GrayKD: Distilling Better Knowledge from Black-Box LLM via Multi-Rationale Injection

Hyeongsoo Lim[1], Hyung Yong Kim[2], Jin Young Kim[1], Min Ho Jang[1], Eun Seo Seo[1], Youshin Lim[2], Shukjae Choi[2], Jihwan Park[2], Yunkyu Lim[2], Hanbin Lee[2], Byeong-Yeol Kim[2], Ji Won Yoon[1]*

[1]Department of AI, Chung-Ang University, Seoul, Republic of Korea
[2]42dot Inc., Seoul, Republic of Korea
{andrew1001, wlsdud338, sunbi8534, jeo0534, jiwonyoon}@cau.ac.kr,
{hyungyong.kim, youshin.lim, shukjae.choi, jihwan.park, yunkyu.lim, hanbin.lee, byeongyeol.kim}@42dot.ai

Abstract

Knowledge distillation (KD) is a promising compression technique for reducing the computational burden of large language models (LLMs). Depending on access to the teacher model's internal parameters, KD is typically categorized into white-box and black-box KD. While white-box KD benefits from full access to intrinsic knowledge such as softmax distributions, black-box KD adopts a black-box LLM (e.g., GPT-4) as the teacher, which provides only text-level outputs via API calls. This limited supervision makes black-box KD generally less effective than its white-box counterpart. To bridge the gap between white-box and black-box KD, we propose GrayKD, a novel framework that can effectively distill text-level knowledge from a black-box LLM in a single-stage manner. In particular, rationales generated by the black-box LLM are injected into the student via a lightweight cross-attention module (teacher mode), enabling the model to approximate the black-box teacher's output distribution without access to internal parameters. The student is then trained with the softmax-level knowledge provided by the teacher mode (student mode). Since both the teacher and student modes share the same backbone, the proposed teacher mode remains highly parameter-efficient, requiring only a small number of additional parameters for rationale injection. Experimental results on instruction-following tasks demonstrate that GrayKD achieves substantial performance improvements over existing KD methods.

Introduction

Recently, large language models (LLMs) have achieved remarkable performance improvements by increasing their parameter scales, following the paradigm of scaling laws (Kaplan et al. 2020). Despite the promise, their intensive computational requirements significantly hinder deployment in resource-constrained environments. To mitigate this burden, several techniques, including knowledge distillation (KD), quantization, and pruning, have been adopted to reduce model size and inference cost. Among these, KD (Hinton, Vinyals, and Dean 2015) is a widely-used compression approach, which is the process of transferring knowledge from a large and powerful teacher model to a smaller, more efficient student model.

In the context of LLMs, KD can be categorized into two types depending on whether the internal parameters of the teacher model are accessible. In white-box KD, an open-source LLM, such as LLaMA (Grattafiori et al. 2024), Qwen (Bai et al. 2023), and SmolLM2 (Allal et al. 2025), serves as the teacher, enabling knowledge to be directly extracted from its checkpoints for distillation (Gu et al. 2024; Kim, Jang, and Yang 2024; Lee, Kim, and Lee 2024). In contrast, black-box KD employs a black-box LLM like GPT-4 and Claude as the teacher, where only text output is available via API calls. Despite their exceptional performance, black-box teachers cannot provide the full output distribution during KD. This limitation reduces the richness of transferable knowledge, making black-box KD less effective than its white-box counterpart. Experimentally, we find that the student distilled solely from GPT-4's text outputs performs worse than the one distilled from SmolLM2-1.7B. This observation aligns with prior studies in the KD literature, which demonstrate that softmax-level knowledge provides more informative supervision for training the student than text-only outputs (Yoon et al. 2023b, 2024). However, a black-box LLM still holds strong potential as a teacher, as it can generate high-quality outputs that encapsulate rich linguistic and reasoning capabilities. If we could approximate its output distribution effectively, the black-box LLM will serve as a more effective teacher by providing high-quality and accurate predictions.

To bridge the performance gap between black-box and white-box KD, we introduce **GrayKD**, a novel framework that can effectively transfer the text-level knowledge of the black-box teacher in a single-stage manner. Firstly, GrayKD injects rationales from the black-box LLM into the student model (teacher mode), guiding the student to approximate the softmax distribution of the black-box teacher. For rationale fusion, we apply a cross-attention mechanism, where the student model's representations serve as queries and the encoded rationales act as key-value pairs. Then, the student is distilled using the knowledge offered by the teacher mode (student mode), operating as the original student without any architectural modifications. Since both

*Corresponding author.

