Coarse-to-Fine Open-Set Graph Node Classification with Large Language Models

Xueqi Ma[1], **Xingjun Ma**[2], **Sarah Monazam Erfani**[1], **Danilo Mandic**[3], **James Bailey**[1]

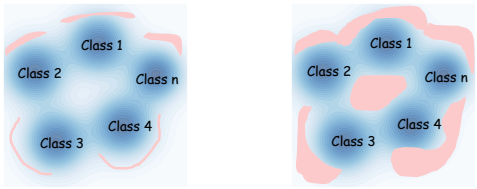
[1]The University of Melbourne, Australia
[2] Fudan University, China
[3] Imperial College London, UK

Abstract

Developing open-set classification methods capable of classifying in-distribution (ID) data while detecting out-of-distribution (OOD) samples is essential for deploying graph neural networks (GNNs) in open-world scenarios. Existing methods typically treat all OOD samples as a single class, despite real-world applications—especially high-stake settings like fraud detection and medical diagnosis—demanding deeper insights into OOD samples, including their probable labels. This raises a critical question: *Can OOD detection be extended to OOD classification without true label information?* To answer this question, we introduce a Coarse-to-Fine open-set Classification (CFC) method that leverages large language models (LLMs) for graph datasets. CFC consists of three key components: 1) A coarse classifier that utilizes LLM prompts for OOD detection and outlier label generation; 2) A GNN-based fine classifier trained with OOD samples from coarse classifier for enhanced OOD detection and ID classification; and 3) Refined OOD classification achieved through LLM prompts and post-processed OOD labels. Unlike methods relying on synthetic or auxiliary OOD samples, CFC employs semantic OOD data-instances that are genuinely out-of-distribution based on their inherent meaning, thus improving interpretability and practical utility. CFC enhances OOD detection by 10% compared to state-of-the-art approaches on graph domain, while achieving up to 70% accuracy in OOD classification on graph datasets.

Introduction

Graph neural networks (GNNs) have demonstrated excellent performance in closed-set scenarios, where the train and test datasets share the same distributions. However, in many real-world applications, models are deployed on data containing previously unseen classes. Traditional GNN methods (Kipf and Welling 2017; Hamilton, Ying, and Leskovec 2017; Ma et al. 2024b) typically classify all unlabeled nodes into known classes, failing to identify nodes that belong to unknown classes, which degrades overall model performance. Addressing this limitation requires the development of models that can accurately classify in-distribution (ID) samples from known classes while effectively rejecting out-of-distribution (OOD) samples from unknown classes. This critical challenge is commonly referred to as the open-set classification problem.



(a) Subspace without Semantic OOD Information

(b) Subspace with Semantic OOD Information

Figure 1: Comparison of subspaces of methods without semantic OOD information and our proposed CFC, which incorporates such information. Blue regions denote ID subspaces; other regions show OOD subspaces. CFC provides larger embedding space (pink), enabling direct OOD identification.

Recent approaches (Hendrycks and Gimpel 2016; Song and Wang 2022) to open-set node classification problem (Wu, Pan, and Zhu 2021) have employed thresholding methods, using the maximum output probability as a confidence score to distinguish OOD samples from ID ones. While intuitive, determining an optimal threshold (Yang, Lu, and Gan 2023) to separate unknown from known classes is both challenging and time-consuming (Perera et al. 2020). Another line of research redefines open-set classification as a closed-set problem by estimating the distribution of unknown classes and adjusting the network's confidence. For example, methods such as (Ge et al. 2017; Neal et al. 2018; Perera et al. 2020; Zhou, Ye, and Zhan 2021) generate synthetic samples as OOD, while others, like (Hendrycks, Mazeika, and Dietterich 2018; Wang et al. 2023), incorporate auxiliary training data—referred to as outlier exposure—to train image classifiers for OOD detection. Building on these works, Zhang et al. (2023) proposed generating proxy unknown nodes to simulate open-set data for graph node OOD detection.

While these approaches have mitigated the problem of OOD detection to some extent, they encounter several challenges, as follows: i) To enable OOD detection, they require the use of a large number of synthetic/auxiliary OOD samples during training, imposing significant computation cost. ii) Using generated samples or auxiliary training data may not accurately reflect real-world OOD variations. Accordingly,

these approaches may lack true semantic understanding and risk overfitting to specific datasets pairs. iii) Without semantic OOD samples that are realistic and meaningful, they fail to accurately represent the true OOD space, resulting in a small subspace and sharp boundaries for OOD detection, as illustrated in Fig. 1 (a). iv) More critically, these methods often group multiple unknown classes into a single OOD category, which significantly reduces their utility. In real-world scenarios, distinguishing between various unknown classes is crucial for informed decision-making, efficient data utilization, and performance in high-stakes applications such as medical diagnosis, autonomous driving, and fraud detection. For instance, in financial networks, grouping diverse fraudulent behaviors—such as phishing attacks, insider threats, and money-laundering schemes—into a single unknown category, oversimplifies their distinctions, hindering the nuanced understanding required for effective mitigation. This challenge raises a fundamental question: *Can we develop a comprehensive classification approach that seamlessly classifies both known and unknown classes, without requiring labeled samples for OODs?*

