

How to Solve Noise and Hard Samples: Filtering and Difficulty-adaptive Augmentation for Fine-grained Negative Emotion Classification

Lianxi Wang^{1,2}, Siting Lin¹, Chao Liang¹, Yuying Li¹, Xinfeng Liao³, Nankai Lin^{1,2†}

¹School of Information Science and Technology, Guangdong University of Foreign Studies

²Guangdong Engineering Research Center of Data Security Governance and Privacy Computing

³College of Information Science and Technology, Jinan University

[†]Corresponding author

wanglianxi@gdufs.edu.cn, neakail@outlook.com

Abstract

Fine-grained negative emotion classification suffers from subjective annotation noise and semantically ambiguous hard samples, which severely degrades model performance. Most existing methods cannot distinguish valid hard samples from noise and lack strategies to improve hard sample learnability. To address these issues, we propose a **Noise Filtering and Difficulty-adaptive Hard Sample Augmentation (NF-DHSA)** framework for fine-grained negative emotion classification. It first mines high-risk candidates using a hybrid score of normalized loss and predictive entropy, then applies a multi-view logic-guided filtering mechanism to reliably separate noise from valid hard samples. Further, difficulty-adaptive dynamic augmentation generates moderate-difficulty, label-stable augmented samples, and prototype-guided dual-granularity contrastive learning aligns original-augmented pairs while strengthening global category boundaries. The results reported on three benchmark datasets show that NF-DHSA outperforms the baseline methods overall, with Macro-F1 scores improved by 1.25%, 2.47%, and 3.53% on the VIRUS, NLPCC, and COVID datasets, respectively.

1 Introduction

Fine-grained negative emotion classification is a core fundamental task in natural language processing (NLP), which aims to automatically identify and distinguish subtle negative emotional states in textual data (Sánchez-Rada and Iglesias, 2019; Demszky et al., 2020). Nevertheless, the inherent diversity and subjectivity of human emotional expression pose two critical challenges to textual negative emotion detection. On the one hand, the overlapping semantic boundaries between different negative emotions lead to a large number of hard boundary samples in real-world datasets, whose ambiguous expression patterns are difficult

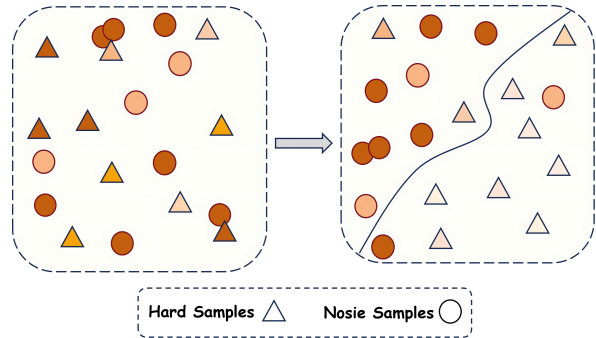


Figure 1: The lighter the color, the easier the sample is for the model to learn. In the left, hard samples are darker. In the right, hard samples become lighter, and hard samples are distinguished from noisy samples.

for models to distinguish and learn effectively. On the other hand, the subjectivity of manual annotation and the ambiguity of emotional definition inevitably produce noisy samples with incorrect or inaccurate labels. The coexistence of hard samples and noisy samples severely degenerates the feature learning and category discrimination ability of emotion detection models, becoming a key bottleneck restricting the performance improvement of fine-grained negative emotion detection.

Existing research (Hu et al., 2023; Yu et al., 2024; Zhang et al., 2024) has made considerable progress in enhancing the discriminative learning capability of negative emotion models, which effectively alleviates the problem of insufficient emotional feature extraction to a certain extent. However, most existing approaches ignore the essential differences between hard boundary samples and noisy samples, and fail to explicitly distinguish these two types of high-risk samples during model training. Hard samples contain valid ambiguous emotional semantic information and are valuable for optimizing category boundaries, while noisy samples are invalid interference data that would mislead model optimization. Undifferentiated training of the two types of samples greatly limits model

performance.

A few recent studies attempt to leverage large language models (LLMs) and semantic similarity metrics to identify and filter noisy samples (Song et al., 2026; Nahum et al., 2025; Sosea and Caragea, 2026). Despite the improved noise detection effect, LLM-based screening methods suffer from inherent limitations. Extensive studies have verified that LLMs are prone to hallucination and unstable semantic judgment in complex fine-grained classification scenarios (Goel et al., 2025; Bavaresco et al., 2025; Shcharbakova et al., 2025). For ambiguous hard boundary samples with blurred emotional features, LLMs often misjudge valid hard samples as noisy data and mistakenly eliminate valuable training samples. In addition, mainstream solutions for hard samples, including loss function optimization and model architecture redesign, only adjust the model learning paradigm externally without optimizing the inherent learning difficulty of hard samples (Wang et al., 2025b; Xiong et al., 2024). These methods cannot fundamentally resolve the ambiguity of hard sample semantics, resulting in persistent insufficient learning of complex negative emotional boundaries. As shown in Figure 1, To address the above research gaps, this paper aims to address two issues:

- *How to effectively distinguish hard samples from noisy samples?*
- *How to convert hard samples into forms that are easy to learn?*

To this end, we propose NF-DHSA, a Noise Filtering and Difficulty-adaptive Hard Sample Augmentation framework for fine-grained negative emotion classification. NF-DHSA first mines high-risk candidates through difficulty-aware screening and separates noisy samples from valid hard samples via multi-view logic-guided filtering. It then generates moderately difficult, label-stable augmented samples for the retained hard samples and further optimizes the representation space through prototype-guided dual-granularity contrastive learning. The main contributions of this paper are summarized as follows: 1) We propose NF-DHSA, a unified framework for noise filtering and hard sample learning in fine-grained negative emotion classification. 2) We design a difficulty-aware multi-view filtering mechanism to distinguish noisy samples from valid hard samples. 3) We introduce difficulty-adaptive dynamic augmen-

tation with prototype-guided contrastive learning to improve hard sample learnability.

2 Related Work

2.1 Text Denoising

Learning from noisy labels remains a central challenge in NLP. Recent methods address this issue through co-regularization and optimization-level noise injection, demonstrating empirical strong robustness on standard noisy benchmarks (Zhou and Chen, 2021; Wu et al., 2022). Another line of work detects label errors through task-loss ranking and self-influence scores, providing useful signals for data cleaning and curriculum design (Chong et al., 2022; Bejan et al., 2023). In weakly supervised text classification, PIEClass addresses pseudo-label noise by combining prompt-based pseudo labeling with noise-robust iterative ensemble training (Zhang et al., 2023).

Nevertheless, real-world label noise is often more complex than simulated corruption. NoiseBench shows that existing noise-robust models still fall short under diverse real annotation errors (Merdjanovska et al., 2024), while LLM and small model collaboration improves efficiency but still struggles with clean and noisy separation (Yuan et al., 2024). These studies suggest that numerical signals alone are insufficient, LLM judgments may be unstable, and hard-but-learnable samples still require more effective supervision.

2.2 Data Augmentation

Data augmentation addresses data scarcity, expression diversity, and class imbalance. Simple perturbations (EDA/AEDA) and sequence-level mixing (SeqMix) already deliver substantial gains on text classification tasks (Wei and Zou, 2019; Karimi et al., 2021; Guo et al., 2020). Recent work explores LLM-based generative augmentation. Diffusion-based augmentation balances diversity and consistency by reconstructing label-critical tokens for low-resource sentiment classification. (Chen et al., 2024), while recent LLM-driven pipelines improve data coverage by generating semantically constrained variants (Rad et al., 2025; Wang et al., 2025a).

Previous studies show that emotion-anchored contrastive learning improves representation separability for similar emotions, while contrastive regularization enhances model robustness by filtering out contextual noise (Yu et al., 2024; Liao

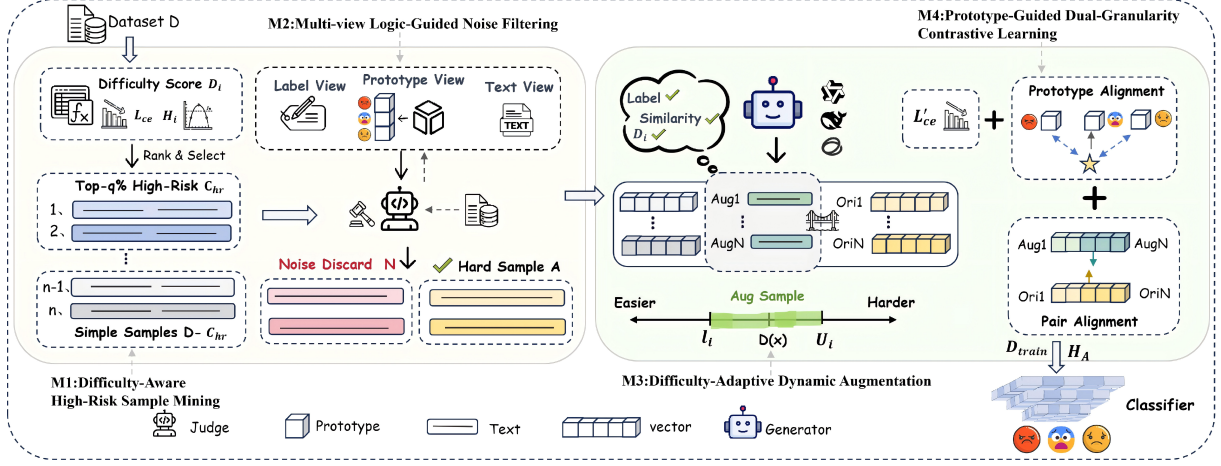


Figure 2: Overview of the proposed method. NF-DHSA consists of four modules. M1 and M2 are designed to address the first question: *How to effectively distinguish hard samples from noisy samples?* M3 and M4 address the second question: *How to convert hard samples into forms that are easy to learn?*

et al., 2026). However, most augmentation methods focus on data expansion and diversity, but overlook difficulty control and hard-sample learnability. Since hard samples often contain valid yet ambiguous emotional cues, we formulate augmentation as a difficulty-adaptive process that generates semantically consistent samples with moderate difficulty for more learnable supervision.

