Streamlining LLMs: Adaptive Knowledge Distillation for Tailored Language Models

Prajvi Saxena
German Research Center for Artificial Intelligence, Saarbrücken, Germany
prajvi.saxena@dfki.de

Sabine Janzen
German Research Center for Artificial Intelligence, Saarbrücken, Germany
sabine.janzen@dfki.de

Wolfgang Maaß
German Research Center for Artificial Intelligence; Saarland University, Saarbrücken, Germany
wolfgang.maass@dfki.de

Abstract

Large language models (LLMs) like GPT-4 and LLaMA-3 offer transformative potential across industries, e.g., enhancing customer service, revolutionizing medical diagnostics, or identifying crises in news articles. However, deploying LLMs faces challenges such as limited training data, high computational costs, and issues with transparency and explainability. Our research focuses on distilling compact, parameter-efficient tailored language models (TLMs) from LLMs for domain-specific tasks with comparable performance. Current approaches like knowledge distillation, fine-tuning, and model parallelism address computational efficiency but lack hybrid strategies to balance efficiency, adaptability, and accuracy. We present ANON - an adaptive knowledge distillation framework integrating knowledge distillation with adapters to generate computationally efficient TLMs without relying on labeled datasets. ANON uses cross-entropy loss to transfer knowledge from the teacher's outputs and internal representations while employing adaptive prompt engineering and a progressive distillation strategy for phased knowledge transfer. We evaluated ANON's performance in the crisis domain, where accuracy is critical and labeled data is scarce. Experiments showed that ANON outperforms recent approaches of knowledge distillation, both in terms of the resulting TLM performance and in reducing the computational costs for training and maintaining accuracy compared to LLMs for domain-specific applications.

1 Introduction

In recent years, Large Language Models (LLMs) have revolutionized the way we interact with technology, setting a dominant trend in the current era of artificial intelligence. Industries are transforming themselves by including LLMs applications ranging from medical diagnostics leveraging interpretable LLM-based solutions (Bisercic et al., 2023), to financial risk analysis and market modeling (Wu et al., 2023), and real-time crisis detection by analyzing text data from news articles and social media (Saxena et al., 2024; Janzen et al., 2024). Despite their impressive capabilities, the deployment of LLMs for domain-specific tasks faces significant challenges. Full fine-tuning of these models requires vast labeled datasets and computational resources, discouraging many organizations, particularly those with constrained budgets. Therefore, effective strategies for model compression are critical to enable broader, practical use of LLMs in resource-constrained environments.

Existing research to address model compression and adaptation include knowledge distillation (KD) (Gu et al., 2023; Sanh et al., 2019), parameter-efficient fine-tuning (PEFT) (Ding et al., 2023), and model pruning (Fan et al., 2021). They essentially streamline a large model into a more efficient version without significant loss of performance. KD transfers knowledge from a larger "teacher" model to a smaller "student" model, preserving performance while reducing computational overhead (Dasgupta et al., 2023; Hsieh et al., 2023; West et al., 2022; Ko et al., 2024). PEFT approaches, such as Adapters (Houlsby et al., 2019), BitFit (Zaken et al., 2021), and LoRA (Hu et al., 2022), optimize a subset of parameters, allowing task-specific adaptation with minimal resource usage. Similarly, prompt-based tuning techniques, including prefix and prompt tuning, inject domain-specific information into model inputs without modifying the core architecture. However, these methods often operate in isolation, lacking hybrid mechanisms that integrate their strengths to address the trade-offs between memory efficiency, computational cost, task-specific performance and data limitation. Recent work, such as adapter distillation (Wang et al., 2023) and language universal adapters (Shen et al., 2023), highlights the potential of combining techniques but leaves room for further exploration of

hybrid approaches optimized for domain-specific applications.

