
Preference-driven Knowledge Distillation for Few-shot Node Classification

Xing Wei¹ Chunchun Chen² Rui Fan^{1,2,3} Xiaofeng Cao⁴ Sourav Medya⁵ Wei Ye^{1,2*}

¹ College of Electronic and Information Engineering, Tongji University, China

² Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, China

³ National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Xi'an Jiaotong University, China

⁴ School of Computer Science and Technology, Tongji University, China

⁵ Department of Computer Science, University of Illinois Chicago, USA

{xing627, c2chen, yew}@tongji.edu.cn, rui.fan@ieee.org,
xiaofeng.cao.uts@gmail.com, medya@uic.edu

Abstract

Graph neural networks (GNNs) can efficiently process text-attributed graphs (TAGs) due to their message-passing mechanisms, but their training heavily relies on the human-annotated labels. Moreover, the complex and diverse local topologies of nodes of real-world TAGs make it challenging for a single mechanism to handle. Large language models (LLMs) perform well in zero-/few-shot learning on TAGs but suffer from a scalability challenge. Therefore, we propose a preference-driven knowledge distillation (PKD) framework to synergize the complementary strengths of LLMs and various GNNs for few-shot node classification. Specifically, we develop a GNN-preference-driven node selector that effectively promotes prediction distillation from LLMs to teacher GNNs. To further tackle nodes' intricate local topologies, we develop a node-preference-driven GNN selector that identifies the most suitable teacher GNN for each node, thereby facilitating tailored knowledge distillation from teacher GNNs to the student GNN. Extensive experiments validate the efficacy of our proposed framework in few-shot node classification on real-world TAGs. Our code is available at <https://github.com/GEXX-Weixing/PKD>.

1 Introduction

Text-attributed graphs (TAGs [1]), such as citation, webpage, and product graphs [2, 3], have nodes associated with text attributes. Graph neural networks (GNNs) [4, 5] have demonstrated excellent performance and efficiency in node classification on TAGs, which are supported by high-quality labels and effective message-passing mechanisms [6]. However, the manual labeling of nodes is undoubtedly a tedious, expensive, and time-consuming task [7, 8]. In many scenarios, only a few node labels are available. Additionally, nodes often have complex and diverse interaction relationships with each other—their local topologies are intricate—which challenge traditional GNNs with fixed message-passing mechanisms. Compared with GNNs, large language models (LLMs) exhibit impressive zero-/few-shot learning capabilities on TAGs [9, 10, 11]. But the large parameter scale considerably hinders their inference efficiency [12].

A natural idea is *to blend their complementary strengths for few-shot node classification on TAGs*. Knowledge distillation (KD) [13] is a feasible solution. However, directly distilling knowledge from the LLM to GNN is impractical. Firstly, the discrepancy of decoder-only (LLMs) and encoder-only

*Corresponding Author

(GNNs) leads to fundamentally different characteristics in their embedding spaces [14]. And the huge embedding-dimension difference needs sophisticated embedding alignment and also brings high training cost [15]. In contrast, conducting prediction distillation from LLMs to GNNs by annotating node labels can efficiently alleviate the label scarcity and scalability dilemma [16]. The critical question is how to select the nodes for the LLM’s label annotation to effectively enhance teacher GNNs. Generally, one may use uncertainty [17] as a selection metric in the embedding space of GNN. However, owing to nodes’ diverse semantic and complex structural attributes (e.g., local topologies), a single GNN cannot capture the essences of nodes completely [18]. Therefore, we investigate the embedding spaces of various-architecture GNNs to effectively mitigate cognitive limitations [19] associated with relying on a single GNN, thereby better selecting nodes for LLM’s label annotations.

Nevertheless, since nodes have intricate local topologies, which need tailored message-passing mechanisms, *how to tailor for each node the most appropriate message-passing mechanism* is another challenge. Different GNNs provide different prediction attributes for each node during the learning process [20], encompassing the understandings of its topologies, its interaction relationships to other nodes, and its latent patterns. These node-specific attribute differences suggest that a single message-passing mechanism cannot fundamentally handle the entire graph. Some studies [21, 18] distill knowledge sequentially or simultaneously from teacher GNNs without taking into account the node-specific local topologies, resulting in no obvious performance improvement or even performance degradation [22]. Therefore, it is essential to identify the GNN message-passing mechanisms that align with the node-specific attributes.

To this end, we propose a preference-driven knowledge distillation (PKD) framework that unites the complementary strengths of LLMs and various-architecture GNNs for few-shot node classification on TAGs. It mainly includes two modules: GNN-preference-driven Node Selector (GNS) and Node-preference-driven GNN Selector (NGS). The prerequisite of GNS is that the LLM should be able to comprehend the graph topology. Thus, we develop the graph topology aware (GTA) prompts to fine-tune the LLM, enhancing its capacity to comprehend graph topology. GNS fully exploits nodes’ prediction discrepancies among various GNNs to decide nodes whose labels are annotated by the LLM will effectively enhance teacher GNNs, facilitating knowledge distillation from the LLM to teacher GNNs. NGS selects for each node the most appropriate GNN message-passing mechanism, facilitating the tailored knowledge distillation from various teacher GNNs to the student GNN. It regards the fine-tuned LLM as the RL-based (reinforcement learning) agent, which treats all textualized node-specific attributes (including node’s semantic, structure, and prediction attributes) as state and the student GNN’s performance as reward. Our contributions can be summarized as follows:

- We introduce a preference-driven knowledge distillation (PKD) framework to synergize the complementary strengths of the LLM and various GNNs ingeniously for few-shot node classification on TAGs;
- We propose a GNN-preference-driven node selector, effectively determining nodes for annotation by the LLM and promoting knowledge distillation from the LLM to teacher GNNs;
- We propose a node-preference-driven GNN selector to tailor for each node the most appropriate message-passing mechanism, promoting knowledge distillation from teacher GNNs to the student GNN;
- We validate the efficacy of PKD for few-shot node classification on nine TAGs. The experiments show that it even defeats some state-of-the-art methods that use more node labels.

2 Related Work

2.1 Graph Neural Networks

The field of graph learning has been dominated by GNNs. Early GCN [4] introduces a spectral-based graph convolution operation to propagate node information through the graph. GAT [23] uses attention mechanisms to weigh neighbors’ contributions, enabling adaptive learning of neighborhood importance. APPNP [24] enhances message passing by using personalized propagation with a power iteration approach, improving label propagation on graphs. H₂GCN [25] extends GCNs by incorporating higher-order neighborhood information to improve representation power. GPRGNN [26]

combines graph convolution with residual connections to improve propagation efficiency, particularly in graphs with diverse node degrees. HoloNets [27] introduces a dual-filter mechanism with spectral response, extending spectral convolutions to directed graphs. DirGNN [28] defines the in-neighbors and out-neighbors and performs separate propagation and aggregation, improving the message passing through the incorporation of edge directionality. To deal with label scarcity, GCNII [29] introduces initial residual connections and identity mapping to construct a deep GNN while EGNN [30] enforces equivariance constraints for the enhancement of data efficiency and generalization. AGST [31] and IceBerg [32] leverage the different self-training [33] methods to effectively utilize unlabeled nodes.

2.2 Knowledge Distillation

KD is not only used for model compression, but for strengthening purposeful abilities of the student model. GFL [34] extracts structural knowledge from a pre-prepared similar auxiliary graph, distilling it to the target graph for enhancing few-shot node classification performance. KDGA [35] utilizes multiple graph augmentation strategies to make student GNN produce robust node representations after distillation. MSKD [36] mitigates the diverse classification situations requiring for different nodes by capturing multi-scale topological semantics distilled from varying layers. However, the capability of an individual teacher is inherently limited. BGNN [21] distills complementary knowledge from multiple GNN teachers sequentially and integrate it by the adaptive temperature parameter and weight boosting modules. MTAAM [22] distills knowledge of multiple teacher GNNs into an MLP-student, offering quick inference speed without compromising accuracy. FairGKD [37] obtains equitable and informative node representations by synergizing multiple GNN experts into a teacher. DMKD [18] harnesses complementary knowledge from various GNNs and conducts layer-level knowledge distillation to mitigate the constraint of a single teacher. Furthermore, [14] is a label-free method that proposes the LLM-GNN. It uses LLMs to get high-quality annotation through active and confidence-awareness node selection, thereby circumventing the difficulty of label annotation by humans. LinguGKD [15] introduces a kind of ingenious contrastive learning to align the LLM’s semantic features with GNN’s structural features to achieve knowledge transfer. Most of the above knowledge distillation methods do not tailor for each node the most appropriate message-passing mechanism and underperform on few-shot node classification.

3 Method: PKD

In this section, we present the preference-driven knowledge distillation (PKD) framework. PKD involves two key modules: GNN-preference-driven Node Selector (GNS) and Node-preference-driven GNN Selector (NGS). The main goal of the former module is to select node groups whose labels are annotated by the LLM will drastically enhance teacher GNNs. The main goal of the latter module is to select the most appropriate teacher GNN for each node, thereby tackling the complication of node-specific local topologies. The PKD framework is illustrated in detail in Figure 1.

3.1 Background

A text-attributed graph (TAG) is denoted by $\mathcal{G}_T = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{A}, \mathbf{T})$, where $\mathcal{V} = \{v_1, \dots, v_N\}$ is a set of nodes with semantic attributes $\mathbf{T} = \{t_1, \dots, t_N\}$ and \mathcal{E} is a set of edges. Each semantic attribute can then be encoded as a sentence embedding $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N] \in \mathbb{R}^{N \times F}$ with the help of language models. $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix. Given the few-shot node classification task, let $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^Q$ ($Q \ll N$) be the set of labeled nodes with \mathbf{y}_i as the one-hot label of the training sample \mathbf{x}_i and \mathcal{D}_U be the set of unlabeled nodes, respectively. The goal is to accurately predict the labels of nodes that belong to \mathcal{D}_U given few labeled nodes in \mathcal{D}_L .

