Pseudo-Label Enhanced Prototypical Contrastive Learning
for Uniformed Intent Discovery

Yimin Deng[1,2]*, Yuxia Wu[3]*, Guoshuai Zhao[1]†, Li Zhu[1]†, Xueming Qian[4]

[1]School of Software Engineering, Xi'an Jiaotong University
[2]City University of Hong Kong, [3]Singapore Management University
[4]School of Information and Communication Engineering, Xi'an Jiaotong University
dymanne@stu.xjtu.edu.cn, yieshah2017@gmail.com
{guoshuai.zhao,zhuli}@xjtu.edu.cn, qianxm@mail.xjtu.edu.cn

Abstract

New intent discovery is a crucial capability for task-oriented dialogue systems. Existing methods focus on transferring in-domain (IND) prior knowledge to out-of-domain (OOD) data through pre-training and clustering stages. They either handle the two processes in a pipeline manner, which exhibits a gap between intent representation and clustering process or use typical contrastive clustering that overlooks the potential supervised signals from the whole data. Besides, they often individually deal with open intent discovery or OOD settings. To this end, we propose a **P**seudo-**L**abel enhanced **P**rototypical **C**ontrastive **L**earning (**PLPCL**) model for uniformed intent discovery. We iteratively utilize pseudo-labels to explore potential positive/negative samples for contrastive learning and bridge the gap between representation and clustering. To enable better knowledge transfer, we design a prototype learning method integrating the supervised and pseudo signals from IND and OOD samples. In addition, our method has been proven effective in two different settings of discovering new intents. Experiments on three benchmark datasets and two task settings demonstrate the effectiveness of our approach.[1]

1 Introduction

Task-oriented conversational systems are typically designed to assist users in completing specific tasks by leveraging predefined ontologies (*e.g.,* intents, slots and values), which face limitations in dynamic environments where novel ontology categories may emerge (Liang et al., 2024a; Mou et al., 2022a; Wu et al., 2024, 2022c; Zhang et al., 2021b). New intent discovery, which aims to uncover and categorize out-of-domain intents absent from the

---

* Co-first authors with equal contribution.
† Corresponding authors.
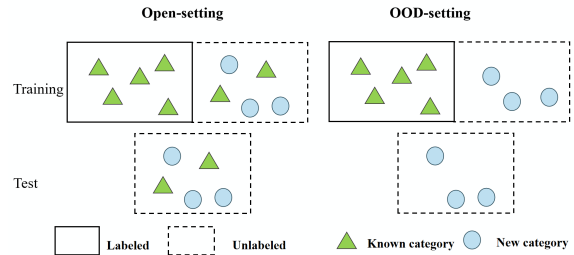[1]The codes and datasets are available at https://github.com/dymanne123/PLPCL

Figure 1: Two basic task settings for uniformed intent discovery. Open-setting: Partially labeled IND data is used for training, and the test data includes both IND and OOD categories. OOD-setting: Fully labeled IND data is used for training, while the test data contains only OOD categories.

training data, has received increasing attention due to its crucial role in dialogue systems (Min et al., 2021; Vedula et al., 2019). Initially, researchers focused on exploring unsupervised clustering methods (Hakkani-Tür et al., 2015; Liu et al., 2021b; Shi et al., 2018). However, real-world scenarios often involve limited labeled data, prompting a shift toward semi-supervised approaches, notably OOD and open intent discovery (Liang and Liao, 2023; Liang et al., 2024b; Lin et al., 2020; Mou et al., 2022a,c; Shen et al., 2021; Zhang et al., 2021b, 2022). OOD intent discovery involves clustering unlabeled OOD intents by utilizing labeled IND data, focusing solely on identifying novel categories absent in the training set. In contrast, open intent discovery seeks to simultaneously recognize both known and new categories from unlabeled data, allowing for a more comprehensive understanding of both in-domain and out-of-domain intents (see Figure 1).

The existing approaches for new intent discovery typically use a two-stage contrastive clustering approach, involving IND pre-training and OOD clustering. For open intent discovery, researchers focus on effectively utilizing the small amount of labeled data for weakly supervised (Lin et al.,

2020) or semi-supervised clustering methods (Shen et al., 2021; Zhang et al., 2021b). For OOD intent discovery, previous works commonly employ a contrastive clustering framework (Li et al., 2021) with approaches such as multi-head contrast learning (Mou et al., 2022c) or neighbor-enhanced contrastive strategies (Mou et al., 2022a).

Although previous methods in intent discovery have achieved notable success, several challenges in the field remain unexplored. (1) One key challenge is efficiently integrating labeled and unlabeled data for joint representation learning and clustering. While some methods focus on contrastive clustering for joint learning (Mou et al., 2022c), they often overlook critical supervised signals from IND samples during the clustering stage. Leveraging labeled information for this process remains underexplored, particularly given the limited availability of labeled data and the risk of overfitting. (2) Another critical issue is devising effective transfer learning mechanisms between IND and OOD data while preventing catastrophic forgetting. Existing methods often discard classifiers trained on prior knowledge, retaining only feature extraction during clustering with OOD samples (Zhang et al., 2021b). This requires additional alignment strategies, potentially introducing noise if suboptimal. There's an urgent need for methods preserving prior knowledge while adapting better to new intent data, ensuring seamless transfer learning. (3) Moreover, the predominant focus of prior research has been either on open intent discovery or OOD intent discovery individually, disregarding the practical need for a unified method capable of handling both scenarios (Mou et al., 2022c; Zhang et al., 2021b). Real-world dialogue systems often encounter situations requiring updates or migration, underscoring the need for a uniform intent discovery approach that can adapt to various system changes.

To address these limitations, we propose a **P**seudo-**L**abel enhanced **P**rotypical **C**ontrastive **L**earning (PLPCL) model which is built upon contrastive clustering for joint representation learning and clustering. Our approach begins by pretraining the contrastive clustering model using labeled IND data. To effectively harness labeled information, we integrate IND samples with unlabeled data using a semi-supervised clustering strategy, employing distinct contrastive learning strategies for labeled and unlabeled data. To prevent overfitting and maximize the utilization of unlabeled data, we iteratively select reliable unlabeled

samples with confident pseudo-labels. These reliable samples serve as potential positive/negative samples during contrastive learning, enhancing the overall contrastive clustering process.

To bridge the gap between IND and OOD data, we introduce a prototype learning strategy. It maintains the prototype matrix by integrating instance and cluster features from both the IND and reliable unlabeled samples. Our method integrates contrastive clustering and prototypical learning, eliminating the need for an extra aligning module. This integration facilitates improved knowledge transfer from IND to OOD without discarding label information from IND samples. Furthermore, our framework is purposefully designed for uniform intent discovery, demonstrating effectiveness in both open-setting and OOD-setting scenarios.

