Knowledge Distillation of Black-Box Large Language Models

Anonymous ACL submission

Abstract

Given the exceptional performance of proprietary large language models (LLMs) like GPT-4, recent research has increasingly focused on boosting the capabilities of smaller models through knowledge distillation (KD) from these powerful yet black-box teachers. While leveraging the high-quality outputs of these teachers is advantageous, the inaccessibility of their internal states often limits effective knowledge transfer. To overcome this limitation, we introduce Proxy-KD, a novel method that uses a proxy model to facilitate the efficient transfer of knowledge from black-box LLMs to smaller models. Our experiments show that Proxy-KD not only enhances the performance of KD from black-box teacher models but also surpasses traditional white-box KD techniques. This approach presents a compelling new avenue for distilling knowledge from advanced LLMs.

1 Introduction

Recently, proprietary large language models (LLMs) like GPT-3.5 (OpenAI, 2022) and GPT-4 (OpenAI, 2023) have demonstrated significant superiority over open-source counterparts such as the LLaMA series (Touvron et al., 2023a,b; MetaAI, 2024). However, their vast number of parameters leads to high inference costs, and they are only accessible via API calls, offering limited customization and transparency. To address these challenges, recent efforts like Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and Orca (Mukherjee et al., 2023) have focused on transferring the capabilities of proprietary LLMs to smaller open-source models through knowledge distillation (Chen et al., 2023; Hsieh et al., 2023; Ho et al., 2022).

Knowledge distillation (KD) (Hinton et al., 2015) is a technique used to enhance the performance of a smaller student model by learning from a larger, more sophisticated teacher model. Depending on the level of access to the teacher

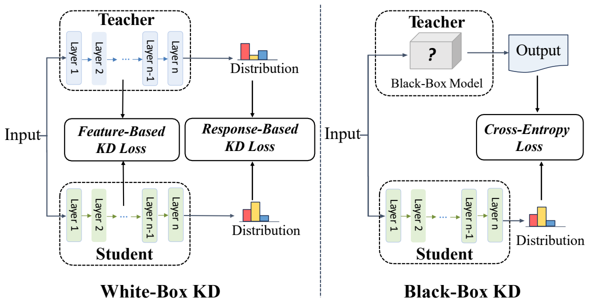

Figure 1: Comparison of white-box knowledge distillation (KD) and black-box knowledge distillation (KD).

model's internals, KD methods can be categorized into two types: KD with black-box teachers and KD with white-box teachers. As illustrated in Figure 1, white-box KD allows the student model to distill more intrinsic knowledge from the teacher by mimicing the teacher model's output distribution (Gu et al., 2023; Wen et al., 2023), hidden states (Jiao et al., 2020; Sun et al., 2019), and attention scores (Wang et al., 2021). Therefore, this method can only be applied when the teacher model's parameters are accessible. On the other hand, black-box KD leverages the high-quality outputs from powerful proprietary LLMs to fine-tune the student model (Hsieh et al., 2023; Fu et al., 2023). Both white-box and black-box KD have their respective drawbacks. While white-box KD is hindered by the limited capacity of the teacher model, which often restricts the distillation performance of the student, black-box KD faces challenges with knowledge transfer due to the inaccessibility of the teacher model's output distribution and internal states.

In this paper, we propose Proxy-based Knowledge Distillation (Proxy-KD) to better transfer knowledge from black-box teacher models. Proxy-KD introduces a proxy model, typically a white-box LLM, between the student and the black-box teacher. The proxy model first aligns with the capabilities of the black-box teacher by leveraging the teacher's outputs. Moreover, preference optimization is performed to further refine and enhance the alignment between the proxy and teacher models.

During the knowledge distillation process, the proxy model generates a dense distribution that closely approximates the black-box teacher's output distribution. This enables the student model to train effectively as if it were using the black-box teacher's guidance. Moreover, the outputs from the black-box teacher are treated as ground-truth labels, similar to traditional white-box knowledge distillation. Introducing the proxy model also mitigates the model capacity gap issue (Cho and Hariharan, 2019; Dong et al., 2023), which typically occurs when there is a significant disparity in strength between the teacher model (e.g., GPT-4) and the student model (e.g., LLaMA-1-7B).

To validate the effectiveness of our method, we conducted comprehensive experiments across a range of well-established benchmarks. The results show that Proxy-KD consistently outperforms both black-box and white-box KD methods. We observed that the alignment between the proxy model and the black-box teacher is crucial; a poorly aligned proxy model significantly diminishes the performance of knowledge distillation. We also found that larger and more robust proxy models are generally more desirable, as they possess stronger foundational capabilities and can align more effectively with the black-box teacher, enhancing the distillation process. Furthermore, we discovered that directly fine-tuning the proxy model with outputs from the black-box teacher is suboptimal for the alignment. These findings highlight the importance of selecting a well-aligned and capable proxy model to fully leverage the benefits of Proxy-KD.

2 Related Work

Existing knowledge distillation methods can be categorized into *white-box knowledge distillation* and *black-box knowledge distillation*.

2.1 White-Box Knowledge Distillation

Traditional knowledge distillation (KD) research predominantly employs white-box teachers and is typically classified into three main branches: feature-based, response-based, and relation-based methods. Feature-based methods seek to replicate the teacher's intermediate representations, such as attention scores (Jiao et al., 2020), attribution maps (Wu et al., 2023), and hidden representations of tokens (Sun et al., 2019). Response-based methods train the student model by minimizing divergences like Kullback–Leibler (KL) divergence (Hinton et al., 2015; Sanh et al., 2019), reverse KL (Gu et al., 2023; Wen et al., 2023), Jensen–Shannon Divergence (JSD) (Fang et al., 2021; Yin et al., 2020), and Total Variation Distance (TVD) (Wen et al., 2023) based on the teacher's output distribution. Relation-based methods train the student model by learning pairwise distances and triple-wise angles among token representations from the teacher (Park et al., 2021), or extracting structural relations from multi-granularity representations (Liu et al., 2022).

