Multi-Level Optimal Transport for Universal Cross-Tokenizer Knowledge Distillation on Language Models

Xiao Cui[1*], **Mo Zhu**[2*], **Yulei Qin**[3], **Liang Xie**[2,4], **Wengang Zhou**[1], **Houqiang Li**[1†]
[1]University of Science and Technology of China
[2]Zhejiang University
[3]Tencent YouTu Lab
[4]Zhejiang University of Technology
cuixiao@mail.ustc.edu.cn, {zhwg,lihq}@ustc.edu.cn, mozhu@zju.edu.cn, yuleiqin@tencent.com, lilydedbb@gmail.com

Abstract

Knowledge distillation (KD) has become a prevalent technique for compressing large language models (LLMs). Existing KD methods are constrained by the need for identical tokenizers (i.e., vocabularies) between teacher and student models, limiting their versatility in handling LLMs of different architecture families. In this paper, we introduce the Multi-Level Optimal Transport (MultiLevelOT), a novel approach that advances the optimal transport for universal cross-tokenizer knowledge distillation. Our method aligns the logit distributions of the teacher and the student at both token and sequence levels using diverse cost matrices, eliminating the need for dimensional or token-by-token correspondence. At the token level, MultiLevelOT integrates both global and local information by jointly optimizing all tokens within a sequence to enhance robustness. At the sequence level, we efficiently capture complex distribution structures of logits via the Sinkhorn distance, which approximates the Wasserstein distance for divergence measures. Extensive experiments on tasks such as extractive QA, generative QA, and summarization demonstrate that the MultiLevelOT outperforms state-of-the-art cross-tokenizer KD methods under various settings. Our approach is robust to different student and teacher models across model families, architectures, and parameter sizes.

Introduction

Large language models (LLMs) such as LLaMA (Touvron et al. 2023a,b; Meta 2024), Mistral (Jiang et al. 2023) and Qwen (Bai et al. 2023; Yang et al. 2024) have set state-of-the-art (SOTA) records on various natural language processing (NLP) tasks. While the scaling laws of LLMs have driven the development of larger models with billions of parameters, their substantial sizes pose significant challenges to deployment under resource-constrained environments. To address this issue, knowledge distillation (KD) has emerged as a cost-efficient technique for its ability to distill smaller models that maintain competitive performance.

Cross-tokenizer knowledge distillation (CTKD) refers to the process of transferring knowledge between models that use different tokenizers (see Figure 1). It is crucial to ensure compatibility for applications such as multi-teacher knowledge transfer, where the student model learns from multiple teacher models with potentially different tokenization schemes. However, most existing KD methods rely on divergence measures such as Kullback–Leibler (KL) divergence (Hinton, Vinyals, and Dean 2015; Park, Kim, and Yang 2021; Agarwal et al. 2024; Wu et al. 2024), reverse KL (RKL) divergence (Tu et al. 2020; Gu et al. 2023b), and Jensen–Shannon (JS) divergence (Wen et al. 2023; Yin et al. 2020; Fang et al. 2021). These measures require a strict point-by-point correspondence across dimensions between the student and teacher, necessitating the use of the same tokenizer and consistent vocabularies, which limits their applicability when different tokenizers are involved.

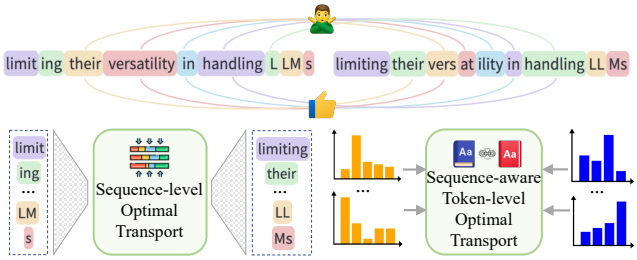

Figure 1: An illustration of vocabulary mismatch resulting from cross-tokenizer discrepancies. Unlike strict token-wise distillation methods that may lead to token misalignment, we employ sequence-level and sequence-aware token-level optimal transport to facilitate effective knowledge transfer.

