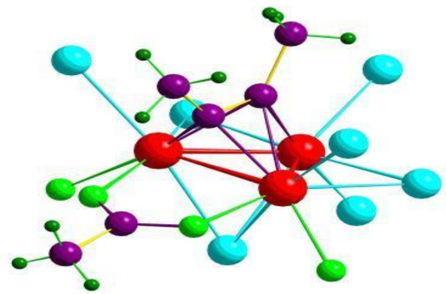
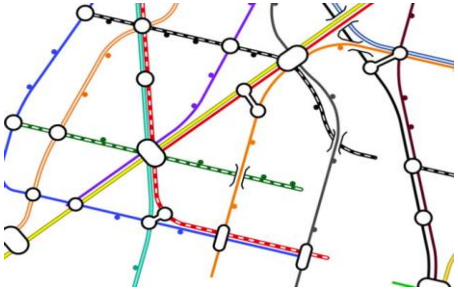
LABEL-FREE NODE CLASSIFICATION ON GRAPHS WITH LARGE LANGUAGE MODELS (LLMS)

Graph data are everywhere

Social Graphs

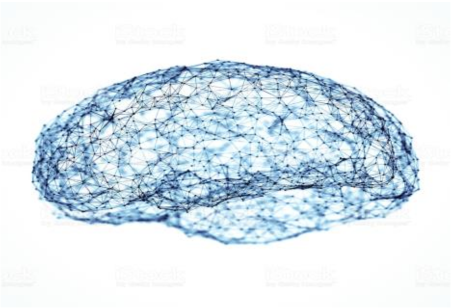


Molecular Graphs

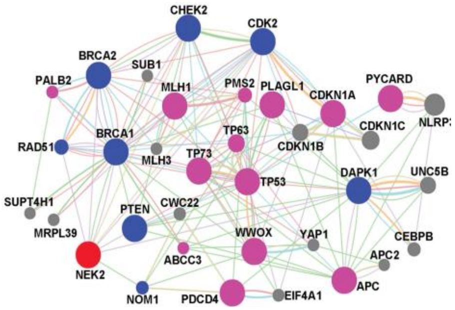


Transportation Graphs

Web Graphs

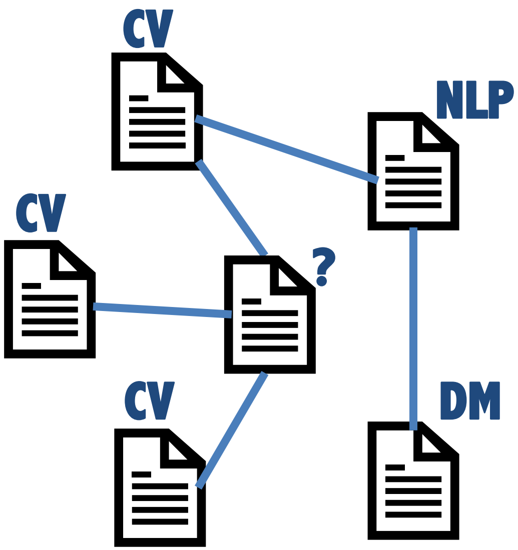


Brain Graphs

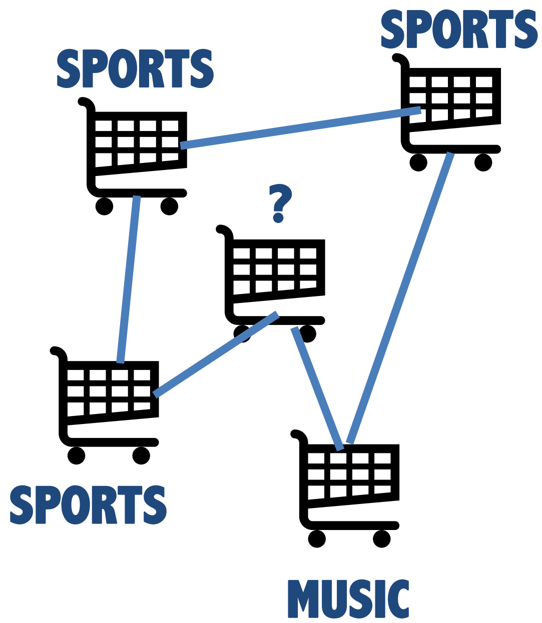


Gene Graphs

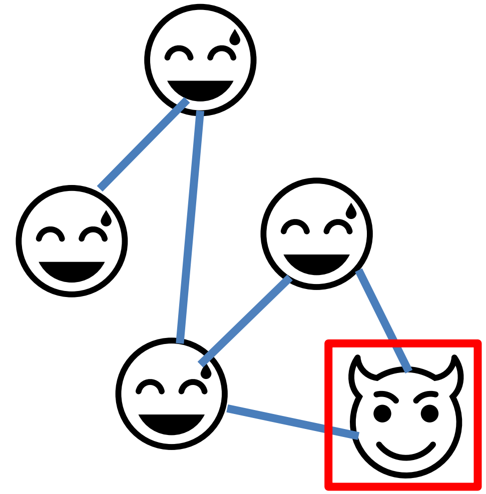
Node Classification is a crucial task for graph



