{c=1}^{C} p_{q,c} = 1, \tag{2}$$

where $n_{q,c}$ denotes the number of annotators assigning class c for sample q, and $\mathbf{p}_q \in \mathbb{R}^C$ represents the corresponding annotator distribution.

Given an input sample x_q, the LLM generates a sampled textual response $o_{(q,i)}$ containing predicted probabilities over the candidate label space:

$$\hat{\mathbf{p}}_{(q,i)} = [\hat{p}_{(q,i),1}, \hat{p}_{(q,i),2}, \ldots, \hat{p}_{(q,i),C}]. \tag{3}$$

$$\hat{p}_{(q,i),c} = P(\hat{y} = c \mid x_q, o_{(q,i)}). \tag{4}$$

We formulate the learning objective within a reward-based optimization framework (Fig. 1b). Instead of optimizing toward a single collapsed label, our goal is to *encourage generated predictions that better align with the underlying distribution of human interpretations*. Based on this formulation, we

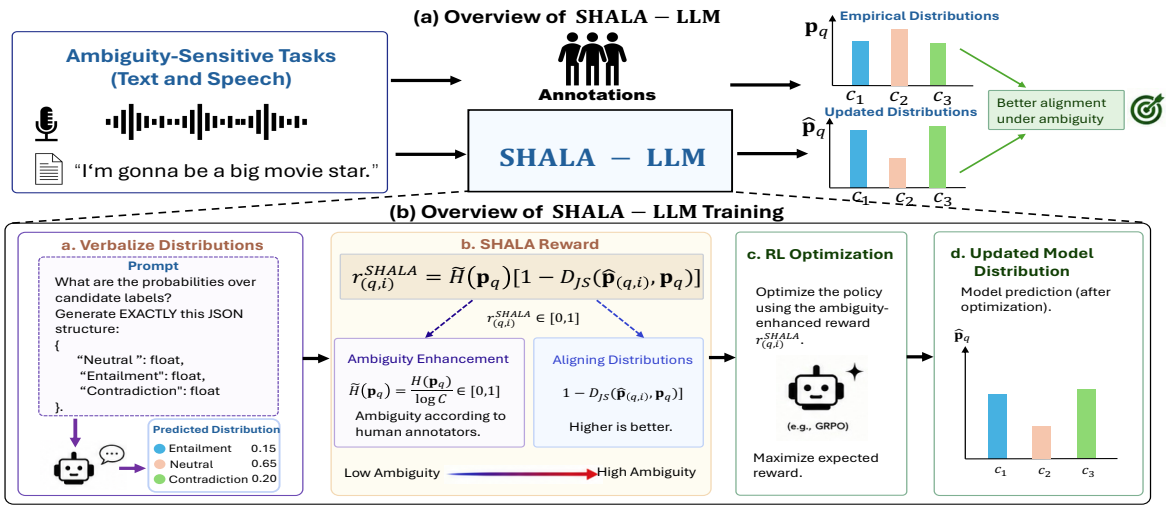

Figure 1: Overview of proposed SHALA-LLM: SMARTLY HANDLING AMBIGUOUS LABELS IN ALIGNING LLMS. Given ambiguity-sensitive tasks such as NLI and ER (a), annotations from multiple annotators are aggregated into empirical label distributions. SHALA-LLM prompts the LLM to verbalize probability distributions over candidate classes (b-a), which are directly aligned with annotator distributions through SHALA reward (b-b). Rollout rewards are dynamically reweighted according to annotator disagreement during GRPO optimization (b-c), producing distributions that better capture uncertainty and diverse human interpretations under ambiguous supervision (b-d).

develop SHALA-LLM that enables LLMs to learn directly from annotator distributions under ambiguous human annotations. We adopt GRPO (Shao et al., 2024) as the optimization backbone due to its flexibility in reward design and its effectiveness in LLM alignment. This formulation allows smart handling of ambiguous labels to be naturally incorporated in reward-based learning.

3.2 Verbalized Distributions from LLM

To obtain distributional outputs from LLMs, we adopt a verbalized distribution prediction formulation (Fig. 1.b-a), where the model directly generates probability estimates over candidate classes in textual form. Such verbalized distributions have been shown to provide an effective mechanism for eliciting uncertainty-aware predictions from LLMs (Radharapu et al., 2025).

The sampled textual response $o_{(q,i)}$ from the input sample x_q is represented as a verbalized distribution over candidate classes:

$$o_{(q,i)} = \{(c_1, \hat{p}_{(q,i),1}), \ldots, (c_C, \hat{p}_{(q,i),C})\}, \tag{5}$$

where $\hat{p}_{(q,i),c}$ denotes the predicted probability assigned to candidate class c for rollout $o_{(q,i)}$. The resulting probability estimates are then parsed into the predicted label distribution $\hat{\mathbf{p}}_{(q,i)}$ defined in Eq. (3).

3.3 Group Relative Policy Optimization

We consider an LLM parameterized by θ as the policy model $\pi_\theta(a \mid s)$, where s denotes the input prompt and a denotes the generated output. Following GRPO (Shao et al., 2024), for each sample q, a rollout group $G_{(q)}$ consisting of multiple sampled outputs $\{o_{(q,i)}\}$ is generated, where $i \in G_{(q)}$ indexes individual rollouts with corresponding rewards $r_{(q,i)}$. GRPO then computes the group-normalized advantage:

$$\hat{A}_{(q,i)} = \frac{r_{(q,i)} - \hat{\mu}_{G_{(q)}}}{\hat{\sigma}_{G_{(q)}} + \varepsilon} \tag{6}$$

where $\hat{\mu}_{G_{(q)}}$ and $\hat{\sigma}_{G_{(q)}}$ are the mean and standard deviation of $\{r_{(q,i)}\}_{i=1}^{|G_{(q)}|}$. With a PPO clipped surrogate $\tilde{A}_{(q,i):k}(\theta)$ constructed from $\hat{A}_{(q,i)}$, GRPO then optimizes $\pi_\theta(a \mid s)$ using a PPO-style trust-region objective[2]:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{(q)\sim\mathcal{D}}\mathbb{E}_{\{o_{(q,i)}\}\sim\pi_{\theta_{\text{old}}}}\Bigg[\frac{1}{|G_{(q)}|}\sum_{i\in G_{(q)}} \frac{1}{n_{o_{(q,i)}}}\sum_{k=1}^{n_{o_{(q,i)}}} \tilde{A}_{(q,i):k}(\theta)\Bigg] - \beta\,\mathbb{E}[D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})] \tag{7}$$

3.4 SHALA Reward

Our hypothesis is that ambiguous samples often contain disagreement structures that provide informative signals regarding nuance, interpretational

[2]Full formulation of GRPO is provided in App. A.

diversity, uncertainty, and conflicting human perceptions. We present a novel way to use this hypothesis to improve LLM alignment: SHALA-LLM, and describe its reward-based optimization algorithm within GRPO.

SHALA-LLM consists of two key components (Fig. 1b-b): First, it aligns LLM predictions with annotator distributions, preserving the structure of human judgments during optimization rather than collapsing supervision into a dominant label. Second, it dynamically reweights rollout contributions according to the degree of annotator disagreement, enabling highly ambiguous samples to exert greater influence during policy optimization.

Following the GRPO formulation in Section 3.3, for each sampled rollout output $o_{(q,i)}$ generated from the LM, we define the ambiguity-enhanced reward:

$$r_{(q,i)}^{\text{SHALA}} = \tilde{H}(\mathbf{p}_q)\left[1 - D_{\text{JS}}\left(\hat{\mathbf{p}}_{(q,i)}, \mathbf{p}_q\right)\right], \tag{8}$$

where $\hat{\mathbf{p}}_{(q,i)}$ denotes the predicted label distribution generated from rollout $o_{(q,i)}$ following Section 3.2, and $\mathbf{p}_q$ denotes the corresponding annotator distribution for sample q defined in Section 3.1.

The component, $1 - D_{\text{JS}}(\hat{\mathbf{p}}_{(q,i)}, \mathbf{p}_q)$, measures agreement between the predicted distribution and the annotator distribution using Jensen-Shannon (JS) Distance. Since $D_{\text{JS}}(\cdot,\cdot) \in [0,1]$, the resulting reward is also bounded within $[0,1]$.

Importantly, not all disagreement structures are equally informative during learning. To explicitly prioritize highly ambiguous samples containing richer uncertainty and interpretational diversity, we introduce ambiguity-enhance reward modulation based on the normalized entropy of the annotator distribution:

$$\tilde{H}(\mathbf{p}_q) = \frac{-\sum_{c=1}^{C} p_{q,c} \log p_{q,c}}{\log C}, \tag{9}$$

where $\tilde{H}(\mathbf{p}_q) \in [0,1]$ quantifies the degree of annotator disagreement in sample q while $p_q c$ refers to the empirical label distribution defined in (2).

Distributional alignment under annotator ambiguity. The proposed reward is incorporated into the GRPO framework through the rollout reward term in Eq. (6). Since GRPO computes policy updates using group-normalized rollout advantages, the reward distribution directly influences rollout contribution dynamics. Unlike discrete majority-label rewards that produce sparse and highly concentrated signals, SHALA-LLM generates continuous rewards proportional to distributional agreement with human annotations. As a result, sampled outputs exhibiting partial alignment with annotator distributions can still contribute positively to policy learning. This reshapes the reward geometry within each rollout group, yielding smoother advantage distributions and more informative policy updates under ambiguous supervision. Compared with conventional majority-label optimization, SHALA-LLM preserves the full structure of human judgments, enabling the model to learn from diverse and potentially conflicting labels rather than only the dominant view.