3 Methodology

As illustrated in Figure 2, we propose **NF-DHSA** for fine-grained negative emotion classification. First, NF-DHSA employs difficulty-aware high-risk sample mining module to construct a high-risk candidate pool. Then, a multi-view logic-guided filtering module is introduced to further separate truly noisy samples from valid hard samples, thereby avoiding the mistaken removal of hard samples that are valuable for boundary learning. Furthermore, a difficulty-adaptive dynamic augmentation module dynamically constructs a sample-level difficulty window according to the difficulty of each hard sample, and generates augmented samples with stable labels, consistent semantics, and moderate difficulty. Finally, a prototype-anchored dual-granularity contrastive optimization module transfers learnable semantics from augmented samples back to original hard samples, strengthening boundary modeling for fine-grained negative emotions.

3.1 Problem Definition

Given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where x_i denotes a text sample and $y_i \in \mathcal{Y}$ denotes its annotated negative emotion label, the objective is to

train a fine-grained negative emotion detector f_θ . Therefore, instead of directly removing all high-loss samples, our method first verifies whether a high-risk sample is truly noisy, and then improves the learnability of the retained hard samples through difficulty-adaptive augmentation.

3.2 Difficulty-Aware High-Risk Sample Mining

We first train a base detector f_θ on the original training set and use it to identify high-risk samples. For each sample x_i , the model outputs the class probability distribution and cross-entropy loss:

$$p_\theta(y_i | x_i) = \text{softmax}(f_\theta(x_i)). \quad (1)$$

$$\ell_i = -\log p_\theta(y_i | x_i). \quad (2)$$

The loss term reflects how difficult a sample is for the detector to fit under its current annotated label. Nevertheless, loss alone is insufficient to distinguish noisy samples from valid hard boundary samples (Yuan et al., 2024). To further characterize the model’s prediction behavior, we introduce the normalized predictive entropy:

$$H_i = -\frac{1}{\log |\mathcal{Y}|} \sum_{c \in \mathcal{Y}} p_\theta(c | x_i) \log p_\theta(c | x_i), \quad (3)$$

where $\mathcal{Y}$ denotes the label set and c is a class index iterating over $\mathcal{Y}$. Intuitively, a sample with high loss and low entropy suggests that the model confidently predicts a label different from the annotation. Such samples are more likely to suffer from semantic-label inconsistency. Based on this

observation, we define the high-risk score as:

$$D_i = \sigma(\alpha \cdot \tilde{\ell}_i)(1 - H_i), \quad (4)$$

where $\tilde{\ell}_i$ denotes the normalized loss and α is a temperature coefficient. We rank training samples according to D_i and select the top $q\%$ samples with the highest scores as the high-risk candidate set:

$$C_{hr} = \text{Top}_q \left(\{(x_i, y_i, D_i)\}_{i=1}^N \right). \quad (5)$$

It is worth noting that C_{hr} is regarded as a mixed set containing both noisy samples and valid hard boundary samples.

3.3 Multi-view Logic-Guided Noise Filtering

To distinguish noisy samples from valid hard samples, we introduce a Multi-view logic-guided noise filtering module. Directly using the LLM as a free-form annotator may introduce hallucinated judgments and thus compromise the reliability of noise identification (Kim et al., 2024). Therefore, we formulate noise detection as a multi-view structured evidence assessment task rather than an unconstrained binary decision. Specifically, we constrain the LLM to construct noise evidence from three complementary perspectives: the label definition, the class prototype, and the text itself.

Label-definition View. This view measures whether a candidate sample is consistent with the basic semantic boundary of its original label. Let d_y denote the label definition of class y . For a candidate sample (x_i, y_i) , the LLM outputs a definition-level consistency judgment, following the scoring criteria detailed in Appendix C.

$$\begin{aligned} r_i^{def} &= \text{LLM}_{def}(x_i, y_i, d_y), \\ r_i^{def} &\in \{\text{HIGH}, \text{MID}, \text{LOW}\}. \end{aligned} \quad (6)$$

When the judgment is LOW, the sample is difficult to be covered by the basic emotional meaning of its original label, thereby providing definition-level noise evidence:

$$e_i^{def} = \mathbb{I}(r_i^{def} = \text{LOW}). \quad (7)$$

Emotion-class Prototype View. Relying only on label definitions may be insufficient to characterize implicit emotions and boundary expressions. Therefore, for each specific emotion category y , we construct a class profile G_y and a multi-level prototype set $\mathcal{T}_y$ as class-level emotional semantic anchors. Prototype samples are selected from

clean instances that are correctly predicted with high confidence:

$$\mathcal{T}_y = \{x_j \mid y_j = \hat{y}_j = y, p_\theta(y \mid x_j) \geq \rho, x_j \notin C_{hr}\}, \quad (8)$$

where $\rho = 0.7$, $\hat{y}_j$ is the label predicted by the base classifier. We select prototypes in a stratified manner to cover low-difficulty, medium-difficulty, and high-difficulty clean samples, thereby improving the coverage of intra-class expression variations. Based on G_y and $\mathcal{T}_y$, the LLM outputs a prototype-level consistency judgment:

$$\begin{aligned} r_i^{proto} &= \text{LLM}_{proto}(x_i, y_i, G_y, \mathcal{T}_y), \\ r_i^{proto} &\in \{\text{HIGH}, \text{MID}, \text{LOW}\}. \end{aligned} \quad (9)$$

If $r_i^{proto} = \text{LOW}$, this indicates that the sample differs from its assigned class under criteria such as emotional trigger logic. The prototype-level noise evidence is defined as:

$$e_i^{proto} = \mathbb{I}(r_i^{proto} = \text{LOW}). \quad (10)$$

Text-only View. The previous two views are both conditioned on the original label. To obtain label-independent evidence, we further ask the LLM to infer the most likely emotion category solely from the text itself:

$$(\hat{y}_i^{llm}, s_i, m_i) = \text{LLM}_{text}(x_i), \quad (11)$$

where $\hat{y}_i^{llm}$ denotes the LLM-predicted emotion label, s_i denotes the confidence score, and m_i denotes the indicator of mixed-emotion judgment. $s_i \in [0, 1]$ denotes the confidence score, and m_i indicates whether mixed emotions are detected. If the LLM predicts a label that is inconsistent with the original annotation with high confidence, text-level noise evidence is formed. We set the confidence threshold as $\tau_t = 0.6$, and define the text-level noise evidence as:

$$e_i^{text} = \mathbb{I}(\hat{y}_i \neq y_i \wedge s_i \geq \tau_t). \quad (12)$$

Finally, to reduce the risk of misjudgment caused by hallucinations from a single view, we adopt a multi-view consistency rule for noise identification. Let $\mathcal{N}$ denote the set of noisy samples. A sample is identified as noisy only when text-level evidence agrees with definition- or prototype-level evidence, or when both label-conditioned views provide consistent negative evidence:

$$\begin{aligned} \mathcal{N} = \left\{ x_i \in C_{hr} \mid \left[\left(e_i^{def} \vee e_i^{proto} \right) \wedge e_i^{text} \right] \right. \\ \left. \vee \left[e_i^{def} \wedge e_i^{proto} \right] \right\}. \end{aligned} \quad (13)$$

The remaining high-risk samples are retained as valid hard boundary samples:

$$\mathcal{A} = \mathcal{C}_{hr} \setminus \mathcal{N}. \quad (14)$$

3.4 Difficulty-Adaptive Dynamic Augmentation

After noise verification, the samples in $\mathcal{A}$ are considered semantically valid, but they remain difficult for the detector to learn directly. To reduce their inherent learning difficulty, we generate an augmented sample x_i^b for each hard boundary sample x_i . Different from conventional LLM-based data augmentation, our objective is neither to simply increase data diversity nor to generate the easiest paraphrase. Instead, the augmented sample is expected to preserve the original emotional semantics while presenting key discriminative cues in a more learnable form. For each valid hard sample $x_i \in \mathcal{A}$, we construct a sample-adaptive target difficulty window $\mathcal{W}_i$

$$\begin{aligned} l_i &= \max(d_{\min}, \gamma_l d_i), \\ u_i &= \min(\gamma_u d_i, d_i - \epsilon), \end{aligned} \quad (15)$$

where $[l_i, u_i]$ denotes the dynamic difficulty interval that the generated augmented sample is allowed to fall into. Here, γ_l and γ_u control the lower and upper difficulty ratios of the generated sample relative to the original sample, respectively, while $d_{\min}$ prevents the generated sample from degenerating into an extremely simple expression. To avoid label drift or over-simplification in LLM-generated samples, we introduce classifier feedback as an immediate dynamic verification signal, where the generated augmented sample is required to satisfy the conditions: $D_i^b \in \mathcal{W}_i$, $\hat{y}_i^b = y_i$, and $p_\theta(y_i | x_i^b) > \tau_b$. An augmented sample is accepted only if it satisfies the difficulty window, semantic consistency, label consistency, and length constraints. Otherwise, the LLM resamples until a valid label-stable sample with moderate difficulty is obtained or the maximum attempts are reached. Finally, we obtain the augmented sample set $\mathcal{A}'$ and the original-augmented hard sample pair set $\mathcal{H}_A$. After noise filtering and difficulty-adaptive augmentation, the final supervised training set is constructed as:

$$\mathcal{D}_{train} = (\mathcal{D} \setminus \mathcal{C}_{hr}) \cup \mathcal{A} \cup \mathcal{A}'. \quad (16)$$

For pair-wise alignment, we further construct the original-augmented hard sample pair set:

$$\mathcal{H}_A = \{(x_i, \hat{x}_i) \mid x_i \in \mathcal{A}, \hat{x}_i \in \mathcal{A}'\}, \quad (17)$$

where each pair consists of an original hard sample x_i and its accepted augmented counterpart $\hat{x}_i$.

