GenDis: Generative-Discriminative Dual-View Co-Training for Generalized Category Discovery

Xi Chen[1,2], **Chuan Qin**[2,3]*, **Jinpeng Li**[2,3], **Shasha Hu**[1], **Chao Wang**[1], **Hengshu Zhu**[2,3], **Hui Xiong**[4]*
[1] University of Science and Technology of China
[2] Computer Network Information Center, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences
[4] The Hong Kong University of Science and Technology (Guangzhou)
{chenxi0401, shashahu}@mail.ustc.edu.cn, {chuanqin0426, zhuhengshu}@gmail.com,
lijinpeng25@mails.ucas.ac.cn, wangchaoai@ustc.edu.cn, xionghui@ust.hk

Abstract

Generalized Category Discovery (GCD) aims to identify both known and novel categories from partially labeled data, reflecting more realistic open-world learning scenarios. However, most existing methods rely solely on one-hot discriminative supervision, leading to overfitting on seen classes and poor generalization to unseen ones. Recent advances introduce large language models (LLMs) to incorporate external semantics, yet they often suffer from semantic–label misalignment and weak semantic integration during training. We propose GenDis, a Generative–Discriminative Dual-View Co-Training framework that unifies discriminative classification and semantic label generation within an LLM. Discriminative pseudo-labels guide the formation of a separable generative latent space, enabling semantically meaningful supervision for novel classes. To ensure consistency between the two views, we employ Canonical Correlation Analysis (CCA)-based alignment and a curriculum-guided, dispersion-aware pseudo-labeling strategy for iterative refinement. Extensive experiments on five GCD benchmarks demonstrate that GenDis substantially outperforms prior methods, validating the effectiveness of dual-view co-training with semantically enriched supervision. Code is available at https://github.com/cx9941/GenDis.

1 Introduction

Although modern machine learning methods have achieved impressive performance on numerous tasks, they often rely on large-scale annotated datasets and operate under a closed-world assumption (Wang et al., 2024), that the data to be classified shares the same set of categories as the labeled training data, limiting their applicability to real-world scenarios. To address this, GCD (Vaze et al., 2022) was proposed, aiming to train models on partially labeled data to classify unlabeled samples that may originate from previously unseen categories. Reflecting the dynamic and continuously evolving nature in the real world, GCD has gained traction in practical applications, such as intent discovery in dialogue systems (Lin et al., 2020; Zhang et al., 2021) and novel product type identification on e-commerce platforms (Gong et al., 2023).

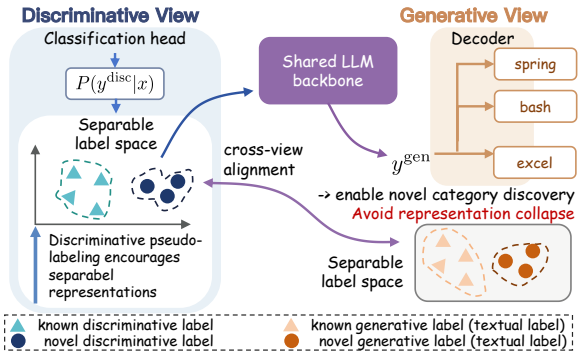

Figure 1: Illustration of our dual-view design.

GCD was first formalized by Vaze et al. (2022) with a two-stage approach: representation learning on labeled data and clustering-based pseudo-labeling for unlabeled data. However, noisy pseudo-labels limit performance. Subsequent works have explored solutions including regularization (Zhang et al., 2022, 2023), calibration (An et al., 2023a; Tang et al., 2024), and prototype-based methods (An et al., 2024b; Shi et al., 2024; An et al., 2025). These methods rely solely on discriminative one-hot labels, lacking explicit semantic guidance, which constrains the model to the training data distribution and hinders its generalization to novel classes. Recent efforts attempt to inject external semantics using LLMs. These approaches sample centroids after clustering and either (1) query pairwise similarities between ambiguous samples and sampled centroids (An et al., 2024a; Liang et al., 2024), or (2) construct pseudo-

*Corresponding Authors

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2330–2351
July 2-7, 2026

labels based on the semantic descriptions of sampled centroids to filter hard examples (Zhang et al., 2024; Zou et al., 2025). These additional semantic signals are crucial because they introduce generalizable information beyond the distribution of training data, thus improving the model's supervision quality and enhancing its generalization for GCD. However, these LLM-enchanced methods rely on sampled centroids, leading to mismatches between queried semantics and true underlying label meanings. Besides, these methods provide only implicit structural supervision to discriminative models but not semantic labels, limiting the effectiveness of semantic transfer and ultimately remaining constrained by the shortcomings of traditional discriminative training on one-hot labels.

This motivates the question: how can we effectively enhance the consistency of provided semantic signals and directly incorporate such generalizable signals into model training? A naive idea is to directly fine-tune LLMs on a conditional text generation task reformulated from the classification objective (Liu et al., 2024b). Through generative fine-tuning, LLMs can perceive the global label space and output semantic pseudo-labels for unlabeled data. However, such LLMs trained only on known categories are prone to representation collapse (Xiao et al., 2024), where representations become overly aligned with known label tokens (Simig et al., 2022; Kumar et al., 2023), leading to poor generation for novel classes. Moreover, generative reformulations reduce the discriminative ability of the model, harming decision precision.

To address these limitations, we propose a novel co-training perspective that jointly learns discriminative and generative views. Specifically, we augment the LLM with an additional MLP head for discriminative classification and construct a dual-view framework that facilitates both semantic novel labels and category separation. As shown in Figure 1, the classification head encourages learning class-separable features, which can also guide the separable latent space for semantic novel label generation. This co-training framework overcomes the constraints of traditional one-hot labeling by providing semantically enriched, generalizable supervision that extends beyond the distribution of existing data, thereby enhancing model generalization for unseen classes. However, implementing such a framework still introduces two major challenges. First, the discriminative and generative objectives promote different feature characteristics—class separation versus semantic fluency. Co-training must align the views without sacrificing their individual strengths. Second, the semantic pseudo-labels are directly involved in model training; thus, mitigating noise in these labels is critical.

To this end, we propose GenDis built upon LLMs for GCD. Specifically, the LLM is initialized via **dual-objective fine-tuning**, with category-aware generative fine-tuning for text generation and token-level discriminative fine-tuning for classification. This equips the model with the capability to both generate and recognize category labels. Subsequently, the model is extended to unlabeled data through **discriminative pseudo-label learning**, which is progressively refined via curriculum-guided labeling. In addition, we employ CCA regularization to achieve view alignment while maintaining functional independence between the generative and discriminative views. Furthermore, we propose **dual-view pseudo-label learning**, including discriminative-guided generative pseudo-labeling based on the aligned latent space with a dispersion-aware filtering mechanism to ensure label reliability. Finally, GenDis enhances GCD by incorporating more generalizable semantically aligned supervision into the pseudo-label learning.

Our main contributions are as follows:

- We propose a novel co-training perspective that integrates more generalizable semantic pseudo-labels into the training process, breaking the limitations of conventional discriminative learning for better representations on novel categories.
- We introduce a comprehensive framework GenDis, which incorporates CCA-based view alignment for effective collaboration between two views and a dispersion-aware pseudo-labeling mechanism to keep label reliability.
- We conduct extensive experiments on five GCD benchmarks, achieving higher ARI than the state-of-the-art (SOTA) methods by around 11.90%, which demonstrates the superiority of GenDis.

2 Related Works

Generalized Category Discovery GCD was first systematically formalized by Vaze et al. (2022), who proposed a two-stage training paradigm for this task: representation learning on known categories, followed by clustering-based pseudo-labeling for novel categories. Prior to this formalization, related works such as CDAC+ (Lin

et al., 2020) and DAL (Zhang et al., 2021) employed pseudo-supervision based on pairwise similarity—either as weak supervision or through alignment strategies as strong supervision—to guide clustering for novel intent discovery. The main limitation of this paradigm is that inaccurate pseudo-labels can significantly degrade model performance. Consequently, subsequent studies have focused on mitigating pseudo-label noise through various techniques, including regularization methods MTP (Zhang et al., 2022), USNID (Zhang et al., 2023) to improve generalization, pseudo-label calibration PTJN (An et al., 2023a) GeoID (Tang et al., 2024) to reduce dominance by known categories, and prototype-based approaches TAN (An et al., 2024b), KTN (Shi et al., 2024), SDC (An et al., 2025) to enhance knowledge transfer between known and novel classes.

LLM for Generalized Category Discovery Recently, LLMs have demonstrated strong capabilities across diverse domains, including talent analytics, scientific discovery, information retrieval, enterprise knowledge bases, and real-world decision-making (Qin et al., 2025c,a; Huang et al., 2026; Tong et al., 2025; Jiang et al., 2024; Song et al., 2026) In scenarios involving open or evolving categories, prior studies have explored LLMs and continual learning techniques for handling unseen classes, such as unknown-aware open-set text classification (Chen et al.), incremental task recognition (Qin et al., 2025b), and datasets supporting evolving semantic categories (Chen et al., 2024).

LLMs have been introduced into GCD to leverage their rich semantic knowledge. Existing methods typically use LLMs as auxiliary modules to assist discriminative models (e.g., BERT) in pseudo-label construction. Representative works include LOOP (An et al., 2024a), which employs neighborhood contrastive learning via scalable LLM querying, ALUP (Liang et al., 2024), which uses comparison-based prompting for category matching, and GLEAN (Zou et al., 2025), which integrates quality-controlled LLM feedback for robust discovery. However, these methods typically use LLMs as auxiliary modules, constrained by input length and a lack of domain adaptation.

In contrast, we are the first to take the LLM as the backbone instead of auxiliary tools and propose an LLM-based framework GenDis for GCD.

3 Preliminaries

3.1 Problem Definition

Models trained on a labeled dataset $\mathcal{D}^l = \{(x, y) \mid y \in \mathcal{Y}^k\}$ are proficient at identifying predefined known categories $\mathcal{Y}^k$. However, in real-world scenarios, these models often encounter unlabeled data $\mathcal{D}^u = \{x \mid y \in \mathcal{Y}^k \cup \mathcal{Y}^n\}$ that encompasses both known categories $\mathcal{Y}^k$ and novel categories $\mathcal{Y}^n$, where $\mathcal{Y}^k \cap \mathcal{Y}^n = \emptyset$, potentially leading to identification failure. To overcome this challenge, GCD requires models to recognize both known and novel categories based on $\mathcal{D}^l$ and $\mathcal{D}^u$. We assume the number of novel categories $|\mathcal{Y}^n|$ is known following previous work (An et al., 2023b; Shi et al., 2024). Finally, the model's performance will be evaluated on a testing set $\mathcal{D}^t = \{(x, y) \mid y \in \mathcal{Y}^k \cup \mathcal{Y}^n\}$ in an inductive manner.

3.2 Discriminative Fine-tuning

Discriminative fine-tuning (Liu et al., 2024c; Qin et al., 2023) is a widely adopted approach for adapting pre-trained language models (PLMs) to classification tasks. Given an input text x and a candidate label set $\mathcal{Y}^k$, the model first encodes the input into hidden representations using a backbone PLM (e.g., BERT). A fixed-length vector representation $\mathbf{h}$ is then extracted—typically the [CLS] token embedding in encoder-based models like BERT or the final token embedding in decoder-only models. A classification head is subsequently applied to compute a score $\mathbf{s}_\theta(x, y)$ for each label y, usually through a linear projection. The model is trained by minimizing the cross-entropy loss, which maximizes the log-likelihood of the correct label y:$\max \log \mathbf{P}(y \mid x) = \log \frac{\exp(\mathbf{s}_\theta(x,y))}{\sum_{i'=1}^{K} \exp(\mathbf{s}_\theta(x,y'))}$, where θ denotes the trainable parameters. This formulation encourages the model to learn discriminative decision boundaries between labels and is particularly effective when the label set is closed.

