Actively Learn from LLMs with Uncertainty Propagation for Generalized Category Discovery

Anonymous ACL submission

Abstract

Generalized category discovery faces a key issue: the lack of supervision for new and unseen data categories. Traditional methods typically combine supervised pretraining with self-supervised learning to create models, and then employ clustering for category identification. However, these approaches tend to become overly tailored to known categories, failing to fully resolve the core issue. Hence, we propose to integrate LLMs' feedback in an active learning paradigm. Specifically, our method innovatively employs uncertainty propagation to select data samples from high-uncertainty regions, which are then labeled using LLMs through a comparison-based prompting scheme. This not only eases the labeling task but also enhances accuracy in identifying new categories. Additionally, a soft feedback propagation mechanism is introduced to minimize the spread of inaccurate feedback. Experiments on various datasets demonstrate our framework's efficacy and generalizability, significantly improving baseline models at a nominal average cost.

1 Introduction

Generalized Category Discovery (GCD) is a crucial task in open-world computing (Lin et al., 2020; Zhang et al., 2021b), where the goal is to automate the classification of partially labeled data. It uniquely challenges systems to not only recognize predefined categories but also to discover entirely new categories from a mix of labeled and unlabeled data (Yang et al., 2021; Zeng et al., 2022). This task mirrors the dynamic and evolving nature of real-world data, where new categories frequently emerge, necessitating models that can adapt and learn continually.

In traditional GCD methods, the initial step often involves supervised pretraining on a labeled dataset to establish a foundational understanding of known categories (Zhong et al., 2021; Vaze et al., 2022). This is followed by self-supervised learning on unlabeled data or even contrastive learning, allowing the model to extract and learn patterns without explicit category labels (An et al., 2023). The final stage typically employs clustering techniques, like KMeans (MacQueen et al., 1967), to group similar data points, aiming to identify categories. However, this sequential process tends to imprint a bias towards the initially learned, known categories, limiting the model's ability to generalize to new, unseen categories (Mou et al., 2022). This overfitting to familiar data restricts the scope of GCD, preventing it from fully embracing the open-world setting it is intended for.

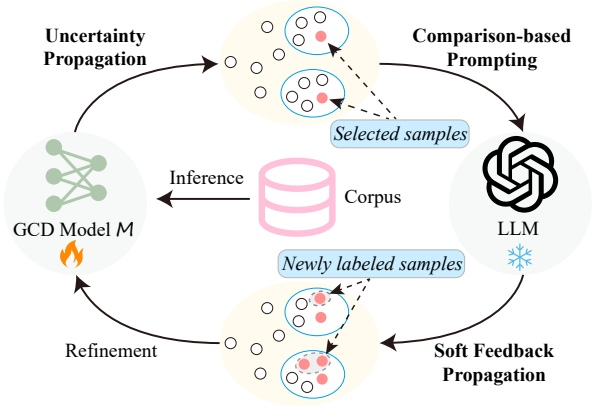

Figure 1: The active learning loop with propagated LLM feedback for model training.

Recently, Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023) have shown extraordinary versatility across a broad range of NLP tasks, providing good quality supervision signals for summarization (Liu et al., 2023), clustering (Zhang et al., 2023c) etc. Their ability to understand and generate nuanced language pat-

terns makes them promising for supplementing the supervision of new categories in GCD. However, the direct application of LLMs in GCD, which typically involves processing and clustering thousands of samples, raises substantial challenges. The intensive computational demands of LLMs could lead to issues with data privacy, high latency, and increased costs, which are particularly problematic in large-scale GCD scenarios.

To circumvent these challenges, integrating LLMs into an active learning framework presents a practical and efficient solution. This approach entails selectively using LLMs to provide supervision signals, especially in cases where the data is most uncertain or the categories are novel. However, this integration brings forth new challenges: optimizing the use of LLMs to ensure cost and time efficiency, and critically, ensuring the reliability of the feedback provided by LLMs. Effective strategies are needed to mitigate the risk of propagating incorrect feedback from LLMs.

Addressing these challenges, our approach ***A***ctively ***L***earns from LLMs with ***U***ncertainty ***P***ropagation for GCD, termed as **ALUP** shown in Figure 1. We begin by employing an uncertainty propagation strategy, which systematically identifies data samples in regions of high uncertainty – these are the areas where the model is least confident and, therefore, where LLM input could be most beneficial. The selected samples are then labeled using LLMs through a sophisticated comparison-based prompting technique. This method leverages the comparative strength of LLMs, making it easier for them to provide accurate feedback, especially for new and complex categories. To further enhance our approach, we incorporate a soft label propagation mechanism. This mechanism carefully extends the LLM-generated feedback to similar, neighboring samples, effectively amplifying the value of each LLM query while minimizing the risk of propagating errors. Rigorous testing on diverse datasets has shown that our method not only significantly improves upon existing baseline models but also does so with a nominal increase in cost, offering a scalable, efficient, and effective solution for the intricate problem of GCD.

In summary, our contributions are threefold:

- We developed an innovative active learning framework integrating LLMs feedback for GCD, addressing the challenge of limited supervision for new data categories.
- We combined uncertainty-region based data selection and comparison-based LLMs prompting, significantly enhancing GCD accuracy and efficiency with soft propagation.
- Experiments demonstrated marked improvements over traditional GCD methods across diverse datasets, affirming the framework's effectiveness and resource efficiency.