Dynamic ambiguity-modulated policy optimization. While distributional alignment preserves the full structure of human judgments, SHALA-LLM further introduces reward modulation that dynamically reweights rollout contributions utilizing annotator ambiguity. Since GRPO computes policy updates through rollout advantage magnitudes, SHALA reward scaling directly modulates the relative contribution strength of sampled outputs. Samples with higher annotator ambiguity receive proportionally amplified rollout rewards and therefore exert greater influence on the resulting policy updates. Compared with uniform distributional supervision, the ambiguity-conditioned optimization by SHALA potentially enables the policy model to better learn variability while maintaining alignment with the underlying annotator distributions.

4 Experimental Setup

Datasets. We evaluate the proposed SHALA-LLM framework on ambiguity-sensitive human-centered tasks exhibiting different forms of annotator disagreement, including semantic ambiguity in NLI and affective ambiguity in ER. For NLI, we conduct experiments on ChaosNLI (Nie et al., 2020), which contains 100 annotations per sample to capture diverse semantic interpretations and disagreement, as well as its underlying ChaosNLI-M (Williams et al., 2018) and ChaosNLI-S (Bowman et al., 2015) subsets. For ER, we evaluate on MSP-Podcast (Busso et al., 2025), one of the largest naturalistic speech emotion corpora, and GoEmotions (Demszky et al., 2020), which contains diverse emotion categories and intentionally ambiguous samples. Both emotion datasets include 5–12 annotations per sample, reflecting subjective emotional perception and interpretation. For MSP-Podcast, we include both speech and text modali-

ties to evaluate SHALA-LLM under multimodal ambiguity.

Model. All experiments are conducted using Qwen2.5-Omni-7B (Jin Xu, 2025) as the LLM. We adopt a unified GRPO-based optimization framework across all experiments following the GRPO-Trainer provided by TRL (von Werra et al., 2020).

Baselines. We compare the proposed SHALA-LLM against a range of baselines and ablation settings. These include: (1) a Zero-Shot (ZS) inference model without task-specific training; (2) Majority-Label Supervision (MLS) (refer to App. D.5), which encourages the model to assign high probability mass to the dominant annotation and serves as a reward-based analogue of conventional majority-label training; (3) recent state-of-the-art methods reported for each dataset, as justified in App. D.6; and (4) an ablation variant without ambiguity enhancement, denoted as SHALA-LLM (w/o) Ambi-En, which removes reward modulation by setting $\tilde{H}(\mathbf{p}_q) = 1$.

Evaluation. To evaluate distributional alignment between model predictions and annotator distributions, we report Jensen-Shannon Distance (JSD) (↓) and Bhattacharyya Coefficient (BC) (↑). Both metrics lie in [0,1]. We additionally report conventional classification metrics, including Accuracy (↑), macro F1-score (↑) and Weighted F1 (W-F1) (↑) to assess whether distributional supervision maintains competitive performance under standard evaluation protocols. We note that the latter metrics do not account for ambiguity[3].

5 Results and Discussion

5.1 Overall Performance Comparison

Tables 1 and 2 show the performance across ambiguity-sensitive tasks under both distributional and conventional single-label evaluation metrics.

Comparison with baselines. As shown in Table 1, SHALA-LLM consistently improves both distributional alignment and conventional classification performance compared with ZS and MLS across all datasets and evaluation metrics. On ChaosNLI, SHALA-LLM reduces JSD from 0.477 to 0.181 (62.1% ↓) while improving BC from 0.751 to 0.966 (28.6% ↑) compared with MLS. SHALA-LLM also substantially improves conventional classification performance, increasing ACC from 0.699 to 0.768 (9.9% ↑), F1 from 0.650 to 0.758 (16.6% ↑), and W-F1 from 0.684 to 0.767 (12.1% ↑). Strong consistent trends are also observed in its subset ChaosNLI-M and ChaosNLI-S and observed across ER tasks. On MSP-Podcast, SHALA-LLM improves BC from 0.580 to 0.544 (6.2%↑) and F1 from 0.233 to 0.301 (29.2%↑) compared to MLS. Although the performance differences on GoEmotion between SHALA-LLM and MLS remain relatively moderate on conventional classification metrics, the overall results demonstrate that SHALA-LLM remains highly effective across ambiguous ER settings. Overall, these findings suggest that preserving annotator disagreement during optimization not only improves alignment with human judgment distributions, but can also benefit dominant-label prediction performance under ambiguous supervision.

Notably, we observe for NLI datasets, although the conventional MLS generally improves conventional classification performance over ZS, it simultaneously leads to substantially worse distributional alignment. For example, on ChaosNLI, MLS increases JSD from 0.374 to 0.477 (27.2% ↑) while reducing BC from 0.851 to 0.751 (11.6% ↓). These findings suggest that MLS encourages the model to collapse ambiguity into dominant labels, thereby discarding variability and disagreement structures embedded within annotator distributions.

Ablation studies. We further compare SHALA-LLM against its ablation variant, SHALA-LLM (w/o) Ambi-En, which removes ambiguity-enhanced reward modulation and optimizes only the distributional alignment objective. While SHALA-LLM (w/o) already demonstrates strong improvements over ZS and MLS, the full SHALA-LLM framework consistently achieves further gains across both distributional and classification metrics. On ChaosNLI, SHALA-LLM further reduces JSD from 0.192 to 0.181 (5.4% ↓) while improving BC from 0.964 to 0.966 (0.2% ↑) compared to SHALA-LLM w/o Ambi-En. SHALA-LLM also improves ACC from 0.736 to 0.768 (4.3% ↑), F1 from 0.686 to 0.758 (10.5% ↑), and W-F1 from 0.721 to 0.767 (6.4% ↑).

On GoEmotions, although the performance differences remain relatively moderate, the full SHALA-LLM framework still demonstrates competitive and consistently stronger performance across several evaluation metrics compared with SHALA-LLM (w/o) Ambi-En, including improve-

[3]Full details of experimental settings are in App. D

Table 1: Performance comparisons of the proposed **SHALA-LLM** framework with baselines across NLI and ER datasets. Relative percentage changes are computed with respect to Zero-shot inference (ZS). Best and second-best results are highlighted in **bold** and underline. Relative performance improvement (+) and degradation (-) compared with the ZS are reported in brackets.

Dataset	Method	JSD (↓)	BC (↑)	Acc (↑)	F1 (↑)	W-F1 (↑)
ChaosNLI	Zero-shot	0.375	0.850	0.603	0.473	0.547
	MLS	0.477 (-27.2%)	0.751 (-11.6%)	0.699 (+15.9%)	0.650 (+37.4%)	0.684 (+25.0%)
	SHALA-LLM (w/o Ambi-En)	0.192 (+48.8%)	0.964 (+13.4%)	0.736 (+22.1%)	0.686 (+45.0%)	0.721 (+31.8%)
	SHALA-LLM	**0.181 (+51.7%)**	**0.966 (+13.6%)**	**0.768 (+27.4%)**	**0.758 (+60.3%)**	**0.767 (+40.2%)**
ChaosNLI-M	Zero-shot	0.376	0.845	0.649	0.511	0.607
	MLS	0.510 (-35.6%)	0.731 (-13.5%)	0.695 (+7.1%)	0.637 (+24.7%)	0.683 (+12.5%)
	SHALA-LLM (w/o Ambi-En)	0.188 (+50.0%)	0.968 (+14.6%)	0.701 (+8.0%)	0.588 (+15.1%)	0.671 (+10.5%)
	SHALA-LLM	**0.173 (+54.0%)**	**0.972 (+15.0%)**	**0.760 (+17.1%)**	**0.737 (+44.2%)**	**0.758 (+24.9%)**
ChaosNLI-S	Zero-shot	0.375	0.855	0.557	0.436	0.490
	MLS	0.445 (-18.7%)	0.770 (-9.9%)	0.703 (+26.2%)	0.654 (+50.0%)	0.686 (+40.0%)
	SHALA-LLM (w/o Ambi-En)	0.195 (+48.0%)	0.960 (+12.3%)	0.769 (+38.1%)	0.746 (+71.1%)	0.763 (+55.7%)
	SHALA-LLM	**0.191 (+49.1%)**	**0.961 (+12.4%)**	**0.775 (+39.1%)**	**0.767 (+75.9%)**	**0.775 (+58.2%)**
MSP-Podcast (Speech + Text)	Zero-shot	0.640	0.508	0.421	0.266	0.388
	MLS	0.580(+10.3%)	0.585(+7.7%)	0.488(+13.7%)	0.233(-3.3%)	0.415(+2.7%)
	SHALA-LLM (w/o Ambi-En)	0.550 (+14.1%)	0.658 (+29.5%)	0.482 (+14.5%)	0.276 (+3.8%)	0.433 (+11.6%)
	SHALA-LLM	**0.544(+15.0%)**	**0.694(+36.7%)**	**0.496(+17.8%)**	**0.301(+13.2%)**	**0.455 (+17.3%)**
GoEmotions (Text)	Zero-shot	0.681	0.480	0.361	0.345	0.377
	MLS	0.542(+20.4%)	0.638(+32.9%)	0.595(+64.8%)	**0.591(+71.3%)**	0.595(+57.8%)
	SHALA-LLM (w/o Ambi-En)	**0.449 (+34.1%)**	0.750 (+56.3%)	**0.611 (+69.3%)**	0.544 (+57.7%)	**0.603 (+59.9%)**
	SHALA-LLM	0.465 (+31.7%)	**0.756 (+57.5%)**	0.600(+66.2%)	0.589(+70.7%)	0.595(+57.8%)

Table 2: Performance comparison between the new **SHALA-LLM** framework and existing ambiguity-aware approaches across NLI and ER datasets. Best results are highlighted in **bold**.