Accurate OOD classification poses significant challenges due to the uncertainty surrounding potential OOD labels. Specifically, the OOD domains—whether proximal or distant from the ID domains—are unknown, and even more so the number of unknown classes. To overcome this, we map the graph into text space and propose a Coarse-to-Fine Classification (CFC) approach to explore the OOD label space. In the first step, leveraging the expert knowledge and reasoning capabilities of LLMs, we design a coarse classifier by creating LLM prompts to detect OOD samples on the test set without prior OOD information and generate potential outlier class labels. The identified OOD data from the coarse classifier provides a potential OOD space with semantic information (i.e., possible meaningful and real-world outlier classes, such as topics of papers, news domains, etc.). Using these noisy coarse OOD data, we proceed to the second step: constructing a GNN-based fine classifier to further detect OOD samples and perform ID classification. More specifically, we use a label propagation method to remove falsely identified OOD samples, and an improved manifold mixup method (Verma et al. 2019) for OOD data augmentation, enabling us to effectively predict the distribution of novel classes. Unlike approaches that rely on additional synthetic or auxiliary data for training, CFC captures semantic OOD samples, reducing the distribution discrepancy (Wang et al. 2023) between the training data and real OOD data. As a result, CFC constructs a larger OOD subspace with a smoother boundary for OOD detection, as shown in Fig. 1 (b). The advantages of semantic OOD are further demonstrated by the improved OOD detection performance achieved with a small number of semantic OOD samples. Finally, we achieve OOD classification using LLM prompts designed with a post-processed OOD label space. It is important to note that we do not have any true OOD label information. Our contributions can be summarized as follows:

- Recognizing the critical need for accurately distinguishing between various unknown classes to ensure safety and enable effective decision-making in unpredictable environments, we introduce a novel challenge in the open-world setting: **OOD classification** on graphs. This task involves not only detecting OOD samples but also classifying them into their respective unknown classes, thereby extending the scope of traditional OOD detection.
- We propose a general coarse-to-fine classification method that integrates semantic OOD samples and a potential OOD label space, enabling the model to effectively perform both ID and OOD classification for the open-set graph node classification problem.
- Our CFC method demonstrates strong performance, achieving up to a 70% improvement in graph OOD classification and a 10% improvement in OOD detection compared to baseline methods. Furthermore, the proposed CFC framework offers a flexible and effective open-set classification solution that can be easily applied to other data types.

Related Works

Open-set classification, which identifies unknown classes while classifying known classes, has been widely studied in images and text. Representative OOD detection methods can be broadly categorized into post-hoc detection (Hendrycks and Gimpel 2016; Liu et al. 2020; Park, Jung, and Teoh 2023), generative model-based approaches (Cai and Li 2023; Kirichenko, Izmailov, and Wilson 2020; Nalisnick et al. 2018; Neal et al. 2018), and outlier exposure techniques (Hendrycks, Mazeika, and Dietterich 2018; Wang et al. 2023; Hu and Khan 2021). Post-hoc models employ various scoring functions (Liang, Li, and Srikant 2017; Sun and Li 2022; Zhu et al. 2023; Liu et al. 2020; Huang, Geng, and Li 2021) to identify OOD samples, but struggle when test label spaces differ from training, often requiring costly retraining. Generative methods and outlier exposure approaches attempt to incorporate synthetic OOD samples or auxiliary data to train models for OOD detection. However, these generated OOD samples or auxiliary data often fail to accurately represent true OOD samples, limiting their effectiveness.

Recently, there has been growing interest in graph open-set classification (Wu, Pan, and Zhu 2020; Um et al. 2025; Chen et al. 2025; Ma et al. 2024a). Many classification-based methods (Song and Wang 2022; Yang, Lu, and Gan 2023; Wu et al. 2023; Ma et al. 2024a) have been proposed to utilize structural information for ID classification and OOD detection. However, these methods are often time-consuming and lack flexibility. To improve efficiency, several works (Gong and Sun 2024; Yang et al. 2025; Wu et al. 2023; Wang et al. 2025) employ energy-based propagation schemes to detect OOD samples. Additionally, Zhang et al. (2023) attempted to generate synthetic samples to approximate the OOD space. However, the generated sample distribution often fails to accurately reflect the true OOD distribution, causing these methods to struggle with identifying challenging OOD samples. Importantly, existing open-set classification methods, including node-level (Bao et al. 2024; Gong and Sun 2024; Zhang et al. 2024b,a) and graph-level (Yin et al. 2024; Shen et al. 2024; Guo et al. 2023b; Liu et al. 2023b), typically

identify multiple unknown classes as one OOD label. In this paper, we propose a more challenging problem: OOD classification, which involves identifying multiple classes.

Problem Definition and Preliminaries

We study open-set node classification in graphs. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$, let $|\mathcal{V}| = N$. Each node $v_i \in \mathcal{V}$ has a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$, forming a feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, and a label vector $y_i \in \{0, 1\}^C$, where C is the number of ID classes. Node connections are represented by the adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{ij} = 1$ if $(v_i, v_j) \in \mathcal{E}$, otherwise $\mathbf{A}_{ij} = 0$. We consider that the full node set $\mathcal{V}$ is partitioned into training set $\mathcal{V}_{\text{train}}$, validation set $\mathcal{V}_{\text{val}}$, and test set $\mathcal{V}_{\text{test}}$. In a typical closed-set node classification task on graph $\mathcal{G}$, with an ID label space $\mathcal{Y} = \{1, \ldots, C\}$, GNN models predict each node in the test set with a certain ID class in $\mathcal{Y}$.

OOD detection in open-set node classification problem. In open-world scenarios, the test set may contain unknown class nodes whose labels fall outside the ID label space. Given a set of ID training samples $T = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the goal of OOD detection is to learn a $(C+1)$-class classifier f_{C+1} using T. This classifier should be capable of: (1) classifying ID samples into their respective C ID classes, and (2) identifying OOD samples as belonging to a single OOD class.

OOD classification in open-set node classification problem. In this paper, we extend the challenge from OOD detection to OOD classification, formulating it as a comprehensive open-set classification problem. Specifically, we aim to learn a $(C+u)$-class classifier f_{C+u} using T to classify (1) ID samples into the corresponding C ID classes, and (2) OOD samples into u distinct OOD classes. Notably, u is not predefined in the open-set scenario.

Methods

To address the challenges of OOD classification, we must tackle two critical questions: (1) How can we approximate the OOD space without labeled information? (2) How can we derive meaningful outlier class labels?