3.5 Prototype-Guided Dual-Granularity Contrastive Learning

Simply treating augmented samples as ordinary augmented data may fail to effectively transfer their learnable semantics back to the original hard boundary samples. Therefore, we introduce prototype-guided dual-granularity contrastive learning to optimize the representation space from both local and global perspectives. Given an encoder g_θ , the representation of sample x_i is denoted as:

$$\mathbf{z}_i = g_\theta(x_i). \quad (18)$$

At the local level, we align each original hard boundary sample with its corresponding augmented sample:

$$\mathcal{L}_{pair} = \frac{1}{|\mathcal{B}_p|} \sum_{(x_i, x_i^b) \in \mathcal{B}_p} \max(0, n - \text{sim}(\mathbf{z}_i, \mathbf{z}_i^b)), \quad (19)$$

where $\mathcal{B}_p$ denotes the pair batch, and n is the similarity margin. The margin mechanism imposes a constraint only when the similarity between the original sentence and its augmented version is insufficient, thus avoiding excessive compression of the representation space.

Furthermore, to avoid restricting the objective to local sample pairs, we introduce class-descriptive prototypes as global semantic anchors. Unlike cluster centers, which may be biased by noisy labels and frequent easy samples, class-descriptive prototypes provide more stable and interpretable label-level guidance. Specifically, we encode the descriptive text of each class c with the same encoder to obtain its prototype representation p_c , and define the prototype-level contrastive loss as follows:

$$\mathcal{L}_{proto} = -\frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{p}_{y_i})/\tau)}{\sum_{c \in \mathcal{Y}} \exp(\text{sim}(\mathbf{z}_i, \mathbf{p}_c)/\tau)}. \quad (20)$$

This loss pulls each sample representation toward its corresponding emotion prototype and pushes it away from the prototypes of other classes, thereby strengthening the global category boundaries among fine-grained negative emotions. The final training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{pair} \mathcal{L}_{pair} + \lambda_{proto} \mathcal{L}_{proto}, \quad (21)$$

where $\mathcal{L}_{ce}$ denotes the standard cross-entropy loss, while λ_{pair} and λ_{proto} control the contributions of

local pair alignment and global prototype alignment, respectively. Through this objective, the model learns from clean samples, transfers learnable semantics from augmented samples to hard boundary samples, and enhances its discrimination ability for complex negative emotion boundaries.

4 Experiments

4.1 Datasets and Baselines

We conduct experiments on three fine-grained negative emotion classification datasets: SMP2020-EWECT-Virus(VIRUS)¹, NLPCC2014(NLPCC)², and India-COVID-X(COVID)³. These datasets cover Chinese and English social media scenarios with different negative emotion label spaces, allowing us to examine cross-language and cross-domain robustness. For all datasets, we keep only single-label samples from the target negative emotion categories. Detailed dataset information is provided in Appendix A.1.

We compare NF-DHSA with representative baselines from multiple perspectives. We first use standard Pre-Trained Language model (PLM)-based classifiers to measure conventional supervised fine-tuning performance. We then include Early Cutting (Yuan et al., 2025), a method that filters Mis-labeled Easy Examples (MEEs), as a robust sample selection baseline against label noise. To evaluate the effect of general augmentation, we compare with EDA (Wei and Zou, 2019), AEDA (Karimi et al., 2021), DiffusionCLS (Chen et al., 2024), and LMTransplant (Wang et al., 2025a), an LLM-based transplant-then-regenerate augmentation paradigm. We also adopt EACL (Yu et al., 2024) to compare emotion-anchor contrastive learning with our prototype-anchored dual-granularity optimization. Finally, we fine-tune Qwen3-4B⁴, Qwen3-1.7B⁵ and Llama3.1-8B-Instruct⁶ as LLM-based classification baselines to compare direct LLM classification with NF-DHSA. Detailed implementation settings are provided in Appendix A.2.

¹<https://smp2020ewect.github.io/>.

²<http://tcci.ccf.org.cn/conference/2014>.

³<https://www.kaggle.com/datasets/surajkum1198/twitterdata>

⁴<https://huggingface.co/Qwen/Qwen3-4B>

⁵<https://huggingface.co/Qwen/Qwen3-1.7B>

⁶<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

4.2 Experimental Setups

We evaluate all methods using macro-F1 and accuracy, and report the average results over three random seeds to minimize the impact of randomness. Experiments without specific annotations default to seed 2025. For the base detector, we use chinese-roberta-wwm-ext⁷ for the NLPCC and VIRUS datasets, and bert-base-uncased⁸ for the COVID dataset. In the difficulty-aware high-risk sample mining stage, we select the top 20% hardest samples as candidate pools. To evaluate whether our framework is robust to different LLM architectures, we further replace GLM-5 with Qwen3.5-Plus and DeepSeek-V3.2 in the LLM-based modules. Additional implementation details are provided in Appendix A.3.

4.3 Results and Analysis

Main Classification Results Table 1 presents the main results on the three fine-grained negative emotion classification datasets. Compared to the standard PLM baseline, NF-DHSA achieves absolute Macro-F1 improvements of 1.25, 2.47, and 3.53 points on the VIRUS, NLPCC, and COVID datasets, respectively.

Against data augmentation baselines, NF-DHSA consistently exhibits superior overall performance. Although LMTransplant achieves a slightly higher Accuracy on the NLPCC dataset, NF-DHSA still obtains the highest Macro-F1 on the same dataset and significantly outperforms all augmentation methods on both VIRUS and COVID. This suggests that while advanced generative augmentation methods like LMTransplant can improve context diversity, they remain insufficient in resolving the inherent learnability bottlenecks of hard boundary samples in fine-grained negative emotion classification.

Furthermore, NF-DHSA remains highly competitive against recent sample selection and representation learning baselines. Although EACL achieves the highest Macro-F1 on the VIRUS dataset, NF-DHSA secures the highest Accuracy on VIRUS and obtains the best overall performance on both NLPCC and COVID. Compared to the noise filtering baseline MEEs, NF-DHSA demonstrates a consistent advantage across all three datasets. This proves that explicitly separating noisy samples

⁷<https://huggingface.co/hfl/chinese-roberta-wwm-ext>

⁸<https://huggingface.co/google-bert/bert-base-uncased>

Method	VIRUS			NLPCC			COVID		
	Macro-F1	Acc	Δ F1	Macro-F1	Acc	Δ F1	Macro-F1	Acc	Δ F1
PLM	77.91 $\pm$ 0.46	80.69 $\pm$ 0.29	-	69.03 $\pm$ 0.91	69.14 $\pm$ 0.22	-	71.11 $\pm$ 0.23	70.96 $\pm$ 0.24	-
EDA	77.85 $\pm$ 0.81	81.15 $\pm$ 0.63	-0.06	67.91 $\pm$ 1.12	68.42 $\pm$ 0.91	-1.12	71.78 $\pm$ 1.27	71.71 $\pm$ 1.44	+0.67
AEDA	78.16 $\pm$ 1.03	81.04 $\pm$ 0.83	+0.25	68.39 $\pm$ 0.61	69.04 $\pm$ 0.50	-0.64	73.73 $\pm$ 0.72	73.60 $\pm$ 0.67	+2.62
Diffusion-CLS	77.55 $\pm$ 0.55	80.58 $\pm$ 0.34	-0.36	67.54 $\pm$ 0.83	67.05 $\pm$ 1.11	-1.49	71.60 $\pm$ 0.81	71.48 $\pm$ 0.81	+0.49
EACL	80.57$\pm$0.75	80.58 $\pm$ 0.73	+2.66	67.73 $\pm$ 0.57	67.35 $\pm$ 0.68	-1.30	65.73 $\pm$ 3.43	66.01 $\pm$ 3.27	-5.38
LMTransplant	75.76 $\pm$ 1.44	79.17 $\pm$ 1.04	-2.15	70.53 $\pm$ 0.50	72.18$\pm$0.78	+1.50	68.19 $\pm$ 2.73	68.46 $\pm$ 2.84	-2.92
MEEs	76.91 $\pm$ 2.42	80.50 $\pm$ 1.50	-1.00	70.96 $\pm$ 1.20	71.72 $\pm$ 1.52	+1.93	73.71 $\pm$ 2.17	73.65 $\pm$ 2.21	+2.60
NF-DHSA	79.16 $\pm$ 1.56	82.34$\pm$1.39	+1.25	71.50$\pm$1.11	71.57 $\pm$ 1.67	+2.47	74.64$\pm$1.03	74.69$\pm$1.27	+3.53

Table 1: Main results on three fine-grained negative emotion classification datasets. Performance is reported as mean $\pm$ standard deviation across 3 random seeds. Δ F1 denotes the absolute improvement in Macro-F1 score compared to the PLM baseline. Red values indicate the best result in each column.