2.2 Black-Box Knowledge Distillation

Given the remarkable performance achieved by proprietary LLMs like GPT-4 (OpenAI, 2023), Claude 3 (Anthropic, 2024), and Gemini (Team et al., 2023), recent studies like Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and Orca (Mukherjee et al., 2023) have focused on transferring diverse capabilities from these black-box teachers into smaller open-source models. For instance, Li et al. (2024) and Liu et al. (2023) improved the mathematical capability of small models by training on tailored rationale samples generated by GPT-3.5-Turbo and GPT-4. To transfer the code generation capability, Azerbayev et al. (2023) prompted Codex (Chen et al., 2021) to create natural language-code pairs and fine-tuned a smaller model on these samples. To transfer the tool usage capability, Gou et al. (2023) utilized GPT-4 to generate interactive tool-use trajectories as training samples for the target model. Other approaches, such as Hsieh et al. (2023); Ho et al. (2022); Chen et al. (2023), utilize rationales generated by black-box teachers as training data to transfer their general reasoning capabilities.

White-box knowledge distillation (KD) efficiently distills knowledge by leveraging the internal states of the teacher model. However, white-box teachers typically possess a more limited capacity compared to their black-box counterparts. In contrast, black-box KD capitalizes on the superior performance of the teacher models but is restricted to fine-tuning on teacher-generated samples. This approach captures input-output patterns without accessing the deeper, intrinsic knowledge of the teacher model. To bridge these gaps, we propose Proxy-KD, a straightforward method that combines the strengths of both white-box and black-box KD while mitigating their respective limitations.

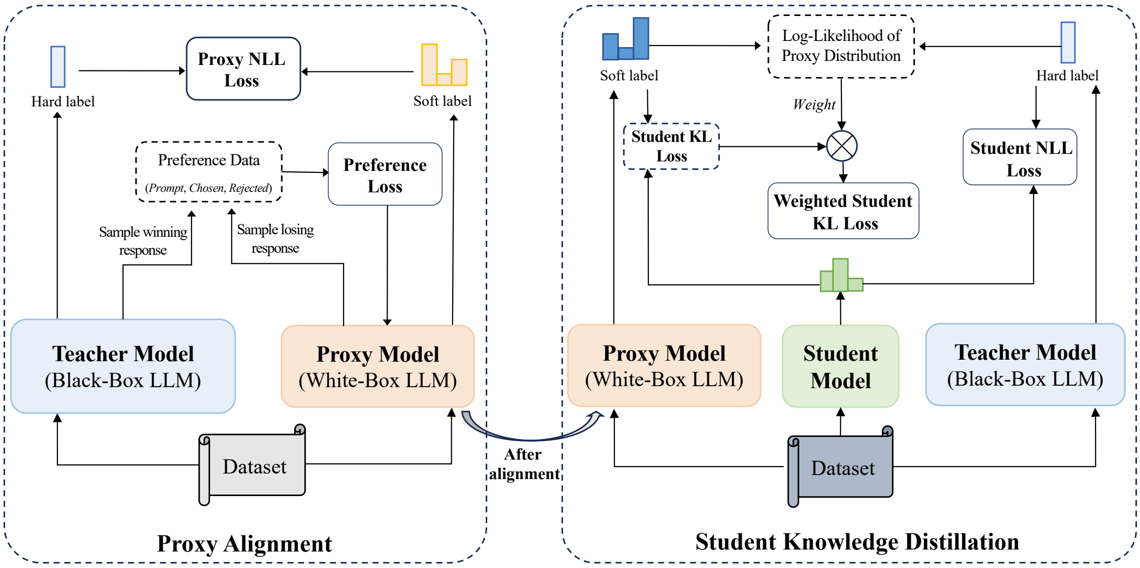

Figure 2: Overview of our proposed Proxy-based Knowledge Distillation (Proxy-KD).

2.3 Connection with Teacher Assistant

The proposed Proxy-KD method draws inspiration from TAKD (Mirzadeh et al., 2020), as both methods use an intermediate network to aid knowledge distillation, but they differ in three significant ways. Firstly, the motivation behind each approach is distinct: TAKD focuses on mitigating the capacity gap between the teacher and student in white-box settings, whereas Proxy-KD addresses the challenges posed by black-box teacher models and seeks to incorporate the benefits found in white-box scenarios. Secondly, they operate in different domains: TAKD is applied in the field of computer vision, while Proxy-KD is specifically designed for natural language processing, targeting the distillation of proprietary large language models (LLMs). Lastly, the methodologies diverge, with Proxy-KD introducing a crucial proxy alignment phase that includes preference optimization to better align the proxy model with the black-box LLM. This step is essential for reducing discrepancies between the proxy and teacher models, thereby improving the effectiveness of the distillation process.

3 Method

In this section, we introduce Proxy-based Knowledge Distillation (Proxy-KD), a simple yet efficient approach for knowledge distillation from black-box LLMs. As illustrated in Figure 2, Proxy-KD introduces a larger white-box LLM as the proxy aiming to capture the black-box teacher's knowledge. The process unfolds in two main stages: (1) proxy model alignment and (2) student knowledge distillation. First, the proxy model is aligned with the teacher through supervised fine-tuning and preference optimization. Once aligned, the student model learns from both the explicit outputs (hard labels) of the black-box teacher and output distributions (soft labels) provided by the aligned proxy.

