Synergizing Large Language Models and Pre-Trained Smaller Models for Conversational Intent Discovery

Anonymous ACL submission

Abstract

In Conversational Intent Discovery (CID), Small Language Models (SLMs) struggle with overfitting to familiar intents and fail to label newly discovered ones. This issue stems from their limited grasp of semantic nuances and their intrinsically discriminative framework. Therefore, we propose ***Syn***ergizing Large Language Models (LLMs) with pre-trained SLMs for CID (**SynCID**). It harnesses the profound semantic comprehension of LLMs alongside the operational agility of SLMs. By utilizing LLMs to refine both utterances and existing intent labels, SynCID significantly enhances the semantic depth, subsequently realigning these enriched descriptors within the SLMs' feature space to correct cluster distortion and promote robust learning of representations. A key advantage is its capacity for the early identification of new intents, a critical aspect for deploying conversational agents successfully. Additionally, SynCID leverages the in-context learning strengths of LLMs to generate labels for new intents. Thorough evaluations across a wide array of datasets have demonstrated its superior performance over traditional CID methods.[1]

1 Introduction

Recognizing user intents within conversational utterances is pivotal for developing intelligent conversational agents (Yilmaz and Toraman, 2020; Shen et al., 2021; Gung et al., 2023). Previous research mainly formulates this problem as a close-world intent classification task (Zhang et al., 2022a; Yehudai et al., 2023). However, in real-world applications, new intents continuously emerge. This spurs increasing interest in the open-world Conversational Intent Discovery (CID) (Zhang et al., 2021c, 2022b; Liang and Liao, 2023), a task that aims to recognize both known and new intents from extensive or even limited amount of user utterances.

[1] https://anonymous.4open.science/r/SynCID-3121

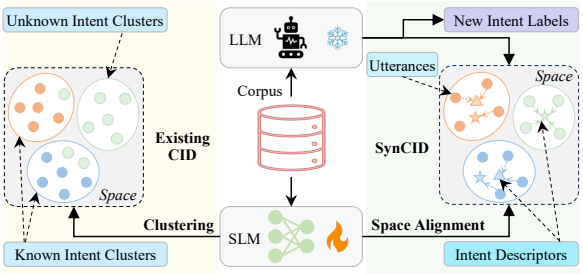

Figure 1: Existing methods rely on SLMs to cluster intents (Left), while our SynCID effectively synergizes LLMs and SLMs via space alignment (Right).

Current attempts at CID primarily rely on pre-trained Small Language Models (SLMs), which fall into two main categories: unsupervised and semi-supervised. Unsupervised methods (Padmasundari and Bangalore, 2018; Shi et al., 2018) firstly train SLMs without using any labeled data to obtain utterance representations, and then cluster them to infer intents. In contrast, semi-supervised methods (Lin et al., 2020; Zhang et al., 2021c; Zhou et al., 2023) leverage the available labeled data for the initial pre-training of SLMs, followed by fine-tuning these models with pseudo supervisory signals on unlabeled utterances for intent recognition. Given the specialized agility of SLMs, these methods can easily fit user utterances and learn discriminative representations for CID.

However, two key challenges persist. The first is *overfitting to known intents*, where these methods struggle to capture the full scope of intents and accurately model known label semantics. This limitation not only biases them towards existing intent categories but also compromises their ability to detect new intents early, a crucial capability for adaptive conversational agents. The second challenge is the *inability to label novel intents*, due to the inherently discriminative architecture of CID models, which falls short in recognizing and labeling emerging intents, marking a critical adaptability gap in current approaches.

Recently, Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023; OpenAI, 2023) have achieved significant breakthroughs in language understanding and generative tasks, including summarization (Liu et al., 2023) and query rewriting (Anand et al., 2023). Their success inspires a potential solution for addressing the above challenges by adapting LLMs to enhance intent discovery. Yet, the context length limitation of LLMs restricts their direct use in CID, which requires clustering thousands of utterances. While integrating user utterances with task-specific prompts to solicit intent labels from LLMs is possible, this prompting method risks generating intent labels without sufficient control, thus leading to unpredictable and uninstructive outcomes (Sun et al., 2023).

To navigate these challenges while leveraging the strengths of both LLMs and SLMs, we introduce SynCID, a framework that synergizes the deep semantic insights of LLMs with the agile, specialized capabilities of SLMs. SynCID employs a dual-prompting mechanism with LLMs to refine both utterances and known intent labels, enhancing the semantic precision of intent descriptors. This refinement process, informed by the nuanced understanding of LLMs, not only clarifies the intent representation but also primes the data for more effective learning. Following this, SLMs are trained through contrastive learning to align the semantic spaces of utterances with those of the intent descriptors. This innovative alignment strategy significantly reduces cluster distortion and improves the system's ability to detect and label new intents early, addressing the primary limitations of current CID approaches. By selecting a limited number of close-to-center utterances from newly formed intent clusters for in-context learning with LLMs, SynCID achieves precise intent labeling.

In summary, our contributions are threefold:

- We propose SynCID, an effective framework that synergizes powerful LLMs with agile SLMs to identify novel user intents and generate corresponding intent labels.
- We introduce a space alignment schema to align the representation spaces of utterances and the intent descriptors, significantly reducing the risk of overfitting to known intents.
- Experiments show that SynCID not only outperforms existing CID methods, but also provides labels for new intent clusters and enables early intent detection.

2 Related Work

2.1 Conversational Intent Discovery

Unsupervised Methods: Early unsupervised methods (Cheung and Li, 2012; Li et al., 2013) primarily extracted statistical features from unlabeled data to cluster queries with similar intents. Later studies (Xie et al., 2016; Yang et al., 2017; Padmasundari and Bangalore, 2018; Caron et al., 2018; Shi et al., 2018; Hadifar et al., 2019) leveraged deep neural networks to learn robust representations for clustering. More recently, the development of LLMs has facilitated their application in unsupervised intent recognition (De Raedt et al., 2023). Despite the progress, none of these unsupervised CID methods can fully harness supervised signals in learning representations and clustering intents.

Semi-supervised Methods: Addressing this limitation, semi-supervised methods (Hsu et al., 2018, 2019; Han et al., 2019; Lin et al., 2020) focus on integrating limited labeled data with extensive unlabeled data to enhance intent identification. For example, Hsu et al. (2018) transferred prior knowledge for clustering via predicting pairwise similarities. Further, several semi-supervised CID methods (Zhang et al., 2021b,c; Wei et al., 2022; Zhang et al., 2023; Zhou et al., 2023; Mou et al., 2023) formulated a two-stage schema for CID, which involves initially pre-training a base SLM and then iteratively fine-tuning it. This schema significantly enhanced CID by utilizing pseudo supervisory signals from the pre-trained SLM. Yet, it often faces issues related to the quality of these pseudo supervisory signals. Thus, there are also efforts (Mou et al., 2022a,b; Zhang et al., 2022b) refined learning objectives, such as contrastive learning, to learn discriminative representations for discerning intents. However, challenges persist in comprehensively grasping the nuanced semantics of both utterances and known intent labels, as well as generating new intent labels, which are addressed by our SynCID by synergizing LLMs and SLMs for CID.