Method	Symmetric 20%		Symmetric 40%		Asymmetric 20%		Asymmetric 40%	
	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc
PLM	66.97 $\pm$ 1.91	66.85 $\pm$ 2.09	61.28 $\pm$ 1.19	61.47 $\pm$ 1.54	66.24 $\pm$ 2.06	66.24 $\pm$ 2.07	54.52 $\pm$ 5.13	55.71 $\pm$ 3.65
Co-Teaching	69.59 $\pm$ 0.56	69.73 $\pm$ 0.67	65.33 $\pm$ 0.66	65.34 $\pm$ 0.41	66.38 $\pm$ 0.82	65.77 $\pm$ 0.81	55.65 $\pm$ 1.93	56.75 $\pm$ 1.13
SCE	67.89 $\pm$ 0.33	68.04 $\pm$ 0.67	59.36 $\pm$ 3.06	60.39 $\pm$ 2.04	65.28 $\pm$ 1.32	65.06 $\pm$ 1.66	53.80 $\pm$ 0.96	53.92 $\pm$ 0.59
ELR	68.60 $\pm$ 0.84	68.93 $\pm$ 0.46	61.76 $\pm$ 1.65	62.42 $\pm$ 1.53	65.26 $\pm$ 0.47	64.73 $\pm$ 0.57	49.32 $\pm$ 7.89	52.31 $\pm$ 3.73
SelfMix	64.66 $\pm$ 3.45	65.44 $\pm$ 3.35	62.76 $\pm$ 3.40	63.83 $\pm$ 2.21	63.39 $\pm$ 1.75	63.50 $\pm$ 1.93	45.95 $\pm$ 4.71	51.89 $\pm$ 2.30
DyGen	73.05 $\pm$ 1.10	73.28 $\pm$ 1.04	61.74 $\pm$ 2.96	62.93 $\pm$ 2.29	71.46 $\pm$ 1.28	71.62 $\pm$ 1.43	60.63 $\pm$ 6.32	61.90 $\pm$ 4.63
MEEs	69.50 $\pm$ 2.24	69.83 $\pm$ 2.09	61.32 $\pm$ 2.03	61.71 $\pm$ 2.02	66.59 $\pm$ 0.27	66.43 $\pm$ 0.57	58.21 $\pm$ 5.02	58.78 $\pm$ 4.50
NF-DHSA	73.60$\pm$0.65	73.61$\pm$0.43	73.31$\pm$1.00	73.18$\pm$1.08	73.78$\pm$0.26	73.70$\pm$0.22	73.10$\pm$0.35	72.95$\pm$0.43

Table 2: Performance comparison with baseline denoising methods under different synthetic noise types (Symmetric and Asymmetric) and noise ratios. Performance is reported as mean $\pm$ standard deviation across 3 random seeds. Red values indicate the best result in each column.

from valid hard samples, and utilizing difficulty-adaptive augmentation to convert hard samples into a more learnable form, are more comprehensive and effective strategy than relying solely on early-stage sample selection.

Robustness Against Synthetic Label Noise To evaluate robustness against controlled label corruption, we introduce symmetric and asymmetric label noise at ratios of 20 and 40, with comparisons to six competitive noisy-label learning baselines: Co-Teaching (Han et al., 2018), SCE (Wang et al., 2019), ELR (Liu et al., 2020), SelfMix (Qiao et al., 2022), DyGen (Zhuang et al., 2023), and MEEs (Yuan et al., 2025). As shown in Table 2, NF-DHSA achieves state-of-the-art results across all four noise configurations. NF-DHSA’s superiority grows substantially with higher noise levels. Specifically, it surpasses the strongest baseline by 7.98 and 12.47 Macro-F1 points under 40 symmetric and 40 asymmetric noise, respectively. Existing noisy-label methods mostly rely on robust losses, early regularization or sample selection, which often treat informative hard samples and mislabeled samples indiscriminately. By contrast, our approach identifies noisy samples via

multi-view evidence, preserves valuable hard samples, and boosts their learnability with difficulty-adaptive augmentation, yielding superior stability under high and class-dependent label noise.

Finally, Figure 3 compares NF-DHSA against direct LLM classification (GLM-5) and supervised fine-tuned (SFT) models, including Qwen3-1.7B, Qwen3-4B, and Llama3.1-8B-Instruct. Across the COVID, NLPCC, and VIRUS datasets, NF-DHSA consistently achieves the highest Macro-F1 and Accuracy. And the standard fine-tuning process is susceptible to noisy labels in the training set, leading to worse classification results. This demonstrates that merely increasing model capacity or conducting conventional fine-tuning is insufficient to meet the challenges of fine-grained negative emotion classification, and explicitly isolating noisy samples and employing difficulty-adaptive augmentation are essential for achieving optimal performance.

4.4 Ablation Study

We conduct ablation studies to evaluate the overall contribution of the main components in NF-DHSA, which is shown in Table 3. Removing multi-view noise filtering or difficulty-adaptive hard sample

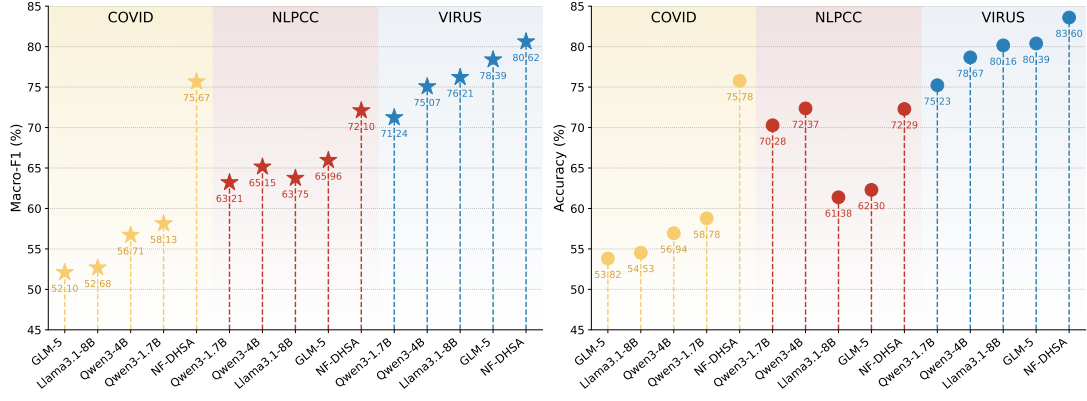


Figure 3: Performance comparison between large language model direct classification (GLM-5) and supervised fine-tuned (SFT) models across three benchmark datasets.

Method	VIRUS		NLPCC		COVID	
	F1	Acc	F1	Acc	F1	Acc
w/o Pair Align.	79.52	82.37	70.40	70.67	73.81	73.94
w/o Global Align.	79.93	82.91	70.69	70.98	72.55	72.66
w/o Multi-view Filter.	76.90	79.59	70.28	70.90	72.96	73.09
w/o Dynamic Aug.	77.99	81.31	69.53	69.66	72.79	72.66
PLM	77.54	80.39	68.99	69.27	70.85	70.68
+ Multi-view Filter.	78.07	81.54	70.78	70.82	72.72	72.80
+ Dynamic Aug.	79.75	82.91	71.70	71.05	74.57	74.65
NF-DHSA	80.62	83.60	72.10	72.29	75.67	75.78

Table 3: Ablation results on three fine-grained negative emotion classification datasets. ‘w/o’ is the abbreviation for without. The ‘+’ symbol indicates adding a new module to the preceding configuration, and ‘NF-DHSA’ equivalently represents the final addition of Dual-Granularity Contrastive Learning.

augmentation results in more noticeable performance drops. Besides, We added a step-by-step experiment in which each supervision source is introduced sequentially under the same backbone and training settings. Specifically, we first add LLM filtering to the PLM baseline, then add LLM augmentation, and finally introduce contrastive training in the full NF-DHSA model. The results show that all three components provide consistent gains across the three datasets.

Overall, the ablation results show that the effectiveness of NF-DHSA does not come from a single module, but from the collaboration among noise filtering, hard sample augmentation, and representation alignment. Noise filtering reduces the interference of mislabeled samples, hard sample augmentation provides more learnable supervision for valid hard samples, and representation alignment further improves the feature space.

Dataset	Pos-Class	Precision	Recall	F1-score	Support
COVID	hard	80.00	83.33	81.63	24
COVID	noise	84.00	80.77	82.35	26
NLPCC	hard	84.00	87.50	85.71	24
NLPCC	noise	88.00	84.62	86.27	26
VIRUS	hard	88.00	84.62	86.27	24
VIRUS	noise	84.00	87.50	85.71	26

Table 4: Human verification results for hard-noise distinction. Pos-Class indicates the positive class.

4.5 Discussions and Visualizations

Human Verification of Hard-Noise Distinction.

To further examine whether NF-DHSA can distinguish noisy samples from valid hard samples, we conduct a human verification study on the outputs of the multi-view logic-guided noise filtering module. For each dataset, we randomly sample 25 predicted *hard* and 25 predicted *noisy* instances, yielding 150 samples in total. Two researchers with sentiment-analysis experience independently annotate each sample as *hard* or *noise*, achieving a Cohen’s Kappa of 0.7735 and an agreement rate of 87.92%. Disagreements are resolved through discussion to obtain the final reference labels.

Based on these human labels, we evaluate the predictions produced by NF-DHSA. As shown in Table 4 and Figure 4, NF-DHSA achieves stable performance on both *hard* and *noise* samples across the three datasets. The confusion matrices show consistently dominant diagonal entries, indicating that most samples are correctly aligned with human judgments. This demonstrates that NF-DHSA does not simply discard high-difficulty instances as noise, but can effectively preserve informative hard samples while filtering truly noisy ones. The remaining errors mainly arise from implicit, mixed,

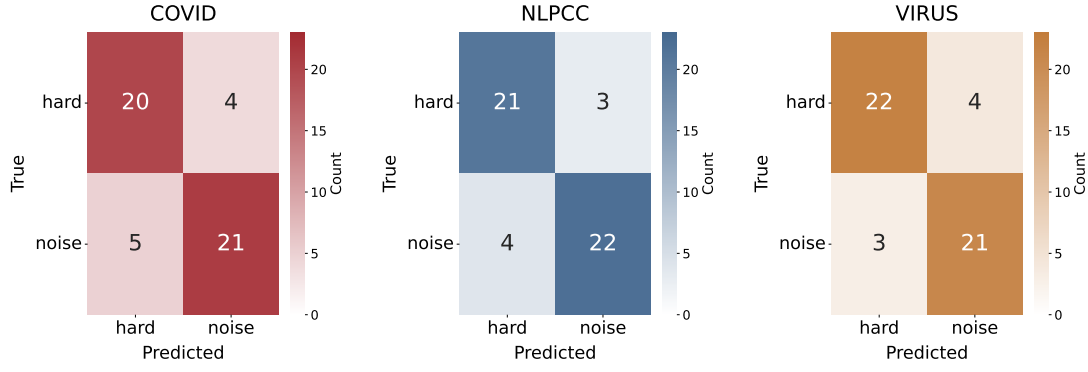


Figure 4: Confusion matrices of hard–noise distinguish on the human-verified samples.