Paper Categorization

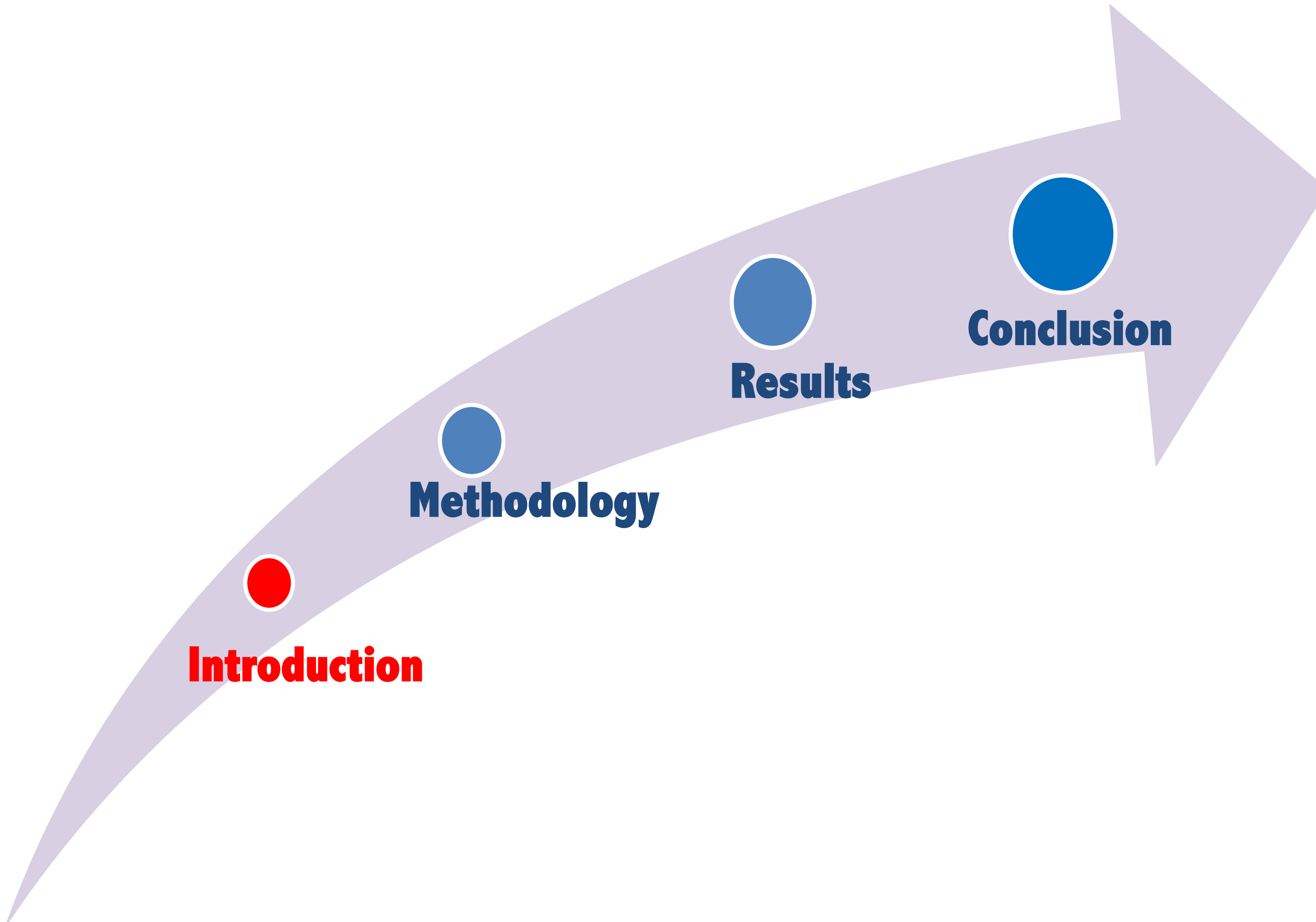


Product Classification

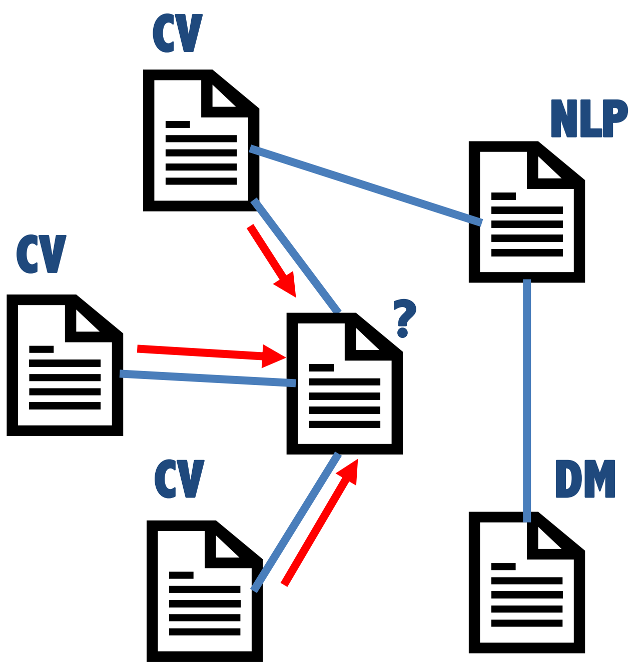


Fraud Detection

Semi-supervised node classification on graphs


Introduction
Methodology
Results
Conclusion

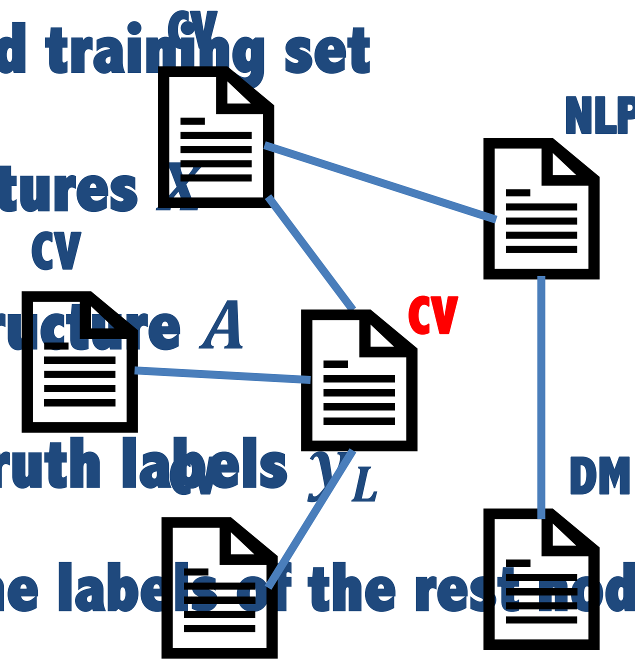
Semi-supervised node classification on graphs



Given a fixed training set

- Node features X
- Graph Structure A
- Ground truth labels y_L

➡ Predict the labels of the rest nodes



Graph neural networks work well for this task with abundant ground truth labels

Two assumptions

Actively select training nodes based on some strategies

A fixed training set

Studied by graph active learning

Overlook the data selection process

Ground tr  **bels**

y_L

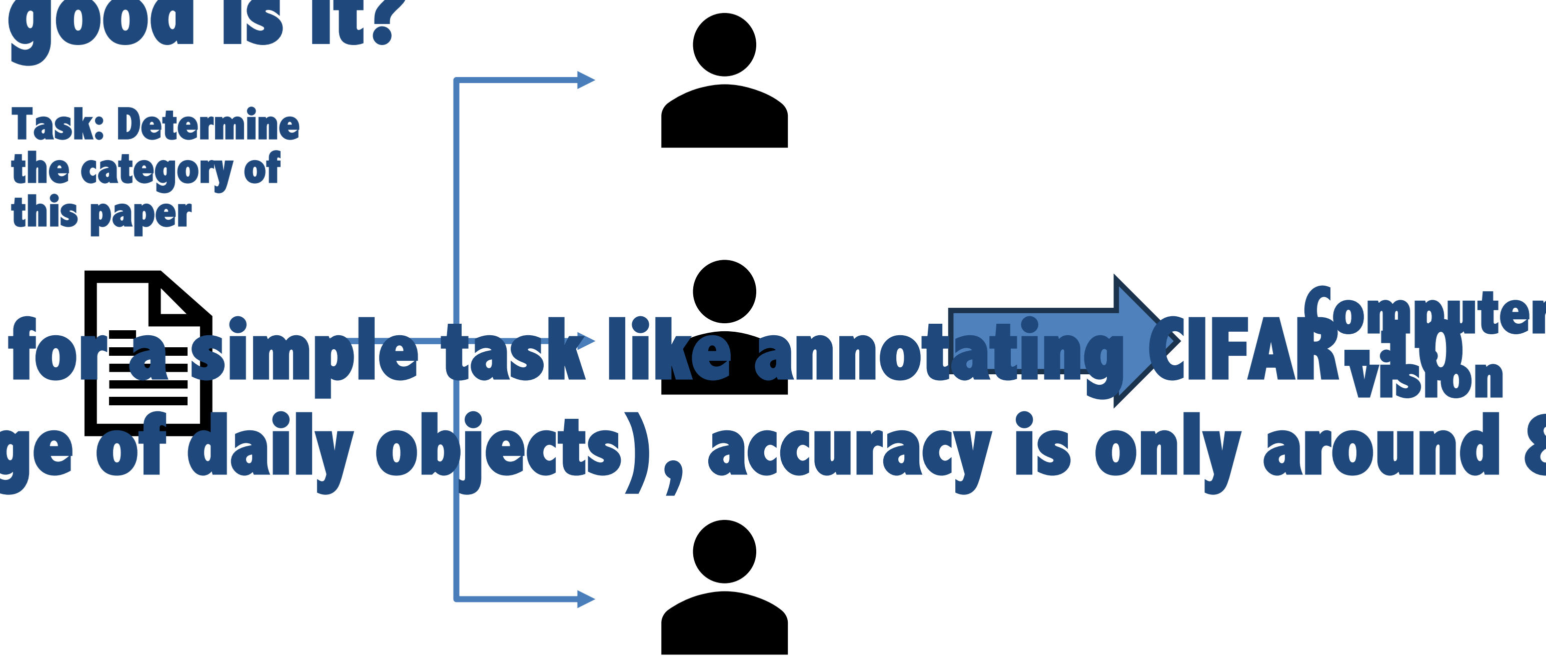
Overlook the intricacy of (graph) data annotation

How can we get high-quality annotations?

