From Noise to Clarity: Filtering Real and LLM-Generated Samples for Enhanced Intent Detection

Junbao Huang[†], **Weizhen Li**[†], **Peijie Huang**[*], **Yuhong Xu**
College of Mathematics and Informatics, South China Agricultural University, China
gumbouh@163.com, lwz1311620816@163.com, pjhuang@scau.edu.cn,
xuyuhong@scau.edu.cn

Abstract

In dialogue intent detection, the challenge of acquiring sufficient corpora and the high cost of manual annotation often lead to incorrectly labeled or unrepresentative samples, which can hinder the generalization ability of classification models. Additionally, as using large language models for generating synthetic samples for data augmentation becomes more common, these synthetic samples may exacerbate the problem by introducing additional noise due to the models' limited prior knowledge. To address this challenge, this paper proposes an interpretable Sample Filter by Topic Modeling (SFTM) framework. By evaluating the diversity and authenticity of the samples, SFTM effectively reduces the quantity of real and synthetic samples while improving the performance of the classification models. Our codes are publicly available at https://github.com/gumbouh/SFTM.

1 Introduction

Intent detection is a fundamental and crucial task in dialogue systems, including utterance-level and dialogue-level intent detection. The typical approach involves converting spoken language into text through automatic speech recognition (ASR), followed by feeding the transcribed text into a model to obtain the corresponding intent label. This process is essential for ensuring that dialogue systems can accurately understand and respond to user intents.

Research in intent detection tasks generally follows two primary directions. The first direction involves fine-tuning large pre-trained language models (PLMs), such as the BERT (Devlin et al., 2019) encoder or auto-regressive models like Llama (Touvron et al., 2023), leveraging their strong semantic capabilities for downstream intent detection. The second direction focuses on specialized modeling approaches tailored to dialogue-specific challenges. For instance, models like AP-HAN (Xu et al., 2022) optimize based on dialogue turns, while HLDIC (Huang et al., 2024b) explores hierarchical structures within dialogue intents.

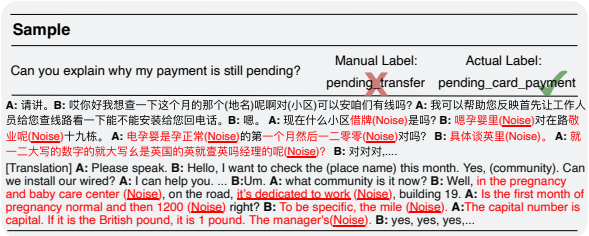

Figure 1: Two examples of low-quality samples are provided. The first mislabeled sample is from the Banking-77 dataset, and the second, noisy and unrepresentative sample is from the CMCC-34. Note that the CMCC-34 is a dialogue-level dataset, and the version shown in the figure is a dialogue fragment with speakers A and B.

While these studies emphasize model development, there has been comparatively less focus on improving the quality and characteristics of dialogue utterances.

Due to the challenges associated with collecting dialogue data, as well as the costs, quality issues, and errors introduced by both manual annotation and ASR, some of the dialogue datasets used for training may lack representativeness and be mislabeled. As shown in Fig. 1, one sample is mislabeled and the other is unrepresentative, both presenting as low-quality. Recent efforts (Lin et al., 2023a; Tang et al., 2023; Gao et al., 2023) have focused on using large language models (LLMs) to generate new samples for data augmentation, aiming to enrich the dataset. However, these LLMs are limited by their inherent prior knowledge and parameter constraints, which often result in synthetic dialogue data that lacks human-like diversity (Sahu et al., 2022; Li et al., 2023). Furthermore, the use of

[†] Equal contribution.
[*] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21736–21746
November 4-9, 2025

such generated data can introduce additional noise, potentially undermining the model's performance. Although recent studies have explored LLM-as-a-Judge (Li et al., 2024), where LLMs are provided with predefined evaluation metrics to score and assess data quality, such black-box methods often lack interpretability and controllability, limiting their practical utility in ensuring robust data curation.

To address this problem, we propose Sample Filter by Topic Modeling (SFTM), an interpretable dual-criteria framework for quality-aware data curation that systematically evaluates both real and synthetic samples through Distribution Diversity and Sample Authenticity metrics. SFTM is based on the Neural Topic Modeling (NTM) framework, specifically the NVDM-GSM (Miao et al., 2017), which leverages a Variational Autoencoder (VAE) (Kingma and Welling, 2014) to optimize and learn an approximate sample distribution, while also incorporating a supervised classification module. Consequently, SFTM can evaluate the quality of both real and synthetic samples from two interpretable dimensions: **Distribution Diversity** and **Sample Authenticity**, allowing it to comprehensively assess the samples from different perspectives.

This careful curation of both real and synthetic samples ensures that the augmented dataset not only remains representative but also supports the model in achieving better performance in intent detection tasks.

Our contributions are as follows:

- We propose Sample Filter by Topic Modeling (SFTM), an interpretable filtering mechanism that innovatively refines classical architectures to select high-quality samples for downstream intent detection.
- We introduce a unified filtering strategy for both real and synthetic samples by leveraging distribution diversity based on KL divergence and sample authenticity based on logits.
- The effectiveness of SFTM is validated through sample quality assessments and downstream intent detection experiments in both full-shot and few-shot settings. The filtered high-quality samples enhance model fitting and improve classification performance.