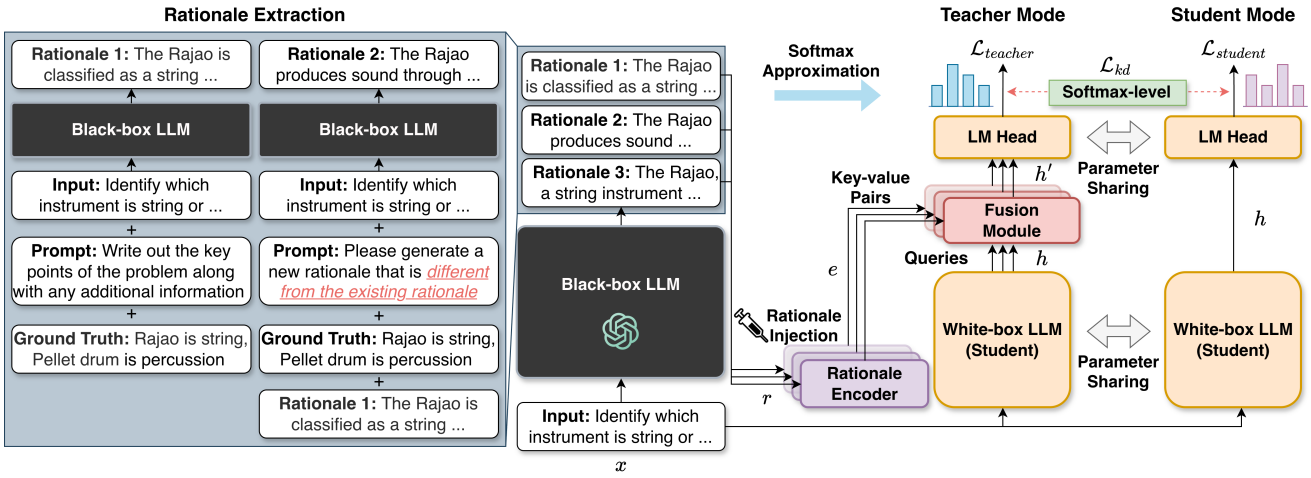

Figure 1: Overview of the GrayKD. The proposed KD consists of teacher and student modes. Rationales generated by the black-box LLM are injected into the student model, guiding the model to approximate the black-box teacher's softmax distribution (teacher mode). The softmax-level knowledge from the teacher mode is then transferred to the student model (student mode).

modes share the same student backbone and the teacher mode is constructed on top of it, the proposed teacher remains parameter-efficient, requiring only a small set of additional parameters. Unlike conventional KD approaches, GrayKD eliminates the need to train a large teacher model. Furthermore, this allows both teacher and student modes to be trained simultaneously in a single stage, thereby improving overall training efficiency.

Experimental results on instruction-following benchmarks demonstrate that GrayKD outperforms existing white-box and black-box KD methods. Notably, GrayKD achieves this superior performance with only 610M total parameters, including the student model itself. In contrast, prior softmax-level KD methods require training a separate 1.7B parameter white-box teacher model. Despite the significantly smaller model size, GrayKD achieves better results without the overhead of large-scale teacher training.

Related Work

Although black-box LLMs often demonstrate superior performance (Achiam et al. 2024; Anil et al. 2025), they tend to serve as relatively poor teachers for student models (Zhang et al. 2024), primarily because their APIs do not expose the full output distribution (Jin, Wang, and Lin 2023; Chen et al. 2024a; Yang et al. 2024). Specifically, the lack of access to model checkpoints severely limits the effectiveness of KD, thus constraining research in this domain. Early black-box KD approaches relied solely on the teacher model's generated text as supervision signal (Kim and Rush 2016; Hsieh et al. 2023), offering limited guidance. More recent efforts aim to enrich the knowledge from black-box LLMs by approximating the unseen softmax-level distribution behind the plain text returned by the API. To achieve this, an auxiliary proxy model is trained to imitate the teacher's logits, enabling softmax-level distillation to be performed through the proxy (Chen et al. 2024a,b). However, this approach still incurs a significant computational burden, as it requires training a large white-box proxy model. In our experiments, proxy-based KD involves training a 1.7B parameter teacher, which results in substantial computational cost. In contrast, our proposed method eliminates the need for any proxy model, yet still approximates black-box LLM logits effectively, while maintaining distillation quality and significantly reducing computational overhead.

Proposed Method

Motivation

In the context of KD for LLMs, both black-box and white-box approaches exhibit inherent limitations. While black-box LLMs demonstrate strong performance, their use in KD remains relatively underexplored due to their inability to expose full softmax distributions. To address this limitation, recent studies propose proxy models that approximate the intrinsic knowledge of black-box teachers (Chen et al. 2024a,b). However, as previously noted, training such proxies introduces substantial computational overhead, limiting their practicality. In the case of white-box KD, transferring softmax-level knowledge requires the teacher and student to share the same tokenizer, a constraint that is rarely satisfied among modern LLMs. Since heterogeneous tokenizers produce fundamentally different token distributions, aligning the softmax outputs between teacher and student becomes extremely challenging when their tokenizers differ (Cui et al. 2025a). This imposes a strict limitation on the flexibility of KD frameworks, particularly restricting the choice of teacher models. Consequently, the applicability of multi-teacher KD (Cui et al. 2025b), which aims to integrate complementary knowledge from multiple teachers, is also limited in settings where tokenizers are not shared.