The old story: Human Annotation

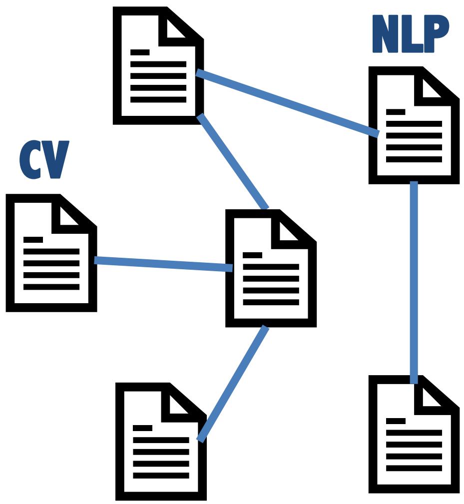
Crowdsourcing platform (like Amazon MTurk) is one of the most popular ways to do annotations

How good is it?



Even for a simple task like annotating CIFAR-10 (image of daily objects), accuracy is only around 80%

Annotating graph data is challenging



- Due to the non-IID nature of the graph, human annotations tend to be **biased and focus on a small group of nodes***

- Annotating some kinds of graph, like OGB-Arxiv (paper), requires related knowledge

- Annotating a massive scale graph, like million-scale OGB-Products, needs lots of time and money

* Zhu, Qi, et al. "Shift-robust gnns: Overcoming the limitations of localized graph training data." *Advances in Neural Information Processing Systems* 34 (2021): 27965-27977.

LLMs as annotators for graphs?

In recent literature*, LLMs present promising zero-shot performance on node classification tasks

Limitations

- Cannot utilize graph structure
- Performance gap to well-trained GNNs
- Expensive & slow for inference

Using LLMs as annotators for GNNs seems a plausible way to harness the strength of both GNNs and LLMs!

Chen, Zhikai, et al. "Exploring the potential of large language models (llms) in learning on graphs." *arXiv preprint arXiv:2307.03393* (2023).