2 Related Work

2.1 Generalized Category Discovery

Unsupervised Approaches: The realm of GCD has been fundamentally shaped by unsupervised methods, focusing on learning cluster-friendly representations. These early approaches (Xie et al., 2016; Yang et al., 2017; Padmasundari and Bangalore, 2018; Caron et al., 2018; Hadifar et al., 2019) laid the groundwork by leveraging unsupervised clustering algorithms to categorize samples based on inherent similarities. Recent advancements, particularly with the emergence of LLMs, have brought a paradigm shift. The integration of LLMs in unsupervised GCD (De Raedt et al., 2023; Zhang et al., 2023d; Viswanathan et al., 2023) represents a novel direction, pushing the boundaries of category identification beyond traditional clustering techniques.

Semi-Supervised Approaches: In contrast, semi-supervised GCD approaches blend limited labeled data with possibly larger unlabeled data to enhance category discovery (Hsu et al., 2018, 2019; Han et al., 2019). Methods like CDAC+ (Lin et al., 2020) utilize labeled data to guide clustering, creating a synergy between supervised knowledge and unsupervised discovery. The two-stage scheme, involving base model pretraining and iterative optimization (Zhang et al., 2021a,b; Wei et al., 2022; Zhang et al., 2023a; Zhou et al., 2023; Mou et al., 2023), has been a popular approach. It benefits from pseudo label signals generated by the pretrained model, although it often struggles with the quality of pseudo labels and sample representations. Efforts to refine learning objectives, such as contrastive learning (Mou et al., 2022; Zhang et al., 2022a), aim to

directly learn discriminative representations for new categories. However, the challenge remains in effectively decoupling pseudo label generation from representation learning, a gap our work addresses by introducing LLMs into the GCD.

2.2 Active Learning in the Era of LLMs

Traditional Active Learning (AL): AL has traditionally been a solution to the data scarcity problem in NLP (Ren et al., 2022; Zhang et al., 2022b), focusing on identifying and annotating informative samples. Various acquisition strategies have been employed, including uncertainty-based (Wang and Shang, 2014; Schröder et al., 2022; Yu et al., 2023), diversity-based (Sener and Savarese, 2018; Gissin and Shalev-Shwartz, 2019; Citovsky et al., 2021), and hybrid methods (Liu et al., 2018; Zhan et al., 2022). While effective, these methods still rely on expensive human expertise for annotation.

LLMs as a Game-Changer in AL: With the advent of LLMs, a new frontier in AL has been explored. LLMs are now being considered as cost-effective alternatives to human experts (Zhang et al., 2023d; Cheng et al., 2023; Zhang et al., 2023b; Margatina et al., 2023). For instance, Xiao et al. (2023) demonstrated the use of LLMs as active annotators, harnessing their ability to distill task-specific knowledge interactively. In our work, we further this exploration by applying AL with LLMs to GCD. Our unique contribution not only lies in the implementation of an uncertainty-driven propagation strategy to maximize the utility of LLMs in a cost-effective manner, but also in the design of a soft feedback propagation scheme to minimize he spread of inaccurate feedback.

3 Methodology

3.1 Problem Formulation

In this work, we study the GCD formulated as follows: Assuming we have a set of known categories $\mathcal{C}_k$ and a set of unknown categories $\mathcal{C}_u$, where $\{\mathcal{C}_k \cap \mathcal{C}_u\} = \varnothing$ and $|\mathcal{C}_k| + |\mathcal{C}_u| = K$. Here K is the total number of categories in dataset. Under the semi-supervised GCD setting, given a set of labeled data $\mathcal{D}_l = \{(x_i, y_i)|y_i \in \mathcal{C}_k\}_{i=1}^{L}$, and a set of unlabeled data $\mathcal{D}_u = \{x_j\}_{j=1}^{U}$ where the category of each x_j belongs to $\{\mathcal{C}_k \cup \mathcal{C}_u\}$, the task is to learn a representation extractor $\mathcal{M}$ to identify all unknown categories from $\mathcal{D}_u$ and perform accurate clustering to classify each sample in $\{\mathcal{D}_l \cup \mathcal{D}_u\}$ into its corresponding category.

3.2 Approach Overview

General GCD models, denoted as $\mathcal{M}$, usually first extract representations $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{|\mathcal{D}_l \cup \mathcal{D}_u|}$ for each data sample x_i and then perform K-means to locate cluster centers $\{\boldsymbol{\mu}_i\}_{i=1}^{K}$ for doing GCD. For our proposed ALUP framework, it builds upon existing GCD models and effectively incorporate LLM feedbacks in active learning scheme.

Figure 2 depicts an overview of the proposed ALUP framework for GCD. It encompasses three key designs: ***Uncertainty Propagation*** for sample selection, ***Comparison-based Prompting*** for soliciting LLM's feedack, and ***Soft Feedback Propagation*** for wisely spreading the feedack. In what follows, we will detail these designs separately.