2.2 The Synergy Between LLMs and SLMs

The emergence of LLMs has recently revolutionized various NLP tasks (Chowdhery et al., 2023; Black et al., 2022; Touvron et al., 2023), spurring research into their synergy with SLMs for boosting performance of small task-specific models. A promising direction in this synergy is using LLMs to create new and high-quality data for training downstream SLMs, enabling them to achieve com-

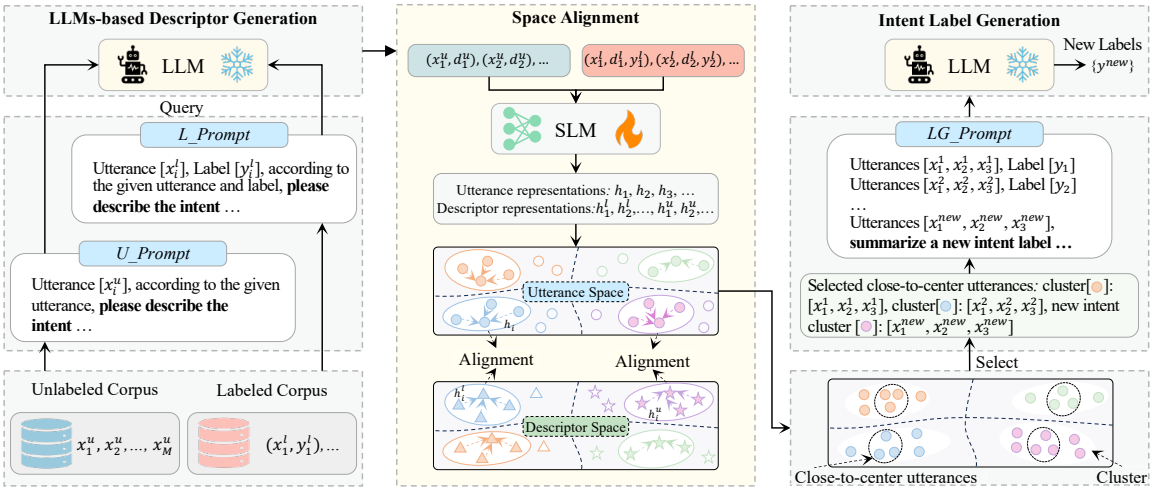

Figure 2: An overview of the proposed SynCID framework. It consists of three stages: LLMs-based Descriptor Generation, Space Alignment, and Intent Label Generation.

petitive performance.(Yang et al., 2020; Ding et al., 2023; Wei and Zou, 2019; Xie et al., 2020). Yet, such a method in CID risks unintentionally altering the semantic meanings of utterances or introducing noise, challenging accurate intent recognition.

Another effective method for synergizing LLMs and SLMs involves distilling task-specific knowledge from LLMs. Wang et al. (2021) showed the potential of GPT-3 as a cost-effective alternative to human labeling. Moreover, researchers like Li et al. (2022), Shridhar et al. (2023), and Hsieh et al. (2023) have utilized LLMs to generate task-specific labels and detailed explanations, facilitating the training of SLMs for reasoning tasks. Nevertheless, all these methods predominantly rely on either using a finite set of labels for annotating data or training generative models for aligning the knowledge from LLMs, which are not applicable in the CID.

In this work, we further the synergy to enhance intent discovery, leveraging a novel space alignment to align LLMs' comprehensive insights with the agility of SLMs and enabling early detection.

3 The SynCID Framework

3.1 Problem Formulation

Here, we study the CID problem as follows: Let $\mathcal{I}_k$ and $\mathcal{I}_{uk}$ represent the sets of known and unknown intents respectively, where $\{\mathcal{I}_k \cap \mathcal{I}_{uk}\} = \varnothing$ and $|\mathcal{I}_k| + |\mathcal{I}_{uk}| = K$. Here K is the total number of the user intents within the dataset. A typical CID task comprises a set of labeled utterance-intent pairs $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^N$, wherein each intent $y_i \in \mathcal{I}_k$, and a set of unlabeled utterances $\mathcal{D}^u = \{(x_i)\}_{i=1}^M$, where the intent of each utterance x_i belongs to $\{\mathcal{I}_k \cup \mathcal{I}_{uk}\}$. The CID task is to learn a SLM $\mathcal{M}$ to recognize all unknown intents $\mathcal{I}_{uk}$ within $\mathcal{D}^u$ and perform accurate clustering to classify each $x_i \in \{\mathcal{D}^l \cup \mathcal{D}^u\}$ into its corresponding intent category.

3.2 Model Overview

Figure 2 depicts an overview of the proposed SynCID framework for CID. It comprises three stages: **LLMs-based Descriptor Generation** (§3.3) for generating accurate and contextually rich intent descriptors, **Space Alignment** (§3.4) for aligning the representation spaces of utterances and intent descriptors to facilitate the synergy between LLMs and SLMs, and **Intent Label Generation** (§3.5) for producing labels for new intent clusters. We detail these stages in the subsequent subsections.

3.3 LLMs-based Descriptor Generation

This stage aims to leverage LLMs to recapitulate utterances and known intent labels into concise, accurate intent descriptors, eliminating irrelevant content in utterances while enriching the semantics of known intent labels. To achieve this, we develop two prompt templates: U_Prompt and L_Prompt, designed to guide the generation of these descriptors. As illustrated in Figure 2, U_Prompt is constructed as (x_i, p_u), prompting LLMs to generate descriptors related to the utterances as follows:

$$d_i^u = \text{LLM}(x_i, p_u), \quad (1)$$

where each $x_i \in \{\mathcal{D}^l \cup \mathcal{D}^u\}$ is a user utterance, and p_u denotes the prompt tokens. Similarly, L_Prompt is defined as (x_i, y_i, p_l) for the generation of label-enriched intent descriptors:

$$d_i^l = \text{LLM}(x_i, y_i, p_l), \quad (2)$$

where (x_i, y_i) is an utterance-intent pair in $\mathcal{D}^l$, and p_l refers to the respective prompt tokens. Crucially, in Equation 2, we integrate each intent label y_i with its corresponding utterance x_i to prompt LLMs for descriptor generation, enhancing the semantics of known user intents. After prompting LLMs to generate corresponding intent descriptors for all utterances and known user intents, we then utilize them to perform space alignment, facilitating the synergy of LLMs and SLMs for recognizing intents. For clarity, we formally redefine the training datasets $\mathcal{D}^l$ and $\mathcal{D}^u$ as follows: $\mathcal{D}^l = \{(x_i, d_i^u, d_i^l, y_i)\}_{i=1}^N$ and $\mathcal{D}^u = \{(x_i, d_i^u)\}_{i=1}^M$. It's noteworthy that we curate the aforementioned prompt templates without deliberation for better generalization.