To address these limitations, we propose ANON, a novel framework that combines KD with adapter-based PEFT for computationally efficient distillation of LLMs into domain-specific task language models (TLMs). ANON transfers knowledge using cross-entropy loss, using the teacher's output distribution and internal representations to retain both high-level abstractions and domain-specific details. The framework employs adaptive prompt engineering to optimize distillation, using data-driven prompts to effectively align teacher and student models effectively (Mishra et al., 2023). Additionally, ANON incorporates a progressive distillation strategy, transferring knowledge in stages from simpler to more complex tasks for comprehensive learning. Lightweight adapter modules, trained independently while freezing the rest of the model, significantly reduce computational costs, making ANON an efficient and scalable solution for domain-specific applications.

We evaluate ANON on a crisis-signaling task, focusing on early detection of potential crises using a corpus of 219,292 news articles. Following the experimental design outlined in (Saxena et al., 2024), we assess ANON's performance using teacher-student pairs from LLaMA-2 (Touvron et al., 2023), OPT (Zhang et al., 2022), and GPT-2 (Radford et al., 2019). These evaluations benchmark ANON against baseline KD methods. The results demonstrate that ANON achieves superior performance with significantly lower resource requirements. For instance, the student model LLaMA-$2_{\mathbf{7B_{ANON}}}$, distilled from the LLaMA-$2_{\mathbf{13B}}$ teacher surpasses the teacher's performance while reducing resource consumption by up to 95.24%. These findings highlight ANON's capacity to balance computational efficiency and domain-specific task performance, offering a scalable solution for resource-constrained AI applications.

2 Adaptive knowledge distillation for domain-specific TLMs

We propose ANON, an adaptive knowledge distillation framework designed to efficiently distill LLMs into domain-specific task language models (TLMs) as shown in Fig:1. ANON integrates lightweight adapter layers into the student model, enabling efficient training by focusing the distillation process on these new parameters while freezing the rest of the architecture. The framework employs cross-entropy loss to align the student model's predictions with the teacher's output distribution, facilitating accurate transfer of knowledge. By leveraging adapters such as LoRA, QLoRA, and Series Adapters (Dettmers et al., 2023), ANON further optimizes training efficiency and reduces computational costs without compromising model performance. The framework also leverages a progressive distillation strategy, where knowledge transfer is conducted in stages, starting with simpler tasks and gradually progressing to more complex ones. This hybrid approach produces a computationally efficient student model, $Student_{ANON}$, that achieves performance comparable to its teacher while significantly reducing resource requirements. The resultant model is well-suited for domain-specific applications such as medical diagnostics, risk management, and customer support, providing scalable and deployable solutions for real-world tasks.

2.1 Prompt Generation

ANON uses task-specific prompts to guide knowledge distillation between teacher and student models. Inspired by PromptAid (Mishra et al., 2023), the prompts follow a general structure with an optional system prompt, a mandatory user instruction describing the task, and a response format specifying machine-readable outputs. Prompts are tailored to the requirements of specific tasks and models. For example, a news article classification task might use a prompt like: "Classify the following news article into one of these categories: 'risk and warning,' 'caution and advice,' or 'safe and harmless.' Input: Energy sector warns of impending shortages and surging bills in upcoming months." These generated prompts serve as inputs to both the teacher and student models, aligning their learning objectives with the task.

2.2 ANON Workflow

Creating computationally efficient, domain-specific task language models (TLMs) requires balancing performance and resource constraints. The ANON framework introduces a comprehensive solution through adaptive knowledge distillation, employing a teacher-student architecture augmented with lightweight adapters. The teacher model, a large pre-trained language model such as LLaMA-$3.1_{(405B;\ 70B)}$ or GPT-4, serves as the source of rich, generalized knowledge. The student model, a smaller, efficient alternative like