3.3 Uncertainty Propagation (UP)

Within the ALUP framework, we design uncertainty propagation to select the most informative unlabeled samples that are representative for high uncertainty regions. Note that given a general GCD model $\mathcal{M}$, we can extract representations $\boldsymbol{z}_i$ for each x_i in the dataset and perform K-means to locate cluster centers $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$. To estimate the model predictive uncertainty, following Xie et al. (2016), we use the Student's t-distribution to compute the probability of assigning the sample x_i to each cluster k:

$$q_{ik} = \frac{(1 + \|\boldsymbol{z}_i - \boldsymbol{\mu}_k\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{k'}(1 + \|\boldsymbol{z}_i - \boldsymbol{\mu}_{k'}\|^2/\alpha)^{-\frac{\alpha+1}{2}}}, \quad (1)$$

where α represents the freedom of the Student's t-distribution. After obtaining the model predictive probabilities, we employ the entropy (Lewis and Gale, 1994) to measure the uncertainty for each sample x_i:

$$u(x_i) = -\sum_{k=1}^{K} q_{ik} \log q_{ik}. \quad (2)$$

Here, a higher $u(x_i)$ can indicate a higher likelihood of the model $\mathcal{M}$ incorrectly assigning x_i to a wrong cluster. However, directly adopting this individual uncertainty score for selecting samples can lead to suboptimal outcomes as it can be sensitive

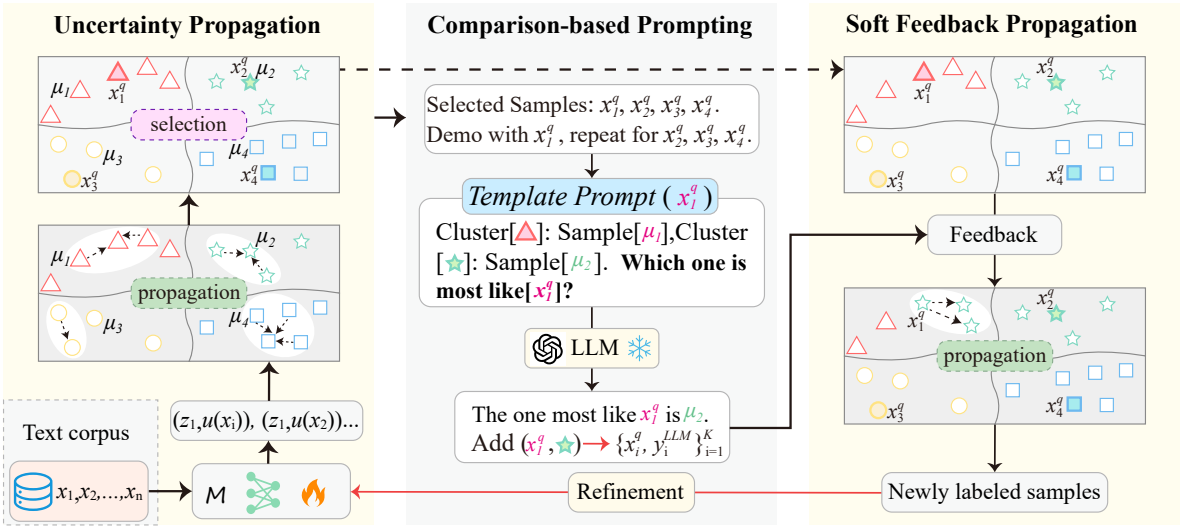

Figure 2: The overall ALUP framework. It consists three main designs: *Uncertainty Propagation* for region-based sample selection, *Comparison-based Prompting* for soliciting more accurate LLM's feedback, and *Soft Feedback Propagation* for wisely spreading the feedback to boost both efficiency and effectiveness.

to outliers (Karamcheti et al., 2021). To address this issue, following Yu et al. (2023), we further measure the similarities between each sample and its neighbors and propagate the individual uncertainty score to neigbors. Specifically, for each data point x_i, we first find its k-nearest neighbors based on the Euclidean distance as:

$$\mathcal{N}(x_i) = \underset{top-k}{\text{KNN}}(\boldsymbol{z}_i, \mathcal{Z}^u), \tag{3}$$

where $\mathcal{Z}^u$ denotes the representations of unlabeled samples and $\mathcal{N}(x_i)$ represents the set of nearest neighbors of x_i. Then, we calculate the similarities between x_i and its neighbors based on the radial basis function (RBF) (Schölkopf et al., 1997):

$$sim(\boldsymbol{z}_i, \boldsymbol{z}_j) = \exp(-\rho \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2^2), \tag{4}$$

where $x_j \in \mathcal{N}(x_i)$ and ρ is a hyper-parameter that regulates the extent of uncertainty propagation. After measuring the similarities, we refine the uncertainty score of sample x_i as:

$$u(x_i) = u(x_i) + \frac{\sum_{x_j \in \mathcal{N}(x_i)} sim(\boldsymbol{z}_i, \boldsymbol{z}_j) \cdot u(x_j)}{|\mathcal{N}(x_i)|}. \tag{5}$$

After several rounds of uncertainty score propagation, we obtain the final uncertainty score $u(x_i)$. Based on which, we greedily select one sample x_i^q from each cluster c_i to form the sample set $\mathcal{Q}$:

$$x_i^q = \underset{x_j \in c_i}{\text{argmax}}(u(x_j)). \tag{6}$$

We emphasize that a sample will exhibit higher propagated uncertainty only when it and its neighboring samples both possess high uncertainty levels. Hence, we are selecting samples from uncertain regions. By actively obtaining feedback from LLMs for such samples in $\mathcal{Q}$, we can significantly improve the model performance in GCD.