We assume B teacher GNNs denoted by $\{T_b\}_{b=1}^B$, and $f_{T_b}^\theta$ is the model parameters of T_b . The B logit outputs of teacher GNNs for node v_i are written as $\mathbf{z}_i^T = [\mathbf{z}_{i,1}^T, \dots, \mathbf{z}_{i,b}^T, \dots, \mathbf{z}_{i,B}^T]$, which is the concatenation of the logit of each teacher $\mathbf{z}_{i,b}^T = [z_{i,b,1}^T, \dots, z_{i,b,c}^T, \dots, z_{i,b,C}^T]$ ($1 \leq b \leq B$), where $z_{i,b,c}^T$ is the probability of v_i belonging to class c ($1 \leq c \leq C$) computed by teacher T_b . Our final objective for the KD from node-preference GNNs to the student GNN can be divided into three parts:

$$\mathcal{L}_{KD} = \alpha \cdot \left(-\frac{1}{N} \sum_i^N \tilde{\mathbf{z}}_i^T \cdot \log f_S^\theta(\mathbf{x}_i) \right) + \beta \cdot \left(-\frac{1}{Q} \sum_i^Q \mathbf{y}_i \cdot \log f_S^\theta(\mathbf{x}_i) \right) + \gamma \cdot \left(\frac{1}{N} \sum_{i=1}^N H(f_S^\theta(\mathbf{x}_i)) \right) \quad (1)$$

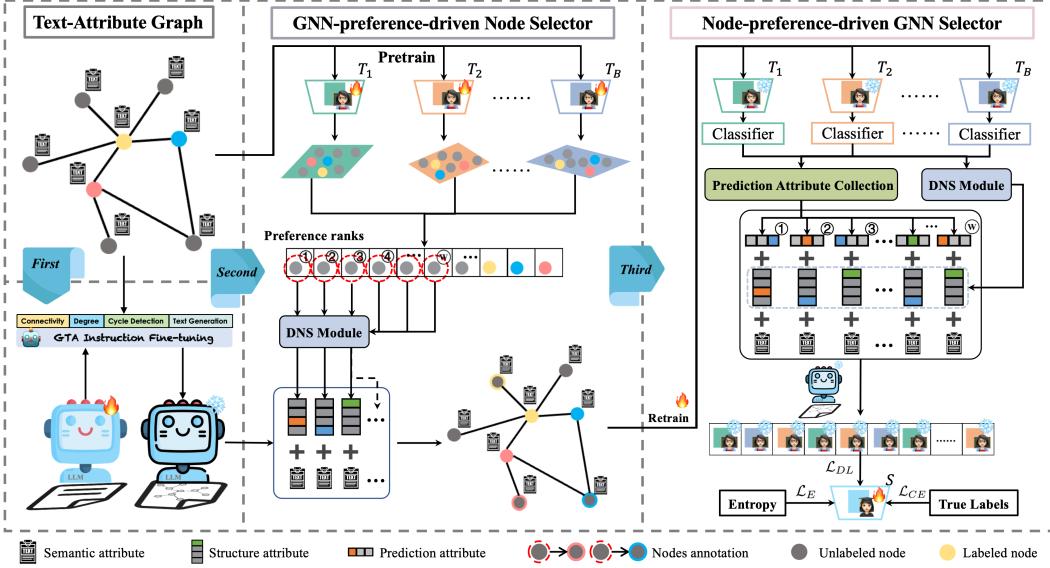


Figure 1: Overview of PKD. The framework has two key modules: GNN-preference-driven Node Selector (GNS) and Node-preference-driven GNN Selector (NGS). Before starting GNS, we first fine-tune the LLM with GTA prompts to enable it to comprehend graph properties. In the GNS module, we exploit the proposed K -uncertainty based on the node prediction uncertainty in each teacher GNN’s embedding space to select nodes. For effectively exploiting the LLM to annotate those selected nodes, we combine the semantic attributes and structure attributes derived from the proposed Distance-based Neighbor Selector (DNS) module on these nodes to construct prompt, promoting the prediction distillation from the fine-tuned LLM to teacher GNNs (T_1, T_2, \dots, T_B). In the NGS module, we select for each node the most appropriate teacher GNN for tailored knowledge distillation. The teacher GNN selection is achieved by reinforcement learning with the fine-tuned LLM as agent.

where α, β, γ are hyper-parameters to balance three losses. For student GNN with parameters f_S^θ , the first loss, distillation loss \mathcal{L}_{DL} , is defined as the cross-entropy between the predictions of the teacher GNNs and that of the student GNN. The $f_S^\theta(\mathbf{x}_i)$ is the Softmax output of student GNN and it denotes the probability distribution of v_i belonging to class c . $\tilde{\mathbf{z}}_i^T = \mathbf{m}_i \otimes \mathbf{z}_i^T$, where \mathbf{m}_i is a one-hot vector denoting which teacher GNN is preferred by v_i . The second loss, \mathcal{L}_{CE} , is the cross-entropy loss in the training of student GNN. Inspired by [38], we add \mathcal{L}_E to the objective as the last part, which makes the logits of student GNN closer to one-hot vectors. The $H(\cdot)$ denotes Shannon entropy.

3.2 LLM Fine-tuning

Recent studies reveal that LLMs possess reasoning abilities [39], but they often underperform compared to even the simple GNNs when tackling graph learning. The key challenge lies in its inability to directly process the raw graph data and understand topology properties, limiting the generalization ability of LLMs in this domain. To address this, we propose GTA prompts fine-tuning.

This method consists of four distinct fine-tuning instruction types, each designed to enhance structural comprehension, such as local connectivity, node degree, cycle structure, and path-based dependencies, by addressing specific tasks: (1) **Connectivity** involves determining whether or not two nodes in an undirected graph are connected; (2) **Degree** requires the LLM to determine the degree of a given node based on the adjacency matrix \mathbf{A} ; (3) **Cycle Detection** requires the LLM to ascertain whether a cycle exists within the given sequence of nodes; (4) **Text Generation** demands

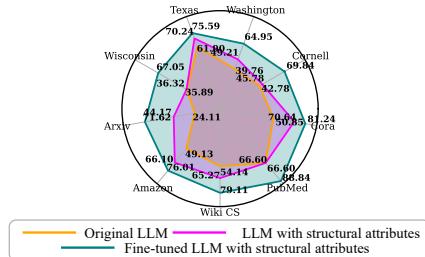


Figure 2: The performance improvements in zero-shot node classification on homophily and heterophily graphs.

the LLM to generate textual contents of given nodes based on the semantic attributes of preceding nodes in the random walk. Through fine-tuning, LLM exhibits significant improvements on the zero-shot node classification task, as demonstrated in Figure 2. More detailed task descriptions and detailed task-specific GTA prompt templates are provided in Appendix B.3.

3.3 GNN-preference-driven Node Selector

After being fine-tuned, the LLM can generate superior node label annotations (as shown in Figure 2). However, how to select nodes for LLM’s label annotation to effectively enhance teacher GNNs (those nodes are assumed to be preferred by GNNs) is a challenging problem. Uncertainty is an essential metric for node selection. It mainly consists of two parts: random uncertainty caused by inherent noise and cognitive uncertainty caused by insufficient observation. The former type is inevitable, so we focus on the latter type. From the perspective of collective consensus [40], we design the GNN-preference-driven Node Selector based on the defined K -uncertainty (δ_K). Specifically, we measure the cognitive disagreement among the teacher GNNs’ SoftMax outputs using the Kullback-Leibler (KL) divergence, and get the preference ranks of all nodes by δ_K , i.e., $\mathcal{V}_{\mathcal{P}\mathcal{R}} = \text{Sort}(\{v_1, \dots, v_N\}, \delta_K(v_1), \delta_K(v_2), \dots, \delta_K(v_N))$. High K -uncertainty of nodes indicates that their prediction uncertainty by GNNs is higher. Those nodes can effectively enhance GNNs if their more accurate labels, annotated by the LLM, are provided to train GNNs, as the following proposition suggests.

Proposition 3.1. *These nodes with higher K -uncertainty (δ_K) are beneficial for GNNs enhancement.*

$$\delta_K(v) \triangleq \sum_{1 \leq i < j \leq B}^B [D_{KL}(f_{T_i}^\theta(v) || f_{T_j}^\theta(v)) + D_{KL}(f_{T_j}^\theta(v) || f_{T_i}^\theta(v))] \propto \delta_v \quad (2)$$

where δ_v is the uncertainty of node v , is defined as $\frac{1}{B} \sum_{i=1}^B D_{KL}(f_{T_i}^\theta(v) || \mathcal{M}(v))$. The $\mathcal{M}(v)$ is the average prediction probability distribution of all B teacher GNNs (See Definition D.1 for details). $D_{KL}(\cdot || \cdot)$ is the function to calculate KL divergence.

$$f_T^{\theta^*}(\tilde{\mathcal{D}}_L) = \arg \min_{v_i \in \{v_{\mathcal{P}\mathcal{R}}^1, v_{\mathcal{P}\mathcal{R}}^2, \dots, v_{\mathcal{P}\mathcal{R}}^W | \delta_K(v_{\mathcal{P}\mathcal{R}}^W) > \tilde{\delta}_K\}} \frac{1}{W} \sum \mathcal{L}(f_T^\theta, v_i) \quad (3)$$

where $\tilde{\mathcal{D}}_L$ is the expanded training dataset. $f_T^{\theta^*}$ is the optimal parameter of teacher GNN. $v_{\mathcal{P}\mathcal{R}}^w$ represents the w -th nodes in the preference rank. W is the number of selected nodes by GNS and the $\tilde{\delta}_K$ is the K -uncertainty threshold depending on the expansion ration.

The proof is given in Appendix D. By selecting these nodes (illustrated in Figure 3), we ensure that the most uncertain and informative nodes are labeled by the LLM to promote the progress of prediction distillation through the cross-entropy function. Correspondingly, GNS also reduces the inference costs associated with LLMs by not querying all nodes in \mathcal{D}_U . To generate high-quality annotations for GNN-preferred nodes, we further design the Distance-based Neighbors Selection (DNS) module, which performs the K-Nearest Neighbor (KNN) search around each selected node across the embedding spaces generated by pretrained teacher GNNs and deletes repeated neighbors. The structure attributes composed of selected neighbors and their textual contents are integrated into the category-induction prompt and inputted into the LLM. Unlike relying solely on neighbors identified by the adjacency matrix (prone to biases from 1-hop homophily), our approach ensures a more robust and diverse selection of high-quality neighbors, facilitating better construction of the category-induction prompt for the LLM. We do not select common KNN neighbors across all the embedding spaces generated by the teacher GNNs, as they may overfit to the adjacency structure.

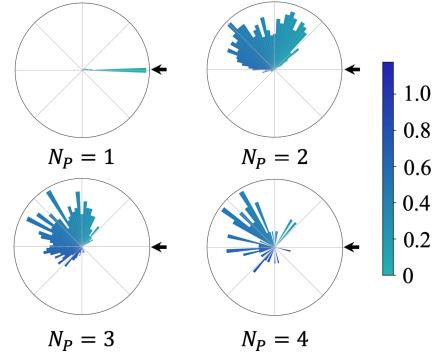


Figure 3: This is exemplified using the CORA dataset. Starting from the arrow and progressing counterclockwise, the KL divergence sum gradually increases, accompanied by a darkening of the triangle colors. The length of each triangle indicates the number of nodes within a specific KL divergence sum range, where N_P denotes the number of classes predicted by the teacher GNNs.

3.4 Node-preference-driven GNN Selector

Distilling knowledge simultaneously from multiple teachers to the student is not a good option since nodes with varying local topologies require distinct message-passing modes for optimal representation updates. To achieve this, we introduce the Node-preference-driven GNN Selector (NGS) to select the most appropriate teacher for each node according to the specific attributes and promote tailored knowledge distillation. For each node in the expanded training data (including the initial few labeled nodes and those selected nodes whose labels are annotated by the LLM), we construct a node-specific prompt by combining its semantic, structural, and prediction attributes derived from the enhanced teacher GNNs. This prompt is then inputted to the fine-tuned LLM to determine the most suitable teacher for this node. The GNN selection task is formulated as a reinforcement learning problem that needs to explore the discrete action space and find a series of assignment actions to get the highest global reward across the expanded training data. Through interaction with the training process, the selector progressively refines its decisions on node-to-teacher assignments, leading to a more efficient and effective assignment strategy. Specifically, the fine-tuned LLM, serving as the agent, selects the most appropriate teacher for each node. The policy is trained to maximize classification accuracy on the expanded training data, with the reward tied to the student’s performance. To address the non-differentiability of the LLM’s decoding process, we add two additional projectors (MLPs) after the logit layer to generate action probabilities and corresponding value estimations, enabling the agent to take discrete teacher-selection actions.