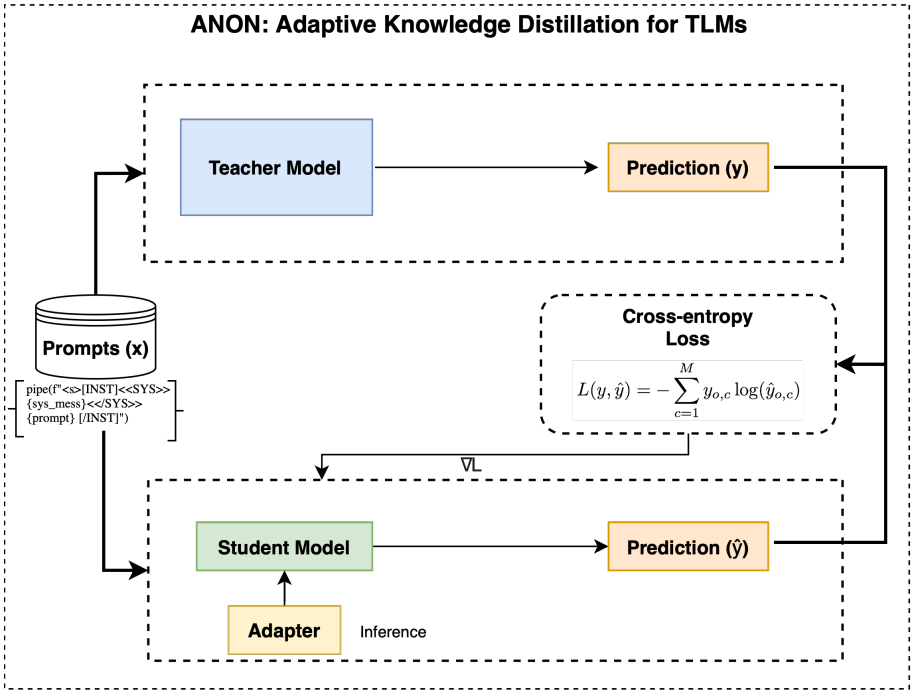

Figure 1: A detailed architecture of ANON the adaptive knowledge distillation for TLMs framework.

LLaMA-2_{7B} or GPT-2, is trained to replicate the teacher's outputs, reducing computational overhead while maintaining comparable performance. The distillation process ensures the student model aligns with the teacher model's output probability distribution. This alignment is achieved by designing prompts (x) that guide both models in generating the desired outputs. The teacher model's predictions (y) serve as ground truth for training the student. The optimization objective is formalized using the cross-entropy loss function:

$$L(y, \hat{y}) = -\sum_{c=1}^{M} y_{o,c} \log(\hat{y}_{o,c}) \qquad (1)$$

Here, M denotes the number of classes, while $y_{o,c}$ and $\hat{y}_{o,c}$ represent the true and predicted probabilities for class c. By minimizing this loss, the student model's predictions ($\hat{y}$) progressively align with those of the teacher, enabling robust performance with reduced computational complexity during inference.

To mitigate the resource demands of the distillation process, ANON integrates adapters within the student model. These adapters are small trainable modules that fine-tune specific components of the model while freezing the rest. By limiting updates to these adapters, ANON minimizes resource consumption during training, addressing the computational overhead associated with recalculating gradients and backpropagating errors for a large number of parameters. This targeted approach ensures that the student model achieves performance comparable to the teacher model while significantly reducing both training and inference costs.

3 Implementation and Evaluation

Based on the proposed framework (cf. Figure1), we implemented ANON for crisis signaling task following the experimental design outlined in (Saxena et al., 2024; Hassanzadeh et al., 2022). In the end, the distilled $Student_{ANON}$ provides domain-specific crisis signals and delivers alerts with confidence and severity levels.

3.1 Data Collection and Processing

An open-domain crisis signaling dataset of 219,292 news articles spanning 42 languages was used for ANON distilling. The dataset covered diverse crises such as supply chain disruptions, refugee movements, and economic instability. The dataset was compiled using keyword expansion and retrieved via the event registry API[1]1. The preprocessing involved standard text cleaning (e.g., removal of special characters and punctuation) and a two-stage filtration pipeline (Saxena et al., 2024). This resulted in a reduced dataset of 137,308 articles, representing 62% of the original corpus.