The contribution of our work is threefold: (1) We introduce a novel method that leverages both labeled and unlabeled data through pseudo-label enhanced semi-supervised contrastive learning. This approach facilitates joint representation learning and clustering by effectively leveraging the whole data. (2) We propose a prototypical contrastive learning model for uniformed intent discovery integrating prototypical learning and contrastive clustering to bridge the gap between IND prior knowledge and OOD categories. (3) We conduct extensive experiments on both OOD and open intent discovery scenarios and the results demonstrate the effectiveness of our proposed method.

2 Related Work

2.1 Intent Discovery

Recent research for intent discovery can be broadly categorized into OOD-setting and open-setting. As shown in Figure 1, open intent discovery involves clustering both IND and OOD intents with IND priori knowledge. Samples with IND intents are not all labeled. Whereas OOD intent discovery focuses on accurately handling OOD intents and assumes that the intents of labeled and unlabeled data do not overlap, which means all IND samples are labeled. Lin et al. (2020) proposed a self-supervised clustering method that utilizes limited labeled data. Zhang et al. (2021b) proposed a k-means-based semi-supervised clustering method that can effectively use prior knowledge in intent discovery. Mou et al. (2022c) proposed a disentangled contrastive learning framework that mainly focuses on OOD intent clustering and decouples instance and cluster-

	Joint representation and cluster	Prior knowledge retention	Reuse of labeled data	OOD setting	Open setting
CDAC+ (Lin et al., 2020)	✓	pair-wise similarity	✓	×	✓
DeepAligned (Zhang et al., 2021b)	×	representation	×	×	✓
DKT (Mou et al., 2022c)	✓	representation	×	✓	×
DPN (An et al., 2023)	×	prototype	✓	×	✓
PLPCL (Ours)	✓	representation, prototype, classifier	✓	✓	✓

Table 1: The differences between our method and prior works

level features to disentangle the knowledge of IND and OOD samples. Han et al. (2019) extended deep embedded clustering to transfer learning setting, incorporating prior knowledge for OOD clustering.

2.2 Contrastive Clustering

Contrastive clustering has been widely used in various clustering scenarios, such as unsupervised semantic segmentation (Hamilton et al., 2022) and generalized self-supervised contrastive learning (Hu et al., 2023). It has also been applied in OOD intent detection and discovery tasks (Kumar et al., 2022; Mou et al., 2022a,c). Li et al. (2021) proposed a contrastive clustering framework with two contrastive learning heads. It provided objective guidance for clustering, avoiding interference from prior knowledge. Mou et al. (2022c) extended contrast clustering to the semi-supervised scenario and designed a two-stage contrastive learning process that includes both supervised pre-training and unsupervised clustering. It achieved state-of-the-art results for OOD intent discovery.

2.3 Prototype Learning.

Prototype learning methods are widely used in clustering analysis and classification problems. In semi-supervised clustering scenarios, coarsely assigned pseudo-labels may result in mismatches between instances and prototypes, introducing noise that significantly affects the clustering performance (An et al., 2023; Huang et al., 2022; Wu et al., 2022b). An et al. (2023) used weighted pseudo-labels to reduce the effect of mismatched prototypes. Huang et al. (2022) proposed the approach of prototype scattering, which enhances the variance between the clusters by maximizing the distances between prototype features, to obtain well-separated clusters. The prototype learning method is robust to noise and outliers. Compared to other clustering methods, it is an intuitive and interpretable approach that can provide references for the entire cluster based on representative examples.

2.4 Novelty Analysis of Our Method

We summarize the differences between our method and prior works and highlight the motivations and advantages of our method in terms of joint representation and clustering, prior knowledge retention, reuse of labeled data, and handling of OOD and open settings in Table 1.

3 Preliminaries

3.1 Problem Statement

Uniformed intent discovery includes both OOD intent discovery and open intent discovery. OOD intent discovery assumes we have a set of labeled IND data D_{IND} and unlabeled OOD data D_{OOD}, with the goal of clustering OOD intents. Notably, there is no overlap between the IND and OOD data. Open intent discovery involves an intent analysis dataset $\{D_l, D_u\}$, where $D_l = \{(x_l, y_l)|y_l \in \mathcal{Y}_k\}$ represents labeled data for known intents, and $D_u = \{(x_u, y_u)|y_u \in \mathcal{Y}_k, \mathcal{Y}_{uk}\}$ includes both known ($\mathcal{Y}_k$) and unknown intents ($\mathcal{Y}_{uk}$). In the extreme case, all samples of the known categories in the training set are labeled, and D_u contains only data with unknown intents, making the prior labeling information same with OOD classification. Since we do not have a priori assumptions that $y_l \cap y_u = \emptyset$, this setting differs from the generalized intent discovery (Mou et al., 2022b).

3.2 Contrastive Clustering

Our model is based on a contrastive clustering framework. It performs instance-level and cluster-level contrastive learning. Specifically, positive/negative instance pairs are formed via data augmentation, followed by their projection into a feature space. Instance-level and cluster-level contrastive learning are performed in the row space and column space, respectively.