2 Related Work

2.1 Intent Detection

Intent detection has evolved from traditional feature engineering approaches to modern deep learning paradigms. Early methods relied on manual feature extraction combined with classifiers like Recurrent Neural Networks (RNNs) (Elman, 1990; Hochreiter and Schmidhuber, 1997). With the advent of pre-trained language models (PLMs), fine-tuning BERT (Devlin et al., 2019) and its variants has become a dominant approach due to their superior semantic representation capabilities (Liu et al., 2019; Henderson et al., 2020). Recent studies further explore dialogue-specific scenarios, and researchers have proposed hierarchical architectures to model multi-turn interactions. AP-HAN (Xu et al., 2022) employs hierarchical attention networks to capture turn-level dependencies, while HLDIC (Huang et al., 2024b) conducts hierarchical modeling based on the structure of intent labels.

As model architectures increasingly converge to the Transformer-like family, researchers have turned to data-level techniques for data augmentation. LRSL (Huang et al., 2024a) reranks the classification logits of hard samples by leveraging label semantic information, aiming to improve the classification performance of hard samples. More studies focus on synthesizing a large number of similar samples from LLMs. During training, these synthetic samples, along with real training samples, are added to enhance the model's fitting ability. Our method also enhances the accuracy of intent detection by synthesizing samples from LLMs. However, considering the noise and uneven quality of the samples synthesized by large models (Sahu et al., 2022; Li et al., 2023), we employ a sample filter to obtain high-quality samples, thereby improving the classification performance of downstream intent detection tasks.

2.2 Data synthesis

Early edit-based methods, such as Easy Data Augmentation (Wei and Zou, 2019), relied on rule-based transformations to generate new samples from existing ones. Another popular approach, back-translation (Sennrich et al., 2016), leverages the semantic drift introduced by translating the text into another language and then back to the original language to create novel samples. However, these methods often fail to produce sentences that are sufficiently challenging or semantically diverse,

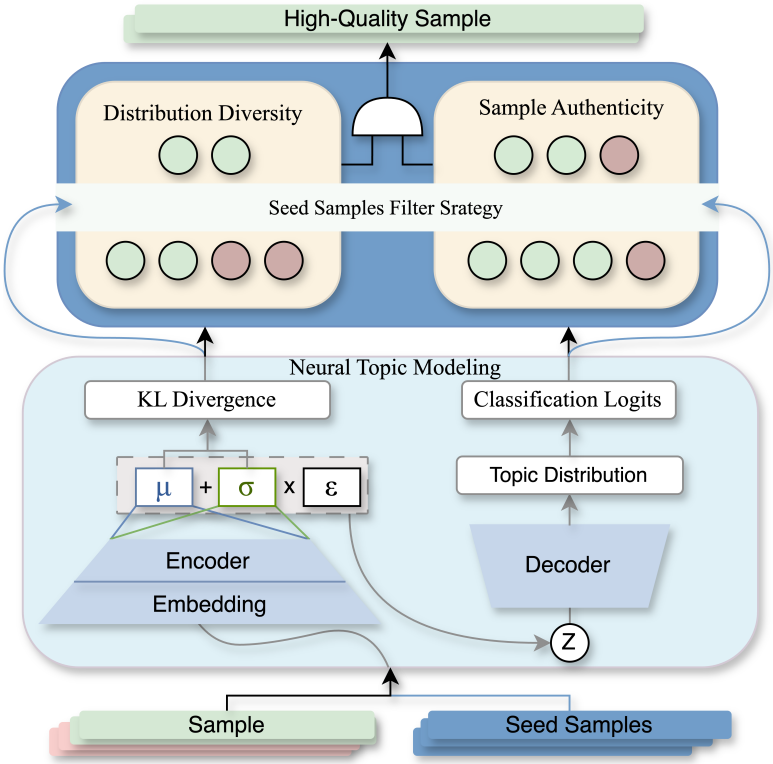

Figure 2: The overall framework of SFTM. The input samples are represented in two colors: green denotes high-quality samples, while red indicates low-quality samples. The filtering strategy leverages both the mean KL divergence and logits of seed samples to jointly evaluate and select high-quality samples.

which are critical for improving the robustness of downstream models.

Recently, there has been a growing trend in utilizing LLMs for data synthesis (Gupta et al., 2023; Choenni et al., 2023; Wang et al., 2024). These LLMs, due to their powerful learning capabilities, can perform few-shot learning by mimicking sample generation when provided with seed examples and corresponding labels. Despite these advantages, LLM-generated samples may suffer from noise (e.g., mislabeled data) and overfitting to the given seed samples, leading to a lack of diversity. This limitation arises from the inherent constraints of LLM parameters and the absence of domain-specific knowledge in their prior training.

To mitigate these issues, several studies have proposed filtering mechanisms to improve sample quality. For instance, ICDA (Lin et al., 2023b) employs Pointwise V-Information to evaluate and filter out low-quality samples. Another approach uses a small set of labeled data to fine-tune PLMs, which then serve as evaluators to score and filter synthetic data.

