NEID: A Semantic-Aware Framework for Discovering New Intents in Educational Dialogues

Anonymous ACL submission

Abstract

Understanding educational intents in teacher-student dialogues is vital for optimizing teaching strategies. However, traditional analysis relies on static coding schemes that fail to adapt to diverse scenarios, while automatically discovering new intents from large-scale unlabeled corpora remains under-explored due to data sparsity and noise. To address these challenges, we propose a New Educational Intent Discovery (NEID) framework. Specifically, we adopt multi-task pre-training with context and speaker information for representation learning. We further propose a reverse knowledge distillation strategy that guides the model to identify noisy words and perform Semantic-Aware Data Augmentation (SADA) for prototypical contrastive learning. Additionally, an LLM-driven Semantic-Aware Cluster Refinement (SACR) method is introduced to optimize cluster structures in both feature and semantic space. Experiments on Mathdial and Talkmoves demonstrate the effectiveness of our framework; notably, it outperforms the strongest baseline by 19.81% in ACC under extreme low-resource settings (KCR=50%, LAR=10%) on Talkmoves.

1 Introduction

Educational intent characterizes the interactional functions underlying teacher-student dialogues, serving as a cornerstone for deconstructing instructional patterns and optimizing teaching strategies. Typically, these intents are analyzed through predefined coding schemes. However, existing coding schemes often face adaptability bottlenecks when applied to diverse educational scenarios. While adapting existing schemes to align with specific research objectives and educational scenarios has proven necessary and rational (Wang et al., 2025), this process unfortunately remains heavily reliant on manual annotation, suffering from high temporal costs (Qi et al., 2024; Wang and Chen, 2024; Wang et al., 2024) and iteration latency. Given the scarcity of labeled data in the educational domain (Qi et al., 2024), leveraging Artificial Intelligence (AI) techniques to automatically discover previously undefined intents (hereafter referred to as new educational intents) from massive unlabeled corpora has emerged as a viable pathway.

General New Intent Discovery (NID) methods aim to optimize the intent recognition capabilities of conversational AI by repeatedly integrating new intents discovered from unlabeled user utterances in real time (Zhang et al., 2022). We abstract the definition of new educational intents into a New Educational Intent Discovery (NEID) task. This task aims to assist educators in dynamically refining schemes for specific scenarios by discovering new intents automatically.

Unlike traditional text classification, intent classification emphasizes the speaker's verbal intent and communicative purpose (Searle, 1969), which is critical in analyzing the discursive functions of educational dialogue. For instance, the teacher utterance "*Can you match the steps in your calculations back to the original question for me?*" would typically be classified under the "*calculation*" category, but corresponds to the "*Focus*" educational intent within the scheme of the Mathdial dataset (Macina et al., 2023). Consequently, for the NEID framework dealing with unknown classes, it is important to guide the model to focus on semantic representations for effective intent mining.

Dialogue context and speaker information are pivotal in characterizing educational intent, as identical utterances often manifest distinct semantics across varying contexts and speakers (Lin et al., 2023a). This information should be explicitly encoded for NEID.

Another prominent challenge in NEID is inherent class imbalance. Intent distributions typically follow a long-tailed pattern (Mou et al., 2022), especially in dialogue systems (Zhang et al., 2024).

For instance, the disparity between the majority and minority classes reaches a ratio of 44:1 in the Talkmoves dataset (Suresh et al., 2022). As Lin et al. (2023b) demonstrate, imbalanced classes in low-resource settings severely degrade intent discovery, a phenomenon inherently present in educational corpora.

Given the background and challenges outlined above, our main contributions are as follows: (1) we present a novel three-stage NEID framework, which aims to discover new educational intents hidden in large-scale unlabeled educational data; (2) we design Semantic-Aware Data Augmentation (SADA), an effective method that captures key semantics and generates augmented views required for NEID; (3) we introduce Semantic-Aware Cluster Refinement (SACR), an LLM-empowered method to optimize cluster structures from a semantic perspective; (4) extensive experiments on Mathdial and Talkmoves demonstrate the effectiveness of our framework.

2 Related work

2.1 New Intent Discovery

The NID task aims to discover latent intents from unlabeled data beyond predefined categories (Al-Tuama and Nasrawi, 2022; Zhang et al., 2023a). Mainstream approaches typically couple supervised pre-training with contrastive clustering on the full corpus. They employ techniques such as multi-task learning (Zhang et al., 2022), pseudo-labeling (Zhang et al., 2023a), or prototype-anchored optimization (Deng et al., 2024) to structure the latent representation space. But most prior works overlook the long-tailed distributions frequently encountered in real-world scenarios (Zhang et al., 2024).

Recently, LLMs have been introduced to the NID task. Training-free methods, like IntentGPT (Rodriguez et al., 2024), directly prompt LLMs for intent generation. While other methods employ LLMs to ture cluster granularity (Zhang et al., 2023c), or to refine ambiguous edge samples during the clustering process (Feng et al., 2024).

However, end-to-end LLM discovery risks hallucination, and instance-level adjustments currently lack synergy with global cluster refinement. Leveraging LLMs for joint instance-cluster optimization in the feature space to resolve majority class collapse caused by data imbalance is a promising direction.

2.2 Educational Intent Detection

Educational Intent Detection predominantly serves Intelligent Tutoring Systems (ITS). Existing solutions range from multi-task architectures (Okur et al., 2023) and supervised learning (Cutler et al., 2025) to zero-/few-shot LLM prompting (Levonian et al., 2025), aiming to optimize interactive experiences.

However, while existing methods perform well on predefined intent categories, no prior work has extended existing paradigms to the NEID task, largely due to the inherent challenges of educational corpora: severe class imbalance and sparse annotations. To bridge this gap, this study proposes a three-stage NEID framework in educational dialogues.

3 Motivation

3.1 SADA Motivation

To fully leverage prior knowledge from labeled data within data-sparse educational domains and to mitigate the distribution shift between labeled and unlabeled data, this study employs Integrated Gradients (IG) (Sundararajan et al., 2017). Unlike BERT's self-attention mechanism, IG explains predictions by calculating the attribution scores of input features. This approach simultaneously reveals feature contributions to the prediction, effectively avoiding the interpretability pitfall that "attention is not explanation".