Unsupervised instance-level contrastive learning (ILCL) is performed on unlabeled data, where the augmented sample of each sample is considered as a positive sample and other samples are considered as negative samples. f_i, f_j refer to the augmented samples that are generated from the same samples
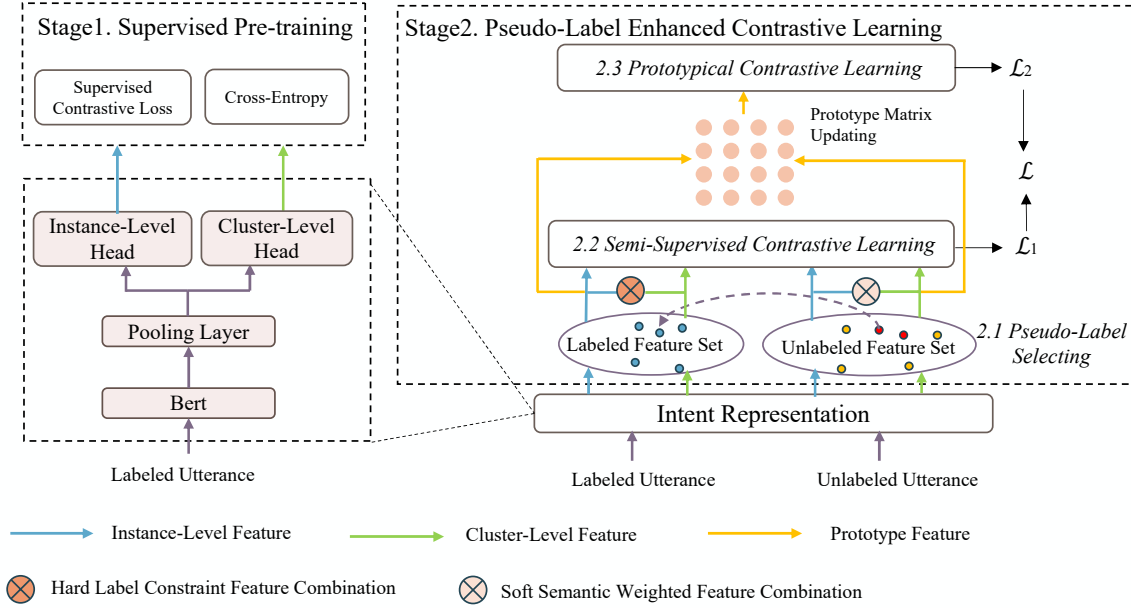
Figure 2: The overall architecture of the proposed PLPCL. (a) Intent representation is achieved based on disentangled instance-level and cluster-level heads. (b) Stage 1 involves supervised contrastive learning on instance-level representation and classification on cluster-level representation. (c) Stage 2 starts from pseudo-label selecting for unlabeled data followed by semi-supervised and prototypical contrastive learning in an iterative manner.

after passing through the dropout layer.

$$\ell_{i,j}^{ins} = -log \frac{\exp\left(sim(f_i, f_j)/\tau\right)}{\sum_{k=1}^{2N} 1_{i \neq k} \exp\left(sim(f_i, f_k)/\tau\right)} \quad (1)$$

On the cluster-level contrastive learning head g, it performs cluster-level contrastive learning (CLCL). The cluster representation of the augment sample is considered as a positive sample, and the other cluster representations are considered as negative samples. y_i refers to the representations of the clusters, which are columns in the cluster-level feature matrix. y_j are the dropout-augment representations for the cluster level.

$$\ell_{i,j}^{clu} = -log \frac{\exp\left(sim(y_i, y_j)/\tau\right)}{\sum_{k=1}^{2N} 1_{i \neq k} \exp\left(sim(y_i, y_k)/\tau\right)} \quad (2)$$

4 Methods

The overall framework of PLPCL is illustrated in Figure 2. Our framework follows a two-stage pipeline. In the pre-train stage, IND-labeled samples are utilized for supervised multi-head contrastive learning to acquire prior knowledge. This model is adapted from contrastive clustering and includes intent representation alongside two independent heads. These heads are instrumental in

decoupling the representation into instance-level and cluster-level spaces, facilitating joint representation learning and clustering.

After pretraining, the prototypes of known categories are obtained, serving as a foundation for efficient knowledge transfer across IND and OOD data. In the second stage, the multi-head contrastive model is further trained on the entire dataset including IND and OOD samples. Specifically, this stage comprises three iterative steps: pseudo-label selecting, semi-supervised contrastive learning, and prototypical contrastive learning. These steps collectively aim to transfer prior knowledge to new categories and enhance the model's adaptability.

4.1 Intent Representation

To facilitate effective knowledge transfer between IND and OOD samples, we aim to achieve joint intent representation and clustering by learning instance-level and cluster-level representations.

Drawing inspiration from (Mou et al., 2022c), we first extract the intent representation using a pre-trained BERT model and a pooling layer to extract text representation. Then we utilize two independent MLPs to map the intent representation z_i into two disentangled latent vectors: $f_i = f(z_i)$ and $g_i = g(z_i)$.

4.2 Supervised Pre-training

To familiarize the model with the prior knowledge obtained from labeled IND samples and to establish initial cluster prototypes, we conduct pre-training on IND samples. Based on the multi-level intent representation, we conduct two-level pre-training.

For instance-level representation, we adopt supervised contrastive learning (SCL) to maximize inter-class variance and minimize intra-class variance within the IND samples.

Formally, for a sample x_i in a mini-batch of size N, the samples within N sharing the same label are considered as positive samples, while the remainder is treated as negative samples. The SCL loss is computed as follows:

$$
\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{|N_{y_i} - 1|} \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \mathbb{1}_{y_i = y_j}
$$
$$
\log \frac{\exp\left(f_i \cdot f_j / \tau\right)}{\sum_{k=1}^{N} \mathbb{1}_{i \neq k} \exp\left(f_i \cdot f_k / \tau\right)} \tag{3}
$$

where y_i, y_j are the labels of samples x_i, x_j and $\mathbb{1}$ is an indicator function. N_{y_i} denotes the number of samples in N with the label y_i. f_i, f_j indicate the instance-level representation of x_i, x_j. τ is the temperature parameter for contrastive learning.

For the cluster-level representation, we apply cross-entropy loss (CE) to learn cluster-friendly features. Note that we use a classifier with both IND and OOD classes for open-setting to better reserve the priori knowledge extracted at this stage, allowing our model to better retain the prior knowledge acquired during the pre-training stage compared to previous works.

4.3 Semi-supervised Training

After pretraining with labeled data, we achieve a good initialization of representation learning and clustering. The challenge now lies in transferring the prior knowledge to new intents. There are two critical problems to be addressed: (1) Effectively utilizing both labeled and unlabeled data to enhance the joint representation and clustering process; (2) Transferring learned knowledge from IND to OOD data while continually refining representation learning to enhance cluster-friendly features without encountering catastrophic forgetting. To tackle these challenges, we introduce a pseudo-label enhanced contrastive learning scheme tailored for iterative clustering and updating. This scheme starts from reliable pseudo-label filtering for unlabeled samples, followed by semi-supervised contrastive learning and prototypical contrastive learning.

4.3.1 Pseudo-label Selecting

It's important to note that the multi-level intent representation heads are pre-trained on limited labeled data. To fully leverage the valuable information embedded within unlabeled data, we employ pseudo-labeling techniques to iteratively select unlabeled samples as weak supervised signals for subsequent contrastive learning processes.

For semi-supervised contrastive learning (4.3.2), reliable pseudo-labeled data are amalgamated with labeled data, augmenting potential contrastive samples. Regarding prototypical contrastive learning (4.3.3), pseudo-labels are employed to enrich the learning process of the prototype matrix. Furthermore, the integration of pseudo-labeled data introduces supplementary constraints to mitigate overfitting and enhance the model's generalization performance to novel unseen categories.

