Knowledge-Enhanced Specific-Domain Reasoning For Pre-Training Language Models*

Xinbai Li

Abstract

The rapid advancement of language models (LMs) has prompted extensive research into enhancing their knowledge and reasoning capabilities. While pre-training language models (PLMs) acquire vast amounts of semantic knowledge from large-scale corpora, their ability to handle domain-specific understanding and reasoning tasks remains limited. Existing knowledge enhancement (KE) methods have been applied to both small-scale models and large language models (LLMs), leveraging external knowledge to improve prediction performance. However, challenges remain in effectively integrating structured knowledge into LLMs and in efficiently training smaller models under limited computational resources.

This thesis explores KE strategies tailored for both LLMs and smaller models. For LLMs, a key challenge is the natural integration of structured data, such as knowledge graphs (KGs), into text-based inputs suitable for LLM processing. To address this, we propose GenKP, a knowledge prompt generation framework that injects knowledge into LLMs via in-context learning (ICL). GenKP utilizes LLMs in combination with KGs to generate knowledge samples, refining them through weighted verification and BM25 ranking to reduce noise and enhance factual accuracy. Experimental results demonstrate that GenKP improves LLM performance, outperforming traditional triple and template-based knowledge injection approaches.

For smaller models, we investigate the incorporation of external knowledge through knowledge distillation (KD). While LLMs can serve as teacher models

*Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 28, 2025.