3.1 Problem Statement

To facilitate the transfer of knowledge from a black-box teacher LLM π_t to a smaller, open-source student LLM π_s, we introduce a proxy model π_p. The training dataset $\mathcal{D}$ consists of input-output pairs (x, y), where x represents the input prompt and y is the output sequence generated by the teacher model π_t. This dataset is strategically divided into three parts: 10% ($\mathcal{D}_w$) for the warm-up phase, 45% ($\mathcal{D}_p$) for aligning the proxy model with the teacher, and the remaining 45% ($\mathcal{D}_s$) for the knowledge distillation training of the student model.

The process begins with a warm-up phase where the proxy model π_p is trained on $\mathcal{D}_w$. This phase helps π_p develop a basic capability to generate responses to input prompts. Following this, the proxy model undergoes alignment with the teacher model π_t using the next dataset, $\mathcal{D}_p$. This alignment is achieved through two methods: hard-label knowledge distillation (KD) and preference learning. These methods enable π_p to approximate the behavior and outputs of the teacher model. Once aligned, π_p acts as an intermediary, facilitating the transfer of knowledge to the student π_s on $\mathcal{D}_s$.

3.2 Preliminary

Hard-Label Knowledge Distillation. In this approach, the student model is trained using the outputs generated by the teacher model by minimizing the negative log-likelihood (NLL) function:

$$\mathcal{L}_{\text{NLL}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\log \pi_s(y|x)\right], \quad (1)$$

where $\pi_s(y|x)$ is the probability of π_s generating y given x. This approach is essentially a form of supervised fine-tuning and typically employed when the teacher is a black-box model.

Soft-Label Knowledge Distillation. In this approach, the student is trained to imitate the token-level probabilities of the teacher, by minimizing the Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{D}_{\text{KL}}(\pi_t(y|x)||\pi_s(y|x))\right]. \quad (2)$$

This knowledge distillation approach is typically employed when the teacher is a white-box model.

While the KL divergence objective provides richer supervision signals by using the token-level output distributions of the teacher model, it cannot be applied to black-box teachers due to the inaccessibility of these distributions. Consequently, current methods (Chiang et al., 2023; Mukherjee et al., 2023) rely on supervised fine-tuning using the outputs generated by black-box models to transfer their knowledge. Proxy-KD addresses this limitation by using a proxy model to incorporate the KL objective. The proxy mimics the black-box teacher, allowing access to its output distributions and enabling a more effective knowledge transfer.

3.3 Proxy Model Alignment

The proxy model π_p is typically a larger white-box LLM than the student model π_s. For effective knowledge transfer, it's crucial to first align the output distribution of the proxy model with that of the black-box teacher model π_t. This alignment ensures that the proxy accurately captures the teacher's behavior.

The proxy model π_p first undergoes supervised fine-tuning on a warm-up dataset $\mathcal{D}_w$. Following this, the proxy is further trained on the $\mathcal{D}_p$ dataset by minimizing the NLL loss:

$$\mathcal{L}_{\text{Proxy-NLL}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_p}\left[-\log \pi_p(y|x)\right]. \quad (3)$$

To enhance the alignment of the proxy model with the teacher, we further introduce a preference learning-based alignment objective, with the hypothesis that the teacher model's responses are of higher quality compared to those from the unaligned proxy model. The objective is to iteratively adjust the proxy model so that it increasingly favors responses similar to those of the teacher while reducing its preference for its own initial outputs. To implement this, we employ the Direct Preference Optimization (DPO) algorithm (Rafailov et al., 2024), which refines the proxy model by systematically preferring the teacher's responses.

Specifically, for a given input x, we iteratively sample a response y from the teacher and $\hat{y}$ from the proxy. These responses form a preference pair $(x, y, \hat{y})$. To train the proxy model to prefer y over $\hat{y}$, we define the following preference loss function:

$$\mathcal{L}^{(i)}_{\text{DPO}}(x, y, \hat{y}) = \log\sigma\left[\beta\log\frac{\pi_p^{(i)}(y|x)}{\pi_p^{(i-1)}(y|x)} - \beta\log\frac{\pi_p^{(i)}(\hat{y}|x)}{\pi_p^{(i-1)}(\hat{y}|x)}\right], \quad (4)$$

where $\pi_p^{(i-1)}$ is the proxy model from the previous training iteration. The overall preference loss over all the preference samples is defined as:

$$\mathcal{L}^{(i)}_{\text{Pref}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_p,\hat{y}\sim\pi_p^{(i)}(x)}\mathcal{L}^{(i)}_{\text{DPO}}(x, y, \hat{y}). \quad (5)$$

At each iteration i, the proxy model is updated based on the combined objective that includes both the NLL loss and the preference loss:

$$\mathcal{L}^{(i)}_{\text{Proxy}} = \mathcal{L}^{(i)}_{\text{Proxy-NLL}} + \mathcal{L}^{(i)}_{\text{Pref}}. \quad (6)$$

This iterative process continues for a fixed number of iterations k or until the proxy model converges. Through this method, the proxy model π_p is aligned to emulate the distribution of the black-box teacher π_t, becoming an effective intermediary for transferring knowledge to the student model.