However, it's important to note that the quality of pseudo-labels is crucial, as noisy or incorrect pseudo-labels can degrade model performance. It is crucial to ensure the accuracy and reliability of pseudo-labels to maintain the effectiveness of the model's training. We treat pseudo-labels with probabilities greater than a confidence threshold as true labels and use them as supervised signals to guide model training:

$$
p(k|\mathbf{x}) > \sigma \tag{4}
$$

where $p(k|\mathbf{x})$ denotes the probability that x belongs to class k, σ represents confidence probability. When a certain probability is greater than the confidence threshold, we have enough confidence to consider it as belonging to this category. The choice of confidence probability will impact the strength of the supervised signal and the introduced noise. A high confidence probability will result in inadequate supervised information, whereas a low confidence probability will introduce erroneous pseudo-labels.

4.3.2 Semi-supervised Contrastive Learning

We utilize distinct contrastive learning strategies for IND labeled data and OOD unlabeled data. We compare confidence probabilities of pseudo-labels with a predefined confidence threshold, and if they are greater than this threshold, we consider them as reliable pseudo-labels. Unlabeled samples with reliable pseudo-labels are considered as labeled data.

Dataset	Classes	Classes-IND	Classes-OOD	Training	Validation	Test
BANKING	77	54	23	9003	1000	3080
STACKOVERFLOW	20	14	6	12000	2000	6000
CLINC	150	105	45	18000	2250	2250

Table 2: Statistics of BANKING, STACKOVERFLOW and CLINC datasets.

SCL is applied on the instance-level contrastive learning head f for labeled data, while unsupervised instance-level contrastive learning (ILCL) is performed on unlabeled data.

On the cluster-level contrastive learning head g, we perform cross-entropy loss for labeled data and perform cluster-level contrastive learning (CLCL) for OOD classes (OOD-setting) or all classes (open-setting). During the training process, the number of unlabeled samples with reliable pseudo-labels will gradually increase.

4.3.3 Prototypical Contrastive Learning

Decoupling knowledge from different levels is beneficial for separating the features of the source domain and target domain, thereby improving the efficiency of transfer learning and reducing overfitting. However, previous work only disentangled the instance-level and cluster-level features and applied constraints on them independently, without considering the inherent connection between the two levels of features. Each sample potentially belongs to a cluster, and each cluster consists of a certain number of samples. In order to extract the relationship between instance features and cluster features and enhance the discrimination between clusters, we propose to maintain a cluster prototype matrix, which is of size $k * m$ and stores the prototype features of each cluster.

For each batch, the output of f is an $n * m$ matrix containing the feature vectors of each sample, n is the batch size and m is the feature vector dimension. The output of g is an $n * k$ matrix, with each row corresponding to the probability that a sentence belongs to each class, *i.e.,* $p(k|x)$, and each column corresponds to the representation of a cluster. The cluster prototype matrix is computed by averaging the instance-level representations over all samples belonging to the class. For labeled data and unlabeled data with reliable pseudo-label, we use ground truth or pseudo-label, referred to as hard label constraint feature combination; for other unlabeled data, we perform probability-weighted calculations, known as soft semantic weighted feature combination. As shown in the equation, G is the cluster-level feature matrix

$[g_1; g_2; \cdots; g_m; 1_{y_{m+1}}, 1_{y_{m+2}}, \cdots, 1_{y_N}]$, F is the instance-level feature matrix $[f_1; f_2; \cdots; f_N]$, m is the number of labeled data in this batch. G' and F' are the cluster-level and instance-level feature matrix of the augmented samples, respectively.

$$M_c = G^T F \qquad (5)$$

The cluster prototype matrix M_c is a $k * m$ matrix consisting of the features of each clustering center and can be written as $[m_1, m_2, \cdots, m_K]$. The obtained vector m_i is normalized and used as the clustering center z_i as shown in the equation.

$$z_c, i = \frac{m_i}{\|m_i\|_2} \qquad (6)$$

After explicitly decoupling the cluster prototype vector z_i, the augmented features of each cluster prototype are used as positive samples, and the rest of the features are used as negative samples for contrastive learning at the cluster prototype level, as shown in the equation. Optimization of prototype contrastive loss (PCL) enables pulling apart different clusters and thus enhancing the discrimination between categories.

$$\ell_{i,j}^{pcl} = -log \frac{\exp\left(sim(z_{c,i}, z_{c,j})/\tau\right)}{\sum_{k=1}^{2N} 1_{i \neq k} \exp\left(sim(z_{c,i}, z_{c,k})/\tau\right)} \qquad (7)$$

The final loss in the training process is obtained by combining SCL, ILCL, CE, CLCL and PCL.[2]

5 Experiments

5.1 Datasets

We conduct experiments on three public datasets BANKING (Casanueva et al., 2020a), STACKOVERFLOW (Xu et al., 2015a) and CLINC (Larson et al., 2019a). BANKING consists of 13,083 queries covering 77 intents in the banking domain, and STACKOVERFLOW dataset contains 20 intents related to the programming domain. CLINC

7554

Method	BANKING-OOD			Stackoverflow-OOD			CLINC-OOD		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
DTC_BERT (Hsu et al., 2017)	45.76	42.88	69.12	57.83	32.31	37.29	51.56	48.33	84.64
KCL_BERT (Han et al., 2019)	47.61	36.5	64.51	41.33	28.74	34.42	57.33	49.45	80.35
MCL_BERT (Hsu et al., 2019)	45.87	34.85	62.83	42.39	27.04	33.71	51.7	43.77	77.69
CDAC+ (Lin et al., 2020)	59.78	44.58	69.19	61.56	28.22	52.76	73.04	64.44	87.90
DeepAligned (Zhang et al., 2021b)	63.86	<u>52.84</u>	<u>73.66</u>	79.68	63.18	65.52	91.56	86.58	94.91
DSSCC (Kumar et al., 2022)	64.67	51.38	71.25	<u>84.28</u>	65.94	64.75	80.89	73.55	89.40
DKT (Mou et al., 2022c)	66.50	52.07	72.22	82.22	61.53	**67.05**	<u>94.96</u>	<u>90.25</u>	<u>95.94</u>
DPN (An et al., 2023)	<u>68.15</u>	49.72	**74.76**	78.9	66.49	66.85	87.33	82.94	95.95
Llama2 (Touvron et al., 2023)	27.82	45.26	3.25	71.24	<u>67.62</u>	48.63	26.37	4.19	56.98
PLPCL	**68.37**	**53.19**	72.04	**86.28**	**69.64**	<u>66.95</u>	**95.11**	**90.71**	**96.15**