[1]https://www.newsapi.ai

Datasets	#Datapoints	Date Range	#Languages	#2-Step Filtration
Bushfires_Australia	9,035	2020 - 2022	23	4,509
Semiconductor_Shortage	19,449	2020 - 2022	7	11,193
Refugee_Crisis	82,671	2017 - 2019	31	53,109
Economic_Crises	107,220	2018 - 2022	34	67,868
Shipping_Port_Issues	917	2020 - 2022	1	629
Sum (Σ)	219,292	2017 - 2022	42	137,308

Table 1: Distribution of extracted and processed news articles across different stages of ANON training

We evaluate ANON's performance using real-world crisis newspaper datasets. (Saxena et al., 2024) provide a comprehensive descriptive analysis of these datasets, including distributions and ranges. For our study, we used 319 human-annotated articles centered on economic recessions and energy-related crises (e.g., supply chain disruptions, energy availability, and costs). These articles serve as a benchmark for model validation.

3.2 Training paradigm

The distillation process begins by generating prompts (x), using the prompt template2.1 for the classification task. Following (Gu et al., 2023), we use three teacher-student pairs: (LLaMA-$2_{\mathbf{13B}}$, LLaMA-$2_{\mathbf{7B}}$; $\text{OPT}_{\mathbf{13B}}$, $\text{OPT}_{\mathbf{1.3B}}$; and GPT-$2_{\mathbf{1.5B}}$, GPT-$2_{\mathbf{124M}}$). Prompts generated classify news articles into risk and warning, caution and advice, and safe and harmless. Few-shot prompting with 20 expert-annotated samples enhances teacher predictions. Once tuned, prompts were passed to teacher and student models for generating the classification predictions y and $\hat{y}$. The teacher model's output y serves as the true label during the distilling process. To minimize the divergence between the predicted probability distribution of the teacher and student models we use the cross-entropy loss function.

To optimize efficiency, we integrate Quantized Low-Rank Adapters (QLoRA), which apply 4-bit quantization and low-rank decomposition to self-attention layers. The weight matrices are factorized into two smaller matrices, A and B, controlled by rank r. After experimenting with 4, 8, 32, and 64 across all models, empirical tuning determined r = 64 as the best trade-off between compression and accuracy, based on the findings of (Hu et al., 2022). We use 4-bit NF4 precision, a cosine learning rate schedule (2e-4) with a 0.03 warmup ratio, and paged AdamW (32-bit) with weight decay (0.001) and max gradient norm (0.3). A dropout rate of 0.1 mitigates overfitting, and gradient checkpointing enhances memory efficiency.

This phased knowledge transfer strategy enables ANON to achieve high accuracy while significantly reducing computational overhead, making it well-suited for real-world crisis monitoring.

4 Results

We evaluated ANON on the (Saxena et al., 2024) benchmark, using accuracy, F1, sensitivity, and specificity (Table 2). Our experiments compare teacher models, standard student models, KD-based students, and ANON-trained students.

In some cases, ANON outperformed standard KD and surpassed the teacher model. Notably, LLaMA-$2_{7\text{B}_{\text{ANON}}}$ achieved 74.22% accuracy, exceeding both its teacher (71.19%) and KD-based student (74.06%), demonstrating enhanced generalization (Furlanello et al., 2018). Despite a 10x parameter reduction in OPT models and a 91.7% reduction in GPT-2, ANON preserved competitive performance even against the traditional KD method despite being far more efficient. Sensitivity generally exceeded specificity due to dataset imbalance, highlighting the need for bias mitigation strategies.