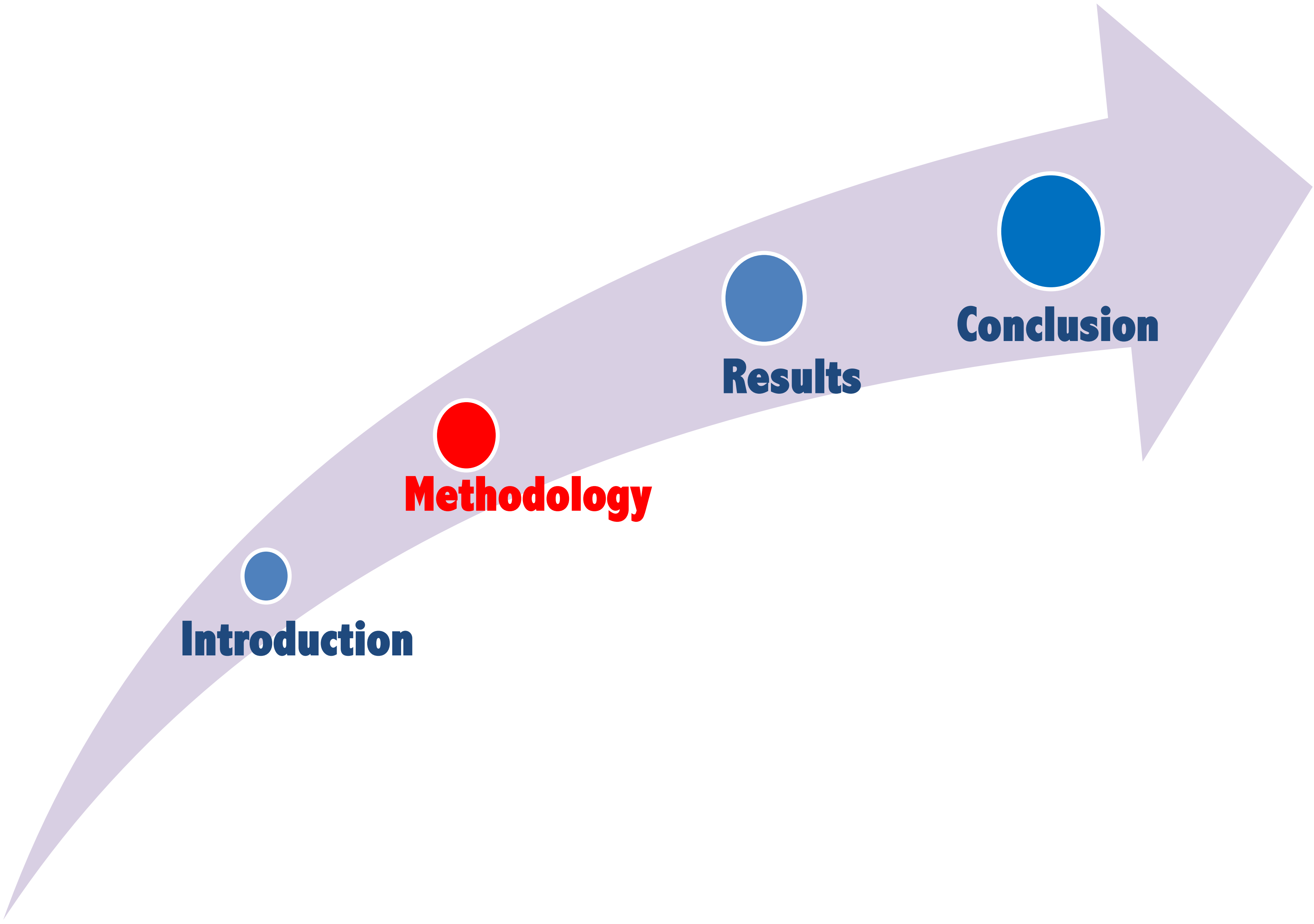
New challenges

Data Selection → LLMs → GNNs

How can we ensure

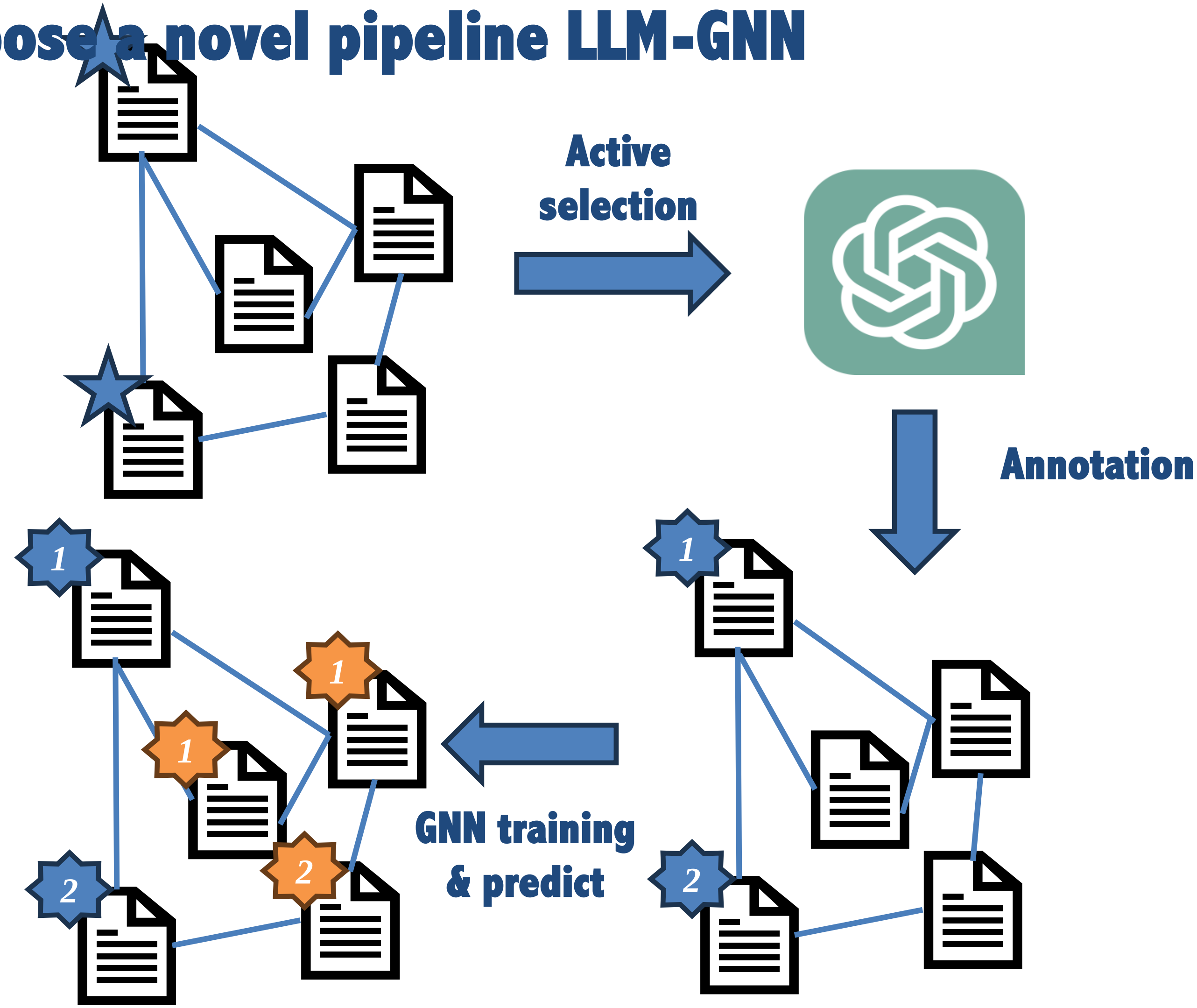
- Diversity & Representativeness
  - **Critical for performance in graph active learning**
- High Annotation quality

→ Optimize the trained GNN performance


Introduction
Methodology
Results
Conclusion

Label-free node classification on graphs with LLMs

We propose a novel pipeline LLM-GNN



Implementation

LLM-GNN supports flexible component design

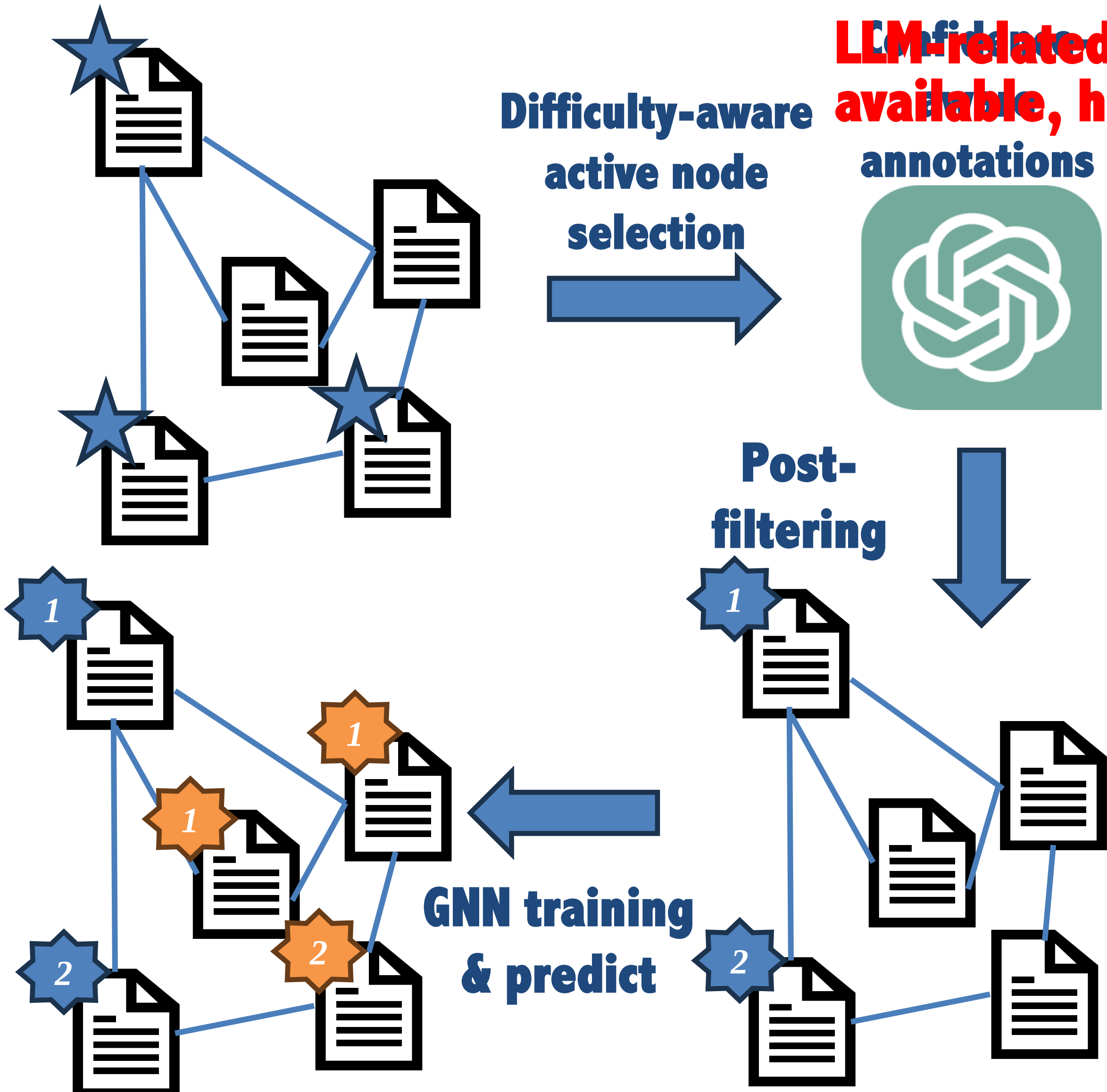
The key part is how to consider the following two factors simultaneously (we show one possible implementation)