However, the label-aware attribution capability of IG necessitates reliance on labeled data. To address its transferability limitations in NEID, we propose a reverse strategy: shifting from "***learning what is important***" to "***learning what is unimportant***". We adopt a knowledge distillation paradigm to train a transformer student network to transfer the prior knowledge from labeled data to unlabeled data. In contrastive learning, the student model applies noise masking to utterances, thereby enabling SADA.

Premise: Ubiquity of Noise. Authentic educational dialogue corpora inherently contain a substantial amount of colloquial redundancies, self-corrections and pedagogical scaffolds, which appear with high frequency across various educational intents to maintain the conversational flow. Consequently, for NEID, learning a generalized noise distribution yields better transfer robustness than fitting sparse feature distributions.

3.2 SACR Motivation

In authentic educational scenarios, utterances with high lexical or structural similarity may convey distinct educational intents. Traditional contrastive learning typically relies solely on the latent feature space to optimize representation distributions. Therefore, we introduce LLMs to guide distribution optimization from a semantic perspective, ensuring that clusters are grouped according to their true educational intents.

4 Methods

4.1 Data Input Preparation

Educational intents exhibit significant context and speaker-role dependency; a single utterance often manifests different educational intents across varying educational scenarios. Therefore, NEID requires integrating context and speaker information. We formalize the input instance as a triplet (X, S, M), representing the token sequence, the segment IDs, and the special tokens mask, respectively.

Let u_t denote the current target utterance. We define `[CLS]` and `[SEP]` as the sequence start and separator tokens, and tok_{spk} as the speaker role (e.g., `[s]`, `[t]`). A binary variable $w \in \{0, 1\}$ indicates whether to include speaker information, giving the augmented utterance representation

$$\tilde{u} = \begin{cases} \text{tok}_{spk} \oplus u & \text{if } w = 1 \\ u & \text{if } w = 0, \end{cases} \tag{1}$$

where $\oplus$ denotes sequence concatenation. We further introduce $\mathcal{I}_{mode} \in \{\textit{Single}, \textit{Context}\}$ for two input strategies. When $\mathcal{I}_{mode} = \textit{Single}$, the Token Sequence is

$$X_{\text{Single}} = \texttt{[CLS]} \oplus \tilde{u}_t \oplus \texttt{[SEP]}, \tag{2}$$

with all-zero Segment IDs $S_{\text{Single}} = \{0\}^{|X_{\text{Single}}|}$. When $\mathcal{I}_{mode} = \textit{Context}$, we construct a dialogue window $\mathcal{D} = \{u_{t-1}, u_t, u_{t+1}\}$ and define

$$\begin{aligned} X_{\text{Context}} = \texttt{[CLS]} \oplus \underbrace{(\tilde{u}_{t-1} \oplus \texttt{[SEP]})}_{\text{Previous}} \\ \oplus \underbrace{(\tilde{u}_t \oplus \texttt{[SEP]})}_{\text{Target}} \oplus \underbrace{(\tilde{u}_{t+1} \oplus \texttt{[SEP]})}_{\text{Next}}, \end{aligned} \tag{3}$$

with $S_{\text{Context}} = \{0\}^{L_{\text{prev}}} \oplus \{1\}^{L_{\text{curr}}} \oplus \{0\}^{L_{\text{next}}}$, where L_{prev}, L_{curr}, and L_{next} are the lengths of the three segments including their special tokens. Assigning segment ID 1 to the target utterance and 0 to the context allows the model to attend to them differently, as elaborated in Sec. 4.3. To preserve the structural integrity of the input during subsequent augmentation, the special tokens mask is defined as

$$M_i = \begin{cases} 1 & \text{if } x_i \in \{\texttt{[CLS]}, \texttt{[SEP]}\} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The proposed input strategies explicitly embed context and speaker tokens, constructing a comprehensive semantic background for the NEID framework.

4.2 Multi-task Pre-training

To address labeled-data sparsity in NEID, we adopt the multi-task pre-training method of Zhang et al.(Zhang et al., 2022), which integrates the task dataset $\mathcal{D}_{\text{task}}$ with an external public intent dataset $\mathcal{D}_{\text{public}}$ to facilitate the transfer of shared intent features (Wu et al., 2026).

To capture domain-specific knowledge and linguistic patterns in educational dialogues, we apply self-supervised masked language modeling (MLM) on the full task dataset $\mathcal{D}_{\text{task}}^{\text{all}}$ (labeled and unlabeled combined). In parallel, supervised classification is applied to the task-labeled set $\mathcal{D}_{\text{task}}^{\text{labeled}}$ and the public-labeled set $\mathcal{D}_{\text{public}}^{\text{labeled}}$ to establish distinct intent decision boundaries. The joint objective is

$$\begin{aligned} \mathcal{L}_{\text{loss}} = \underbrace{\mathcal{L}_{\text{ce}}(\mathcal{D}_{\text{public}}^{\text{labeled}}; \theta) + \mathcal{L}_{\text{ce}}(\mathcal{D}_{\text{task}}^{\text{labeled}}; \theta)}_{\text{supervised}} \\ + \underbrace{\mathcal{L}_{\text{mlm}}(\mathcal{D}_{\text{task}}^{\text{all}}; \theta)}_{\text{self-supervised}}, \end{aligned} \tag{5}$$

where θ denotes the model parameters.

4.3 Knowledge Distillation

Segment IDs are designed to provide structural boundary information for utterances. Setting them uniformly to zero hinders the model from distinguishing token attribution at the sentence level and degrades token-level prediction (Van Der Goot et al., 2022). In low-resource scenarios, differentiated segment IDs project distinct utterances into distinct feature spaces, producing pronounced attention-guidance effects.