3.4 Comparison-based Prompting (CP)

After selecting the most informative unlabeled samples based on the UP strategy, we need to query external LLMs to obtain pseudo category labels for these samples. However, since the category labels of newly emerged categories remain unknown, it is infeasible to request LLMs to directly generate possibly a brand new label for the selected sample. To overcome this, we design a comparison-based prompting method from the clustering perspective, which prompts LLMs to classify a sample by comparing it with other samples representing distinct categories.

Specifically, for each category cluster, we need to find a representative sample for it. Hence, we first compute the distances of various samples within the cluster to its center $\boldsymbol{\mu}_i$, and then select the sample closest to $\boldsymbol{\mu}_i$ to represent this cluster. We denote the close-to-center sample as μ_i. Given these close-to-center samples $\mathcal{S} = \{\mu_i\}_{i=1}^K$, we construct the prompt to query LLMs as:

Cluster [c_1]: Sample [μ_1]; Cluster [c_2]: Sample [μ_2]; ... ; Cluster [c_p]: Sample [μ_p]. Above is a list of samples representing distinct categories. Please identify one sample that shares the same or similar underlying category as the input sample from the provided list.

Here, p is the number of representative samples used for the comparison. In our experiments, for each x_i^q in $\mathcal{Q}$, we empirically incorporate $p = |\mathcal{Q}|/2$ representative samples that are close to x_i^q into the prompt. With this design, we can effectively utilize LLMs to classify the selected samples into their corresponding categories, denoted as $\mathcal{Q} = \{x_i^q, y_i^{LLM}\}_{i=1}^K$, bypassing the requirement for explicit labels of unknown categories.

3.5 Soft Feedback Propagation (SFP)

By querying LLMs using the CP method, we can endow the selected unlabeled samples with their respective pseudo labels to augment the GCD models for discerning new categories. However, a performance gap persists between the partially and fully LLM-augmented GCD models. Given that the selection of the unlabeled samples is based on their model predictive uncertainty and neighboring uncertainty, and samples distributed close to each other are more likely to share the same category, we thus propose a Label Propagation mechanism to propagate the pseudo labels generated by LLMs across their similar neighbors, amplifying the utility of the feedback from LLMs without any additional cost. Specifically, for each x_i^q in $\mathcal{Q}$, we refine the model prediction $\boldsymbol{q}_j$ of its uncertain neighbor $x_j \in \mathcal{N}(x_i^q)$ in Equation (1) to propagate the LLM-generated pseudo label y_i^{LLM}:

$$\boldsymbol{q}_j = (1 - sim(\boldsymbol{z}_j, \boldsymbol{z}_i^q)) \cdot \boldsymbol{q}_j + sim(\boldsymbol{z}_j, \boldsymbol{z}_i^q) \cdot \boldsymbol{y}^{LLM}, \quad (7)$$

$$y_j^{prop} = \begin{cases} y_i^{LLM}, & \text{if argmax}(\boldsymbol{q}_j) = y_i^{LLM} \\ -1, & \text{otherwise} \end{cases}, \quad (8)$$

where $sim(\cdot, \cdot)$ denotes the similarity function defined in Equation (4). $\boldsymbol{y}^{LLM}$ is an one-hot vector where the value of position y_i^{LLM} is set to 1. To interpret the Equation (8), we argue that when the uncertain neighbor $x_j \in \mathcal{N}(x_i^q)$ is assigned to the same cluster as the LLM-labeled sample x_i^q according to the refined $\boldsymbol{q}_j$, the pseudo label y_i^{LLM} will be propagated to the x_j. Otherwise, the x_j will reject the pseudo label y_i^{LLM} and remain as an unlabeled sample.

4 Experiments

4.1 Datasets

We conduct experiments on three popular GCD 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. In our experiments, we keep the same train, development, and test splits as previous work (Liang and Liao, 2023). More experimental details are provided in the Appendix A.2.

4.2 Evaluation Metrics

We use the following metrics adopted in previous work (Zhang et al., 2022a; Zhou et al., 2023; De Raedt et al., 2023) to evaluate the GCD performance: Accuracy (**ACC**) based on Hungarian algorithm, Adjusted Rand Index (**ARI**), and Normalized Mutual Information (**NMI**). The specific definitions are presented in Appendix A.3. It is worth noting that **ACC** is regarded as the primary metric for evaluation, with higher values indicating better GCD performance.

4.3 Baselines

We compare with the following SOTA GCD methods: **DTC** (Han et al., 2019), **CDAC+** (Lin et al., 2020), **DeepAligned** (Zhang et al., 2021b), **ProbNID** (Zhou et al., 2023), **DCSC** (Wei et al., 2022), **MTP-CLNN** (Zhang et al., 2022a), **USNID** (Zhang et al., 2023a), and the best-performing method **CsePL** (Liang and Liao, 2023). We leave the details of these baselines in Appendix A.4.