In this paper, we propose a Coarse-to-Fine open-set Classification (CFC) framework to progressively achieve advanced OOD classification. First, we design LLM-based prompts specifically tailored for coarse-grained graph node OOD identification. In this step, we leverage the expert knowledge and reasoning capabilities of LLMs to detect OOD samples relevant to the test domain and generate a candidate OOD label space. Next, based on the semantic OOD samples identified by the LLM, we introduce a GNN-based fine-grained classifier for ID classification and precise OOD detection. This step enhances granularity by denoising and OOD data augmentation. Furthermore, we provide a theoretical analysis demonstrating the benefits of integrating semantic OOD samples and the refined augmentation method, which expand the OOD subspace and smooth decision boundaries. Finally, we conduct OOD classification by employing LLM prompts with the refined OOD label space.

A Coarse-Classifier with LLMs

Large Language Models, with their extensive knowledge, have demonstrated impressive zero-shot and few-shot capabilities, particularly for node classification tasks on text-attributed graphs (TAGs) (Chen et al. 2023; Guo et al. 2023a; Chen et al. 2024a), where each node and edge in the graph is associated with a text sentence. In an open-set setting, we explore the OOD space by leveraging the capabilities of LLMs. Using ID labels from the training set, the LLM is employed to predict whether the label of a test node belongs to the provided ID label space, acting as a binary classifier to differentiate between ID and OOD samples.

According to ID label space, we categorize identification tasks into two types: **Easy-Reject** and **Hard-Reject**, as described below. We then elaborate on the corresponding LLM prompts designed to facilitate confidence-aware OOD identification and to generate the potential OOD label space.

Easy-Reject. This refers to scenarios where the ID classes in the label space contain a small proportion of their respective major categories, making it easier to reject ID samples as OOD class. Building on the existing ID label space, we prompt the LLM to determine whether the label of the input test node belongs to the provided ID classes using confidence-aware prompts (Chen et al. 2023). The confidence score associated with this identification is essential, as LLM annotations, similar to human annotations, can exhibit a degree of label noise. This confidence score helps assess the quality of detection and filter out noisy labels. Since these samples are likely to be rejected as OOD, we design the LLM prompt with a restriction: annotate samples as OOD only when the LLM is highly confident. If the test node is identified as an ID sample, we prompt the LLM to provide its category within the specified ID label space. Otherwise, the LLM will offer an outlier class label beyond the ID label space. Finally, we obtain the label (ID or OOD), the LLM's confidence score, and the category for each test sample, as illustrated in Fig. 2.

Hard-Reject. This refers to cases where the ID classes in the label space encompass a large proportion of their respective main categories, making it easier to accept OOD samples as ID ones. Building upon the existing ID label space, we first guide the LLM to summarize these classes and identify their respective major categories. Next, we prompt the LLM to provide possible outlier class labels that fall within the major categories but are not included in the ID class labels, creating a candidate OOD label space. Subsequently, input with the candidate OOD classes, we use the LLM to determine the label of the input test node, generate a confidence score, and provide a predicted category (as illustrated in Fig. 2).

In general, Easy-Reject is used for small coverage and far-OOD cases, while Hard-Reject is used for large coverage and near-OOD cases.

GNN-based Fine-Classification

Assume we have obtained a coarse-grained OOD set $\mathcal{V}_{\text{ood}}$ through LLM-based OOD detection. The samples in $\mathcal{V}_{\text{ood}}$ are semantic OODs with potential categories, and are closely aligned with the true test OOD space, providing a more representative and structured foundation for OOD detection.

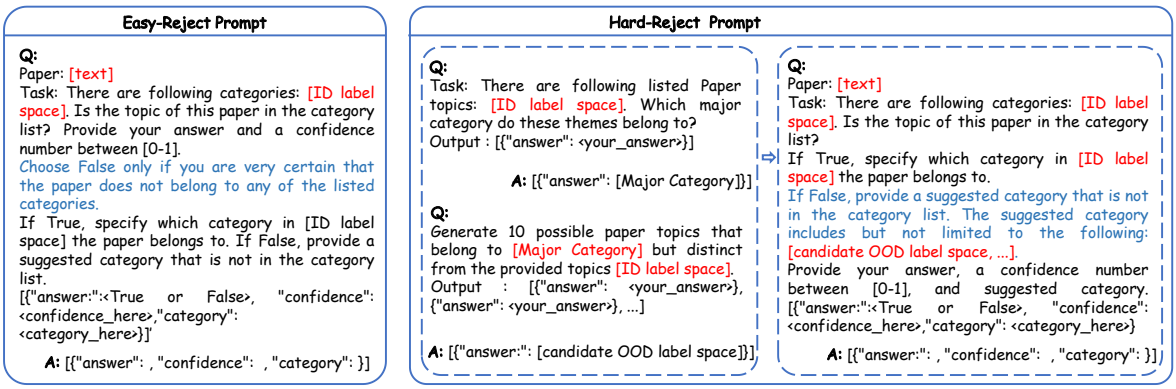


Figure 2: LLM prompts for Easy-Reject and Hard-Reject OOD detection include both Q(uestion) and A(nswer) contents. The inputs are [text] (describing the graph node) and [ID label space] (a list of ID categories, e.g., [machine learning, neural networks, ...]). For Hard-Reject OOD detection, we first determine the [Major Category] of ID classes and the [candidate OOD label space], then use [text], [ID label space], and [candidate OOD label space] for OOD detection and category generation.

We further construct a GNN-based classifier with $(C+1)$ labels to perform ID classification and OOD detection. Given that the OOD samples identified by the LLM may contain some noise (i.e., misclassified ID samples) or be insufficient in number, we employ a label propagation method for denoising and utilize an improved mixup method (Verma et al. 2019; Han et al. 2022) for OOD data augmentation.