We also verified the performance of ANON for resource consumption. Our finding, detailed in Table 3 reveals that adding adapter modules into each student model leads to a remarkable decrease in computational demand. For the LLaMA-$2_{\mathbf{7B}_{\mathbf{ANON}}}$ model, there was a drastic reduction in memory requirements from approximately 84Gb to 4Gb when transitioning from standard KD to ANON, marking a 95.24% decrease. This result showcased the ANON's ability to maintain a comparable performance (cf. Table 2) while substantially lowering the memory requirements (cf. Table 3). Furthermore, ANON also reduced the number of trainable parameter counts by 99.43% for the LLaMA family case. In the case of the OPT and GPT-2

Model	#Params	Method	Accuracy	F1	Sensitivity	Specificity
LLaMA-2	13B	Teacher Model	71.19	68.45	**78.39**	62.72
	7B	Student Model	66.23	64.88	69.1	58.49
	$7B_{KD}$	KD	74.06	**72.37**	77.8	62.29
	$7B_{ANON}$	ANON	**74.22**	71.02	73.59	**62.8**
OPT	13B	Teacher Model	**62.31**	**61.94**	**70.72**	**58.06**
	1.3B	Student Model	46.92	41.03	42.5	39.38
	$1.3B_{KD}$	KD	59.7	58.2	61.58	57.46
	1.3_{ANON}	ANON	56.38	57.95	54.37	55.71
GPT-2	1.5B	Teacher Model	**53.89**	**51.76**	**51.93**	**48.47**
	124M	Student Model	34.70	33.72	40.68	38.07
	$124M_{KD}$	KD	42.92	40.8	47.61	41.06
	$124M_{ANON}$	ANON	40.68	40.02	47.33	38.8

Table 2: Result of the teacher and student models using ANON approach on crisis test datasets, including accuracy, F1 score, sensitivity, and specificity (Legend: KD = Knowledge Distillation; ANON = Adaptive Knowledge Distillation for Tailored Language Models)

	LLaMA-2_{7B}		**$OPT_{1.3B}$**		**GPT-2_{124M}**	
	16-bit float	**4-bit float**	**16-bit float**	**4-bit float**	**16-bit float**	**4-bit float**
Model Weights	14Gb	3.5Gb	2.6Gb	0.65Gb	0.24Gb	0.06Gb
Gradients	14Gb	0.08Gb	2.6Gb	0.04Gb	0.24Gb	0.0014Gb
Optimizer States	28Gb	0.16Gb	5.2Gb	0.08Gb	0.49Gb	0.0028Gb
gradients copy (fp32)	28Gb	0.16Gb	5.2Gb	0.08Gb	0.49Gb	0.0028Gb
Total	~84Gb	**~4Gb**	~15.6Gb	**~0.85Gb**	~1.48Gb	**~0.066Gb**

Table 3: Result of the memory consumption for LLaMA-2_{7B}, $OPT_{1.3B}$, and GPT-2_{124M} models after applying ANON framework using QLoRA as an adapter.

model families, similar efficiency gains are evident, which shows the ANON adaptability across different model sizes and architectures. In summary, the ANON framework enabled considerable computational savings without compromising the model performance.

5 Conclusion

In this work, we present ANON, adaptive knowledge distillation for tailored language models (TLMs). ANON addresses the challenges of limited training data and significant computational constraints associated with training and deploying LLMs for specific use cases. ANON leverages adapters and knowledge-distilling approach to achieve high performance and parameter efficiency in domain-specific applications. It can manage the complexities of dealing with a large corpus of data, supporting multilingual data processing without the burdensome costs associated with fine-tuning LLMs for downstream tasks. Additionally, it also addresses the issues of transparency, explainability, and maintaining accuracy in the complex high-parameter count model. To evaluate our approach we experimented with three different language model families for teacher-model distilling using a QLoRA adapter for crisis signaling task. The results showcased ANON's capability in terms of accuracy and resource consumption for practical scenarios of crisis signaling tasks. It achieved comparable and even exceeded the performance of teacher models, while significantly lowering memory usage by up to 95.24% and reducing parameters by 99.43% for some cases. Our framework not only advances the application of LLMs in crisis management but also lays a solid foundation for future research across various domains.

Acknowledgement

This work was partially funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the research project ESCADE (grant number: 01MN23004A).

References

Aleksa Bisercic, Mladen Nikolic, Mihaela van der Schaar, Boris Delibasic, Pietro Lio', and Andrija Petrović. 2023. Interpretable medical diagnostics with structured data extraction by large language models. *ArXiv*, abs/2306.05052.