Table 3: The OOD-setting results on three datasets. Overall 1st/2nd in **bold**/<u>underline</u>. We randomly sample 30% of all classes as OOD intents for both datasets. Results are averaged over three random runs. ($p < 0.05$ under t-test)

Method	BANKING-Open			Stackoverflow-Open			CLINC-Open		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
DTC_BERT (Hsu et al., 2017)	42.56	31.72	69.12	52.7	35.19	49.3	50.22	39.72	79.51
KCL_BERT (Han et al., 2019)	64.87	54.52	80.07	63.43	50.42	61.83	69.02	62.98	88.77
MCL_BERT (Hsu et al., 2019)	65.39	55.21	79.53	63.55	47.51	57.18	68.4	60.68	87.78
CDAC+ (Lin et al., 2020)	45.00	33.10	69.49	67.05	48.66	66.03	51.64	35.89	79.96
DeepAligned (Zhang et al., 2021b)	<u>74.84</u>	<u>64.37</u>	84.86	<u>76.77</u>	<u>59.42</u>	71.97	**88.57**	**83.71**	**95.51**
DSSCC (Kumar et al., 2022)	69.55	63.13	85.37	70.55	56.89	68.92	83.51	79.81	94.91
DKT (Mou et al., 2022c)	70.38	61.16	83.71	72.57	58.6	69.12	75.44	71.20	92.69
DPN (An et al., 2023)	70.23	60.44	<u>85.7</u>	71.45	62.61	**77.86**	82.76	79.43	<u>95.4</u>
Llama2 (Touvron et al., 2023)	25.13	43.21	2.06	69.26	66.00	40.64	26.49	1.79	52.21
PLPCL	**76.50**	**67.13**	**85.99**	**77.63**	**63.58**	72.2	<u>86.38</u>	<u>81.65</u>	95.02

Table 4: The open-setting results on three datasets. Overall 1st/2nd in **bold**/<u>underline</u>. We randomly sample 30% of all classes as OOD intents for both datasets. Results are averaged over three random runs. ($p < 0.05$ under t-test)

is a cross-domain intent dataset covering 150 intents across 10 domains. Detailed statistics are shown in Table 2, the division of the training set, validation set and test set remains consistent with previous works. We take 30% categories as unknown categories in both datasets, and all data with known intent is labeled.

5.2 Baselines

We employ a series of semi-supervised methods as baselines for comparing OOD intent discovery and open intent discovery: DTC_BERT (Hsu et al., 2017), KCL_BERT (Han et al., 2019), MCL_BERT (Hsu et al., 2019), CDAC+ (Lin et al., 2020), DeepAligned (Zhang et al., 2021b), DSSCC (Kumar et al., 2022), DKT (Mou et al., 2022c), DPN (An et al., 2023). We also evaluate the classification ability of large language models with Llama2 (Touvron et al., 2023). To ensure lightweight implementation and reduce reliance on external data, we excluded MTP-CLNN (Zhang et al., 2022), as it requires extensive use of externally labeled data during the pre-training phase. All baselines use the

same BERT backbone to ensure a fair comparison.

5.3 Evaluation Matrics

We use three cluster evaluation metrics ACC, ARI, and NMI to evaluate the model effect, followed by DeepAligned (Zhang et al., 2021b). To obtain the results of ACC, we use the Hungarian algorithm to map prediction categories to ground-truth.

5.4 Implementation Details

We use the pre-trained bert-base-uncased model as the backbone consistent with the previous work, and pooling the context embeddings for each token using GRU and dense layers. The feature vector dimension is 768, the dropout probability is 0.1, and the GRU layer number is 1. In OOD discovery, the batch size of IND pre-training is 128, in the OOD clustering stage, the batch size of STACKOVERFLOW-OOD and BANKING-OOD are both 400, and the batch size of CLINC-OOD is 512. For open intent discovery, the batch_size is 128 for each dataset for the pre-training and training stages. As with DKT, the learning rate of

the pre-training process is set to 5e-5 of the training process is set to 0.0003, and the instance-level feature dimension is 128. Therefore, the cluster prototype feature dimension is also 128. The training epochs for training stage are 100. The experiment was conducted on an RTX 2080Ti GPU, and the running process takes 4 hours. We set $k = 2$ for pseudo-label threshold selection and we also analyze the performance of different confidence thresholds. We reproduce DSSCC (Kumar et al., 2022) and DPN (An et al., 2023) under the settings outlined in the original papers. For BANKING-OOD and CLINC-OOD, the results of CDAC+, DeepAligned and DKT are obtained from (Mou et al., 2022c), and others are obtained from the text open intent recognition platform (Zhang et al., 2021a). For open-setting, the results of baselines except DKT, DSSCC and DPN are obtained from (Zhang et al., 2021a).

5.5 Main Results

Table 3 and 4 show the performance of different models under the two task settings of three datasets. Our method outperformed previous approaches, especially on the BANKING and STACKOVER-FLOW datasets, indicating its strong adaptability to single-domain intent classification tasks and superior discriminability for professional intents with semantically similar meanings. For the CLINC dataset, which is a cross-domain dataset with significant differentiation between categories, previous methods also exhibit strong clustering capabilities.

5.6 Comparison with Large Language Model

In the penultimate row of the experiment results Table 3 and 4, we compared our results with Llama2-13B model (Touvron et al., 2023). Taking into account the input tokens' limitations and their relatively weaker clustering abilities, we employed large models for classification tasks with the provision of category names. The example of a prompt template is shown in Appendix A.3. The results indicate that the performance of LLM is inferior to our method in both settings of the three datasets. On the STACKOVERFLOW dataset with few categories, LLM outperforms some previous methods. However, on the BANKING and CLINC datasets with a larger number of categories, LLM clustering shows poor performance.

	ACC	ARI	NMI
ILCL,CLCL	46.07	36.79	37.00
+SCL,CE	70.38	61.16	83.71
+PCL	74.29	65.90	85.36
+PL	**76.50**	**67.13**	**85.99**

Table 5: Ablation study of different training objectives on BANKING-open.

σ	ACC	ARI	NMI
1	74.29	65.90	85.36
0.99	**76.50**	**67.13**	85.99
0.9	76.43	66.25	**86.16**
0.8	74.38	65.25	85.48
0.7	73.18	64.75	85.35
0.5	70.78	62.39	84.19
0	66.17	17.66	80.72

Table 6: Results under different confidence thresholds on BANKING-open.