3.4 Intent Discovery with Space Alignment

Given the intent descriptors from LLMs, we propose Space Alignment (SA) to synergize LLMs and SLMs for intent recognition. It comprises two sub-strategies: (1) SA with Contrastive Learning, which directly aligns the semantic spaces of utterances and intent descriptors, fostering robust utterance representation learning. (2) SA with Neighbor Filtering, which utilizes intent descriptors to refine neighborhood relationships between utterances, filtering out noise and promoting the formation of compact intent clusters.

SA with Contrastive Learning. Utilizing LLMs' strength in understanding and generation, we derive intent descriptors that offer more reliable and enriched insights into user intents. To effectively synergize LLMs and SLMs, we align the semantic spaces of utterances and LLM-generated intent descriptors via two contrastive learning objectives. Given the specialized agility of SLMs, this alignment can adeptly fit them into LLMs' insights, mitigating cluster distortion and enhancing the identification of new intents. Specifically, given a general pre-trained SLMs based CID model $\mathcal{M}$, we initially extract representations $\boldsymbol{x}_i$ and $\boldsymbol{d}_i^u$ for each utterance x_i and its corresponding intent descriptor d_i^u. Since d_i^u is derived from x_i using LLMs, $\boldsymbol{x}^i$ and $\boldsymbol{d}_i^u$ naturally form a positive pair. Following Gao et al. (2021), we compute an unsupervised contrastive loss between $\boldsymbol{x}_i$ and $\boldsymbol{d}_i^u$ as follows:

$$\mathcal{L}^{ucl} = -\log \frac{e^{\text{sim}(\boldsymbol{x}_i, \boldsymbol{d}_i^u)/\tau_1}}{\sum \mathbb{1}_{[k \neq i]} e^{\text{sim}(\boldsymbol{x}_i, \boldsymbol{d}_k^u)/\tau_1}}, \quad (3)$$

where $\text{sim}(\boldsymbol{x}_i, \boldsymbol{d}_i^u) = \frac{{\boldsymbol{x}_i}^T \boldsymbol{d}_i^u}{\|\boldsymbol{x}_i\| \|\boldsymbol{d}_i^u\|}$ is the cosine similarity and τ_1 is the temperature. The $\mathcal{L}^{ucl}$ aims to pull the representation of x_i close to the representation of its associated intent descriptor while maintaining distinction from others.

Additionally, for labeled utterances in $\mathcal{D}^l$, we further utilize the high-quality supervisory signals to optimize the SynCID. On the one hand, we utilize the supervised contrastive loss to align the extracted representations $\boldsymbol{x}_i$ and $\boldsymbol{d}_i^l$ for utterance x_i and its label-enriched intent descriptor d_i^l, facilitating discriminative representation learning as below:

$$\mathcal{L}^{scl} = -\sum_{j=1}^{\mathcal{Y}_{x_i}} \log \frac{e^{\text{sim}(\boldsymbol{x}_i, \boldsymbol{d}_j^l)/\tau_2}}{\sum \mathbb{1}_{[k \neq i]} e^{\text{sim}(\boldsymbol{x}_i, \boldsymbol{d}_k^l)/\tau_2}}, \quad (4)$$

where τ_2 is the temperature. $\mathcal{Y}x_i$ is the index set of data sharing the same label as x_i.

On the other hand, we compute a standard cross-entropy loss $\mathcal{L}^{ce}$ for the labeled utterances in $\mathcal{D}^l$ to regulate the training of SynCID. It optimizes the model $\mathcal{M}$ to distinguish the target intent classes of utterances from all known intent classes, enhancing the learning of utterance representations. Specifically, we map the utterance representation $\boldsymbol{x}_i$ into a probability distribution using a classifier and maximize the likelihood of its corresponding ground truth class (equation omitted for space). As a result, the overall loss $\mathcal{L}_{SACL}$ is formulated as follows:

$$\mathcal{L}_{SACL} = \mathcal{L}^{ce} + \lambda \mathcal{L}^{ucl} + \eta \mathcal{L}^{scl}, \quad (5)$$

where λ and η denote hyper-parameters that modulate the respective contributions of distinct losses.

SA with Neighbor Filtering. Upon training with $\mathcal{L}_{SACL}$, SynCID can learn some compact utterance representations for clustering. Yet, these representations are inevitably affected by the utterance noise from either the use of the unsupervised contrastive loss $\mathcal{L}^{ucl}$ or the limited comprehension capability of the model $\mathcal{M}$. To more effectively synergize LLMs with SLMs and amplify LLMs' insights for discerning intents, we further enhance SynCID by implementing neighbor utterance filtering, aiming for a more consistent alignment between the semantic spaces of the utterances and the intent descriptors from LLMs. Specifically, for each utterance x_i and its intent descriptor d_i^u, we first retrieve their nearest neighboring utterances $\mathcal{N}_{x_i}$ and intent descriptors $\mathcal{N}_{d_i^u}$ respectively. Owing to the accurate comprehension of LLMs, the intent descriptor d_i^u is anticipated to have cleaner neighbors. Thus, we filter out noisy utterance neighbors by omitting any $x_j \in \mathcal{N}_{x_i}$ where its paired $d_j^u \notin \mathcal{N}_{d_i^u}$, retaining

a purified neighbor set $\mathcal{N}'_{x_i}$ for x_i. During training, we update SynCID via a contrastive learning objective to pull together all filtered neighboring utterances and push apart non-neighbors as follows:

$$\mathcal{L}_{SANF} = -\sum_{j=1}^{\mathcal{N}'_{x_i}} \log \frac{e^{\text{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j)/\tau_3}}{\sum \mathbb{1}_{[p \neq i]} e^{\text{sim}(\boldsymbol{x}_i, \boldsymbol{x}_p)/\tau_3}}, \quad (6)$$

where τ_3 is the temperature. Here, we update the neighbor sets $\mathcal{N}_{x_i}$ and $\mathcal{N}_{d_i^u}$ every several epochs for filtering out noisy utterances during training.

3.5 Intent Label Generation

After training models to learn discriminative representations, existing CID methods (Zhang et al., 2022b, 2023) typically utilize clustering algorithms like K-means to group utterances into distinct clusters for inferring intents. Yet, it remains challenging to assign accurate labels for newly identified intent clusters. SynCID addresses this by utilizing the in-context learning capability of LLMs to generate suitable labels for new intent clusters. Specifically, we devise a label generation prompt (LG_Prompt) for extracting labels from LLMs. As illustrated in Figure 2, the LG_Prompt is constructed as:

$$\text{LG\_Prompt} = (\textit{ICDs}, \textit{Center Utterances}, p_c),$$

where $\textit{ICDs} = \{(x_1^j, ..., x_k^j, y_j)\}_{j=1}^n$ is a set of n in-context demonstrations. We can set the number n ranging from 1 to L considering the context size of LLMs. Each demonstration comprises a known intent label y_j and the top-k utterances near the y_j cluster center. *Center Utterances* is a set of utterances $(x_1, ..., x_k)$ located around the same unknown intent cluster center. p_c is the task description. For each unknown intent cluster, we integrate the top-k utterances allocated to it into the LG_Prompt, prompting LLMs to generate a new intent label y specific to it.