In the RL framework, the elements are structured as $(State, Action, Reward)$. During each iteration, the agent interacts with the environment by receiving all attributes of one node in the expanded training data. The agent then takes an action on which teacher is more appropriate.

State: Each state corresponds to the prompt \mathcal{P}_i of a node, including node-specific semantic, structural, and prediction attributes. These prompts are detailed in Appendix B.2. The size of the expanded training data is denoted as W .

Action: The Policy Model (the fine-tuned LLM combined with an MLP projector) generates a text-related output to indicate its selection from multiple teachers, formulated as a probability distribution vector $\pi_T = [\pi_{T_1}, \pi_{T_2}, \dots, \pi_{T_B}]$, where π_{T_b} denotes the probability of selecting the b -th teacher T_b . The action is determined through sampling.

Reward: The function is correlated with the performance of the student GNN, which is trained by distilling knowledge from the selected teacher for each node. The reward function consists of three key parts: classification accuracy, cross-entropy loss, and distillation loss. It can be written as follows:

$$R = \eta * (\mathcal{L}'_{DL} - \mathcal{L}_{CE}) + (1 - \eta) * A_{cc} \quad (4)$$

where A_{cc} represents the classification accuracy of the student GNN on the expanded training data, η is a hyper-parameter to balance the three parts, where $\mathcal{L}'_{DL} = -\frac{1}{W} \sum_i^W \tilde{\mathbf{z}}_i^T \cdot \log f_S^\theta(\mathbf{x}_i)$, and $\mathcal{L}_{CE} = -\frac{1}{Q} \sum_i^Q \mathbf{y}_i \cdot \log f_S^\theta(\mathbf{x}_i)$.

To effectively optimize the agent’s actions for better knowledge distillation, we employ the simplified version of Proximal Policy Optimization (PPO) [41] algorithm, which retains the core principles. Specifically, we do not instantiate the Reward Model explicitly and calculate the reward based on the performance of the student GNN. The Reference Model is also not explicitly referenced, because the parameter update objective function we utilize involves a comparison with the previous strategy. To avoid large fluctuations between the current and old policies, we adopt the CLIP strategy [41] to limit the update margin. During the KD process, the parameters f_A^θ of NGS, remain fixed, while the parameters f_S^θ of the student GNN are trained. During the NGS process, the parameters f_S^θ of the student GNN are kept fixed to compute the reward, while the parameters f_A^θ of NGS based on the collected rewards from all episodes are optimized. The pseudocode, detailed implementations, and time complexity analysis are provided in Appendix C.

4 Experiments

4.1 Experimental Setup

Datasets In order to assess the few-shot node classification performance of our method on TAGs, we conduct a comprehensive series of experiments across 9 real-world datasets: CORNELL, WASH-

Table 1: Node classification accuracies (%) on real-world datasets. T_1 , T_2 , T_3 , and T_4 denote the teacher GNNs for homophily or heterophily graphs (refer to the descriptions in **Baselines** for more details of the teacher and student GNNs). The OOM stands for Out-Of-Memory. The best results are highlighted in dark gray, while the runner-up results are marked in light gray.

Methods	Dataset	CORNELL	WASHINGTON	TEXAS	WISCONSIN	AMAZON RATINGS	OGBN-ARXIV	WIKI CS	PUBMED	CORA
T_1		58.04 \pm 1.1	57.84 \pm 2.1	53.43 \pm 4.1	59.32 \pm 2.1	41.22 \pm 6.6	56.51 \pm 1.2	81.57 \pm 0.7	83.34 \pm 2.4	87.79 \pm 1.6
T_2		46.29 \pm 0.9	65.00 \pm 2.5	82.83 \pm 2.0	48.30 \pm 2.5	36.69 \pm 0.2	59.19 \pm 5.2	79.08 \pm 1.8	82.52 \pm 2.1	87.59 \pm 0.8
T_3		44.62 \pm 4.3	55.27 \pm 1.7	45.19 \pm 2.5	61.49 \pm 0.6	37.41 \pm 2.2	56.71 \pm 3.6	80.17 \pm 1.6	79.57 \pm 2.3	88.38 \pm 0.6
T_4		32.73 \pm 2.8	58.33 \pm 1.7	63.64 \pm 0.1	62.89 \pm 1.3	48.93 \pm 0.5	53.64 \pm 1.3	72.01 \pm 2.5	55.15 \pm 1.4	77.07 \pm 4.0
GCNII [29] / # LN 5		57.82 \pm 2.8	64.17 \pm 3.1	68.79 \pm 4.3	60.94 \pm 1.5	48.22 \pm 4.3	35.14 \pm 5.6	58.29 \pm 2.8	67.83 \pm 7.7	77.74 \pm 3.7
EGNN [30] / # LN 5		53.38 \pm 7.8	63.33 \pm 1.2	71.72 \pm 2.9	55.97 \pm 3.6	49.03 \pm 8.6	36.15 \pm 3.9	63.97 \pm 6.6	66.12 \pm 9.3	72.85 \pm 0.7
LLMGNN [14] / # LN 5		52.63 \pm 4.3	41.09 \pm 2.2	62.82 \pm 3.6	46.54 \pm 0.9	47.64 \pm 2.0	44.11 \pm 2.5	66.09 \pm 0.4	78.84 \pm 1.1	76.23 \pm 1.7
GAugLLM [48] / # LN 5		62.98 \pm 3.3	65.13 \pm 1.1	73.81 \pm 2.2	62.20 \pm 0.9	42.42 \pm 6.0	53.47 \pm 0.5	83.10 \pm 1.7	85.98 \pm 0.6	79.48 \pm 4.5
Self-training [33] / # LN 5		61.90 \pm 6.1	65.89 \pm 0.5	72.62 \pm 2.4	66.29 \pm 0.9	41.99 \pm 5.0	33.40 \pm 2.5	74.99 \pm 0.9	83.11 \pm 0.4	83.19 \pm 1.7
AGST [31] / # LN 5		71.43 \pm 0.7	70.09 \pm 0.8	68.45 \pm 0.8	70.08 \pm 0.7	43.11 \pm 0.4	OOM	72.49 \pm 3.1	73.75 \pm 0.5	77.25 \pm 5.6
IceBerg [32] / # LN 5		33.33 \pm 11.9	67.76 \pm 2.9	50.00 \pm 4.9	41.53 \pm 2.0	25.99 \pm 1.5	33.63 \pm 1.2	84.88 \pm 0.2	62.41 \pm 9.3	76.23 \pm 2.6
KDGA [35]		54.39 \pm 2.9	60.00 \pm 0.1	66.67 \pm 1.5	58.74 \pm 3.9	38.06 \pm 1.2	OOM	65.03 \pm 4.1	OOM	68.87 \pm 0.8
MSKD [36]		51.27 \pm 4.2	50.39 \pm 0.2	62.63 \pm 2.0	41.51 \pm 0.2	35.60 \pm 5.8	58.27 \pm 1.0	62.73 \pm 2.9	45.86 \pm 0.3	51.61 \pm 0.6
BGNN [21]		58.60 \pm 3.3	56.67 \pm 0.8	56.66 \pm 2.0	59.12 \pm 6.9	37.53 \pm 2.0	46.67 \pm 8.1	56.96 \pm 4.3	76.12 \pm 0.7	71.28 \pm 3.7
MTAAM [22]		72.68 \pm 1.0	73.33 \pm 0.8	80.81 \pm 4.0	71.69 \pm 1.9	39.54 \pm 0.2	32.32 \pm 5.5	65.24 \pm 3.3	83.42 \pm 2.3	79.16 \pm 4.0
FairGKD [37]		61.05 \pm 2.4	60.00 \pm 4.1	84.85 \pm 1.1	57.11 \pm 0.9	43.93 \pm 0.5	42.03 \pm 2.1	60.25 \pm 1.2	70.40 \pm 0.3	69.85 \pm 2.7
RANDOM / # LN 5		54.31 \pm 1.7	58.04 \pm 1.2	58.93 \pm 1.1	58.04 \pm 2.7	57.95 \pm 2.5	54.97 \pm 3.3	65.27 \pm 1.8	66.60 \pm 2.9	70.64 \pm 2.6
VOTING / # LN 5		44.97 \pm 3.0	58.88 \pm 3.5	61.31 \pm 2.0	46.97 \pm 3.6	58.64 \pm 2.1	58.53 \pm 2.0	72.28 \pm 2.2	70.64 \pm 3.1	74.32 \pm 3.1
PKD _{Llama}	# LN 1	74.60 \pm 2.1	76.64 \pm 0.9	80.36 \pm 1.3	69.32 \pm 2.8	64.11 \pm 1.7	53.67 \pm 1.6	79.31 \pm 0.8	83.75 \pm 1.1	85.64 \pm 2.1
	# LN 3	76.72 \pm 0.9	81.36 \pm 1.0	83.33 \pm 0.7	71.49 \pm 1.5	65.64 \pm 0.9	58.65 \pm 2.2	80.01 \pm 0.6	84.34 \pm 0.9	86.18 \pm 1.7
	# LN 5	80.95 \pm 1.1	83.74 \pm 0.4	86.31 \pm 0.5	76.89 \pm 0.9	66.79 \pm 0.3	61.03 \pm 0.7	81.39 \pm 0.4	85.69 \pm 0.3	91.14 \pm 0.3

INGTON, TEXAS, WISCONSIN [25], AMAZON RATINGS [42], OGBN-ARXIV [43], WIKI CS [44], PUBMED, CORA [45]. They have various 1-hop homophily ratios [46] and additional details of the datasets can be found in Appendix A. For the KD-baselines, we partition the nodes of each graph into training, validation, and test sets, allocating 48%, 32%, and 20%, respectively, based on the proportion division mentioned in [47]. For PKD and other baselines, we randomly select 1, 3, and 5 labeled nodes per class as the initial training data and then expand the dataset to 48% of the total using the GNS module. The remaining data is randomly split into 32% for validation and 20% for testing, with the preserved indices for the baselines. This operation is repeated 5 times. We report the average test classification accuracy and standard deviation of each model with parameters that lead to the peak validation accuracy.

Baselines We compare our method against the following baseline models: (i) Advanced GNNs: GCNII [29] and EGNN [30]; (ii) GNNs enhanced by LLMs: LLMGNN [14] and GAugLLM [48]; (iii) self-training for graph learning: Self-training [33], AGST [31] and IceBerg [32]; (iv) Knowledge Distillation (KD) for GNNs: KDGA [35], MSKD [36], BGNN [21], MTAAM [22], and FairGKD [37]. For homophily graphs, the teacher GNNs used are: GCN [4] (T_1), GAT [23] (T_2), APPNP [24] (T_3), H₂GCN [25] (T_4), and the student is GCN; for heterophily graphs, the teacher GNNs employed are: DirGNN [28] (T_1), GPRGNN [26] (T_2), HoloNets [27] (T_3), H₂GCN (T_4) and the student is H₂GCN. The LLM used in the experiments is Llama-3.1-8B-Instruct [49].

4.2 Performance Analysis and Discussion

Notably, # LN 1, # LN 3, # LN 5 indicate only 1, 3, 5 labeled nodes per class are used for training PKD, while the results of the teacher GNNs ($\{T_i\}_{i=1}^4$) and other baselines are trained under the data splitting of 48%/32%/20% as mentioned above. According to Table 1, our method almost achieves the best or second-best accuracy results.