3.3 Generative Fine-tuning

Generative fine-tuning (Liu et al., 2024c) harnesses the inherent generative capabilities of LLMs for text classification by formulating the task as a conditional text generation problem. Given an input x, the input is first embedded into a prompt template to facilitate label generation. A typical format is: "## Input: x. ## Output: y". The objective is to maximize the likelihood of generating the label tokens y:$\max \mathbf{P}(y \mid x) = \prod_{i=1}^{|y|} \mathbf{P}_\theta(\mathrm{y} \mid x, \mathrm{y}_{<i})$,

where $|\mathrm{y}|$ is the number of label tokens, and $\mathrm{y}_{<i}$ represents generated tokens prior to y_i.

4 Method

In this section, we introduce the proposed GenDis for GCD. We begin by presenting the generative-discriminative dual-view representation network built upon an LLM, and then describe the co-training pipeline.

4.1 Dual-View Representation Network

To simultaneously enable LLMs to explicitly distinguish class boundaries and retain the ability to understand and generate label semantics, we first construct a dual-view representation network based on LLMs. As illustrated in Figure 2.a, before the final classification layer (LLM Head), the hidden representation from the LLM is extracted as the original generative representation $\mathbf{h}^{\text{gen}}$. This representation is subsequently passed through a multi-layer perceptron (MLP) to obtain the discriminative representation $\mathbf{h}^{\text{dis}}$. While $\mathbf{h}^{\text{gen}}$ supports the model's generative capability, $\mathbf{h}^{\text{dis}}$ is used for classifying samples via a task-specific classifier.

4.2 Dual-Objective Fine-Tuning

As the initial stage of training, this phase aims to initialize the LLM with training data, an awareness of text formatting and discourse style, and the ability to differentiate representation spaces across categories. It serves as a warm-up phase for subsequent pseudo-label learning.

Category-Aware Generative Fine-Tuning. We design a general prompt template for the GCD task (see Appendix B.1), forming a conversation set $\mathcal{Z} = \{z = Prompt(x, y) | (x, y) \in \mathcal{D}^l \cup \mathcal{D}^u\}$, where y is a placeholder for unlabeled samples in $\mathcal{D}^u$. This template serves as input for LLM fine-tuning. Unlike prior approaches that design task-specific prompts for particular scenarios (e.g., intent detection), our prompt template is unified and adaptated to diverse text classification tasks.

Subsequently, we fine-tune the LLM on the constructed conversion set $\mathcal{Z}$ using a generative loss $\mathcal{L}_{\text{gen}}$, aiming to inject initial semantic grounding and encourage the model to generate appropriate category labels based on input text. For each instance, we define two objectives. First, an autoregressive loss is defined as $\mathcal{L}_{\text{auto}}(z) = -\sum_{i=1}^{|\mathrm{z}|} \log \mathbf{P}_\theta(\mathrm{z}_i \mid \mathrm{z}_{<i})$, where z denotes the tokens of the conversion z. Second, to emphasize learning the label generation process, we additionally incorporate a conditional generation loss over the label tokens: $\mathcal{L}_{\text{con}}(x, y^{\text{gen}}) = -\sum_{i=1}^{|\mathrm{y}^{\text{gen}}|} \log \mathbf{P}_\theta(\mathrm{y}_i \mid x, \mathrm{y}^{\text{gen}}_{<i})$, where y^{gen} is the generative label of y, and y^{gen} is the tokens of the label y. The final generative loss $\mathcal{L}^{(1)}_{\text{gen}}$ is defined as follows:

$$\mathcal{L}_{\text{gen}}(x, y^{\text{gen}}) = \mathcal{L}_{\text{auto}}(z) + \mathcal{L}_{\text{con}}(x, y^{\text{gen}}). \quad (1)$$

Token-level Discriminative Fine-Tuning. To ensure that the learned representations are semantically discriminative, we leverage the hidden representation at the token position immediately preceding label generation, denoted as $\mathbf{h}^{\text{gen}}$. This representation is projected into a discriminative space $\mathbf{h}^{\text{dis}}$ via a lightweight multi-layer perceptron (MLP), followed by a classification layer that computes the class scores $\mathbf{s}_\theta(x^l, y^{\text{dis}})$ over the known label set $\mathcal{Y}^k$. For each labeled instance $(x^l, y^{\text{dis}}) \in \mathcal{D}^l$, the model is trained to minimize the standard cross-entropy loss $\mathcal{L}^{(1)}_{\text{dis}}$:

$$\mathcal{L}_{\text{dis}}(x^l, y^{\text{dis}}) = -\log \frac{\exp(\mathbf{s}_\theta(x^l, y^{\text{dis}}))}{\sum_{y' \in \mathcal{Y}^k} \exp(\mathbf{s}_\theta(x, y'))}. \quad (2)$$

Finally, two objectives are intergrated as $\mathcal{L}_{\text{DualFT}} = \mathcal{L}^{(1)}_{\text{dis}} + \lambda_{\text{gen}} \mathcal{L}^{(1)}_{\text{gen}}$, where λ_{gen} is a scaling hyperparameter for LLM fine-tuning.

4.3 Discriminative Pseudo-Label Learning

Since the initial classification supervision covers only a subset of known classes, the LLM suffers from representation collapse (Xiao et al., 2024), where the representations of known and novel classes become entangled, causing the model to generate only known labels. To address this issue, this stage aims to enforce a clear separation not only among known categories but also among unknown ones at both the generative and discriminative representation levels.

Curriculum-Guided Discriminative Pseudo-Labeling. We extend training to the unlabeled set $\mathcal{D}^u$ by generating discriminative pseudo-labels $\hat{y} \in \mathcal{Y}^k \cup \mathcal{Y}^u$ (i.e., cluster indices) via K-means clustering in the learned discriminative latent space $\mathbf{h}^{\text{dis}}$ and Hungarian algorithm (Kuhn, 2004) for alignment between each epoch.

To mitigate the influence of noisy pseudo-labels, we estimate sample confidence through distance-based uncertainty and adopt a curriculum strategy that gradually incorporates high-confidence

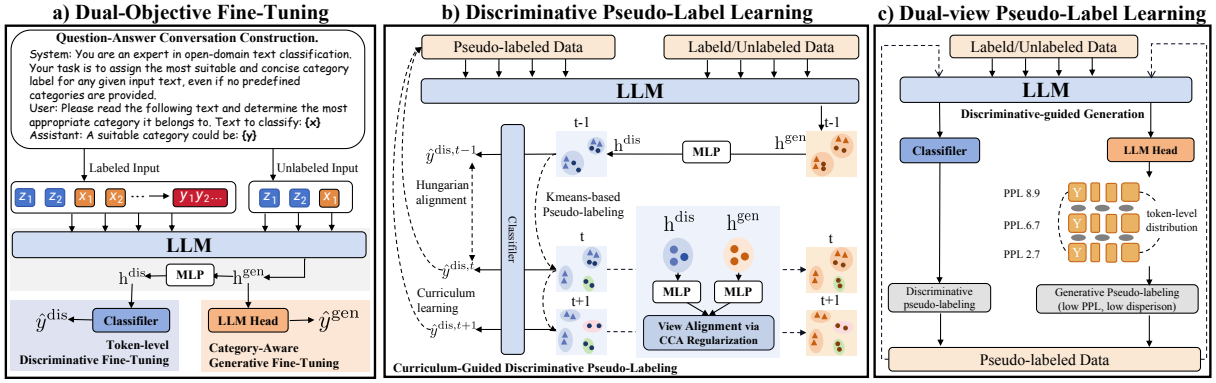

Figure 2: Illustration of the proposed GenDis framework.

instances. For each cluster with prototype $\mathbf{p}$, we measure the cosine similarity between instance $\mathbf{h}_i^{\text{dis}}$ and $\mathbf{p}_j$: $\mathrm{M}_{i,j} = \frac{\langle \mathbf{h}_i^{\text{dis}}, \mathbf{p}_j \rangle}{|\mathbf{h}_i^{\text{dis}}| \cdot |\mathbf{p}_j|}$. Top-$k$ confident samples per cluster are progressively added to the training set, forming an easy-to-hard curriculum that stabilizes optimization. The loss $\mathcal{L}_{\text{dis}}^{(2)}$ is defined as:

$$\mathcal{L}_{\text{dis}}^{(2)} = \mathcal{L}_{\text{dis}}(x^l, y^{\text{dis}}) + \mathcal{L}_{\text{dis}}(x^u, \tilde{y}^{\text{dis}}), \quad x^u \in \mathcal{D}^u, \tag{3}$$

where high-confidence unlabeled samples $x^u \in \mathcal{D}^u$ are iteratively incorporated.

View Alignment via CCA Regularization. While the discriminative representation $\mathbf{h}^{\text{dis}}$ enables category separation, the generative representation $\mathbf{h}^{\text{gen}}$ primarily captures linguistic fluency, leading to potential semantic misalignment between the two. To align them at a category-relevant level while preserving their distinct roles, we introduce a Canonical Correlation Analysis (CCA)–based regularization term (Tang et al., 2018). This term projects $\mathbf{h}^{\text{dis}}$ and $\mathbf{h}^{\text{gen}}$ into a shared latent space and maximizes their cross-view correlation, encouraging the two views to encode consistent yet complementary semantics. The overall loss combines the discriminative, generative, and alignment objectives:

$$\mathcal{L}_{\text{DisPL}} = \mathcal{L}_{\text{dis}}^{(2)} + \lambda_{\text{gen}} \mathcal{L}_{\text{gen}}^{(1)} + \lambda_{\text{cca}} \mathcal{L}_{\text{cca}}, \tag{4}$$

where λ_{cca} controls the alignment strength. This joint optimization ensures semantic consistency between $\mathbf{h}^{\text{dis}}$ and $\mathbf{h}^{\text{gen}}$, allowing the LLM to discover novel categories that are both structurally separable and linguistically coherent. The detailed CCA computation is provided in Appendix B.4.

4.4 Dual-view Pseudo-Label Learning

Building on the structured generative space, we generate semantically meaningful pseudo-text labels for unlabeled samples. A dispersion-aware generative pseudo-labeling strategy with multi-sample inference is adopted to filter unreliable labels. Finally, generative and discriminative pseudo-supervision are jointly optimized to enhance semantic generalization in GCD.

Dispersion-aware Generative Pseudo-label Learning To capture uncertainty, we perform multi-sample inference, generating d candidate labels $\{Y^{(i)}\}_{i=1}^{d}$ for each input instance x. Each candidate label $Y^{(i)}$ is associated with token-level probability distributions $P^{(i)} = (p_1^{(i)}, \ldots, p_T^{(i)})$, where $p_t^{(i)} \in \Delta^{|\mathcal{V}|}$ denotes the softmax output at position t over a vocabulary of size $|\mathcal{V}|$. Then we compute Wasserstein distance (Matsubara, 2024) between for any token pair $(Y^{(i)}, Y^{(j)})$:

$$W(P^{(i)}, P^{(j)}) = \inf_{\gamma \in \Pi(P^{(i)}, P^{(j)})} \mathbb{E}_{(u,v)\sim\gamma} \left[\|u - v\|_2 \right], \tag{5}$$

where $\Pi(P^{(i)}, P^{(j)})$ denotes the set of all valid joint distributions with marginals $P^{(i)}$ and $P^{(j)}$, and $\|\cdot\|_2$ is the Euclidean distance.

In addition, to assess the fluency and confidence of each generated label, we compute its perplexity (PPL) (Brown et al., 2020), where a lower PPL indicates a more fluent and confident generation. By jointly considering the average pairwise Wasserstein distance and the perplexity score, we select the generative pseudo-label with the lowest PPL among those whose semantic dispersion falls within a predefined threshold.