Denoising. We correct falsely identified OOD samples from LLM-based OOD detection using a label propagation method. We assume the initial label matrix $\mathbf{Y}^{l(0)} = [y_1^{l(0)}, y_2^{l(0)}, \cdots, y_N^{l(0)}]$ consists of one-hot label indicator vectors for ID training nodes in $\mathcal{V}_{\text{train}}$ and OOD nodes in $\mathcal{V}_{\text{ood}}$, while having zero vectors for unlabeled nodes. By propagating the labels with the normalized adjacency $\mathbf{D}^{-1}\mathbf{A}$, the k^{th} iteration of label propagation (Zhu 2005; Wang and Leskovec 2020) is formulated as $\mathbf{Y}^{l(k)} = \mathbf{D}^{-1}\mathbf{A}\mathbf{Y}^{l(k-1)}$. At each iteration, the ID training samples are reset to their initial labels: $y_i^{l(k)} = y_i^{l(0)}, \forall i \in \mathcal{V}_{\text{train}}$. This is to maintain the label information of the ID training nodes so that the other nodes do not overpower the original labeled ones, as the initial labels would otherwise fade away.

After K-order label propagation, we can obtain the label matrix $\mathbf{Y}$. For candidate OOD samples, their labels are updated using the maximum probability in Y^K. We discard OOD samples that are predicted as ID in $\mathcal{V}_{\text{ood}}$, and achieve new OOD set $\mathcal{V}_{\text{ood}}^{'}$.

OOD Data Augmentation. We consider the practical case in which LLM just identifies a small number of samples as OOD. Having a sufficient number of semantic OOD samples is crucial for representing the OOD space and improving open-set classification. How to obtain more stable OOD samples? Manifold mixup (Verma et al. 2019), as a data augmentation method, has been theoretically and empirically shown to improve the generalization and robustness of deep neural networks for images, by training neural networks on linear combinations of hidden representations of training examples. In this work, we extend manifold mixup to augment OOD data for improved performance.

For a well-trained classifier, features of nodes that belong to the same classes are close to each other, while those from different classes are distant. Ideally, clear boundaries separate the different classes. Since nodes whose features are close to the boundaries are more likely to be less representative to their own classes, we generate OOD samples using nodes near the boundary regions.

We collect K nodes with low classification confidence in the training set. Then, the manifold mixup is applied on these near boundary nodes and the center of the OOD samples in $\mathcal{V}_{\text{ood}}^{'}$ as

$$\begin{cases} \tilde{x}_i = \alpha \boldsymbol{h}_i^k + (1-\alpha)\boldsymbol{h}_c^k, i \leq K \\ \tilde{y}_i = C+1 \end{cases} \tag{1}$$

where $\boldsymbol{h}^k = \text{GNN}\left(\mathbf{A}, \boldsymbol{h}^{k-1}\right)$ is the hidden embedding with a GNN encoder, $\boldsymbol{h}_c^k$ is the center embedding of OOD samples, $\alpha > 0$ is a hyperparameter to control the distance between the generated samples and the existing OOD samples. Finally, we obtain an augmented OOD set $\mathcal{V}_{\text{ood}}^a = \{\mathcal{V}_{\text{ood}}^{'}, \mathcal{V}_a\}$ where $\mathcal{V}_a$ is the generated OOD set.

ID Classification and OOD Detection. We train a GNN-based classifier f_{C+1} on training set $\mathcal{V}_{\text{train}}$ and the augmented OOD set $\mathcal{V}_{\text{ood}}^a$ for ID classification and OOD detection. Taking GCN as an example, a two-layer GCN model can be formulated as

$$\boldsymbol{Z} = \text{softmax}\left(\hat{\mathbf{A}}\,\text{ReLU}(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}^{(0)})\mathbf{W}^{(1)}\right), \tag{2}$$

where, $\boldsymbol{Z}$ is the GCN's output predictions, $\hat{\mathbf{A}} = \hat{\mathbf{D}}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I}_n)\hat{\mathbf{D}}^{-\frac{1}{2}}$ is the normalized $\mathbf{A}+\mathbf{I}_n$ matrix by the degree matrix $\hat{\mathbf{D}}$, and $\mathbf{W} = (\mathbf{W}^{(0)}, \mathbf{W}^{(1)})$ are the weights of the two-layer GCN model. For graph node classification, the objective function $\mathcal{L}$ is

$$\mathcal{L} = -\frac{1}{|\mathcal{V}_{\text{train}} \cup \mathcal{V}_{\text{ood}}^a|} \sum_{v_i \in \mathcal{V}_{\text{train}} \cup \mathcal{V}_{\text{ood}}^a} \boldsymbol{y}_i^\top \log(\boldsymbol{z}_i), \tag{3}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ are the label and prediction of node v_i. We use the trained GNN-based open-set classifier to predict test

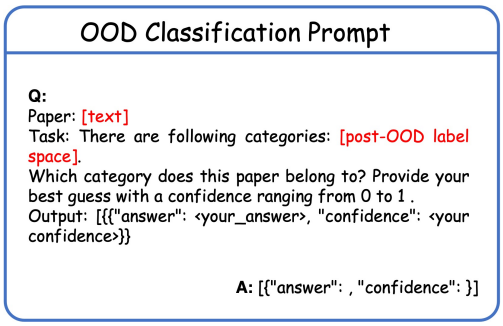


Figure 3: LLM prompts with [text] and [post-OOD label space] for OOD classification.

labels and obtain the final predicted OOD set, denoted as $\mathcal{V}_{\text{ood}}^{f}$.

OOD Classification

After detecting the OOD samples, we proceed with OOD classification for the nodes in $\mathcal{V}_{\text{ood}}^{f}$, by leveraging the potential OOD label space (consisting of the possible OOD categories) discussed in subsection Coarse-Classifier with LLMs.

We employ similarity measures (such as word-level or semantic-level comparisons, e.g., TF-IDF (Ramos et al. 2003)) to merge similar categories and filter out categories with too few samples. After post-processing, we obtain a set of OOD categories, $\{l_1, l_2, \cdots, l_u\}$, forming the post-OOD label space. We then use LLMs to generate annotations for the OOD samples in $\mathcal{V}_{\text{ood}}^{f}$ based on this post-OOD label space, as illustrated in Fig. 3.