Due to the extreme insufficiency of labels, GCNII and EGNN are restricted in further improvement, although they have distinctive network architectures. Lacking carefully designed fine-tuning and enough cognition makes LLMGNN fail to produce high-quality pseudo labels and is dramatically defeated by our method PKD. Although GAugLLM harnesses LLM for feature and structure augmentations to benefit GNN, its self-training depends only on SoftMax scores to identify candidate nodes to assign pseudo-labels, a method that can sometimes be unreliable. GAugLLM achieves the best result on the PUBMED dataset, but it is outperformed by PKD on other datasets. AGST is excessively dependent on the original graph topology for label propagation, rendering it vulnerable

to structural noise and facing significant challenges when transferred to large-scale graphs, such as OGBN-ARXIV. IceBerg does not perform well on heterophily graphs because its capacity to disseminate information across longer distances is hampered by the proliferation of noise edges. MTAAM shows satisfactory performance on most datasets, due to its ability to autonomously identify the most valuable knowledge from each teacher during training. FairGKD achieves runner-up results on some datasets. The poor performances of KDGA and BGNN result from their excessive sensitivity to GNN selection. MSKD is equipped with the fixed message-passing mechanism, showing that the single message-passing mechanism underperforms on all the datasets compared to PKD. The RANDOM / # LN 5 approach refers to randomly selecting node predictions from 4 teachers, utilizing 5 labeled nodes per class to train teacher GNNs. The VOTING / # LN 5 method selects the most frequently predicted label from 4 teachers as the annotation label. We can see that these two simple and intuitive strategies are defeated by PKD on all datasets.

Our PKD consistently achieves superior node classification results across all datasets, irrespective of the specific type of LLM. The few-shot node classification results after replacing Llama-3.1-8B-Instruct with Qwen2.5-7B-Instruct [50] and Mixtral-7B-Instruct-v0.3 [51] are shown in Table 2.

Table 2: Few-shot node classification accuracy (%) on eight TAGs using three different LLMs. The # LN 1, # LN 3, # LN 5 represent 1, 3, 5 labeled nodes per class, respectively. The best results are highlighted in dark gray, while the runner-up results are marked in light gray.

Methods	Dataset	CORNELL	WASHINGTON	TEXAS	WISCONSIN	AMAZON RATINGS	OGBN-ARXIV	WIKI CS	PUBMED	CORA
PKD _{Qwen}	# LN 1	73.54 \pm 2.6	75.70 \pm 1.1	82.14 \pm 0.8	72.59 \pm 1.3	74.58 \pm 1.1	54.17 \pm 2.2	79.49 \pm 1.2	82.81 \pm 0.8	86.45 \pm 1.0
	# LN 3	77.25 \pm 1.4	77.35 \pm 0.9	84.52 \pm 0.4	73.86 \pm 0.7	75.46 \pm 0.8	60.63 \pm 1.0	80.01 \pm 0.6	83.61 \pm 1.1	87.74 \pm 0.8
	# LN 5	79.84 \pm 0.6	79.63 \pm 0.6	85.71 \pm 0.2	74.24 \pm 0.2	77.69 \pm 0.6	62.62 \pm 2.1	81.21 \pm 0.2	85.96 \pm 0.6	90.07 \pm 0.4
PKD _{Mixtral}	# LN 1	76.31 \pm 2.2	74.42 \pm 1.5	79.41 \pm 3.3	69.81 \pm 0.5	70.02 \pm 1.2	57.69 \pm 0.4	80.56 \pm 0.8	82.42 \pm 0.9	84.87 \pm 2.4
	# LN 3	78.95 \pm 1.2	76.74 \pm 3.1	82.86 \pm 1.7	75.47 \pm 0.8	71.50 \pm 2.6	61.17 \pm 0.6	81.96 \pm 1.3	83.19 \pm 2.7	87.64 \pm 1.1
	# LN 5	81.58 \pm 2.1	81.39 \pm 2.5	85.29 \pm 1.9	77.36 \pm 3.1	73.96 \pm 1.9	62.44 \pm 0.6	83.33 \pm 1.4	84.71 \pm 1.6	88.56 \pm 0.7

Furthermore, to evaluate the quality of LLM-generated pseudo-labels, we compare the node classification performance of PKD and three baselines under different label settings (# LN 5, 48% training ratio expanded by the annotated labels and real labels, respectively). The experiments are conducted on four datasets (CORA, WIKI CS, WASHINGTON, and WISCONSIN). The results are presented in Table 3. For GCNII and IceBerg, they are proposed to tackle the challenge of sparse labels, using the LLM-annotated node labels can improve their performance on all datasets. However, using the same number of real labels achieves better performance.

Table 3: Classification accuracy comparison under different label configurations. The best results are highlighted in dark gray, while the runner-up results are marked in light gray.

Models	Labels configuration	CORA	WIKI CS	WASHINGTON	WISCONSIN
GCNII	# LN 5	77.74	56.29	64.17	60.94
	48% LLM-generated labels	76.69	51.18	70.83	62.50
	48% real labels	81.54	59.17	71.79	65.98
IceBerg	# LN 5	76.23	84.88	67.76	41.53
	48% LLM-generated labels	78.66	71.23	70.12	42.22
	48% real labels	81.94	86.49	72.04	45.43
MSKD	# LN 5	43.91	46.81	45.29	33.33
	48% LLM-generated labels	45.89	54.06	48.17	39.50
	48% real labels	51.61	62.73	50.39	41.51
PKD _{Llama}	# LN 5	90.27	81.39	83.74	76.89

4.3 Ablation Study

Generally, the fine-tuned LLM using our proposed GTA prompts also demonstrates pretty zero-shot node classification performance, surpassing some semi-supervised GNNs from the values in Figure 2.

We assess the significance of the GTA prompts, DNS and $\mathcal{V}_{\mathcal{P}\mathcal{R}}$ with the following default parameter settings: # LN = 3, K = 4. Here, K denotes the number of selected neighbors surrounding the node, to be annotated, within each embedding space of the teacher GNNs structure attributes. In the absence of DNS, neighbors are selected according to the adjacency matrix directly; in the non-use of $\mathcal{V}_{\mathcal{P}\mathcal{R}}$, we expand the training data by random selection.

Table 4: Ablation study for GTA, DNS, and \mathcal{V}_{PR} . \uparrow denotes an accuracy (%) increment. The three components play different roles in the improvement of the performance of our method.

Dataset/Module	GTA	DNS	\mathcal{V}_{PR}	Accuracy	Dataset/Module	GTA	DNS	\mathcal{V}_{PR}	Accuracy
CORA	\times	\times	\times	45.02	AMAZON	\times	\times	\times	42.01
	\times	\times	\checkmark	\uparrow 26.94		\times	\times	\checkmark	\uparrow 13.02
	\times	\checkmark	\checkmark	\uparrow 30.99		\times	\checkmark	\checkmark	\uparrow 16.96
	\checkmark	\checkmark	\checkmark	\uparrow 41.14		\checkmark	\checkmark	\checkmark	\uparrow 23.97

As shown in Table 4, the implementations of GAT prompts, DNS, and \mathcal{V}_{PR} result in varying degrees of performance improvement. Supported by fine-tuning with GTA prompts, the LLM’s enhanced logical reasoning ability, combined with high-quality neighboring nodes, substantially enhances zero-shot node classification capability, leading to superior classification performance improvement. Additionally, we also assess the methods without using reinforcement learning in the teacher selection process, including entropy-based ranking, i.e., selecting the teacher GNN with the highest prediction confidence, random selection, and end-to-end learning. Their relevant results are provided in the Appendix E.2.

To assess the effectiveness of each part in the reward function (Eqn. (4)), we visualize the training processes of three variants in Figure 4: (a) R_1 : The reward function for teacher GNN selection depends solely on the classification accuracy of the student GNN on the expanded training data; (b) R_2 : In addition to classification accuracy, the reward function also incorporates the negative cross-entropy loss ($-\mathcal{L}_{CE}$); (c) R_3 : Building upon R_2 , the reward function also includes the negative knowledge distillation loss ($-\mathcal{L}_{DL}$). As shown in Figure 4, both the three parts contribute to the improved classification performance.

4.4 Sensitivity Analysis

We investigate the impact of the hyper-parameter K on the zero-shot node classification performance. We vary the value of K within the range $\{1, 2, 3, 4, 5\}$ for homophily graphs and heterophily graphs to observe the variation in zero-shot node classification accuracy. As illustrated in Figure 5, accuracy exhibits significant fluctuations as K changes. When $K = 4$, the fine-tuned LLM demonstrates strong performance on most graphs.

To further explore the relationship between the parameter scale of LLM and PKD’s performance, we evaluated Qwen2.5-7B-Instruct with three different parameter scales: 7B/14B/32B parameters. The results are shown in Table 5. Obviously, the classification performance of PKD basically gets better with the increase of parameter scale. This is mainly related to the LLMs with larger parameter-scale have richer knowledge storage and better ability handling complex tasks.

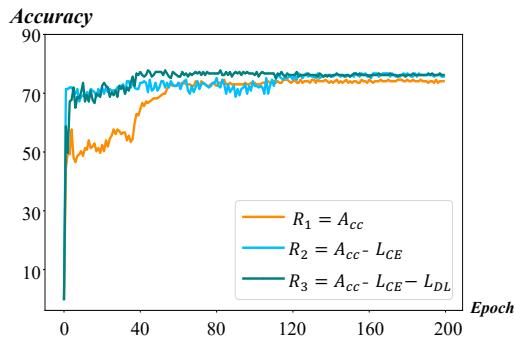


Figure 4: The comparison of different Rewards. When including all three parts simultaneously, our method (the curve in green) performs the best.

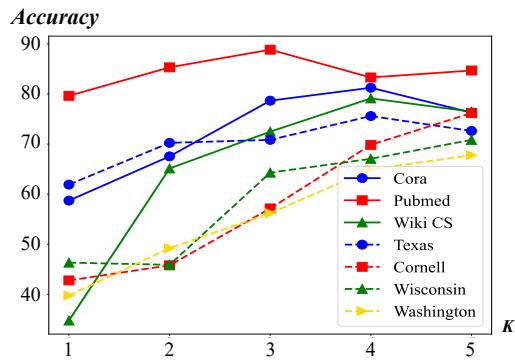


Figure 5: The effects of K on homophily and heterophily graphs. When $K = 4$, zero-shot node classification accuracy of the fine-tuned LLM is the highest on most graphs.

Next, we investigate the ratios of nodes selected for annotating their labels by the LLM as a means to expand the training set. The results are given in Table 6. Increasing the expansion ratio can enhance the performance of PKD. This improvement can be attributed not only to the high-quality label annotation generated by the fine-tuned LLM, but also to the characteristic of PKD that is underpinned by selecting each node the most appropriate teacher GNN for knowledge distillation.

Furthermore, we perform the hyperparameter sensitivity analysis over the loss-weight coefficients $\alpha, \beta, \gamma, \eta$. For the sensitivity analysis of α , we set $\beta = 1, \gamma = 1, \eta = 0.5$. This strategy also applies to the sensitivity analysis of β and γ . As for η , we fix α, β , and γ to their respective best values. The β and η are more sensitive than the other two hyperparameters, because β is related to the supervision signal provided by the LLM-generated annotation labels and η is related to the student GNN’s performance. All the results are reported in the Appendix E.3.

Table 6: The results with different node annotation ratios. The best results are highlighted in dark gray, while the runner-up results are marked in light gray.