4 Experiments

4.1 Datasets

We conduct experiments on three CID datasets: **BANKING** (Casanueva et al., 2020), **CLINC** (Larson et al., 2019), and **StackOverflow** (Xu et al., 2015). The detailed statistics are reported in Appendix A.1. We keep the same train, development, and test splits as previous work (Zhang et al., 2023). To avoid randomness, we average the experimental results in five random runs. More experimental details are provided in the Appendix A.2.

4.2 Evaluation Metrics

We adopt three standard metrics for evaluating the CID performance: Accuracy (**ACC**) based on the Hungarian algorithm, Adjusted Rand Index (**ARI**), and Normalized Mutual Information (**NMI**). The specific definitions are shown in Appendix A.4. Note that **ACC** is considered as the primary metric, with higher values indicating better performance.

4.3 Baselines

We mainly compare our SynCID with the following SOTA baselines in our experiments:

Unsupervised: (1) **DEC** (Xie et al., 2016), (2) **DCN** (Yang et al., 2017), (3) **SCCL** (Zhang et al., 2021a), (4) **IDAS** (De Raedt et al., 2023).

Semi-supervised: (1) **DTC** (Han et al., 2019), (2) **CDAC+** (Lin et al., 2020), (3) **DeepAligned** (Zhang et al., 2021c), (4) **ProbNID** (Zhou et al., 2023), (5) **DCSC** (Wei et al., 2022), (6) **MTP-CLNN** (Zhang et al., 2022b), (7) **USNID** (Zhang et al., 2023), (8) **CsePL** (Liang and Liao, 2023). More details are provided in Appendix A.5.

4.4 Main Results

4.4.1 CID Performance Comparison

We report the main CID results in Table 1, with the highest performance highlighted in **bold**. We analyze the results as follows:

SynCID consistently outperforms CID baselines by large margins: Table 1 shows that SynCID exceeds all baseline methods in performance across three CID datasets and various KIR settings. For example, SynCID surpasses the top baseline CsePL by averages of 4.35% in ACC, 5.04% in ARI, and 2.07% in NMI on BANKING-25%. Moreover, SynCID shows stronger robustness in relation to the ratio of labeled data available. From BANKING-50% to BANKING-25%, SynCID's performance merely drops 2.42% in ACC, 2.27% in ARI, and 0.94% in NMI. In contrast, the corresponding metrics for CsePL diminish by 5.88%, 6.30%, and 2.33%, respectively. This suggests that SynCID, leveraging the nuanced understanding from LLMs, learns more robust utterance representations for recognizing intents and effectively alleviates the issue of overfitting to known user intents.

SynCID provides a better way to unleash the power of LLMs for CID. We can observe that our SynCID consistently demonstrates superior performance over previous unsupervised leading baseline IDAS. Specifically, SynCID surpasses IDAS

KIR	Methods	BANKING			CLINC			StackOverflow		
		ACC	ARI	NMI	**ACC**	ARI	NMI	**ACC**	ARI	NMI
0%	DEC	38.60	25.32	62.65	48.77	31.71	74.83	59.49	36.23	58.76
	DCN	38.59	25.36	62.72	48.69	31.68	74.77	59.48	36.23	58.75
	SCCL	40.54	26.98	63.89	50.44	38.14	79.35	68.15	34.81	69.11
	USNID	54.83	43.33	75.30	75.87	68.54	91.00	69.28	52.25	72.00
	IDAS	67.43	57.56	82.84	85.48	79.02	93.82	83.82	72.20	81.26
	SynCID	**72.89†**	**62.42†**	**84.20†**	**86.80†**	**80.85†**	**94.23***	**86.90†**	**74.42†**	**81.95***
25%	DTC	31.75	19.09	55.59	56.90	41.92	79.35	29.54	17.51	29.96
	CDAC+	48.00	33.74	66.39	66.24	50.02	84.68	51.61	30.99	46.16
	DeepAligned	49.08	37.62	70.50	74.07	64.63	88.97	54.50	37.96	50.86
	ProbNID	55.75	44.25	74.37	71.56	63.25	89.21	54.10	38.10	53.70
	DCSC	60.15	49.75	78.18	79.89	72.68	91.70	-	-	-
	MTP-CLNN	65.06	52.91	80.04	83.26	76.20	93.17	74.70	54.80	73.35
	USNID	65.85	56.53	81.94	83.12	77.95	94.17	75.76	65.45	74.91
	CsePL	71.06	60.36	83.32	86.16	79.65	94.07	79.47	64.92	74.88
	SynCID	**75.41†**	**65.40†**	**85.39†**	**87.85†**	**82.39†**	**94.85†**	**87.86†**	**76.11†**	**82.46†**
50%	DTC	49.85	37.05	69.46	64.39	50.44	83.01	52.92	37.38	49.80
	CDAC+	48.55	34.97	67.30	68.01	54.87	86.00	51.79	30.88	46.21
	DeepAligned	59.38	47.95	76.67	80.70	72.56	91.59	74.52	57.62	68.28
	ProbNID	63.02	50.42	77.95	82.62	75.27	92.72	73.20	62.46	74.54
	DCSC	68.30	56.94	81.19	84.57	78.82	93.75	-	-	-
	MTP-CLNN	70.97	60.17	83.42	86.18	80.17	94.30	80.36	62.24	76.66
	USNID	73.27	63.77	85.05	87.22	82.87	95.45	82.06	71.63	78.77
	CsePL	76.94	66.66	85.65	88.66	83.14	95.09	85.68	71.99	80.28
	SynCID	**77.83†**	**67.67†**	**86.33†**	**90.64†**	**85.96†**	**95.91***	**88.40†**	**77.24†**	**83.34†**

Table 1: Main performance results on CID across three public datasets. KIR denotes the ratio of known intents. Results are averaged over five random runs. († and * denote p-value$<$0.01 and p-value$<$0.05 under t-test respectively.)

by margins of 5.46% in ACC, 4.86% in ARI, and 1.36% in NMI on the BANKING-0%. On the multi-domain CLINC dataset, SynCID records improvements of 1.32% in ACC, 1.83% in ARI, and 0.41% in NMI. It is noteworthy that IDAS utilizes LLMs to refine a frozen pre-trained encoder for discerning intents. Our SynCID, by contrast, dynamically synergizes LLMs and SLMs through the alignment between original utterances and intent descriptors. This observation suggests that our SynCID can effectively unleash LLMs' nuanced comprehension capability to synergize them with SLMs for CID, guiding the SLMs in learning clarified utterance representations for intent identification.