GrayKD

To address these limitations, we propose the GrayKD framework, which effectively leverages text-level knowledge from a black-box teacher model while also supporting a multi-teacher setup. As shown in Figure 1, GrayKD framework comprises teacher and student modes, both built upon a shared student model backbone. This design achieves parameter efficiency by obviating the need for a large white-box teacher model required in conventional KD. In teacher mode, multiple rationale encoders receive the rationales extracted from the black-box LLM. By fusing these encoded rationales, GrayKD enables the teacher mode to approximate the black-box LLM's softmax distribution using only text-level outputs. Since the teacher mode shares the student's backbone, the approximated logits are naturally aligned with the student model, as they are derived from the same internal representations. Moreover, the diverse rationales extracted from the black-box LLM are injected into separate rationale encoders, which serve as a different teacher. This allows GrayKD to perform multi-teacher KD while requiring only 250M additional parameters.

Student Mode. The student mode operates exactly as the original student model without any architectural modification. We therefore consider an LLM that maps an input text sequence x into the last hidden state h via transformer layers of student mode $\mathcal{T}$, and subsequently predicts the next token y through the LM head $\mathcal{H}$. Formally, this process is defined as follows:

$$\mathcal{T}(x) \rightarrow h, \quad \mathcal{H}(h) \rightarrow y. \tag{1}$$

This sequence of operations constitutes the forward pass of the student mode within our KD framework. It is important to note that, during inference, only the original parameters of the transformer layers $\mathcal{T}$ and the LM head $\mathcal{H}$ of the LLM are utilized.

Teacher Mode. In the GrayKD framework, teacher mode introduces an additional rationale encoder and a fusion module into the student architecture. Specifically, the rationale encoder $\mathcal{R}$ first transforms the rationale text r, extracted from the black-box LLM, into rationale embeddings e. At this point, the rationale encoder is composed of several parallel sets, enabling the model to accept multiple rationales simultaneously as shown below:

$$\mathcal{T}(x) \rightarrow h, \quad \mathcal{R}(r) \rightarrow e. \tag{2}$$

Then, the rationale embeddings e serve as the key–value pairs for the fusion module $\mathcal{F}$, whereas the original hidden states h from the student's transformer layers are used as queries. Then, the fusion module $\mathcal{F}$ employs a cross-attention mechanism (Vaswani et al. 2017; Yoon et al. 2023a,b) to effectively integrate rationale with the original input of the student as follows:

$$\begin{aligned} h' = \mathcal{F}(h, e) &= \text{CrossAttention}(h, e, e) \\ &= \text{softmax}\left(\frac{he^\top}{\sqrt{d_k}}\right) e. \end{aligned} \tag{3}$$

Here, d_k denotes the dimensionality of the keys. Finally, the output head produces the prediction logits as shown below:

$$\mathcal{H}(h') \rightarrow y. \tag{4}$$

Since teacher mode shares the parameters of student mode, the only additional parameters come from the rationale encoder $\mathcal{R}$ and the fusion module $\mathcal{F}$, resulting in a parameter-efficient design.

Multi-rationale. In the proposed framework, multiple rationales are generated by the black-box LLM using a prompt that includes the question, the answer, and all previously generated rationales. The prompt is carefully designed to help the black-box LLM utilize the target answer effectively, while reducing overlap with previously generated rationales. Full prompt details will be provided in the Appendix. This setup ensures that each newly generated rationale overlaps minimally with existing ones, thereby maximizing informational diversity. We observe that greater informational diversity among the rationales yields larger performance gains in the proposed multi-teacher setup, which will be further explored in the analysis section. Once generated, each rationale is injected into its corresponding rationale encoder, enabling the model to integrate diverse perspectives during distillation. After injection, a fixed proportion (e.g., 15%) of each rationale is randomly masked during training. Since we cannot extract infinite rationale data, we vary the masking positions across rationales, obtaining benefits comparable to those of data augmentation. Specifically, two distinct rationales with different masking patterns can produce four unique rationale combinations, and three rationales can yield six combinations, enabling richer learning signals without requiring additional rationale annotations. Through this architecture, teacher mode naturally operates in a multi-teacher configuration, enabling multi-teacher KD.