Node annotation ratios	10% / # LN 5	20% / # LN 5	30% / # LN 5	40% / # LN 5	48% / # LN 5
AMAZON RATINGS	50.27 ± 6.6	55.91 ± 1.8	62.07 ± 3.5	63.93 ± 1.1	66.79 ± 0.3
CORA	73.37 ± 5.1	77.51 ± 1.1	81.49 ± 2.5	83.27 ± 2.2	91.14 ± 0.3

4.5 Running Time

We also study the training efficiencies of PKD and all baselines. The running times on CORA are shown in Table 7. There is a trade-off between accuracy and time complexity. The incorporation of the LLM undoubtedly boosts the few-shot classification accuracy of GNNs on TAGs, but the training time increases. When applied to the bigger graphs, the time increase will be more obvious.

Table 7: Running time (second per epoch) of each method, including the pretraining process.

Datasets / Methods	T_1	T_2	T_3	T_4	GCNII	EGNN	LLMGNN	GAugLLM	-
CORA	0.006	0.197	0.247	0.035	0.014	0.366	0.630	0.402	-
Datasets / Methods	Self-training	AGST	IceBerg	KDGA	MSKD	BGNN	MTAAM	FairGKD	PKD _{Llama}
CORA	0.016	0.018	0.011	3.911	2.289	1.001	3.318	4.100	7.314

5 Conclusions, Limitations & Future Work

In this work, we have proposed a preference-driven knowledge distillation (PKD) framework for few-shot node classification on TAGs, consisting of GNN-preference-driven Node Selector (GNS) and Node-preference-driven GNN Selector (NGS). Fine-tuned with our proposed GTA prompts, the refined LLM generates high-quality annotations. The GNS effectively determines nodes for the fine-tuned LLM to annotate and promotes knowledge distillation from the LLM to teacher GNNs. The NGS tailors for each node the most appropriate message-passing mechanism, promoting knowledge distillation from teacher GNNs to the student GNN. On various real-world TAGs, our method PKD outperforms almost all advanced GNNs and KD methods for few-shot node classification while using only a few node labels. One limitation of our method is that it is designed for TAGs. Moving forward, we plan to further explore more efficient mechanism of synergizing LLM and GNN to address the limitation of training efficiency as well as datasets beyond TAGs.

6 Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their valuable and constructive comments. This work was supported partially by the National Natural Science Foundation of China under Grants 62176184, 62476109, and 62206108, and the Fundamental Research Funds for the Central Universities.

Table 5: The few-shot node classification accuracy (%) of PKD with different parameter-scale LLM.

Datasets	CORA		
Parameter scales	7B	14B	32B
PKD _{Qwen}	90.07	90.58	91.54
Datasets			
Parameter scales	7B	14B	32B
PKD _{Qwen}	85.96	86.64	87.16

References

- [1] Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. Graphformers: Gnn-nested transformers for representation learning on textual graph. In *Proceedings of the Advances in neural information processing systems*, volume 34, pages 28798–28810, 2021.
- [2] Jiarui Feng, Hao Liu, Lecheng Kong, Mingfang Zhu, Yixin Chen, and Muhan Zhang. Taglas: An atlas of text-attributed graph datasets in the era of large graph and language models. *arXiv preprint arXiv:2406.14683*, 2024.
- [3] Shujie Li, Yuxia Wu, Chuan Shi, and Yuan Fang. Hetgb: A comprehensive benchmark for heterophilic text-attributed graphs. *arXiv preprint arXiv:2503.04822*, 2025.
- [4] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [5] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. Graph transformer networks. In *Proceedings of the Advances in neural information processing systems*, volume 32, 2019.
- [6] Jason Zhu, Yanling Cui, Yuming Liu, Hao Sun, Xue Li, Markus Pelger, Tianqi Yang, Liangjie Zhang, Ruofei Zhang, and Huasha Zhao. Textgnn: Improving text encoder via graph neural network in sponsored search. In *Proceedings of the Web Conference 2021*, pages 2848–2857, 2021.
- [7] Xiangting Shi, Yakang Zhang, Abinash Pujahari, and Sambit Kumar Mishra. When latent features meet side information: A preference relation based graph neural network for collaborative filtering. *Expert Systems with Applications*, 260:125423, 2025.
- [8] Xiaofeng Cao, Mingwei Xu, Xin Yu, Jiangchao Yao, Wei Ye, Shengjun Huang, Minling Zhang, Ivor W Tsang, Yew Soon Ong, James T Kwok, et al. Analytical survey of learning with low-resource data: From analysis to investigation. *arXiv preprint arXiv:2510.08962*, 2025.
- [9] Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61, 2024.
- [10] Duo Wang, Yuan Zuo, Fengzhi Li, and Junjie Wu. Llms as zero-shot graph learners: Alignment of gnn representations with llm token embeddings. In *the Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pages 5950–5973, 2024.
- [11] Yuxia Wu, Shujie Li, Yuan Fang, and Chuan Shi. Exploring the potential of large language models for heterophilic graphs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5198–5211, 2025.
- [12] Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Harnessing explanations: LLM-to-LM interpreter for enhanced text-attributed graph representation learning. In *The Twelfth International Conference on Learning Representations*, 2024.
- [13] G Hinton. Distilling the knowledge in a neural network. In *Deep Learning and Representation Learning Workshop in Conjunction with NIPS*, 2014.
- [14] Zhikai Chen, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, Hui Liu, and Jiliang Tang. Label-free node classification on graphs with large language models (llms). In *The Twelfth International Conference on Learning Representations*, 2024.
- [15] Shengxiang Hu, Guobing Zou, Song Yang, Shiyi Lin, Yanglan Gan, Bofeng Zhang, and Yixin Chen. Large language model meets graph neural network in knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17295–17304, 2025.

[16] Bo Pan, Zheng Zhang, Yifei Zhang, Yuntong Hu, and Liang Zhao. Distilling large language models for text-attributed graph learning. In *the Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1836–1845, 2024.

[17] Renée G Fox. Training for uncertainty. In *The student-physician: Introductory studies in the sociology of medical education*, pages 207–242. Harvard University Press, 1957.

[18] Jing Liu, Tianai Yue, Chuanguang Yang, Yuqi Li, Qinfen Hao, Xiang Li, and Shiping Wen. Distilling multi-teacher knowledge from distinct graph neural networks. *Available at SSRN* 5084903, 2025.

[19] Kai Wang, Yu Liu, Qian Ma, and Quan Z Sheng. Mulde: Multi-teacher knowledge distillation for low-dimensional knowledge graph embeddings. In *the Proceedings of the Web Conference 2021*, pages 1716–1726, 2021.

[20] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24276–24285, 2023.

[21] Zhichun Guo, Chunhui Zhang, Yujie Fan, Yijun Tian, Chuxu Zhang, and Nitesh V Chawla. Boosting graph neural networks via adaptive knowledge distillation. In *the Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7793–7801, 2023.

[22] Bo-Wei Yang, Ming-Yi Chang, Chia-Hsun Lu, and Chih-Ya Shen. Two heads are better than one: Teaching mlps with multiple graph neural networks via knowledge distillation. In *the Proceedings of the International Conference on Database Systems for Advanced Applications*, pages 452–462, 2024.

[23] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[24] Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*, 2019.

[25] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *the Proceedings of the Advances in neural information processing systems*, volume 33, pages 7793–7804, 2020.

[26] Eli Chien, Jianhao Peng, Pan Li, and Olgica Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations*, 2021.

[27] Christian Koke and Daniel Cremers. Holonet: Spectral convolutions do extend to directed graphs. In *The Twelfth International Conference on Learning Representations*, 2024.

[28] Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael M Bronstein. Edge directionality improves learning on heterophilic graphs. In *the Proceedings of the Learning on Graphs Conference*, pages 25–1, 2024.

[29] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *the Proceedings of the International Conference on Machine Learning*, pages 1725–1735, 2020.

[30] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *the Proceedings of the International Conference on Machine Learning*, pages 9323–9332, 2021.

[31] Kaize Ding, Elnaz Nouri, Guoqing Zheng, Huan Liu, and Ryen White. Toward robust graph semi-supervised learning against extreme data scarcity. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[32] Zhixun Li, Dingshuo Chen, Tong Zhao, Daixin Wang, Hongrui Liu, Zhiqiang Zhang, Jun Zhou, and Jeffrey Xu Yu. Iceberg: Debiased self-training for class-imbalanced node classification. In *Proceedings of the ACM on Web Conference 2025*, pages 3160–3170, 2025.

[33] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[34] Huaxiu Yao, Chuxu Zhang, Ying Wei, Meng Jiang, Suhang Wang, Junzhou Huang, Nitesh Chawla, and Zhenhui Li. Graph few-shot learning via knowledge transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6656–6663, 2020.

[35] Lirong Wu, Haitao Lin, Yufei Huang, and Stan Z Li. Knowledge distillation improves graph structure augmentation for graph neural networks. In *the Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pages 11815–11827, 2022.

[36] Chunhai Zhang, Jie Liu, Kai Dang, and Wenzheng Zhang. Multi-scale distillation from multiple graph neural networks. In *the Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4337–4344, 2022.

[37] Yuchang Zhu, Jintang Li, Liang Chen, and Zibin Zheng. The devil is in the data: Learning fair graph neural networks via partial knowledge distillation. In *the Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1012–1021, 2024.

[38] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *the Proceedings of the Advances in neural information processing systems*, volume 31, 2018.

[39] Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? In *the Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2024.

[40] Dongyang Fan, Celestine Mendler-Dünner, and Martin Jaggi. Collaborative learning via prediction consensus. In *Proceedings of the Advances in neural information processing systems*, volume 36, pages 1988–2009, 2023.

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[42] Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *The Eleventh International Conference on Learning Representations*, 2023.

[43] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

[44] Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

[45] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pages 40–48, 2016.

[46] Jiayi Yang, Sourav Medya, and Wei Ye. Incorporating heterophily into graph neural networks for graph classification. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1544–1551, 2024.

[47] Langzhang Liang, Sunwoo Kim, Kijung Shin, Zenglin Xu, Shirui Pan, and Yuan Qi. Sign is not a remedy: Multiset-to-multiset message passing for learning on heterophilic graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 29621–29643, 2024.

[48] Yi Fang, Dongzhe Fan, Daochen Zha, and Qiaoyu Tan. Gaugilm: Improving graph contrastive learning for text-attributed graphs with large language models. In *the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 747–758, 2024.

[49] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[50] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[51] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[52] Leskovec Jure. Snap datasets: Stanford large network dataset collection. *Retrieved October 2025 from <http://snap.stanford.edu/data>*, 2014.

[53] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[54] Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

[55] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.

[56] Wenbo Shang, Xuliang Zhu, and Xin Huang. Path-lm: A shortest-path-based lm learning for unified graph representation. *arXiv preprint arXiv:2408.05456*, 2024.

[57] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[58] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *the Proceedings of the Advances in neural information processing systems*, volume 32, 2019.

[60] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

[61] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15:201–221, 1994.

[62] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Proceedings of the Advances in neural information processing systems*, volume 32, 2019.

[63] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

A Detailed Description of Datasets

Table 8: Statistics of datasets. The **Hom. ratio** means 1-hop homophily ratio.