Dataset	Method	JSD↓	BC↑	Acc↑	F1↑	WF1↑
ChaosNLI	LLM-Explain (Chen et al., 2025)	0.207	–	–	–	0.645
	LLM-MJD (Chen et al., 2024a)	0.208	–	–	–	0.621
	SHALA-LLM	**0.181**	**0.966**	**0.768**	**0.758**	**0.767**
ChaosNLI-M	Chaos-Benchmark (Nie et al., 2020)	0.306	–	0.635	–	–
	Dist. NLI (Zhou et al., 2022)	0.197	–	0.637	–	–
	AmbiNLI (Meissner et al., 2021)	0.223	–	0.584	–	–
	Flan-T5 (Lee et al., 2023)	0.260	–	0.726	–	–
	SHALA-LLM	**0.173**	**0.972**	**0.760**	**0.737**	**0.758**
ChaosNLI-S	Chaos-Benchmark (Nie et al., 2020)	0.220	–	0.787	–	–
	Dist. NLI (Zhou et al., 2022)	**0.181**	–	**0.794**	–	–
	AmbiNLI (Meissner et al., 2021)	0.189	–	0.755	–	–
	Flan-T5 (Lee et al., 2023)	0.231	–	0.738	–	–
	SHALA-LLM	0.191	**0.961**	0.775	**0.767**	**0.775**
MSP-Podcast	TTS-Benchmark (Jia et al., 2026)	**0.285**	0.621	0.423	0.253	–
	SHALA-LLM	0.544	**0.649**	**0.496**	**0.301**	**0.455**
GoEmotions	AER-LLM (ZS) (Hong et al., 2025)	0.49	0.54	0.371	–	0.357
	AER-LLM (FS) (Hong et al., 2025)	**0.44**	0.70	0.505	–	0.511
	SHALA-LLM	0.47	**0.76**	**0.60**	**0.59**	**0.60**

ments on BC (0.8%↑) and F1 (8.3%↑). These findings suggest that ambiguity-enhanced reward modulation provides additional optimization benefits beyond distributional alignment alone, while the primary gains are already largely achieved through ambiguity-aware distributional supervision.

Comparison with state of the art. We further compare SHALA-LLM with previously reported ambiguity-aware approaches. While direct comparisons should be interpreted cautiously due to differences in model architectures and experimental settings, SHALA-LLM achieves competitive or superior performance on both distributional and conventional metrics. For example, on ChaosNLI, SHALA-LLM achieves lower JSD and substantially higher W-F1 than recent ambiguity-aware methods, including LLM-Explain (Chen et al., 2025) and LLM-MJD (Chen et al., 2024a). Similar trends are observed on MSP-Podcast and GoEmotions. These findings further support the effectiveness and generalizability of SHALA-LLM across diverse ambiguity-sensitive tasks.

Collectively, these findings suggest that highly ambiguous samples provide particularly informative supervision signals during optimization. By dynamically prioritizing samples with substantial annotator disagreement, SHALA-LLM encourages the model to better capture uncertainty and diverse human interpretations embedded in ambiguous annotations. Overall, the results demonstrate that ambiguity-aware optimization in reinforce-

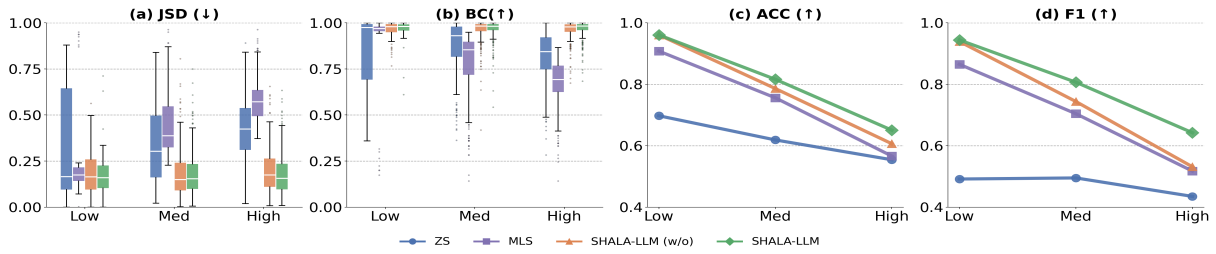

Figure 2: Performance comparison across different ambiguity levels on the ChaosNLI dataset.

ment learning–based alignment leads to stronger agreement with human judgment distributions under subjective supervision.

5.2 Analysis at Different Ambiguity Levels

To gain deeper insight into the robustness and effectiveness of ambiguity-aware optimization, we further analyze model performance across varying levels of annotator disagreement. Specifically, we partition samples according to the entropy of annotator distributions and evaluate different supervision strategies under low-, medium-, and high-ambiguity settings. Details of the data partition procedure can be found in App.B.

Overall performance at all ambiguity levels. Fig. 2 presents the results on the ChaosNLI dataset. Overall, the proposed SHALA-LLM framework consistently achieves superior performance compared with ZS and MLS baselines across all ambiguity levels. Specifically, for distributional evaluation (Fig. 2a-b), both SHALA-LLM and its ablation variant SHALA-LLM (w/o) exhibit consistently better median performance together with reduced variance across ambiguity levels compared with baseline methods. Similar observations are also seen for conventional classification metrics (Fig. 2c-d), where SHALA-LLM (green lines) consistently achieves the strongest overall performance across all ambiguity levels.

Robustness as ambiguity level increases. Importantly, across the boxplots, the performance of SHALA-LLM and SHALA-LLM (w/o) remains relatively stable as ambiguity increases, with no statistically significant degradation observed across ambiguity levels ($p > 0.05$). In contrast, the baseline methods exhibit substantially larger degradation under higher ambiguity conditions. Similarly, although all methods show lower performance in highly ambiguous settings, the degradation remains smaller for SHALA-LLM . For example, the F1 of SHALA-LLM decreases by 32.0% from low- to high-ambiguity, whereas MLS shows a larger drop of 40.2%. These findings suggest that preserving annotator disagreement during optimization improves robustness under increasingly subjective supervision conditions.

Interestingly, MLS (purple boxplots) demonstrates competitive and occasionally superior performance under low-ambiguity settings where dominant consensus labels are clearer. However, its performance drops substantially under medium- and high-ambiguity conditions, for example, its BC decreases by 28.56% from low (0.970) to high-ambiguity (0.693) with even worse performance at high-ambiguity level compared to ZS (0.844). This suggests that conventional majority-label optimization struggles to generalize once supervision becomes increasingly subjective and disagreement structures become more prominent.

Finally, comparing SHALA-LLM against its ablation variant SHALA-LLM (w/o), we observe that SHALA-LLM consistently achieves stronger robustness under higher ambiguity levels, reflected by both higher overall performance and smaller performance degradation as ambiguity increases. These findings indicate that dynamically prioritizing highly ambiguous samples during optimization further strengthens the model's ability to learn from rich disagreement structures embedded within human annotations.

5.3 Interpreting Model Behavior under Ambiguous Supervision

We further investigate how ambiguity-aware optimization influences model behavior under ambiguous supervision. We analyze category-level performance patterns and representative reasoning examples to better understand how SHALA-LLM captures disagreement structures and multiple plausible interpretations in ambiguous settings.

Quantitative analysis across different semantic categories. Fig. 3 presents a category-level anal-

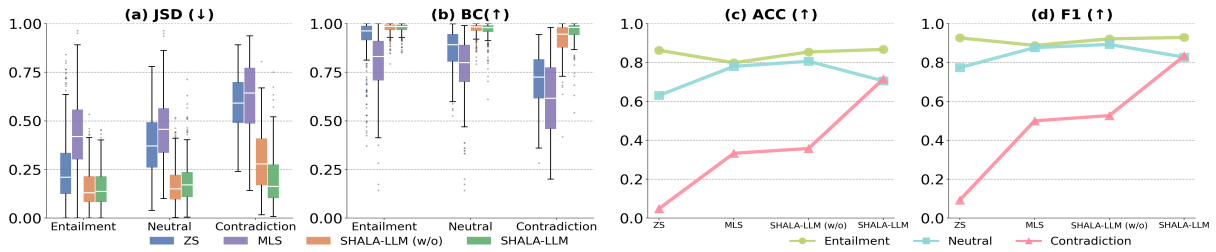

Figure 3: Category-level performance analysis on the ChaosNLI dataset across different classes.

ysis on the ChaosNLI dataset across the *Entailment*, *Neutral*, and *Contradiction* classes. Overall, both SHALA-LLM and its ablation variant SHALA-LLM (w/o) consistently improve performance across all semantic categories compared with ZS and MLS baselines. The largest improvements are observed for the *Contradiction* class, which also exhibits the highest ambiguity and annotator disagreement.

While all methods achieve relatively strong performance on *Entailment* and *Neutral*, baseline approaches struggle substantially on *Contradiction* samples. In contrast, SHALA-LLM achieves dramatic improvements across all evaluation metrics for contradiction cases (Fig. 3c-d), improving F1 from 0.09 (ZS) and 0.50 (MLS) to 0.83, while ACC increases from 0.05 (ZS) and 0.33 (MLS) to 0.72. Similar trends are also observed on distributional metrics with substantially lower JSD and higher BC metrics.

These findings suggest that ambiguity-aware optimization is particularly beneficial for semantic categories with greater uncertainty and conflicting interpretations. By preserving disagreement structures during optimization, SHALA-LLM better captures multiple plausible semantic relationships under ambiguous supervision.

Interpreting reasoning behavior under high ambiguity. We further examine the reasoning behavior of the model under SHALA-LLM training. Table 3 presents a representative highly ambiguous example from the ChaosNLI dataset. While the ZS baseline gives an overconfident prediction, SHALA-LLM generates probability estimates and reasoning patterns that more closely reflect the underlying annotator disagreement (refer to the bold text in the table). Rather than collapsing the sample into a single interpretation, SHALA-LLM reasons over multiple plausible semantic relationships, producing a predicted distribution that better aligns with human judgment distributions. These findings suggest that ambiguity-aware optimization influences not only final predictions, but also the model's underlying reasoning behavior.

Table 3: **Premise**: "In short, we all got tired of clever analyses of what might happen; and throughout economics there was a shift in focus away from theorizing, toward data collection and careful statistical analysis"; **Hypothesis**: "We all got tired of data collection and clever analyses of what might happen";Labels: Entailment (E), Neutral (N), or Contradiction (C).