Theoretical Analysis

In CFC, we integrate augmented semantic OOD samples by proposing a method to mix the hidden embeddings of samples from ID classes and coarse OOD samples obtained through coarse-grained OOD detection. We theoretically demonstrate the advantages of this approach in extending and flattening the OOD subspace, which leads to improved OOD detection.

Definition 3.1 (Space Dimension)**.** *Given a hypothesis space $\mathcal{H}$, the space dimension is the rank of the matrix formed by the set of vectors that span the subspace.*

Assumption 3.1 (Feature Space Dimension)**.** *Given a feature space $\mathcal{X} \subset \mathbb{R}^d$, assume that the features of $\mathcal{X}$ belong to C distinct classes. Then, the dimension of the feature space is $d - C$.*

Theorem 3.1 *Given an open-set classification task with an ID label space $\mathcal{Y}_{id} = \{y_1, y_2, \ldots, y_C\}$ and an OOD label y_{ood}, assume the existence of the ID feature space $\mathcal{H}$ and the OOD space $\mathcal{H}^{'}$ related to practical test domain. The proposed CFC, trained on both ID samples and semantic OOD samples related to the OOD space $\mathcal{H}^{'}$, forms a larger subspace with dimension $dim(\mathcal{H} + \mathcal{H}^{'}) - (C + 1)$, compared to general methods that lie within $dim(\mathcal{H}) - (C + 1)$. Consequently, CFC results in a smoother and flatter decision boundary for OOD detection.*

Experiments

In this section, we evaluate the performance of our proposed CFC method for graph node classification in an open-set setting by investigating the following questions: **Q1.** How does the performance of CFC compare to other classification methods? **Q2.** What is the effect of different prompts and LLMs on coarse-grained OOD detection? **Q3.** How do denoising and OOD data augmentation in fine-grained OOD detection affect the performance of CFC? **Q4.** What is the impact of the potential OOD label space on OOD classification?

Experimental Settings. In this paper, we utilize widely used graph datasets from different domains—textual graphs (Cora (McCallum et al. 2000), Citeseer (Giles, Bollacker, and Lawrence 1998), WikiCS (Mernyei and Cangea 2020), and DBLP (Ji et al. 2010)) and non-textual graphs (Amazon-Computer and Amazon-Photo (Ni, Li, and McAuley 2019))—for open-set node classification. For each dataset, multiple classes are designated as out-of-distribution (OOD) classes (i.e., $u = 2$), while the remaining classes are considered in-distribution (ID) classes. For the ID classes, 50% of the nodes are sampled for training. The remaining ID samples and all OOD samples are split as 40% for validation and 60% for testing.

For all datasets, we adopt the text-attributed graph versions from (Chen et al. 2023; Liu et al. 2023a; Chen et al. 2024b). We use four popular large language models (LLMs), including e5-large-v2 (e5, (Wang et al. 2022)), Sentence Transformer (ST, (Reimers 2019)), Llama2-7b and Llama2-13b (Touvron et al. 2023), to generate embeddings as original input features. We utilize GPT-4o (Achiam et al. 2023) for detecting OOD samples and generating potential outlier class labels for coarse-grained OOD identification, as well as for the final OOD classification. In the coarse classifier with LLM, we apply a confidence threshold of 0.7 to identify OOD samples, using Easy-Reject for Cora, DBLP, and WikiCS, and Hard-Reject for Citeseer, Computer, and Photo. In fine-grained classification, we generate 100 OOD samples for text datasets, and over 2000 for Computers and Photo datasets using improved manifold mixup. We compare our method with popular closed-set classification methods and state-of-the-art open-set classification methods, including GCN_Poser (Zhou, Ye, and Zhan 2021), G^2Pxy (Zhang et al. 2023), GNNSafe (Wu et al. 2023), NodeSafe (Yang et al. 2025), and GOLD (Wang et al. 2025).

Comparison With Other Node Classification Methods

We conducted two tasks: the traditional open-set graph node classification (ID classification and OOD detection) and the newly proposed OOD classification (mean accuracy (%) over 5 different runs).

OOD Detection Table 1 presents a comparison of closed-set and recent state-of-the-art open-set graph node classification methods for ID classification and OOD detection across four text-attributed graph datasets. Here, CFC adopts e5-large-v2 as the feature encoder. We observe that the proposed CFC method outperforms all other baselines across

Methods	Cora			Citeseer			WikiCS			DBLP		
	ID	OOD	overall	ID	OOD	overall	ID	OOD	overall	ID	OOD	overall
GCN_softmax	90.25	0.0	62.76	76.15	0.00	38.60	73.67	0.0	53.01	91.11	0.0	56.62
GCN_sigmoid	90.64	0.0	63.03	76.07	0.00	38.56	60.18	0.00	43.30	91.54	0.00	56.89
GCN_softmax_τ	81.13	66.98	76.84	57.64	81.97	69.60	41.91	58.84	46.68	79.15	45.14	66.28
GCN_sigmoid_τ	85.17	62.18	78.16	67.52	75.41	71.41	54.13	67.03	57.74	71.11	61.54	66.31
GCN_PROSER	84.21	71.18	80.65	71.25	76.01	72.74	48.13	44.19	46.74	63.01	68.75	65.18
G^2Pxy	85.65	72.46	81.63	71.52	77.30	74.36	58.40	57.09	58.03	65.88	62.63	64.65
GNNSafe	79.06	62.92	74.14	71.43	36.18	53.64	83.56	73.29	79.59	93.85	47.25	76.21
NodeSafe	87.93	80.63	85.71	73.97	53.81	64.03	86.66	42.99	70.03	94.52	42.50	74.83
GOLD	87.48	66.54	81.11	70.33	37.75	53.26	73.60	32.83	58.12	94.05	43.68	74.98
GTP-4o	72.23	58.08	68.62	46.82	48.18	47.50	67.33	67.65	67.43	84.26	56.78	66.11
CFC (wo / D/M)	85.44	94.50	88.20	76.59	72.92	74.77	79.18	81.76	79.91	75.18	87.62	83.40
CFC (wo / M)	88.63	91.75	89.58	79.29	67.49	73.44	80.19	77.13	79.32	76.25	85.00	82.03
CFC (wo / D)	86.23	94.91	88.87	72.40	**83.00**	**77.65**	75.15	**89.95**	79.34	75.37	**88.37**	83.96
CFC	87.49	**95.74**	**90.00**	73.92	80.57	77.21	80.19	81.89	**80.44**	78.47	86.89	**84.03**

Table 1: Comparison of different methods for ID classification and OOD detection across four datasets with two OOD classes. Note that CFC (w/o D/M) refers to the CFC without the Denoising and Manifold Mixup data augmentation techniques.CFC uses GCN as the backbone for fine classification.