- **Diversity & Representativeness** ➡ **Can be addressed by graph active learning**
- **Annotation quality** ➡ **We propose**

1. Difficulty-aware active selection

2. Confidence-aware prompt + Post filtering

Implementation



LLM-related information is not available, heuristic-based methods

Use LLMs' outputs to further optimize the selection

Difficulty-aware active node selection

In the selection stage, only feature and structure is available

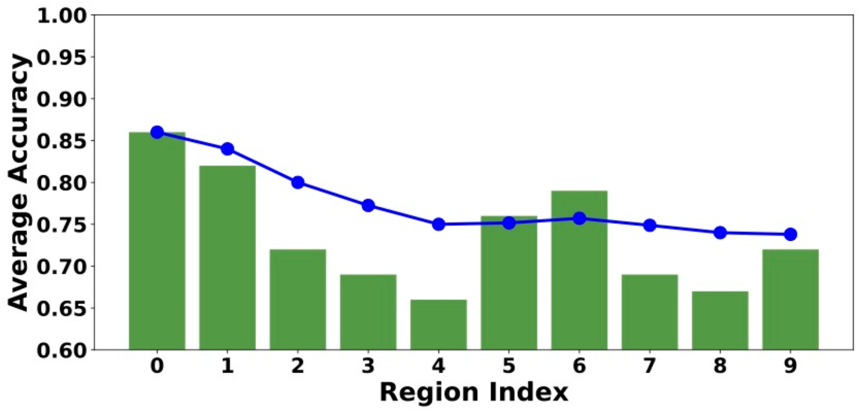
We induce the difficulty of annotation by the rule of thumb

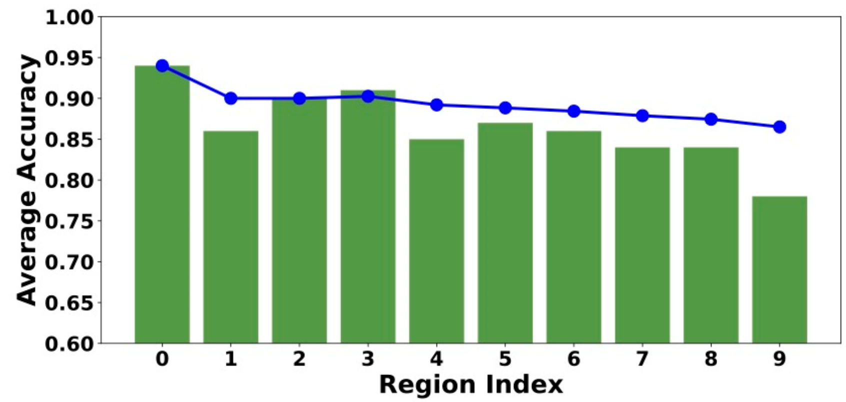
The difficulty of annotation can be induced from density of nodes in the feature space

Distance of nodes to their closest clustering centers (CC)

Difficulty-aware active node selection

If we group and sort nodes with their distances to each one's CC





LLMs present better annotation quality (lower difficulty) to those nodes closer to their CC

Intuition: Closer to CC indicating nodes with more "common" features, it may be easier for LLMs to annotate "common" nodes

Difficulty-aware active node selection

Our methods: Combining difficulty-aware metrics with traditional graph active learning metrics

$$f_{act}(v_i) \qquad CDensity(v_i) = \frac{1}{1 + ||x_{v_i} - x_{CC_i}||}$$

Then, use ranking aggregation to combine metrics considering, more robust to scale differences

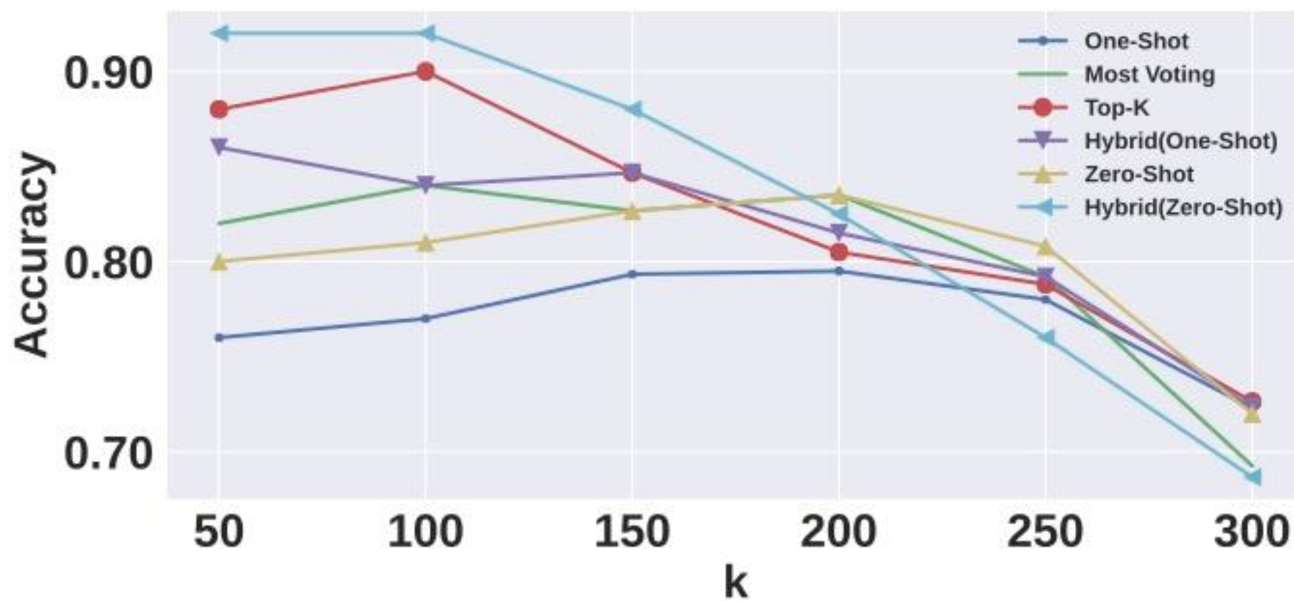
$$f_{act}(v_i) = \alpha_0 r_{f_{act}(v_i)} + \alpha_1 r_{CDensity(v_i)}$$

With proper hyper-parameters, we can get a good trade-off between diversity/representativeness and annotation difficulty

Confidence aware prompts + post filtering

We
filt

For
the



results

To
cor

2. Do k time queries and aggregate results

We sort nodes with their confidence and the higher confidence, the higher annotation quality, which shows the effectiveness of our hybrid strategy

Post filtering

We may further use confidence to post-select nodes

However, directly selecting nodes with top confidence may cause problems

➡ Diversity of the selected set is overlooked

In post-filtering stage, we can directly use label distributions to measure diversity