Training Objective. Our framework includes three training objectives. First, the sequential KD loss ($\mathcal{L}_{\text{student}}$), which serves as the warm-up stage, aligns the student mode with GPT outputs at the text level, thereby improving agreement with rationales extracted from the black-box LLM (Chen et al. 2024a; Shrestha et al. 2025; Sun et al. 2025). During the initial training epochs ($i < N$), where N is the number of warm-up epochs, we train only the student mode. Second, the teacher mode loss ($\mathcal{L}_{\text{teacher}}$), a cross-entropy language-modeling loss evaluated on data that contains both ground-truth labels and GPT-generated outputs; and Third, the distillation loss ($\mathcal{L}_{kd}$), transferring knowledge from teacher mode to student mode,

$$\mathcal{L}_{\text{kd}} = D_{\text{KL}}\big(p_\tau^{\text{T}} \parallel p_\tau^{\text{S}}\big). \tag{5}$$

After this warm-up stage ($i \geq N$), we optimize the teacher mode. Overall, we have

$$\mathcal{L}_{main} = \lambda_1 \, \mathcal{L}_{\text{teacher}} + (1 - \lambda_1) \, \mathcal{L}_{\text{kd}}, \tag{6}$$

where λ_1 is a tunable hyperparameter. Additionally, $\mathcal{L}_{\text{student}}$ is used only during the warm-up stage ($i < N$).

Teacher								
KD Teacher	**Trained Model** (SmolLM2-1.7B)	**Method**	**Dolly**	**Self-Inst**	**Vicuna**	**S-NI**	**UnNI**	**Avg.**
None	White Teacher 1	SFT	25.72	19.23	19.12	36.37	31.76	26.44
Black Teacher (GPT)	White Teacher 2	SeqKD	23.76	17.41	23.61	38.62	34.45	27.57
Black Teacher (GPT)	White Teacher 3	ProxyKD	23.63	18.24	24.05	40.46	34.57	28.19
Student								
KD Teacher	**Student**	**Method**	**Dolly**	**Self-Inst**	**Vicuna**	**S-NI**	**UnNI**	**Avg.**
None	SmolLM2 (360M)	SFT	21.54	16.24	16.39	24.89	26.11	21.03
White Teacher 1		KD	23.88	16.46	17.28	27.14	28.49	22.65
		MiniLLM	26.11	19.23	19.39	32.19	33.63	26.11
		PromptKD	26.02	17.72	18.91	31.49	33.30	25.49
White Teacher 2		KD	21.36	15.19	18.82	26.22	26.13	21.54
		MiniLLM	22.48	16.16	23.39	34.05	28.83	24.98
		PromptKD	23.49	16.25	24.36	33.94	29.48	25.50
White Teacher 3		KD	21.74	15.34	19.07	27.01	26.32	21.90
		MiniLLM	23.41	16.99	24.36	35.44	30.59	26.16
		PromptKD	24.25	17.02	**25.37**	34.55	30.99	26.44
Black Teacher (GPT)		SeqKD	20.92	14.67	21.67	32.44	28.24	23.59
Black Teacher (GPT) + White Teacher 3		ProxyKD	21.68	15.20	21.39	32.77	28.57	23.92
		KD + ProxyKD	21.21	14.85	21.70	31.24	26.05	23.01
		MiniLLM + ProxyKD	21.01	14.95	20.93	31.97	28.79	23.53
		PromptKD + ProxyKD	22.72	15.32	22.46	33.40	28.35	24.45
Black Teacher (GPT)		**GrayKD (Dual)**	**26.60**	19.21	20.81	**35.66**	**34.91**	**27.44**
		GrayKD (Triple)	**26.55**	**20.08**	21.20	**35.75**	**34.61**	**27.64**

Table 1: Evaluation results of the proposed method and baseline models (Rouge-L, averaged over three random seeds: 10, 20, 30). For each benchmark, the student model that achieved the highest performance is highlighted in bold. GrayKD (Dual) and GrayKD (Triple) refer to configurations in which the teacher mode uses two and three rationales, respectively.

Experiments

Experimental Setup

Following the experimental setup of MiniLLM (Gu et al. 2024) and PromptKD (Kim, Jang, and Yang 2024), we adopted instruction-following (Ouyang et al. 2022) as our conditional text-generation task, where models generate responses conditioned on given instructions. Specifically, we constructed our training dataset following the MiniLLM approach, using Databricks-Dolly-15K (Conover et al. 2023), which comprises 15 K human-written instruction–response pairs. Samples that exceeded the models' maximum context length were filtered out, after which the dataset was randomly split into validation (1 K) and test (0.5 K) sets, leaving approximately 12.5 K examples for training. We evaluated our trained models using five instruction-following benchmarks: Dolly, our 500-sample test set derived from Databricks-Dolly-15K; Self-Inst (Wang et al. 2023), containing 252 user-oriented instruction-following examples; Vicuna (Peng et al. 2023), composed of 80 challenging instruction questions; the S-NI dataset (Wang et al. 2023), comprised of 9 K samples across 119 tasks, which were further split by ground-truth response-length ranges $[0, 5]$, $[6, 10]$, $[11, +\infty)$ as in prior works (Gu et al. 2024; Kim, Jang, and Yang 2024); and UnNI, 10 K samples which were randomly selected (Honovich et al. 2023), similarly subdivided by response length. Our primary evaluations used the $[11, +\infty)$ subsets of S-NI and UnNI, with a detailed analysis of all subsets provided in subsequent sections. To quantitatively assess generated responses, we employed the Rouge-L metric (Lin 2004), which is commonly used to evaluate the precision of instruction-following generation (Gu et al. 2024; Kim, Jang, and Yang 2024; Li et al. 2025).