Methods	Amazon-Computer			Amazon-Photo		
	ID	OOD	overall	ID	OOD	overall
softmax	81.87	0.0	41.98	82.46	0.0	70.03
sigmoid	73.85	0.0	37.51	67.55	0.0	57.37
softmax_τ	81.84	0.08	41.60	82.43	0.60	70.10
sigmoid_τ	14.15	93.11	52.33	41.20	**86.66**	47.92
GNNSafe	70.62	42.66	56.86	69.10	12.85	60.62
NodeSafe	86.73	66.64	76.90	84.00	48.09	78.58
GOLD	73.18	32.74	52.87	69.58	3.50	59.61
CFC	78.15	**86.54**	**82.28**	82.81	76.14	**81.81**

Table 2: Comparison of different methods for ID classification and OOD detection across two nontextual datasets.

all datasets by significantly large margins. Specifically, CFC achieves over or around a 10% improvement over the second-best in terms of overall accuracy on Cora, WikiCS, and DBLP. Furthermore, even without denoising and OOD data augmentation, CFC (wo / D/M) delivers comparable or better results on all datasets compared to other baselines. This suggests that incorporating semantic OOD information related to true OOD domain during training benefits the model, which is a reasonable outcome. Additionally, we observe that GPT-4o recognizes only about half of OOD samples when provided with the ID label space in most cases, making it unsuitable for high-stakes applications. The effectiveness of CFC on non-textual graph datasets is demonstrated in Table 2.

Table 1 provides detailed classification accuracy for both known (ID) and unknown (OOD) classes. While ID classification performance slightly decreases compared to closed-set methods (e.g., from 90.64% to 87.49% on Cora when comparing CFC to GCN_sigmoid), OOD detection accuracy significantly improves from 0% to 95.74%, which is remarkable. Compared to open-set classification methods such as G^2Pxy, CFC improves OOD detection accuracy from 72.46% to 95.74%, while also enhancing ID classification from 82.65% to 87.49% on Cora. The same trend is observed in other datasets. This shows that CFC better separates known and unknown classes by using semantic OOD.

OOD Classification Without extra label information, it is evident that existing closed-set and open-set classification methods are not equipped to handle OOD classification when multiple OOD classes are present. Although some methods, like GCN_Poser and G^2Pxy, generate OOD samples during training, they struggle to differentiate between various OOD classes without access to OOD label information. Leveraging the annotation capabilities of LLMs and special designed prompt, our proposed CFC method successfully classifies unknown samples into different OOD labels. Notably, using the post-OOD label space and GPT-4o, we achieve accuracies of 69.76%, 70.30%, 57.96%, and 48.45% on Cora, Citeseer, WikiCS, and DBLP, respectively, for two OOD classes.

Impact of Different Prompts and LLMs in Coarse-Classifier

LLM Prompts. We investigated the effectiveness of using a constraint (rejecting with high confidence for **Easy-Reject** detection and integrating a candidate OOD label space for **Hard-Reject** detection) when designing LLM prompts. The rest of the LLM prompt content remains unchanged. Specifically, the **reject with high confidence** LLM prompt instructs the model to classify an input sample as OOD only when it is highly confident. The LLM prompt with the **candidate OOD label space** provides the model with additional label choices and encourages consideration of extra labels when making decisions. As shown in Fig. 4 (a), without the constraint, OOD performance degrades on AUROC metrics, underscoring the importance of the proposed constraint (Cora utilizes the **reject with high confidence** LLM prompt, and Citeseer uses the **candidate OOD label space** LLM prompt).

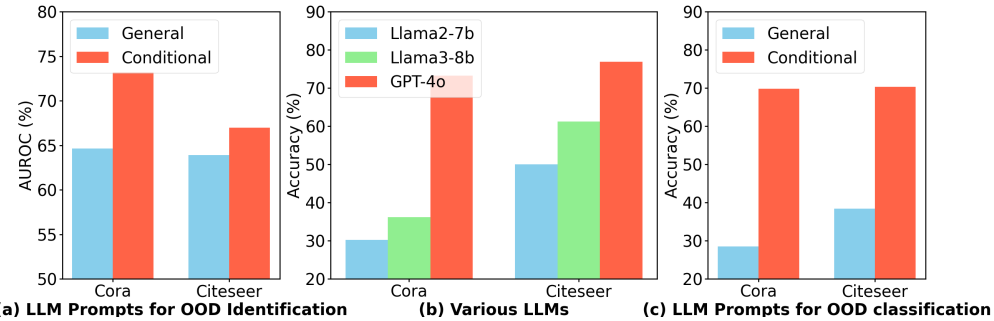


Figure 4: Ablation study on (a) LLM prompts for OOD identification, (b) Various LLM for OOD identification, and (c) LLM prompts for OOD classification on Cora and Citeseer.

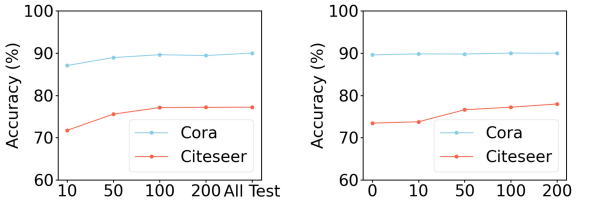


Figure 5: Study on the effect of (a) the number of identified OOD samples from coarse-grained classification, and (b) the number of generated OOD samples by manifold mixup method for the CFC performance on Cora and Citeseer.