3.4 Knowledge Distillation

To transfer knowledge from the black-box teacher to the student model π_s, we define the first training objective using teacher-generated sequences and the hard-label knowledge distillation objective:

$$\mathcal{L}_{\text{Student-NLL}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_s}\left[-\log \pi_s(y|x)\right]. \quad (7)$$

Based on the proxy model aligned with the black-box teacher, which delivers accessible output distributions, we define another training objective for the student through soft-label knowledge distillation:

$$\mathcal{L}_{\text{Student-KL}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_s}\left[\mathbb{D}_{\text{KL}}(\pi_p(y|x)||\pi_s(y|x))\right]. \quad (8)$$

In this process, the proxy model functions as an intermediary for the black-box teacher, facilitating the transfer of knowledge to the student model. However, as illustrated in Figure 6 in Appendix, discrepancies between the teacher's and the proxy's output distributions persist even after aligning the proxy model, potentially degrading the effectiveness of knowledge distillation. To address these discrepancies, we propose a weighted approach to the soft-label knowledge distillation objective. By introducing weights, we dynamically adjust the influence of each sample based on the alignment quality between the proxy and the black-box teacher. This approach ensures that the student model prioritizes samples where the proxy's distribution closely matches the teacher's distribution and reduces focus on samples where it does not. The weights are calculated based on the log-likelihood of the teacher's output generated by the proxy, normalized by the mean and variance of these log-likelihoods:

$$
\begin{aligned}
&w(x,y) = \sigma\left[\frac{\log \pi_p(y|x) - \mu}{\gamma}\right], \\
&\mu = \mathbb{E}_{(x,y)\sim\mathcal{D}_s}[\log \pi_p(y|x)], \\
&\gamma^2 = \mathbb{V}\mathrm{ar}_{(x,y)\sim\mathcal{D}_s}[\log \pi_p(y|x)],
\end{aligned} \tag{9}
$$

where $w(x,y)$ is a weight reflecting the quality of the proxy's prediction for the sample (x,y), $\mathbb{V}\mathrm{ar}(\cdot)$ is the variance operation, γ is the standard deviation, σ is the sigmoid function that projects the score to nonzero. Based on Equation (8), we derive the weighted version of $\mathcal{L}_{\text{Student-KL}}$ as follow:

$$
\begin{aligned}
&\mathcal{L}_{\text{Weight-KL}} = \\
&\mathbb{E}_{(x,y)\sim\mathcal{D}_s}\left[w(x,y)\mathbb{D}_{\text{KL}}(\pi_p(y|x)||\pi_s(y|x))\right].
\end{aligned} \tag{10}
$$

Therefore, the overall objective for student knowledge distillation can be derived as:

$$
\mathcal{L}_{\text{Student}} = \mathcal{L}_{\text{Student-NLL}} + \alpha\mathcal{L}_{\text{Weight-KL}}, \tag{11}
$$

where α is a hyperparameter utilized to adjust the strength of the weighted KL loss.

This knowledge distillation strategy effectively blends the advantages of both black-box and white-box knowledge distillation methods, employing the proxy model to bridge the gap between black-box LLMs and open-source student LLMs.

4 Experimental Setup

In this section, we introduce the experimental settings of models, datasets, and method baselines.

4.1 Models and Datasets

Teacher/Proxy/Student Models. In Proxy-KD, we choose GPT-4 (OpenAI, 2023) as the teacher, which is a powerful proprietary large language model. We select LLaMA-2-70b (Touvron et al., 2023b) and LLaMA-2-13b (MetaAI, 2024) as the proxy, respectively. Our student models come from two model types: LLaMA-1-7B (Touvron et al., 2023a) and LLaMA-2-7B (Touvron et al., 2023b).

Training Corpus. We combine the OpenOrca (Lian et al., 2023) and Nectar (Zhu et al., 2023) datasets as our training corpus, containing a total of 1M output sequences generated by the block-box teacher GPT-4. The OpenOrca dataset consists of instruction-following tasks, where GPT-4 is prompted to generate responses based on diverse input instructions. Nectar is a 7-wise comparison dataset, we filter and select those responses derived from GPT-4. Following Li et al. (2024), we also incorporate synthetic data generated by GPT-4, based on existing benchmark training sets. We split the original training corpus $\mathcal{D}$ into three parts: 10% as $\mathcal{D}_w$ with 100K samples, 45% as $\mathcal{D}_p$ with 450K samples, and 45% as $\mathcal{D}_s$ with 450K samples.

Evaluation Benchmarks. Evaluation benchmarks include complex reasoning dataset BBH (Suzgun et al., 2022), knowledge-based datasets AGIEval (Zhong et al., 2023), ARC-challenge (Clark et al., 2018), and MMLU (Zeng, 2023), commonsense reasoning dataset CSQA (Talmor et al., 2019), and mathematical reasoning dataset GSM8K (Cobbe et al., 2021). All evaluated models apply a zero-shot greedy decoding strategy.

4.2 Training Configurations

All experiments are conducted on 8×A100 Nvidia GPUs with 80GB memory. All proxy and student models are trained for only one epoch. We use a constant learning rate of 1e-5 and the Adam optimizer, with a max sequence length of 1024. We set hyperparamter $\alpha = 100$ in Equation (11), and $k = 16$ for the number of proxy alignment iterations. All models are trained using LoRA (Hu et al., 2021) with mixed-precision: frozen parameters in bfloat16 and LoRA-trained parameters in float32.

4.3 Baselines

We compare Proxy-KD with different white-box KD and black-box KD methods.