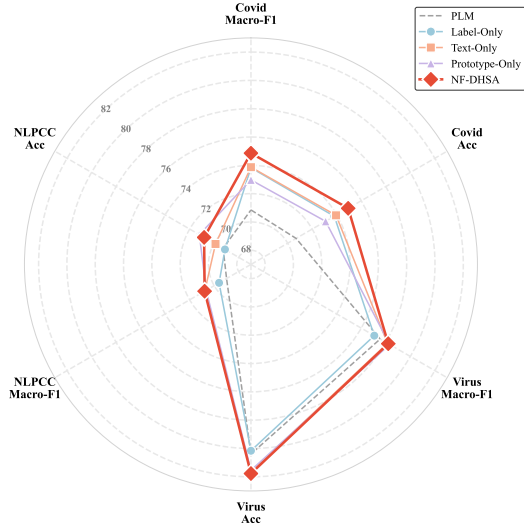


Figure 5: Performance comparison between Single-View and Multi-View noise filtering on three datasets.

or context-dependent expressions, which are ambiguous even for human annotators.

Effect of Multi-view Noise Filtering. To isolate the effect of multi-view noise verification, we compared Label-Definition-Only, Text-Only, Prototype-Only, and the full multi-view verification methods across the three datasets. During this process, the subsequent augmentation and contrastive learning modules were removed to avoid interference from other components. As shown in the Figure 5, the baseline PLM consistently forms the innermost polygon, reflecting its vulnerability across all metrics. Although single-view methods improve upon the baseline, their performance tends to fluctuate depending on the dataset. For instance, the Prototype-Only method achieves the highest Macro-F1 (78.31) on VIRUS and the highest Accuracy (71.21) on NLPCC, but its performance drops noticeably on COVID. This indicates that

a single LLM-based judgment is still unstable in fine-grained negative emotion classification, where noisy samples and valid hard boundary samples can be easily confused. In contrast, the proposed NF-DHSA framework integrates these complementary evidence sources, forming the most encompassing and stable outer boundary. This geometric stability confirms that synthesizing label definitions, text semantics, and class prototypes effectively mitigates the inherent biases of any single view, thereby ensuring consistent and robust classification capability across diverse domains. Other details of the experiment sensitivity studies are provided in Appendix B.

5 Conclusion

In this paper, we propose NF-DHSA, a noise filtering and difficulty-adaptive hard sample augmentation framework for fine-grained negative emotion classification. NF-DHSA first separates noisy samples from valid hard samples through difficulty-aware mining and multi-view logic-guided filtering. It then generates more learnable augmented samples for verified hard samples and transfers their useful semantics back through prototype-guided dual-granularity contrastive learning. Experiments on three fine-grained negative emotion classification datasets demonstrate that NF-DHSA achieves robust performance across different domains, and label spaces. Further ablation studies and sensitivity analyses confirm the effectiveness of the proposed hard-noise distinguish and hard sample learning mechanisms. Overall, NF-DHSA provides a practical framework for reducing noisy-label interference while preserving and exploiting informative hard samples in fine-grained negative emotion classification.

Limitations

Although NF-DHSA achieves promising results, it still has several limitations. First, the current framework focuses on samples selected by the difficulty-aware high-risk mining module. This design improves efficiency by applying LLM-based verification only to the most suspicious candidates, but it may overlook noisy samples outside the high-risk pool. Second, NF-DHSA relies on LLMs for semantic verification and augmented sample generation. Although our experiments show that the framework is stable across different strong LLMs, the quality of verification and augmentation may still be affected by the instruction-following ability, semantic reasoning quality, and generation stability of the selected LLM.

Ethics Statement

All experiments are conducted on existing benchmark datasets released for research purposes, and we do not collect any new private user data. We only use the textual content and emotion labels provided by the datasets, and do not attempt to infer user identities, demographic attributes, or other sensitive personal information. Since negative emotion classification may involve subjective and sensitive emotional expressions, we emphasize that the proposed method is intended for dataset-level research and model robustness analysis, rather than individual-level psychological diagnosis, surveillance, or decision making. As for the AI assistant, we utilize ChatGPT and Claude to identify textual errors and polish paper.

Acknowledgments

This work was supported by the National Social Science Fund of China (No. 22BTQ045).

References

- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Irina Bejan, Artem Sokolov, and Katja Filippova. 2023. [Make every example count: On the stability and utility of self-influence for learning from noisy NLP datasets](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10107–10121, Singapore. Association for Computational Linguistics.
- Zhuowei Chen, Lianxi Wang, Yuben Wu, Xinfeng Liao, Yujia Tian, and Junyang Zhong. 2024. [An effective deployment of diffusion LM for data augmentation in low-resource sentiment classification](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1844–1856, Miami, Florida, USA. Association for Computational Linguistics.
- Derek Chong, Jenny Hong, and Christopher Manning. 2022. [Detecting label errors by using pre-trained language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9074–9091, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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.
- Aman Goel, Daniel Schwartz, and Yanjun Qi. 2025. [Zero-knowledge LLM hallucination detection and mitigation through fine-grained cross-model consistency](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1982–1999, Suzhou, China. Association for Computational Linguistics.
- Demi Guo, Yoon Kim, and Alexander Rush. 2020. [Sequence-level mixed sample data augmentation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5547–5552, Online. Association for Computational Linguistics.
- Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. [Co-teaching: Robust training of deep neural networks with extremely noisy labels](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. [Supervised adversarial contrastive learning for emotion recognition in conversations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.
- Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. [AEDA: An easier data augmentation technique for](#)

- text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. [MEGAnno+: A human-LLM collaborative annotation system](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176, St. Julians, Malta. Association for Computational Linguistics.
- Xinfeng Liao, Xuanqi Chen, Lianxi Wang, Ziyang Rong, Jiahuan Yang, Jie Li, and Zhuowei Chen. 2026. Feature-level enhanced syntactic-semantic graph networks via optimal transport for aspect-based sentiment analysis. *Knowledge and Information Systems*, 68(1):212.
- Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. 2020. [Early-learning regularization prevents memorization of noisy labels](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 20331–20342. Curran Associates, Inc.
- Elena Merdjanovska, Ansar Aynedinov, and Alan Akbik. 2024. [NoiseBench: Benchmarking the impact of real label noise on named entity recognition](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18182–18198, Miami, Florida, USA. Association for Computational Linguistics.
- Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szepktor, and Roi Reichart. 2025. [Are LLMs better than reported? detecting label errors and mitigating their effect on model performance](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26782–26809, Suzhou, China. Association for Computational Linguistics.
- Dan Qiao, Chenchen Dai, Yuyang Ding, Juntao Li, Qiang Chen, Wenliang Chen, and Min Zhang. 2022. [SelfMix: Robust learning against textual label noise with self-mixup training](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 960–970, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Melissa Kazemi Rad, Alberto Purpura, Himanshu Kumar, Emily Chen, and Mohammad Shahed Sorower. 2025. [GRAID: Synthetic data generation with geometric constraints and multi-agentic reflection for harmful content detection](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30059–30077, Suzhou, China. Association for Computational Linguistics.
- Hanna Shcharbakova, Tatiana Anikina, Natalia Skachkova, and Josef van Genabith. 2025. [When](#)
- scale meets diversity: Evaluating language models on fine-grained multilingual claim verification. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*, pages 69–84, Vienna, Austria. Association for Computational Linguistics.
- Tianrui Song, Wen-Shuo Chao, and Hao Liu. 2026. [Hard vs. noise: Resolving hard-noisy sample confusion in recommender systems via large language models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(18):15743–15751.
- Tiberiu Sosea and Cornelia Caragea. 2026. [LANE: Label-aware noise elimination for fine-grained text classification](#). In *The Fourteenth International Conference on Learning Representations*.
- J. Fernando Sánchez-Rada and Carlos A. Iglesias. 2019. [Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison](#). *Information Fusion*, 52:344–356.
- Guangzhan Wang, Hongyu Zhang, Beijun Shen, and Xiaodong Gu. 2025a. [Transplant then regenerate: A new paradigm for text data augmentation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13906–13920, Suzhou, China. Association for Computational Linguistics.
- Haonan Wang, Minbin Huang, Runhui Huang, Lanqing Hong, Hang Xu, Tianyang Hu, Xiaodan Liang, Zhen-guo Li, Hong Cheng, and Kenji Kawaguchi. 2025b. [Getting more juice out of your data: Hard pair refinement enhances visual-language models without extra data](#). 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 7854–7873, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. [Symmetric cross entropy for robust learning with noisy labels](#). In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 322–330.
- Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.
- Tingting Wu, Xiao Ding, Minji Tang, Hao Zhang, Bing Qin, and Ting Liu. 2022. [STGN: an implicit regularization method for learning with noisy labels in natural language processing](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7587–7598, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Sishi Xiong, Yu Zhao, Jie Zhang, Li Mengxiang, Zhongjiang He, Xuelong Li, and Shuangyong Song. 2024. [Dual prompt tuning based contrastive learning for hierarchical text classification](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12146–12158, Bangkok, Thailand. Association for Computational Linguistics.
- Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. [Emotion-anchored contrastive learning framework for emotion recognition in conversation](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4521–4534, Mexico City, Mexico. Association for Computational Linguistics.
- Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. 2024. [Hide and seek in noise labels: Noise-robust collaborative active learning with LLMs-powered assistance](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10977–11011, Bangkok, Thailand. Association for Computational Linguistics.
- Suqin Yuan, Lei Feng, Bo Han, and Tongliang Liu. 2025. [Enhancing sample selection against label noise by cutting mislabeled easy examples](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 43942–43970. Curran Associates, Inc.
- Pinyi Zhang, Jingyang Chen, Junchen Shen, Zijie Zhai, Ping Li, Jie Zhang, and Kai Zhang. 2024. [Message passing on semantic-anchor-graphs for fine-grained emotion representation learning and classification](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2771–2783, Miami, Florida, USA. Association for Computational Linguistics.
- Yunyi Zhang, Minhao Jiang, Yu Meng, Yu Zhang, and Jiawei Han. 2023. [PIEClass: Weakly-supervised text classification with prompting and noise-robust iterative ensemble training](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12655–12670, Singapore. Association for Computational Linguistics.
- Wenxuan Zhou and Muhao Chen. 2021. [Learning from noisy labels for entity-centric information extraction](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5381–5392, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yuchen Zhuang, Yue Yu, Ling kai Kong, Xiang Chen, and Chao Zhang. 2023. [Dygen: Learning from noisy labels via dynamics-enhanced generative modeling](#). In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3674–3686.