5.7 Ablation Study and Further Analysis

Table 5 shows the effects of each module in our model, experimenting on BANKING-open. The results show that including SCL and CE during training helps to fully utilize the supervised signal. The absence of the supervised signal will result in a partial loss of pre-training information and a significant decrease in effectiveness. The addition of PCL improves the model's performance by 3.91% (ACC), 4.74% (ARI), and 1.65% (NMI), indicating that explicitly decoupling and separating the cluster centers is beneficial for distinguishing and separating different category features in the feature space. The addition of confident pseudo-labels (PL) improved the model's performance by 2.21% (ACC), 1.23% (ARI), and 0.63% (NMI), indicating that gradually including samples with sufficiently high confidence in the supervised signal during model iteration is beneficial for obtaining prior information, compensating for the limitations of simple sample scattering in unsupervised contrastive learning.

Table 6 shows the effect of different confidence thresholds on the effectiveness of the model on BANKING-open. When the threshold is 1, no pseudo-labels are used. When the threshold is 0, pseudo-labels are used for all samples. The confidence threshold analysis of on other datasets and settings can be found in Appendix A.1.

Figure 3 illustrates the clustering performance (ACC) of various models at different labeled ratios for IND intents while maintaining an IND category ratio of 70%. The results demonstrate the robust performance of our models across different labeling ratios. More results showing the effects of labeled ratio and known cluster ratio can be checked in

Method	BANKING-Open			Stackoverflow-Open			CLINC-Open		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
DeepAligned (Zhang et al., 2021b)	66.79	55.58	82.17	60.48	42.09	67.04	75.96	72.92	93.33
DPN (An et al., 2023)	69.77	62.22	**86.38**	65.6	45.91	66.85	78.36	**75.14**	**94.53**
Ours	**71.17**	**63.58**	84.74	**69.27**	**55.96**	**68.30**	**80.58**	74.56	94.16

Table 7: Clustering performance with predicted K

Appendix A.1.
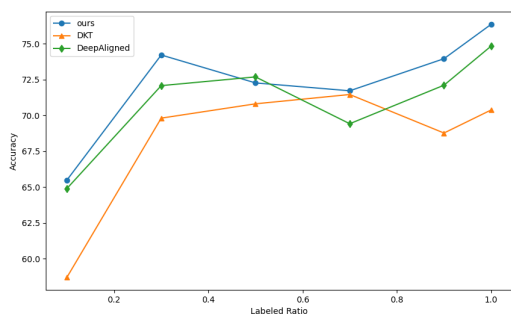
Figure 3: Influence of the labeled ratio on BANKING-open.

5.8 K-estimate Experiments

In real-world scenarios, it is difficult to determine the number of clusters K in advance. Therefore, we conduct K-estimation experiments by evaluating both the accuracy of the predicted K value and the performance using the predicted K following the approach in DeepAligned (Zhang et al., 2021b).

We compared the performance of our method with the DeepAligned and DPN (An et al., 2023) algorithms and show the results in Table 8. Our predicted K values are closer to the ground truth compared to those of DeepAligned and DPN, demonstrating that the features extracted by our PLPCL model are more clustering-friendly, resulting in a more accurate estimation of the number of intents. Based on our predicted K, we conducted

K	BANKING	Stackoverflow	CLINC
Ground Truth	77	20	150
DeepAligned	66	15	130
DPN	67	18	137
Ours	68	19	138

Table 8: Estimation of the number of categories.

experiments on three datasets to evaluate the impact of discrepancies between the estimated and actual number of unknown intents on clustering performance. The experimental results presented in Table 7 indicate that our approach outperforms the baseline methods when the number of clusters is unknown.

6 Conclusion

In this paper, we propose a pseudo-label enhanced prototypical contrastive learning approach for both open intent discovery and OOD intent discovery. The pseudo-label filtering strategy enhances supervised signal during the training process, while the prototypical contrastive learning module addresses the isolation issue between two independent contrastive learning heads. Experiments on two task settings and three benchmark datasets demonstrate the effectiveness of our proposed method. We hope to explore more self-supervised methods for OOD and open intent discovery in the future.

Limitations

In this work, we employ BERT-style models as our backbone, while the disadvantage lies in lacking the ability to generate coherent and contextually relevant text. While they can handle known intents well, BERT-style models may struggle with completely novel or OOD intents as they rely heavily on patterns seen during training. The generative LLMs are excellent at generating human-like text, making them suitable for creating responses and discovering new intents in open-ended queries. These models can adapt to new or unseen intents more effectively by generating diverse responses based on the input context, which is valuable for open intent discovery. To leverage the strengths of both BERT-style models and generative LLMs, some hybrid approaches can be employed in the future.

Acknowledgments

References

Wenbin An, Wenkai Shi, Feng Tian, Haonan Lin, QianYing Wang, Yaqiang Wu, Mingxiang Cai, Luyan Wang, Yan Chen, Haiping Zhu, and Ping Chen. 2024. Generalized category discovery with large language models in the loop. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8653–8665, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wenbin An, Feng Tian, Ping Chen, Siliang Tang, Qinghua Zheng, and QianYing Wang. 2022. Fine-grained category discovery under coarse-grained supervision with hierarchical weighted self-contrastive learning. *Preprint*, arXiv:2210.07733.

Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, QianYing Wang, and Ping Chen. 2023. Generalized category discovery with decoupled prototypical network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12527–12535.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020a. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020b. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Ajay Chatterjee and Shubhashis Sengupta. 2020. Intent mining from past conversations for conversational agent. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4140–4152, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhihong Chen, Jiawei Wu, Chenliang Li, Jingxu Chen, Rong Xiao, and Binqiang Zhao. 2022. Co-training disentangled domain adaptation network for leveraging popularity bias in recommenders. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 60–69.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2019. Continual learning: A comparative study on how to defy forgetting in classification tasks. *arXiv preprint arXiv:1909.08383*, 2(6):2.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dilek Hakkani-Tür, Yun-Cheng Ju, Geoffrey Zweig, and Gokhan Tur. 2015. Clustering novel intents in a conversational interaction system with semantic parsing. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. 2022. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8401–8409.

Keqing He, Jinchao Zhang, Yuanmeng Yan, Weiran Xu, Cheng Niu, and Jie Zhou. 2020. Contrastive zero-shot learning for cross-domain slot filling with adversarial attack. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1461–1467, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2017. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125*.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. In *International Conference on Learning Representations*.

Tianyang Hu, Zhili LIU, Fengwei Zhou, Wenjia Wang, and Weiran Huang. 2023. Your contrastive learning is secretly doing stochastic neighbor embedding. In *The Eleventh International Conference on Learning Representations*.

Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. 2022. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1836–1853.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A.