4.4 Main Results

4.4.1 GCD Performance Comparison

Table 1 presents the main GCD results of our proposed ALUP compared to the existing baselines, where the peak performance is highlighted in **bold**. Generally speaking, the proposed ALUP framework consistently outperforms all existing baselines across three datasets with large margins. Here, we analyze the results from the following aspects:

Comparison of different methods in GCD: As shown in Table 1, it is observed that the ALUP has

KCR	Methods	BANKING			CLINC			StackOverflow		
		ACC	ARI	NMI	**ACC**	ARI	NMI	**ACC**	ARI	NMI
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
	ALUP	**74.61**	**62.64**	**84.06**	**88.40**	**82.44**	**94.84**	**82.20**	64.54	**76.58**
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
	ALUP	**79.45**	**68.78**	**86.79**	**90.53**	**84.84**	**95.97**	**86.70**	**73.85**	**81.45**

Table 1: Main performance results on the generalized category discovery across three public datasets. KCR denotes the known category rate.

a significant improvement compared with existing top-performing baselines, *i.e.*, CsePL and USNID. For example, the proposed ALUP surpasses previous SOTA CsePL by margins of 2.51% in ACC, 2.12% in ARI, and 1.14% in NMI on BANKING-50%. Moreover, it is noteworthy that performance gains are more pronounced when a larger number of categories remain unknown. For example, the ACC performance of our ALUP improves 3.55% on BANKING-25%. It proves that the ALUP can acquire effective supervision signals from LLMs to enhance the model performance of discovering new categories.

Comparison of different datasets: We evaluate the performance of the proposed ALUP on different datasets, including the single-domain, fine-grained BANKING dataset and the multi-domain CLINC dataset. From Table 1, we can notice that all existing methods exhibit significantly lower performance on the BANKING dataset compared to the CLINC dataset, indicating that the single-domain fine-grained scenario is more challenging for GCD. However, the ALUP achieves a more significant improvement of 1%~3% on BANKING-50% compared with the CsePL, while only 0.8%~2% on CLINC-50%. This observation further strengthens the benefits of our ALUP in providing effective supervision signals to cope with the challenges in fine-grained category discovery.

KCR	Methods	BANKING		
		ACC	ARI	NMI
25%	ALUP	**74.61**	**62.64**	**84.06**
	- *w/o* UP	73.41	61.30	83.42
	- *w/o* SFP	72.73	60.97	83.68
	- *w* HP	70.24	59.08	82.32
50%	ALUP	**79.45**	**68.78**	**86.79**
	- *w/o* UP	78.64	67.16	86.05
	- *w/o* SFP	77.66	67.04	86.43
	- *w* HP	75.60	64.33	84.72

Table 2: Ablation results on BANKING dataset.

4.5 In-depth Analyses

In this subsection, we conduct further detailed analyses to explore the impact of each key component within the proposed ALUP framework.

4.5.1 Effect of Uncertainty Propagation

Table 2 presents the experimental results of removing the Uncertainty Propagation in Equation (5) from the ALUP on the BANKING dataset. We can observe that the GCD performance of the ALUP substantially diminishes across various known category ratios. Especially, the ACC of the ALUP decreases by 1.20% while the ARI and NMI drop

1.34% and 0.64% on BANKING-25% respectively. This observation indicates that the Uncertainty Propagation can accurately identify the most informative samples for querying LLMs to enhance the GCD model performance, which notably avoids selecting outliers with high model uncertainty but are less beneficial for the model learning.

4.5.2 Effect of Soft Feedback Propagation

We also explore the contribution of the Soft Feedback Propagation by comparing the model performance when omitting the feedback propagation from LLMs Equation (8) with the standard ALUP. As presented in Table 2, we find that the model performance significantly decreases without propagating the feedback from LLMs, with ACC drops by 1.88%, ARI by 1.67%, and NMI by 0.38%. Nevertheless, ALUP *w/o* SFP still slightly outperforms the best-performing baseline CsePL. We argue the reasons for this observation lie in: (1) Acquiring supervision signals from LLMs for the informative samples is beneficial for improving the model performance in discovering new categories. (2) The Soft Feedback Propagation strategy can effectively filter out and propagate the accurate supervision signals from LLMs, amplifying the utility of LLM's feedback while concurrently minimizing the risk of propagating errors.

Compared to the Soft Feedback Propagation, we also investigate the Hard Propagation (ALUP *w* HP) in the ALUP, which directly extends the LLMs' feedback to the neighboring samples without any control. As presented in Table 2, we can observe that the model performance significantly decreases with Hard Propagation, falling below even that of the CsePL. This is probably due to the propagation of inaccurate supervision signals from LLMs,which introduces considerable noise into the model learning.

4.5.3 Number of Propagated Neighbors

To delve deeper into the effectiveness of the Uncertainty Propagation within the proposed ALUP, we further conduct experiments on the BANKING dataset to explore how varying the number of propagated neighbors for selecting unlabeled samples influences model performance. Figure 3 presents the performance curves corresponding to various numbers of propagated neighbors. When increasing the number of propagated neighbors in Equation (3), there is a gradual improvement in the performance of the proposed ALUP, culminating in peak performance with 25 propagated neighbors. However, when the number of propagated neighbors exceeds 25, there is a decline in model performance. This decrease can be attributed to the inclusion of less uncertain samples, which potentially introduces significant noise into the process of unlabeled sample selection.