A Dataset and Baseline Details

A.1 Dataset Details

To evaluate the robustness of our framework, we conduct experiments on three fine-grained negative emotion classification datasets: SMP2020-EWECT-Virus (VIRUS), NLPCC2014 (NLPCC), and India-COVID-X (COVID).

SMP2020-EWECT-Virus (VIRUS). SMP2020-EWECT-Virus is the pandemic-related subset of SMP2020-EWECT, consisting of Chinese Weibo posts collected during the COVID-19 outbreak. The original dataset contains multiple emotion categories, such as *neutral*, *happy*, *angry*, *sad*, *fear*, and *surprise*. In this work, we focus on fine-grained negative emotion classification and retain three negative categories: *sad*, *fear*, and *angry*. Since the texts are written in the context of a public health emergency, they contain many event-specific expressions, implicit emotional cues, informal social media language, and potentially noisy annotations. Therefore, this dataset is suitable for evaluating hard-noise distinction in Chinese pandemic-related social media scenarios.

India-COVID-X (COVID). India-COVID-X is an English social media dataset collected from Twitter/X during the COVID-19 pandemic. The texts mainly discuss topics such as coronavirus, COVID-19, and lockdown in the Indian context. We retain three negative emotion categories: *sad*, *fear*, and *anger*. Compared with VIRUS, this dataset shares the pandemic-related domain but differs in language, platform, and cultural context. It is therefore used to evaluate whether the proposed framework can generalize to English social media texts and cross-lingual pandemic-related emotion classification.

NLPCC2014 (NLPCC). NLPCC2014 is a Chinese social media emotion classification dataset containing short user-generated texts from general online scenarios. Different from the two COVID-19-related datasets, NLPCC covers broader topics such as daily life, social events, and personal opinions. Following the fine-grained negative emotion setting, we retain four negative categories: *sadness*, *fear*, *anger*, and *disgust*. This dataset has a more complex negative emotion label space, where semantically close categories such as *anger* and *disgust*, or *sadness* and *fear*, often exhibit ambiguous boundaries. Thus, it is suitable for examining

the model’s ability to distinguish subtle negative emotional states in general Chinese social media texts. Table 5 reports the statistics of the original processed datasets.

Dataset	Train	Test	Labels	Language	Domain
VIRUS	2526	872	3	Chinese	Virus
COVID	1657	706	3	English	Virus
NLPCC	7806	1292	4	Chinese	General

Table 5: Statistics of the original processed datasets.

For training, we further construct different training sets according to the difficulty-aware candidate mining ratio. Specifically, we consider three difficulty ratios, i.e., *Top10%*, *Top15%*, and *Top20%*. For each ratio, the training set consists of simple samples, retained hard samples after noise filtering, and generated augmented samples. The original-augmented pairs are used for pair-wise alignment. The detailed statistics for each dataset are shown in Tables 6–8.

It should be noted that the number of retained hard samples is not necessarily equal to the number of finally accepted augmented samples. In NF-DHSA, a hard sample is first identified as semantically valid and potentially valuable for learning, but this does not imply that every hard sample can be successfully transformed into a qualified augmented sample. For each valid hard sample, we only accept candidate augmented samples that simultaneously satisfy semantic consistency, label consistency, length-ratio constraints, non-duplication constraints, and the sample-level dynamic difficulty window. If no candidate satisfies all these constraints within the maximum number of generation attempts, the corresponding sample is not used to construct an original-augmented pair.

This design reflects a conservative quality-control strategy. On the one hand, hard samples often contain implicit emotions, mixed emotional cues, event-dependent contexts, or ambiguous category boundaries, making it inherently difficult to rewrite them into augmented samples that are easier to learn while still preserving the original label semantics. On the other hand, forcing every hard sample to have an augmented counterpart may introduce low-quality generations with oversimplification, semantic drift, or label shift, thereby weakening the reliability of augmented learning. Therefore, we discard augmented candidates that fail to satisfy the constraints rather than introduc-

ing potentially low-quality augmented samples into pair-wise alignment.

In our experiments, the number of successfully augmented samples accounts for approximately 85% of the valid hard samples. Further manual inspection shows 96% failed cases are samples for which it is difficult to simultaneously satisfy the dynamic difficulty constraint and the semantic consistency constraint. Therefore, the mismatch between the number of hard samples and augmented samples does not indicate that hard samples are discarded. Instead, it is a consequence of our strict augmentation filtering mechanism. For hard samples without valid augmented counterparts, we still retain them in standard supervised training, while only augmented samples that pass all quality constraints are used for original–augmented alignment. This strategy enables NF-DHSA to exploit the learning value of hard samples while reducing the negative impact of low-quality augmented samples on model training.

Difficulty Ratio	Simple	Hard	Augmented	Total	Ori-Aug Pairs
10%	2274	106	92	2472	92
15%	2148	186	158	2492	158
20%	2020	275	221	2516	221

Table 6: Training data statistics of SMP2020-EWECT-Virus under different difficulty ratios.

Difficulty Ratio	Simple	Hard	Augmented	Total	Ori-Aug Pairs
10%	1491	59	50	1601	50
15%	1408	107	82	1598	82
20%	1326	153	103	1582	103

Table 7: Training data statistics of India-COVID-X under different difficulty ratios.

Difficulty Ratio	Simple	Hard	Augmented	Total	Ori-Aug Pairs
10%	7025	144	125	7294	125
15%	6635	308	279	7222	279
20%	6245	537	490	7272	490

Table 8: Training data statistics of NLPCC2014 under different difficulty ratios.

A.2 Baseline Details

We provide details about the backbone models and baseline methods used in our experiments.

PLM. For all PLM-based methods, we use chinese-roberta-wwm-ext as the backbone for the Chinese datasets and bert-base-uncased for the English dataset. We adopt standard PLM-based classifiers as basic supervised fine-tuning baselines.

The classifier is trained with the standard cross-entropy objective on the original training set. This baseline is used to evaluate the basic performance of conventional PLM fine-tuning in fine-grained negative emotion classification.

EDA (Wei and Zou, 2019). EDA is a rule-based textual data augmentation method that generates augmented samples through simple word-level operations, including synonym replacement, random insertion, random swap, and random deletion. It is used to examine whether lightweight lexical augmentation can improve hard sample learning.

AEDA (Karimi et al., 2021). AEDA augments text by randomly inserting punctuation marks into the original sentence. Compared with EDA, it avoids modifying original words and provides a simpler perturbation strategy. We use it as another rule-based augmentation baseline.

DiffusionCLS (Chen et al., 2024). DiffusionCLS is a diffusion language model based data augmentation method for sentiment classification. It generates pseudo samples by reconstructing strong label-related tokens and aims to balance label consistency and sample diversity. We include it to compare NF-DHSA with a recent model-based augmentation method and to examine whether general-purpose augmentation can effectively improve fine-grained negative emotion classification.

EACL (Yu et al., 2024). EACL is an emotion-anchored contrastive learning framework originally proposed for emotion recognition in conversation. It uses textual emotion labels as semantic anchors to guide representation learning and improve the separability of similar emotions. We adopt EACL to compare emotion-anchor contrastive learning with the prototype-anchored dual-granularity optimization in NF-DHSA.

LMTransplant (Wang et al., 2025a). LMTransplant is a text data augmentation paradigm that leverages large language models (LLMs) to generate high-quality variants. It employs a "transplant-then-regenerate" approach, embedding the original seed text into an LLM-expanded context and prompting the model to regenerate a new variant. We include this to compare NF-DHSA with a generative augmentation strategy

MEEs (Yuan et al., 2025). A robust sample selection strategy designed to identify and filter out Misabeled Easy Examples (MEEs)—misabeled samples that a model correctly predicts early in training and that disproportionately harm generalization. It utilizes the model’s later training state

Category	Hyperparameter	Setting
LLM Generation	Initial temperature	0.70
	Top-p	0.95
Retry Policy	Maximum API retries	3
	Maximum generation attempts per sample	25
	Length ratio	[0.30, 2.00]
Acceptance Criteria	Difficulty interval	$0.1D_{\text{orig}} < D_{\text{aug}} < 0.9D_{\text{orig}}$
	Minimum true-label probability	0.55
	Minimum Semantic similarity	0.5
	Learning rate	$1e-5/2e-5$
Classifier Training	Optimizer	AdamW
	Dropout	0.5

Table 10: Hyperparameter settings.

to recalibrate early confident subsets, avoiding misleading confidence. We include it to compare our noise filtering mechanism with a representative dynamic-based sample selection baseline.