Finally, the selected generative pseudo-label $\hat{y}^{\text{gen}}$ is combined with the input x using the same template to form a QA-style input for generative fine-tuning. The corresponding loss is computed as: $\mathcal{L}_{\text{gen}}^{(2)} = \mathcal{L}_{\text{gen}}(x, y^{\text{gen}}) + \mathcal{L}_{\text{gen}}(x^u, \hat{y}^{\text{gen}})$. The overall objective of this component integrates both the

generative and discriminative pseudo-label loss and the CCA regularization term:

$$\mathcal{L}_{\text{DualPL}} = \mathcal{L}_{\text{dis}}^{(2)} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}}^{(2)} + \lambda_{\text{cca}}\mathcal{L}_{\text{cca}}. \quad (6)$$

4.5 Test-time Inference

During inference, the LLM extracts both discriminative and generative representations. For each representation view, we perform K-means clustering, followed by Sinkhorn-Knopp (SK) normalization (Cuturi, 2013) to transform the raw distance matrix into a balanced probability distribution over classes. The detailed procedure is provided in Appendix B.3. To determine which view produces better clustering quality, we evaluate the clustering accuracy (e.g., K-ACC) of both views on a held-out validation set containing only known-class labels. The view with higher clustering performance is then used to produce the final cluster assignments.

5 Experiments

5.1 Experimental Settings

Datasets. We evaluate our proposed method on five widely-used GCD benchmarks. **BANKING** (Casanueva et al., 2020) is a fine-grained intent classification dataset comprising 77 distinct intents collected from real-world banking dialogues. **CLINC** (Larson et al., 2019) is a public dataset designed for intent detection across various domains. **HWU** (Liu et al., 2021) focuses on personal assistant queries, offering diverse intent classes. **MCID** (Levy and Wang, 2020) is a compact dataset built around COVID-19 related services. **StackOverflow** (Xu et al., 2015) is a dataset derived from technical question titles on Stack Overflow.

Dataset Split. To simulate open-world scenarios, we define two key parameters: the Known Class Ratio (KCR), representing the proportion of categories treated as known during training, and the Labeled Ratio (LAR), denoting the fraction of labeled samples provided for each known category. All labeled examples are randomly sampled from the training set. To ensure fair comparisons, our dataset partitioning strictly follows the settings established in prior works (An et al., 2024a). We report experimental results under varying KCR values of {25%, 50%, 75%} with a fixed LAR of 10%. For each run, we designated the KCR portion of classes as known, and the sample LAR portion of training data per known class as labeled. The LLM was first trained on this partially labeled training set (comprising both labeled and unlabeled data). It is then tuned using a validation set split from the training set containing only known categories, and finally evaluated on the full test set. Detailed dataset statistics are provided in Appendix Table S1.

Baselines. We compare GenDis against a comprehensive set of strong and diverse baselines. These include a wide range of representative methods, covering both **LLM-free methods** (e.g., DeepAligned (Zhang et al., 2021), DPN (An et al., 2023b), PTJN (An et al., 2023a), GeoID (Tang et al., 2024), TAN (An et al., 2024b), KTN (Shi et al., 2024), SDC (An et al., 2025)) and recent **LLM-enhanced methods** (e.g., LOOP (An et al., 2024a), ALUP (Liang et al., 2024), GLEAN (Zou et al., 2025)). For fair comparison, we adopt BERT as the backbone for discriminative models, and Qwen2.5-7B-Instruct (Qwen Team, 2024) as the backbone for generative models. In particular, we utilize Deepseek-V3-0324 (Liu et al., 2024a), one of the current state-of-the-art LLMs, as the auxiliary model for all LLM-enhanced methods. Due to space limitations, detailed descriptions of each baseline method can be found in Appendix D.3, and additional experimental results using different LLM backbones are provided in Appendix E.2

Evaluation Metrics. Following prior works (An et al., 2024a), we adopt six evaluation metrics to comprehensively assess clustering performance in the GCD setting: (1) **ARI** assesses pairwise consistency between the clustering and ground-truth partitions; (2) **NMI** evaluates the mutual information between predicted clusters and ground-truth labels; (3) **K-ACC** and (4) **N-ACC** measure clustering accuracy on known and novel classes respectively; (5) **H-score** computes the harmonic mean of K-ACC and N-ACC; and (6) **ACC** indicates overall clustering accuracy across all classes.

Due to page limitations and metric redundancy, we report only **ARI**, **NMI**, **K-ACC**, and **N-ACC** in the main text, for H-score and ACC are derived from K-ACC and N-ACC through different averaging strategies and thus offer limited additional insight. Detailed results with H-Score and ACC are provided in Appendix Table S2.

5.2 Overall Performance.

Table 1 presents the performance comparison between our proposed GenDis and a range of SOTA baselines across five datasets and three different KCRs. The results demonstrate that

Table 1: Performance of various methods across five datasets under different KCRs. Metrics include ARI, NMI, K-ACC, and N-ACC. All results are averaged over four independent runs to ensure fairness. The best results are highlighted in bold, while the second-best results are underscored.

Dataset		BANKING				CLINC				HWU				MCID				StackOverflow			
KCR	Method	ARI	NMI	K-ACC	N-ACC	ARI	NMI	K-ACC	N-ACC	ARI	NMI	K-ACC	N-ACC	ARI	NMI	K-ACC	N-ACC	ARI	NMI	K-ACC	N-ACC
0.25	DeepAligned	34.54	68.68	58.67	42.44	65.96	89.40	86.46	72.00	42.52	75.22	63.86	50.61	18.56	44.64	38.95	35.30	42.98	54.91	82.83	50.49
	DPN	34.49	73.15	57.64	41.81	65.08	92.32	85.79	70.22	41.23	77.12	70.59	43.81	23.64	50.05	53.20	38.67	47.04	60.10	85.04	52.83
	PTJN	39.75	71.94	61.33	48.88	70.82	91.20	87.68	76.36	45.19	76.29	66.70	53.97	22.59	47.70	40.70	40.20	55.92	68.96	80.07	68.59
	GeoID	24.30	57.74	39.66	34.12	77.72	93.54	90.00	82.81	54.13	80.31	67.83	67.39	26.60	50.92	29.65	50.41	55.47	69.67	75.97	71.40
	TAN	28.35	64.19	59.50	36.32	46.45	81.71	80.10	50.86	29.01	66.70	64.02	37.56	18.95	45.55	39.77	38.94	28.04	41.68	82.65	33.53
	KTN	27.91	66.08	66.98	33.55	59.00	88.17	93.79	62.60	43.50	76.20	71.80	54.92	13.89	39.25	21.40	37.14	61.59	70.47	88.13	74.63
	SDC	44.06	74.09	78.37	48.02	64.80	89.79	89.40	68.09	53.33	79.75	75.82	62.46	14.49	39.64	26.74	35.67	57.28	75.65	84.86	77.27
	LOOP	58.98	82.36	76.23	66.35	82.05	94.85	94.49	84.81	60.32	83.80	75.82	69.77	34.92	58.27	39.53	57.63	64.42	75.55	85.30	76.52
	ALUP	61.92	83.99	70.78	72.78	83.46	95.16	92.24	87.77	63.76	84.57	77.05	73.25	49.08	68.69	47.91	72.33	66.39	75.40	88.81	78.84
	Glean	56.91	82.02	71.15	64.80	80.36	94.67	89.01	84.70	58.02	83.24	70.49	69.76	48.97	68.80	46.12	69.52	64.09	77.15	83.29	76.65
	GenDis	**65.54**	**86.32**	**82.49**	**72.80**	**88.40**	**97.12**	**96.62**	**88.45**	**74.62**	**90.05**	**87.59**	**80.74**	**58.24**	**75.26**	**70.56**	**72.97**	**82.88**	**86.50**	**95.83**	**88.08**
0.50	DeepAligned	47.90	76.76	65.50	53.20	73.06	91.92	86.36	75.97	51.59	79.36	59.64	65.56	31.36	55.19	61.83	39.65	59.13	69.06	77.46	70.97
	DPN	46.96	79.48	74.87	39.33	73.90	94.26	90.80	69.29	50.75	81.46	67.77	56.47	27.64	54.08	62.04	32.40	64.31	73.16	85.04	74.76
	PTJN	51.74	78.64	74.19	54.42	76.76	93.08	89.38	78.84	54.94	80.69	66.65	67.84	33.17	57.18	52.01	53.11	58.25	72.91	78.12	67.50
	GeoID	57.46	81.48	77.10	59.70	81.77	94.74	92.78	82.53	58.73	82.74	71.48	67.79	28.77	52.32	45.06	50.44	62.20	75.17	82.22	71.31
	TAN	45.12	74.68	69.53	43.87	64.44	89.05	85.62	60.66	38.79	73.17	62.08	46.32	24.84	50.04	50.25	41.07	42.74	56.14	82.80	37.88
	KTN	49.97	78.13	81.39	44.51	74.69	93.37	93.46	68.34	58.06	82.16	73.32	68.17	14.06	39.64	36.05	30.77	61.85	74.79	83.75	74.51
	SDC	53.82	80.31	76.35	53.80	72.78	92.60	88.13	74.59	60.52	83.27	72.10	73.79	49.73	69.02	63.21	70.89	61.82	75.11	82.62	75.31
	LOOP	62.71	84.34	80.77	64.77	83.04	95.20	93.05	82.92	63.55	84.76	74.34	73.49	34.83	57.22	44.08	58.34	68.62	76.76	85.41	80.28
	ALUP	64.57	85.28	84.03	66.22	83.79	95.26	91.93	86.52	65.33	85.64	76.91	75.71	57.64	73.75	71.97	76.81	73.59	78.03	88.62	84.17
	Glean	63.64	85.39	82.31	63.36	75.61	93.74	88.09	79.23	65.98	85.93	74.27	80.28	53.56	71.84	55.56	81.46	63.60	77.18	81.58	73.63
	GenDis	**67.76**	**87.26**	**85.98**	**67.00**	**91.78**	**97.79**	**97.56**	**90.56**	**75.92**	**90.26**	**84.48**	**85.16**	**67.93**	**81.78**	**79.17**	**81.66**	**86.64**	**88.64**	**93.92**	**92.75**
0.75	DeepAligned	54.37	80.16	67.82	59.50	80.70	94.21	89.73	78.33	59.24	82.25	72.95	64.09	48.42	67.50	73.33	34.98	57.71	70.63	78.64	66.86
	DPN	61.33	84.30	79.10	45.18	83.76	95.97	93.23	74.34	66.18	86.46	79.82	66.95	47.22	70.53	68.60	37.97	63.60	75.38	82.86	80.82
	PTJN	60.25	82.03	77.19	59.99	81.70	94.66	91.06	76.36	60.20	82.55	74.34	67.70	43.76	64.14	67.60	43.82	59.20	73.06	75.71	69.54
	GeoID	65.80	85.52	78.69	66.56	87.52	96.27	95.02	83.42	66.11	85.61	75.60	79.83	28.86	52.56	51.20	41.05	65.21	76.21	80.17	78.23
	TAN	56.83	80.98	75.02	50.92	75.55	92.70	89.36	62.17	55.72	80.81	74.67	49.44	35.32	62.37	64.88	39.26	51.83	64.36	80.34	34.35
	KTN	65.46	85.02	81.80	58.89	85.97	96.06	95.29	75.96	67.55	85.99	80.52	73.73	14.39	40.08	31.44	40.00	62.42	75.51	80.54	87.18
	SDC	61.51	83.48	77.60	68.81	79.75	94.43	88.24	85.51	61.26	83.18	72.53	78.43	54.67	73.45	73.92	71.11	60.61	74.10	80.01	82.90
	LOOP	64.39	85.12	75.07	70.42	86.28	95.97	93.26	83.68	63.67	85.06	74.02	72.96	35.61	59.38	53.92	50.37	71.88	78.45	83.69	84.42
	ALUP	70.78	87.39	85.25	66.67	87.34	96.26	93.17	87.30	69.99	87.22	80.25	79.23	60.68	76.41	79.36	64.44	76.06	79.87	88.41	85.87
	Glean	71.96	88.26	82.72	74.49	73.93	93.54	89.39	79.36	70.04	87.54	80.48	79.54	64.47	80.38	78.27	77.37	51.59	72.70	74.20	85.33
	GenDis	**75.26**	**90.69**	**85.41**	**79.14**	**92.49**	**98.05**	**96.60**	**87.63**	**76.68**	**90.54**	**84.45**	**88.30**	**69.36**	**82.42**	**79.70**	**84.26**	**86.51**	**88.93**	**91.81**	**94.62**