To achieve precise intent attribution within concatenated contexts, we therefore propose a

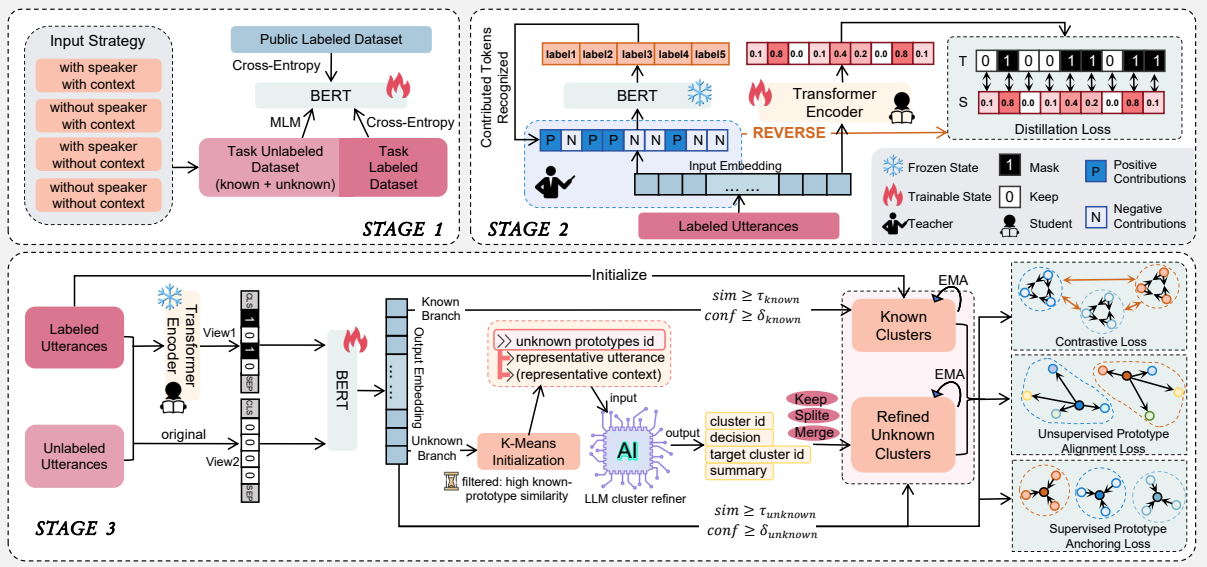

Figure 1: The overall NEID framework. Legends are interpreted in the lower-right corner of Stage 2. In stage 1, we perform multi-task pre-training on both public and task datasets (Sec. 4.2). In stage 2, we perform knowledge distillation, where the student model internalizes noise detection capabilities under the guidance of the teacher model (Sec. 4.3). In stage 3, we perform prototypical contrastive learning empowered by SADA and SACR (Sec. 4.4). After training, a non-parametric clustering algorithm is applied to obtain the final clusters, each of which corresponds to one type of educational intent.

structure-aware IG algorithm that fully exploits the model's attention on segment 1. Specifically, we assign segment ID 1 to the target utterance and 0 to the surrounding context. This differentiated structure incorporates essential contextual cues while imposing structural constraints that prevent attention divergence in long sequences, yielding superior intent recognition compared with the default all-zero input.

In this stage, we re-align tokens into words and freeze the teacher F_T pre-trained in Sec. 4.2. F_T constructs a word-level binary noise mask y^{mask} from word attribution scores $\Phi = \{\phi_1, \ldots, \phi_N\}$ under a noise ratio τ. We sort Φ in ascending order and take the $\lfloor N\tau \rfloor$-th value as the ratio δ_τ, so that δ_τ represents the attribution intensity at the lowest-τ proportion of Φ. We adopt this proportion-based strategy rather than an absolute value to ensure balanced noise identification across samples with varying attribution score magnitudes. For non-special tokens (i.e., $M_i = 0$), the noise mask is defined as

$$y_i^{\text{mask}} = \begin{cases} 0 \text{ (Keep)} & \text{if } \phi_i > \delta_\tau \\ 1 \text{ (MASK)} & \text{otherwise,} \end{cases} \tag{6}$$

where $y_i^{\text{mask}} = 1$ marks x_i as a noisy word to be masked and 0 marks it to be retained; special tokens always have $y_i^{\text{mask}} = 0$. The implementation of F_T is partially based on Liao et al.'s (2024) method.

The student F_S fits the mask distribution produced by F_T. Its output noise probability for x_i is $p_i = \sigma(F_S(f_i)/T)$, where f_i is the embedding feature of x_i, and $\sigma(\cdot)$ is the sigmoid function. F_S is optimized end-to-end by minimizing the binary cross-entropy between p_i and y_i^{mask}. Through this objective, F_S acquires a generalizable noise-discrimination capability grounded in semantic features rather than positional cues.

4.4 Prototypical Contrastive Learning

Application of SADA. In this stage, we first deploy the student model F_S (Sec. 4.3) to perform SADA. To prevent semantic drift, F_S inherits the noise ratio of F_T. For each utterance, we construct dual views $(\mathbf{v}_1, \mathbf{v}_2)$, where $\mathbf{v}_1$ is augmented via SADA and $\mathbf{v}_2$ is the unaltered original text. These views are processed by a BERT encoder to yield their projected feature vectors $\mathbf{f}_1$ and $\mathbf{f}_2$.

Prototypical Routing Mechanism. Next, we employ prototypical contrastive learning to jointly align instance-level and cluster-level representations (Deng et al., 2024). Governed by a gating mechanism, the features are routed through two parallel branches to either known or unknown prototypes. Concurrently, these prototypes are dynamically updated using the routed features, thereby refining the cluster structures.

Formally, we maintain a set of known prototypes $\mathcal{P}_{\text{known}} = \{\mathbf{p}_k\}_{k=1}^{K}$ and unknown prototypes $\mathcal{P}_{\text{unknown}} = \{\mathbf{p}_u\}_{u=1}^{U}$. Each known prototype $\mathbf{p}_k$ is initialized as the feature centroid of its corresponding labeled samples. To initialize the unknown prototypes $\mathbf{p}_u$, we first exclude 15% of the unlabeled samples whose features are closest to any known prototype. We then apply K-means clustering on the remaining samples, utilizing the normalized cluster centroids as the initial $\mathbf{p}_u$.