White-Box KD. For knowledge distillation with

Method	#Params	AGIEval	ARC	BBH	CSQA	GSM8K	MMLU	Avg
Black-Box Teacher								
GPT-4	-	56.40	93.26	88.0	-	92.0	86.4	-
White-Box KD								
Forward KL ♣	7B	35.16	66.87	35.68	74.40	44.12	51.42	51.27
Reverse KL ♣	7B	35.37	66.93	35.94	74.45	44.11	51.23	51.34
Black-Box KD								
GPT-3 (Ho et al., 2022)	6.7B	-	-	-	56.76	6.75	-	-
FlanT5-XL (Fu et al., 2023)	3B	-	-	39.0	-	22.4	-	-
FlanT5-XXL (Fu et al., 2023)	11B	-	-	47.20	-	27.10	-	-
MCC-KD (Chen et al., 2023)	11B	-	-	-	84.93	33.99	-	-
MCC-KD (Chen et al., 2023)	7B	-	-	-	76.41	41.58	-	-
Orca-1 (Mukherjee et al., 2023)	13B	41.7	74.74	49.7	-	26.46	53.80	-
Orca-2 (Mitra et al., 2023)	7B	45.10	78.41	45.93	-	47.23	53.70	-
Orca-2 (Mitra et al., 2023)	13B	49.93	83.36	50.18	-	59.14	57.73	-
WizardLM (Xu et al., 2023)	13B	38.25	74.74	38.47	-	48.60	55.00	-
Vicuna (Chiang et al., 2023)	13B	29.3	-	23.3	-	-	-	-
Vanilla Black-Box KD ♣	7B	34.71	66.85	46.68	74.43	49.51	49.82	53.66
Proxy-KD	7B	36.59	71.09	53.40	75.18	53.07	51.35	**56.78**

Table 1: Overall results on evaluated benchmarks. The superscript ♣ represents our own implemented methods. GPT-4 is used as the black-box teacher, the chat version of LLaMA-2-70B is used as the white-box teacher, the base version of LLaMA-2-70B is used as the proxy model, student models are based on the LLaMA-2-7B backbone. All models utilize a zero-shot greedy decoding strategy for evaluation. Other results are from their original papers.

white-box teachers, we compare forward KL methods (Hinton et al., 2015; Agarwal et al., 2024) and reverse KL methods (Gu et al., 2023; Wen et al., 2023). The chat version of LLaMA-2-70b is utilized as the white-box teacher.

Black-Box KD. For knowledge distillation with black-box teachers, we compare the vanilla black-box KD methods (Mukherjee et al., 2023; Mitra et al., 2023; Xu et al., 2023), which directly fine-tunes the student model on the data generated by the black-box teacher.

For baselines implemented by us, we start from the same student checkpoint as Proxy-KD and use the same input prompts. In white-box KD, output sequences are generated by the white-box teacher, while in black-box KD, output sequences are generated by the black-box teacher.

5 Result and Analysis

In this section, we present the main results and additional experiments of Proxy-KD.

5.1 Overall Results

We show the overall comparison of Proxy-KD against baseline methods in Table 1. Overall, the performance of black-box KD methods outperforms that of white-box KD methods, demonstrating the efficacy of distilling knowledge from powerful black-box models. Notably, Proxy-KD further enhances the performance, consistently achieving higher scores across most evaluated benchmarks compared to the white-box KD methods and the vanilla black-box KD method. Specifically, the improvement is particularly pronounced in the challenging datasets like BBH and GSM8K, where Proxy-KD obtains scores of 53.40 and 53.07, respectively, outperforming even larger models trained using traditional black-box KD methods. This demonstrates the effectiveness of Proxy-KD in leveraging both hard and soft labels through a well-aligned proxy, thereby facilitating more accurate knowledge transfer.

We also present the performance changes of student models during the distillation process in Figure 3. We show the accuracy curves of students on the benchmark test sets for every 40K training steps. We compare three methods: vanilla black-box KD, Proxy-KD, and white-box KD (forward KL). The results show that Proxy-KD stands out with the most significant enhancements, indicating its superior capability to efficiently transfer the comprehensive knowledge of black-box teachers to student models. The steeper and more consistent improvement curves of Proxy-KD across bench-

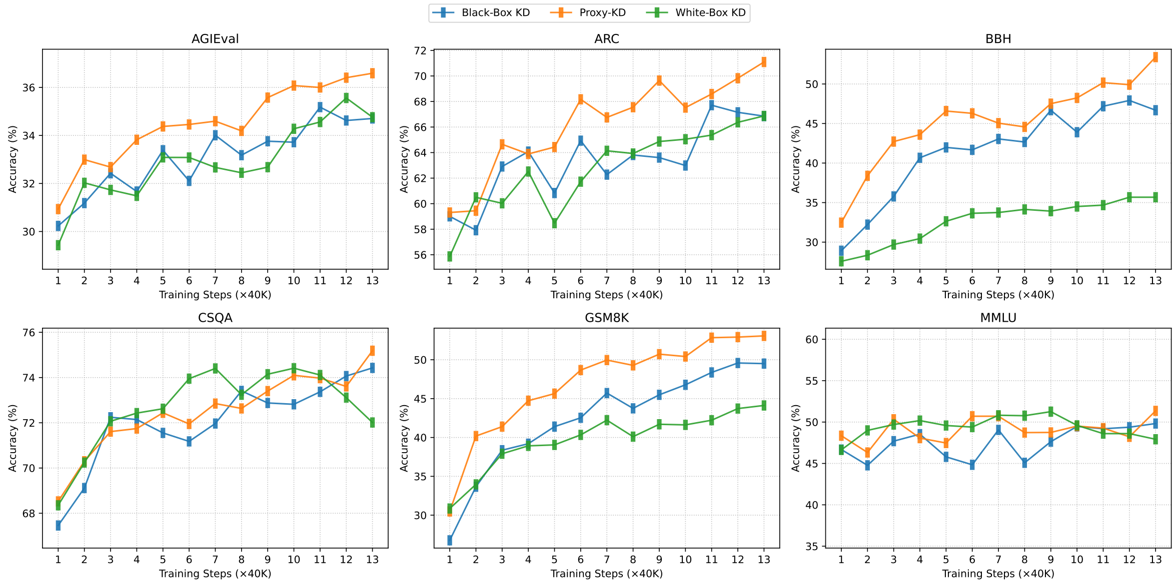

Figure 3: Accuracy curves for student models during knowledge distillation process. The y-axis is the accuracy of the students on the benchmark test sets, and the x-axis is the number of training steps. We compare Proxy-KD with black-box KD (vanilla black-box KD) and white-box KD (forward KL) baselines. Notably, Proxy-KD did not show sign of saturation on some benchmarks, such as AGIEval, ARC, and BBH benchmarks.