Dataset	CORNELL	WASHINGTON	TEXAS	WISCONSIN	AMAZON RATINGS	OGBN-ARXIV	WIKI CS	PUBMED	CORA
Hom. ratio	0.1504	0.1545	0.1989	0.2109	0.4777	0.6542	0.6588	0.7924	0.8252
# Node	189	214	168	264	5068	169343	11701	19717	2708
# Edge	166	182	91	388	17334	1166243	216123	88648	10556
# Features	1703	1703	1703	1703	300	128	300	500	1433
# Classes	5	5	5	5	5	40	10	3	7
Domain	Web page	Web page	Web page	Web page	Co-purchase	Co-citation	Wikipedia page	Co-citation	Co-citation

CORNELL, WASHINGTON, TEXAS, and WISCONSIN:

These four datasets are derived from the WEBKB webpage dataset, collected from the computer science departments of various universities. In these datasets, nodes represent web pages, while edges denote hyperlinks connecting them. All words from the given web pages are collected as the features for the nodes. The webpage categories can be listed as following: Student, Project, Course, Staff, Faculty.

AMAZON RATINGS:

This dataset is derived from the AMAZON product co-purchasing network metadata, sourced from the SNAP datasets [52]. Nodes represent products (Books, Music CDs, DVDs, Videos) and edges signify relationships between products that are frequently co-purchased. The task involves predicting the average rating assigned to each product by reviewers. The possible rating values are grouped into five distinct classes. For node features, we utilize the NV-Embed-v2 [53] embeddings generated from the product descriptions. To reduce the size of the graph, we only consider the largest connected component of the 5-core of the graph.

WIKI CS:

WIKI CS is a graph derived from the Wikipedia platform. The nodes in WIKI CS represent Wikipedia page descriptions, while the edges correspond to hyperlinks between distinct pages. The WIKI CS dataset and its raw text [44] are sourced from OFA [54]. The graph consists of 11,701 nodes and 216,123 edges. The WIKI CS dataset is suitable for node classification tasks. The WIKI CS dataset is categorized into 10 distinct categories: Computational Linguistics, Databases, Operating Systems, Computer Architecture, Computer Security, Internet Protocols, Computer File Systems, Distributed Computing Architecture, Web Technology, Programming Language Topics.

CORA, PUBMED, and OGBN-ARXIV:

The CORA dataset represents a co-citation graph of computer science research papers. The dataset is sourced from OFA [54], with the original data derived from [9]. In [9], the authors recollect the dataset due to the commonly employed bag-of-words features in the widely used CORA dataset within the GNN community, where raw text is difficult to retrieve. The revised CORA dataset contains 2,708 nodes and 10,556 edges, matching the specifications of the original dataset. The dataset is divided into 7 categories: Theory, Reinforcement Learning, Genetic Algorithms, Neural Networks, Probabilistic Methods, Case-Based, Rule Learning.

The PUBMED dataset represents a co-citation graph of biomedical research papers focused on diabetes mellitus. The source and processing procedure of PUBMED are identical to those of CORA. After processing, the dataset consists of 19,717 nodes and 88,648 edges. The dataset is classified into 3 categories: Experimentally Induced Diabetes, Type 1 Diabetes, Type 2 Diabetes.

The OGBN-ARXIV dataset is a citation graph of papers from the arXiv platform. It is collected from the Arxiv dataset and its raw text as OGB[43] and OFA [54]. There are 169,343 nodes and 1,166,243 edges in the graph. It contains 40 sub-categories of compute science.

B Detailed Prompts

We provide all specific prompt templates in the following for zero-shot node classification, Node-preference-driven GNN Selector and GTA Prompts, respectively.

B.1 Prompts for Zero-shot Node Classification

The complete prompts for zero-shot node classification are provided as below. Similarly, for each dataset, we refine specific descriptions to ensure contextual coherence.

Table 9: The prompt template for zero-shot node classification.

Role	Prompt
System Prompt	<p>Papers in this field can be divided into 7 categories: [Case Based, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning, Theory]. You will serve as an assistant to help me to classify this target paper into the 7 categories above according to its description and related papers' descriptions, who may be of the same category as this target paper. I will provide you with the descriptions of this target paper and its related papers.</p> <p>Here are the instructions:</p> <p>I will provide you with information in the form of a JSON string that describes the target paper:</p> <p>Title: the title of this target paper. Abstract: the abstract of this target paper.</p> <p>Related Title: the title of the related paper. Related Abstract: the abstract of the related paper.</p> <p>Related Title: the title of the related paper. Related Abstract: the abstract of the related paper.</p> <p>.....</p> <p>Requirements:</p> <p>① Please provide your response in JSON format, following this structure:</p> <p>Reasoning: Briefly explain your reasoning process for the predicted category.</p> <p>Category: The best category you predict for this paper, this category must belong to these 7 categories: [Case Based, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning, Theory];</p> <p>② There are 2000 words limits for the reasoning;</p> <p>③ Do not provide any other text outside the JSON string;</p> <p>④ Focus only on content in the actual text and avoid making false associations;</p> <p>⑤ The output can only contain category and reasoning.</p>
User Prompt	<p>Title: t_{title}. Abstract: $t_{abstract}$.</p> <p>Related Title: t_{title}^{r1}. Related Abstract: $t_{abstract}^{r1}$.</p> <p>Related Title: t_{title}^{r2}. Related Abstract: $t_{abstract}^{r2}$.</p> <p>Related Title: t_{title}^{r3}. Related Abstract: $t_{abstract}^{r3}$.</p> <p>...</p>

B.2 Prompts for Node-preference-driven GNN Selector

Unlike the prompts used zero-shot node classification described above, we do not collect responses from the LLM; instead, we focus solely on the outputs generated by the subsequent projector. Similarly, for each dataset, we refine certain descriptions to maintain contextual consistency.

Table 10: The prompt template for Node-preference-driven GNN Selector.

Role	Prompt
System Prompt	<p>There are four names of teacher networks: [APPNP, GCN, H₂GCN, GAT]. We need to perform knowledge distillation for each node in this graph consist of nodes (papers) and edges (citation relationships). You will serve as an assistant to help me to assign the best teacher network for the target node (paper) based on the following information. I will provide you with three kinds of attributes of the target node (paper).</p> <p>Here are the instructions:</p> <p>I will provide you with information in the form of a JSON string that describes the node (paper):</p> <p>Semantic attributes: the title and abstract of this paper.</p> <p>Structure attributes: four teacher networks' logit output of this target node.</p> <p>Prediction attributes: important neighbors (papers), which are closely related the target node (paper) and their contents.</p> <p>Requirements:</p> <ul style="list-style-type: none"> ❶ Please provide your response in JSON format, following this structure: Reasoning: Briefly explain your reasoning process for the selected teacher network. Teacher network: The best teacher network you assign for this node (paper), this result must belong to these 4 teachers: [APPNP, GCN, H₂GCN, GAT]; ❷ There are 2000 words limits for the reasoning; ❸ Do not provide any other text outside the JSON string; ❹ Focus only on content in the actual text and avoid making false associations; ❺ The output can only contain teacher network and reasoning.
User Prompt	<p>Semantic attributes: It is the content description of this target paper: t.</p> <p>Structure attributes: It has following important neighbors (papers), which are closely related the target paper. Their content descriptions are: ...</p> <p>Prediction attributes:</p> <p>The APPNP's logits output of this target paper is $\text{str}(\mathbf{z}_{APPNP})$, The GCN's logits output of this target paper is $\text{str}(\mathbf{z}_{GCN})$, The H₂GCN's logits output of this target paper is $\text{str}(\mathbf{z}_{H_2GCN})$, The GAT's logits output of this target paper is $\text{str}(\mathbf{z}_{GAT})$</p> <p>...</p>

B.3 Graph Topology Aware (GTA) Prompts

Generating effective prompts for graph-based tasks can be challenging for LLMs, due to the inherent complexity of graph structures and relationships that must be accurately represented. To address this challenge, we propose structured-tasks text for graph topology aware, designed specifically for fine-tuning LLMs.

TASK 1: Connectivity This task is determining whether or not two nodes v_i and v_j in an undirected graph are connected. Specifically, we randomly select node pairs $v_i, v_j \in \mathcal{V}$ and ask whether or not an edge exists between them in the graph, answering with a "True/False" response. To ensure prompt diversity, only one-third of the possible node pairs are selected for each graph.

TASK 2: Degree The degree of a node, D , is the number of nodes directly connected to it. In this task, we group nodes based on their degree and select a node v_i from a group. The LLM is then given the node's local structure according to the adjacency matrix \mathbf{A} , and is asked for the degree of the node. To prevent repetitive prompts, only one-third of the nodes from each degree group are selected.

TASK 3: Cycle Detection A cycle in an undirected graph without self-loop is a path where the first and last nodes are the same. This task requires the LLM to answer whether a cycle exists in the given sequence of nodes, $\{v_1, \dots, v_l, \dots, v_1\}$. We generate random walks [55] of length greater than

10 and arrange them into node sequences. After describing their neighbors information (derived from the adjacency matrix \mathbf{A}), the LLM is then asked whether or not any sequence of nodes forms a cycle.

TASK 4: Text Generation We randomly select a node set $\mathcal{W} = \{v_i\}_{i=1}^{N/3}$ as the source nodes, and a breadth-first search (BFS) is conducted from each source node to identify nodes in graph at a distance greater than t edges from v_i , which are collected as target nodes. Redundant nodes are removed via the long-to-short path conversion module [56]. The LLM is tasked with generating textual descriptions of target nodes based on the semantic attributes of the preceding nodes in the path.

Specifically, **TASK 1** enhances the LLM’s ability to identify neighboring nodes and understand the structure of local neighborhoods; **TASK 2** strengthens the LLM’s ability to recognize the significance of node degrees within the graph context; **TASK 3** reinforces the LLM to reason about complex graph topologies, such as cycles and long-range node dependencies; **TASK 4** improves path-based reasoning and contextualization of nodes in the local graph structure.

Table 11: The prompt template for the **TASK 1: Connectivity**.

Role	Prompt
System Prompt	<p>You will serve as a graph machine learning expert in connectivity detection to help me to determine whether the edge exists between the given two targeted nodes. There is a undirected graph consisting of papers (nodes) and the citation relationships (edges) between them. I will provide the information of the two targeted nodes and their neighbors, consisting of indexes, textual content.</p> <p>Here are the instructions:</p> <p>I will provide you with information in the form of a JSON string that describes the target papers:</p> <p>The first targeted paper:</p> <p>Node index: ...; Title: ...; Abstract: ...;</p> <p>The k_{th} neighbor: Index:...; Title: ...; Abstract: ...;</p> <p>...</p> <p>The second targeted paper:</p> <p>Node index: ...; Title: ...; Abstract: ...;</p> <p>The k_{th} neighbor: Index:...; Title: ...; Abstract: ...;</p> <p>...</p> <p>Requirements:</p> <p>① Please provide your response in JSON format, following this structure:</p> <p>Reasoning: Briefly explain your reasoning process for the selected teacher network.</p> <p>Answer: You only can select one from [True, False] as the best answer;</p> <p>② There are 2000 words limits for the reasoning;</p> <p>③ Do not provide any other text outside the JSON string;</p> <p>④ Focus only on content in the actual text and avoid making false associations;</p> <p>⑤ The output can only contain answer and reasoning.</p>
User Prompt	<p>The first targeted paper:</p> <p>Node index: i; Title: \mathbf{t}_{title}^i; Abstract: $\mathbf{t}_{abstract}^i$</p> <p>The k_{th} neighbor’s node index: I_k^i Title: $\mathbf{t}_{title}^{I_k^i}$ Abstract: $\mathbf{t}_{abstract}^{I_k^i}$...</p> <p>...</p> <p>The second targeted paper:</p> <p>Node index: j; Title: \mathbf{t}_{title}^j; Abstract: $\mathbf{t}_{abstract}^j$</p> <p>The k_{th} neighbor’s node index: I_k^j Title: $\mathbf{t}_{title}^{I_k^j}$ Abstract: $\mathbf{t}_{abstract}^{I_k^j}$...</p> <p>...</p>

The full prompts for **Connectivity** is presented above. When generating prompts for different datasets, we adjust certain descriptions to better align with the specific context. For example, when constructing prompts for **TEXAS**, the background description should be adapted to reflect web pages, and the relationship should be revised to hyperlinks, along with other context-specific adjustments.