For each unlabeled feature vector $\mathbf{f}$, let s_{k_1} and s_{k_2} denote its top-1 and top-2 cosine similarities with the known prototypes $\mathcal{P}_{\text{known}}$. Similarly, let s_{u_1} and s_{u_2} denote those with the unknown prototypes $\mathcal{P}_{\text{unknown}}$. We define the routing confidence scores as $c_k = s_{k_1} - s_{k_2}$ and $c_u = s_{u_1} - s_{u_2}$. The gating condition for routing the feature to $\mathcal{P}_{\text{known}}$ is formulated as:

$$g_k = \begin{cases} 1 & \text{if } s_{k_1} \geq \tau_{\text{sim},k} \text{ and } c_k \geq \tau_{\text{con},k} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\tau_{\text{sim},k}$ and $\tau_{\text{con},k}$ denote the similarity and confidence thresholds, respectively. The gate g_u for the $P_{unknown}$ follows the similar formulation. We adopt an asymmetric threshold design, assigning lower thresholds for $\mathcal{P}_{\text{unknown}}$ to counter the inherent feature-space bias toward known classes. When both $g_k = 1$ and $g_u = 1$, the sample is routed to $\mathcal{P}_{\text{unknown}}$ if $s_{u_1} + b \geq s_{k_1}$, where b is a bias. Under low-resource settings, this bias prevents cluster collapse and facilitates NEID. Once routed, each prototype is refined via an exponential moving average (EMA) to smoothly track the evolving feature distribution of its assigned samples.

Application of SACR. While the above routing rules allocate samples based on feature-space geometry, we introduce SACR to further refine these clusters at the semantic level. As Feng et al.'s (2024) illustrate, LLMs possess powerful text understanding capabilities that can significantly improve clustering accuracy.

We leverage an LLM during training to perform SACR. Specifically, the LLM takes representative utterances of each unknown cluster as input, optionally incorporating contextual utterances and speaker identities depending on the input strategy. Prompted to evaluate based on discourse functions, the LLM decides whether a cluster should be split, merged, or kept intact. These operations directly modify cluster assignments.

Specifically, for a merge decision, members of the involved clusters are aggregated into a single cluster, with its centroid recomputed as the mean of all members' features; for a split decision, the target cluster is partitioned into two sub-clusters via K-means ($K = 2$), which naturally yields the updated centroids.

For a training batch of Z samples, the SADA yields a total of $2Z$ projected representations. We construct a binary adjacency mask $\mathbf{M} \in \{0,1\}^{2Z \times 2Z}$, where $\mathbf{M}_{ij} = 1$ if representations i and j share a ground-truth label or are routed to the same prototype, and 0 otherwise. The joint instance-level and cluster-level contrastive loss is defined as:

$$\mathcal{L}_{\text{cl}} = -\sum_{i=1}^{2Z} \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \log \frac{\exp(\mathbf{f}_i^\top \mathbf{f}_j / \tau)}{\sum_{a \neq i} \exp(\mathbf{f}_i^\top \mathbf{f}_a / \tau)}, \tag{8}$$

where $\mathcal{N}(i) = \{j : \mathbf{M}_{ij} = 1, j \neq i\}$. Through this loss, sample i is pulled toward all samples with $\mathbf{M}_{ij} = 1$ and pushed away from those with $\mathbf{M}_{ij} = 0$. We additionally introduce a supervised prototype loss that aligns labeled samples with their respective prototypes:

$$\mathcal{L}_{\text{spl}} = -\sum_{i \in \mathcal{D}_l} \log \frac{\exp(\mathbf{f}_i^\top \mathbf{p}_{y_i} / \tau)}{\sum_{a=1}^{P} \exp(\mathbf{f}_i^\top \mathbf{p}_a / \tau)}, \tag{9}$$

where $\mathcal{D}_l$ is the set of labeled samples, $\mathbf{p}_{y_i}$ is the ground-truth prototype of sample i, P denotes the total number of prototypes.

For unlabeled samples, we apply a symmetric unsupervised prototype loss $\mathcal{L}_{\text{upl}}$. It shares an identical formulation to $\mathcal{L}_{\text{spl}}$, but operates over the unlabeled set $\mathcal{D}_u$ and aligns each sample with its routed prototype $\mathbf{p}_{r_i}$. Crucially, $\mathcal{L}_{\text{upl}}$ is scaled by a normalized confidence score $\tilde{c}_i = c_i / \sum_{j \in \mathcal{D}_u} c_j$. This weight down-weights the influence of low-confidence unlabeled samples on model training. The overall training objective is

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cl}} + \lambda_2 \mathcal{L}_{\text{spl}} + \lambda_3 \mathcal{L}_{\text{upl}}, \tag{10}$$

where λ_1, λ_2, and λ_3 denote the parameters.

After training, a K-means is applied to the features; each resulting cluster represents one educational intent. Unlabeled samples are either absorbed into known intents or form independent new clusters, signifying the discovery of new educational intents.

Method	Talkmoves				Mathdial			
	KCR=0.5 LAR=0.1	KCR=0.5 LAR=0.25	KCR=0.75 LAR=0.1	KCR=0.75 LAR=0.25	KCR=0.5 LAR=0.1	KCR=0.5 LAR=0.25	KCR=0.75 LAR=0.1	KCR=0.75 LAR=0.25
USNID (Zhang et al., 2023b)	37.27	36.07	41.34	45.00	44.31	46.18	50.05	47.60
MTP-CLNN (Zhang et al., 2022)	38.58	36.76	38.48	40.13	48.85	45.16	46.78	52.35
NEID (base)	47.56	50.21	48.91	52.28	49.19	50.14	51.97	52.68
NEID (full)	**59.39**	**61.89**	**68.04**	**70.36**	**49.54**	**50.42**	**52.86**	**53.61**

Table 1: ACC of baselines and the proposed NEID framework on the Talkmoves and Mathdial datasets. NEID (base) indicates $\mathcal{I}_{\text{mode}}$ = Single, $w = 0$ and $\mathcal{C}_{\text{refine}}$ = None. **NEID (full)** indicates $\mathcal{I}_{\text{mode}}$ = Context, $w = 1$ and $\mathcal{C}_{\text{refine}}$ = LLM. The best results are **bolded**, and the second-best are underlined.

5 Experimental Details

Datasets. We evaluated the proposed method on two educational datasets. Mathdial (Macina et al., 2023), a dataset of one-on-one teacher-student dialogues focused on mathematical reasoning. Talkmoves (Suresh et al., 2022), a large-scale dataset of transcripts from K-12 science classroom instruction. Following the approach of Zhang et al. (Zhang et al., 2022), we selected CLINC150 (Larson et al., 2019), which covers 150 intents across the general domain, as the external public intent dataset. In our experiments, utterances labeled as "none" were filtered out, but were retained as context during training.