Method	AGIEval	ARC	BBH	CSQA	GSM8K	MMLU
			Studnet Model Distillation			
$\mathcal{L}_{\text{Student}}$	36.59	71.09	53.40	75.18	53.07	51.35
w/o π_p	34.71 (-1.88)	66.85 (-4.24)	46.68 (-6.72)	74.43 (-0.75)	49.51 (-3.56)	49.82 (-1.53)
w/o $\mathcal{L}_{\text{Proxy}}$	35.05 (-1.54)	67.18 (-3.91)	43.0 (-10.40)	76.04 (+0.86)	47.54 (-5.53)	48.09 (-3.26)
w/o $\mathcal{L}_{\text{Pref}}$	35.38 (-1.21)	66.11 (-4.98)	52.51 (-0.89)	75.51 (+0.33)	52.49 (-0.58)	48.79 (-2.56)
w/o $\mathcal{L}_{\text{Weight-KL}}$	33.99 (-2.60)	71.81 (+0.72)	51.50 (-1.90)	75.11 (-0.07)	52.91 (-0.16)	49.47 (-1.88)
			Proxy Model Alignment			
$\mathcal{L}_{\text{Proxy}}$	49.12	87.67	66.04	82.18	78.24	68.62
w/o $\mathcal{L}_{\text{Pref}}$	48.31 (-0.81)	86.93 (-0.74)	62.16 (-3.88)	80.95 (-1.23)	79.15 (+0.91)	66.38 (-2.24)

Table 2: Ablation studies of Proxy-KD. We examine the impact of the proxy model π_p, proxy model alignment loss $\mathcal{L}_{\text{Proxy}}$, proxy preference loss $\mathcal{L}_{\text{Pref}}$, and weighted KL loss $\mathcal{L}_{\text{Weight-KL}}$ on the performance of the student model training, as well as the impact of the proxy preference loss $\mathcal{L}_{\text{Pref}}$ on the performance of the proxy model alignment.

marks such as AGIEval, ARC, and particularly in complex tasks like BBH and GSM8K, underscore its robust and effective approach in leveraging proxy models for knowledge distillation.

5.2 Ablation Studies

In this section, we perform ablation studies to examine the impact of different components within our method. LLaMA-2-7B and LLaMA-2-70B are utilized as the backbones of the student and the proxy models, respectively.

Effect of the Proxy Model. The proxy model π_p is crucial for the effectiveness of Proxy-KD. Removing the proxy model forces the distillation process to revert to hard-label knowledge distillation, leading to significant performance drops across multiple benchmarks: a decrease of 4.24 on ARC, 6.72 on BBH, and 3.56 on GSM8K, as shown in Table 2. These declines underscore the proxy model's essential role in capturing and transferring the distributional knowledge from the black-box teacher, which is particularly important for tasks involving complex reasoning and mathematical challenges. Without the proxy, the student model fails to benefit from the detailed distributional guidance, resulting in markedly lower performance.

Effect of Proxy Model Alignment. The proxy model alignment, facilitated by the loss $\mathcal{L}_{\text{Proxy}}$, is vital for effective knowledge transfer. Table 2 shows that when the proxy is initialized directly from the LLaMA-2-70B checkpoint without alignment, the performance drops notably on BBH (-

10.40), GSM8K (-5.53), and MMLU (-3.26). This decline illustrates the adverse effect of an unaligned proxy, which fails to approximate the teacher's distribution and consequently underperforms compared to models directly fine-tuned on teacher data. The slight increase on CSQA (+0.86) when skipping alignment might be attributed to the simplicity of the task, indicating potential overfitting to teacher outputs without proxy guidance. This reinforces the necessity of the alignment process to ensure the proxy effectively bridges the knowledge transfer from the black-box teacher to the student model across diverse and complex tasks.

Effect of Preference Optimization. Table 2 illustrates the significant role of preference optimization in enhancing the performance of both the proxy and student models. When the proxy preference loss $\mathcal{L}_{\text{Pref}}$ is removed, reducing the proxy alignment loss to $\mathcal{L}_{\text{Proxy-NLL}}$, we observe notable performance drops across various benchmarks. Specifically, the alignment of the proxy model with the black-box teacher deteriorates, as evidenced by decreases in scores on benchmarks like BBH and MMLU, which subsequently impacts the student model. The overall trend confirms that preference optimization is crucial for refining the proxy model's ability to emulate the teacher effectively.

Effect of Weighted KL. When $\mathcal{L}_{\text{Weight-KL}}$ is replaced with the standard KL loss $\mathcal{L}_{\text{Student-KL}}$, we also observe declines in performance across most benchmarks, indicating that the effectiveness of the distillation process diminishes. The results shown in Table 2 highlight that focusing on high log-likelihood distributions from the proxy, as facilitated by the weighted KL loss, significantly enhances the quality of knowledge transfer. The overall declines underscore that this weighting mechanism significantly improves the quality of knowledge distillation, enhancing the student's ability to learn from a well-aligned proxy.