Post filtering

We propose a new metric *Change of Entropy*

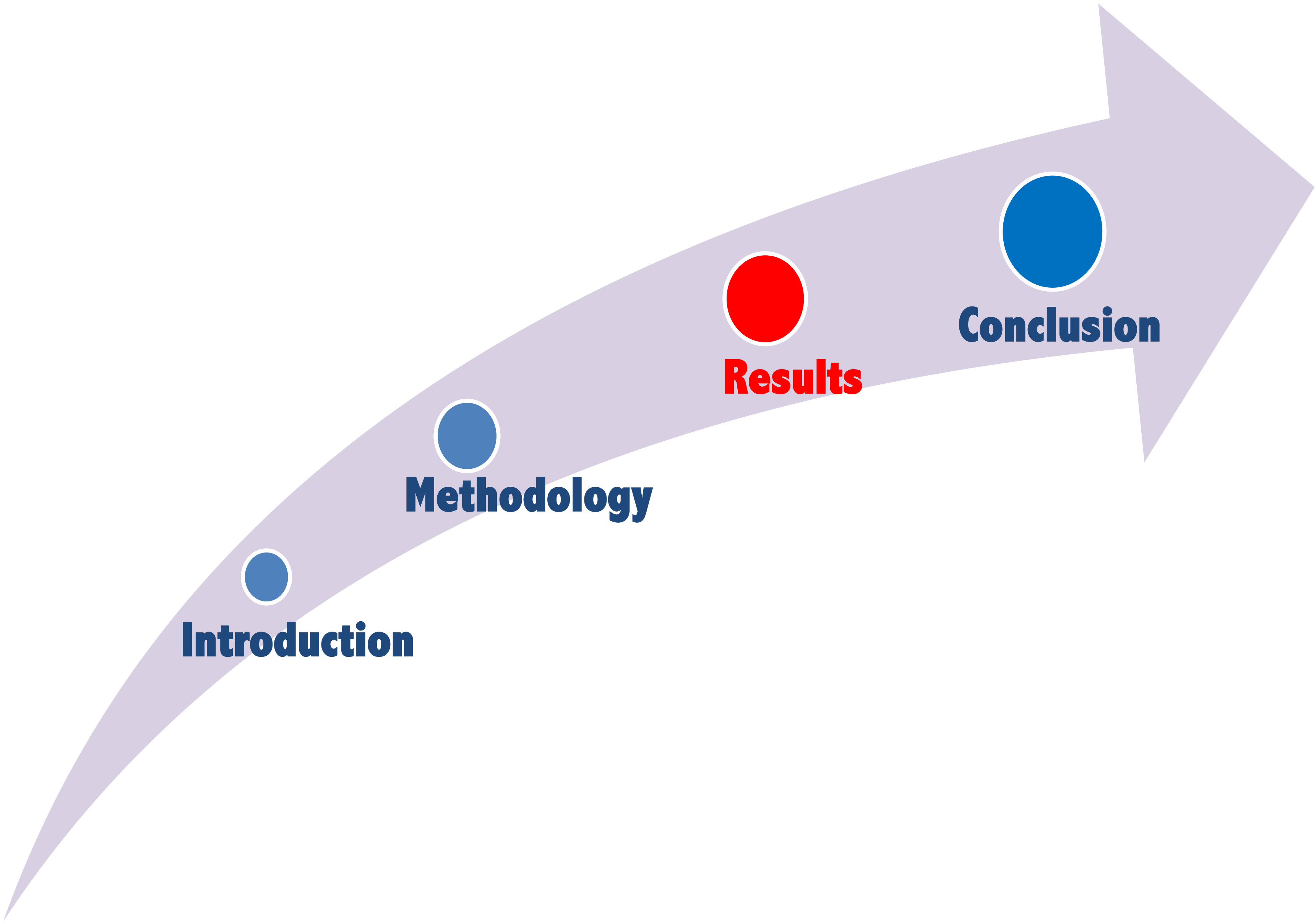
$\tilde{y}$**: LLMs' annotations;** V_{sel} **: the selected set of nodes;**H**: Shannon's entropy**

$$COE(\mathrm{v}_i) = H\left(\tilde{y}_{V_{sel}-\{v_i\}}\right) - H\left(\tilde{y}_{V_{sel}}\right)$$

Then, we still use rank aggregation to combine COE and LLMs' confidence

$$f_{filter}(v_i) = \beta_0 r_{f_{conf}(v_i)} + \beta_1 r_{COE(v_i)}$$

Each time, the node with the smallest f_{filter} **will be dropped and** COE **will be recomputed, until a ratio** γ **is reached**


Conclusion
Results
Methodology
Introduction

Overview of the experimental results

- RQ1: Difficulty-aware active selection (DA), Post filtering (PS), and combining DA with PS, which one is most effective?

- RQ2: How does our pipeline compare to other label-free classification methods

- RQ3: Advantages and limitations of our methods

RQ1 effectiveness

How to combine DA, PS, and graph active learning to achieve the best results?

Observation 0: Directly using C-Density to select nodes will suffer from the diversity problem

	CORA	CITESEER	PUBMED	WIKICS	OGBN-ARXIV	OGBN-PRODUCTS
Random	70.48 ± 0.73	65.11 ± 1.12	75.64 ± 2.15	62.30 ± 1.73	64.59 ± 0.16	70.59 ± 0.60
C-Density	42.22 ± 1.59	64.98 ± 1.15	39.76 ± 0.00	57.77 ± 0.85	44.08 ± 0.39	8.29 ± 0.00
PS-Random	69.83 ± 0.81	66.62 ± 0.72	73.77 ± 4.08	62.92 ± 2.18	64.18 ± 0.08	71.60 ± 0.34

C-Density can get very good annotation quality: near 90% accuracy across datasets, however, the label distribution will be imbalanced

RQ1 effectiveness

Observation 1: integrating DA and PS can both enhance the performance of graph active learning methods

	Cora	Citeseer	PubMed	WikiCS	ogbn-arxiv	ogbn-products
Pagerank	70.31 ± 0.42	61.21 ± 0.11	68.58 ± 0.14	67.13 ± 0.46	59.52 ± 0.03	69.87 ± 0.32
DA-Pagerank	72.79 ± 0.29	60.44 ± 0.40	75.02 ± 0.77	67.13 ± 0.80	58.82 ± 0.52	48.11 ± 0.13
PS-Pagerank	72.92 ± 0.26	61.87 ± 0.15	67.57 ± 0.31	[illegible]	[illegible]	[illegible]
PS-DA-Pagerank	72.19 ± 0.37	[illegible]	73.41 ± 0.19	[illegible]	60.12 ± 0.29	51.48 ± 0.39
AGE	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
DA-AGE	71.56 ± 0.37	57.18 ± 0.72	62.81 ± 1.84	58.67 ± 0.31	48.21 ± 0.80	60.03 ± 0.11
PS-AGE	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PS-DA-AGE	71.53 ± 0.19	56.38 ± 0.14	64.61 ± 0.29	59.74 ± 0.19	[illegible]	67.21 ± 0.39
RIM	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	OOT
DA-RIM	75.00 ± 0.35	[illegible]	63.97 ± 0.94	66.95 ± 0.01	OOT	OOT
PS-RIM	72.06 ± 0.35	62.43 ± 0.25	76.97 ± 0.29	68.56 ± 0.39	OOT	OOT
PS-DA-RIM	72.34 ± 0.19	65.21 ± 0.17	[illegible]	[illegible]	OOT	OOT
GraphPart	[illegible]	[illegible]	[illegible]	[illegible]	OOT	OOT
DA-GraphPart	69.34 ± 2.08	69.39 ± 1.05	67.36 ± 4.31	71.32 ± 0.81	OOT	OOT
PS-GraphPart	[illegible]	[illegible]	[illegible]	[illegible]	OOT	OOT
PS-DA-GraphPart	66.64 ± 2.26	63.57 ± 4.14	66.78 ± 4.14	69.10 ± 2.46	OOT	OOT
FeatProp	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	74.04 ± 0.15
PS-FeatProp	75.54 ± 0.34	69.06 ± 0.32	74.98 ± 0.35	66.09 ± 0.35	66.14 ± 0.27	74.91 ± 0.17

Observation 2: combining DA and AGE together may not result in better performance, since the role of C-Density and confidence is similar (assuming using fixed hyper-parameter, if we tune parameters, DA and PS can be included in DA-PS)

Two methods may present different effectiveness across datasets. DA uses C-Density for computation, mainly leveraging the good separability in the feature space, hence requiring high-quality features; PS uses LLM for filtering, so it requires LLM to handle the corresponding task effectively.