Baselines. We evaluate our proposed framework against two baselines: **USNID** (Zhang et al., 2023b), a semi-supervised intent discovery framework aiming to capture the complex semantics of discrete text representations; and **MTP-CLNN** (Zhang et al., 2022), which learns semantic utterance representations from abundant in-domain and out-of-domain data while exploiting self-supervisory signals in unlabeled data for better clustering.

Experimental Design. We adopted the pre-trained bert-base-uncased model (Wolf et al., 2020) (110M parameters) as our base model. For clarity, we define two variables. The proportion of known intents is defined as $|C_k|/(|C_k| + |C_u|)$ and referred to as the "Known Class Ratio" (KCR), where C_k and C_u represent the sets of known and unknown classes, respectively. Additionally, the proportion of labeled examples for each known intent is denoted as the "Labeled Ratio" (LAR). The labeled data are randomly sampled from the original training set. We set KCR $\in \{50\%, 75\%\}$ and LAR $\in \{10\%, 25\%, 50\%\}$. For the data input strategy, we set $\mathcal{I}_{mode} \in \{\text{Single}, \text{Context}\}$ and use a binary variable $w \in \{0, 1\}$ to indicate whether speaker information is included. We extend the base prototypical contrastive learning by incorporating four cluster refinement strategies: $\mathcal{C}_{\text{refine}} \in \{\text{None}, \text{Random}, \text{Geometric}, \text{LLM}\}$.

Experimental Implementation. The noise ratio τ is kept consistent between F_T and F_S and is set to 35%, which proved to be the optimal ratio among $\{25\%, 30\%, 35\%, 40\%\}$. Across the three stages, the learning rates were set to 5e-5, 1e-3, and 1e-5, respectively. The default similarity and confidence ratios are 0.5 and 0.15 for known prototypes, and 0.35 and 0.10 for unknown prototypes. We employ GPT-5.4 as the LLM in our experiments. For model optimization, we used AdamW provided by Wolf et al. (Wolf et al., 2020). The batch sizes were chosen based on available GPU memory. All experiments were conducted on four NVIDIA RTX-4090 GPUs with distributed training. The total computational budget for a single complete training run (across all three stages and one random seed) is approximately 1.5 GPU hours. The results were averaged over 5 different random seeds.

Evaluation Metrics. We adopt three evaluation metrics for clustering: normalized mutual information (NMI), adjusted rand index (ARI), and accuracy (ACC).

6 Results and Analysis

6.1 Main Results

To align with typical low-resource settings in the NID task, the KCR and LAR are set to $\{50\%, 75\%\}$ and $\{10\%, 25\%\}$, respectively. As shown in Table 1, our NEID framework consistently outperforms all baselines in ACC across both datasets under four different settings. Taking the Talkmoves dataset as an example, the base framework achieves an average improvement of 9.82% over USNID. Notably, at KCR = 50% and LAR = 25%, ACC surges by 14.14%, highlighting the superiority of the framework when the proportion of known classes is limited.

Compared with MTP-CLNN, the base framework replaces the RTR augmentation (the top-performing strategy in MTP-CLNN) with SADA and adopts prototypical contrastive learning in place of traditional contrastive learning, better addressing the challenges of short utterances and imbalanced distributions inherent in educational data. This yields an average improvement of 11.25%.

Most notably, by incorporating +Speaker and +Context to optimize data input and applying SACR to refine cluster structures from a semantic perspective, our full framework achieves optimal performance, surpassing all baselines by average margins of 24.75% and 26.18%, respectively.

			Talkmoves		Mathdial	
	Spe	**Con**	**ACC**	**ARI**	**ACC**	**ARI**
KCR=0.5 LAR=0.1			47.56	22.36	52.28	17.89
	✓		59.17	39.15	65.89	19.20
		✓	49.72	26.53	64.04	18.65
	✓	✓	**59.19**	**41.63**	**70.38**	**19.54**
KCR=0.75 LAR=0.25			49.19	34.77	52.68	21.60
	✓		47.87	50.10	53.38	21.10
		✓	48.44	44.06	53.58	20.13
	✓	✓	**49.96**	**55.03**	**53.60**	**22.24**

Table 2: Performance comparison of four input strategies within the NEID framework across two datasets. *Spe* and *Con* denote the inclusion of speaker and context information, respectively. No cluster refimement strategy has been applied. The best results are **bolded**, and the second-best are underlined.

6.2 Input Strategy Utilities

As shown in Table 2, we evaluate four input strategies: Base ($\mathcal{I}_{\text{mode}} = \text{Single}, w = 0$), +Context ($\mathcal{I}_{\text{mode}} = \text{Context}, w = 0$), +Speaker ($\mathcal{I}_{\text{mode}} = \text{Single}, w = 1$), and +Both ($\mathcal{I}_{\text{mode}} = \text{Context}, w = 1$). The +Both strategy yields substantial gains across all settings. Under the restrictive setting of KCR = 50% and LAR = 10%, it achieves ACC improvements of 11.63% and 18.10% on Talkmoves and Mathdial, respectively, indicating superior robustness in low-resource scenarios.

The effectiveness of input strategies proved to vary by dataset.

- **Mathdial**: Characterized by high semantic density and notable similarity between teacher and student utterances, the model is highly sensitive to both strategies, especially at KCR = 50% and LAR = 10%. Introducing +Context or +Speaker effectively mitigates semantic homogeneity, empowering the model to capture underlying intent variations.
- **Talkmoves**: Featuring short and fragmented utterances typical of K-12 classrooms, +Speaker serves as a crucial disambiguation cue for NEID. Specifically, at KCR = 50% and LAR = 10%, +Speaker alone yields a 9.61% ACC gain, whereas +Context provides relatively marginal improvement.

Overall, regardless of the dataset, +Both consistently leverages the complementary strengths of speaker and context information, yielding the best performance.