	Model Reasoning	**E**	**N**	**C**
Ground Truth		.45	.17	.38
Zero-shot	The premise and hypothesis are in direct contradiction. . . . There are **no multiple valid interpretations** here. . . .	.00	.00	1.00
SHALA-LLM	**There are multiple valid interpretations**: 1. The premise and hypothesis could be seen as contradictory because . . . 2. Alternatively, the premise could be interpreted as a shift . . . Given these interpretations, the probability of entailment is relatively **low**, as... The probability of neutrality is **moderate**, as... The probability of contradiction is **high**, as...	.20	.50	.30

6 Conclusion

In this work, we introduced SMARTLY HANDLING AMBIGUOUS LABELS IN ALIGNING LLMS (SHALA-LLM), a reinforcement learning based alignment framework that enables large language models to learn directly from annotator distributions under ambiguous supervision. By preserving disagreement structures and dynamically prioritizing highly ambiguous samples during optimization, SHALA-LLM consistently improves both distributional alignment and conventional classification performance across natural language inference and emotion recognition tasks, while demonstrating stronger robustness under highly ambiguous conditions. Overall, our findings highlight the value of treating human disagreement and ambiguity as informative supervision signals rather than annotation noise, suggesting a promising direction toward more robust and human-centered large language models under subjective real-world settings.

Limitations

One limitation of this work is that SHALA-LLM is primarily evaluated on tasks with relatively structured label distributions. While our approach demonstrates improved alignment with distributions of human judgments, the current experiments do not yet capture more complex forms of ambiguity that arise in open-ended generative tasks, long-form reasoning, or real-world multi-turn interactions. In addition, our reward formulation assumes that annotator disagreement can be adequately represented through empirical label distributions, which may not fully reflect underlying factors such as annotator expertise, demographic variation, or contextual uncertainty. Moreover, our experiments on Qwen2.5-Omni-7B disabled stochastic sampling to improve output consistency and parsing stability for verbalized distribution generation. Future work should further investigate the uncertainty of model-generated distributions and how sampling variability may influence ambiguity-aware optimization under subjective supervision. Finally, although we focus on GRPO-based reasoning alignment, the generalization of the proposed ambiguity-aware reward design to other reinforcement learning and preference optimization frameworks remains future work.

Ethical Considerations

This work focuses on ambiguity-aware alignment for subjective human-centered tasks, including natural language inference and emotion recognition, where disagreement between annotators often reflects diverse interpretations and perceptions rather than annotation noise. By preserving annotator disagreement distributions during optimization, the proposed framework aims to better capture uncertainty and interpretational diversity instead of enforcing a single dominant label. Nevertheless, the underlying datasets may still contain societal, cultural, or demographic biases inherited from human annotations, which could influence model behavior and potentially amplify biased disagreement patterns. Our framework partially mitigates this issue by explicitly modeling disagreement distributions rather than collapsing annotations into majority labels, thereby preserving diverse interpretations under subjective supervision. Additionally, outputs from ambiguity-aware models should not be interpreted as definitive representations of human emotions, intentions, or beliefs, particularly in high-stakes decision-making settings.

Potential risks of this work include the possibility that societal, cultural, or demographic biases embedded within human annotations may also be preserved or amplified through ambiguity-aware optimization. Additionally, outputs from ambiguity-aware models could be misinterpreted as definitive representations of human emotions, intentions, or beliefs, particularly in high-stakes decision-making settings. Our framework partially mitigates these risks by explicitly modeling disagreement distributions rather than collapsing annotations into majority labels, thereby preserving diverse interpretations and uncertainty under subjective supervision.

Acknowledgments

References

Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, and Yushi Aono. 2019. Speech Emotion Recognition Based on Multi-Label Emotion Existence Model. In *Interspeech 2019*, pages 2818–2822.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI magazine*, 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.

Deboshree Bose, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2024. Continuous emotion ambiguity prediction: Modeling with beta distributions. *IEEE Trans. Affect. Comput.*, 15(3):1684–1695.

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*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Carlos Busso, Reza Lotfian, Kusha Sridhar, Ali N Salman, Wei-Cheng Lin, Lucas Goncalves, Srinivas Parthasarathy, Abinay Reddy Naini, Seong-Gyun Leem, Luz Martinez-Lucas, and 1 others. 2025. The msp-podcast corpus. *arXiv preprint arXiv:2509.09791*.

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*, pages 10777–10802, Vienna, Austria. Association for Computational Linguistics.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024a. "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*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024b. "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*, pages 14396–14419.

Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Ali N. Salman, Chi-Chun Lee, and Carlos Busso. 2025. Minority views matter: Evaluating speech emotion classifiers with human subjective annotations by an all-inclusive aggregation rule. *IEEE Transactions on Affective Computing*, 16(1):41–55.

Huang-Cheng Chou and Chi-Chun Lee. 2019. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5886–5890. IEEE.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Ali Pourramezan Fard, Mohammad Mehdi Hosseini, Timothy D Sweeny, and Mohammad H Mahoor. 2025. Affectnet+: A database for enhancing facial expression recognition with soft-labels. *IEEE Transactions on Affective Computing*.

Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 890–897, New York, NY, USA. Association for Computing Machinery.

Xin Hong, Yuan Gong, Vidhyasaharan Sethu, and Ting Dang. 2025. Aer-llm: Ambiguity-aware emotion recognition leveraging large language models. In *ICASSP*, pages 1–5.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *Preprint*, arXiv:2503.24290.

Chathuri Jayaweera and Bonnie 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.

Hong Jia, Weibin Li, Jingyao Wu, Xiaofeng Yu, Yan Gao, Jintao Cheng, Xiaoyu Tang, Feng Xia, and Ting Dang. 2026. Decoding ambiguous emotions with test-time scaling in audio-language models. *arXiv preprint arXiv:2602.03873*.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based label set generation for multi-modal multi-label emotion detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 512–520, New York, NY, USA. Association for Computing Machinery.

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

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *Preprint*, arXiv:2503.20783.

Junyu Lu, Kai Ma, Kaichun Wang, Kelaiti Xiao, Roy Ka-Wei Lee, Bo Xu, Liang Yang, and Hongfei Lin. 2025. Is llm an overconfident judge? unveiling the capabilities of llms in detecting offensive language with annotation disagreement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5609–5626.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 862–869, Online. Association for Computational Linguistics.

Nicole Meister, Carlos Guestrin, and Tatsunori B Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 24–49.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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)*, pages 9131–9143, Online. Association for Computational Linguistics.

Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. 2025. Rethinking emotion annotations in the era of large language models. *IEEE transactions on affective computing*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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

Bhaktipriya Radharapu, Manon Revel, Megan Ung, Sebastian Ruder, and Adina Williams. 2025. Arbiters of ambivalence: Challenges of using llms in no-consensus tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4677–4731.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. 2025. Omni-r1: Do you really need audio to fine-tune your audio llm? *Preprint*, arXiv:2505.09439.

Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth Narayanan. 2019. The ambiguous world of emotion representation. *arXiv preprint arXiv:1909.00360*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

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

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Adina Williams, Nikita Nangia, and Samuel 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, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2022. A novel sequential monte carlo framework for predicting ambiguous emotion states. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8567–8571.

Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2024a. Dual-constrained dynamical neural odes for ambiguity-aware continuous emotion prediction. *arXiv preprint arXiv:2407.21344*.

Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2024b. Emotion recognition systems must embrace ambiguity. In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 166–170. IEEE.

Jingyao Wu, Grace Lin, Yinuo Song, and Rosalind Picard. 2026. Amber2: Dual ambiguity-aware emotion recognition applied to speech and text. *arXiv preprint arXiv:2601.18010*.

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*. Association for Computational Linguistics.

A Full GRPO Formulation

For completeness, we provide the full formulation of GRPO, including the surrogate objective and importance-sampling formulation. Accordingly, for task m and sample q, GRPO samples a rollout group $G_{(m,q)}$ of responses $\{o_{(m,q,i)}\}$, where $i \in G_{(m,q)}$ indexes individual rollouts (i.e., a sampled response) with rewards $r_{(m,q,i)}$, computing the group-normalized advantage:

$$\hat{A}_{(m,q,i)} = \frac{r_{(m,q,i)} - \hat{\mu}_{G_{(m,q)}}}{\hat{\sigma}_{G_{(m,q)}} + \varepsilon}, \tag{10}$$

where $\hat{\mu}_{G_{(m,q)}}$ and $\hat{\sigma}_{G_{(m,q)}}$ are the empirical mean and standard deviation of $\{r_{(m,q,i)}\}_{i=1}^{|G_{(m,q)}|}$. GRPO then optimizes $\pi_\theta(a \mid s)$ by performing a PPO-style trust-region update. At token position k of response $o_{(m,q,i)}$, $\varphi_{(m,q,i):k}(\theta)$ denotes the importance sampling ratio between π_θ and the old policy $\pi_{\theta_{\text{old}}}$, $\tilde{A}_{(m,q,i):k}(\theta)$ denotes the PPO-clipped surrogate using $\hat{A}_{(m,q,i)}$, and $J_{\text{GRPO}}(\theta)$ averages this surrogate over tokens and rollout samples with an optional KL penalty to a reference policy π_{ref} (with weight β). We summarize these with a compact objective:

$$\begin{aligned}
\varphi_{(m,q,i):k}(\theta) &= \frac{\pi_\theta\big(o_{(m,q,i):k} \mid q,\, o_{(m,q,i):<k}\big)}{\pi_{\theta_{\text{old}}}\big(o_{(m,q,i):k} \mid q,\, o_{(m,q,i):<k}\big)} \\
\tilde{A}_{(m,q,i):k}(\theta) &= \min\Big(\varphi_{(m,q,i):k}(\theta)\,\hat{A}_{(m,q,i)}, \\
&\qquad \text{clip}\big(\varphi_{(m,q,i):k}(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\,\hat{A}_{(m,q,i)}\Big) \\
J_{\text{GRPO}}(\theta) &= \mathbb{E}_{q\sim\mathcal{D}_m}\mathbb{E}_{\{o_{(m,q,i)}\}\sim\pi_{\theta_{\text{old}}}}\Bigg[\frac{1}{|G_{(m,q)}|}\sum_{i=1}^{|G_{(m,q)}|} \\
&\frac{1}{n_{o_{(m,q,i)}}}\sum_{k=1}^{n_{o_{(m,q,i)}}}\tilde{A}_{(m,q,i):k}(\theta)\Bigg] - \beta\,\mathbb{E}[D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})]\,.
\end{aligned} \tag{11}$$

B Dataset Description

B.1 ChaosNLI Dataset

ChaosNLI (Nie et al., 2020) is a text-based dataset designed to capture human disagreement in natural language inference tasks. It consists of premise–hypothesis pairs drawn from established NLI benchmarks and reannotated with 100 independent human raters per example. Raters are asked to determine whether the hypothesis is entailed by, contradicted by, or neutral with respect to the premise. In this work, we restrict our analysis to examples originating from the Stanford Natural Language Inference Corpus (SNLI) (Bowman et al., 2015) and Multi-Genre Natural Language Inference Corpus (MNLI) (Williams et al., 2018).