4.4.2 Generated New Intent Labels

To study the quality of intent labels produced by SynCID, we conduct a comparative analysis between the gold standard labels and SynCID-generated intent labels on the CLINC dataset. Table 2 presents the comparison across different categories of intent labels. We can observe that for those clusters with specific and well-rounded user intent information, SynCID can accurately generate their corresponding intent labels, such as *Book hotel* and *Flight status*. Regarding the clusters that describe general user questions, SynCID can provide intent labels by condensing the user questions into high-level intents. For example, the intents

Gold Intent Label	**Generated Intent Label**
Book hotel	Book hotel
Flight status	Flight status
Who do you work for	Employer
Do you have pets	Pet ownership
Application status	Credit card application status
Oil change when	Oil change schedule

Table 2: Examples of generated new labels on CLINC.

Who do you work for and *Do you have pets* are succinctly transformed into *Employer* and *Pet ownership*, respectively. As for the clusters with overly general gold labels, *i.e.*, *Application status* and *Oil change when*, the SynCID is able to integrate additional cluster details to construct more specific and accurate intent labels. This analysis indicates that SynCID, leveraging the capabilities of LLMs, can effectively capture the intrinsic intents conveyed within utterances and generate high-quality intent labels for newly identified intents clusters.

4.4.3 Early Detection of New Intents

Effectively identifying new intents at their initial emergence is vital for developing adaptive conversational agents. To meet this practical demand, we evaluate the performance of SynCID in the early discovery of new intents, comparing it with existing top-performing baselines. Table 3 showcases ex-

Shots	Methods	BANKING		
		ACC	ARI	NMI
5	MTP-CLNN	45.72	33.56	69.07
	USNID	43.64	33.00	69.78
	CsePL	47.44	37.34	70.98
	SynCID	**56.06**	**44.40**	**74.54**
10	MTP-CLNN	46.00	35.69	70.54
	USNID	47.29	37.61	72.73
	CsePL	52.31	39.85	73.12
	SynCID	**59.01**	**46.25**	**75.67**
20	MTP-CLNN	50.08	40.15	73.90
	USNID	50.17	40.66	74.77
	CsePL	61.43	49.16	77.33
	SynCID	**66.09**	**54.06**	**80.22**

Table 3: Results of early new intent detection on BANKING-25%. Shots denote the number of utterances within each unknown intent.

KIR	Methods	BANKING		
		ACC	ARI	NMI
25%	SynCID-*Flan-T5-XXL*	73.47	61.90	83.84
	SynCID-*gpt-3.5-turbo*	74.29	62.86	84.21
	SynCID-*davinci-003*	75.41	65.40	85.39
	SynCID-*gpt-4*	**77.79**	**65.95**	**85.46**

Table 4: Effect of different LLMs on BANKING.

KIR	Methods	BANKING		
		ACC	ARI	NMI
0%	SynCID-*BERT*	72.89	62.42	84.20
	SynCID-*E5*	**74.06**	**63.91**	**85.34**
25%	SynCID-*BERT*	75.41	65.40	85.39
	SynCID-*E5*	**77.34**	**68.16**	**86.70**
50%	SynCID-*BERT*	77.83	67.67	86.33
	SynCID-*E5*	**79.71**	**70.27**	**87.84**

Table 5: Effect of different SLMs on BANKING.

perimental results in scenarios with a limited number of utterances per unknown intent, specifically at {5, 10, 20} shots. It is observed that existing baselines demonstrate a notable decrease in performance compared to their prior evaluations. In contrast, our SynCID, despite the reduction in performance, consistently surpasses other leading baselines. For example, with 20 utterances per unknown intent on BANKING-25%, SynCID achieves improvements over the baseline CsePL by 4.66% in ACC, 4.90% in ARI, and 2.89% in NMI. Additionally, it is noted that SynCID's performance gains over existing baselines progressively amplify as the number of utterance shots decreases. With only 5 utterance shots available for each unknown intent, SynCID attains improvements of 8.62% in ACC, 7.06% in ARI, and 3.56% in NMI. We hypothesize this observation can be explained by two main points: (1) Existing methods, which predominantly rely on SLMs, necessitate a sufficient quantity of utterances to cluster intents for reaching competitive performance. (2) In contrast, our SynCID synergizes LLMs and SLMs by aligning the semantic spaces of utterances with intent descriptors, providing a nuanced semantic understanding that compensates for limited data and thus enhancing the early discovery of new intents.

4.5 Detailed Analysis

4.5.1 Effect of Different LLMs

In addition to utilizing *text-davinci-003* as our basic LLM in the experiments, we further explore the use of various distinct LLMs, including the open-sourced *Flan-T5-XXL* (Chung et al., 2022) and the close-sourced *gpt-3.5-turbo* and *gpt-4*, for deriving intent descriptors within SynCID. As shown in Table 4, integrating SynCID with different LLMs for intent descriptor generation consistently surpasses the top-performing baseline CsePL, in all three evaluation metrics on the BANKING-25% dataset. Notably, utilizing *gpt-4* for intent descriptor generation yields further enhancements over *text-davinci-003*. We hypothesize that this enhancement is attributable to the superior quality of intent descriptors generated by the more advanced LLM, which are more constructive in accurately fulfilling user intent discovery.

4.5.2 Effect of Different Pre-trained SLMs

The proposed SynCID primarily synergizes the agile responsiveness of the pre-trained SLMs and LLMs' reliable insights for effectively discovering new intents. We inspect the contribution of different pre-trained SLMs, such as the BERT-based model and the more recent E5 model (Wang et al., 2022), to our SynCID, as detailed in Table 5. We can observe that integrating the E5 model into SynCID leads to further performance enhancements across various known intent rates when compared to the standard SynCID. It suggests that our SynCID framework stands to gain from synergizing LLMs and more advanced pre-trained SLMs.

4.5.3 Effect of Space Alignment

To verify the impact of different contrastive learning objectives within the space alignment on SynCID's performance, we conduct a comprehensive ablation study on BANKING-25%, with the results detailed in Table 6. Specifically, we selectively re-

KIR	Methods	BANKING		
		ACC	ARI	NMI
25%	SynCID	**75.41**	**65.40**	**85.39**
	- *w/o* $\mathcal{L}^{scl}$	72.69	62.42	84.12
	- *w/o* $\mathcal{L}^{ucl}$	72.86	62.47	84.11
	- *w/o* $\mathcal{L}_{SANF}$	68.20	57.28	81.38

Table 6: Ablation results on BANKING.

move three distinct contrastive losses from SynCID for analysis, where *w/o* denotes the model without the corresponding loss. Findings from Table 6 show a performance decline in CID when any contrastive loss is excluded. For example, removing $\mathcal{L}^{scl}$ results in SynCID's performance dropping by 2.72% in ACC, 2.98% in ARI, and 1.27% in NMI. Yet, despite these reductions, SynCID variants still maintain competitive performance compared to existing top-performing baselines. This underscores the efficacy of the contrastive learning objectives in the space alignment, highlighting their effectiveness in synergizing the powerful LLMs and the agile SLMs to learn discriminative representations, thereby facilitating the new intent identification.