Sayantan Dasgupta, Trevor Cohn, and Timothy Baldwin. 2023. Cost-effective distillation of large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7346–7354, Toronto, Canada. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Ning Ding, Yuxiao Qin, Guang Yang, et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5:220–235.

Angela Fan, Pierre Stock, Benjamin Graham, Edouard Grave, Remi Gribonval, Herve Jegou, and Armand Joulin. 2021. Training with quantization noise for extreme model compression. *Preprint*, arXiv:2004.07320.

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International Conference on Machine Learning*.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models. *arXiv*.

Oktie Hassanzadeh, Parul Awasthy, Ken Barker, Onkar Bhardwaj, Debarun Bhattacharjya, Mark Feblowitz, Lee Martie, Jian Ni, Kavitha Srinivas, and Lucy Yip. 2022. Knowledge-based news event analysis and forecasting toolkit. pages 5870–5873.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sabine Janzen, Prajvi Saxena, Sebastian Baer, and Wolfgang Maass. 2024. "listening in": Social signal detection for crisis prediction. In *Hawaii International Conference on System Sciences*.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. Distillm: Towards streamlined distillation for large language models. *Preprint*, arXiv:2402.03898.

Aditi Mishra, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2023. Promptaid: Prompt exploration, perturbation, testing and iteration using visual analytics for large language models. *Preprint*, arXiv:2304.01964.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv*, abs/1910.01108.

Prajvi Saxena, Sabine Janzen, and Wolfgang Maass. 2024. Newspaper signaling for crisis prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 166–173, Mexico City, Mexico. Association for Computational Linguistics.

Zhijie Shen, Wu Guo, and Bin Gu. 2023. Language-universal adapter learning with knowledge distillation for end-to-end multilingual speech recognition. *Preprint*, arXiv:2303.01249.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Junjie Wang, Yicheng Chen, Wangshu Zhang, Sen Hu, Teng Xu, and Jing Zheng. 2023. AdapterDistillation: Non-destructive task composition with knowledge distillation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 194–201, Singapore. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. *Preprint*, arXiv:2110.07178.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *ArXiv*, abs/2303.17564.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *Preprint*, arXiv:2402.13116.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

A Example Appendix

A.1 Background on Knowledge Distillation and QLoRA

Knowledge Distillation (KD) transfers knowledge from a large teacher model to a smaller student model by training the student to mimic the teacher's output distributions (Gou et al., 2021). It enables efficient deployment of Large Language Models (LLMs) by reducing computational overhead while preserving performance. KD is categorized into offline, online, and self-distillation (Xu et al., 2024). We adopt offline distillation, where a pre-trained LLM acts as a teacher to guide a smaller student model.

QLoRA (Dettmers et al., 2023), an extension of Low-Rank Adaptation (LoRA), integrates quantization into adaptation to enhance training and inference efficiency. By reducing weight precision from Float32 to int4, QLoRA significantly lowers memory usage and accelerates computation, making it well-suited for parameter-efficient fine-tuning (PEFT). It also improves memory efficiency through three key innovations. First, it introduces 4-bit NormalFloat (NF4), optimized for weights with a normal distribution, reducing the memory footprint. Second, Double Quantization applies quantization not only to model weights but also to quantization constants, further compressing storage. Third, paged optimizers dynamically manage memory, mitigating spikes during large-scale model training.

For a quantized base model with a LoRA adapter, the output of a linear layer is:

$$\mathbf{Y}^{BF16} = X^{BF16} \cdot doubleDequant(c_1^{FP32}, c_2^{k-bit}, W^{NF4}) + X^{BF16} L1^{BF16} L2^{BF16}$$

Here, W is stored in NF4, c_2 in FP8, with block sizes of 64 and 256, respectively, to balance quantization accuracy and memory efficiency. Parameter updates focus on adapter weights ($\frac{\partial E}{\partial L_i}$) rather than 4-bit weights ($\frac{\partial E}{\partial W}$). Conversion from W^{NF4} to W^{BF16} enables gradient computation in BF16 precision.