Experimental Results

We compared GrayKD against six approaches: supervised fine-tuning (SFT), vanilla knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015), sequence-level KD (SeqKD)

Method		Cosine Sim	Dolly	Self-Inst	Vicuna	S-NI	UnNI	Avg.
GrayKD (Triple)	Baseline	**0.8254**	**26.55**	**20.08**	21.20	**35.75**	**34.61**	**27.64**
	Sim 1	0.8618	26.11	19.16	**21.70**	34.99	34.47	27.29
	Sim 2	0.8715	25.89	18.72	20.74	34.10	33.03	26.50

Table 2: Cosine similarity on the rationale dataset using mean-pooled last hidden state embeddings from the QwQ-32B model. Results corresponded to the GrayKD (Triple) and were averaged over three random seeds (10, 20, 30).

(Kim and Rush 2016), MiniLLM (Gu et al. 2024), PromptKD (Kim, Jang, and Yang 2024), and ProxyKD (Chen et al. 2024a). Specifically, we employed five teacher settings in our experiments. The black-box teacher was GPT-4o-mini. The original white-box teacher, referred to as White Teacher 1, was a SmolLM2-1.7B (Allal et al. 2025) fine-tuned via SFT. In addition, we constructed two proxy-based white-box teachers: White Teacher 2 and White Teacher 3, distilled from GPT-4o-mini using SeqKD and ProxyKD, respectively. The Rouge-L performance of the white-box teacher (White Teacher 1) and the proxy-based teachers (White Teacher 2 and 3) is reported in Table 1. Since the original ProxyKD setup (Chen et al. 2024a) involved both the black-box teacher and the white-box proxy (White Teacher 3) jointly supervising the student, we also included a combined teacher setting: Black Teacher + White Teacher 3. Across all configurations, we used SmolLM2-360M as the student model. For GrayKD, it operated in Dual and Triple configurations, depending on whether two or three rationales were injected into the teacher mode.

The Rouge-L results on the Dolly, Self-Inst, Vicuna, S-NI, and UnNI benchmarks are shown in Table 1. Despite GPT-4o-mini's inherently superior performance, both SeqKD and ProxyKD performed worse than conventional white-box KD methods, failing to provide effective guidance to student. The gap was especially noticeable for SeqKD, which relied solely on text outputs without using any proxy model. This suggests that effectively leveraging the black-box LLM as the teacher remains a particularly challenging problem. However, even when leveraging the same black-box teacher (GPT-4o-mini), GrayKD consistently outperformed all other methods across the five benchmark datasets, clearly demonstrating its effectiveness in practical settings. This trend was especially pronounced in the Triple configuration, where three rationales were utilized during distillation. In this setting, GrayKD outperformed all other student models in most configurations, recording an average Rouge-L score of 27.64. This represents a 1.2-point improvement over PromptKD with White Teacher 3, the strongest white-box KD baseline in our experiments. As noted earlier, White Teacher 3 was a 1.7B proxy model distilled from the black-box teacher via ProxyKD. This highlights a key strength of GrayKD: it can successfully approximate the softmax distribution of the black-box teacher without relying on any white-box proxy. Thus, the proposed method removes the need for large-scale teacher model training while maintaining or even surpassing performance levels of proxy-based or fully white-box distillation methods. This result is particularly notable, as it demonstrates that effective knowledge distillation can be achieved even in a purely black-box setting.

Analysis

Rationale Diversity. To better understand the role of rationale diversity, we conducted an additional set of experiments. As part of this analysis, we computed the cosine similarity between each rationale to quantify how similar the rationales were to one another. A higher similarity indicates lower rationale diversity. The table 2 shows how the performance of GrayKD's triple-teacher variant changes as we artificially increase the pairwise cosine similarity among rationales. In the proposed framework, we sampled three rationales per training instance. To maximize diversity, each rationale was generated conditioned on the previously sampled ones, with an explicit prompt instructing the black-box LLM to avoid overlap with earlier rationales. This configuration yielded the lowest similarity among rationales and the highest average score across benchmarks, as shown in Table 3. In Sim 1, we still used three rationales per training instance but modified the prompt to encourage more similar rationales, which slightly reduced the average performance. In Sim 2, diversity was further reduced by reusing a single rationale across all encoders, resulting in the largest performance drop. These observations indicated that performance decreased as the rationales became more similar, reinforcing the hypothesis that greater diversity among rationale encoders benefits multi-teacher KD.