4.5.4 Impact of Descriptor Shots

To further validate the efficacy of the intent descriptors within the proposed SynCID, we explore the impact of varying intent descriptor shots on SynCID's performance in intent discovery. We conduct experiments on StackOverflow-25%, where the improvement observed with SynCID is most pronounced, thus providing a solid foundation for this investigation. Figure 3 showcases a comparison of the CID performance corresponding to different intent descriptor shots within the SynCID. It can be observed that increasing the quantity of the intent descriptors for optimizing the SynCID does not yield substantial improvements in identifying new intents. We hypothesize this can be attributed to the propensity of LLMs to generate similar intent descriptors, even when prompted to generate multiple descriptors for a single utterance. These analogous intent descriptors do not provide enough supplementary information for the SynCID while increasing computation costs.

4.6 Visualisation of Alleviating Overfitting

For a more intuitive analysis of the effect of our SynCID on utterance representation learning, we present the t-SNE visualizations comparing the SynCID framework with the top baseline CsePL, as illustrated in Figure 4. We can observe that the SynCID framework performs space alignment to align the original utterance semantic space with LLMs' intent descriptor space for representation learning, thereby facilitating the formation of more compact and distinct intent clusters. Additionally, we can notice that SynCID effectively segregates the intertwined intent clusters, *i.e.*, *UK_automatic_top_up* and *K_top_up_limits*, *K_unable_to_verify_identity* and *K_why_verify_identity*, compared with the CsePL. The visualization of utterance representations demonstrates the proficiency of SynCID in alleviating the issue of overfitting to known intents.

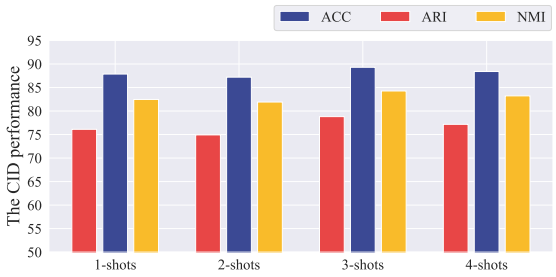

Figure 3: Effect of descriptor shots on StackOverflow.

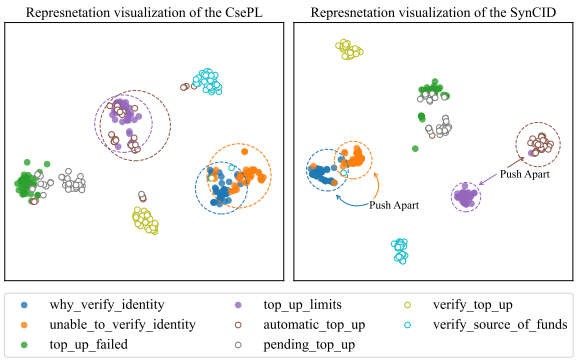

Figure 4: T-SNE visualization. The prefix "*UK_*" and "*K_*" denote unknown intents (hollow circles) and known intents (solid circles) respectively.

5 Conclusion

In this paper, we introduced SynCID, a novel framework that can effectively synergize LLMs and pretrained SLMs for conversational intent discovery. By aligning LLMs' reliable insights with the agile responsiveness of specialized SLMs, SynCID effectively alleviates the risk of overfitting to known intents in CID. Furthermore, SynCID enables the LLMs with in-context learning to skillfully produce labels for newly identified intent clusters. Through extensive experiments, our findings confirm SynCID's effectiveness. Deeper analysis reveals that SynCID not only sets new benchmarks in CID but also generates appropriate intent labels and enables early detection of new intents.

Limitations

Despite the promising results obtained by our SynCID, it is important to acknowledge several limitations: (1) The SynCID's reliance on LLMs subjects it to LLMs' inherent flaws, including biases in the training data and the propensity for hallucinating incorrect information. (2) The financial cost of utilizing commercial LLM APIs, such as OpenAI's, for experiments is significant. In our case, accessing APIs of LLMs such as *gpt-4*, *gpt-3.5-turbo*, and *davinci-003* for getting all the experimental results incurred a cost of approximately $510. (3) Our SynCID, similar to existing baselines, assumes a known ground-truth number of intents for clustering utterances — a condition that diverges from real-world applications where the exact number of intents remains unknown. To validate SynCID's effectiveness and robustness, we conduct further experiments with an estimated number of intents and explore the impact of various intent numbers around it on the CID performance of our SynCID. The findings from these additional experiments are detailed in Appendix B.

References

Avishek Anand, Venktesh V, Abhijit Anand, and Vinay Setty. 2023. Query understanding in the age of large language models. *CoRR*, abs/2306.16004.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *BigScience*, pages 95–136.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 139–156.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *NLP4ConvAI@ACL*, pages 38–45.

Jackie Chi Kit Cheung and Xiao Li. 2012. Sequence clustering and labeling for unsupervised query intent discovery. In *WSDM*, pages 383–392.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, pages 2292–2300.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023. IDAS: Intent discovery with abstractive summarization. In *NLP4ConvAI@ACL*, pages 71–88.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *ACL*, pages 11173–11195.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910.

James Gung, Raphael Shu, Emily Moeng, Wesley Rose, Salvatore Romeo, Arshit Gupta, Yassine Benajiba, Saab Mansour, and Yi Zhang. 2023. Intent induction from conversations for task-oriented dialogue track at DSTC 11. In *DSTC-WS*, pages 242–259.

Amir Hadifar, Lucas Sterckx, Thomas Demeester, and Chris Develder. 2019. A self-training approach for

short text clustering. In *RepL4NLP@ACL*, pages 194–199.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, pages 8400–8408.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of ACL*, pages 8003–8017.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2018. Learning to cluster in order to transfer across domains and tasks. In *ICLR*.

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

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. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *EMNLP-IJCNLP*, pages 1311–1316.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Yanen Li, Bo-June Paul Hsu, and ChengXiang Zhai. 2013. Unsupervised identification of synonymous query intent templates for attribute intents. In *CIKM*, pages 2029–2038.

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

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *AAAI*, pages 8360–8367.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023. On learning to summarize with large language models as references. *CoRR*, abs/2305.14239.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

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 *EMNLP*, pages 1517–1529.

Yutao Mou, Keqing He, Yanan Wu, Zhiyuan Zeng, Hong Xu, Huixing Jiang, Wei Wu, and Weiran Xu. 2022b. Disentangled knowledge transfer for OOD intent discovery with unified contrastive learning. In *ACL*, pages 46–53.