In this paper, we propose a modified VAE-based Neural Topic Model to comprehensively assess sample quality. Our model leverages the VAE framework to learn the latent distribution of the data while simultaneously scoring samples based on their authenticity. This dual-dimensional evaluation offers a robust mechanism for filtering synthetic data, ensuring higher quality and greater diversity in the resulting dataset.

3 Method

As shown in Fig. 2, our proposed method SFTM for filtering real and synthetic samples in intent detection consists of four interconnected components: the Neural Topic Modeling Module, the Distribution Diversity Evaluation Module, the Sample Authenticity Evaluation Module, and the Sample Filtering Strategy. The following sections detail the design and functionality of each module.

3.1 Neural Topic Modeling Module

The core of our model is the NTM module, which leverages a VAE enhanced by PLMs to capture the latent semantic structure of input samples. We utilize the semantic capabilities of PLM embeddings (Reimers and Gurevych, 2019) as the foundation for topic modeling.

Given an input sample, the NTM module employs the BERT to project the input data X into a latent space, producing a mean vector μ and a variance vector σ, which parameterize a Gaussian distribution:

$$z \sim \mathcal{N}(\mu, \sigma^2), \tag{1}$$

where z represents the latent variable, which is challenging to sample directly.

Therefore, we need to use reparameterization, which reformulates the sampling operation of the random variable $\hat{z}$ by combining μ, σ and a noise term ϵ drawn from a standard normal distribution. The formula is as follows:

$$\hat{z} = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1). \tag{2}$$

This approach makes the sampling process differentiable, enabling the optimization of model parameters through backpropagation.

Following prior research (Miao et al., 2017), the latent representation $\hat{z}$ is fed into a decoder, a simple multilayer perceptron (MLP) for input reconstruction. This design allows the model to capture both the intrinsic structure of samples and their topic distributions in a low-dimensional latent space.

3.2 Distribution Diversity Evaluation Module

The Distribution Diversity Evaluation Module aims to assess and preserve the diversity of the input samples, which evaluates the diversity of the sample distribution using two metrics: Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) and reconstruction loss.

KL Divergence. This metric quantifies the difference between the approximate posterior distribution $q(\mathbf{z}|\mathbf{X})$, inferred by the encoder, and the prior distribution $p(\mathbf{z})$, which is typically a standard Gaussian. The KL divergence term encourages the learned latent space to align with the prior distribution, ensuring that the representations remain diverse and regularized, which helps prevent overfitting and mode collapse. By doing so, it promotes the exploration of a broad range of possible latent codes. The KL divergence is computed as follows:

$$\mathrm{KL}(q(\mathbf{z}|\mathbf{X})\|p(\mathbf{z})) = -\frac{1}{2}\sum_{i=1}^{N}\left[1+\log(\sigma_i^2)-\mu_i^2-\sigma_i^2\right], \tag{3}$$

where μ_i is the i-th element of the posterior mean vector μ generated after BERT projects the input data X into the latent space; σ_i^2 corresponds to the variance component capturing distribution dispersion around μ_i; N is the dimensionality of the latent topic space.

Reconstruction Loss. This loss measures the discrepancy between the original input and its reconstruction by the decoder. A lower reconstruction loss indicates that the latent representation $\mathbf{z}$ effectively captures the essential features of the input sample. Reconstruction loss is typically calculated using the Mean Squared Error (MSE) between the reconstructed output R and the original input O:

$$\mathrm{MSE}(\mathrm{R}, \mathrm{O}) = \frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{R}_i - \mathrm{O}_i\right)^2, \tag{4}$$

where O represents the original input sample, and R represents the output after reconstruction by the decoder.

By optimizing these two metrics, the Distribution Diversity Evaluation Module can assess the distributional differences between samples, thereby providing an evaluation metric for sample diversity.

3.3 Sample Authenticity Evaluation Module

Given that the samples are labeled, it is essential to assess their authenticity by examining the relationship between the labels and the samples themselves. To address this, we incorporate a supervised classification task. Specifically, we employ R for classification, enabling us to evaluate how well the reconstructed samples align with their corresponding labels. We employ a single linear layer to classify, resulting in logits:

$$\mathrm{Logits} = \mathrm{Classifier}(\mathrm{R}), \tag{5}$$

where Classifier denotes a MLP used to classify the corresponding label. These logits are then used to evaluate the authenticity of the samples.

3.4 Sample Filtering by Seed Samples

Based on the evaluation metrics of Distribution Diversity and Sample Authenticity, SFTM adaptively applies distinct filtering strategies to different types of samples by introducing seed samples as dynamic evaluation thresholds. To ensure consistency and eliminate biases from seed samples, we select them from the training sets to minimize potential errors arising from manual selection.

KL-based Distribution Diversity. SFTM employs KL divergence to measure the discrepancy between the latent distribution of samples and a Gaussian

prior. Lower KL values indicate distributions closer to the prior, often signifying samples that are overly simplistic, uninformative, or unrepresentative. Retaining such samples may cause two key issues: 1) The model may overfit to these uninformative or noisy samples, negatively impacting its generalization ability. 2) Simpler samples may overshadow complex examples, limiting the model's capacity to learn diverse features. By filtering these samples, SFTM enhances training efficiency by prioritizing high-information data, thereby improving model performance.

The KL-based thresholding mechanism for distribution diversity filtering is formally defined in Algorithm 1, where class-specific thresholds dynamically adapt to seed sample distributions, ensuring representative diversity characteristics are preserved.