ChaosNLI-S (Bowman et al., 2015) is a large-scale natural language inference dataset constructed from image captions in the Flickr30k corpus. Human annotators are presented with a text caption describing an image and asked to write three corresponding hypotheses: one that is definitely true given the caption (entailment), one that may be true but is uncertain (neutral), and one that is definitely false (contradiction). This process produces sentence pairs with relatively simple linguistic structure and forces the data to be balanced among these classes.

ChaosNLI-M (Williams et al., 2018) is modeled on the SNLI corpus but extends the task to a broader range of spoken and written text. It collects premise sentences from ten distinct genres including fiction, government reports, travel guides, telephone speech, and spoken conversations. Annotators are then asked to label hypothesis–premise pairs according to the same entailment, neutral, and contradiction categories, resulting in a dataset with more diverse language and more challenging inference phenomena.

B.2 MSP Podcast

The MSP Podcast (v1.12) (Busso et al., 2025) dataset is a large-scale naturalistic speech corpus constructed from podcast audio. Each short audio segment is annotated by at least five human raters for perceived emotion labels. To ensure consistency across samples, we follow prior work in restricting the label space to a standard set of eight emotion categories: anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise (Hong et al., 2025; Wu et al., 2026). Examples containing annotations outside this label set are discarded.

B.3 GoEmotions

GoEmotions (Demszky et al., 2020) is a large-scale text emotion dataset consisting of English Reddit comments annotated with fine-grained emotion labels. Each example is initially annotated by 3–5 human raters, where annotators are allowed to assign multiple emotion labels to a single example. If no two annotators agree on any emotion label, two additional raters are assigned to the example. Annotators may also choose to assign no label if they consider the example too difficult to classify. Rather than using the aggregated binary labels released with the original dataset, we utilize the raw

annotator-level responses to construct human annotation distributions.

We apply several preprocessing steps before constructing these distributions. First, we discard examples where fewer than three annotators provide valid emotion labels to ensure a reliable estimate of an underlying distribution. Second, the original dataset contains 27 fine-grained emotion categories and a neutral label, making distribution prediction substantially more difficult and sparse. Following the Ekman taxonomy proposed in (Demszky et al., 2020), we map the original labels into six broader emotion categories as shown in Table 4. For each example, annotator labels are mapped to the corresponding Ekman categories, aggregated across raters, and normalized to produce a probability distribution over emotions.

Finally, analysis of the majority label of resulting annotation distributions revealed substantial class imbalance. We observed that a large proportion of examples were concentrated in the joy and neutral classes. To reduce bias toward high-frequency classes and improve evaluation across emotions, we apply stratified sampling to select approximately equal numbers of examples for each dominant emotion category.

Table 4: Mapping of GoEmotions 27 fine-grained emotion labels to Ekman's six basic emotion categories (Demszky et al., 2020).

Ekman Category	GoEmotions Fine-Grained Labels
Joy	Admiration, Amusement, Approval, Caring, Desire, Excitement, Gratitude, Joy, Love, Optimism, Pride, Relief
Sadness	Disappointment, Embarrassment, Grief, Remorse, Sadness
Anger	Anger, Annoyance, Disapproval
Surprise	Confusion, Curiosity, Realization, Surprise
Fear	Fear, Nervousness
Disgust	Disgust
Neutral	Neutral

C Data Processing

C.1 Dataset statistics

C.1.1 ChaosNLI

The ChaosNLI dataset consists of 3,113 examples, including 1,514 examples sourced from SNLI and 1,599 examples sourced from MNLI. Prior to training, the dataset is randomly shuffled and partitioned into training, validation, and test sets using a 65/15/20 split. Table 5 summarizes the distribution of the majority label across the dataset, where the majority label for an example is defined as the label receiving the largest proportion of annotations.

Table 5: ChaosNLI dataset statistics by majority label.

Dataset	Majority Label	Count
ChaosNLI	Entailment	1,168
	Neutral	1,397
	Contradiction	548
ChaosNLI-S	Entailment	424
	Neutral	811
	Contradiction	279
ChaosNLI-M	Entailment	744
	Neutral	586
	Contradiction	269

C.1.2 MSP Podcast

After preprocessing and filtering, the resulting dataset contains 12,955 examples. Following prior work (Wu et al., 2026), we adopt an even five-fold partitioning strategy rather than the standard speaker-based splits to achieve a more balanced distribution of ambiguous examples across data partitions. We divide the dataset into five equal-sized folds, using three folds for training, one for validation, and one for testing, resulting in an approximate 60/20/20 split. Table 6 summarizes the distribution of majority emotion labels across the dataset.

Table 6: MSP Podcast dataset statistics by majority label.

Majority Label	Count
Angry	1,053
Contempt	1053
Disgust	806
Fear	603
Happy	3519
Neutral	4609
Sad	432
Surprise	880

C.1.3 GoEmotions

After preprocessing and filtering, we collect 6567 examples in the GoEmotions dataset. Examples

are randomly shuffled and partitioned into training, validation, and test sets using a 70/15/15 split. Table 7 summarizes the distribution of the majority emotion label across the dataset.

Table 7: GoEmotions dataset statistics by majority label.

Majority Label	Count
Anger	1,000
Disgust	1000
Fear	640
Joy	927
Sadness	1000
Surprise	1000
Neutral	1000

C.2 Data partition across different ambiguity levels

In addition to overall dataset statistics, we analyze how examples are distributed across different levels of ambiguity.

C.2.1 ChaosNLI

For ChaosNLI, ambiguity is quantified using the entropy of the ground-truth annotation distribution. Lower entropy corresponds to stronger agreement among annotators, while higher entropy indicates greater disagreement. We partition examples into three ambiguity levels: low, medium, and high, using normalized entropy thresholds of $[0, 0.33)$, $[0.33, 0.66)$, and $[0.66, 1]$, respectively. Table 8 summarizes the resulting distribution of examples across these ambiguity levels.

Table 8: ChaosNLI dataset statistics by ambiguity level.

Dataset	Ambiguity Level	Count
ChaosNLI	Low	377
	Medium	1493
	High	1243
ChaosNLI-S	Low	327
	Medium	865
	High	322
ChaosNLI-M	Low	50
	Medium	628
	High	921

An interesting observation is that ChaosNLI-M contains a larger proportion of examples concentrated in the high-ambiguity level compared to ChaosNLI-S. This difference likely reflects characteristics of the underlying source datasets. ChaosNLI-S is constructed from image captions and therefore tends to contain shorter, more concrete descriptions with relatively straightforward relationship structures. In contrast, ChaosNLI-M spans a broader range of genres and linguistic styles, introducing more complex language and inference patterns that naturally lead to greater annotator disagreement.

C.2.2 MSP-Podcast

For MSP-Podcast, we characterize ambiguity using the number of active labels present in the ground-truth distribution rather than entropy directly. We define an active label as an emotion category receiving non-zero probability mass in the annotation distribution. Intuitively, the lowest ambiguity occurs when all annotators agree on a single emotion label. Examples with two active labels correspond to cases where annotators only identify two distinct emotion categories, while progressively larger numbers of active labels indicate greater disagreement among raters. We partition examples into categories of one, two, three, and four-or-more active labels. Table 9 presents the distribution of examples across these groups along with the average normalized entropy of the ground-truth distribution.

Table 9: MSP Podcast dataset statistics by ambiguity level.

Active Labels	Count	Average Entropy
1 label	2963	0.0000
2 labels	2410	0.2537
3 labels	3152	0.4661
4+ labels	4430	0.6604

C.2.3 GoEmotions

Similar to MSP-Podcast, ambiguity in GoEmotions is characterized using the number of active labels within the ground-truth distribution. Examples with a single active label represent strong annotator agreement, whereas examples with multiple active labels indicate increasing diversity in annotation responses. We group examples into categories containing one, two, three, and four-or-more active labels. Table 10 summarizes the distribution of examples across these ambiguity levels and the average entropy of the corresponding ground-truth

distributions.

Table 10: GoEmotions dataset statistics by ambiguity level.

Active Labels	Count	Average Entropy
1 label	1434	0.0000
2 labels	1828	0.3140
3 labels	1714	0.5138
4+ labels	1591	0.6875

D Experimental Setup

D.1 Artifact Usage and Licensing.

All datasets and pretrained models used in this work are publicly available for research purposes and were used in accordance with their intended usage conditions and licenses.