Case Study: Rationale. In Table 3, rationale 1 in each case provided only a general statement, whereas rationale 2 and rationale 3 deliberately added more specific explanatory detail. For the instrument classification task (Case #1), rationale 1 merely said, "The Rajao is classified as a string instrument, while the Pellet drum falls under percussion," thereby restating the labels without explanation. By contrast, rationale 2 and rationale 3 specified the physical sound production mechanisms, "vibrating strings" versus "being struck/impact", that justified those labels. A similar pattern appeared in the coin-counting task (Case #2). Rationale 1 simply repeated the answer ("3 dimes amount to 30 cents"), whereas rationale 2 and rationale 3 spelled out the arithmetic: each dime equals 10 cents, so $3 \times 10 = 30$. These examples show that rationale 1 is intentionally brief and high-level, whereas rationales 2 and 3 are generated with access to all previous rationales and therefore enumerate progressively more specific information. By contrast, to examine the

Case #1	
Instruction	Identify which instrument is string or percussion: Pellet drum, Rajao
Output	Rajao is string, Pellet drum is percussion
GPT Output	The pellet drum is a percussion instrument, while the rajao is a string instrument.
Rationale 1	The Rajao is classified as a string instrument, while the Pellet drum falls under the category of percussion instruments.
Rationale 2	The Rajao produces sound through vibrating strings, whereas the Pellet drum creates sound by being struck, placing it in the percussion category.
Rationale 3	The Rajao, a string instrument, operates through the vibration of its strings, while the Pellet drum, categorized as percussion, generates sound through impact.
Case #2	
Instruction	How many cents do I have if I have 3 dimes?
Output	If you have 3 dimes, you have 30 cents.
GPT Output	If you have 3 dimes, you have 30 cents. Each dime is worth 10 cents, so 3 dimes are worth 3 x 10 = 30 cents.
Rationale 1	From the prompt, we can infer that having 3 dimes amounts to a total of 30 cents.
Rationale 2	Having 3 dimes means you possess 30 cents, as each dime contributes 10 cents to the total.
Rationale 3	Each dime is equivalent to 10 cents, so having 3 dimes results in a total of 30 cents when calculated.

Table 3: Rationales generated with the GPT-4o-mini model. Each prompt contained both the original instruction and its answer so that the model could produce a reasoned justification. For the second and third rationales, all previously generated rationales were included in the prompt to ensure that the new rationales did not overlap with earlier ones.

Method	**Teacher (1.7B)**	**Student (360M)**	**Add. Params.**	**Total Params. Used**
SFT	-	✓	-	360M
KD	✓	✓	-	2.06B
MiniLLM	✓	✓	-	2.06B
PromptKD	✓	✓	-	2.06B
ProxyKD	✓	✓	-	2.06B
GrayKD	-	✓	✓	610M

Table 4: Parameter comparison across different KD methods. The GrayKD entry refers to the triple variant that utilizes three encoders to process three rationales.

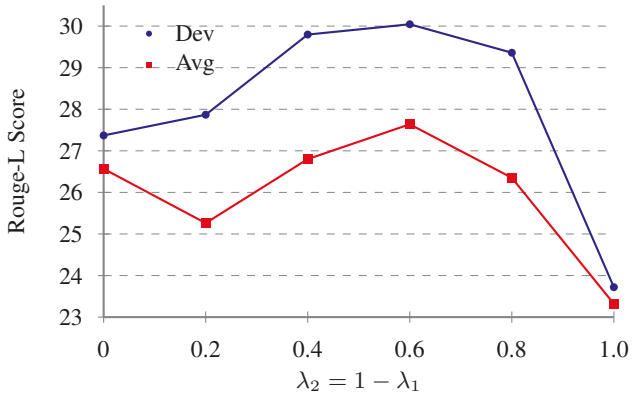

Figure 2: Performance comparison across different λ_3 values.

effect of diversity, we first generated rationale 1 and then deliberately prompted rationales 2 and 3 to resemble it, thereby artificially reducing diversity and enabling us to observe how performance varied with information richness. Detailed results are reported in the Appendix, and, for reproducibility, we also provide there the full GPT-4o-mini prompt used to generate these highly similar rationales.

Computational Cost Comparison. Conventional KD methods, including not only white-box KD approaches but also black-box KD methods based on proxy models, incurred substantial computational overhead. This overhead stems from their dependence on training the large-scale white-box teacher model. In our experiments, 1.7B-parameter SmolLM2 model was used as the white-box teacher, which required considerable additional computational resources for supervised fine-tuning prior to the student distillation stage. As shown in Table 4, existing KD methods trained the full 1.7B white-box teacher, resulting in a total parameter usage of over 2 billion. In contrast, GrayKD eliminated the need for extra teacher training and instead introduced a lightweight additional module. This results in a significantly lower total parameter count of just 610M, of which 360M belongs to the student model. Such efficiency makes GrayKD a highly scalable and resource-friendly approach.