Logits-based Sample Authenticity. Logits quantify the alignment between samples and their assigned labels. High logits value from SFTM indicate samples that are easily classifiable, suggesting they are simplistic and offer limited training value. Conversely, if the logits are particularly low, it suggests potential label errors, as these samples may be incorrectly labeled, which could interfere with the model's training process.

As shown in Algorithm 2, the Sample Authenticity filtering strategy defines a range as $\hat{p} \pm \beta$ (where $\hat{p}$ is the mean logits of seed samples) and retains samples within this interval.

Algorithm 1 KL-based Distribution Diversity Filtering

Require: Seed sample $\{\mathcal{S}_j\}_{j=1}^m$, synthetic sample $\{\tilde{\mathcal{X}}_j\}_{j=1}^m$

for each class $j \in \{1, 2, \ldots, m\}$ **do**
  Compute $\overline{KL}_j^{(\text{seed})}$ via Eq.3
  for each sample $\tilde{x}_k \in \tilde{\mathcal{X}}_j$ **do**
    Calculate $\text{KL}_k^{(\text{synth})}$ via Eq.3
    if $\text{KL}_k^{(\text{synth})} > \gamma \cdot \overline{KL}_j^{(\text{seed})}$ **then**
      Add $\tilde{x}_k$ to $\mathcal{F}_j$
    end if
  end for
end for

Ensure: Filtered sample $\{\mathcal{F}_j\}_{j=1}^m$

Algorithm 2 Logits-based Sample Authenticity Filtering

Require: Seed samples $\{\mathcal{S}_j\}_{j=1}^m$, synthetic samples $\{\tilde{\mathcal{X}}_j\}_{j=1}^m$, margin β

for each class $j \in \{1, \ldots, m\}$ **do**
  Compute mean logits: $\hat{p}_j$ via Eq.5
  for each $\tilde{x}_k \in \tilde{\mathcal{X}}_j$ **do**
    Compute logits: p_k via Eq.5
    if $p_k \in [\hat{p}_j - \beta, \hat{p}_j + \beta]$ **then**
      $\mathcal{G}_j \leftarrow \mathcal{G}_j \cup \{\tilde{x}_k\}$
    end if
  end for
end for

Ensure: Filtered samples $\{\mathcal{G}_j\}_{j=1}^m$

4 Experimental Setup

4.1 Datasets

Banking-77[1]: The dataset provides a very fine-grained set of intents in a banking domain. It comprises 13,083 customer service queries labeled with 77 intents. It focuses on fine-grained single-domain intent detection (Mehri et al., 2020).

CMCC-34[2]: This is a long-text, dialogue-level intent detection dataset for Chinese multi-turn customer service interactions, transcribed from recordings between users and service representatives. This dataset contains significant amounts of noise due to being transcribed from speech. It is considered a relatively realistic dataset in the field of intent detection. Details are shown in Appendix A.

[1]https://huggingface.co/datasets/banking77

[2]http://www.cips-cl.org/static/CCL2018/call-evaluation.html

4.2 Experiment Settings

4.2.1 Synthetic Sample Settings

The dataset used in our experiments comprises the full set of real training samples as well as synthetic samples generated by GPT-4o-mini (OpenAI, 2023), with 100 synthetic samples per class. The prompt is shown in Appendix B.

4.2.2 Filtering Settings

For seed sample selection, we used the 5-shot training set as seed samples for full-shot experiments, with 5 samples per class. In few-shot settings, respective real-sample training sets served as seed samples. We chose 0.2 as the value of β in the Sample Authenticity filtering strategy.

4.2.3 Downstream Training Settings.

We train the SFTM to filter the samples, resulting in a set of filtered real and synthetic sam-

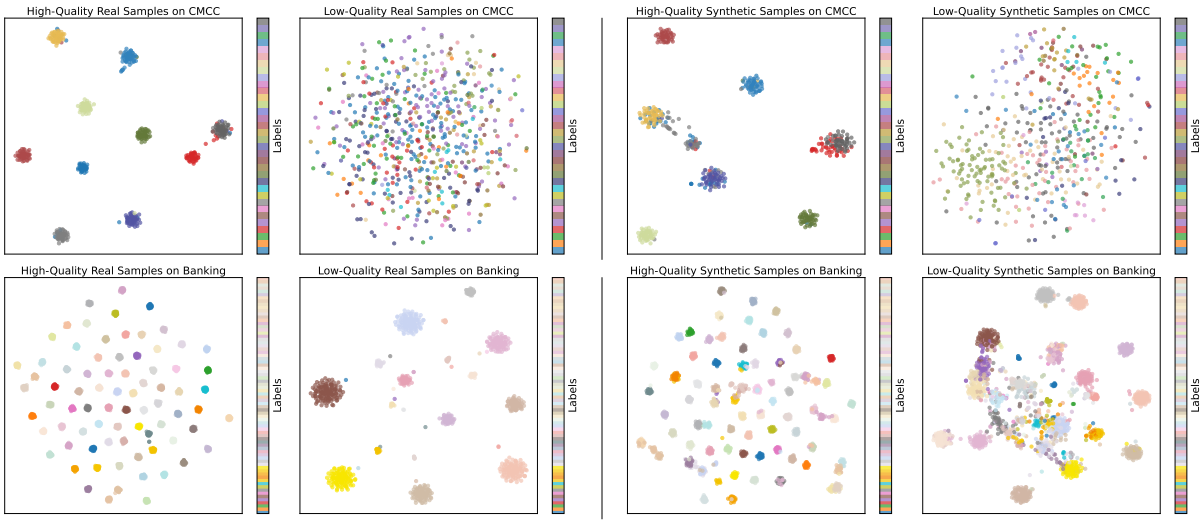

Figure 3: The relationship between the topic distribution of samples and their labeled classes, visualized using t-SNE.