D.2 Model and GRPO Configuration

We conduct all experiments using the Qwen2.5-Omni-7B model (Jin Xu, 2025) and implement training with the GRPOTrainer from the TRL framework (von Werra et al., 2020). We follow the GRPO configuration settings of (Rouditchenko et al., 2025), which also performs GRPO fine-tuning on Qwen2.5-Omni, particularly for improving the model's audio interpretation capabilities. We utilize the AdamW optimizer with an initial learning rate of 1×10^{-6}. Following the GRPO setup, we set the number of rollouts to 4, temperature to 1.2, and maximum completion length to 128 tokens. We additionally set $\beta = 0$, removing the KL divergence regularization term from the original GRPO formulation (Shao et al., 2024). This choice is motivated by recent studies (Liu et al., 2025; Hu et al., 2025) showing that the KL divergence term is not necessary effective GRPO training. All experiments were conducted using the Qwen2.5-Omni-7B model backbone, which contains approximately 7 billion parameters.

For generation, we set do_sample to False, resulting in deterministic decoding where the model selects the highest-probability token at each generation step. This improves reproducibility and ensures that evaluation results are not affected by sampling variation. Training is performed on a single compute node with two NVIDIA H200 GPUs and 400GB of memory. The batch size per GPU is set to 1 with gradient accumulation over two steps, resulting in an effective batch size of 4 prompts per optimization step. We employ DeepSpeed ZeRO Stage 3 optimization for efficient distributed training and memory management. Training progress and evaluation metrics are logged using WandB. Unless otherwise specified, all remaining hyperparameters follow the default settings defined in GRPOConfig within TRL (von Werra et al., 2020). Results are reported as single-run experiments due to the computational cost of GRPO-based LLM fine-tuning.

D.3 GRPO Configuration

D.4 Prompts

Below, we provide the prompt templates used for training across each dataset. Table 11 presents the prompt template used for ChaosNLI, Table 12 shows the template for MSP Podcast, and Table 13 shows the template used for GoEmotions. While the task-specific context and label spaces differ across datasets, all prompts follow a common structure consisting of background information, the target utterance to evaluate, task instructions, and output formatting constraints.

Table 11: Prompt template used for ChaosNLI with example target utterance.

Section	Prompt Content
Background	
Target Utterance	Premise: A child in a red jacket, waist deep in a pit on the beach Hypothesis: A child is building a sandcastle on the beach
Task	Given a premise and a hypothesis, predict the probability of the relationship between them from the following options: entailment, neutral, contradiction. 1. entailment: the hypothesis logically follows from the premise 2. neutral: neither entailment nor contradiction can be determined 3. contradiction: the hypothesis conflicts with the premise You MUST produce a calibrated probability distribution.
Output Constraints	Before outputting, check if the format of your output is in accordance with the requirements I provided. 1. 1. Generate the label probabilities in EXACTLY this JSON structure: {{"entailment": float, "neutral": float, "contradiction": float}}. 2. The sum of all probabilities must be exactly 1.0. 3. Do not include any explanations or text besides the dictionary.

Table 12: Prompt template used for MSP Podcast with example target utterance.

Section	Prompt Content
Background	Two speakers are having a conversation.
Target Utterance	that's right. spilling my load of liberty all over your faces. it's the golden stallion of the tech-
Task	Predict the probability of the emotion in the target utterance from the following options: angry, contempt, disgust, fear, happy, neutral, sad, surprise. You MUST produce a calibrated probability distribution. Identify any emotional cues expressed by the speaker, even subtle ones. For each of the following emotions from angry, contempt, disgust, fear, happy, neutral, sad, surprise, evaluate whether the emotion is present. If only one emotion present, assign 1.0 to that emotion. If multiple emotions are present, estimate the relative strength and assign a probability to each emotion. The probability should reflect how much time, intensity, and presence each emotion has in the conversation.
Output Constraints	Before outputting, check if the format of your output is in accordance with the requirements I provided. 1. Generate the emotion probabilities in EXACTLY this JSON structure: {{"Angry": float, "Contempt": float, "Disgust": float, "Fear": float, "Happy": float, "Neutral": float, "Sad": float, "Surprise": float}}. 2. The sum of all probabilities must be exactly 1.0. 3. Do not include any explanations or text besides the dictionary.

Table 13: Prompt template used for GoEmotions with example target utterance.

Section	Prompt Content
Background	This is a text comment extracted from Reddit.
Target Utterance	Dear [NAME] man! (Irony intended) You didńt say "proof" you said "evidence"!
Task	Assign a calibrated probability distribution over the following emotion categories: anger, disgust, fear, joy, sadness, surprise, neutral. These categories follow the Ekman taxonomy: • anger: annoyance, disapproval, hostility • disgust: contempt, revulsion • fear: nervousness, anxiety, dread • joy: happiness, admiration, gratitude, excitement, love, optimism, pride, relief, amusement • sadness: disappointment, grief, remorse, embarrassment • surprise: curiosity, confusion, realization • neutral: no clear emotion expressed Predict the probability of the emotion in the target utterance from the following options: anger, disgust, fear, joy, sadness, surprise, neutral. You MUST produce a calibrated probability distribution. For each of the following emotions from anger, disgust, fear, joy, sadness, surprise, neutral, evaluate whether the emotion is present. If only one emotion present, assign 1.0 to that emotion. If multiple emotions are present, estimate the relative strength and assign a probability to each emotion. The probability should reflect how much intensity and presence each emotion has in the comment.
Output Constraints	Before outputting, check if the format of your output is in accordance with the requirements I provided. 1. Generate the emotion probabilities in EXACTLY this JSON structure: {{"anger": float, "disgust": float, "fear": float, "joy": float, "sadness": float, "surprise": float, "neutral": float}}. 2. The sum of all probabilities must be exactly 1.0. 3. Do not include any explanations or text besides the dictionary.

D.5 Majority-label supervision (MLS) Baseline

As a reference baseline, we first consider conventional Majority-label supervision (MLS), where the annotator distribution is reduced to its majority label:

$$y_q^* = \arg\max_c p_{q,c}, \tag{12}$$

where $p_{q,c}$ denotes the annotator distribution defined in Eq. 2.

Given the predicted distribution $\hat{\mathbf{p}}_{(q,i)}$ defined in Eq. 3, the rollout reward is defined as the probability assigned to the majority label:

$$r_{(q,i)}^{\text{maj}} = \hat{p}_{(q,i),y_q^*}. \tag{13}$$

This formulation encourages the model to assign high probability mass to the dominant annotation and serves as a reward-based analogue of conventional MLS.

D.6 Justification of Included Baselines

We include a diverse range of recent studies and strong baselines for comparison across both ambiguity-aware learning and large language model alignment settings. On ChaosNLI (including ChaosNLI-M and ChaosNLI-S), we include the original benchmark framework proposed together with the dataset collection process, denoted as Chaos-Benchmar (Nie et al., 2020). We further include recent studies that explicitly model annotator disagreement or distributional supervision under the same benchmark settings, including distribution-aware LLM alignment and ambiguity-aware learning approaches (Chen et al., 2025, 2024a; Lee et al., 2023).

Additionally, we include prior ambiguity-aware fine-tuning methods such as (Zhou et al., 2022; Meissner et al., 2021), which are based on BERT-style architectures rather than LLM backbones. Although these methods do not perform reinforcement learning based alignment, they are included because they optimize directly on annotator distributions and report evaluation results using the same distributional metrics adopted in our work,

enabling meaningful comparison under ambiguity-aware supervision.

For emotion recognition datasets, there remain relatively limited prior studies explicitly addressing ambiguity-aware distributional supervision. On MSP-Podcast, we include the recent benchmark study on decoding ambiguous emotions using test-time scaling, denoted as TTS-Benchmark (Jia et al., 2026). On GoEmotions, we include recent LLM-based ambiguity-aware emotion recognition approaches (Hong et al., 2025) that report distributional emotion outputs under both zero-shot (ZS) and few-shot (FS) settings, enabling direct comparison under ambiguity-aware evaluation protocols.

E Evaluation Metrics

Jensen-Shannon Distance Jensen-Shannon distance (JS) is a metric that measures the similarity between two probability distributions. It is a symmetric and smoothed version of the Kullback–Leibler divergence (KL) with finite values bounded between 0 and 1 when using base 2 logarithms. A smaller JS distance would indicate more similarity between two probability distributions, with zero indicating two identical distributions. Given probability distributions P and Q, let $M = \frac{P+Q}{2}$ and $D_{\mathrm{KL}}(P\|Q)$ be the KL divergence given by

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

The JS distance is defined as

$$D_{\mathrm{JS}}(P\|Q) = \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M)}.$$

By comparing comparing both distributions to a central mixture distribution, the JS distance metric is more stable than KL divergence. In our analysis, we set Q to be the ground truth probability distribution from the annotations and P to be the model's predicted probability distribution.

Bhattacharyya Coefficient The Bhattacharyya Coefficient (BC) measures the amount of overlap or statistical similarity between two probability distributions. It is bounded between 0 and 1, where a value of 1 indicates identical distributions and 0 indicates no overlap. For discrete probability distributions, BC is defined by

$$BC(P,Q) = \sum_{x\in\mathcal{X}} \sqrt{P(x)\cdot Q(x)}.$$

Accuracy For classification accuracy, we first derive a single label, $\hat{y}$, from the model's predicted probability distribution, P, by taking the argmax over the label space $\mathcal{Y}$

$$\hat{y} = \arg\max_{y\in\mathcal{Y}} P(y).$$

The ground truth label, y^*, is similarly derived from the ground truth distribution, Q, by selecting the label with maximum probability mass.

$$\hat{y^*} = \arg\max_{y\in\mathcal{Y}} Q(y).$$

Given N evaluation examples, accuracy is then defined as

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i^*].$$

F1 Score The F1 score is the harmonic mean of precision and recall for a given class c

$$\text{F1}_c = \frac{2\cdot\text{Precision}_c\cdot\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c},$$

where

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

and TP_c, FP_c, and FN_c denote the true positives, false positives, and false negatives for class c, respectively. Hard predicted labels $\hat{y}$ and ground truth labels y^* are derived from Q and P via argmax as described above. The macro-averaged F1 score averages F1_c uniformly across all $|\mathcal{Y}|$ classes:

$$\text{F1}_{\text{macro}} = \frac{1}{|\mathcal{Y}|}\sum_{c\in\mathcal{Y}} \text{F1}_c.$$

Weighted F1 Score The weighted F1 score extends the macro-averaged F1 by accounting for class imbalance through support-weighted averaging. Specifically, the F1 score for each class c is weighted by its support N_c, defined as the number of ground-truth instances belonging to class c. The weighted F1 is then computed as:

$$\text{F1}_{\text{weighted}} = \frac{\sum_{c\in\mathcal{Y}} N_c\cdot\text{F1}_c}{\sum_{c\in\mathcal{Y}} N_c},$$

where $N_c = TP_c + FN_c$ represents the total number of true instances of class c. This metric therefore reflects both per-class performance and the empirical class distribution, giving higher influence to more frequent classes in the dataset.