LLM-based classifiers. We further fine-tune Qwen3-4B, Qwen3-1.7B and Llama3.1-8B-Instruct as LLM-based classification baselines. Specifically, we perform LoRA-based supervised fine-tuning, and use the fine-tuned LLMs to directly predict emotion labels. These baselines are used to evaluate whether directly relying on the classification capability of LLMs is sufficient, compared with the explicit hard-noise distinction and difficulty-adaptive hard sample augmentation in NF-DHSA.

A.3 Implementation Details

We ran our experiments on an NVIDIA RTX 3090 GPU. The software environment is Python 3.11 with CUDA 13. All LLM experiments are implemented using LlamaFactory with Low-Rank Adaptation (LoRA) for efficient fine-tuning. The detailed hyperparameter settings are summarized in Table 9. The maximum sequence length is set to 1024 with sequence packing enabled. We use bf16 precision throughout training, and the number of training epochs is set to 15 for the COVID and VIRUS datasets, and 8 for the NLPCC dataset.

The detailed generation and training configuration parameters are shown in Table 10, including LLM sampling parameters, retry counts and generation limits, acceptance criteria for augmented samples, and classifier hyperparameters, etc.

Hyperparameter	Qwen3-1.7B	Qwen3-4B	Llama3.1-8B-Instruct
LoRA rank	64	64	64
LoRA alpha	128	128	128
LoRA dropout	0.05	0.05	0.05
Batch size (per device)	2	2	2
Gradient accumulation	4	4	4
Learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}

Table 9: Hyperparameter settings for the LLM LoRA SFT experiments.

B Experiment Sensitivity Study

B.1 Start Epoch for Local Alignment S

We further analyze the effect of start epoch for local alignment S, which determines when the original-augmented sample alignment loss is activated during training. Different from the pair-wise alignment weight that controls the strength of consistency regularization, start epoch for local alignment S controls the temporal scheduling of such regularization. Activating pair-wise alignment too early may force the model to align original hard samples with augmented samples before the basic class boundaries are sufficiently learned, which may suppress useful semantic diversity introduced by augmentation. In contrast, delaying the alignment allows the classifier to first learn basic decision boundaries from standard supervision, and then use the pair-wise constraint to refine the consistency between original and augmented samples.

As shown in Figure 6, the optimal start epoch varies across datasets, indicating that the alignment schedule should be adapted to dataset characteristics. For Virus, the best performance is obtained when start epoch for local alignment S is set to 4, and several later settings yield close results. This suggests that Virus benefits from a short warm-up before applying pair-wise alignment. Since this dataset contains many event-specific contexts, repost structures, and implicit emotional expressions, introducing alignment after the model has learned initial category cues can help augmented samples support original hard samples more effectively.

For NLPCC, the best Macro-F1 is achieved when start epoch for local alignment S is set to 8, while the best accuracy appears at epoch 4. This discrepancy suggests that the effect of pair-wise alignment is different on overall prediction correctness and class-balanced performance. Since NLPCC contains four negative emotion categories, including semantically close labels such as *anger* and *disgust*, the model first needs sufficient standard supervised training to learn coarse decision boundaries. If pair-wise alignment is introduced too early, the original and augmented samples may be forced together before the fine-grained class structure is sufficiently formed, which can limit the model’s ability to distinguish minority or easily confused categories. A later activation at epoch 8 allows the model to first build a more stable representation space, and then uses the original-augmented pair alignment to refine hard boundary samples. There-

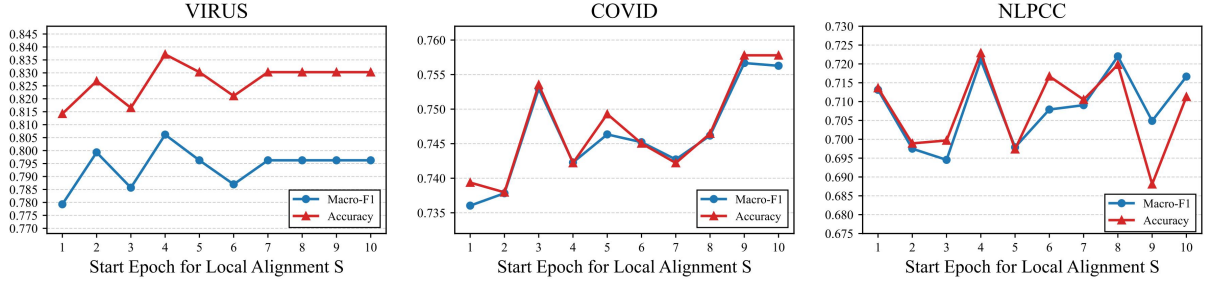


Figure 6: Sensitivity analysis of Start Epoch for Local Alignment S on COVID, VIRUS, and NLPCC.

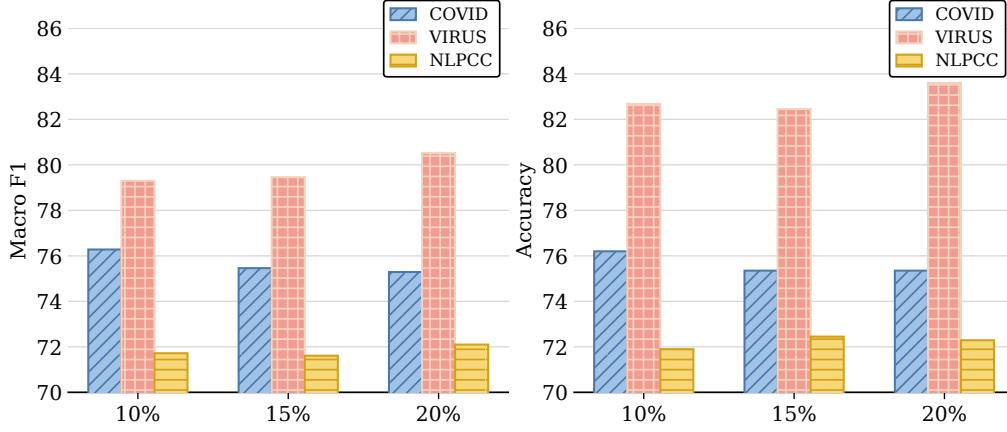


Figure 7: Performance comparison under different difficulty ratios.

fore, although epoch 4 yields slightly better Accuracy, epoch 8 provides better Macro-F1, indicating more balanced improvements across emotion categories.

For COVID, the best performance is achieved when start epoch for local alignment S is set to 9 or 10. This dataset has the smallest training size, making early representations less stable. If pair-wise alignment is imposed too early, the model may overfit unstable augmented pairs before forming reliable decision boundaries. A later activation therefore allows the model to first learn from standard supervision and then use original-augmented pair alignment as a consistency refinement signal.

Overall, start epoch for local alignment S reflects a temporal trade-off between diversity learning and consistency regularization. The results show that an appropriate warm-up before pair-wise alignment is generally necessary, but the optimal start epoch is dataset-dependent. Thus, the alignment schedule should be selected according to supervision scale, expression diversity, and category boundary complexity.

B.2 Sensitivity Analysis of Difficulty Ratio q

We further analyze the influence of the difficulty ratio q , which determines the proportion of samples selected into the high-risk candidate pool before multi-view logic-guided noise filtering. As shown in Figure 7, the best overall performance is achieved when $q = 20\%$. Compared with 10% and 15%, selecting the top 20% difficult samples provides a broader candidate pool, allowing the multi-view noise filtering module to capture more truly noisy samples as well as informative hard boundary samples. This is especially beneficial for VIRUS and NLPCC, where the emotional expressions are more diverse or the label boundaries are more complex. For these two datasets, $q = 20\%$ achieves the best Macro-F1, indicating that a larger high-risk pool improves the recall of hard-noise distinction.

For India-COVID-X, $q = 10\%$ obtains the highest score. This dataset has the smallest training size, and selecting too many high-risk samples may introduce more unstable boundary samples into subsequent verification and augmentation, which slightly weakens performance. However, $q = 20\%$ still achieves the best average performance across