ples. Subsequently, we conduct intent detection experiments using representative PLMs: BERT-base (Devlin et al., 2019), RoBERTa-Large (Liu et al., 2019), and Qwen-2.5-7B-Instruct (Qwen) (Hui et al., 2024). BERT and RoBERTa undergo full fine-tuning, while Qwen is fine-tuned using the LoRA (Hu et al., 2022) for efficient adaptation. For SFTM and BERT, the batch size is set to 24, and early stopping is employed with a patience strategy of 3. To prevent overfitting, a dropout with a probability of 0.1 is applied. The parameters are updated using the Adam algorithm, with the learning rate initialized to 2e-5. For LoRA, we configure the LoRA rank (r) to 8 and the LoRA alpha to 32. The Qwen is trained with a batch size of 8 over 3 epochs.

For comprehensive evaluation, we establish experimental comparisons under both full-shot and few-shot learning paradigms.

Full-shot Experiments. In the case of real samples, we filter out 5% of low-quality samples from both the CMCC-34 and Banking-77 datasets. Regarding synthetic samples, SFTM demonstrates significant filtering effectiveness. In the CMCC-34 dataset, it removes 80% of low-quality data, leaving 20% (680 high-quality synthetic samples). For the Banking-77 dataset, 68.2% of synthetic samples are filtered out, resulting in 2,448 high-quality samples.

Few-shot Experiments. When using SFTM to filter synthetic samples, notable results are obtained. In the Banking-77 dataset, 86.29% of relatively low-quality synthetic samples are eliminated, leaving 1,056 high-quality samples. As for the CMCC-34 dataset, 27.24% of low-quality samples are removed, leaving 2,474 high-quality samples.

4.3 Baselines

4.3.1 Baselines for full-shot experiments.

We benchmark against three representative baselines: **LRSL:** (Huang et al., 2024a): A semantic label-guided data augmentation approach. **DA-GPT:** Direct augmentation using GPT-4o-mini generated synthetic samples. **DA-Judge:** Black-box data curation with systematic evaluation metrics, employing GPT-4o as the scoring model. More details are shown in Appendix C.

4.3.2 Baselines for few-shot experiments

Following standard practice for previous methods (Lin et al., 2023b), we employ RoBERTa-Large and Qwen as the PLMs for few-shot experiments. We compare the following methods: **DA-GPT** and **DA-Judge** as above. **ICDA** (Lin et al., 2023b): Current state-of-the-art few-shot intent detection method.

5 Results and Discussion

5.1 Sample Quality Assessment

To quantitatively assess sample quality, we analyze the distribution patterns of high- and low-quality filtered samples using t-SNE (Van der Maaten and Hinton, 2008) visualization, as presented in Fig. 3.