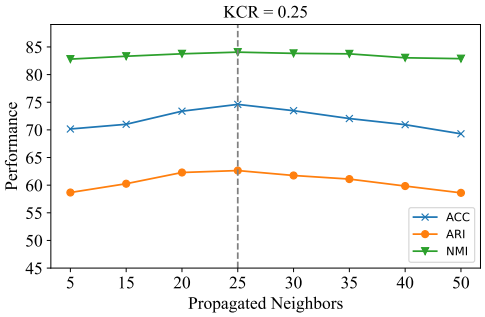

Figure 3: Effect of the number of propagated neighbors.

KCR	p	BANKING		
		ACC	ARI	NMI
25%	19	73.70	61.40	83.58
	38	**74.61**	**62.64**	**84.06**
	57	72.01	60.86	83.04
	77	71.36	59.58	82.64
50%	19	78.44	67.46	86.25
	38	**79.45**	**68.78**	**86.79**
	57	77.56	66.04	85.93
	77	76.66	65.23	85.59

Table 3: Effect of the number of representative samples in Comparison-based Pompting.

4.5.4 Effect of Representative Samples

To evaluate the effectiveness of the Comparison-based Prompting method, we also analyze how the model performance varies with different numbers of representation samples p incorporated into the prompt for querying LLMs. We conduct experiments using values of p set to {19, 38, 57, 77}, where 19 is about one quarter of the total number of clusters. As reported in Table 3, we can notice that the optimal GCD performance is attained by including 38 representative samples to prompt LLMs for acquiring supervision signals for the unlabeled samples. The main reasons for this observation come from two aspects: (1) A smaller p may potentially omit the representative samples

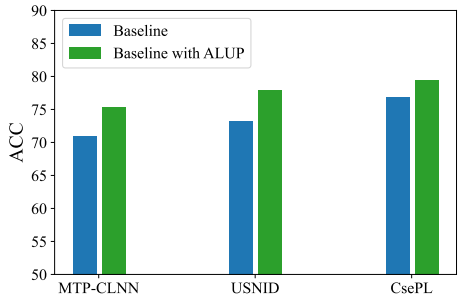

Figure 4: GCD Performances of different base models in ALUP on the BANKING-50%.

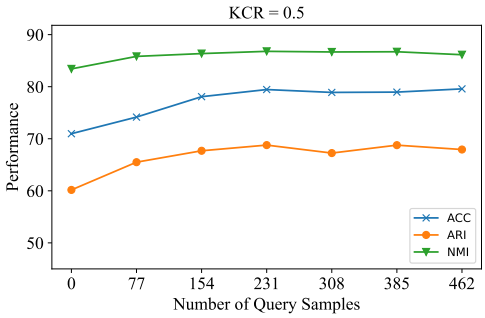

Figure 5: Effect of number of query samples on the BANKING-50%.

sharing the same underlying category as the selected samples, leading to an inability for LLMs to provide necessary supervision signals when comparing with the incorporated representative samples. (2) A larger number of representative samples used in the Comparison-based Prompting leads to a longer prompt. This can potentially cause LLMs to misclassify the selected unlabeled samples into inaccurate categories, which further degrade the model performance.

4.6 Impact of Different Base GCD Models

In our experiments, we select the most informative unlabeled samples based on the existing GCD models. To validate the effectiveness of the proposed ALUP, we also examine how its performance varies when different GCD models are integrated within ALUP on the BANKING-50% dataset. As depicted in Figure 4, we can observe consistent and significant improvements with the proposed ALUP. It indicates that the proposed ALUP framework is effective in acquiring supervision signals from LLMs to enhance the model performance of discovering new categories and is adaptable to other GCD models.

KCR	Methods	BANKING		
		ACC	ARI	NMI
25%	ALUP-*gpt-3.5-turbo*	**74.61**	**62.64**	**84.06**
	ALUP-*FlanT5-XXL*	73.38	62.29	83.76

Table 4: Effect of different LLMs.

4.7 Influence of Query Sample Number

We study the effect of varying the number of selected unlabeled samples for querying LLMs in Figure 5. It is observed that there is an increase in model performance corresponding to the rise in the number of samples selected for querying LLMs. However, this growth rate progressively diminishes as the LLMs' feedback is propagated, and selecting informative samples becomes more challenging with the increasing number of selected samples.

4.8 Effect of Different LLMs

In addition to employing close-sourced *gpt-3.5-turbo* as the basic LLM in our experiments, we also conduct experiments on the BANKING-25% to explore the use of open-sourced LLM *FlanT5-XXL* (Chung et al., 2022) to derive supervision signals for the ALUP. As shown in Figure 4, we can notice that, when coupling the *FlanT5-XXL* in the ALUP framework, there is a slight decrease in GCD performance compared to other close-sourced *gpt-3.5-turbo* model. However, it still surpasses the top-performing baseline CsePL.

5 Conclusion

In summary, our ALUP framework innovatively integrates Large Language Models with uncertainty propagation in generalized category discovery, marking a significant leap in the field. By employing comparison-based LLM prompting and a novel soft feedback propagation mechanism, ALUP adeptly identifies and categorizes new data with enhanced accuracy and efficiency. This approach not only surpasses traditional GCD methods but also minimizes the risk of error propagation, a critical advancement in handling real-world, dynamic datasets with LLMs. Future endeavors will focus on refining LLM integration, extending our methods to multi-modal data, and enhancing scalability and data privacy measures, furthering ALUP's potential in diverse and evolving open-world computing.

