Appendix A. Implement Details

Regarding the relevant details of reproduction, our code can be found at https://github.com/xerrors/Labelprompt. Particularly in aspects such as the fine-tuning strategy, exemplar selection, and the reasoning behind certain modules (for example, the entity-aware module).

In our few-shot experiments, we fine-tuned all the model parameters rather than using in-context learning. This is similar to the approach used in KnowPrompt. Specifically, the model is fine-tuned on a small labeled dataset with a few examples per class. To ensure consistency, we randomly select them from the dataset using a specific random seed. The random seed for sampling is consistent with the random seed for model training, and the values are (1, 2, 3, 4, 5), ensuring that the same examples are picked each time the experiment is run with that seed.

Appendix B. Thinking about Large Language Models

B.1. Introduction to Large Language Models (LLMs)

In recent years, Large Language Models (LLMs) have witnessed rapid growth, with architectures like OpenAI's GPT-4 and Meta's LLaMA 3.1 representing cutting-edge developments. Their ability to learn from few examples, alongside the rise of prompt engineering, has sparked a shift in how NLP tasks are approached. LLMs now often replace fine-tuned models, showcasing high versatility and generalization abilities.

In our study, we utilized an encoder-only model (RoBERTa), which is primarily designed for Natural Language Understanding (NLU) tasks. These models process text bidirectionally, giving them an advantage in comprehension-based tasks like relation classification. In contrast, decoder-only models like GPT, LLaMA specialize in generative tasks. However, the computational costs and infrastructure needed to operate LLMs are significantly higher than those for small language models (SLMs), such as RoBERTa with 110M parameters, which are more suitable for scenarios requiring low-latency inference and specialized task optimization.

Despite LLMs' impressive capabilities in various tasks, the simplicity, efficiency, and precision of encoder-only models remain invaluable, particularly in tasks like intent recognition in vertical domains. Here, the need for fine-tuned, domain-specific performance outweighs the generalization offered by LLMs.

B.2. Performance

To better compare our method with the popular LLM-based approaches, we use in-context learning with Meta-LLaMA3.1-8B-Instruct, we randomly added 8/16 examples to the prompt per request. Take TACRED and ReTACRED as examples:

dataset	method	precision	recall	f1
retacred (sample 16)	meta-llama3.1-8b-instruct	30.28	30.25	30.26
retacred (sample 8)	meta-llama3.1-8b-instruct	32.47	32.44	32.45

Our experimental results highlight that RoBERTa outperformed LLMs in relation classification tasks (e.g., TACRED, ReTACRED). While LLMs can provide broad generalization, domain-specific tasks with fewer examples or lower variance between classes can benefit more from smaller, well-tuned models. For instance, in intent recognition, where specificity is key, encoder-only models deliver faster and more accurate predictions.