		Talkmoves		
	Strategy	ACC	ARI	NMI
KCR=0.5 LAR=0.1	Base	59.19	38.63	47.75
	+Random	59.15	41.22	51.83
	+Geometric	**59.43**	40.56	49.76
	+LLM	59.39	**41.26**	**52.34**
KCR=0.5 LAR=0.25	Base	62.08	47.17	53.68
	+Random	60.96	42.81	52.03
	+Geometric	**62.82**	**47.46**	**54.30**
	+LLM	61.89	47.12	53.70
KCR=0.75 LAR=0.1	Base	67.75	50.71	55.74
	+Random	67.41	50.62	55.79
	+Geometric	65.66	**51.36**	55.05
	+LLM	**68.04**	50.68	**56.85**
KCR=0.75 LAR=0.25	Base	70.38	55.03	61.34
	+Random	70.34	56.64	61.84
	+Geometric	**70.84**	57.47	62.38
	+LLM	70.36	**57.57**	**62.81**

Table 3: Performance comparison of different cluster refinement strategies within the NEID framework. *Base*: no cluster structure optimization is applied. *+Random*: keep, merge, or split operations are applied randomly. *+Geometric*: operations are determined based on geometric structure. *+LLM*: a LLM decides and executes the operations. The parameters are set to $\mathcal{I}_{\text{mode}} = \text{Context}$ and $w = 1$. The best results are **bolded**.

6.3 LLM Utilities

Table 3 reports the performance of four refinement strategies on Talkmoves under four different settings. We highlight three key observations:

First, +Random improves NMI and ARI over the base model in two settings (e.g., NMI increases by 4.08 at KCR=50% and LAR=10%), yet degrades ACC. This suggests that the initial K-means clusters are suboptimal and even random reorganization improves the overall cluster structure. The ACC drop indicates that random perturbations destroy

the strict cluster-to-label alignment required for downstream evaluation.

Second, +Geometric achieves the best ACC in three settings, operating with a merge threshold of centroid cosine similarity$\geq$0.85 and a split threshold of top-25% intra-cluster variance. These geometric signals prove effective when strong supervision yields a well-structured feature space. However, under weak supervision (LAR=10%, KCR=75%), +Geometric tends to over-split small, compact clusters, pulling ACC down to 65.66%.

Third, +LLM is the strongest strategy for structural metrics. It achieves the best NMI in three of four settings and the best ARI in two of four, while maintaining the base ACC. This stems from the LLM's ability to regroup utterances based on actual discourse function rather than predefined labels, capturing semantic patterns that geometric rules miss.

In summary, the necessity of cluster reorganization is confirmed by +Random's structural improvements over the base model. +Geometric excels under strong supervision but struggles in low-resource settings. +LLM uniquely enhances overall cluster structure without sacrificing label alignment, making SACR the most robust strategy.

6.3.1 LLM Decision

Context	True Label
[T] rory *[S] what if you started out with bug and tried to smaller* [T] that would work	Providing Evidence
[S] two thirds are bigger *[S] but you have to do it like this* [S] ok here we go	Providing Evidence
[S] a fourth *[S] a fourth* [S] so then it's just a to the fourth	Making a Claim
LLM Rationale: The representative utterances of this cluster contain two distinct student discourse functions: proposing problem-solving strategy and stating fractional answers. This cluster mixes unrelated functions and needs to be split into separate clusters for each distinct function.	
LLM Decision: Split (confidence: strong, 3/3 votes)	

Table 4: A representative LLM cluster-refinement decision on the Talkmoves dataset. The context row shows the target utterance (in *italics*) alongside its preceding and following utterances. [S] and [T] indicate student and teacher speakers, respectively. The parameters are $\mathcal{I}_{\text{mode}} = \text{Context}$ and $w = 1$.

Table 4 shows an example of LLM-based semantic cluster refinement. Given representative utterances from a cluster, the LLM identifies a mixture of distinct educational intents: "proposing problem-solving strategy" aligns with the true label *providing evidence*, while "stating fractional answers" aligns with *making a claim*. Since three independent LLM calls yield the same conclusions, the confidence is rated as "strong", leading to a "split" decision. As training stabilizes, all decisions converge to "keep", at which point the LLM rationale can assist educators in naming the newly discovered educational intents.

7 Conclusion

This study introduces the NEID framework to tackle class imbalance and high noise in educational datasets, helping educators adapt coding schemes for diverse educational scenarios. Specifically, we propose the structure-aware Integrated Gradients algorithm to accommodate input structures enriched with context and speaker information. By introducing a knowledge distillation method, we transfer attribution capabilities from F_T (trained on labeled data) to F_S, empowering the latter to identify general semantic noise in semi-supervised scenarios. Furthermore, SADA and SACR are introduced for data augmentation and LLM-driven cluster refinement, respectively, strengthening prototypical contrastive learning while preserving core semantics. Extensive experiments on two educational datasets demonstrate the effectiveness of our proposed framework.

Limitations

A key limitation of our NEID framework is the dataset-dependent effectiveness of SACR. While SACR achieves substantial improvements on Talkmoves, its benefits diminish significantly on Mathdial, where the None strategy achieves the best ACC, NMI, and ARI in most settings. We attribute this to three dataset-specific properties: (1) its restricted vocabulary causes severe semantic overlap across intents; (2) high linguistic similarity between teacher and student utterances weakens role signals; and (3) dialogue variation is driven by continuous reasoning steps rather than categorical intent shifts. These properties push the base K-means structure close to the performance ceiling, and any reorganization likely introduces additional noise.

Therefore, our method yields greater benefits on corpora featuring diverse and role-asymmetric interactions, such as Talkmoves. Future work will explore three directions: (1) incorporating more

diverse samples into LLM inputs to improve representativeness; (2) integrating cross-epoch decision memory to reduce label oscillation; and (3) examining how different LLM backbones influence cluster refinement quality. We aim to unlock further performance potential of NEID.

References

Alaa T Al-Tuama and Dhamyaa A Nasrawi. 2022. Intent classification using machine learning algorithms and augmented data. In *2022 International Conference on Data Science and Intelligent Computing*, pages 234–239.

Ella Cutler, Zachary Levonian, and S Thomas Christie. 2025. Detecting student intent for chat-based intelligent tutoring systems. *arXiv preprint arXiv:2502.15096*, pages 1–6.