F Ambiguity Analysis Results for GoEmotions and MSP-Podcast

In this section, we expand on the results in Table 1 by providing an analysis of model performance at different ambiguity levels for each dataset as defined in App. C.2

We provide an in-depth analysis of model performance across different ambiguity levels on GoEmotions and MSP-Podcast in Fig. 5 and 4. Both SHALA-LLM and its variant, SHALA-LLM (w/o Ambi-En), consistently outperform the ZS and MLS baselines across all ambiguity levels and evaluation metrics, particularly on the distributional metrics JSD and BC. Importantly, as ambiguity increases, SHALA-LLM exhibits the smallest performance degradation, demonstrating the robustness of ambiguity-aware training under highly ambiguous conditions. Although SHALA-LLM and SHALA-LLM (w/o Ambi-En) do not show substantial gains over MLS on conventional classification metrics compared with the improvements over ZS, this is likely because MLS is primarily optimized for dominant single-label classification. Overall, SHALA-LLM demonstrates strong capability in both capturing annotator distributions, as evidenced by JSD and BC, and maintaining competitive performance on dominant-label classification, further supporting the discussion in Section 5.2.

G Per-Class Analysis Results

In this section, we expand on the results in Table 1 by providing an analysis of model performance at the per-class level.

G.1 MSP Podcast Per-Class Analysis

Fig. 6 presents the per-class analysis across different models. It is observed that, on distributional evaluation metrics, both SHALA-LLM and its ablation variant SHALA-LLM (w/o) consistently outperform the baseline ZS and MLS models, while still maintaining relatively strong performance on conventional classification metrics. Importantly, the largest improvements are concentrated in the *Happy* class, likely due to class imbalance, as this class contains substantially more training samples. Interestingly, the *Contempt* class exhibits relatively poor performance under MLS and SHALA-LLM (w/o), whereas the full SHALA-LLM framework substantially improves both ACC and F1. This suggests that ambiguity-enhanced optimization helps the model better capture highly ambiguous or underrepresented emotional categories by amplifying informative disagreement structures during learning.

G.2 GoEmotions Per-Class Analysis

Fig. 7 presents the per-class analysis on the GoEmotions dataset. Both SHALA-LLM and SHALA-LLM (w/o Ambi-En) demonstrate strong performance compared with the ZS and MLS baselines on distributional evaluation metrics, while also maintaining generally competitive classification performance. Notably, improvements are particularly evident for highly ambiguous categories such as *Surprise*, which was intentionally designed by the dataset creators to exhibit ambiguity. Although these improvements are less apparent on conventional classification metrics due to the inherently ambiguous nature of the category, the distributional metrics reveal that SHALA-LLM better captures the underlying annotator disagreement.

H Model Reasoning Under Different Ambiguity Levels

We further include model reasoning outputs for representative low-, medium-, and high-ambiguity examples in Tables 14 to 16. For low- and medium-ambiguity cases, we do not observe substantial differences between SHALA-LLM and the baseline. However, for highly ambiguous cases, SHALA-LLM demonstrates a stronger capability to reason over multiple valid interpretations, as discussed in detail in Section 5.3 of the main paper.

We include the following prompt during inference to encourage the model to explain its reasoning.

> Before producing your final answer, explain

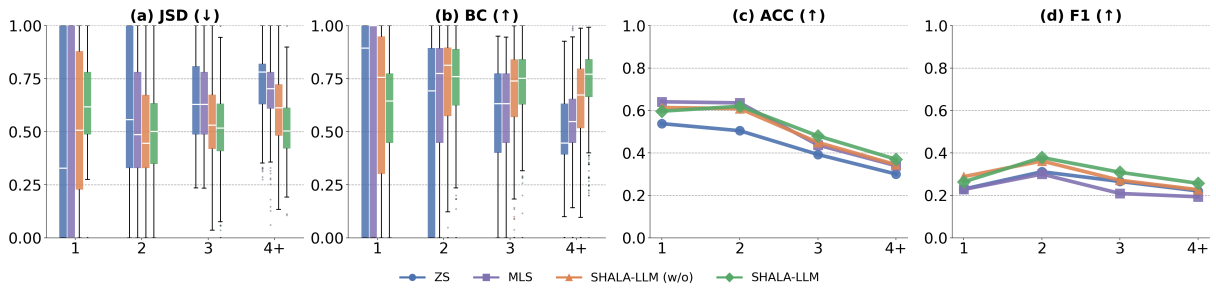

Figure 4: Performance comparison across different ambiguity levels on the **MSP Podcast** dataset. x-axis denotes the number of active labels in the ground truth distribution which reflects ambiguity levels from low (left) to high (right) explained in App. C.2.2

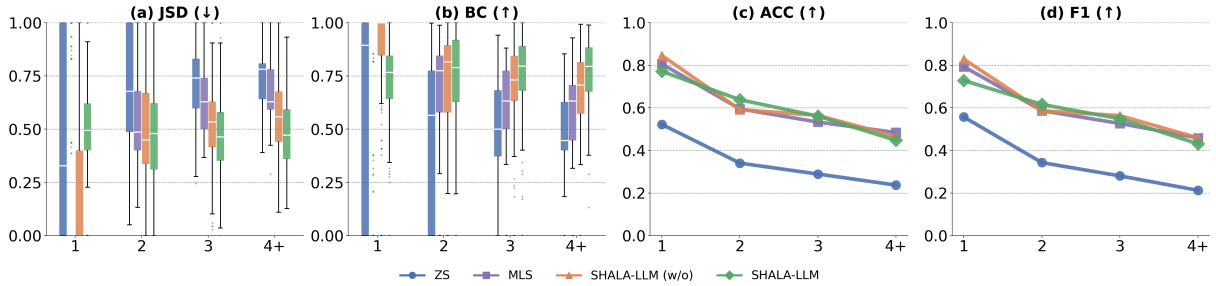

Figure 5: Performance comparison across different ambiguity levels on the **GoEmotions** dataset. x-axis denotes the number of active labels in the ground truth distribution which reflects ambiguity levels from low (left) to high (right) explained in App. C.2.3

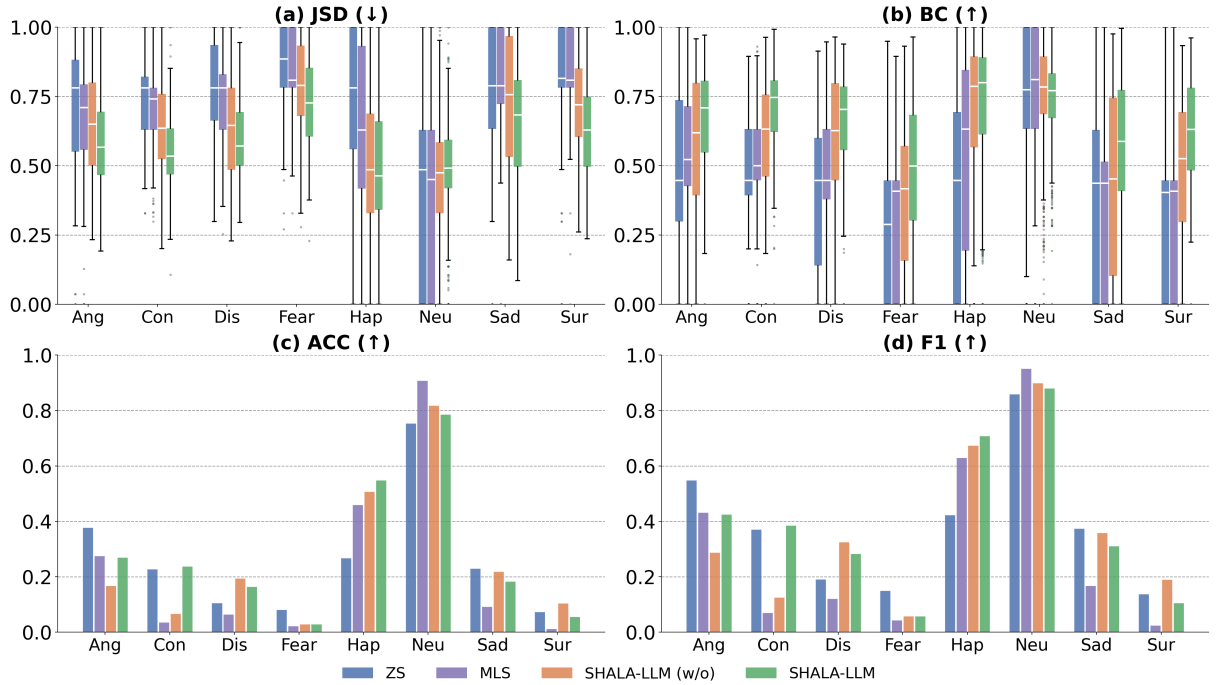

Figure 6: Performance comparison across different emotion classes on the **MSP Podcast** dataset.