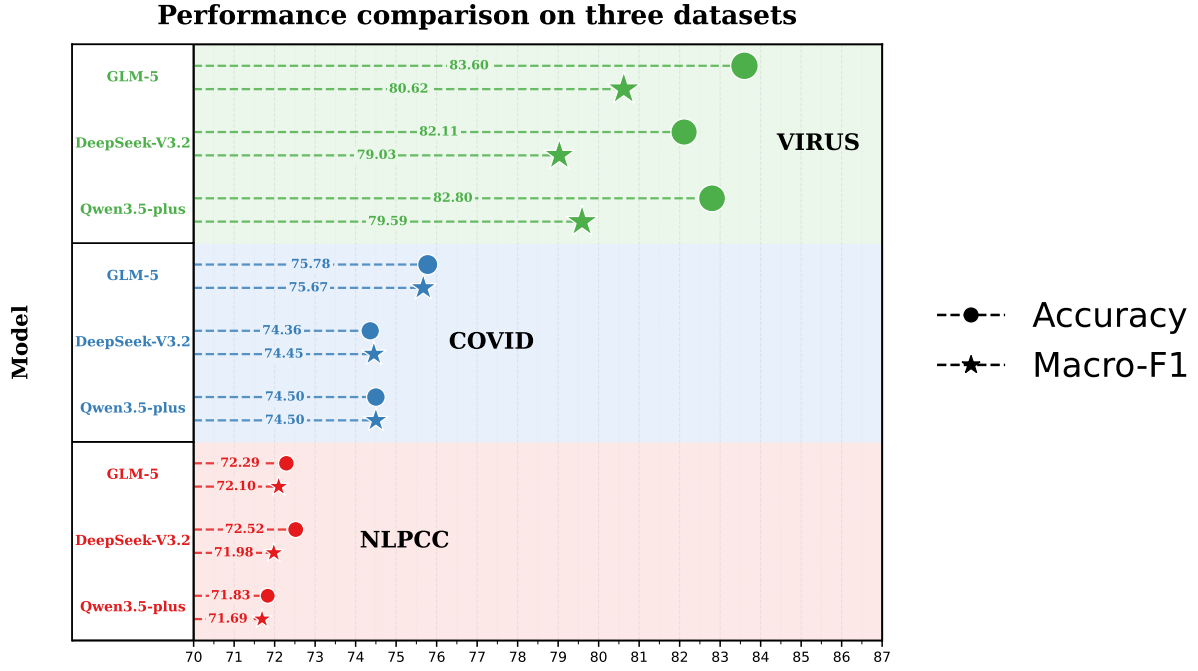


Figure 8: Effect of different LLM choices on NF-DHSA. The results show that NF-DHSA maintains stable performance across GLM-5, DeepSeek-v3.2, and Qwen3.5-plus, indicating that the proposed framework is robust to the choice of LLM.

the three datasets and performs best on two of them. Therefore, we set $q = 20\%$ as the default value in our main experiments, since it provides a better balance between high-risk sample coverage and overall robustness.

B.3 Effect of LLM Choice

As shown in Figure 8, we further analyze the influence of different LLMs on NF-DHSA. Specifically, we employ GLM-5, DeepSeek-v3.2, and Qwen3.5-plus as semantic evaluators in the multi-view logic-guided noise filtering module and the difficulty-adaptive dynamic augmentation module. NF-DHSA maintains stable and competitive performance across the three datasets under different LLM settings. This indicates that the effectiveness of our method does not stem from the incidental outputs of a particular model. Instead, NF-DHSA can transform semantic judgments from different LLMs into stable and effective training signals through structured verification and classifier-feedback constraints.

The outputs of the LLM are further constrained by label definitions, class prototypes, text-level inference, and difficulty feedback. As a result, semantic judgments produced by different LLMs can be integrated into the same verification and filter-

ing framework. This design mitigates the influence of model-specific biases and unstable generations, thereby improving the robustness of the overall pipeline with respect to LLM choice.

From the results, GLM-5 achieves the best overall performance and is therefore adopted as the default LLM in our main experiments. Meanwhile, DeepSeek-v3.2 and Qwen3.5-plus also obtain comparable results, further demonstrating that NF-DHSA can consistently absorb semantic capabilities from different LLMs. These findings suggest that, compared with directly relying on LLMs for end-to-end classification, the proposed method can more effectively leverage the semantic reasoning ability of LLMs and convert it into useful supervision for fine-grained negative emotion classification.

C Multi-view Logic-Guided Noise Filtering Scoring Criteria

Table 11 summarizes the scoring criteria for the three views used in multi-view logic-guided noise filtering. The label-definition view evaluates whether the sample matches the basic meaning of its annotated label, while the prototype view examines its consistency with class profiles and prototype samples in terms of trigger logic, seman-

View	High	Mid	Low
Label-definition view	The text clearly matches the basic definition of its annotated label. For example, <i>sad</i> expresses low mood, loss, helplessness, or disappointment, <i>angry</i> expresses blame, dissatisfaction, sarcasm, or accusation and <i>fear</i> expresses worry, anxiety, risk expectation, or concern about negative consequences.	The text is partially related to the annotated label, but the emotional signal is weak, implicit, mixed with other emotions, or requires additional context to determine. The original label is plausible but not strongly supported by the label definition.	The text is difficult to cover by the basic definition of its annotated label. Its dominant emotion is clearly closer to another class, or the text lacks the emotional meaning required by the annotated label.
Prototype view	The text is consistent with the class profile and prototype samples in terms of emotional trigger logic, semantic stance, and expression scope. Lexical similarity is not required and implicit expressions should still be rated as HIGH if their emotional cause and stance match the annotated class.	The text shares some properties with the class profile or prototypes, but the trigger logic, stance, or expression pattern is incomplete, ambiguous, or overlaps with another negative emotion class. It may be a boundary hard sample.	The text deviates from the annotated class under the scoring criteria, such as emotional trigger logic, semantic stance, or expression scope. It is more consistent with another class profile or falls outside the representative expression range of the annotated class.
Text-only view	Ignoring the original label, the text itself is predicted as the annotated class with relatively high confidence, and no strong mixed-emotion signal is observed. This indicates that the text-level semantics support the original label.	The text has mixed or ambiguous emotional cues. The predicted label may be the annotated class with moderate confidence, or another class appears as a plausible second label. The text-level evidence is insufficient to determine clear noise.	Ignoring the original label, the text is predicted as a different class with high confidence, or its dominant emotion clearly conflicts with the annotated label. This provides label-independent evidence that the sample may be noisy.

Table 11: Scoring criteria for the three views in multi-view logic-guided noise filtering. The label-definition view evaluates whether the sample matches the basic meaning of its annotated label. The prototype view evaluates whether the sample is consistent with class profiles and prototype samples in terms of trigger logic, semantic stance, and expression scope. The text-only view ignores the original label and judges the dominant emotion from the text itself.

tic stance, and expression scope. The text-only view ignores the original label and judges the dominant emotion directly from the text. For each view, HIGH indicates strong consistency, LOW indicates clear semantic-label inconsistency, and MID denotes ambiguous or mixed cases. These complementary criteria help the model distinguish truly noisy samples from valid but difficult hard samples.

D Case Study

Figure 9 and Figure 10 present representative hard and noisy samples from the three datasets. For each sample, we provide multi-view judgments from the label-definition view, prototype/profile view, and text-only view. The hard samples are consistently supported by all three views, even though their emotional cues are implicit, context-dependent, or expressed in a relaxed tone. This indicates that they are semantically valid but difficult-to-learn samples rather than label noise.

By contrast, the noisy samples exhibit clear conflicts between the annotated label and the semantic evidence. Their label-definition and prototype/profile scores are low, while the text-only view usually predicts another dominant emotion with high confidence. For instance, samples annotated as *angry* or *fear* may actually express risk expectation, self-isolation fear, or sadness/helplessness. These cases demonstrate that NF-DHSA can distinguish valid hard samples from truly noisy samples, rather than blindly discarding all high-difficulty instances.

HARD SAMPLE ID 73 Gold label: sad SMP2020-EWECT-Virus dataset Original text 不知道这样备受煎熬、在希望与失望之间反复挣扎的日子还要持续多久……但愿来年樱花盛开的时候一切都有好转。 Key rationale The dominant emotion is prolonged suffering, helplessness, and low mood, although it also contains a hopeful ending.	Multi-view scores Def.: High Proto.: High Text: Sad/ High LLM_final_reson: All three views support the annotated sad label, but the expression is long and context-dependent, making it a valid hard sample.
HARD SAMPLE ID 1532 Gold label: fear India-COVID-X dataset Original text ok so i do not really take stress about coronavirus but it is kind of annoying to be on the verge of getting a regular flu and your mind just lingers in the “but what if you also get corona would you notice” Key rationale The tone is casual, but the core concern is infection uncertainty and anxiety about missing symptoms.	Multi-view scores Def.: High Proto.: High Text: Fear/ High LLM_final_reson: The label is semantically valid, while the relaxed wording makes the fear signal implicit and difficult to learn.
HARD SAMPLE ID 139 Gold label: fear NLPCC2014 Original text 明天的辩论赛，在因学业而不得不放下一个学期之后，终于又能以攻辩的身份站在台上，莫名的紧张，甚至开始怀疑自己…… Key rationale The sample expresses nervousness and self-doubt before an uncertain event, matching fear/anxiety semantics.	Multi-view scores Def.: High Proto.: High Text: Fear/ High LLM_final_reson: All views support fear, but the cue is contextual and implicit rather than a direct fear word.

Figure 9: Case study of multi-view logic-guided noise filtering: hard samples.

NOISE SAMPLE ID 321 Gold label: angry SMP2020-EWECT-Virus dataset Original text 屋外火树银花，噼里啪啦，我在被窝自我隔离，有点害怕 Key rationale The text mainly describes self-isolation and fear; no clear blame, accusation, or hostile stance is expressed.	Multi-view scores Def.: Low Proto.: Low Text: Fear/ High LLM_final_reson: The label-conditioned views conflict with the annotated angry label, while the text-only view points to fear.
NOISE SAMPLE ID 1147 Gold label: anger India-COVID-X dataset Original text surgeon general has coronavirus warning “this week it is going to get bad” Key rationale The sentence reports a warning about worsening risk; it does not contain accusation, dissatisfaction, or anger.	Multi-view scores Def.: Low Proto.: Low Text: Fear/ High LLM_final_reson: The annotated anger label is weakly supported, while the dominant semantics are risk expectation and fear.
NOISE SAMPLE ID 197 Gold label: fear NLPCC2014 Original text 看着满屋子需要收藏的厚衣服和满屋子需要洗的脏衣服脏袜子，无从下手，情何以堪，幸好今天晚上没有课，只有大洗了，唉…… Key rationale The text mainly conveys fatigue, helplessness, and a sigh toward chores, rather than danger or future risk.	Multi-view scores Def.: Low Proto.: Mid Text: Sadness/ High LLM_final_reson: The fear label lacks definition-level support; the text-only judgment is closer to sadness/helplessness.

Figure 10: Case study of multi-view logic-guided noise filtering: noisy samples.