Tunable Parameter. As shown in Figure 2, we performed a grid search by first defining $\lambda_2 = 1 - \lambda_1$ and then sweep-

Method	Dolly	Self-Inst	Vicuna	S-NI	UnNI	Avg.
GrayKD (Triple)	26.55	20.08	21.20	35.75	34.61	27.64
w/o Mix sample	26.53 **(-0.02)**	19.57 **(-0.51)**	19.41 **(-1.79)**	34.17 **(-1.58)**	34.55 **(-0.06)**	26.85 **(-0.79)**
w/o Masking	26.66 (+0.11)	19.40 **(-0.68)**	20.92 **(-0.28)**	35.76 (+0.01)	34.58 **(-0.03)**	27.46 **(-0.18)**
w/o Mix sample & Masking	26.51 **(-0.04)**	18.51 **(-1.57)**	18.62 **(-2.58)**	33.34 **(-2.41)**	34.70 (+0.09)	26.34 **(-1.30)**

Table 5: Ablation study on GrayKD with respect to (i) mix-sample strategy and (ii) masking strategy. Rouge-L scores were averaged over three random seeds (10,20,30).

Case #1	
Instruction	Did dinosaurs have lips?
SFT w/o KD	No, they did not.
KD	No, they did not.
MiniLLM	No, but they didn't have teeth either. They had a beak-like structure that was used for taking in food, much like the teeth of a carnivore.
PromptKD	No, dinosaurs did not have lips. However, they had a lot of similarities to humans in terms of how they used their lips in order to speak. They had a tongue that was cause for a lot of pressure when they spoke.
GrayKD	Yes, dinosaurs may not have had lips like we do, but they could have had a similar appearance. They could have had a thin skin covering their lips, which would have allowed them to expose a small amount of skin. This would have been similar to how modern-day birds have a similar appearance.
GrayKD (Sim1)	No, dinosaurs did not have lips. They were a group of animals that lived long before humans and had different body types, including those that may have had some features that resemble lips.

Table 6: Model responses to the Dolly instruction "Did dinosaurs have lips?" compared across SFT, KD, MiniLLM, PromptKD, GrayKD, and GrayKD (Sim1), where GrayKD (Sim1) denotes the variant trained with only minimal rationale information.

ing λ_2 over $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Development-set Rouge-L peaked at $\lambda_2 = 0.6$ ($\lambda_1 = 0.4$; 30.04), with a similar optimum on the held-out evaluation set ("Avg." curve). Consequently, we fixed $\lambda_1 = 0.4$ for all subsequent experiments. The sharp drop at $\lambda_2 = 1.0$ ($\lambda_1 = 0$) confirmed that relying solely on the distillation signal, without any teacher mode loss, hurt generalization, whereas too small a distillation contribution ($\lambda_1 \approx 1$) likewise under-utilized the richer teacher representations. Thus, $\lambda_1 = 0.4$ provided the best trade-off between preserving teacher knowledge and guiding the student, aligning with the observed performance gains. After the warm-up stage, we applied a mix-sample strategy that replaced 20% of ground-truth answers in D with fixed GPT-4o-mini outputs, and a masking strategy that randomly masked 15% of rationale tokens. Table 5 presents the results when each strategy was removed. As the table shows, Rouge-L scores dropped noticeably whenever a strategy was omitted. Detailed ablation results for the remaining tunable hyperparameters (e.g., N) are reported in the Appendix. Also, after the warm-up stage, the weight λ_1 balanced the teacher mode language-modeling loss ($\mathcal{L}_{\text{teacher}}$) against the distillation loss ($\mathcal{L}_{kd}$).

Qualitative Evaluation. In Table 6, addressing the Dolly Eval prompt *"Did dinosaurs have lips?"*, the SFT and KD each produced a terse "No," whereas MiniLLM committed the factual error that dinosaurs lacked teeth and PromptKD contradictorily claimed they "used their lips to speak." In contrast, GrayKD advanced a nuanced hypothesis that dinosaurs lacked human-like lips yet might have borne a thin, skin-covered margin and substantiated it by analogy with the keratinized beak tissue of modern birds, thus performing two-step reasoning of conditional affirmation plus anatomical justification. When trained with sparser rationales, GrayKD(Sim) regressed to a plain denial and lost evidence, comparison, and logical structure. This contrast highlighted how rationale richness governed specificity and coherence, and it demonstrated the qualitative superiority of GrayKD, which delivered the most persuasive explanatory answer.

Conclusion

In this paper, we proposed GrayKD, a novel framework that can effectively transfer text-level knowledge from a black-box teacher. By injecting rationales from the black-box LLM directly into the student architecture, the proposed method could approximate the black-box teacher's intrinsic knowledge in a more student-friendly manner. Our experimental results demonstrated that GrayKD achieved substantial performance gains over existing KD methods, while eliminating the need for a separate proxy teacher.