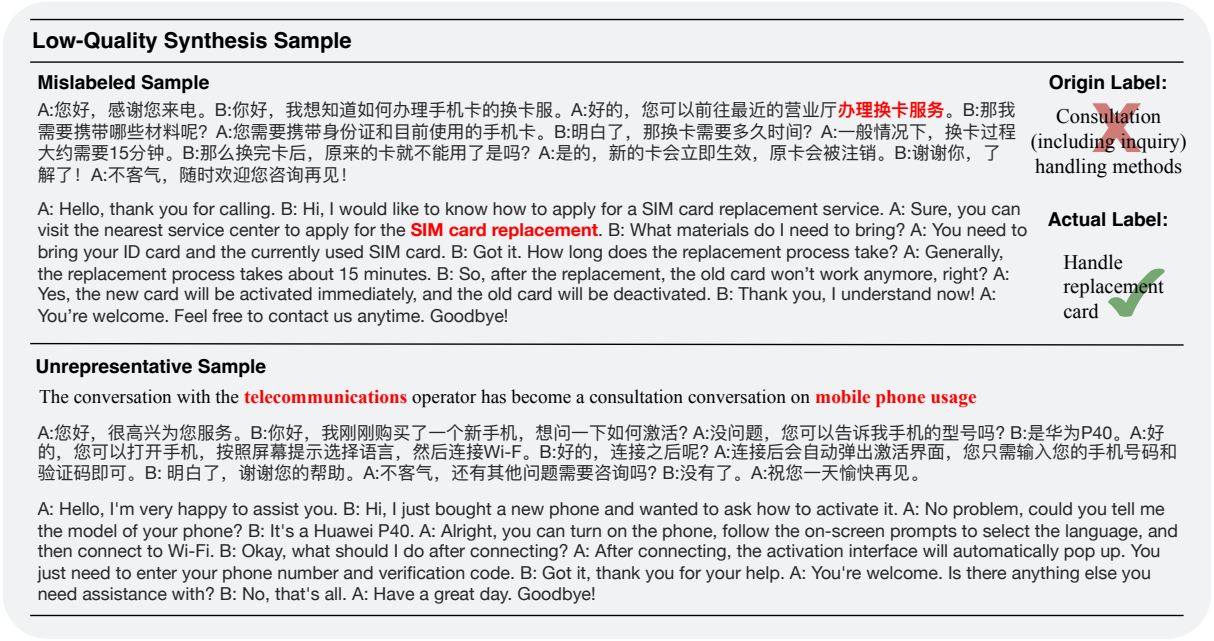

Figure 4: Two representative cases of LLM-generated low-quality samples from the CMCC-34 dataset. The figure presents dialogue fragments with interactions between speakers A and B.

5.1.1 Real Samples Analysis

High-quality samples from both datasets form distinct, tight clusters with well-separated class boundaries, indicating their strong representational capacity for downstream classification. In contrast, low-quality CMCC-34 samples exhibit disordered dispersion patterns, confirming the substantial noise mentioned in Fig. 1 - a manifestation of *unrepresentative samples* effectively identified by SFTM. While Banking-77's low-quality samples retain partial cluster structures, significant inter-cluster overlap reveals label confusion issues, corresponding to Fig. 1's *mislabeled samples*.

5.1.2 Synthetic Samples Analysis

Synthetic samples exhibit fundamentally inferior quality compared to real samples, necessitating rigorous filtering. SFTM filtered high-quality synthetic samples demonstrate clearer separation boundaries than their low-quality counterparts, which show exacerbated dispersion and overlap exceeding real samples' noise levels. This confirms SFTM's superior effectiveness in synthetic data curation - it successfully identifies relatively higher-quality candidates from predominantly mediocre synthetic samples, thereby amplifying data augmentation benefits while mitigating adverse effects from low-quality instances.

5.1.3 Case Study

We conduct a case study on two representative low-quality samples from CMCC-34 filtered by SFTM to elucidate LLM-generated samples, as visualized in Fig. 4.

Case 1: Mislabeled Sample. The first case demonstrates *intent hallucination* despite few-shot prompting: A sample annotated as *Consultation (including inquiry) handing methods* actually expresses *Handle replacement card* intent. This reveals LLMs' tendency to generate semantically inconsistent samples even when provided with clear intent definitions and demonstrations, highlighting the necessity of post-generation verification.

Case 2: Unrepresentative Sample. The second case exhibits *domain deviation*: While instructed to generate customer service dialogues for telecom operators, the LLM produces samples about smartphone activation - a divergent theme lacking domain relevance. Such samples introduce distributional shifts that degrade model performance, constituting precisely the noise SFTM aims to eliminate.

5.1.4 Key Insight

The visualization quantitatively validates SFTM's dual capability in 1) preserving semantically coherent samples with discriminative cluster structures, and 2) eliminating noisy instances that disrupt class separability - a crucial mechanism for enhancing

model generalization.

5.2 Downstream Full-shot Experiments

As shown in Table 1, our method demonstrates significant performance improvements across two representative datasets and distinct model architectures. Specifically, SFTM achieves maximum gains of 1.5% on sentence-level Banking-77 and 3.59% on dialogue-level CMCC-34. The empirical results suggest that SFTM exhibits stronger effectiveness in dialogue-level scenarios, where it effectively selects higher-quality samples that enable downstream models to better fit data distributions and enhance classification performance.

Less is More. Compared with DA-GPT's data augmentation approach using synthetic samples (employing 7,700 samples for Banking-77 and 3,400 for CMCC-34), SFTM achieves superior performance with substantially fewer samples (2,448 for Banking-77 and 680 for CMCC-34). This demonstrates that for intent detection tasks, model performance depends more critically on sample quality than quantity. The quality of LLM-generated samples is inherently constrained by model parameters and domain-specific prior knowledge, making systematic filtering imperative. As an efficient and interpretable filtering method, SFTM employs dual-dimensional evaluation to effectively select high-quality samples for performance enhancement.

Synthetic Samples Outperform Feature Augmentation. In contrast to LRSL of leveraging text embeddings to enhance classification confidence for hard samples, SFTM adopts a more fundamental strategy through meticulous data curation. This simple method achieves superior results without modifying the classification paradigm, presenting a more generalizable and efficient solution.

Ablation Study. To further investigate SFTM's effectiveness, we conduct ablation studies in Table 2. **For Real Samples**, applying SFTM filters out a portion of the dataset, consistently improving performance compared to using the full training set, with gains of up to 1%. This demonstrates that SFTM effectively removes low-quality real samples, thereby enhancing the representativeness of the remaining data. **For Synthetic Samples**, the improvements after applying SFTM are similarly significant. Compared to using the full set of synthetic data, the performance gain after filtering reaches up to 2.09%. Additionally, when compared to using only the full set of real training data without synthetic samples, the performance improvement can be as high as 2.77%. **Moreover**, combining the filtered real samples with the filtered synthetic samples results in even more pronounced performance gains, with an improvement of 3.59%. This shows that fewer, higher quality samples outperform unfiltered ones, even when the latter have larger quantities, as noise in the unfiltered data can degrade the model's performance.