GenDis achieves consistent and substantial improvements under all settings. Specifically, for the lowest KCR (0.25), our model improves ARI, K-ACC, N-ACC, and NMI by **14.46%**, **12.34%**, **4.73%**, and **6.56%**, respectively. For KCR = 0.5, the improvements are **13.03%** (ARI), **6.51%** (K-ACC), **4.47%** (N-ACC), and **6.88%** (NMI). Even under the high label regime (KCR = 0.75), where performance gains are typically saturated, GenDis continues to outperform all baselines, achieving improvements of **8.21%**, **2.14%**, **6.93%**, and **4.38%** on the four metrics, respectively. These results confirm the generalization of our method under varying supervision levels. Beyond quantitative gains, we derive several key observations:

Co-training LLM unlocks their full potential beyond LLM-enhanced paradigms: LLM-enhanced baselines (LOOP, ALUP, and Glean) already outperform traditional methods, particularly under low KCR settings, underscoring the potential of LLMs in GCD tasks. Among these, LOOP, which relies on instance-level similarity, provides a weak and indirect supervision signal, whereas ALUP and Glean offer stronger pseudo-label supervision through explicit class labels. However, none of these methods fine-tune the LLM itself, thereby limiting their capacity to align the LLM with label semantics and model for generalization. In contrast, GenDis takes the LLM as the backbone and jointly fine-tunes the LLM with both discriminative and generative supervision signals with pseudo-label learning, facilitating more generalizable representations for GCD, resulting in significant performance.

Strong generalization with limited label supervision: Our model exhibits strong generalization capabilities when only a small portion of known class labels is available, effectively addressing the challenge of discovering novel categories under scarce supervision. As the known class ratio decreases from 0.75 to 0.25, performance degradation is observed across most methods, reflecting the increased difficulty of discovering novel categories with limited prior knowledge. This effect is particularly pronounced in traditional and weakly supervised models, which are prone to overfitting and noise amplification when guided by unreliable pseudo-labels. GenDis, however, mitigates these issues through its co-training of two strong supervision signals, enabling robust and semantically meaningful representation learning even in extremely low-label regimes.

Reliable performance in low-resource datasets: GenDis maintains superior performance even on small-scale datasets, further highlighting

Table 2: Ablation study on HWU dataset with metrics K-ACC and N-ACC. Detailed results across all metrics are put in Appendix E.3 due to the page limitations.

KCR	0.25		0.50		0.75	
Metric	K-ACC	N-ACC	K-ACC	N-ACC	K-ACC	N-ACC
Onlygen	76.84	71.82	72.43	82.69	76.78	83.05
Onlydis	77.87	79.82	75.36	77.63	78.97	85.41
GenDis w/o semi	85.30	78.30	81.06	79.48	80.98	74.11
Gen+semi	72.95	77.47	76.17	77.63	74.22	83.26
Dis+semi	87.30	76.65	76.99	80.41	80.10	79.83
GenDis w/o GPL	85.86	78.30	81.06	77.82	83.35	82.83
GenDis w/o $\mathcal{L}_{cca}$	84.02	78.43	82.54	84.14	83.60	78.11
GenDis w/o curr	76.92	70.48	75.97	73.57	82.48	69.53
GenDis w/o was	84.36	73.81	80.91	81.84	82.23	77.25
GenDis	**87.59**	**80.74**	**84.48**	**85.16**	**84.45**	**88.30**

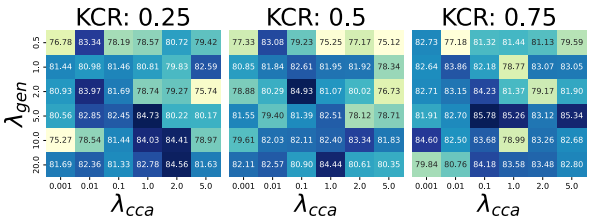

Figure 3: Grid search results of λ_{gen} and λ_{cca} on the HWU dataset under different KCRs with H-Score.

its adaptability to low-resource scenarios. Our method achieves the best results on the CLINC dataset, where the large data volume and rich category hierarchy facilitate effective learning. In contrast, performance on the MCID dataset is relatively lower for all methods due to its limited size and category sparsity, which pose challenges for semi-supervised learning. Nonetheless, our method maintains a clear advantage in this low-resource setting, demonstrating strong generalization across diverse data distributions.

5.3 Abalation Study

To assess the contribution of each component in our framework, we conduct a comprehensive ablation study on the HWU dataset under three different KCRs. Table 2 reports the performance of several variants of GenDis with specific modules removed or altered. We group the variants into three categories: **Non-semi-supervised variants.** We consider three variants that do not incorporate pseudo-label learning: Onlygen, Onlydis, and GenDis w/o semi, corresponding to using only the generative objective $\mathcal{L}_{\text{gen}}$, only the discriminative objective $\mathcal{L}_{\text{dis}}$, or both jointly without any pseudo-label learning. **Single-view semi-supervised variants.** To isolate the contribution of pseudo-labels, we examine gen+semi and dis+semi, where pseudo-labels are introduced only in the generative or discriminative view. **Component-level ablations of our framework.** We further evaluate GenDis by ablating four key modules: w/o GPL disables the generative pseudo-labeling; w/o $\mathcal{L}_{\text{cca}}$ removes the CCA regularization, w/o curr removes the curriculum strategy; w/o was disables the Wasserstein-based filtering.

The complete model consistently outperforms all ablated variants across KCRs, demonstrating that each component contributes meaningfully to overall performance. Furthermore, we draw the following conclusions: **(1) Pseudo-label learning is beneficial but insufficient when isolated.** gen+semi and dis+semi show improvements over non-semi variants but still underperform the full model, highlighting that single-view pseudo-labeling introduces helpful signals but lacks robustness. **(2) Curriculum learning and dispersion-based filtering mitigate noise.** Removing curriculum learning (w/o curr) or Wasserstein filtering (w/o was) results in clear performance degradation, especially under low KCRs, suggesting that these modules effectively reduce noise and promote label quality during training. **(3) CCA loss enhances alignment between views.** The ablation of the CCA regularization reduces performance, showing that $\mathcal{L}_{\text{cca}}$ helps align representations across the generative and discriminative views for better generalization.

5.4 Parameter Analysis

To better understand the influence and interaction of the generative and consistency constraints, we perform a grid search over λ_{gen} and λ_{cca} on the HWU dataset under three different KCRs. The results are visualized in Figure 3, using H-Score as the metric, which provides a balanced measure of performance across known and novel classes and is well-suited for hyperparameter selection.

Based on the results, we select the best-performing hyperparameters (λ_{gen}, λ_{cca}) for each KCR: (5.0, 1.0) for KCR = 0.25, (2.0, 0.1) for KCR = 0.5, and (5.0, 0.01) for KCR = 0.75. Jointly increasing both λ_{gen} and λ_{cca} generally leads to improved performance (i.e., bottom-right regions), demonstrating the benefit of combining semantic supervision with consistency regularization. Moreover, as KCR decreases, the known label space becomes increasingly sparse, making the model more susceptible to overfitting when relying solely on generative or discriminative signals. Consequently, stronger consistency regularization (i.e., CCA loss) becomes particularly important in low-supervision

scenarios. We also provide the parameter analysis of CCA projection dimension d and the number of generated responses k in Appendix E.4.

6 Conclusion

In this work, we proposed GenDis, a Generative-Discriminative Dual-View Co-Training framework for GCD. Our approach jointly optimizes a LLM for both discriminative classification and semantic label generation, enabling the learning of semantically enriched pseudo-labels alongside conventional one-hot supervision. To ensure effective coordination between two views, we employed CCA-based alignment and introduced a curriculum-guided discriminative pseudo-labeling strategy as well as a dispersion-aware generative pseudo-labeling mechanism. Extensive experiments on five benchmarks demonstrated that GenDis consistently outperforms existing SOTA methods.

Limitations

Our study focuses on the textual domain of GCD and does not extend experiments to multimodal scenarios such as vision–language or audio–text tasks. This choice aligns with the current research scope of GCD, where most prior works are designed and evaluated within purely textual settings. Moreover, our framework already integrates a wide range of established GCD benchmarks and evaluation metrics in Natural Language Processing (NLP), covering the major datasets and paradigms in this field. Exploring the applicability of our dual-view co-training paradigm to multimodal GCD remains an interesting direction for future research.

Ethical Considerations

We will abide by the laws, rules, and regulations of our community, school, work, and country. We will conduct ourselves with integrity, fidelity, and honesty. We will openly take responsibility for our actions and only make agreements that we intend to keep. All data used in this study are publicly available datasets obtained from GitHub and intended for research purposes. No human participants were directly involved, and no personally identifiable information (PII) was collected.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) (Grant No. 62506352), in part by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB1350102), in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by the guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057), in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities (2024ZDZX1007), in part by the Natural Science Foundation of Anhui Province (Grant No. 2508085QF211), the National Natural Science Foundation of China (Grant No. 62506348), New Generation Artificial Intelligence-National Science and Technology Major Project (Grant No. 2025ZD0122601), the Opening Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK (COGOS-2025HE02).

References

Wenbin An, Haonan Lin, Jiahao Nie, Feng Tian, Wenkai Shi, Yaqiang Wu, Qianying Wang, and Ping Chen. 2025. Unleashing the potential of model bias for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

Wenbin An, Feng Tian, Ping Chen, Qinghua Zheng, and Wei Ding. 2023a. New user intent discovery with robust pseudo label training and source domain joint training. *IEEE Intelligent Systems*.

Wenbin An, Feng Tian, Wenkai Shi, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. 2024b. Transfer and alignment network for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, QianYing Wang, and Ping Chen. 2023b. Generalized category discovery with decoupled prototypical network. In *Proceedings of the AAAI conference on artificial intelligence*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are

few-shot learners. *Advances in neural information processing systems*.

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*.

Mingcai Chen, Yuntao Du, Yi Zhang, Shuwei Qian, and Chongjun Wang. 2022. Semi-supervised learning with multi-head co-training. In *Proceedings of the AAAI conference on artificial intelligence*.

Xi Chen, Chuan Qin, Chuyu Fang, Chao Wang, Chen Zhu, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024. Job-sdf: A multi-granularity dataset for job skill demand forecasting and benchmarking. *Advances in Neural Information Processing Systems*, 37:129329–129356.

Xi Chen, Chuan Qin, Ziqi Wang, Shasha Hu, Chao Wang, Hengshu Zhu, and Hui Xiong. Beyond the known: An unknown-aware large language model for open-set text classification. In *The Fourteenth International Conference on Learning Representations*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

W Dong-DongChen and ZH WeiGao. 2018. Tri-net for semi-supervised deep learning. In *Proceedings of twenty-seventh international joint conference on artificial intelligence*.

Shansan Gong, Zelin Zhou, Shuo Wang, Fengjiao Chen, Xiujie Song, Xuezhi Cao, Yunsen Xian, and Kenny Zhu. 2023. Transferable and efficient: Unifying dynamic multi-domain product categorization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xiaohan Huang, Meng Xiao, Chuan Qin, Qingqing Long, Jinmiao Chen, Yuanchun Zhou, and Hengshu Zhu. 2026. Scihorizon-gene: Benchmarking llm for life sciences inference from gene knowledge to functional understanding. *arXiv preprint arXiv:2601.12805*.

Feihu Jiang, Chuan Qin, Kaichun Yao, Chuyu Fang, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024. Enhancing question answering for enterprise knowledge bases using large language models. In *International Conference on Database Systems for Advanced Applications*, pages 273–290. Springer.