Various LLMs. We conducted experiments with various LLMs to gain a more comprehensive understanding of their ability to detect OOD samples. Specifically, we use Llama (Llama2-7b and Llama3-8b) (Touvron et al. 2023) and GPT-4o for OOD detection. Llama models use few-shot learning with one ID and one OOD example; GPT-4o uses zero-shot learning. All prompts follow the constriction strategy. Results on Cora and Citeseer (Fig. 4 (b)) show GPT-4o outperforms Llama, with Llama3-8b slightly better than Llama2-7b.

Impact of Different Strategies in Fine-Classification

We conducted ablation studies to evaluate different strategies, including denoising and data augmentation methods for GNN-based fine-grained classification under the case $u = 2$. As shown in Tables 1, both CFC (wo / D/M) and CFC (wo / D) models, which exclude the denoising, consistently achieve lower ID accuracies across all datasets compared to their counterparts, CFC (wo / M) and CFC with denoising. The CFC (wo / D) model, which employs the improved manifold mixup method to generate more OOD data, demonstrates significant improvement over CFC (wo / D/M) across all datasets. Manifold mixup serves as a regularizer, encouraging the neural network to make less confident predictions on interpolated hidden representations. By integrating ID classes with coarse OOD classes, CFC produces a GNN-based classifier with smoother decision boundaries across multiple representation levels. We also found that data augmentation has a greater impact than denoising, since the LLM-designed prompt already helps reduce noise.

In fine-grained detection, the OOD samples identified from coarse-grained detection play a crucial role. Given LLMs' high cost on large test sets, we evaluate how the number of identified OOD samples affects CFC's performance. As shown in Fig. 5 (a), CFC performs well even with few identified OOD samples, highlighting the advantages of incorporating semantic OOD samples. This demonstrates CFC's scalability to large datasets, with more identified OOD samples further improving accuracy on Cora and Citeseer.

Manifold mixup-based data augmentation method is a key component of CFC. To assess its impact, we examine how the generated OOD samples influence CFC's performance. Fig. 5 (b) shows that increasing generated OOD samples improves performance on Citeseer and maintains stable performance on Cora.

Impact of OOD Label Space for OOD Classification

We examine the impact of the post-OOD label space in LLM prompts for OOD classification. Specifically, we evaluate the performance of LLM annotations with and without the potential OOD label space on the predicted OOD samples from fine-grained OOD detection. In coarse-grained detection, the LLM generates potential OOD labels using a general prompt without the post-OOD label space as a baseline. For datasets like Citeseer, even when candidate OOD labels are provided, the label space remains broad and comparable to the general prompt. Fig. 4 (c) shows that a conditional prompt with a post-OOD label space substantially improves accuracy.

Conclusion

We address open-world OOD classification with CFC. LLM prompts enable coarse OOD detection and construct a candidate OOD label space, from which semantic OOD samples are generated to train a GNN-based fine-grained classifier. A refined LLM prompt performs final OOD classification. Experiments demonstrate state-of-the-art OOD detection and strong OOD classification performance on graph datasets. CFC assumes graph nodes can be described by text; if not, LLMs may struggle with OOD detection and classification. It also depends on the LLM's knowledge of ID categories. To reduce reliance on large LLMs, fine-tuning smaller models on domain-specific data and integrating retrieval-augmented generation (RAG) into CFC is a promising direction.

Acknowledgments

This work is in part supported by the National Natural Science Foundation of China (Grant No. 62276067) and Australian Research Council Discovery project DP230101534.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bao, T.; Wu, Q.; Jiang, Z.; Chen, Y.; Sun, J.; and Yan, J. 2024. Graph out-of-distribution detection goes neighborhood shaping. In *ICML*.

Cai, M.; and Li, Y. 2023. Out-of-distribution detection via frequency-regularized generative models. In *WACV*, 5521–5530.

Chen, Y.; Luo, Y.; Song, Y.; Dai, P.; Tang, J.; and Cao, X. 2025. Decoupled graph energy-based model for node out-of-distribution detection on heterophilic graphs. *arXiv preprint arXiv:2502.17912*.

Chen, Z.; Mao, H.; Li, H.; Jin, W.; Wen, H.; Wei, X.; Wang, S.; Yin, D.; Fan, W.; Liu, H.; et al. 2024a. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2): 42–61.

Chen, Z.; Mao, H.; Liu, J.; Song, Y.; Li, B.; Jin, W.; Fatemi, B.; Tsitsulin, A.; Perozzi, B.; Liu, H.; et al. 2024b. Text-space graph foundation models: Comprehensive benchmarks and new insights. *NeurIPS*, 37: 7464–7492.

Chen, Z.; Mao, H.; Wen, H.; Han, H.; Jin, W.; Zhang, H.; Liu, H.; and Tang, J. 2023. Label-free node classification on graphs with large language models (llms). *arXiv preprint arXiv:2310.04668*.

Ge, Z.; Demyanov, S.; Chen, Z.; and Garnavi, R. 2017. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*.

Giles, C. L.; Bollacker, K. D.; and Lawrence, S. 1998. CiteSeer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, 89–98.

Gong, Z.; and Sun, Y. 2024. An Energy-centric Framework for Category-free Out-of-distribution Node Detection in Graphs. In *KDD*, 908–919.

Guo, J.; Du, L.; Liu, H.; Zhou, M.; He, X.; and Han, S. 2023a. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*.

Guo, Y.; Yang, C.; Chen, Y.; Liu, J.; Shi, C.; and Du, J. 2023b. A Data-centric Framework to Endow Graph Neural Networks with Out-Of-Distribution Detection Ability. In *KDD*, 638–648.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *NeurIPS*, volume 30.

Han, X.; Jiang, Z.; Liu, N.; and Hu, X. 2022. G-mixup: Graph data augmentation for graph classification. In *ICML*, 8230–8248. PMLR.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hu, Y.; and Khan, L. 2021. Uncertainty-aware reliable text classification. In *KDD*, 628–636.