Model	Dataset	
	Banking-77	CMCC-34
BERT-base	92.22	56.53
+ DA-GPT	92.65	57.21
+ DA-Judge	93.42	58.82
+ LRSL	<u>93.66</u>	<u>59.44</u>
+ SFTM(ours)	**93.72**	**60.12**
Qwen2.5-7B-LoRA	91.92	58.70
+ DA-GPT	92.42	59.55
+ DA-Judge	<u>92.72</u>	60.25
+ LRSL	92.32	<u>60.40</u>
+ SFTM (ours)	**93.07**	**61.01**

Table 1: Full-shot experimental results on Banking-77 and CMCC-34. Accuracy (%) is used as the evaluation metric, with **bold** indicating the best performance and <u>underlined</u> values denoting the second-best results.

Model	Dataset	
	Banking-77	CMCC-34
BERT-base + All Synthetic	92.65	57.21
BERT-base	92.22	56.53
− **Filtered Real**	92.81	57.40
+ **Filtered Synthetic**	<u>93.30</u>	<u>59.30</u>
− **Filtered Real** + **Filtered Synthetic**	**93.72**	**60.12**
Qwen2.5-7B-LoRA + All Synthetic	92.42	59.55
Qwen2.5-7B-LoRA	91.92	58.70
− **Filtered Real**	92.92	59.58
+ **Filtered Synthetic**	<u>93.02</u>	<u>60.65</u>
− **Filtered Real** + **Filtered Synthetic**	**93.07**	**61.01**

Table 2: Ablation Study on Full-shot settings.

5.3 Downstream Few-shot Experiments

Enhanced Verification of SFTM's Filtering Efficacy. To rigorously validate SFTM's effectiveness in filtering LLM-generated samples, we conduct few-shot intent detection experiments where all augmented data originates from synthetic samples. The experimental results presented in Table 3 demonstrate that under few-shot settings, our filtered samples consistently achieve superior performance in downstream tasks. Specifically, our method surpasses the previous SOTA approach ICDA on Banking-77 under equivalent configura-

Model	Dataset			
	Banking-77		CMCC-34	
	5-shot	10-shot	5-shot	10-shot
RoBERTa-Large	76.36	86.17	19.43	29.53
+ ICDA*	81.95	87.37	-	-
+ DA-GPT	83.60	86.75	29.48	32.08
+ DA-Judge	82.42	87.25	29.72	32.58
+ SFTM(ours)	**83.77**	**87.69**	**30.45**	**33.60**
Qwen2.5-7B-LoRA	73.05	84.90	17.75	24.65
+ DA-GPT	80.71	86.70	26.90	33.38
+ DA-Judge	79.15	86.24	27.42	33.98
+ SFTM(ours)	**81.68**	**86.72**	**32.50**	**35.35**

Table 3: Few-shot experimental results. Due to the non-released implementation of **ICDA***, we adopt **ICDA-S** from the original paper for comparison, as its synthetic sample size (1,540) closely aligns with ours (1,056) under identical experimental settings.

tions, while showing significant improvements over DA-Judge on CMCC-34, with a maximum gain of 5.08%. DA-Judge, as a widely used black-box data curation method, demonstrates negative effects on Banking-77 in few-shot settings compared to DA-GPT. In contrast, SFTM, as a white-box interpretable method, consistently achieves stable and effective performance across both datasets.

Counterintuitive LLM Behavior Analysis. Notably, in dialogue-level CMCC-34 scenarios, we observe the unexpected phenomenon where Qwen2.5-7B's few-shot performance initially underperforms RoBERTa-Large - a counterintuitive finding given LLMs' renowned generalization capabilities. We attribute this to the substantial noise present in CMCC-34's original samples, including frequent misspellings that induce semantic misinterpretation. Crucially, when employing SFTM-filtered samples for data augmentation, Qwen2.5-7B achieves greater performance gains than RoBERTa-Large. This suggests that high-quality filtered samples can significantly enhance LLMs' generalization potential by mitigating noise-induced learning biases.

6 Conclusion

We propose SFTM, an interpretable framework for filtering both real and synthetic samples using two key metrics: Distribution Diversity and Sample Authenticity. Extensive experiments show that SFTM effectively filters out low-quality samples, resulting in significant performance improvements in classification models.

7 Acknowledgements

This work was supported by the National Natural Science Foundation of China (62306119 and 71472068) and the Science and Technology Projects in Guangzhou (2025A04J3436).

Limitations

This work primarily focuses on data augmentation for intent detection tasks, while the exploration of this paradigm for other Natural Language Understanding or Generation tasks remains an open research direction. We conduct comparative studies against commonly used baseline methods to validate the effectiveness of our approach. Additionally, alternative methods may exist to evaluate the quality of samples, which could be investigated in future work.

References

Sunil Choenni, Tony Busker, and Mortaza S. Bargh. 2023. Generating synthetic data from large language models. In *15th International Conference on Innovations in Information Technology, IIT 2023, Al Ain, United Arab Emirates, November 14-15, 2023*, pages 73–78.

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

Jeffrey L. Elman. 1990. Finding structure in time. *Cogn. Sci.*, 14(2):179–211.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. 2023. Targen: Targeted data generation with large language models. *CoRR*, abs/2310.17876.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Peijie Huang, Junbao Huang, Yuhong Xu, Weizhen Li, and Xisheng Xiao. 2024a. Logits reranking via semantic labels for hard samples in text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11250–11262, Miami, Florida, USA.