RQ1 effectiveness

Observation 3: combining FeatProp with PS presents promising performance and efficiency

	CORA	CITESEER	PUBMED	WIKICS	OGBN-ARXIV	OGBN-PRODUCTS
Pagerank	70.31 ± 0.42	61.21 ± 0.11	68.58 ± 0.14	67.13 ± 0.46	59.52 ± 0.03	69.87 ± 0.32
DA-Pagerank	72.79 ± 0.29	60.44 ± 0.40	75.02 ± 0.77	67.13 ± 0.80	58.82 ± 0.52	48.11 ± 0.13
PS-Pagerank	72.92 ± 0.26	63.87 ± 0.15	67.57 ± 0.31	70.22 ± 0.41	59.30 ± 0.21	70.57 ± 0.38
PS-DA-Pagerank	72.19 ± 0.37	60.36 ± 0.14	73.41 ± 0.19	69.14 ± 0.29	60.12 ± 0.29	51.48 ± 0.39
AGE	69.15 ± 0.38	54.25 ± 0.31	74.55 ± 0.54	55.51 ± 0.12	46.68 ± 0.30	65.63 ± 0.15
DA-AGE	71.56 ± 0.37	57.18 ± 0.72	62.81 ± 1.84	58.67 ± 0.31	48.21 ± 0.80	60.03 ± 0.11
PS-AGE	72.30 ± 0.13	63.04 ± 0.18	70.84 ± 0.76	64.00 ± 0.37	50.63 ± 0.19	68.69 ± 0.13
PS-DA-AGE	71.53 ± 0.19	56.38 ± 0.14	64.61 ± 0.29	59.74 ± 0.19	50.55 ± 0.39	67.21 ± 0.39
RIM	68.28 ± 0.38	63.06 ± 0.11	76.48 ± 0.16	67.06 ± 0.16	OOT	OOT
DA-RIM	75.00 ± 0.35	60.33 ± 0.40	63.97 ± 0.94	66.95 ± 0.01	OOT	OOT
PS-RIM	72.96 ± 0.35	62.43 ± 0.25	76.97 ± 0.29	68.56 ± 0.39	OOT	OOT
PS-DA-RIM	72.34 ± 0.19	65.21 ± 0.17	71.76 ± 0.29	63.23 ± 0.29	OOT	OOT
GraphPart	69.54 ± 2.18	66.59 ± 1.34	78.52 ± 1.34	67.28 ± 0.87	OOT	OOT
DA-GraphPart	69.34 ± 2.08	69.39 ± 1.05	67.36 ± 4.31	71.32 ± 0.81	OOT	OOT
PS-GraphPart	69.26 ± 0.19	70.00 ± 0.35	78.45 ± 1.12	67.74 ± 0.32	OOT	OOT
PS-DA-GraphPart	66.64 ± 2.26	63.57 ± 4.14	66.78 ± 4.14	69.10 ± 2.46	OOT	OOT
FeatProp	72.82 ± 0.08	66.61 ± 0.55	76.28 ± 0.13	64.17 ± 0.18	66.06 ± 0.07	74.04 ± 0.15
PS-FeatProp	75.54 ± 0.34	69.06 ± 0.32	74.98 ± 0.35	66.09 ± 0.35	66.14 ± 0.27	74.91 ± 0.17

RQ1 effectiveness

Observation 4: PS is more robust to hyper-parameter selection compared to DA

In this table, we select identical weight for each part, which means $\alpha_0 = \alpha_1 = 1, \beta_0 = \beta_1 = 1$ since there's no validation set

	CORA	CITESEER	PUBMED	WIKICS	OGBN-ARXIV	OGBN-PRODUCTS
Pagerank	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	69.87 ± 0.32
DA-Pagerank	72.79 ± 0.29	60.44 ± 0.40	75.02 ± 0.77	67.13 ± 0.80	58.82 ± 0.52	48.11 ± 0.13
PS-Pagerank	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PS-DA-Pagerank	72.19 ± 0.37	60.36 ± 0.14	73.41 ± 0.19	69.14 ± 0.29	60.12 ± 0.29	51.48 ± 0.39
AGE	69.15 ± 0.38	54.25 ± 0.31	74.55 ± 0.54	55.51 ± 0.12	46.68 ± 0.30	65.63 ± 0.15
DA-AGE	71.56 ± 0.37	57.18 ± 0.72	62.81 ± 1.84	58.67 ± 0.31	48.21 ± 0.80	60.03 ± 0.11
PS-AGE	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
PS-DA-AGE	71.53 ± 0.19	56.38 ± 0.14	64.61 ± 0.29	59.74 ± 0.19	50.55 ± 0.39	67.21 ± 0.39
RIM	[illegible]	[illegible]	[illegible]	[illegible]	OOT	OOT
DA-RIM	75.00 ± 0.35	60.33 ± 0.40	63.97 ± 0.94	66.95 ± 0.01	OOT	OOT
PS-RIM	72.96 ± 0.35	62.43 ± 0.25	76.97 ± 0.29	68.56 ± 0.39	OOT	OOT
PS-DA-RIM	72.34 ± 0.19	65.21 ± 0.17	71.76 ± 0.29	63.23 ± 0.29	OOT	OOT
GraphPart	[illegible]	[illegible]	[illegible]	[illegible]	OOT	OOT
DA-GraphPart	69.34 ± 2.08	69.39 ± 1.05	67.36 ± 4.31	71.32 ± 0.81	OOT	OOT
PS-GraphPart	[illegible]	[illegible]	[illegible]	[illegible]	OOT	OOT
PS-DA-GraphPart	66.64 ± 2.26	63.57 ± 4.14	66.78 ± 4.14	69.10 ± 2.46	OOT	OOT
FeatProp	72.82 ± 0.08	66.61 ± 0.55	76.28 ± 0.13	64.17 ± 0.18	66.06 ± 0.07	74.04 ± 0.15
PS-FeatProp	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

We can see that PS is more stable across datasets, while DA often needs proper hyper-parameter to work well

We find for proper hyper-parameters (tuned on test dataset), DA-AGE and PS-DA-AGE can achieve good performance across datasets

How to find proper parameter without a validation set for DA is a future direction

RQ2 Comparison to other label-free baselines

Observation 5: Compared to traditional label-free baselines based on GNNs and smaller-scale LMs, our pipeline present much better performance