Laurenzano, Lingjia Tang, and Jason Mars. 2019a. 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, Hong Kong, China. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019b. 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, Hong Kong, China. Association for Computational Linguistics.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8547–8555.

Jinggui Liang and Lizi Liao. 2023. Clusterprompt: Cluster semantic enhanced prompt learning for new intent discovery. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10468–10481.

Jinggui Liang, Lizi Liao, Hao Fei, and Jing Jiang. 2024a. Synergizing large language models and pre-trained smaller models for conversational intent discovery. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14133–14147.

Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. 2024b. Actively learn from llms with uncertainty propagation for generalized category discovery. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7838–7851.

Haonan Lin, Wenbin An, Jiahao Wang, Yan Chen, Feng Tian, Mengmeng Wang, Guang Dai, Qianying Wang, and Jingdong Wang. 2024. Flipped classroom: Aligning teacher attention with student in generalized category discovery. *arXiv preprint arXiv:2409.19659*.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8360–8367.

Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, and Xianchao Zhang. 2021a. An explicit-joint and supervised-contrastive learning framework for few-shot intent classification and slot filling. *CoRR*, abs/2110.13691.

Pengfei Liu, Youzhang Ning, King Keung Wu, Kun Li, and Helen Meng. 2021b. Open intent discovery through unsupervised semantic clustering and dependency parsing. *arXiv preprint arXiv:2104.12114*.

Yu Liu and Tinne Tuytelaars. 2022. Residual tuning: Toward novel category discovery without labels. *IEEE Transactions on Neural Networks and Learning Systems*.

Qingkai Min, Libo Qin, Zhiyang Teng, Xiao Liu, and Yue Zhang. 2021. Dialogue state induction using neural latent variable models. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Yutao Mou, Keqing He, Pei Wang, Yanan Wu, Jingang Wang, Wei Wu, and Weiran Xu. 2022a. Watch the neighbors: A unified k-nearest neighbor contrastive learning framework for ood intent discovery. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1517–1529.

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

Yutao Mou, Keqing He, Yanan Wu, Zhiyuan Zeng, Hong Xu, Huixing Jiang, Wei Wu, and Weiran Xu. 2022c. Disentangled knowledge transfer for ood intent discovery with unified contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 46–53.

Yutao Mou and Heyang Xu. 2023. Bridge pre-training and clustering: A unified contrastive learning framework for ood intent discovery. *IEEE Access*.

Xiang Shen, Yinge Sun, Yao Zhang, and Mani Najmabadi. 2021. Semi-supervised intent discovery with contrastive learning. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 120–129.

Chen Shi, Qi Chen, Lei Sha, Sujian Li, Xu Sun, Houfeng Wang, and Lintao Zhang. 2018. Auto-dialabel: Labeling dialogue data with unsupervised learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 684–689.

Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2023. Large language models meet open-world intent discovery and recognition: An evaluation of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10291–10304, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Carel van Niekerk, Christian Geishauser, Michael Heck, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2023. Camell: Confidence-based acquisition model for efficient self-supervised active learning with label validation. *arXiv preprint arXiv:2310.08944*.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2019. Towards open intent discovery for conversational text. *arXiv preprint arXiv:1904.08524*.

Yanan Wu, Keqing He, Yuanmeng Yan, QiXiang Gao, Zhiyuan Zeng, Fujia Zheng, Lulu Zhao, Huixing Jiang, Wei Wu, and Weiran Xu. 2022a. Revisit overconfidence for OOD detection: Reassigned contrastive learning with adaptive class-dependent threshold. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4165–4179, Seattle, United States. Association for Computational Linguistics.

Yuxia Wu, Tianhao Dai, Zhedong Zheng, and Lizi Liao. 2024. Active discovering new slots for task-oriented conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yuxia Wu, Lizi Liao, Xueming Qian, and Tat-Seng Chua. 2022b. Semi-supervised new slot discovery with incremental clustering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6207–6218.

Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. 2022c. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia*.

Zhenhe Wu, Xiaoguang Yu, Meng Chen, Liangqing Wu, Jiahao Ji, and Zhoujun Li. 2023. Enhancing new intent discovery via robust neighbor-based contrastive learning. In *INTERSPEECH 2023*, pages 740–744.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015a. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015b. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, Denver, Colorado. Association for Computational Linguistics.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. 2021. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–878, Online. Association for Computational Linguistics.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021a. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 167–174, Online. Association for Computational Linguistics.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021b. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14365–14373.

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, Dublin, Ireland. Association for Computational Linguistics.

Yunhua Zhou, Jiawei Hong, and Xipeng Qiu. 2023. Towards open environment intent prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2226–2240.

A Appendix

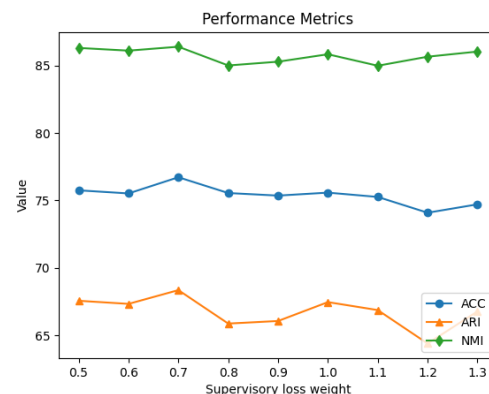## A.1 Further Analysis of loss weights and settings

Figure 4: Influence of the supervisory loss weight on BANKING dataset

Figure 4 demonstrates the performance of our model on the BANKING dataset under different weights of supervised contrastive loss, showing that our model is insensitive to loss weights. Figure 5 illustrates the impact of different labeled ratios and known cluster ratios on the model performance.

σ	ACC	ARI	NMI
1	74.33	60.66	70.92
0.99	**77.63**	**63.58**	72.2
0.9	71.08	57.74	68.38
0.8	69.37	56.76	68.8
0.7	66.93	54.24	67.64
0.5	66.9	57.15	70.18
0	66.8	37.73	**72.7**

Table 9: Results under different confidence thresholds on STACKOVERFLOW-open.

σ	ACC	ARI	NMI
1	67.52	**55.05**	69.27
0.99	**68.37**	53.19	**72.04**
0.9	63.09	49.68	70.19
0.8	53.26	38.04	62.4
0.7	47.93	36.88	63.02
0.5	34.24	21.7	49.64

Table 10: Results under different confidence thresholds on BANKING-OOD.

σ	ACC	ARI	NMI
1	85.17	67.58	65.73
0.99	**86.28**	**69.64**	**66.95**
0.9	71	53.67	56.12
0.8	71.67	43.63	49.54
0.7	53.39	25.06	41.03
0.5	61.72	43.61	51.77

Table 11: Results under different confidence thresholds on STACKOVERFLOW-OOD.