Yutao Mou, Xiaoshuai Song, Keqing He, Chen Zeng, Pei Wang, Jingang Wang, Yunsen Xian, and Weiran Xu. 2023. Decoupling pseudo label disambiguation and representation learning for generalized intent discovery. In *ACL*, pages 9661–9675.

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

Padmasundari and Srinivas Bangalore. 2018. Intent discovery through unsupervised semantic text clustering. In *INTERSPEECH*, pages 606–610.

Yilin Shen, Yen-Chang Hsu, Avik Ray, and Hongxia Jin. 2021. Enhancing the generalization for intent classification and out-of-domain detection in SLU. *CoRR*, abs/2106.14464.

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 *EMNLP*, pages 684–689.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of ACL*, pages 7059–7073.

Jiao Sun, Yufei Tian, Wangchunshu Zhou, Nan Xu, Qian Hu, Rahul Gupta, John Frederick Wieting, Nanyun Peng, and Xuezhe Ma. 2023. Evaluating large language models on controlled generation tasks. In *EMNLP*, pages 3155–3168.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv*, abs/2212.03533.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In *Findings of EMNLP*, pages 4195–4205.

Feng Wei, Zhenbo Chen, Zhenghong Hao, Fengxin Yang, Hua Wei, Bing Han, and Sheng Guo. 2022. Semi-supervised clustering with contrastive learning for discovering new intents. *CoRR*, abs/2201.07604.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*, pages 6381–6387.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, pages 478–487.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *NeurIPS*.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *VS@HLT-NAACL*, pages 62–69.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*, pages 3861–3870.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. G-daug: Generative data augmentation for commonsense reasoning. In *Findings of EMNLP*, pages 1008–1025.

Asaf Yehudai, Matan Vetzler, Yosi Mass, Koren Lazar, Doron Cohen, and Boaz Carmeli. 2023. QAID: question answering inspired few-shot intent detection. In *ICLR*.

Eyup Halit Yilmaz and Cagri Toraman. 2020. KLOOS: KL divergence-based out-of-scope intent detection in human-to-machine conversations. In *SIGIR*, pages 2105–2108.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen R. McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021a. Supporting clustering with contrastive learning. In *NAACL-HLT*, pages 5419–5430.

Hanlei Zhang, Xiaoteng Li, Hua Xu, Panpan Zhang, Kang Zhao, and Kai Gao. 2021b. TEXTOIR: An integrated and visualized platform for text open intent recognition. In *ACL-IJCNLP*, pages 167–174.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021c. Discovering new intents with deep aligned clustering. In *AAAI*, pages 14365–14373.

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

Haode Zhang, Haowen Liang, Yuwei Zhang, Li-Ming Zhan, Xiao-Ming Wu, Xiaolei Lu, and Albert Y. S. Lam. 2022a. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization. In *NAACL-HLT*, pages 532–542.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2022b. New intent discovery with pre-training and contrastive learning. In *ACL*, pages 256–269.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. 2023. A probabilistic framework for discovering new intents. In *ACL*, pages 3771–3784.

A Experimental Details

A.1 Dataset Statistics

Table 7 reports the detailed statistics for the BANKING, CLINC, and StackOverflow datasets. Specifically, the BANKING dataset includes over 13,000 carefully curated customer queries from the banking domain, categorized into 77 unique intents. The CLINC dataset encompasses a diverse collection of 22,500 labeled utterances distributed across 150 intents, covering multiple domains. StackOverflow, sourced from Kaggle.com, is a specialized dataset featuring 20,000 technical questions, organized into 20 distinct categories.

A.2 Implementation Details

For the dataset configuration, we randomly select a portion of intents to be designated as known intents, defining this portion as the known intent rate (KIR) at levels of 0%, 25%, and 50%. The KIR = 0% indicates the unsupervised setting to CID, whereas the KIR $>$ 0% implies the semi-supervised CID setting. From each intent selected as known, we sample 10% of the labeled utterances to create the labeled dataset $\mathcal{D}^l$. The remaining utterances are considered unlabeled, forming the basis of the unlabeled dataset $\mathcal{D}^u$.

For the LLMs-based Descriptor Generation and Intent Label Generation, our experiments are conducted with *text-davinci-003* serving as the basic LLM. To ensure deterministic outputs during descriptor generation, the temperature parameter is fixed at 0, and the output is limited to a maximum of 256 tokens. All other parameters are maintained at their default settings.

Within the Space Alignment, we utilize the pre-trained BERT model (*bert-uncased*), featuring a 12-layer transformer architecture, as the foundational SLM for training. The optimization of model parameters is conducted using the AdamW optimizer (Loshchilov and Hutter, 2019). During the SA with Contrastive Learning, the learning rate is set to 5×10^{-5}. The model outputs are projected from a 768-dimensional space to a 128-dimensional space for computing the contrastive loss. The temperatures $\{\tau_1, \tau_2\}$ for Equation 3 and 4 are uniformly set to 0.07. Furthermore, to achieve a balanced integration of $\mathcal{L}^{ucl}$ and $\mathcal{L}^{scl}$, we apply λ and η values of 1.0. A more detailed analysis of these hyper-parameters is available in Section A.3.

We leverage an early stopping mechanism with a patience setting of 20 epochs on the development set to train the model. For the SA with Neighbor Filtering, the learning rate is set to 1×10^{-5}. We set the temperature τ_3 in Equation 6 to 0.07 similarly. Regarding the selection of neighborhood sizes $\{|\mathcal{N}_{x_i}|, |\mathcal{N}_{d_i^u}|\}$, following Zhang et al. (2022b), we empirically assign the values {100, 50} for the BANKING dataset, {120, 50} for the CLINC dataset, and {1000, 500} for the StackOverflow dataset.

Dataset	Domain	Intents	Utterances
BANKING	banking	77	13,083
CLINC	multi-domain	150	22,500
StackOverflow	question	20	20,000

Table 7: Statistics of datasets used in the experiments.

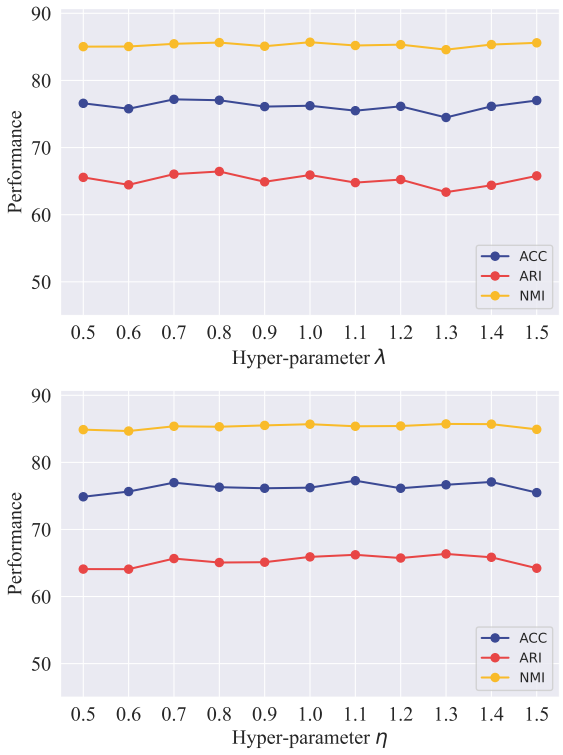

Figure 5: Impact of hyper-parameters λ and η on CID performance.