Yimin Deng, Yuxia Wu, Guoshuai Zhao, Li Zhu, and Xueming Qian. 2024. Pseudo-label enhanced prototypical contrastive learning for uniformed intent discovery. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7549–7562.

Zijin Feng, Luyang Lin, Lingzhi Wang, Hong Cheng, and Kam-Fai Wong. 2024. LLMEdgeRefine: Enhancing text clustering with LLM-based boundary point refinement. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18455–18462.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, and et al. Tang, Lingjia. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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 1311–1316.

Zachary Levonian, Owen Henkel, Chenglu Li, and et al. Postle, Millie-Ellen. 2025. Designing safe and relevant generative chats for math learning in intelligent tutoring systems. *Journal of Educational Data Mining*, 17(1):66–97.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Yuzhong Chen, Xi Jiang, David Liu, and et al. Zhu, Dajiang. 2024. Mask-guided bert for few-shot text classification. *Neurocomputing*, 610:128576.

Jionghao Lin, Wei Tan, Lan Du, Wray Buntine, David Lang, Dragan Gašević, and Guanliang Chen. 2023a. Enhancing educational dialogue act classification with discourse context and sample informativeness. *IEEE Transactions on Learning Technologies*, 17:258–269.

Jionghao Lin, Wei Tan, Ngoc Dang Nguyen, David Lang, Lan Du, Wray Buntine, Richard Beare, Guanliang Chen, and Dragan Gašević. 2023b. Robust educational dialogue act classifiers with low-resource and imbalanced datasets. In *International Conference on Artificial Intelligence in Education*, pages 114–125.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621.

Yutao Mou, Keqing He, Yanan Wu, Pei Wang, Jingang Wang, Wei Wu, Yi Huang, Junlan Feng, and Weiran Xu. 2022. Generalized intent discovery: Learning from open world dialogue system. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 707–720.

Eda Okur, Roddy Fuentes Alba, Saurav Sahay, and Lama Nachman. 2023. Assessing spoken language understanding pipeline of a multimodal dialogue system for kids learning math at home. In *Proceedings of the Neural Conversational AI Workshop-What's left to TEACH chatbots*, pages 1–18.

Changyong Qi, Longwei Zheng, Yuang Wei, Haoxin Xu, Peiji Chen, and Xiaoqing Gu. 2024. Edudcm: A novel framework for automatic educational dialogue classification dataset construction via distant supervision and large language models. *Applied Sciences*, 15(1):154.

Juan A Rodriguez, Nicholas Botzer, David Vazquez, Christopher Pal, Marco Pedersoli, and Issam Laradji. 2024. Intentgpt: Few-shot intent discovery with large language models. *arXiv preprint arXiv:2411.10670*.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press, Cambridge, England.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H Martin, and Tamara Sumner. 2022. The talkmoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 4654–4662.

Rob Van Der Goot, Max Müller-Eberstein, and Barbara Plank. 2022. Frustratingly easy performance improvements for low-resource setups: A tale on bert and segment embeddings. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1418–1427.

Deliang Wang and Gaowei Chen. 2024. Making ai accessible for stem teachers: Using explainable ai for unpacking classroom discourse analysis. *IEEE Transactions on Education*, 67(6):907–918.

Deliang Wang, Dapeng Shan, Ran Ju, Ben Kao, Chenwei Zhang, and Gaowei Chen. 2025. Investigating dialogic interaction in k12 online one-on-one mathematics tutoring using ai and sequence mining techniques. *Education and Information Technologies*, 30(7):9215–9240.

Deliang Wang, Yang Tao, and Gaowei Chen. 2024. Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, 123:102275.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, and et al. Funtowicz, Morgan. 2020. Huggingface's transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Di Wu, Liming Feng, and Xiaoyu Wang. 2026. New intent discovery with syntactic structure masking pre-training and density-aware contrastive learning. *Neurocomputing*, 659:131760.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2023a. A clustering framework for unsupervised and semi-supervised new intent discovery. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5468–5481.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2023b. Usnid: A framework for unsupervised and semi-supervised new intent discovery. *CoRR, abs/2304.07699*.

Shun Zhang, Yan Chaoran, Jian Yang, Jiaheng Liu, Ying Mo, Jiaqi Bai, Tongliang Li, and Zhoujun Li. 2024. Towards real-world scenario: Imbalanced new intent discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3949–3963.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023c. Clusterllm: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2022. New intent discovery with pre-training and contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 256—269.

A Analysis of Noise Threshold

To determine the optimal noise threshold τ, we conducted a sensitivity analysis within the range of $\{0.25, 0.30, 0.35, 0.40\}$. The results identify $\tau = 0.35$ as the optimal balance between task difficulty and semantic integrity. Specifically, the model exhibits a distinct rise-and-fall trend in ACC and ARI, peaking at $\tau = 0.35$, while NMI remains relatively robust with a flatter trend. Consequently, we set $\tau = 0.35$ to achieve optimal generalization performance across all experiments.

B Cluster Packet Examples for SACR

To illustrate the LLM input format during Semantic-Aware Cluster Refinement (SACR), we present a representative cluster from an intermediate training epoch on the Talkmoves dataset, specifically displaying only four sampled utterances within this single cluster. Tables 5 and 6 detail the packets constructed under the Base ($\mathcal{I}_{mode} = \textit{Single}$, $w = 0$) and Full ($\mathcal{I}_{mode} = \textit{Context}$, $w = 1$) input strategies, respectively.

Notably, when $\mathcal{I}_{mode} = \textit{Context}$, if a target utterance appears at the dialogue's boundary, its preceding or following utterance is left empty (e.g., utterance 4 in Table 6).

Attribute	Content
Fingerprint	`b3860647fb78b073553616267ba35653 23507246bbd224d0700921ae95427407`
Dataset	Talkmoves (Seed: 44)
Progress	Current Epoch: 20 / Total Epoch: 50

Cluster ID: 0
Representative:

1.Utterance: i had to know that you needed to just round the hundredths actually
2.Utterance: yes but mrs reevess are on the top this time
3.Utterance: and i just put it in the tens place
4.Utterance: normally you subtract what youre adding

Table 5: Example cluster packet for LLM-based SACR under the Base input strategy ($\mathcal{I}_{mode} = \textit{Single}$, $w = 0$). Each representative sample contains only the target utterance, without speaker tokens or contextual utterances.