Similarly, for each task, the prompts must also be modified to correspond to the specific content described in Section 3.2.

C Implementation Details and Time Complexity Analysis

Algorithm 1: The training of PKD.

Input: $\mathcal{G}_T = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{A}, \mathbf{T})$, training dataset with true labels \mathcal{D}_L , teacher GNNs $\{T_b\}_{b=1}^4$ with parameters $\{f_{T_b}^\theta\}_{b=1}^4$, student GNN S with parameter f_S^θ , fine-tuned LLM LLM^θ , Policy Model f_A^θ , Value Model f_V^ϕ , epoch number of RL L_1

Output: The expanded training dataset $\tilde{\mathcal{D}}_L$, optimized parameters $LLM^{\theta^*}, f_S^{\theta^*}, f_A^{\theta^*}, f_V^{\phi^*}$ and predicted labels \tilde{y} .

- 1 $\tilde{\mathcal{D}}_L \leftarrow \mathcal{D}_L$;
- 2 $LLM^{\theta^*} \leftarrow LLM^\theta$;
- 3 Filter out GNN-preference nodes based on the preference rank $\mathcal{V}_{\mathcal{P}\mathcal{R}}$ and get their annotations from LLM^{θ^*} ;
- 4 Conduct prediction distillation from LLM^{θ^*} to $\{f_{T_b}^\theta\}_{b=1}^4$ for retrain them ;
- 5 **for** $l_1 \leftarrow 1$ **to** L_1 **do**
 - 6 Shuffle $\tilde{\mathcal{D}}_L$ to get a new training sequence;
 - 7 Complete prompts $\{\mathcal{P}_i\}_{i=1}^W$ for each selected nodes;
 - 8 **for** each node $v_{\mathcal{P}\mathcal{R}} \in \tilde{\mathcal{D}}_L$ **do**
 - 9 NSG select teacher GNN for $v_{\mathcal{P}\mathcal{R}}$ and get one-hot vector \mathbf{m}_i ;
 - 10 Update the parameter f_S^θ and get reward R_i by Eqn. (1);
 - 11 Store $(\mathcal{P}_i, \mathbf{m}_i, R_i)$ to the episode history \mathcal{F} ;
 - 12 Update the parameter f_A^θ and f_V^ϕ by Eqn. (5) and Eqn. (6) ;
- 13 **return** $\tilde{\mathcal{D}}_L, LLM^{\theta^*}, f_S^{\theta^*}, f_A^{\theta^*}, f_V^{\phi^*}$;

First of all, we outline the training setup employed for the experiments detailed in Section 4.2. Uniform training hyper-parameters are applied across all baseline models and datasets. Specifically, the following hyper-parameter values are utilized: the hidden dimension is set to 128. We use ReLU activation functions in all our baseline models. The Adam optimizer is utilized with a learning rate of $1 \times e^{-2}$ and weight decay of $5 \times e^{-4}$. We train each baseline for 600 steps and select the best step based on the validation accuracy. In our proposed method, we train the student 5 epochs after GNN selection driven by node attributes every time and train the agent 200 epochs. The other weight hyper-parameters are set as follows: $\alpha = 0.5, \beta = 1, \gamma = 0.1, \eta = 0.3, c_1 = 0.5, c_2 = 0.01, \epsilon = 0.2$.

Additionally, the parameters of Action Model and Value Model are updated as follows:

$$f_A^\theta \leftarrow f_A^\theta - \rho_A \nabla_{f_A^\theta} (\mathcal{L}_A + c_1 \mathcal{L}_V - c_2 H(\pi_T)) \quad (5)$$

$$f_V^\phi \leftarrow f_V^\phi - \rho_V \nabla_{f_V^\phi} \mathcal{L}_V \quad (6)$$

where f_A^θ and f_V^ϕ represent the trainable parameters of the Policy Model and Value Model, respectively. ρ_A and ρ_V are their learning rates and $\nabla_{f_A^\theta}$ and $\nabla_{f_V^\phi}$ are the gradients of their parameters. \mathcal{L}_A and \mathcal{L}_V are objective functions belonging to the Policy Model and Value Model, respectively. c_1, c_2 are hyper-parameters to balance weights. $H(\pi_T)$ is employed to enhance the entropy of the policy and promote sufficient exploration. Based on the CLIP strategy [41], the final objective function of the Policy Model is:

$$\mathcal{L}_A = -\mathbb{E}_i [\min(r_i(f_A^\theta) \hat{A}_i, \text{clip}(r_i(f_A^\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_i)] \quad (7)$$

where \mathbb{E}_i represents the expectation in the time step i . $r_i(f_A^\theta)$ is the ratio of the i -th policy to the $(i-1)$ -th policy. \hat{A}_i is the advantage estimation in the current step, denoting how good or bad the Action is. ϵ is a hyper-parameter, which determines the range of the CLIP operation.

The objective functions of the Value Model and $H(\pi_T)$ are:

$$\mathcal{L}_V = \mathbb{E}_i[(f_V^\phi(\mathcal{P}_i) - \hat{R}_i)^2] \quad (8)$$

$$H[\pi_T] = -\mathbb{E}_i[\pi_{f_A^\theta}(A_T|\mathcal{P}_i) \log \pi_{f_A^\theta}(A_T|\mathcal{P}_i)] \quad (9)$$

where $f_V^\phi(\mathcal{P}_i)$ and \hat{R}_i denote the Value Model's estimation of State \mathcal{P}_i and the target value of real Reward R_i , respectively. A_T denotes the specific action and $\pi_{f_A^\theta}(A_T|\mathcal{P}_i)$ is the probability that Policy f_A^θ takes action A_T in state \mathcal{P}_i .

The detailed training procedure is shown in **Algorithm 1**.

The specific analysis of the time complexity of PKD training and testing are provided below:

The time complexity of PKD training is mainly divided into three parts: LLM fine-tuning (Line 2), GNN-preference-driven Node Selector (Line 3-4) and Node-preference-driven GNN Selector (Line 5-12). The GNN-preference-driven Node Selector also can be divided into the annotations generation and prediction distillation.

Table 12: The GTA fine-tuning configurations on Llama-3.1-8B-Instruct.

Model Name	Dataset Size	Epoch	lora_r	lora_alpha	Optimizer	Learning Rate	Time Cost
Llama-3.1-8B-Instruct	53,617	2	4	4	AdamW [57]	$1e^{-4}$	9h 41m 48s

First, we use Low-Rank Adaptation (LoRA) strategy [58] for efficient parameter training, with hyperparameters set to $r = 4$, $\alpha = 4$, $epoch = 2$ (as shown in Table 12), and the rest are set according to the default settings of llama-factory². Weight merge is also involved. In general, the time complexity of this part is $\mathcal{O}(nLdr + Ld^2r)$, where n is the number of instructions, L is the number of layers applying Lora, and d is the dimension of the LLM hidden layer. $r \ll d$, so the time complexity is bound by $\mathcal{O}(NLd + Ld^2)$.

Table 13: The time costs on **CORA** and **OGBN-ARXIV** of annotations generation by Llama-3.1-8B-Instruct.

Dataset	CORA	OGBN-ARXIV
Time / GPU-hours	0.11	1.68

The process of annotations generation includes sorting the selected nodes and the reasoning process of LLM, and its time complexity is $\mathcal{O}(W \log W)$ and $\mathcal{O}(WL'(l^2d + ld^2))$, where W is the number of selected nodes, L' is the number of transformer layers in LLM, and l is the input sequence length. Generally, $W \ll l$, $L' \ll l$, then the time complexity is $\mathcal{O}(l^2d + ld^2)$.

Table 14: The time costs of retraining teacher GNNs on **CORA** and **OGBN-ARXIV**.

Datasets	Teacher GNNs (T_1, T_2, T_3, T_4)	Total running time (seconds)
CORA	GCN, GAT, APPNP, H2GCN	2.4716
OGBN-ARXIV	GCN, GAT, APPNP, H2GCN	18.2017

The time complexity of teacher GNN (2-layers) re-training is bound by $\mathcal{O}((NF + M)D)$, where N is the number of nodes, F is the node feature dimension, M is the number of edges, and D is the GNN hidden layer dimension.

The time complexity of Node-preference-driven GNN Selector is $\mathcal{O}(W(l^2d + ld^2 + dd' + d'a))$, where d' is the dimension of the MLP hidden layer, a is the number of action categories, $W \ll l$, $a \ll d$, so the time complexity is bound by $\mathcal{O}(l^2d + ld^2 + dd')$. The training time complexity of student GNN is $\mathcal{O}((NF + M)D)$. Therefore, the overall time complexity is bound by $\mathcal{O}((L + 2l)d^2 + (d' + nL)d + 2l^2 + 2D(NF + M))$.

²<https://github.com/hiyouga/LLaMA-Factory>

Table 15: Peak memories and running times of Node-preference-driven GNN Selector on **CORA** and **OBGN-ARIXV**. "m" and "s" denote minute and second.

Dataset	Peak memory	Running time / epoch
CORA	454.62 MB	7.9s
OBGN-ARIXV	1655.18 MB	44m 8.3s

The inference time complexity of PKD is determined by the testing process of the student GNN. So its time complexity is bound by $\mathcal{O}((NF + M)D)$.

Specifically, We implement our proposed PKD with PyTorch (2.5.1) [59], PyTorch Geometric (2.6.1) [60], Python (3.10.16), Transformers (4.50.3), and vllm (0.7.0). We conduct all experiments on the NVIDIA A800-SXM4-80GB GPU and Intel(R) Xeon(R) CPU Max 9468.

The time costs for GTA fine-tuning, the Distance-based Neighbor Selector, LLM annotation, teacher re-training, and the PPO loop on Cora and Ogbn-Arixv are presented in Tables 12, 13, 14 and 15, respectively. The peak memories of performing PKD on the Cora and Ogbn-Arixv are also listed in Table 15.

D Proofs for Propositions 3.1

The uncertainty usually refers to a measure of the confidence of a model in predicting a certain sample. From the perspective of collective consensus [40], we define the K -uncertainty of one node as the deviation of each teacher GNN's prediction probability distribution from the overall prediction probability distribution. From the **Proposition 3.1**, we can get that, the larger δ_K of one node, the stronger the uncertainty of this node, which is more beneficial to teacher GNNs training.