Harold W Kuhn. 2004. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*.

Vaibhav Kumar, Hana Koorehdavoudi, Masud Moshtaghi, Amita Misra, Ankit Chadha, and Emilio Ferrara. 2023. Controlled text generation with hidden representation transformations. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Hunter Lang, Monica N Agrawal, Yoon Kim, and David Sontag. 2022. Co-training improves prompt-based learning for large language models. In *International Conference on Machine Learning*.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Sharon Levy and William Yang Wang. 2020. Cross-lingual transfer learning for COVID-19 outbreak alignment. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. 2024. Actively learn from LLMs with uncertainty propagation for generalized category discovery. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bo Liu, Li-Ming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. 2024b. How good are LLMs at out-of-distribution detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Bo Liu, Li-Ming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. 2024c. How good are llms at out-of-distribution detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing naturalness and flexibility in spoken dialogue interaction: 10th international workshop on spoken dialogue systems*.

Takuo Matsubara. 2024. Wasserstein gradient boosting: A framework for distribution-valued supervised learning. *Advances in Neural Information Processing Systems*.

Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. 2018. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*.

Chuan Qin, Xin Chen, Chengrui Wang, Pengmin Wu, Xi Chen, Yihang Cheng, Jingyi Zhao, Meng Xiao, Xiangchao Dong, Qingqing Long, and 1 others. 2025a. Scihorizon: Benchmarking ai-for-science readiness from scientific data to large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5754–5765.

Chuan Qin, Chuyu Fang, Kaichun Yao, Xi Chen, Fuzhen Zhuang, and Hengshu Zhu. 2025b. Cotr: Efficient job task recognition for occupational information systems with class-incremental learning. *ACM Transactions on Management Information Systems*, 16(2):1–30.

Chuan Qin, Le Zhang, Yihang Cheng, Rui Zha, Dazhong Shen, Qi Zhang, Xi Chen, Ying Sun, Chen Zhu, Hengshu Zhu, and 1 others. 2025c. A comprehensive survey of artificial intelligence techniques for talent analytics. *Proceedings of the IEEE*.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2023. T5Score: Discriminative fine-tuning of generative evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Qwen Team. 2024. Qwen2.5-7b-instruct. https://qwenlm.github.io/blog/qwen2.5/ and Hugging Face repository "Qwen/Qwen2.5-7B-Instruct".

Wenkai Shi, Wenbin An, Feng Tian, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. 2024. A unified knowledge transfer network for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Daniel Simig, Fabio Petroni, Pouya Yanki, Kashyap Popat, Christina Du, Sebastian Riedel, and Majid Yazdani. 2022. Open vocabulary extreme classification using generative models. In *Findings of the Association for Computational Linguistics*.

Zhiheng Song, Jingshuai Zhang, Chuan Qin, Chao Wang, Chao Chen, Longfei Xu, Kaikui Liu, Xiangxiang Chu, and Hengshu Zhu. 2026. Mobilitybench: A benchmark for evaluating route-planning agents in real-world mobility scenarios. *arXiv preprint arXiv:2602.22638*.

Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. 2024. Learning geometry-aware representations for new intent discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Lulu Tang, Zhi-Xin Yang, and Kui Jia. 2018. Canonical correlation analysis regularization: An effective deep multiview learning baseline for rgb-d object recognition. *IEEE Transactions on Cognitive and Developmental Systems*.

Zhenyu Tong, Chuan Qin, Chuyu Fang, Kaichun Yao, Xi Chen, Jingshuai Zhang, Chen Zhu, and Hengshu Zhu. 2025. From missteps to mastery: Enhancing low-resource dense retrieval through adaptive query generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1373–1384.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2022. Generalized category discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Hongjun Wang, Sagar Vaze, and Kai Han. 2024. SPTNet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *The Twelfth International Conference on Learning Representations*.

Ruixuan Xiao, Lei Feng, Kai Tang, Junbo Zhao, Yixuan Li, Gang Chen, and Haobo Wang. 2024. Targeted representation alignment for open-world semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

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*.

Xiaohang Zhan, Ziwei Liu, Junjie Yan, Dahua Lin, and Chen Change Loy. 2018. Consensus-driven propagation in massive unlabeled data for face recognition. In *Proceedings of the European conference on computer vision (ECCV)*.

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

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2023. A clustering framework for unsupervised and semi-supervised new intent discovery. *IEEE Transactions on Knowledge and Data Engineering*.

Shun Zhang, Chaoran Yan, Jian Yang, Wei Zhang, Changyu Ren, Tongliang Li, Jiaqi Bai, and Zhoujun Li. 2024. Tinid: A transfer and interpretable llm-enhanced framework for new intent discovery. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2022. New intent discovery with pre-training and contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. 2024. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. In *European Conference on Computer Vision*.

Henry Peng Zou, Siffi Singh, Yi Nian, Jianfeng He, Jason Cai, Saab Mansour, and Hang Su. 2025. Glean: Generalized category discovery with diverse and quality-enhanced llm feedback. *arXiv preprint arXiv:2502.18414*.

A The Use of Large Language Models Statement

The authors use LLM (ChatGPT) as an assistive tool in the preparation of this manuscript. We use LLMs to proofread, check grammar, and refine the language in the manuscript for improved clarity and readability.

B Additional Method Details

B.1 Prompt Design

We design a general prompt template for the GCD task, consisting of a system prompt, a user prompt, and an assistant prompt. These components together form a conversation set $\mathcal{Z} = \{z = Prompt(x, y) | (x, y) \in \mathcal{D}^l \cup \mathcal{D}^u\}$, where y is a placeholder for unlabeled samples in $\mathcal{D}^u$. This template serves as input for LLM fine-tuning.

> **System**: You are an expert in open-domain text classification. Your task is to assign the most suitable and concise category label for any given input text, even if no predefined categories are provided.
> **User**: Please read the following text and determine the most appropriate category it belongs to. *Text to classify:* {**x**}.
> **Assistant**: A suitable category could be: {**y**}

Unlike prior approaches that design task-specific prompts for particular scenarios (e.g., intent detection), our prompt template is unified and adaptated to diverse text classification tasks.

B.2 Hungarian Algorithm

GenDis expand training to the unlabeled set $\mathcal{D}^u$ by generating discriminative pseudo-labels $\hat{y} \in \mathcal{Y}^k \cup \mathcal{Y}^u$ (cluster index) through K-means clustering in the learned latent space $\mathbf{h}^{\text{dis}}$. However, the indices after K-means are permuted randomly in each training epoch, so the classifier parameters have to be reinitialized before each training epoch (Zhang et al., 2021). Thus, we adopt the Hungarian algorithm (Kuhn, 2004) as the alignment strategy to tackle the assignment inconsistency problem. Specifically, Let $\hat{y}^{(t-1)}$ denote the pseudo-labels from the previous generation and $\hat{y}^{(t)}$ the current generation's cluster assignments. We first construct a contingency matrix $C \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$, where $C_{i,j}$ counts the number of samples assigned to cluster i in $\hat{y}^{(t)}$ and to cluster j in $\hat{y}^{(t-1)}$. This matrix captures the co-occurrence statistics between clusters of two successive generations. To find the optimal one-to-one mapping that maximizes overall alignment, we apply the Hungarian algorithm to the cost matrix $\tilde{C} = C_{\max} - C$, where $C_{\max}$ is the largest entry in C. This ensures that minimizing the cost corresponds to maximizing the agreement between matched clusters. Once the optimal assignment is obtained, we remap the cluster indices in $y^{(t)}$ according to the alignment. This strategy produces a label sequence $\tilde{y}^{(t)}$ that is aligned with prior training epochs, thus preserving temporal consistency and improving the stability of discriminative pseudo-supervision.

B.3 Sinkhorn Normalization

The Sinkhorn-Knopp (SK) algorithm (Cuturi, 2013) is an iterative matrix scaling method that transforms a non-negative matrix into a doubly stochastic matrix—i.e., a matrix where each row and each column sums to one. It is widely used in optimal transport and entropy-regularized matching problems due to its efficiency and stability.

Given an input cost or similarity matrix $D \in \mathbb{R}^{n \times k}$ (e.g., distance between n samples and k cluster centroids), the SK algorithm applies exponentiation with a temperature-controlled scaling factor, followed by alternating row and column normalizations:

$$P = \text{Sinkhorn}(D) = \text{Normalize}_{\text{col}}\left(\text{Normalize}_{\text{row}}\left(\exp\left(-\frac{D}{\tau}\right)\right)\right), \tag{7}$$

where τ is a temperature parameter controlling the sharpness of the resulting distribution, and $\text{Normalize}_{\text{row}}(\cdot)$ and $\text{Normalize}_{\text{col}}(\cdot)$ denote row-wise and column-wise normalization steps, respectively.

The iterations proceed until the matrix P approximates a balanced assignment, where each class

Table S1: Dataset statistics under different known class ratios. The leftmost block (ALL) reports the full dataset statistics, including the entire training, validation, and test sets with all class labels. For each ratio setting (0.25/0.50/0.75), we report the number of labeled training and validation instances (|Train| and |Eval|), and the number of known classes (|Label|). All training and validation sets are constructed by sampling 10% labeled data from the selected known classes.

KCR	ALL				0.25			0.50			0.75		
Split	|Train|	|Eval|	|Test|	|Label|	|Train|	|Eval|	|Label|	|Train|	|Eval|	|Label|	|Train|	|Eval|	|Label|
BANKING	9,000	3,073	999	77	193	23	19	430	50	38	682	79	58
CLINC	17,995	2,250	2,250	150	456	76	38	900	150	75	1,344	224	112
HWU	7,712	1,032	933	64	183	25	16	365	52	32	595	83	48
MCID	1,198	331	170	16	29	4	4	59	8	8	92	12	12
StackOverflow	11,996	5,991	1,998	20	300	50	5	600	100	10	900	150	15

(column) receives nearly uniform assignment mass, and each sample (row) is assigned a normalized soft probability over classes. In our framework, SK normalization is applied to the distance matrix obtained from K-means clustering over both the generative and discriminative representations. This ensures that the resulting soft class assignment distribution is not only aligned with the distance structure but also encourages balanced cluster usage, which is essential in scenarios with limited labeled data or open-category settings.

B.4 Canonical Correlation Analysis Regularization

We introduce a CCA regularization term (Tang et al., 2018) into the dual-view representation learning process, enforcing alignment between $\mathbf{h}^{\text{dis}}$ and $\mathbf{h}^{\text{gen}}$ while maintaining their functional independence. Specifically, we first project the two views into a shared k-dimensional latent space using separate MLPs:

$$\mathbf{e}^{\text{dis}} = \mathbf{MLP}^{\text{dis}}(\mathbf{h}^{\text{dis}}), \quad \mathbf{e}^{\text{gen}} = \mathbf{MLP}^{\text{gen}}(\mathbf{h}^{\text{gen}}). \tag{8}$$

Let $\mathbf{E}_1$ and $\mathbf{E}2$ denote the batch-wise matrices of projected embeddings from the discriminative and generative views, respectively. We compute their covariance matrices as:

$$\begin{aligned}\Sigma_{11} &= \frac{\mathbf{E}_1\mathbf{E}_1^\top}{b-1} + r_1\mathbf{I},\\ \Sigma_{22} &= \frac{\mathbf{E}_2\mathbf{E}_2^\top}{b-1} + r_2\mathbf{I},\\ \Sigma_{12} &= \frac{\mathbf{E}_1\mathbf{E}_2^\top}{b-1},\end{aligned} \tag{9}$$

where r_1 and r_2 are regularization coefficients to ensure numerical stability and b is the batch size. To compute the canonical correlations, we perform whitening transformations by computing the root inverse of Σ_{11} and Σ_{22} through eigen decomposition, and derive the cross-view transformation matrix $\mathbf{T}$ and quantify the sum of the top-k singular values of $\mathbf{T}$ as the correlation $\text{corr}(\mathbf{E}_1, \mathbf{E}_2)$:

$$\begin{aligned}\mathbf{T} &= \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2},\\ \text{corr}(\mathbf{E}_1, \mathbf{E}_2) &= \left(\sum_{i=1}^{d}\log\sigma_i\right)/d,\end{aligned} \tag{10}$$

where $\{\sigma_i\}$ are the eigenvalues of $\mathbf{T}^\top\mathbf{T}$ and d is a hypermeter for choosing top-d eigenvalues. We define the CCA regularization term as the negative total canonical correlation: $\mathcal{L}_{\text{cca}} = -\text{corr}(\mathbf{E}_1, \mathbf{E}_2)$. This regularization loss encourages the two views to encode maximally correlated features in the shared subspace. To further stabilize the correlation computation, we apply spectral clipping and small-value filtering to remove unstable components near zero.