Limitations

While our ALUP framework marks a significant advance in Generalized Category Discovery using LLMs, it does have some limitations. The reliance on LLMs can introduce biases and inaccuracies, particularly in areas where these models have limited training data or exposure. Although our propagation method effectively reduces overall costs, the initial computational demands of LLMs may still pose scalability challenges, especially for resource-limited environments. Additionally, the framework currently focuses on textual data, which could limit its applicability in multi-modal data scenarios. Moreover, while our soft feedback propagation mechanism aims to minimize error spread, it is not immune to the risk of amplifying initial inaccuracies from LLM feedback. Finally, data privacy and security remain critical concerns in the use of external LLMs, necessitating ongoing vigilance and adaptation.

References

Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, Qianying Wang, and Ping Chen. 2023. Generalized category discovery with decoupled prototypical network. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12527–12535. AAAI Press.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, pages 139–156.

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

Qinyuan Cheng, Xiaogui Yang, Tianxiang Sun, Linyang Li, and Xipeng Qiu. 2023. Improving contrastive learning of sentence embeddings from AI feedback. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11122–11138.

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, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. 2021. Batch active learning at scale. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11933–11944.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2292–2300.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023. IDAS: Intent discovery with abstractive summarization. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 71–88.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. *CoRR*, abs/1907.06347.

Amir Hadifar, Lucas Sterckx, Thomas Demeester, and Chris Develder. 2019. A self-training approach for short text clustering. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 194–199.

K. Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to discover novel visual categories via deep transfer clustering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8400–8408.

Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2018. Learning to cluster in order to transfer across domains and tasks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind your outliers! investigating the negative impact of outliers on active learning for visual question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7265–7281.

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. Association for Computational Linguistics.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 3–12.

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, Singapore. Association for Computational Linguistics.

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.

Ming Liu, Wray L. Buntine, and Gholamreza Haffari. 2018. Learning how to actively learn: A deep imitation learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1874–1883.

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

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. *ArXiv*, abs/2305.14264.

Yutao Mou, Keqing He, Pei Wang, Yanan Wu, Jingang Wang, Wei Wu, and Weiran Xu. 2022. 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, 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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 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 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 606–610.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2022. A survey of deep active learning. *ACM Comput. Surv.*, pages 180:1–180:40.

Bernhard Schölkopf, Kah Kay Sung, Christopher J. C. Burges, Federico Girosi, Partha Niyogi, Tomaso A. Poggio, and Vladimir Vapnik. 1997. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.*, pages 2758–2765.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2194–2203.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

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

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2022. Generalized category discovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7482–7491.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering.

Dan Wang and Yi Shang. 2014. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks, IJCNN 2014, Beijing, China, July 6-11, 2014*, pages 112–119.

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. *arXiv preprint arXiv:2201.07604*.

Rui Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. Freeal: Towards human-free active learning in the era of large language models. *ArXiv*, abs/2311.15614.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 478–487.

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

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3861–3870.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. Generalized out-of-distribution detection: A survey. *CoRR*, abs/2110.11334.

Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen, and Chao Zhang. 2023. Cold-start data selection for better few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. In *ACL*, pages 2499–2521.

Weihao Zeng, Keqing He, Zechen Wang, Dayuan Fu, Guanting Dong, Ruotong Geng, Pei Wang, Jingang Wang, Chaobo Sun, Wei Wu, and Weiran Xu. 2022. Semi-supervised knowledge-grounded pre-training for task-oriented dialog systems. In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 39–47.

Xueying Zhan, Qingzhong Wang, Kuan-Hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B. Chan. 2022. A comparative survey of deep active learning. *CoRR*, abs/2203.13450.

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.

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.

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

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023b. Llmaaa: Making large language models as active annotators. *ArXiv*, abs/2310.19596.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023c. Clusterllm: Large language models as a guide for text clustering. *arXiv preprint arXiv:2305.14871*.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023d. Clusterllm: Large language models as a guide for text clustering.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2022a. 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.

Zhisong Zhang, Emma Strubell, and Eduard H. Hovy. 2022b. A survey of active learning for natural language processing. In *EMNLP*, pages 6166–6190.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. 2021. Neighborhood contrastive learning for novel class discovery. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10867–10875.

Yunhua Zhou, Guofeng Quan, and Xipeng Qiu. 2023. A probabilistic framework for discovering new intents. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3771–3784.

A Appendix

A.1 Dataset Statistics

We show the detailed statistics of BANKING, CLINC and StackOverflow datasets in Table 5. Specifically, BANKING is a fine-grained category discovery dataset collected from user dialogues in banking domain. It contains over 13K user utterances that span over 77 distinct categories. CLINC is a multi-domain dataset, which encompasses 150 distinct categories and 22,500 utterances across 10 domains. StackOverflow is a technical question dataset collected from Kaggle.com, which includes 20K questions with 20 categories.

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

Table 5: Statistics of datasets used in the experiments.

