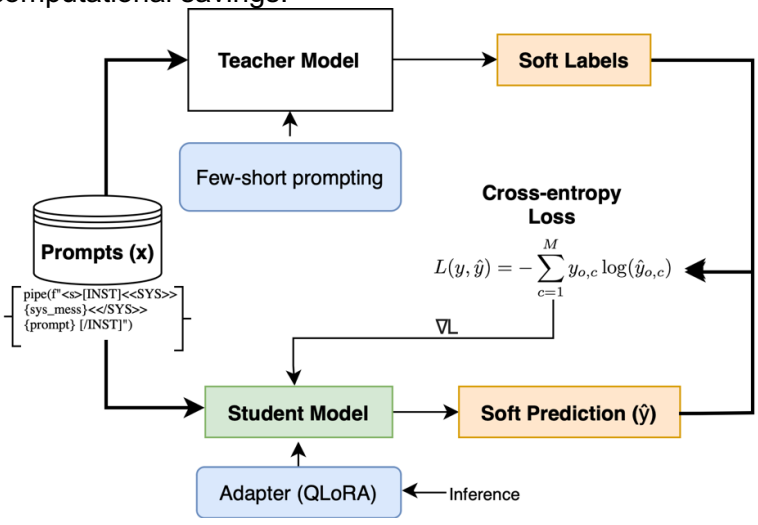
Adaptive Knowledge Distillation for Efficient Domain-Specific Language Models.

Large Language Models (LLMs) such as GPT, LLaMA, Mistral, etc have demonstrated remarkable capabilities across a wide range of tasks. However, customizing these pre-trained models for domain-specific applications requires significant computational and memory demands making them impractical for deployment in resource-constrained environments. Existing techniques, such as Knowledge Distillation (KD) [1, 5], Parameter-Efficient Fine-Tuning (PEFT) [2], and model parallelism address these issues by reducing the model size and the number of trainable parameters. KD compresses the knowledge of a larger, more complex model (teacher) into a smaller, more efficient model (student) while attempting to preserve the accuracy. In contrast, PEFT selectively tunes a small subset of parameters in large pre-trained models, freezing the rest, to reduce computational overhead. Key methods in PEFT include Adapters, BitFit, LoRa, Compacter, and Soft Prompts, each distinguished by their strategies to integrate and optimize a small set of parameters within large pre-trained models. Despite their advantages, these methods often fail to maintain the performance required in specialized domains. Moreover, current approaches of KD in LLM typically rely on black-box distillation, which uses hard labels and fixed architectures, limiting the flexibility and effectiveness of knowledge transfer in the models.

We introduce AKD, an Adaptive Knowledge Distillation framework that addresses these limitations by integrating adapters with white-box distillation. AKD uses soft target cross-entropy loss to transfer knowledge, exposing the student model to the teacher's output distribution and its internal representations, thus preserving critical information for domain-specific tasks, an improvement over black box distillation of LLMs. By integrating adapters (e.g., QLoRA [3]) into the student model, AKD focuses distillation on these newly added adapters while freezing the rest of the parameters. Additionally, our pipeline employs an adaptive prompt engineering optimization mechanism motivated by PromptAid [4], which allows exploring, perturbing, testing, and iterating over prompts to prompt a language model better. These prompts, refined through few-shot learning techniques, will enable the teacher model to produce more accurate and context-aware outputs, improving the quality of knowledge transfer to the student. AKD also features a progressive distillation strategy, where knowledge is transferred in phases from simpler to more complex tasks. This incremental approach ensures that the student model captures both high-level abstractions and domain-specific representations from the teacher. Overall, the AKD hybrid approach not only addresses the limitations of black-box distillation but also improves computational efficiency, achieving performance comparable to traditional KD methods while reducing training overhead.

We evaluated the AKD framework on domain-specific applications, such as crisis prediction and risk assessment, using a comprehensive dataset of 219,292 newspaper articles to assess its effectiveness in identifying potential signals for emerging open-domain crises. Our experiments involved training and distilling AKD with three open-source model families as teacher-student pairs: LLaMA-2 (13B; 7B), OPT (13B; 1.3B), and GPT- 2 (1.5B; 124M). We adopted the MiniLLM [5] training setup and evaluated AKD's performance for crisis prediction and risk assessment application. Our results highlighted, that AKD outperforms recent KD methods in terms of performance and as well as computational efficiency. We measured accuracy, F1 score, sensitivity, and specificity using a human-annotated dataset, where the LLaMA-2 teacher model with 13 billion parameters set a performance benchmark. Remarkably, our AKD student model based on the LLaMA-2 with 7 billion parameters and using QLoRA as an adapter not only matched but outperformed the teacher model's performance, demonstrating up to a 95% reduction in computational costs and a 99% reduction in trainable parameters compared to traditional knowledge distillation methods. Whereas the case of OPT and GPT-2 model families was more critical due to the drastic parameter reduction in the choice of teacher and student models. Despite this substantial decrease in parameters, the AKD achieved comparable performance to its teacher with up to 9% overall accuracy loss. Our results demonstrate that AKD offers a scalable and efficient framework for deploying LLMs in practical, resource-constrained environments, offering both high performance and significant computational savings.



[1] Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., & Zhou, T. (2024). A survey on knowledge distillation of large language models. arXiv.

[2] Ding, N., Qin, Y., Yang, G., & others. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 5, 220–235.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLORA: efficient finetuning of quantized LLMs. In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 441, 10088–10115.

[4] Mishra, Aditi & Soni, Utkarsh & Arunkumar, Anjana & Huang, Jinbin & Kwon, Bum Chul & Bryan, Chris. (2023). PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models.

[5] Gu, Y., Dong, L., Wei, F., & Huang, M. (2024). MiniLLM: Knowledge distillation of large language models. In Proceedings of the Twelfth International Conference on Learning Representations.