C Additional Related Works

Co-training Co-training is a semi-supervised learning paradigm that leverages unlabeled data by training classifiers on two conditionally independent "views" of the data (Blum and Mitchell, 1998). Early approaches attempted to enforce view diversity through architectural separation. For example, Deep Co-Training (Qiao et al., 2018) and Tri-net (Dong-DongChen and WeiGao, 2018) utilized dual or triple network architectures to encourage divergence. Subsequent methods, such as Consensus-Driven Propagation (Zhan et al., 2018), improved pseudo-label robustness with committees of heterogeneous models, while Multi-Head Co-Training (Chen et al., 2022) attached multiple heads to a shared backbone with a disagreement loss for diverse prediction. More recent research has focused on exploiting inherent model or data

disparities to construct views, including pairing large and small language models (Lang et al., 2022) or utilizing vision and language as distinct modalities for co-teaching (Zheng et al., 2024). To the best of our knowledge, we are the first to extend the co-training paradigm to the GCD task by introducing generative and discriminative views within a single LLM.

D Additional Experiemental Settings

D.1 Detailed Dataset Description

To verify the effectiveness and universality of our proposed method, we conducted exhaustive experiments on five widely used text classification datasets. These datasets are summarized as:

- **BANKING** (Casanueva et al., 2020) is a kind of dataset about the banking business, with 77 categories. The dataset is derived from real-world banking queries and is known for its class imbalance and lexical similarity between different intents, which makes it challenging for classification models.

- **CLINC** (Larson et al., 2019) is a very popular dataset, which encompasses a broad range of intents, totaling 150 across 10 domains, such as travel, banking, utilities, and weather.

- **HWU** (Liu et al., 2021) is a dataset derived from real-world dialogue systems, containing 64 intent classes. The data is collected from real-world voice assistant interactions and spans multiple domains, including information retrieval, home automation, and entertainment, reflecting realistic user behavior.

- **MCID** (Levy and Wang, 2020) is a dataset created for COVID-19 chatbot intent classification. It includes 16 intent categories in total, originally collected in four languages; in our study, we use the English subset only.

- **StackOverflow** (Xu et al., 2015) is a dataset collected from the StackOverflow developer community. It contains short text questions or posts labeled with their corresponding programming language tags. This dataset is often used to evaluate models on short-text intent detection or tag classification, covering categories such as *python*, *java*, *c#*, etc.

All datasets are publicly available for academic research and widely used in the NLP community. We strictly adhere to their usage protocols and licenses.

D.2 More Details on Dataset Split

We present the dataset split statistics across different KCRs in Table S1. The training and validation sets contain only known samples, while the test set includes both known and novel categories.

In prior work, the known and novel classes are randomly re-sampled in each experimental run based on the specified KCR. While this strategy enables class diversity across runs, it introduces significant randomness into the evaluation process and undermines the fairness of comparisons. As a result, previously reported performance metrics often vary considerably—even for the same model—making it difficult to reproduce and compare results consistently. To address this issue, we fix a single class split for each KCR setting and use it consistently across all experiments. The exact class split configurations are made publicly available via an anonymous link for reproducibility. All baselines are evaluated on these fixed splits to ensure fair and stable comparisons.

D.3 Detailed Baseline Descriptions

To comprehensively evaluate our proposed framework, we compare it against two categories of representative baselines: traditional discriminative models without LLM involvement, and recent LLM-enhanced approaches.

Non-LLM Baselines. We include several strong baselines that do not incorporate large language models but have demonstrated competitive performance in GCD:

- **DeepAligned** (Zhang et al., 2021): A two-stage method that uses alignment strategies (e.g., the Hungarian algorithm) to generate pseudo labels for all unlabeled instances, enabling unified learning of known and novel categories.

- **DPN** (An et al., 2023b): A prototype-based framework that constructs dual prototypes for known and novel classes to improve feature alignment and clustering effectiveness.

- **PTJN** (An et al., 2023a): A method that leverages model ensembling and iterative refinement to improve the robustness of pseudo-label assignments.

- **GeoID** (Tang et al., 2024): A method that improves pseudo-label distribution by enforcing uniformity through the Sinkhorn-Knopp algorithm, enhancing robustness in open-set scenarios.

- **TAN** (An et al., 2024b): A transfer-based method that reweights known class representations to construct refined novel class prototypes via similarity matrices.

- **KTN** (Shi et al., 2024): An approach that enhances knowledge transfer by incorporating similarity-based weights into pseudo-label distribution learning.

- **SDC** (An et al., 2025): A calibration-based framework that integrates entropy-based uncertainty estimation to mitigate the impact of noisy pseudo labels.

LLM-Enhanced Baselines. We also compare with recent methods that incorporate large language models as auxiliary tools to assist the GCD process:

- **LOOP** (An et al., 2024a): Utilizes LLMs to identify semantically similar neighbors through few-shot prompting, and applies neighborhood contrastive learning to improve representation quality.

- **ALUP** (Liang et al., 2024): Introduces a comparison-based prompting strategy, where LLMs classify samples by comparing them with exemplars from known and novel categories, enhancing pseudo-supervision quality.

- **GLEAN** (Zou et al., 2025): Enhances the discovery of ambiguous samples by incorporating quality-controlled feedback from LLMs at both the instance and label levels, thereby improving robustness and interpretability.

All baseline methods are re-implemented or adapted based on their official code or descriptions to ensure consistency.

D.4 Detailed Evaluation Metrics

Following prior works (Zhang et al., 2021, 2023), we adopt six evaluation metrics to comprehensively assess clustering performance in the GCD setting:

- **Known Accuracy (K-ACC)**: The clustering accuracy computed on instances from known categories. Given a clustering assignment, we use the Hungarian algorithm (Kuhn, 2004) to find the optimal one-to-one mapping between predicted clusters and ground-truth labels. Let $\mathcal{Y}^k$ be the known class set and $\mathcal{C}^k$ be the predicted clusters assigned to them, K-ACC is computed as the proportion of correctly assigned samples in known classes: $\text{K-ACC} = \frac{1}{|\mathcal{Y}^k|}\sum_{y\in\mathcal{Y}^k}\frac{|\hat{C}_y\cap C_y|}{|C_y|}$, where C_y is the set of ground-truth samples with label y, and $\hat{C}_y$ is the assigned cluster.

- **Novel Accuracy (N-ACC)**: Similar to K-ACC, but calculated exclusively on novel categories, measuring the model's ability to discover new classes without supervision. It reflects clustering performance on the unknown portion of the data.

- **H-Score**: To avoid biased evaluation toward either known or novel classes, we report the harmonic mean of K-ACC and N-ACC, defined as: $\text{H-Score} = \frac{2\cdot\text{K-ACC}\cdot\text{N-ACC}}{\text{K-ACC}+\text{N-ACC}}$, This score penalizes models that perform well only on one subset (e.g., known classes) and encourages balanced performance across both.

- **Overall Accuracy (ACC)**: The standard clustering accuracy over all samples, again based on Hungarian matching between predicted clusters and all ground-truth labels. While intuitive, it tends to overemphasize known categories due to class imbalance, making it less reliable in open-world settings.

- **Normalized Mutual Information (NMI)**: NMI measures the mutual dependence between the predicted cluster assignment U and the ground-truth labels V. It is defined as: $\text{NMI}(U,V) = \frac{2\cdot I(U;V)}{H(U)+H(V)}$, where $I(U;V)$ is the mutual information, and $H(U), H(V)$ are the entropies of the cluster and label distributions, respectively. NMI ranges from 0 to 1, with 1 indicating perfect agreement.

- **Adjusted Rand Index (ARI)**: ARI evaluates the similarity between two data partitions (clusters and labels) by counting pairwise agreements. It adjusts for chance and is defined as: $\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right] - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}$, where n_{ij} is the number of data points in both predicted cluster i and ground-truth class j; a_i and b_j are the row and column sums of the contingency table. ARI ranges from -1 to 1, with 1 indicating perfect clustering and 0 corresponding to random assignments.

Table S2: Performance of various methods across five datasets at different ratios under metric ACC and H-Score.

Dataset		BANKING		CLINC		HWU		MCID		StackOverflow	
KCR	Metric	ACC	H-Score	ACC	H-Score	ACC	H-Score	ACC	H-Score	ACC	H-Score
0.25	DeepAligned	46.44	49.21	75.67	78.55	53.75	56.44	36.26	36.68	58.58	62.73
	DPN	45.72	48.11	74.17	77.22	50.15	53.54	42.45	44.56	60.89	64.61
	PTJN	51.95	54.38	79.22	81.61	56.98	59.54	40.34	39.94	71.46	73.88
	GeoID	35.48	36.52	84.63	86.24	67.49	67.5	45.02	37.06	72.54	73.57
	TAN	42.04	45.05	58.27	62.16	43.82	47.24	39.15	39.08	45.82	47.35
	KTN	41.81	44.65	70.5	75.08	58.91	62.2	33.05	27.05	78.0	80.81
	SDC	55.52	59.54	73.49	77.3	65.62	68.41	33.35	30.19	79.16	80.88
	LOOP	68.79	70.86	87.26	89.38	71.2	72.59	52.93	45.53	78.72	80.64
	ALUP	72.29	71.71	88.91	89.94	74.15	75.08	65.98	57.18	81.33	83.52
	Glean	66.36	67.81	85.79	86.8	69.93	70.11	63.45	54.87	78.31	79.76
	GenDis	**74.69**	**77.28**	**90.52**	**92.34**	**82.36**	**83.98**	**72.58**	**71.52**	**90.02**	**91.78**
0.50	DeepAligned	59.25	58.68	81.16	80.8	62.74	62.45	50.5	48.24	74.21	73.97
	DPN	56.82	51.48	80.04	78.57	61.84	61.58	46.9	42.22	79.9	79.28
	PTJN	64.15	62.78	84.11	83.77	67.27	67.2	52.57	52.4	72.81	72.3
	GeoID	68.26	67.26	87.66	87.35	69.55	69.58	47.81	47.59	76.76	76.2
	TAN	56.5	53.78	73.14	70.97	53.82	53.02	45.56	44.99	60.34	51.71
	KTN	62.67	57.54	80.9	78.94	70.62	70.53	33.35	33.04	79.13	78.83
	SDC	64.9	63.11	81.36	80.79	72.98	72.93	67.13	66.8	78.96	78.72
	LOOP	72.65	71.83	87.98	87.67	73.9	73.9	51.36	50.02	82.85	82.74
	ALUP	74.99	74.06	89.23	89.08	76.28	76.28	74.44	73.92	86.4	86.34
	Glean	72.69	71.53	83.66	83.41	77.43	77.15	68.78	66.06	77.6	77.32
	GenDis	**76.22**	**75.29**	**94.06**	**93.92**	**84.88**	**84.79**	**80.44**	**80.16**	**93.33**	**93.32**
0.75	DeepAligned	65.77	63.31	86.84	83.61	70.95	68.22	63.95	46.67	75.7	72.04
	DPN	70.74	57.38	88.44	82.69	76.91	72.7	61.1	48.23	82.35	81.74
	PTJN	72.95	67.31	87.33	83.04	72.84	70.82	61.78	52.52	74.17	72.46
	GeoID	75.7	72.08	92.08	88.82	76.55	77.57	48.72	45.23	79.68	79.02
	TAN	69.08	60.54	82.47	73.28	68.97	59.32	58.61	48.01	68.85	48.06
	KTN	76.15	68.28	90.39	84.48	78.99	76.97	33.53	35.18	82.19	83.73
	SDC	75.43	72.85	87.55	86.83	73.61	75.36	73.23	72.44	80.73	81.29
	LOOP	73.93	72.59	90.84	88.2	73.78	73.42	53.05	51.9	83.87	83.91
	ALUP	80.67	74.76	91.68	90.13	80.02	79.7	75.71	70.82	87.78	87.12
	Glean	80.69	78.38	86.84	84.06	80.26	80.0	78.05	77.75	76.98	79.25
	GenDis	**83.49**	**82.04**	**94.32**	**91.87**	**85.32**	**86.29**	**80.82**	**81.79**	**92.51**	**93.15**

Table S3: Performance of various backbones based on GenDis across 5 datasets at different ratios. Metrics include ARI, NMIC, and H-Score. The best results are highlighted in bold, while the second-best results are underscored.