Simin Huang, Peijie Huang, Yuhong Xu, Jingzhou Liang, and Jingde Niu. 2024b. Exploring label hierarchy in dialogue intent classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11511–11515.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, An Yang, Rui Men, Fei Huang, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. Qwen2.5-coder technical report. *CoRR*, abs/2409.12186.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

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

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *CoRR*, abs/2411.16594.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore.

Yen Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023a. Selective in-context data augmentation for intent detection using pointwise v-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1455–1468.

Yen Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023b. Selective in-context data augmentation for intent detection using pointwise v-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1455–1468.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tür. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *CoRR*, abs/2009.13570.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990.

Gaurav Sahu, Pau Rodríguez, Issam H. Laradji, Parmida Atighehchian, David Vázquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI, ConvAI@ACL 2022, Dublin, Ireland, May 27, 2022*, pages 47–57.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *CoRR*, abs/2303.04360.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(86):2579–2605.

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, and Yunhong Wang. 2024. A survey on data synthesis and augmentation for large language models. *CoRR*, abs/2410.12896.

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.

Jiabao Xu, Peijie Huang, Youming Peng, Jiande Ding, Boxi Huang, and Simin Huang. 2022. Adjacency pairs-aware hierarchical attention networks for dialogue intent classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7622–7626.

A Details of Datasets

Dataset	Type	Split of train/dev/test	Average Token	Label
CMCC-34	Dialogue Level	12799 / 3200 / 4000	379	34
Banking-77	Utterance Level	8011 / 2006 / 3084	14	77

Table 4: Details of CMCC-34, and Banking-77 datasets.

In this section, we provide detailed information about the datasets used in our experiments. The datasets include CMCC-34, and Banking-77. The following table 4 summarizes the split of the train, development (dev), and test sets, along with the type, the average token length and the number of intents for each dataset.

B Prompt for Synthesizing Samples from LLM

To generate synthetic samples that closely resemble real data from the outset, we employ few-shot prompting, where the LLM learns to mimic patterns by being provided with a small number of authentic examples. As illustrated in Fig. 5, this few-shot prompting approach is widely adopted due to its effectiveness in guiding LLMs to produce higher-quality synthetic samples. Specifically, the provided real samples serve as contextual anchors, enabling the model to better capture domain-specific linguistic patterns and intent representations, thereby reducing the generation of low-quality or irrelevant outputs. This method strikes a balance between sample diversity and authenticity, making it a practical choice for data augmentation tasks.

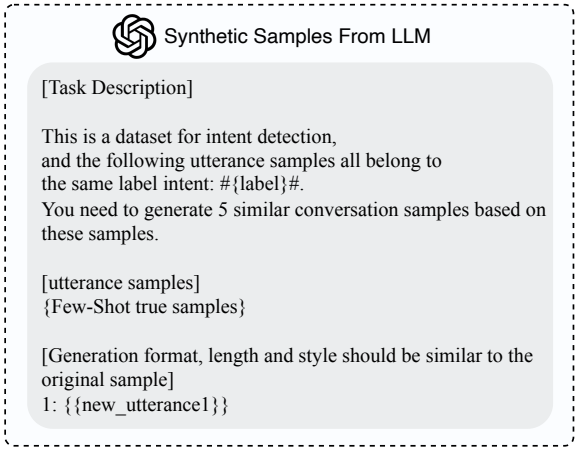

Figure 5: The prompt of generating samples from LLM.

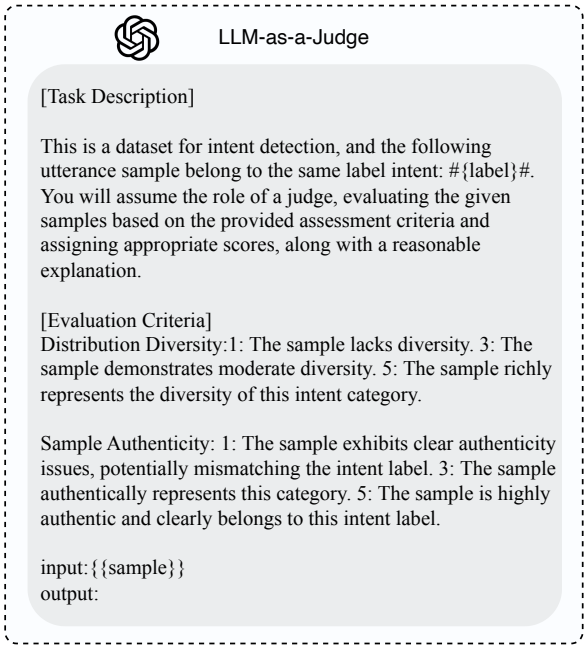

Figure 6: The prompt of judging samples by LLM.

C Prompt for DA-Judge

We adopt the LLM-as-a-Judge paradigm (DA-Judge), where the LLM evaluates samples based on predefined, well-designed metrics. As shown in Fig.6, we utilize the same two criteria as SFTM—Distribution Diversity and Sample Authenticity—to ensure a fair and consistent comparison.