A.2 Implementation Details

For the dataset setup, following (Zhang et al., 2023a), we randomly select a specified ratio {25%, 50%} of categories, denoted as known category rate (KCR), to serve as known categories. For each known category, 10% of labeled samples are selected to constitute a labeled dataset $\mathcal{D}_l$, while the remaining samples are deemed as the unlabeled data, thereby forming the unlabeled dataset $\mathcal{D}_u$.

For the Uncertainty Propagation, we set the freedom α in Equation (1) to 1.0. The number of propagated neighbors is specifically set to 25 for all datasets. The ρ for calculating similarities in Equation (4) is set to 1.0.

For the Comparison-based Prompting, we employ the *gpt-3.5-turbo* as the basic LLM in our experiments. While acquiring supervision signals, the temperature is set to 0 for deterministic outputs, and the maximum tokens are constrained to 256. The default values are retained for the rest of parameters. The number of representative samles is specifically set to 38 for the BANKING dataset, 75 for the CLINC dataset, and 20 for the StackOverflow dataset.

A.3 Evaluation Metrics

In the experiments, we employ three standard evaluation metrics: ACC, ARI, and NMI to evaluate the GCD performance. Specifically, ACC measures the performance of GCD by comparing the predicted labels with the ground-truth labels. The definition of ACC is as follows:

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

where $\{\hat{y}_i, y_i\}$ denote the predicted label and the ground-truth label for a given sample x_i respectively. $map(\cdot)$ is a mapping function that maps each predicted label $\hat{y}_i$ to its corresponding ground-truth label y_i by Hungarian algorithm.

ARI calculates the similarity between the predicted and ground-truth clusters, assessing the accuracy of clustering on a pairwise basis. ARI is defined as:

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

where $u_i = \sum_j n_{i,j}$, and $v_j = \sum_i n_{i,j}$. N denotes the number of all samples. $n_{i,j}$ is the number of sample pairs that are both assigned to i^{th} predicted cluster and j^{th} ground-truth cluster.

NMI computes the normalized mutual information to quantify the agreement between the predicted and ground-truth clusters, providing a measure of clustering consistency. It can be calculated as follows:

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

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.4 Baselines

In this work, we compare the proposed ALLUP with the following representative baselines:

- **DTC** (Han et al., 2019): A semi-supervised deep clustering approach with a novel mechanism for estimating the number of intents based on labeled data.
- **CDAC+** (Lin et al., 2020): A pseudo-labeling approach that employs pairwise constraints and a target distribution as guiding factors in the learning of new categories.
- **DeepAligned** (Zhang et al., 2021b): A semi-supervised approach that addresses the clustering inconsistency problem by using an alignment strategy for learning utterance embeddings.

Cluster Num	Methods	Banking77		
		ACC	ARI	NMI
$K = 77$ (gold)	USNID	65.85	56.53	81.94
	CsePL	71.06	60.36	83.32
	ALUP	**74.61**	**62.64**	**84.06**
$K = 74$ (predicted)	USNID	60.72	49.18	78.11
	CsePL	69.75	56.70	81.30
	ALUP	**72.55**	**61.04**	**82.78**

Table 6: Effect of estimating cluster number K.

- **ProbNID** (Zhou et al., 2023): A probabilistic framework that capitalizes on the expectation-maximization algorithm, conceptualizing intent assignments as probable latent variables.
- **DCSC** (Wei et al., 2022): A pseudo-labeling method involving the dual-task, which uses the SwAV algorithm and Sinkhorn-Knopp (Cuturi, 2013) to assign soft clusters.
- **MTP-CLNN** (Zhang et al., 2022a): A two-stage method that enhances representation learning via a multi-task pre-training and a nearest neighbor contrastive learning for identifying new categories.
- **USNID** (Zhang et al., 2023a): A framework supports both unsupervised and semi-supervised new intent discovery, incorporating an effective centroid initialization strategy designed to learn cluster representations by utilizing historical clustering information.
- **CsePL** (Liang and Liao, 2023): A method that utilizes two-level contrastive learning with label semantic alignment to enhance the cluster semantics, and a soft prompting strategy for discovering new intents.

We re-run the released code of ProbNID to get its results. The other baselines' results are retrieved from Zhang et al. (2023a).

B Estimate the Category Number K

In the complex task of generalized category discovery in real-world scenarios, accurately predicting the total number of categories, represented as K, remains a significant challenge. Drawing from the methodologies proposed by Zhang et al. (2021b), our research leverages pre-initialized intent features to autonomously determine K. We begin by assigning an initially large number of clusters, K', and then utilize a refined model to extract feature representations from our training dataset. These representations are grouped into distinct clusters using the K-means algorithm. Clusters that are densely populated and demonstrate well-defined boundaries are recognized as valid category clusters. Conversely, smaller, less distinct clusters are considered less relevant and subsequently discarded. The selection criteria for this process can be outlined as follows.

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

where $|S_i|$ is the i-th grouped cluster size, ρ is the filtering threshold. $\delta(\cdot)$ denotes the indicator function, whose output is 1 if the condition is satisfied.

Experimental results are reported in Table 6. The comparative results show that the proposed ALUP incurs only a minor performance decline with predicted category number. This indicates that our ALUP exhibits robustness in handling inaccurately predicted category number.