Dataset		BANKING			CLINC			HWU			MCID			StackOverflow		
KCR	Backbone	ARI	NMI	H-Score	ARI	NMI	H-Score	ARI	NMI	H-Score	ARI	NMI	H-Score	ARI	NMI	H-Score
0.25	Llama3-8B-R1	57.46	82.13	66.94	84.33	95.86	89.28	68.72	87.49	80.02	66.81	79.73	82.50	78.82	84.57	88.08
	Llama3.1-8B	64.81	86.30	77.49	87.72	96.87	91.54	68.06	87.50	79.80	52.78	73.28	58.15	82.89	86.03	92.57
	Qwen2.5-7B	65.54	86.32	77.28	88.40	97.12	92.34	74.62	90.05	83.98	58.24	75.26	71.52	82.88	86.50	91.78
	Qwen3-8B	67.46	87.40	78.55	89.36	97.02	92.98	75.02	91.24	84.76	51.41	72.47	65.41	84.99	87.62	93.76
0.50	Llama3-8B-R1	65.95	86.10	72.20	89.80	97.26	92.18	64.56	85.71	71.89	59.50	76.88	68.41	83.03	87.11	91.80
	Llama3.1-8B	68.38	87.06	75.61	89.53	97.29	91.56	73.64	89.08	82.52	67.21	81.26	80.38	83.53	87.27	92.07
	Qwen2.5-7B	67.76	87.26	75.29	91.78	97.79	93.92	75.92	90.26	84.79	67.93	81.78	80.16	86.64	88.64	93.32
	Qwen3-8B	67.64	86.95	73.83	89.75	97.50	91.27	72.37	88.76	80.32	63.44	80.45	70.53	83.03	88.40	87.76
0.75	Llama3-8B-R1	75.82	89.90	77.87	92.73	97.89	92.74	73.50	89.16	79.49	68.93	82.47	76.61	84.26	86.52	92.04
	Llama3.1-8B	72.71	88.79	74.54	92.80	98.11	92.67	70.46	87.72	76.26	73.78	85.03	85.35	85.79	87.87	92.81
	Qwen2.5-7B	75.26	90.69	82.04	92.49	98.05	91.87	76.68	90.54	86.29	69.36	82.42	81.79	86.51	88.93	93.15
	Qwen3-8B	73.04	88.98	75.31	91.70	97.90	90.75	74.88	89.69	80.88	71.00	83.40	69.40	83.65	88.33	89.61

D.5 Implementation Details

To ensure fair comparisons, we employ bert-base-uncased* as the backbone for the discriminative

*https://huggingface.co/google-bert/bert-base-uncased

Table S4: Performance of various backbones based on GenDis across 5 datasets at different ratios. Metrics include ACC, K-ACC and N-ACC. The best results are highlighted in bold, while the second-best results are underscored.

Dataset		BANKING			CLINC			HWU			MCID			StackOverflow		
KCR	Backbone	ACC	K-ACC	K-ACC	ACC	K-ACC	K-ACC	ACC	K-ACC	K-ACC	ACC	K-ACC	K-ACC	ACC	K-ACC	K-ACC
0.25	Llama3-R1	65.70	69.70	64.39	87.16	94.39	84.70	77.03	87.30	73.86	82.48	82.56	82.45	85.81	93.53	83.24
	Llama3	74.03	87.22	69.71	89.78	95.61	87.80	75.97	90.16	71.57	65.86	48.84	71.84	91.12	95.80	89.56
	Qwen2.5-7B	74.69	82.49	72.80	90.52	96.62	88.45	82.36	87.59	80.74	72.58	70.56	72.97	90.02	95.83	88.08
	Qwen3-8B	76.93	82.21	75.19	92.31	94.39	91.61	83.20	88.07	81.70	66.16	63.95	66.94	92.41	96.73	90.96
0.50	Llama3-R1	73.45	83.74	63.46	92.27	95.11	89.42	71.90	71.69	72.09	69.79	60.49	78.70	91.80	92.05	91.56
	Llama3	76.28	84.14	68.65	91.82	96.71	86.93	82.56	81.87	83.18	80.66	76.54	84.62	92.07	91.92	92.22
	Qwen2.5-7B	76.22	85.98	67.00	94.06	97.56	90.56	84.88	84.48	85.16	80.44	79.17	81.66	93.33	93.92	92.75
	Qwen3-8B	74.88	84.47	65.58	91.64	97.51	85.78	80.43	78.82	81.89	72.21	61.73	82.25	87.82	85.54	90.09
0.75	Llama3-R1	83.44	87.86	69.92	95.20	97.50	88.42	80.81	81.85	77.25	78.85	80.80	72.84	92.10	92.17	91.91
	Llama3	79.66	83.76	67.15	94.98	97.14	88.60	79.26	81.48	71.67	86.10	86.80	83.95	92.89	92.97	92.65
	Qwen2.5-7B	83.49	85.41	79.14	94.32	96.60	87.63	85.32	84.45	88.30	80.82	79.70	84.26	92.51	91.81	94.62
	Qwen3-8B	78.62	81.43	70.05	93.64	96.31	85.79	82.85	84.36	77.68	81.27	89.20	56.79	88.57	87.47	91.84

language model, and Qwen2.5-7B-Instruct[†] as the backbone for the generative model. For the BERT-based model, we apply full-parameter fine-tuning. For the large language model (LLM) (Touvron et al., 2023), we adopt a parameter-efficient fine-tuning (PEFT) strategy using LoRA (Hu et al., 2022), which inserts trainable low-rank adapters into each Transformer layer while freezing the original weights. We optimize all models using the AdamW optimizer with a learning rate of 5×10^{-5}, a batch size of 32, and 15 training epochs. The hyperparameters for our proposed method, $(\lambda_{\text{gen}}, \lambda_{\text{cca}})$, are selected based on the Known Class Ratio (KCR): (5.0, 1.0) for KCR = 0.25, (2.0, 0.1) for KCR = 0.5, and (5.0, 0.01) for KCR = 0.75. We set the CCA subspace dimension to $d = 32$, and generate $k = 4$ candidate labels. For the Sinkhorn-Knopp (SK) algorithm, we follow the settings used in prior baselines. Specifically, we perform 3 iterations of alternating row and column normalization to obtain a doubly stochastic transport matrix. The entropy regularization coefficient is set to 0.1, which controls the smoothness of the transport plan—balancing between numerical stability and solution sparsity. Additionally, we set the imbalance factor to 1.0, indicating no explicit adjustment for class imbalance. Each experiment are conducted on a single NVIDIA RTX 5880 GPU with 48GB memory.

E Additional Experiemental Results

E.1 Detailed Experimental Results

Table S2 presents the results under overall ACC and H-Score, reflecting the holistic performance and balance across categories. These results confirm that our proposed GenDis achieves consistent improvements across all metrics and settings, demonstrating strong generalization to unseen classes while maintaining robust performance on known categories.

E.2 Detailed Backbone Results

To further understand the impact of different LLM backbones on the overall performance of GenDis, we report a comprehensive evaluation using four representative backbones: LLaMA3-8B-R1[‡], LLaMA3.1-8B[§], Qwen2.5-7B[¶], and Qwen3-8B[‖]. These models vary in parameter scale and training quality, representing a diverse range of foundation model capabilities. Table S3 and Table S4 present detailed results under multiple metrics, including ARI, NMI, H-Score, ACC, K-ACC, and N-ACC, across five benchmark datasets and three KCR settings (0.25, 0.50, 0.75). Overall, these results validate the robustness of GenDis across various LLM backbones and indicate that our framework can adapt to different foundation model capabilities.

E.3 Detailed Ablation Studies

Table S5 presents a comprehensive ablation study on the HWU dataset across five evaluation metrics under different KCRs. We analyze the contribution of each major component in GenDis by incrementally removing or replacing them with simplified variants.

[†] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[‡] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

[§] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[¶] https://huggingface.co/Qwen/Qwen2.5-7B

[‖] https://huggingface.co/Qwen/Qwen3-8B

Table S5: Ablation study on HWU datasets across all five metrics.

KCR	0.25						0.50						0.75					
Metric	ARI	NMI	H-Score	ACC	K-ACC	N-ACC	ARI	NMI	H-Score	ACC	K-ACC	N-ACC	ARI	NMI	H-Score	ACC	K-ACC	N-ACC
Onlygen	64.35	85.80	74.25	73.02	76.84	71.82	70.71	88.47	77.21	77.81	72.43	82.69	70.67	88.32	79.76	78.19	76.78	83.05
Onlydis	71.98	88.96	78.83	79.36	77.87	79.82	68.22	87.66	76.48	76.55	75.36	77.63	73.52	89.49	82.06	80.43	78.97	85.41
GenDis w/o semi	72.02	88.94	81.19	80.43	85.30	78.30	71.12	88.78	80.26	80.23	81.06	79.48	71.19	85.93	77.40	79.43	80.98	74.11
Gen+semi	68.72	87.48	75.15	76.40	72.95	77.47	68.69	88.23	76.88	76.94	76.17	77.63	69.64	88.54	78.47	76.26	74.22	83.26
Dis+semi	70.36	88.08	81.63	79.17	87.30	76.65	71.35	88.79	78.66	78.78	76.99	80.41	71.29	88.50	79.96	80.04	80.10	79.83
GenDis w/o GPL	71.60	88.74	81.90	80.09	85.86	78.30	73.39	89.45	79.41	79.36	81.06	77.82	75.41	90.15	83.09	83.24	83.35	82.83
GenDis w/o $\mathcal{L}_{cca}$	70.04	88.61	81.13	79.75	84.02	78.43	73.51	86.32	82.84	83.85	82.54	84.14	74.39	89.73	80.76	82.36	83.60	78.11
GenDis w/o curr	64.31	84.34	73.56	72.00	76.92	70.48	66.98	87.02	74.75	74.71	75.97	73.57	71.41	87.98	75.45	79.55	82.48	69.53
GenDis w/o was	68.37	86.60	78.40	76.42	84.36	73.81	73.56	87.77	81.37	81.40	80.91	81.84	72.50	89.11	79.66	81.10	82.23	77.25
GenDis	**74.62**	**90.05**	**83.98**	**82.36**	**87.59**	**80.74**	**75.92**	**90.26**	**84.79**	**84.88**	**84.48**	**85.16**	**76.68**	**90.54**	**86.29**	**85.32**	**84.45**	**88.30**