Huang, R.; Geng, A.; and Li, Y. 2021. On the importance of gradients for detecting distributional shifts in the wild. *NeurIPS*, 34: 677–689.

Ji, M.; Sun, Y.; Danilevsky, M.; Han, J.; and Gao, J. 2010. Graph regularized transductive classification on heterogeneous information networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 570–586. Springer.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proc. 5th Int. Conf. Learn. Representations*.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2020. Why normalizing flows fail to detect out-of-distribution data. *NeurIPS*, 33: 20578–20589.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Liu, H.; Feng, J.; Kong, L.; Liang, N.; Tao, D.; Chen, Y.; and Zhang, M. 2023a. One for all: Towards training one graph model for all classification tasks. *arXiv preprint arXiv:2310.00149*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *NeurIPS*, 33: 21464–21475.

Liu, Y.; Ding, K.; Liu, H.; and Pan, S. 2023b. Good-d: On unsupervised graph out-of-distribution detection. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 339–347.

Ma, L.; Sun, Y.; Ding, K.; Liu, Z.; and Wu, F. 2024a. Revisiting Score Propagation in Graph Out-of-Distribution Detection. In *NeurIPS*.

Ma, X.; Ma, X.; Erfani, S.; and Bailey, J. 2024b. Training Sparse Graph Neural Networks via Pruning and Sprouting. In *SDM*, 136–144. SIAM.

McCallum, A. K.; Nigam, K.; Rennie, J.; and Seymore, K. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3: 127–163.

Mernyei, P.; and Cangea, C. 2020. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*.

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; Gorur, D.; and Lakshminarayanan, B. 2018. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*.

Neal, L.; Olson, M.; Fern, X.; Wong, W.-K.; and Li, F. 2018. Open set learning with counterfactual images. In *ECCV*, 613–628.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP-IJCNLP*, 188–197.

Park, J.; Jung, Y. G.; and Teoh, A. B. J. 2023. Nearest neighbor guidance for out-of-distribution detection. In *ICCV*, 1686–1695.

Perera, P.; Morariu, V. I.; Jain, R.; Manjunatha, V.; Wigington, C.; Ordonez, V.; and Patel, V. M. 2020. Generative-discriminative feature representations for open-set recognition. In *CVPR*, 11814–11823.

Ramos, J.; et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, 29–48. Citeseer.

Reimers, N. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

Shen, X.; Wang, Y.; Zhou, K.; Pan, S.; and Wang, X. 2024. Optimizing ood detection in molecular graphs: A novel approach with diffusion models. In *KDD*, 2640–2650.

Song, Y.; and Wang, D. 2022. Learning on graphs with out-of-distribution nodes. In *KDD*, 1635–1645.

Sun, Y.; and Li, Y. 2022. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 691–708. Springer.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Um, D.; Lim, J.; Kim, S.; Yeo, Y.; and Jung, Y. 2025. Spreading Out-of-Distribution Detection on Graphs. In *ICLR*.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, 6438–6447. PMLR.

Wang, D.; Qiu, R.; Bai, G.; and Huang, Z. 2025. GOLD: Graph Out-of-Distribution Detection via Implicit Adversarial Latent Generation. *arXiv preprint arXiv:2502.05780*.

Wang, H.; and Leskovec, J. 2020. Unifying graph convolutional neural networks and label propagation. *arXiv preprint arXiv:2002.06755*.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wang, Q.; Fang, Z.; Zhang, Y.; Liu, F.; Li, Y.; and Han, B. 2023. Learning to augment distributions for out-of-distribution detection. *NeurIPS*, 36: 73274–73286.

Wu, M.; Pan, S.; and Zhu, X. 2020. Openwgl: Open-world graph learning. In *ICDM*, 681–690. IEEE.

Wu, M.; Pan, S.; and Zhu, X. 2021. Openwgl: open-world graph learning for unseen class node classification. *Knowledge and Information Systems*, 63(9): 2405–2430.

Wu, Q.; Chen, Y.; Yang, C.; and Yan, J. 2023. Energy-based out-of-distribution detection for graph neural networks. *arXiv preprint arXiv:2302.02914*.

Yang, L.; Lu, B.; and Gan, X. 2023. Graph Open-Set Recognition via Entropy Message Passing. In *ICDM*, 1469–1474. IEEE.

Yang, S.; Liang, B.; Liu, A.; Gui, L.; Yao, X.; and Zhang, X. 2025. Bounded and uniform energy-based out-of-distribution detection for graphs. *arXiv preprint arXiv:2504.13429*.

Yin, N.; Wang, M.; Chen, Z.; Shen, L.; Xiong, H.; Gu, B.; and Luo, X. 2024. DREAM: Dual structured exploration with mixup for open-set graph domain adaption. In *ICLR*.

Zhang, Q.; Li, X.; Lu, J.; Qiu, L.; Pan, S.; Chen, X.; and Chen, J. 2024a. ROG_PL: Robust Open-Set Graph Learning via Region-Based Prototype Learning. In *AAAI*, volume 38, 9350–9358.

Zhang, Q.; Lu, J.; Li, X.; Wu, H.; Pan, S.; and Chen, J. 2024b. CONC: complex-noise-resistant open-set node classification with adaptive noise detection. In *IJCAI*. ICJAI.

Zhang, Q.; Shi, Z.; Zhang, X.; Chen, X.; Fournier-Viger, P.; and Pan, S. 2023. G2Pxy: generative open-set node classification on graphs with proxy unknowns. In *IJCAI*, 4576–4583.

Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2021. Learning placeholders for open-set recognition. In *CVPR*, 4401–4410.

Zhu, J.; Li, H.; Yao, J.; Liu, T.; Xu, J.; and Han, B. 2023. Unleashing mask: Explore the intrinsic out-of-distribution detection capability. *arXiv preprint arXiv:2306.03715*.

Zhu, X. 2005. *Semi-supervised learning with graphs*. Carnegie Mellon University.