A.3 Hyper-parameter Analysis

We conduct extensive hyper-parameter exploration experiments on BANKING-25% for selecting the proper λ and η to optimize the proposed SynCID. In the experiments, We carefully considered a range of values λ and η, ranging from 0.5 to 1.5. Figure 5 illustrates the effect of different settings of these hyper-parameters on the overall performance of SynCID. It is observed that varying these hyper-

parameters, either by increasing or decreasing their values, does not result in a significant change in the model performance, which demonstrates the robustness and stability of our SynCID.

A.4 Evaluation Metrics

During our experimental analysis, we utilize three metrics for evaluating CID performance: ACC, ARI, and NMI. Specifically, ACC is employed to assess the CID effectiveness by comparing the model's predicted labels against the actual ground-truth labels. The calculation of ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^{N} \mathbb{1}_{y_i = map(\hat{y}_i)}}{N}$$

where $\{\hat{y}_i, y_i\}$ represent the predicted and ground-truth labels for an input utterance x_i, respectively. The function $map(\cdot)$ aligns each predicted label $\hat{y}_i$ with its associated ground-truth label y_i, utilizing the Hungarian algorithm for this mapping process.

ARI measures the concordance of the predicted and actual clusters through an assessment of pairwise accuracy within clusters. The formulation of ARI is as follows:

$$ARI = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - [\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{u_i}{2} + \sum_j \binom{v_j}{2}] - [\sum_i \binom{u_i}{2} \sum_j \binom{v_j}{2}]/\binom{N}{2}}$$

where $u_i = \sum_j n_{i,j}$, and $v_j = \sum_i n_{i,j}$. The total number of samples is given by N, and $n_{i,j}$ indicates the count of sample pairs concurrently classified into the i^{th} predicted and the j^{th} actual cluster.

NMI is calculated to gauge the degree of concordance between the predicted and actual clusters by quantifying the normalized mutual information between them, as delineated below:

$$NMI(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{2 \cdot I(\hat{\boldsymbol{y}}, \boldsymbol{y})}{H(\hat{\boldsymbol{y}}) + H(\boldsymbol{y})}$$

where $\{\hat{\boldsymbol{y}}, \boldsymbol{y}\}$ denote the predicted labels and the ground-truth labels respectively. $I(\hat{\boldsymbol{y}}, \boldsymbol{y})$ is the mutual information between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$. $H(\cdot)$ represents the entropy function.

A.5 Baselines

In this work, we compare the SynCID with the following SOTA baselines in our experiments:

Unsupervised Methods: (1) **DEC** (Xie et al., 2016): An unsupervised intent discovery method that iteratively learns and refines features by optimizing a clustering objective based on an auxiliary distribution. (2) **DCN** (Yang et al., 2017): A method that combines nonlinear dimensionality reduction with k-means clustering to learn cluster-friendly representations for CID. (3) **SCCL** (Zhang et al., 2021a): An end-to-end clustering framework that jointly optimizes a top-down clustering loss with a bottom-up instance-wise contrastive loss. (4) **IDAS** (De Raedt et al., 2023): An unsupervised method that utilizes LLMs to refine a frozen pre-trained encoder for identifying intents.

Semi-supervised Methods: (1) **DTC** (Han et al., 2019): A semi-supervised deep learning methodology for clustering, featuring an innovative mechanism for estimating the number of intents by leveraging labeled data. (2) **CDAC+** (Lin et al., 2020): An approach based on pseudo-labeling employs pairwise constraints and a target distribution strategy to facilitate the learning process in intent recognition. (3)**DeepAligned** (Zhang et al., 2021c): A semi-supervised technique that addresses inconsistencies in clustering through an alignment strategy, enhancing the learning of utterance embeddings. (4) **ProbNID** (Zhou et al., 2023): A probabilistic framework employs the expectation-maximization technique, considering intent categorizations as potential latent variables. (5) **DCSC** (Wei et al., 2022): An approach for discovering intents through pseudo-labeling incorporates a dual-task mechanism, utilizing the SwAV algorithm alongside the Sinkhorn-Knopp method (Cuturi, 2013) for the assignment of soft clusters. (6) **MTP-CLNN** (Zhang et al., 2022b): A two-stage approach that improves the learning of utterance representations for discovering novel intents by integrating an initial multi-task pre-training with a subsequent nearest neighbor contrastive learning. (7) **USNID** (Zhang et al., 2023): A framework for both unsupervised and semi-supervised intent discovery, featuring a novel strategy for initializing centroids effectively to derive cluster representations using historical clustering information. (8) **CsePL** (Liang and Liao, 2023): A method that employs two-level contrastive learning with label semantic alignment for enhancing the cluster semantics, alongside a soft prompting strategy for identifying new intents.

B Estimate the Intent Number K

Predicting the precise number of intent clusters in conversational intent discovery systems presents a significant challenge in real-world applications. Leveraging the approach presented by Zhang et al. (2021c), our research utilizes the pre-initialized in-

Cluster Num K	BANKING		
	ACC	ARI	NMI
K = 74 (predicted)	75.13	63.74	84.74
K = 77 (gold)	**75.41**	**65.40**	**85.39**
K = 71	73.90	62.43	84.17
K = 73	74.94	63.57	84.88
K = 75	75.10	63.94	84.78
K = 79	75.32	64.73	85.30
K = 81	75.39	65.19	85.39

Table 8: Experimental results of different cluster number K under the BANKING-25% setting.

tent features to autonomously ascertain the optimal number of intent clusters, represented as K. Initially, we assign a larger estimated number of clusters, K', and extract feature representations for the training dataset using a meticulously trained model. Subsequent clustering via the K-means algorithm divides these features into distinct groups. From this division, we distinguish between substantive intent clusters, characterized by their density and distinct boundaries, and smaller, less consequential clusters, which are then disregarded. The criteria for discerning between these cluster types can be outlined as follows:

$$K = \sum_{i=1}^{K'} \delta(|S_i| > \rho)$$

where $|S_i|$ is the size of the i^{th} grouped cluster, and ρ serves as the threshold for filtering. The function $\delta(\cdot)$ acts as an indicator, yielding a value of 1 when a specified condition is met.

Results of the experiments are reported in Table 8, where, in addition to the predicted number of clusters K, we examine the performance across a range of intent numbers proximal to it. The comparative results reveal that SynCID experiences merely marginal reductions in performance when confronted with inaccurate numbers of intents, indicating the robustness of SynCID in adapting to variations in the prediction of intent numbers.