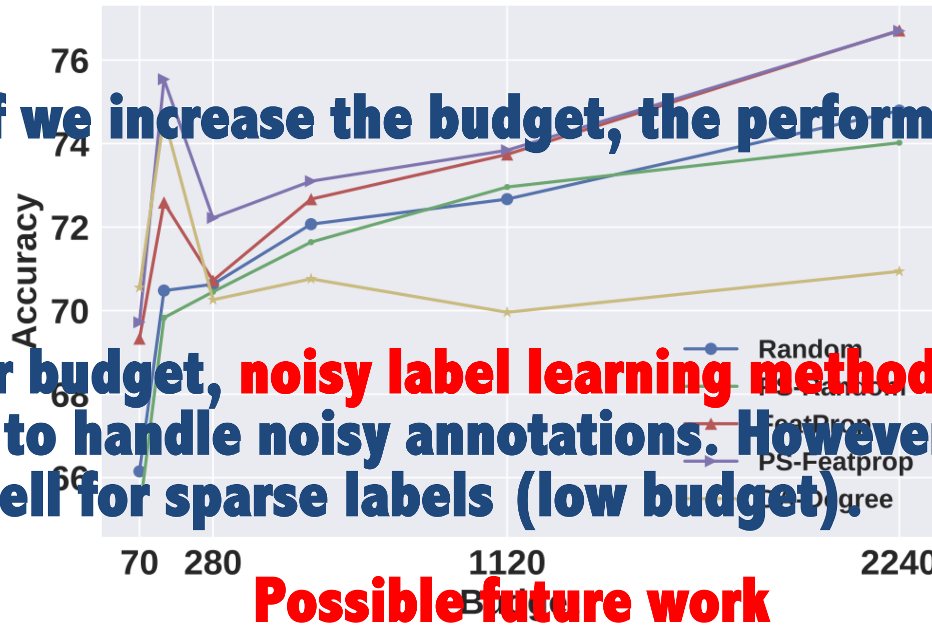
Compared to LLM-based baselines, our pipeline can achieve similar with results with much lower costs and scale to massive datasets

Table 5: Comparison of label-free node classification methods. The cost is computed in dollars. The performance of methods with * are taken from Li & Hooi (2023). Notably, the time cost of LLMs is proportional to the expenses.

	OGBN-ARXIV		OGBN-PRODUCTS	
Methods	Acc	Cost	Acc	Cost
SES(*)	13.08	N/A	6.67	N/A
TAG-Z(*)	37.08	N/A	47.08	N/A
BART-large-MNLI	13.2	N/A	28.8	N/A
LLMs-as-Predictors	73.33	79	75.33	1572
LLM-GNN	66.14	0.63	74.91	0.74

RQ3 Advantages and limitations of our methods

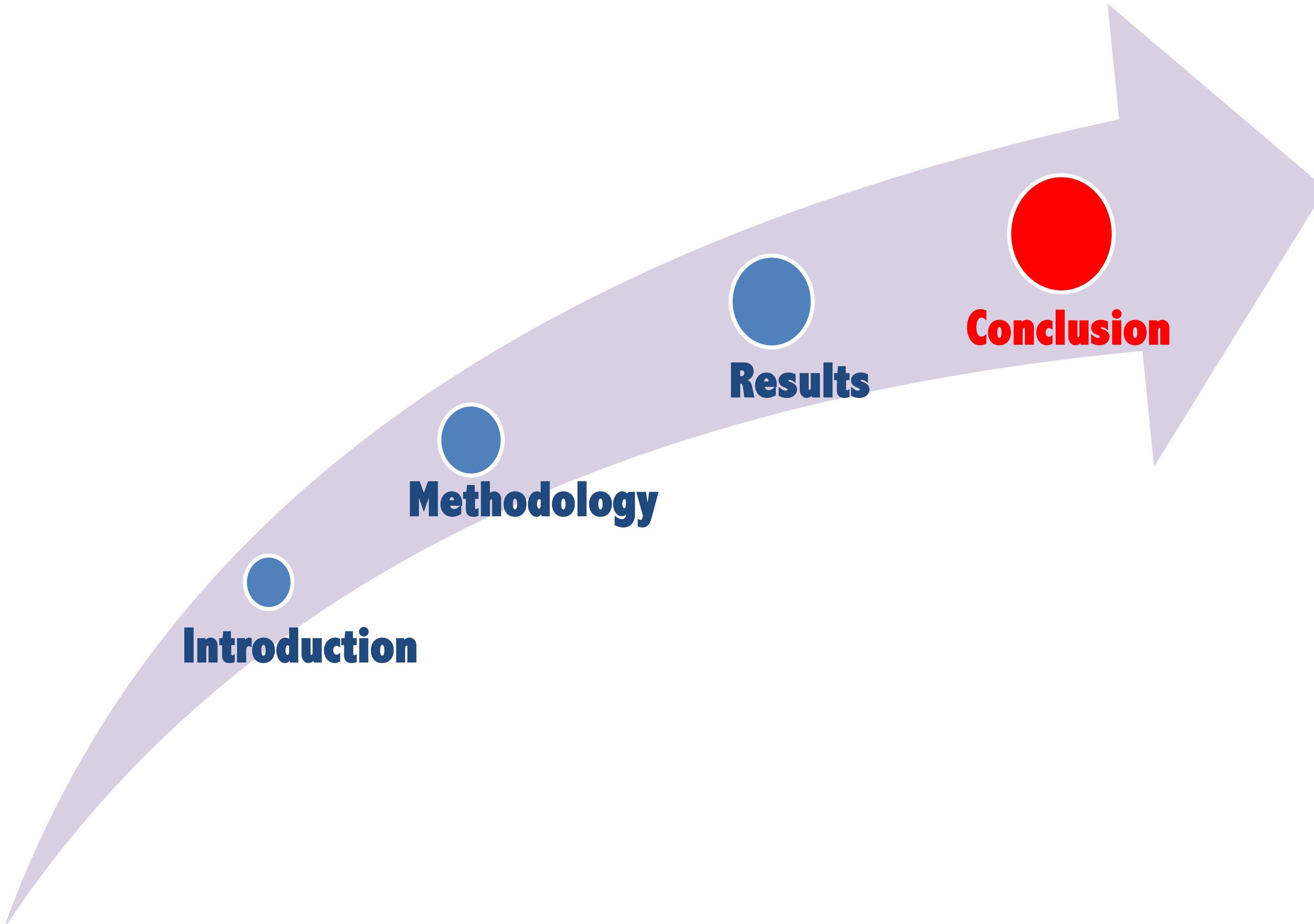
Observation 5: The characteristic of our method is that it can achieve a decent model with a very low annotation cost.



However, if we increase the budget, the performance gain is limited.

For a larger budget, noisy label learning method may be better way to handle noisy annotations. However, they may not work well for sparse labels (low budget).

Possible future work


Introduction
Methodology
Results
Conclusion

Conclusion

- We propose a new pipeline LLM-GNN which can harness the strength of both LLMs and GNNs
- We present an implementation of the pipeline with difficulty-aware selection, confidence-aware prompts, and post-filtering
- Our methods present promising effectiveness across different datasets and scale to large datasets with very low costs

Future Directions

- Extended to more types of graphs without text attributes. (A recent paper* provides a solution)
- Combined with weak supervision to handle more challenging annotations tasks
- Hybrid annotation with both human beings and LLMs

* Zhao, Jianan et al. "GraphText: Graph Reasoning in Text Space." (2023).