A.2 Prompts Examples

Fig. 2 illustrate the prompts used in our experiments. For all experiments, we employ teacher-student pairs such as LLaMA-2 (13B $\rightarrow$ 7B), OPT (13B $\rightarrow$ 1.3B), and GPT-2 (1.5B $\rightarrow$ 124M). These prompts are designed to provide clear and precise guidance for the distilling process. The customization of prompts for fine-tuning is dependent on the specific requirements of different models, although a general structure is commonly observed (Mishra et al., 2023). This structure typically includes an optional system prompt, such as 'Below is an instruction that describes a task', followed by a mandatory instruction detailing the task, for

system_message = """
You are an expert multilingual news analyst with advanced skills in understanding and interpreting news articles across multiple languages. Your job is to carefully classify each article into one of three categories: 'safe and harmless,' 'caution and advice,' or 'risk and warning.' For each classification, provide a concise and well-reasoned justification in English, explaining your decision based on the content of the article.
Examples:

1. Title: "Thousands of flights canceled as German airport staff strike - KION546"
Article: "BERLIN (AP) -- Thousands of flights to and from German airports were canceled Friday as workers walked out to press their demands for inflation-busting pay increases. The strikes...."
Classification: risk and warning
Reason: This news article falls under the category of 'risk and warning' because it highlights significant disruptions in air travel due to strikes, indicating potential ongoing issues and a warning of a "summer of chaos" if demands are not met.

2. Title: "Dezember bringt wenig Erbauliches für die Euro-Wirtschaft | Börsen-Zeitung"
Article: "Das Jahr 2022 endet für die Euro-Wirtschaft mit einem eher trüben Bild: Das Handelsbilanzdefizit hat sich im Dezember ausgeweitet und die Industrie hat die Produktion ...."
Classification: caution and advice
Reason: The news article presents a mixed outlook for the Euro economy, highlighting challenges such as an expanding trade deficit and reduced industrial production, while also mentioning signs of optimism like improved global trade and high order backlogs. This indicates a cautious tone, advising readers to be aware of economic difficulties while also recognizing potential recovery signs.

3. Title: "dpa-AFX: DZ Bank hebt fairen Wert für Munich Re auf 360 Euro - 'Kaufen'"
Article: "FRANKFURT (dpa-AFX Analyser) - Die DZ Bank hat den fairen Wert für Munich Re von 345 auf 360 Euro angehoben und die Einstufung auf 'Kaufen' belassen. Dass der Rückversicherer in einem durch Inflation, Hurrikan Ian, auslaufende Pandemie und Ukraine-Krieg geprägten Umfeld das ursprüngliche Gewinnziel übertroffen hat...."
Classification: safe and harmless
Reason: The news article provides a positive update on Munich Re's stock valuation and a recommendation to "buy," indicating confidence in the company's performance despite challenging external factors. Therefore, it falls under the category of 'safe and harmless.'
"""

Figure 2: Detailed prompt to extract ground truth labels from the teacher model.

instance, 'Classify the article into one of these categories: 'risk and warning', 'caution and advice', and 'safe and harmless''. User prompts are also incorporated to provide explicit instructions. The process concludes with the addition of the input article and, for fine-tuning purposes, the ground truth in terms of the output. For example, an input 'Energy sector warns of impending shortages and surging bills in upcoming months.....' would have an output 'risk and warning'. Thus, a comprehensive prompt might be formulated as: "Below is an instruction that describes a task. Instruction:the crisis article into one of these categories 'risk and warning', 'caution and advice', and 'safe and harmless'; Input: Energy sector warns of impending shortages and surging bills in upcoming months; Output: 'risk and warning'". To enhance outcomes, incorporating a few manually curated input examples for few-shot prompting with domain-specific samples is recommended. This approach underscores the pivotal importance of precise and thorough prompt design in facilitating effective training and knowledge distillation.