5.3 Impact of Proxy Model's Capability

How well the proxy aligned with the teacher can directly affect the performance of the student. The final alignment effectiveness of the proxy model depends on two factors: the design of the alignment algorithm and the inherent alignment capability of the proxy backbone model itself. In this section, we investigate the impact of the latter. We hypothesize that the size of the proxy model's parameters is crucial for its capacity to align with the black-box teacher's capability, especially when the teacher's parameter size is significantly larger than the proxy's. Experiments are conducted with LLaMA-2-70B and LLaMA-2-13B as the proxy backbone models. We show the performance of these aligned proxy models. As depicted in Figure 4, the proxy model based on LLaMA-2-70B performs better than the one based on LLaMA-2-13B, the latter has fewer parameters. We also examine the impact of proxy models with different capacities on student performance. We observe that the stronger proxy based on LLaMA-2-70B yields better student performance than the weaker proxy based on LLaMA-2-13B. Furthermore, when using a proxy based on a backbone model with a larger capacity, the student demonstrates a greater potential for achieving higher performance.

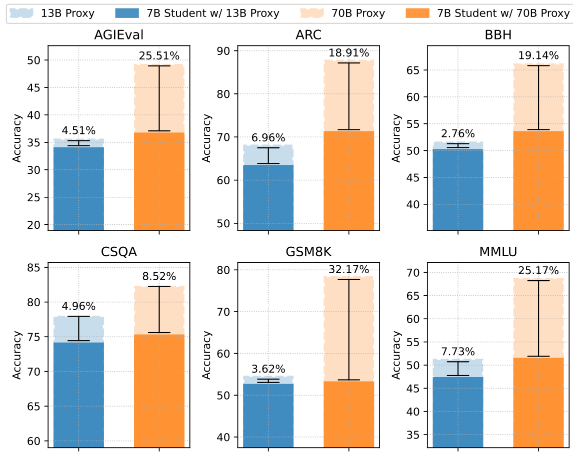

Figure 4: Performance of student models under different proxy models. We also show the ratio of performance gap between the proxy models and the student models.

6 Conclusion

This paper aims to tackle the challenge of knowledge distillation for black-box large language models (LLMs), where we can only access the outputs generated by the teacher model. Given the inaccessibility of the internal states of these black-box models, we introduce Proxy-KD, a novel approach that leverages a proxy model to enhance the distillation process. The proxy model is first aligned with the black-box teacher, closely mimicking its behavior. Then, the student model is trained using the combined knowledge from both the black-box teacher and the proxy model. Extensive experiments and analyses across a variety of well-established benchmarks demonstrate that Proxy-KD significantly outperforms existing black-box and white-box knowledge distillation methods.

Limitations

The limitations of this work include the training time overhead associated with proxy model alignment, particularly when the proxy model has a large number of parameters. Additionally, the proposed preference optimization requires online sampling from the proxy model, further increasing the training time overhead. Another limitation is the type of experimental backbone models used. Due to resource constraints, this work only conducts experiments with the LLaMA model series, without including other model backbones such as Qwen (Bai et al., 2023) or Mistral (Jiang et al., 2023).

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Anthropic. 2024. Claude 3 family. Accessed: 2024-06-04.

Zhangir Azerbayev, Ansong Ni, Hailey Schoelkopf, and Dragomir Radev. 2023. Explicit knowledge transfer for weakly-supervised code generation.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. 2023. MCC-KD: Multi-CoT consistent knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6805–6820, Singapore. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, et al. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ximing Dong, Olive Huang, Parimala Thulasiraman, and Aniket Mahanti. 2023. Improved knowledge distillation via teacher assistants for sentiment analysis. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 300–305.

Gongfan Fang, Yifan Bao, Jie Song, Xinchao Wang, Donglin Xie, Chengchao Shen, and Mingli Song. 2021. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems*, 34:11920–11932.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *ArXiv*, abs/2309.17452.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

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, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. 2024. Common 7b language models already possess strong math capabilities. *arXiv preprint arXiv:2403.04706*.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`.

Bingbin Liu, Sébastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. 2023. Tinygsm: achieving >80% on gsm8k with small language models. *ArXiv*, abs/2312.09241.

Chang Liu, Chongyang Tao, Jiazhan Feng, and Dongyan Zhao. 2022. Multi-granularity structural knowledge distillation for language model compression. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1001–1011, Dublin, Ireland. Association for Computational Linguistics.

MetaAI. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5191–5198.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *ArXiv*, abs/2306.02707.

OpenAI. 2022. Introducing chatgpt. Technical report.

OpenAI. 2023. Gpt-4 is openai's most advanced system, producing safer and more useful responses. Technical report.

Geondo Park, Gyeongman Kim, and Eunho Yang. 2021. Distilling linguistic context for language model compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 364–378, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332, Hong Kong, China. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed Huai hsin Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Annual Meeting of the Association for Computational Linguistics*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10817–10834, Toronto, Canada. Association for Computational Linguistics.

Siyue Wu, Hongzhan Chen, Xiaojun Quan, Qifan Wang, and Rui Wang. 2023. AD-KD: Attribution-driven knowledge distillation for language model compression. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8449–8465, Toronto, Canada. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *ArXiv*, abs/2304.12986.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied Sanosi Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv*, abs/2304.06364.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness and harmlessness with rlaif.