C Known-Category Guidance Prompt

To enhance intent discovery, we append a known-category guidance block to the main SACR prompt. By outlining existing discourse categories and examples, this block establishes a semantic reference frame that enables the LLM to make informed refinement decisions. Table 7 details this guidance prompt for the Talkmoves dataset.

Attribute	Content
Fingerprint	`b3860647fb78b073553616267ba35653 23507246bbd224d0700921ae95427407`
Dataset	Talkmoves (Seed: 44)
Progress	Current Epoch: 20 / Total Epoch: 50

Cluster ID: 1
Representative:

1. Utterance: [T] you see what that means
Context: [PREV] [T] they used adapt and change so you are adapting the math problem by changing the alignment | [CURR] *[T] you see what that means* | [NEXT] [T] okay

2. Utterance: [T] you get the picture
Context: [PREV] [S] oh right | [CURR] *[T] you get the picture* | [NEXT] [T] so earlier you calculated the surface area first then you used the multiplication table

3. Utterance: [T] whos got direction for hospitals to city hall
Context: [PREV] [T] okay how about hospital to city hall | [CURR] *[T] whos got direction for hospitals to city hall* | [NEXT] [T] maxine

4. Utterance: [T] 16 dots right
Context: [CURR] *[T] 16 dots right* | [NEXT] [S] and theres another square in the middle of it

Table 6: Example cluster packet for LLM-based SACR under the Full input strategy ($\mathcal{I}_{mode} = \textit{Context}$, $w = 1$). Each representative sample includes the target utterance along with its preceding and following utterances, all prefixed with speaker tokens.

D LLM Prompt for SACR

We design a structured prompt to guide the LLM in evaluating clusters during Semantic-Aware Cluster Refinement (SACR). Based on underlying discourse functions, the LLM decides whether each cluster should be *kept*, *merged*, or *split*. Table 9 presents the prompt for the Full input strategy ($\mathcal{I}_{mode} = \textit{Context}$, $w = 1$).

E LLM Output Format

To ensure the LLM's refinement decisions are automatically parsable, we append a structured output specification to the SACR prompt. As detailed in Table 8, this block defines the allowable decisions (*keep*, *merge*, *split*), required fields, and execution rules for inter-cluster consistency. By mandating a unified JSON schema and explicit evidence-based summaries, we guarantee that all LLM responses can be directly integrated into the downstream cluster update procedure (Section 4.4) without ambiguity.

Prompt Instruction
For reference only — the following are KNOWN discourse categories already labeled in the dataset. Do NOT assign any cluster to these categories; they are provided solely to help you understand what discourse functions have already been identified, so you can focus on discovering NEW ones.
Known Categories Provided
- **Keeping Everyone Together**: Does everyone see that?; Are you listening?; All eyes on me - **Getting Students to Relate**: Do you agree with what she said?; Can you build on his idea?; What do you think about what Jamie said? - **Restating**: So you're saying the area is 24; You think we need to divide by 2; So the answer is 12

Table 7: The known-category guidance appended to the SACR prompt for the Talkmoves dataset. By detailing existing discourse categories, it serves as a semantic reference to help the LLM accurately evaluate the intents of clusters.

Instruction	**Content**
Decisions	`keep`, `merge`, `split`
Required	`cluster_id`, `decision`, `target_cluster_id`, `summary`
Notes	**1.** Operate at the cluster level only. **2.** Use `target_cluster_id` only for `merge`; set to `null` for `keep` and `split`. **3.** If the decision is `merge`, the `decision` field of both clusters must be set to `merge`, and the `target_cluster_id` must be marked with the other's `cluster_id`. **4.** The `summary` field must explicitly state the strongest supporting evidence for this decision. **5.** Return a JSON array only. No text outside the JSON.

Table 8: Output format constraints for the SACR prompt. These rules ensure the LLM generates parseable, evidence-grounded refinement decisions within a unified JSON schema.

Prompt Section	Content
System Role	You are an expert in classroom discourse analysis, responsible for refining the clustering results of teacher and student utterances from classrooms.
Core Rules	**1.** Each cluster must correspond to exactly ONE coherent discourse function; clusters that fail to meet this requirement must be corrected via "merge" or "split". **2.** For all utterances provided to you, the "utterance" field refers to the current utterance, and the "context" field indicates the position of the current utterance in the dialogue context. **3.** You must analyze the utterance based on its discourse function in context. The same utterance may have different discourse functions in different contexts. **4.** "[s]" denotes a student utterance and "[t]" denotes a teacher utterance.
Input Fields	**1. cluster_id**: Cluster identifier. **2. representative-utterance**: The samples closest to the cluster center with the highest typicality. **3. representative-context**: The representative utterance concatenated with its preceding and following utterances.
Decision Rules	**1. keep**: The cluster corresponds to only one clear discourse function and has distinct functional differences from neighboring clusters. **2. merge**: Two clusters correspond to identical or highly similar discourse functions. **3. split**: The cluster clearly mixes two or more distinct discourse functions.
Decision Basis	**1.** If the "representative utterances" of a cluster obviously contain more than one distinct discourse function, this shall be regarded as strong evidence for "split". **2.** Compare "representative utterances" across all other clusters; consistent discourse functions of "representative utterances" shall be regarded as strong evidence for "merge". **3.** If neither of the above two conditions is satisfied, set the decision to "keep".
Requirements	**1.** Judgments based merely on superficial wording are prohibited; decisions must focus on the underlying discourse function. **2.** Different speaker roles can be taken as a reference for discourse function judgment, but "split" cannot be adopted merely because of different speakers. **3.** Defaulting to an all-keep decision is prohibited. The all-keep solution is valid when every cluster features unified internal discourse functions and clear functional differentiation from neighboring clusters. **4.** You must have clear evidence for your judgment.

Table 9: The complete LLM prompt template for SACR under the Full input strategy ($\mathcal{I}_{mode} = \textit{Context}$, $w = 1$). It directs the LLM to evaluate representative utterances per cluster and output a refinement decision (*keep*, *merge*, or *split*) alongside its rationale.