Table 9 ,10, 11 shows the impact of different confidence thresholds on the model's effectiveness across various datasets and task settings. The experiments show that selecting an appropriate confidence threshold is crucial for optimizing model performance across different datasets and tasks. Higher confidence thresholds generally lead to better classification results, indicating that the model can more accurately identify and leverage high-confidence samples for effective learning. This effect is even more pronounced in the OOD setting, where a low threshold can significantly degrade performance, emphasizing the need for careful threshold selection.

A.2 Visualization

Figure 6 shows the visualization results of the confusion matrix for DKT and our model on BANKING-open, with a total of 77 categories in the test set, and we show the first 20 categories. It can be found that the DKT model may completely confuse two certain categories, i.e. the samples of two certain categories are grouped into the same cluster. However, our model avoids this problem well, and rarely there is no correct sample in a certain category. Figure 7 shows the clustering visualization results of DKT and our model on BANKING-open and BANKING-OOD. For fair comparison, we use the same representation after the pooling layer. We can find that after adding contrastive learning for prototype and reliable pseudo label, while keeping the samples of the same cluster compact, the distance between different clusters is widened, and the different clusters become scattered on the whole feature space.Unlike scattering of unlabeled samples, the premise for contrastive learning in cluster prototype is that each cluster has its own unique features, and cluster center scattering aims to separate these features.

A.3 Prompt Template of LLM

The prompt template is shown in Table12.

Below is an instruction that describe a task. Write a response
that appropriately completes the request.
###instruction:
Please give the intent label for the following sentences.
Select one label in the set {...}
For example:
Input:
Can I exchange currencies with this app? .
Output:
{ intent_label:"exchange_via_app"}
###question:
Input:
{data sample}
Provide intent label in JSON format with the following keys: intent_label
###Response:

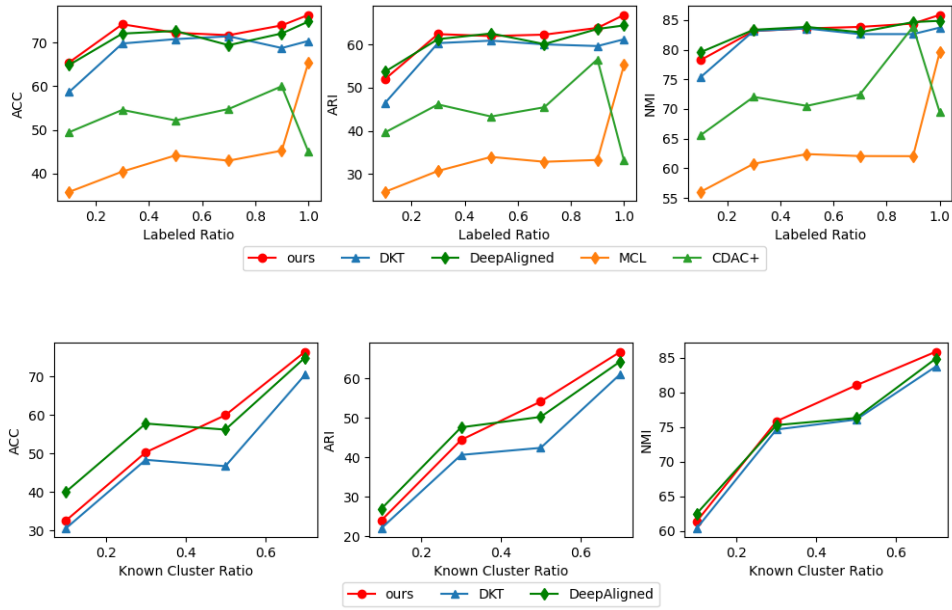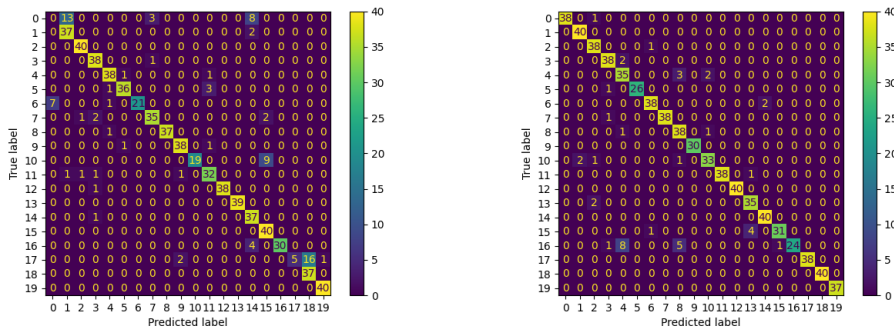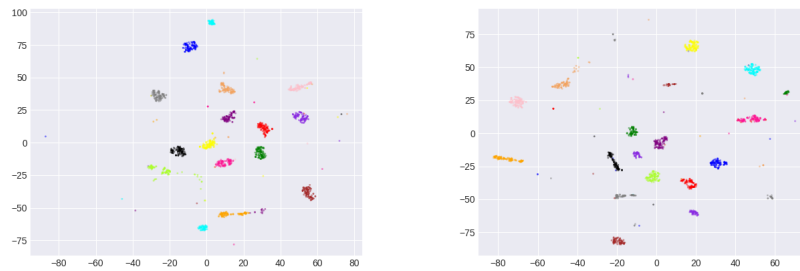Table 12: An example of the prompt templates we used.

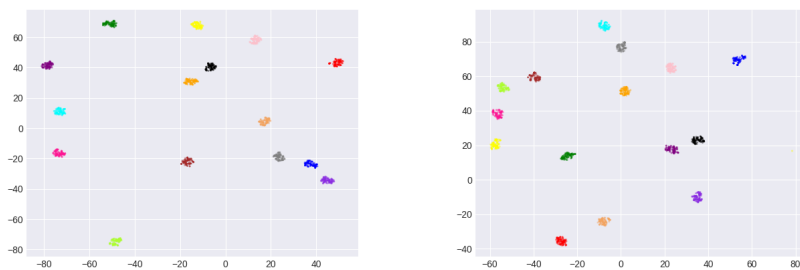Figure 5: Influence of the labeled ratio and known cluster ratio on BANKING-open.

(a) DKT

(b) Ours

Figure 6: Confusion matrix visualization of different models. We use the same test set of BANKING-open.

(a) DKT_OOD

(b) Ours_OOD

(c) DKT_open

(d) Ours_open

Figure 7: Cluster visualization of different models. We use the same test set of BANKING.