Appendix

A Datasets

B Baselines

C Detailed Introductions of Prompts

D Comparison of PLs and APLs

E Impact of Fine-tuning

F Impact of ST

Comparison of GSSL with and without ST

G Theoretical Motivation

H Design Philosophy Behind GAPL-LLM

H.1 Why SSL?

In many practical applications, collecting labeled data is an expensive and time - consuming process. Semi-Supervised Learning (SSL) leverages a small amount of labeled data and a large amount of unlabeled data to train models. Compared with traditional supervised learning, SSL can effectively utilize unlabeled data, thus improving the model's performance in data-scarce situations.

Its main advantages include: (i) It can utilize a large amount of unlabeled data for training, thereby enhancing the model's generalization ability; (ii) It can reduce the cost of labeling data and improve learning efficiency; (iii) It can achieve better prediction results with limited labeled data. Meanwhile, its main disadvantages include: (i) Due to incomplete data labeling, the model's prediction performance may be suboptimal; (ii) Appropriate algorithms need to be designed to enable effective learning with limited labeled data.

This paper aims to enable the proposed algorithm to fully explore the certain knowledge of labeled data and the potential knowledge of unlabeled data as much as possible.

H.2 Why LLM for SSL?

Compared with traditional models, the advantages of Large Language Models (LLMs) mainly contains: (i) **Stronger learning ability:** LLMs possess a larger number of parameters and more complex architectures, enabling them to fit data better and demonstrating great applicability across multiple domains. (ii) **Superior generalization ability:** The knowledge learned by LLMs during the training process is more general. They can better generalize to unseen data, reducing the reliance on a large amount of labeled data. They can learn more powerful feature representations and pattern recognition capabilities from a vast amount of data, thus effectively withstanding the impacts of noise and interference. (iii) **Higher efficiency computing ability:** By adopting technologies such as hierarchical design and distributed training, LLMs can be trained efficiently on existing hardware devices.

Traditional data annotation methods are usually carried out manually, which is not only costly but also inefficient. In contrast, LLMs can quickly adapt to specific tasks through techniques such as transfer learning and fine-tuning, greatly improving the efficiency and accuracy of data annotation. The advantages of LLM-based annotation are as follows: (i) **High efficiency:** LLMs have powerful feature extraction and generalization capabilities, enabling them to complete the annotation of a large amount of data in a short time and thus enhancing the annotation efficiency. (ii) **High accuracy:** Through transfer learning and fine-tuning, LLMs can better adapt to specific tasks, reducing annotation errors and improving annotation accuracy. (iii) **Automation:** LLMs have a certain degree of self-adaptability and can, to some extent, automate the annotation process, alleviating the burden on human annotators.

Existing pseudo-labeling methods generate pseudo-labels based on the predictions of unlabeled data by GNNs trained on labeled data. These methods suffer from issues such as unstable predictions, data distribution biases, and low data utilization rates. By incorporating the rich knowledge of LLMs, it is expected to generate stable and high-quality pseudo-labels.

H.3 Why Dynamic Threshold?

Corresponding content shown in Figure 1.(b) and Section 4.3.

H.4 Why Consistency Learning?

Deep learning models are prone to overfitting. When subjected to small perturbations (noise), the prediction results can be significantly affected. To mitigate over-fitting, in supervised learning, new loss terms are added, while in SSL, consistency regularization is employed. Its core idea is to constrain the features learned by the model by comparing the similarity between pairs of unlabeled data with the same label. The advantages include: improving model performance, reducing the dependence on labeled data, and being easy to implement.

Specific Description: Based on the smoothness assumption and cluster assumption in semi-supervised learning, data points with distinct labels are separated by low-density regions, while similar data points exhibit consistent outputs. When practical perturbations are applied to unlabeled data instances, the underlying principle dictates that their predictions should not exhibit significant deviations, thereby ensuring output consistency. This approach typically leverages prediction vectors from model outputs rather than explicit labels, making it inherently suitable for SSL frameworks. By constructing an unsupervised regularization loss term that measures the discrepancy between perturbed predictions $\hat{y}$ and original predictions y on unlabeled data, the model's generalization capabilities are systematically enhanced.