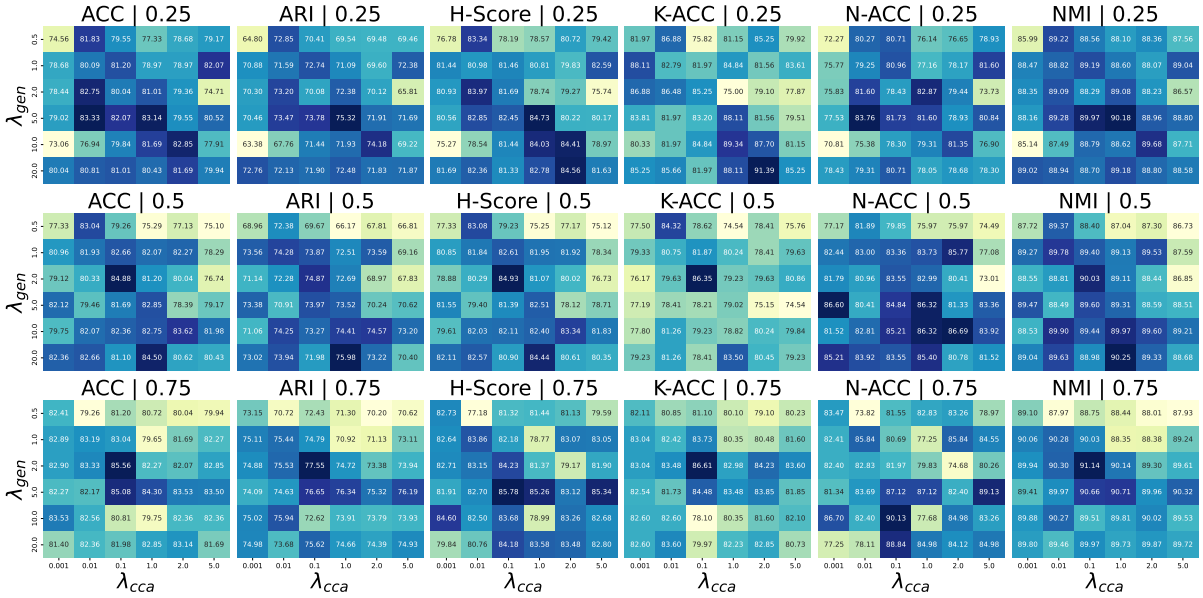

Figure S1: Parameter on λ_{gen} and λ_{cca}

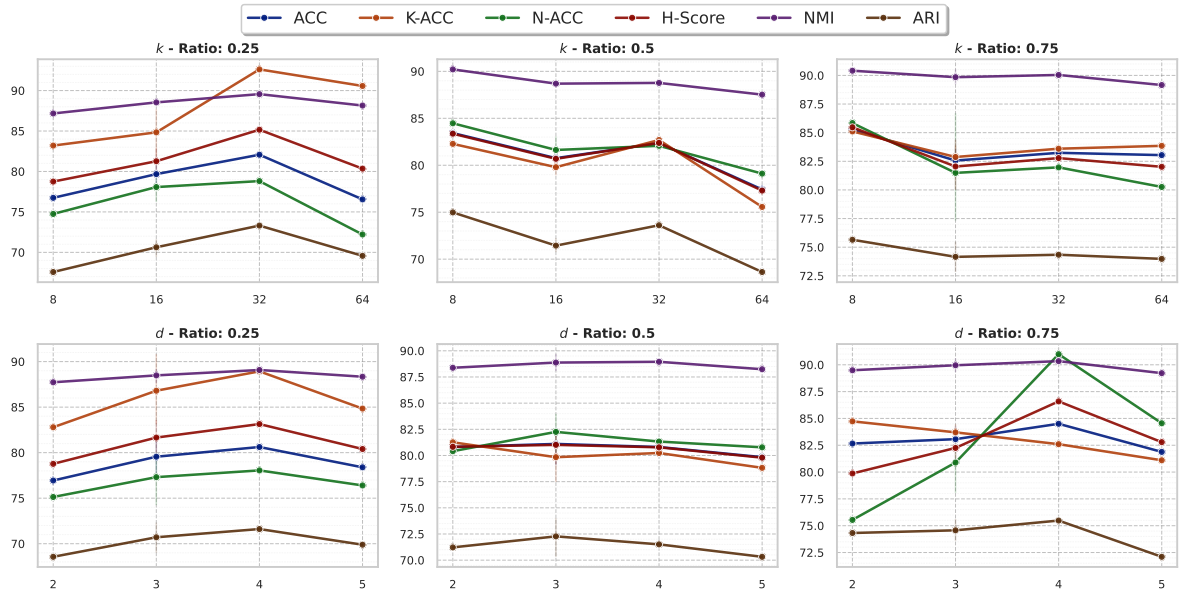

Figure S2: Parameter on CCA projection dimension (d) and the number of generated responses (k)

E.4 Detailed Parameter Analysis

We conduct two groups of parameter sensitivity analyses to better understand the impact of key

hyperparameters in our framework.

The first analysis (Figure S1) investigates the joint effect of the loss weights λ_{gen} and λ_{cca} on model performance. We observe a clear trade-off between **K-ACC** (performance on known classes) and **N-ACC** (performance on novel classes): increasing one loss weight often improves one metric at the expense of the other. Notably, the **H-Score**, which harmonizes these two aspects, offers a more balanced and informative performance measure. Accordingly, we select the optimal hyperparameter combination based on the highest H-Score.

The second analysis (Figure S2) examines the effect of the CCA projection dimension d and the number of generated responses k used in multi-inference during label generation. The results indicate that setting the CCA subspace dimension to $d = 32$ and generating $k = 4$ candidate responses achieves the best trade-off across all metrics, offering a balance between representational expressiveness and generation diversity.

E.5 Details of Case Study

T-SNE visualization We visualize the learned representations on the StackOverflow test set using T-SNE to compare GenDis with a representative baseline (ALUP) under different KCRs. As shown in Figure S3, ALUP produces entangled and overlapping clusters, especially for unseen classes under low KCR settings. In contrast, GenDis consistently forms more compact and well-separated clusters, even when labeled data is scarce, showing that GenDis learns more robust and discriminative representations, particularly for novel categories.

To further verify that the latent space guided by discriminative pseudo-labeling can effectively support semantic label generation, we extract the pseudo-labels generated by GenDis on the StackOverflow dataset, which is an intent detection benchmark for programming-related questions. We observe that GenDis consistently produces concise label names that align closely with ground-truth categories (e.g., *apache*, *drupal*, *wordpress*). In contrast, baseline methods such as LOOP and Glean tend to generate verbose or noisy phrases that are often irrelevant to the original label semantics. These results demonstrate the superiority of our approach in capturing precise, label-aligned semantics, which is critical for guiding pseudo-label learning and improving generalization. More details and generated labels are provided in Appendix E.5.

Generated Labels by GenDis To better understand the quality of semantic signals provided by our method, we conduct a comparative case study on the interpreted label descriptions at different KCRs (KCR = 0.25, 0.50, 0.75). Specifically, we generate the textual label for each instance in the unlabeled dataset using the sample prompt described in Section 4.2, and then apply the Hungarian algorithm to align the generated labels with the ground-truth labels for evaluation. As shown in Tables S6–S8, we compare our model (GenDis) with two representative LLM-enhanced baselines (LOOP and Glean). It is worth noting that another recent method, ALUP (Liang et al., 2024), does not generate semantic label names. Instead, it relies on querying representative samples and uses retrieved neighbors as a form of weak supervision signal.

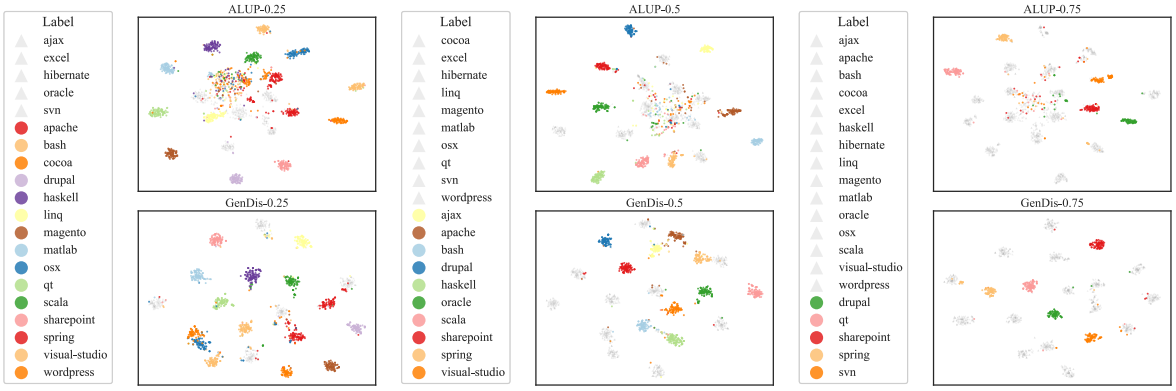

Figure S3: T-SNE visualization under different KCRs. As known categories typically exhibit better clustering quality, we render them in gray and use distinct colors to highlight novel classes, thereby emphasizing the clustering behavior of novel categories.

Table S6: Case study of label interpretation results at KCR = 0.25.

Label	GenDis	LOOP	Glean
ajax	–	–	–
apache	apache	Web server configuration	Apache and HTAccess Rewrites
bash	bash	Bash scripting techniques	Excel Queries
cocoa	cocoa	Qt widget customization	Querying and Collections
drupal	drupal	Content Management System Inquiry	Subversion Management
excel	–	–	–
haskell	haskell	understanding system workings	Programming Languages and Tools
hibernate	–	–	–
linq	linq	LINQ Operations	Programming Queries
magento	magento	technical issue	Development Tasks
matlab	matlab	MATLAB Help Request	Querying and Programming
oracle	–	–	–
osx	mac	Drupal question	SVN Configuration and Usage
qt	qt	how to configure or use external tools/functions	AJAX Programming
scala	scala	Haskell concepts inquiry	Excel Programming and Automation
sharepoint	sharepoint	Drupal issues and feature requests	Oracle SQL and PL/SQL
spring	spring	Web Technology Concepts	Hibernate Troubleshooting
svn	–	–	–
visual-studio	vs2008	Seeking guidance on licensing and usage	Code Interpreter
wordpress	wordpress	seeking guidance on implementation/configuration	CMS and Plugin Management

Table S7: Case study of label interpretation results at KCR = 0.50.

Label	GenDis	LOOP	Glean
ajax	ajax	Programming Inquiry	AJAX and ASP.NET
apache	apache	Scripting Commands	AJAX and Web Server Issues
bash	–	Bash Scripting	Excel Formulas and Manipulations
cocoa	–	–	–
drupal	drupal	SharePoint Integration Issues	Mac OS X Configuration and Automation
excel	–	–	–
haskell	haskell	Programming Issues	Haskell Programming
hibernate	–	–	–
linq	–	–	–
magento	–	–	–
matlab	–	–	–
oracle	oracle	Seeking Clarification	–
osx	python	–	–
qt	–	–	–
scala	scala	Bash Commands	Linq to Sql Queries and Operations
sharepoint	sharepoint	Request for modification or retrieval of specific content	Web Development and Server Management
spring	spring	Troubleshooting Guides	Hibernate Troubleshooting
svn	–	–	–
visual-studio	visual-studio	how to configure external tools or elements	Programming Queries
wordpress	–	–	–

Table S8: Case study of label interpretation results at KCR = 0.75.

Label	GenDis	LOOP	Glean
ajax	–	–	–
apache	–	–	–
bash	–	–	–
cocoa	–	–	–
drupal	drupal	Drupal troubleshooting	Programming Queries
excel	–	–	–
haskell	–	–	–
hibernate	–	–	–
linq	–	–	–
magento	–	–	–
matlab	–	–	–
oracle	–	–	–
osx	–	–	–
qt	qt	Drupal module functionality	Qt Programming Issues
scala	–	–	–
sharepoint	sharepoint	Custom Field Configuration	Development Tools and Integrations
spring	spring	Drupal form customization	Programming Issues
svn	subversion	how to upgrade/transition	–
visual-studio	–	–	–
wordpress	–	–	–