References

Achiam, J.; Adler, S.; Agarwal, S.; and et al., L. A. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Allal, L. B.; Lozhkov, A.; Bakouch, E.; Blázquez, G. M.; Penedo, G.; Tunstall, L.; Marafioti, A.; Kydlíček, H.; Lajarín, A. P.; Srivastav, V.; Lochner, J.; Fahlgren, C.; Nguyen, X.-S.; Fourrier, C.; Burtenshaw, B.; Larcher, H.; Zhao, H.; Zakka, C.; Morlon, M.; Raffel, C.; von Werra, L.; and Wolf, T. 2025. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model. *arXiv preprint arXiv:2502.02737*.

Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; et al. 2025. Gemini: A Family of Highly Capable Multimodal Models. Accessed July 31, 2025, arXiv:2312.11805.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; and Wenbin. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

Chen, H.; Chen, R.; Yi, Y.; Quan, X.; Li, C.; Yan, M.; and Zhang, J. 2024a. Knowledge Distillation of Black-Box Large Language Models. *arXiv preprint arXiv:2401.07013*.

Chen, H.; Quan, X.; Chen, H.; Yan, M.; and Zhang, J. 2024b. Knowledge Distillation for Closed-Source Language Models. In *Proc. CoRR*.

Conover, M.; Hayes, M.; Mathur, A.; Xie, J.; Wan, J.; Shah, S.; Ghodsi, A.; Wendell, P.; Zaharia, M.; and Xin, R. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

Cui, X.; Zhu, M.; Qin, Y.; Xie, L.; Zhou, W.; and Li, H. 2025a. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. In *In Proc. AAAI*, volume 39, 23724–23732.

Cui, X.; Zhu, M.; Qin, Y.; Xie, L.; Zhou, W.; and Li, H. 2025b. Multi-Level Optimal Transport for Universal Cross-Tokenizer Knowledge Distillation on Language Models. In *Proc. AAAI*.

Grattafiori, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2024. MiniLLM: Knowledge Distillation of Large Language Models. In *Proc. ICLR*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Honovich, O.; Scialom, T.; Levy, O.; and Schick, T. 2023. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In *Proc. ACL*, 14409–14428.

Hsieh, C.-Y.; Li, C.-L.; Yeh, C.-k.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.-Y.; and Pfister, T. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 8003–8017. Toronto, Canada: Association for Computational Linguistics.

Jin, Y.; Wang, J.; and Lin, D. 2023. Black-box Knowledge Distillation.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Kim, G.; Jang, D.; and Yang, E. 2024. PromptKD: Distilling Student-Friendly Knowledge for Generative Language Models via Prompt Tuning. In *Proc. EMNLP (Findings)*, 6266–6282.

Kim, Y.; and Rush, A. M. 2016. Sequence-Level Knowledge Distillation. In *Proc. EMNLP*.

Lee, H.; Kim, J.; and Lee, S. 2024. Mentor-KD: Making Small Language Models Better Multi-step Reasoners. In *Proc. EMNLP*, 17643–17658.

Li, Y.; Gu, Y.; Dong, L.; Wang, D.; Cheng, Y.; and Wei, F. 2025. Direct Preference Knowledge Distillation for Large Language Models. In *Proc. ICML*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out (ACL Workshop)*, 74–81.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Shrestha, S.; Kim, M.; Nepal, A.; Shrestha, A.; and Ross, K. 2025. Warm Up Before You Train: Unlocking General Reasoning in Resource-Constrained Settings. *arXiv preprint arXiv:2505.13718*.

Sun, Z.; Liu, Y.; Meng, F.; Chen, Y.; Xu, J.; and Zhou, J. 2025. Warmup-Distill: Bridge the Distribution Mismatch between Teacher and Student before Knowledge Distillation. *arXiv preprint arXiv:2502.11766*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Proc. NIPS*.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proc. ACL*.

Yang, C.; Zhu, Y.; Lu, W.; Wang, Y.; Chen, Q.; Gao, C.; Yan, B.; and Chen, Y. 2024. Survey on knowledge distillation for large language models: methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*.

Yoon, E.; Yoon, H. S.; Eom, S.; Han, G.; Nam, D. W.; Jo, D.; On, K.-W.; Hasegawa-Johnson, M. A.; Kim, S.; and Yoo, C. D. 2024. TLCR: Token-Level Continuous Reward for Fine-grained Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2407.16574*.

Yoon, J. W.; Ahn, S.; Lee, H.; Kim, M.; Kim, S. M.; and Kim, N. S. 2023a. EM-Network: Oracle Guided Self-distillation for Sequence Learning. In *Proc. ICML*.

Yoon, J. W.; Kim, H. Y.; Lee, H.; Ahn, S.; and Kim, N. S. 2023b. Oracle Teacher: Leveraging Target Information for Better Knowledge Distillation of CTC Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 2974–2987.

Zhang, S.; Zhang, X.; Sun, Z.; Chen, Y.; and Xu, J. 2024. Dual-space knowledge distillation for large language models. *arXiv preprint arXiv:2406.17328*.