Proof. For each node v , the prediction probability distributions of B teacher GNNs can be denoted by P_1, P_2, \dots, P_B . The K -uncertainty of node v is defined as:

$$\delta_K(v) \triangleq \sum_{1 \leq i < j \leq B} [D_{KL}(P_i(v) \parallel P_j(v)) + D_{KL}(P_j(v) \parallel P_i(v))] \quad (10)$$

Here, we define the average prediction probability distribution as following:

Definition D.1. The average prediction probability distribution \mathcal{M} is the benchmark for the overall prediction probability distribution to measure both the models confidence and the consistency of each GNN with the overall probability distribution.

$$\mathcal{M}(v) = \frac{1}{B} \sum_{i=1}^B P_i(v) \quad (11)$$

Then, the uncertainty of node v is,

$$\delta_v = \frac{1}{B} \sum_{i=1}^B D_{KL}(P_i(v) \parallel \mathcal{M}(v)) \quad (12)$$

According to Jenson's inequality, we have

$$\delta_K(v) \geq N\delta_v \quad (13)$$

For any probability distribution P , we have

$$D_{KL}(P \parallel \mathcal{M}) = H(P, \mathcal{M}) - H(P) \quad (14)$$

Then,

$$\delta_K = \frac{1}{N} \sum [H(P_i, \mathcal{M}) - H(P_i)] \quad (15)$$

As the K -uncertainty increases, the entropy of the GNN's prediction probability distribution P_i increases, and the cross-entropy $H(P_i, \mathcal{M})$ grows significantly due to the larger probability distribution differences. So there is,

$$\delta_K(v) \propto \sum_{i=1}^B [H(P_i, \mathcal{M}) - H(P_i)] \propto \delta_v \quad (16)$$

That is,

$$\delta_K(v) \propto \delta_v \quad (17)$$

From the **Proposition 3.1**, we also can get that, selecting high-uncertainty nodes to expand the training set benefits GNNs training.

For a GNN with prediction probabilities $P(y = c|v)$, the entropy of an unlabeled node v is

$$H(v) = - \sum_{c=1}^C P(y = c|v) \log P(y = c|v) \quad (18)$$

To maximize information gain, we select the node v^* with the highest uncertainty (entropy):

$$v^* = \arg \max_v H(v) \quad (19)$$

After expanding v^* to the training dataset, the loss function becomes:

$$\mathcal{L}_{new} = \mathcal{L}_{old}(\theta) + \mathcal{L}(f_\theta(v^*), y^*) \quad (20)$$

Here, y^* is considered the true label based on the fine-tuned LLM. The GNN parameters are updated as:

$$\theta_{new} = \theta_{old} - r \cdot \nabla_\theta \mathcal{L}(f_\theta(v^*), y^*) \quad (21)$$

Since the prediction probability distribution of v^* is close to uniform (due to high entropy) [61], the gradient more effectively corrects the GNN parameters [62], reducing the error. According to the preference rank: $\mathcal{V}_{\mathcal{PR}} = \text{Sort}(\{v_1, \dots, v_N\}, \delta_K(v_1), \delta_K(v_2), \dots, \delta_K(v_N))$, we can get the follows:

$$f_T^{\theta^*}(\tilde{\mathcal{D}}_L) = \arg \min_{v_i \in \{v_{\mathcal{PR}}^1, v_{\mathcal{PR}}^2, \dots, v_{\mathcal{PR}}^W | \delta_K(v_{\mathcal{PR}}^W) > \tilde{\delta}_K\}} \frac{1}{W} \sum \mathcal{L}(f_T^\theta, v_i) \quad (22)$$

where $\tilde{\mathcal{D}}_L$ is the expanded training dataset. $f_T^{\theta^*}$ is the optimal parameter of teacher GNN. $v_{\mathcal{PR}}^w$ represents the w -th nodes in the preference rank.

□

E Other Experimental Results

E.1 Visualization

Figure 6 presents the outstanding node classification performance we mentioned in Section 4.2, which is illustrated by the t-SNE [63] visualization of the embedding spaces for CORA. Notably, Figure 6(a) illustrates the results of the student GNN (GCN) under the # LN 5 condition.

From the Figure 6, we can see that, some KD methods fail to enable the student GNN to learn discriminative node representations, as evidenced by the absence of clustered structures in the embedding space, exemplified by MSKD, BGNN, and MTAAM. GCNII and KDGA struggle to form well-separated clusters, whereas methods like LLMGNN, GAugLLM, and FairGKD yield clusters with limited purity. Compared to these baselines, our method generates embeddings with significantly enhanced inter-class separability and high cluster purity, resulting in improved few-shot node classification performance.

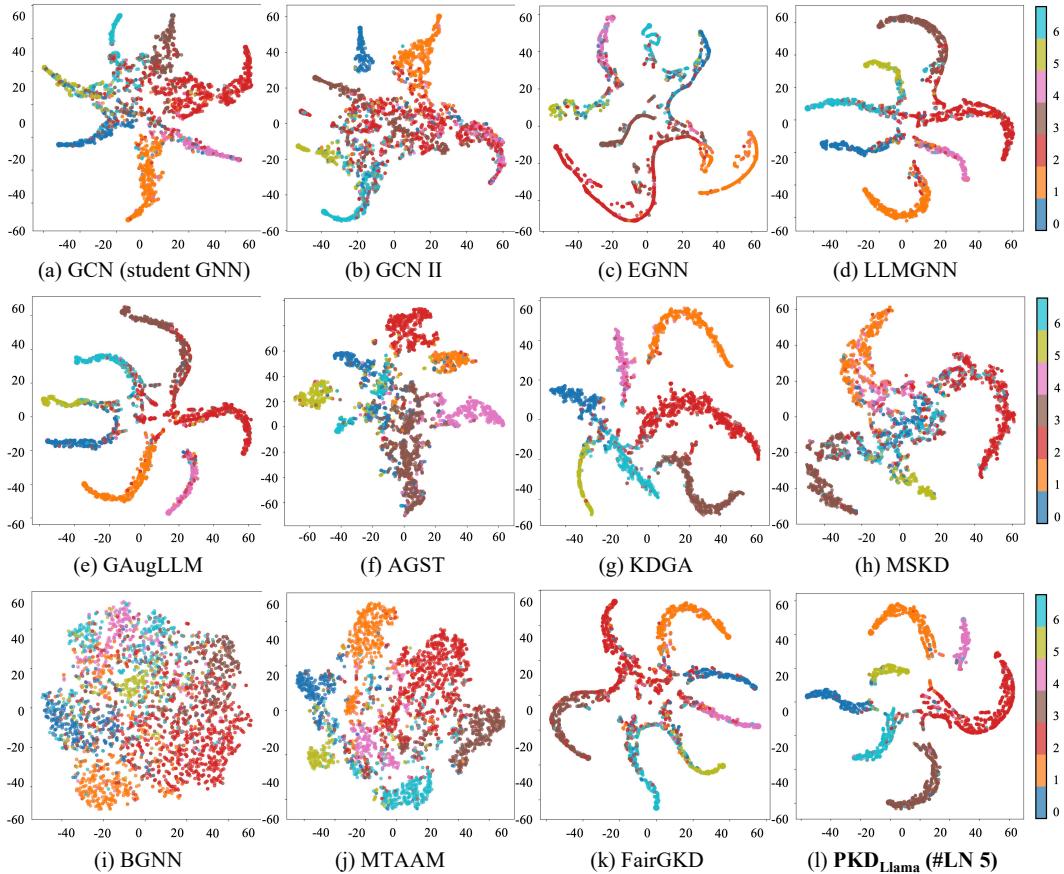


Figure 6: T-SNE [63] visualizations on CORA.

E.2 Ablation Study

The results on Cora and Amazon Ratings using our $\text{PKD}_{\text{Llama}}$ (RL-based method) and the other three teacher selection methods (Entropy-based ranking, i.e., selecting the teacher GNN with the highest prediction confidence, Random selection, and End-to-end learning) are shown in Table 16. It is obvious that the RL-based method significantly outperforms the other three methods.

Table 16: Performance comparison of Entropy-based ranking, Random selection, End-to-end learning and RL-based approach on **CORA** and **AMAZON RATINGS**.

Methods	CORA	AMAZON RATINGS
Entropy-based ranking	75.70	55.74
Random selection	62.80	63.05
End-to-end learning	60.29	60.39
PKD _{Llama} (RL-based)	90.27	65.93

E.3 Hyperparameters Sensitivity Analysis

As mentioned in Sec. 4.4, we perform the hyperparameter sensitivity analysis over the loss-weight coefficients $\alpha, \beta, \gamma, \eta$ on two datasets, and report the results in Tables 17, 18, 19 and 20. As result, the proposed PKD can achieve much better performance when $\alpha = 0.5, \beta = 1, \gamma = 0.1, \eta = 0.3$.

Table 17: The influence of α .

α	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	2
CORA	73.78	74.70	75.66	73.15	72.41	69.05	72.71	69.49	70.12
AMAZON RATINGS	61.81	62.86	63.83	62.76	61.30	60.28	60.47	61.66	61.74

Table 18: The influence of β .

β	0.25	0.5	1	2
CORA	64.10	67.17	75.36	74.15
AMAZON RATINGS	62.94	63.23	63.91	63.93

Table 19: The influence of γ .

γ	0.05	0.1	0.2	0.5	1	2
CORA	68.90	70.12	63.77	64.40	65.06	62.66
AMAZON RATINGS	62.58	64.01	63.71	63.63	63.93	63.93

Table 20: The influence of η .

η	0.1	0.3	0.5	0.7	0.9
CORA	89.96	90.27	88.43	86.67	89.42
AMAZON RATINGS	64.87	65.93	64.79	63.75	63.81

F Broader Impact

The proposed PKD offers significant broader impacts by enhancing few-shot node classification on TAGs. By combining the strengths of LLM and GNN, it improves learning efficiency, reducing the need for expensive and time-consuming manual annotation. This can benefit industries like social media, recommendation systems, and network analysis, enabling more accurate and scalable models for personalized services, fraud detection, and dynamic optimization.

Additionally, PKD can tailor message-passing mechanisms to node-specific attributes can lead to more adaptive and efficient machine learning models. It also democratizes access to advanced machine learning, allowing smaller organizations and researchers with limited resources to develop effective models. However, ethical considerations, such as privacy and fairness, must be prioritized to ensure responsible deployment.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: **[Yes]**

Justification: In this paper, we propose a preference-driven knowledge distillation (PKD) framework that unites LLMs and various-architectures GNNs for few-shot node classification on TAGs. We claim the contributions and scope in the abstract and introduction sections (See Abstract and Introduction Section).

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: **[Yes]**

Justification: In this work, we discuss the limitations of our research and outline directions for future work (See Conclusion).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: In this work, we provide the **Proposition 3.1** and its complete proof(See Method and Appendix. **D**).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the code necessary for replicating the studies described in this paper via an anonymous link, and we detail the experimental setup for the replication in the article itself (See Experiments and Appendix. **C**).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: For the datasets disclosed in the article, we have provided information regarding their sources and origins (See Appendix. A).

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: we have specified all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results (See Experiments and Appendix. C).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In this paper, we have reported the standard deviation of the experiments (See Experiments and Appendix. E).

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: In this paper, we provide detailed information about the experimental resources, including GPU configurations used in our studies and running time costs about all methods (See Experiments).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: The study presented in this paper conforms to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: We have provided the societal impacts of the work (See Appendix F).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper does not address issues related to this aspect.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All creators and original owners of the assets used in our paper, such as code, data, and models, have been properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The research presented in this paper is not concerned with new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve experiments or research related to human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not address potential risks incurred by study participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs is an important component of the core methods in this research and we has describe the usage in detail (See Method, Experiments and Appendix C).

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.