your reasoning:

- What is the relationship between the

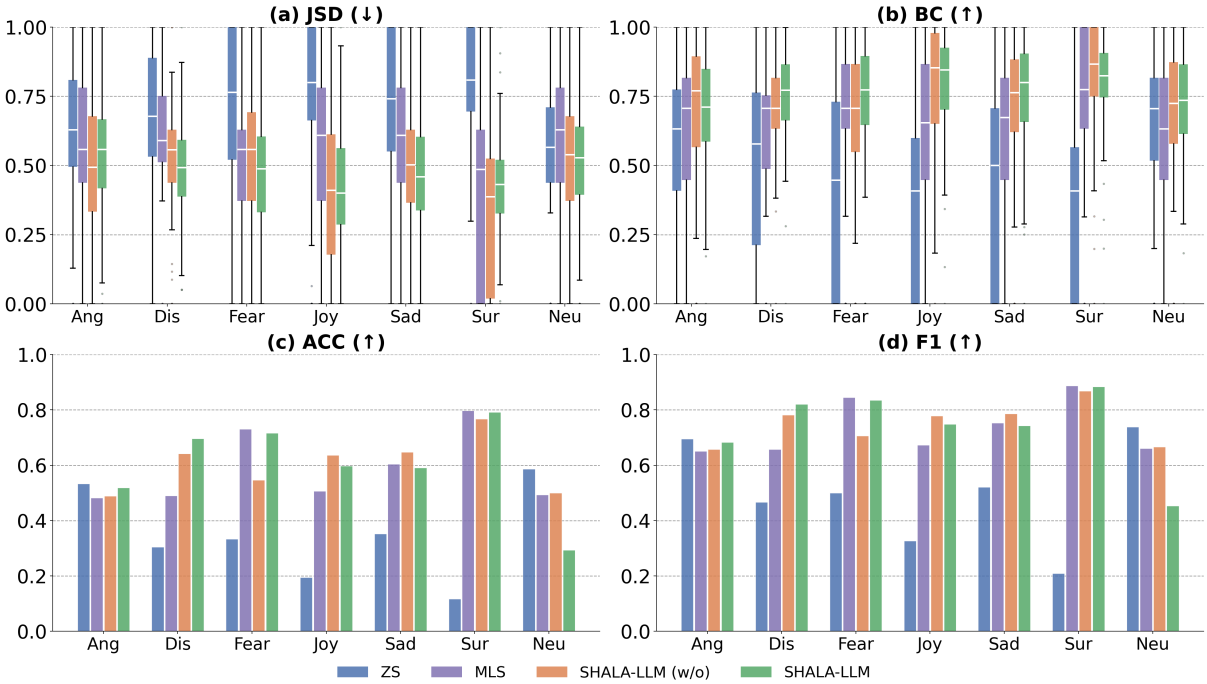

Figure 7: Performance comparison across different emotion classes on the **GoEmotions** dataset.

premise and hypothesis?"

- Are there multiple valid interpretations? If so, describe them.
- How does your reasoning inform the probability you assign to each label?

Then output EXACTLY this JSON on its own line: {{"entailment": float, "neutral": float, "contradiction": float}}

I Use of AI Assistants

AI assistants were used for language refinement, editing support during manuscript preparation and debugging for codes. All technical content, experimental design, implementation, and scientific claims were developed and verified by the authors.

Table 14: Observing model reasoning on a ChaosNLI example with a **High** level of ambiguity. Given the **Premise**: "In short, we all got tired of clever analyses of what might happen; and throughout economics there was a shift in focus away from theorizing, toward data collection and careful statistical analysis" and **Hypothesis**: "We all got tired of data collection and clever analyses of what might happen", annotators are asked to give a vote on entailment (E), neutral (N), or contradiction (C).

	Model Reasoning	**E**	**N**	**C**
Ground Truth		.45	.17	.38
Baseline	1. **Premise Analysis**: - The premise states that there was a shift in focus away from theorizing and clever analyses towards data collection and careful statistical analysis. - It implies that people got tired of clever analyses and moved towards data collection. 2. **Hypothesis Analysis**: - The hypothesis states that people got tired of data collection and clever analyses of what might happen. - This is the opposite of what the premise suggests. 3. **Relationship Between Premise and Hypothesis**: - The premise and hypothesis are in direct contradiction. The premise suggests a shift towards data collection, while the hypothesis suggests a shift away from data collection. 4. **Multiple Valid Interpretations**: - There are no multiple valid interpretations here. The premise and hypothesis are clearly in conflict. 5. **Probability Assignment**: - Since the premise and hypothesis are in direct contradiction, the probability of entailment is 0. - The probability of neutrality is also 0 because the relationship is not neutral. - The probability of contradiction is 1 because the hypothesis directly contradicts the premise.	.00	.00	1.00
R_1	The relationship between the premise and the hypothesis is somewhat complex. The premise states that there was a shift in focus away from theorizing and towards data collection and statistical analysis, which implies that the focus on clever analyses of what might happen was lessened. However, the hypothesis suggests that we got tired of data collection and clever analyses of what might happen, which is somewhat contradictory to the premise. There are multiple valid interpretations: 1. The premise and hypothesis could be seen as contradictory because the premise indicates a shift away from clever analyses, while the hypothesis suggests tiredness with both data collection and clever analyses. 2. Alternatively, the premise could be interpreted as a shift in focus, which might imply that the tiredness with clever analyses was a result of the shift towards data collection and statistical analysis. Given these interpretations, the probability of entailment is relatively low, as the hypothesis suggests a contradiction to the premise. The probability of neutrality is moderate, as the premise and hypothesis could be seen as having different focuses. The probability of contradiction is high, as the hypothesis directly contradicts the premise.	.20	.50	.30

Table 15: Observing model reasoning on a ChaosNLI example with a **Medium** level of ambiguity. Given the **Premise**: A child in a red jacket, waist deep in a pit on the beach and **Hypothesis**: A child is building a sandcastle on the beach, annotators are asked to give a vote on entailment (E), neutral (N), or contradiction (C).

	Model Reasoning	**E**	**N**	**C**
Ground Truth		.00	.77	.23
Baseline	1. **Premise Analysis**: - The premise describes a child in a red jacket, waist deep in a pit on the beach. This suggests the child is engaged in some form of activity in the sand, but it does not specify what that activity is. 2. **Hypothesis Analysis**: - The hypothesis states that a child is building a sandcastle on the beach. This is a specific activity that involves creating a sandcastle. 3. **Relationship Between Premise and Hypothesis**: - The premise does not provide enough information to confirm or deny the hypothesis. The child could be building a sandcastle, but they could also be doing other activities like digging, playing, or just standing in the sand. 4. **Multiple Valid Interpretations**: - There are multiple valid interpretations of the premise. The child could be building a sandcastle, but they could also be doing other activities. This lack of specificity in the premise means that the hypothesis is neither entailed nor contradicted by the premise. 5. **Probability Assignment**: - Since the premise does not provide enough information to confirm or deny the hypothesis, the relationship between the two statements is neutral. The probability of entailment is low because the premise does not logically follow the hypothesis. The probability of contradiction is also low because the premise does not conflict with the hypothesis. The probability of neutrality is high because the premise does not provide enough information to determine the relationship.	.05	.90	0.05
R_1	The relationship between the premise and the hypothesis is neutral. The premise describes a child in a red jacket, waist deep in a pit on the beach, but it does not provide enough information to determine if the child is building a sandcastle. The child could be doing various activities such as playing, digging, or even just standing in the pit. There are multiple valid interpretations of the premise, which makes it difficult to assign a high probability to either entailment or contradiction. Therefore, the probability of entailment is low, the probability of neutrality is high, and the probability of contradiction is also low.	.10	.80	.10

Table 16: Observing model reasoning on a ChaosNLI example with a **Low** level of ambiguity. Given the **Premise**: Linda Tripp was indicted for illegally taping telephone conversations with Monica Lewinsky and **Hypothesis**: Monica Lewinsky was illegally taped by Linda Tripp, annotators are asked to give a vote on entailment (E), neutral (N), or contradiction (C).

	Model Reasoning	**E**	**N**	**C**
Ground Truth		.95	.03	.02
Baseline	1. **Premise Analysis**: - The premise states that "Linda Tripp was indicted for illegally taping telephone conversations with Monica Lewinsky." - This implies that Linda Tripp was the one who performed the illegal act of taping. 2. **Hypothesis Analysis**: - The hypothesis states that "Monica Lewinsky was illegally taped by Linda Tripp." - This suggests that Monica Lewinsky was the one who was taped by Linda Tripp. 3. **Relationship Between Premise and Hypothesis**: - The premise and hypothesis are describing the same event but from different perspectives. - The premise focuses on Linda Tripp as the tapper, while the hypothesis focuses on Monica Lewinsky as the taped person. - Since the premise explicitly states that Linda Tripp was the one who taped, the hypothesis directly contradicts this by suggesting that Monica Lewinsky was the one who was taped. 4. **Multiple Valid Interpretations**: - There are no multiple valid interpretations here. The premise and hypothesis are in direct conflict. 5. **Probability Assignment**: - Given the direct contradiction between the premise and the hypothesis, the probability of the hypothesis being true given the premise is 0. - The probability of the hypothesis being neutral or entailment is also 0 because the hypothesis directly contradicts the premise.	.00	.00	1.00
R_1	The relationship between the premise and the hypothesis is that the hypothesis logically follows from the premise. The premise states that Linda Tripp was indicted for illegally taping telephone conversations with Monica Lewinsky, which directly implies that Monica Lewinsky was the subject of the illegal taping by Linda Tripp. **There is no ambiguity in this relationship**, and the hypothesis is a direct consequence of the premise. Given this clear relationship, the probability of entailment is high. The hypothesis is neutral in terms of additional information not provided in the premise, and there is no contradiction since the hypothesis is a logical outcome of the premise.	.85	.10	.05