Models	#GPUs	Hours/Round
LLaMA-7B-SFT	4	1.0
LLaMA-7B-Distill	4	2.0
LLaMA-13B-SFT	8	1.8
LLaMA-13B-Pref	8	9.0
LLaMA-70B-SFT	8	5.5
LLaMA-70B-Pref	8	28.0

Table 3: Training time overhead. We show the training hours per round for different methods. SFT is the supervised fine-tuning method, Distill is the knowledge distillation method, Pref is the preference optimization method. Each round contains 40K training samples

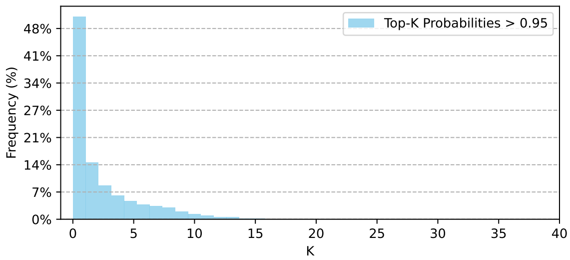

Figure 5: The statistics of the cumulative probability within the Top K exceeding 0.95. The x-axis represents different values of K, while the y-axis shows the percentage of instances meeting this threshold.

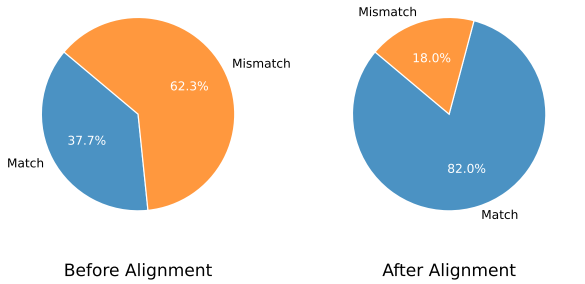

Figure 6: The match ratio between the proxy and teacher's output tokens before and after alignment. If the top-1 token given by the proxy equals the token given by the teacher in a current step, it is considered a match; otherwise, it is considered a mismatch..

A Experimental Analysis

A.1 Analysis of Training Efficiency

We show the training time overhead for different methods in Table 3. We show the training hours per round for supervised fine-tuning, knowledge distillation, and preference optimization methods across various model sizes. Each round contains 40K training samples. We note that preference optimization is the main time overhead due to online sampling from the proxy model. In Proxy-KD, we obtain the proxy model's output distribution offline during student distillation. As Figure 5 shows, most probability mass is concentrated on a few tokens. To save memory, only the top 10 token indices and their logits are retained.

A.2 Output Token Agreement

To serve as a stand-in for the teacher model's output distribution, it's important for the proxy model's output to align with the teacher model's output distribution, which is achieved through proxy model alignment. We measure the change in agreement between the top-1 token given by the proxy and the token provided by teacher in current step, before and after alignment. To visualize this alignment, at each step, consider the top-1 token given by the proxy's output distribution and the token given by the teacher. If the top-1 token given by the proxy matches the token given by the teacher at the current step, it is considered a match; otherwise, it is considered a mismatch. As shown in Figure 6, We find that after the proxy model alignment, the matched portions show a significant upward trend, indicating a trend towards alignment.

A.3 Additional Results

We present the performance changes of student models during the distillation in Figure 7. The student models are based on LLaMA-1-7B backbone, and the proxy model is based on LLaMA-2-70B backbone. We test the accuracy of students on benchmarks for every 20K training steps. We compare Proxy-KD with vanilla black-box KD method. We observe Proxy-KD consistently outperform vanilla black-box KD.

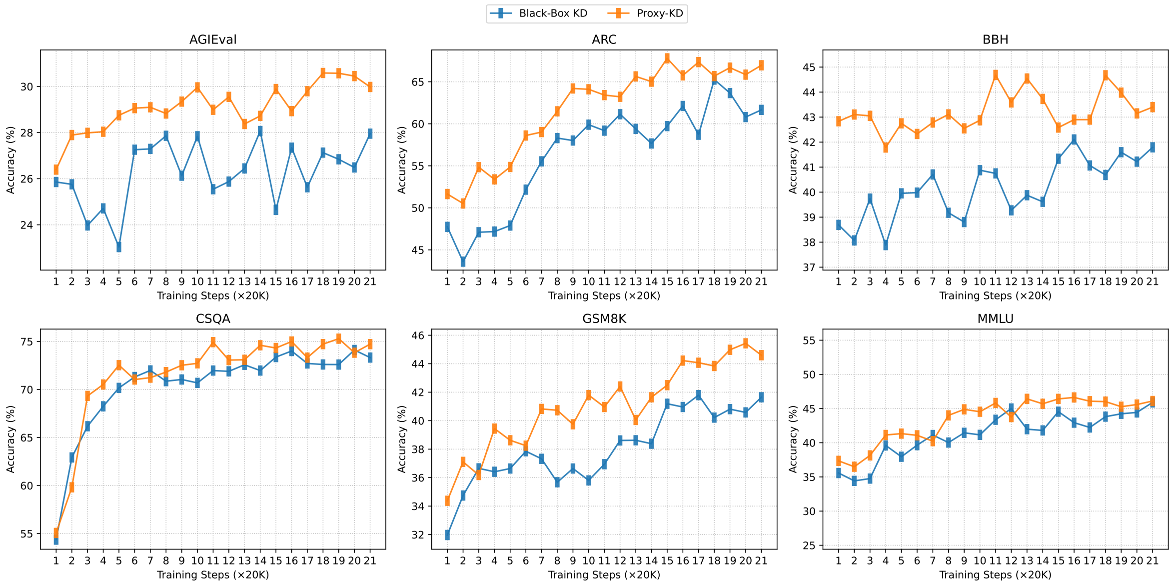

Figure 7: Accuracy curves for student models during knowledge distillation process. The y-axis is the accuracy of students on the benchmark test sets, and x-axis is the number of training steps. We compare Proxy-KD with vanilla black-box KD. The students are based on LLaMA-1-7B, and the proxy is based on LLaMA-2-70B.