Very few studies notice such deficiency in directly applying existing KD techniques on LLMs, for the simple reason that most KD methods are developed for few mainstream open-source models. ULD (Boizard et al. 2024), the first attempt ever to tackle this issue, aligns the distributions of individual tokens between the teacher and the student using token-wise optimal transport (OT). However, ULD focuses solely on the internal information of individual tokens without considering the global context for robust matching. Additionally, its reliance on zero padding introduces noise and hinders the effective use of logarithmic cost matrices. DSKD (Zhang et al. 2024b), another token-wise alignment

*These authors contributed equally.
†Corresponding author: Houqiang Li.

method, tries to transform the hidden states of one model to the space of another one bidirectionally via learnable projectors. Despite its efforts in alignment for a unified output space, DSKD fails to effectively leverage the distribution information as the transformed distribution often exhibits low accuracy. Also, although these methods avoid strict dimensional correspondence, they assume a rigid token-by-token correspondence, which is often not the case in practice.

To address these shortcomings, we propose the Multi-Level Optimal Transport (MultiLevelOT) for cross-tokenizer knowledge distillation on LLMs. Our method comprehensively measures the discrepancy between teacher and student logit distributions by calculating the optimal transport distance both within and across tokens in each sequence. Such a dual-level approach ensures that both token-level and sequence-level relationships are incorporated into the distillation process, effectively eliminating the need for dimensional or token-by-token correspondence.

At the token level, we jointly optimize all tokens within a sequence by minimizing token-level discrepancies within the context of the entire sequence. This is achieved by applying a sequence-level ranking process, which enables the same optimal transport plan for all tokens and effectively selects the important dimensions. To eliminate noise from redundant dimensions, we truncate the logits, focusing only on the most impactful logit dimensions for each sequence. This truncation ensures that the teacher and student logits share a common support size, making each dimension meaningful and applicable for a logarithmic form cost matrix. To capture both the fine-grained, token-wise nuances and the holistic, sequence-scale context view, we employ two types of cost matrices: one in the form of absolute difference and the other in the form of logarithm-based likelihood difference. The absolute difference cost matrix captures the direct discrepancies in logits, providing a straightforward and interpretable measure of distance. Conversely, the logarithmic cost matrix accounts for the relative differences to offer a more nuanced and scalable measure. It is particularly effective in handling logits with a wide range of magnitudes.

At the sequence level, which has not been considered in previous studies, we utilize token-to-token OT distances to construct the sequence-level cost matrix. Since optimal transport automatically finds the corresponding relationships between tokens, this is particularly crucial for addressing token order misalignment caused by varying tokenization of long words across different tokenizers. Unlike token-level transport, which deals with individual logit values, sequence-level transport requires calculating the optimal transport between vectors of tokens. Given the computational intensity of directly computing the Wasserstein distance for this purpose, we employ the Sinkhorn distance as an efficient approximation. This approach retains the benefits of the Wasserstein distance while significantly reducing computational complexity. Importantly, we achieve all the improvements without introducing additional modules or modifying output formats specific to NLP tasks.

Extensive experiments are conducted in view of 1) **comparability**, 2) **validity**, and 3) **generalizability**. For comparability, we test our method on different tasks under both labeled and unlabeled distillation settings. Our method consistently outperforms the state-of-the-art CTKD methods. For validity, we provide a comprehensive analysis through ablation studies and hyper-parameter tuning, which corroborate the effectiveness of each component. For generalizability, the proposed method is validated on different students across families, architectures, and sizes. We also experiment with diverse teachers to demonstrate its robustness across various model choices. In summary, our contributions are:

- We propose the MultiLevelOT, a cross-tokenizer knowledge distillation approach that leverages both sequence-aware token-level and sequence-level optimal transport for comprehensive distribution matching.
- We enhance the robustness of our method by jointly optimizing all tokens and using varied cost matrices, effectively capturing both global and local information.
- We demonstrate the superiority of MultiLevelOT over existing methods through extensive experiments, validating its comparability, validity, and generalizability.

Related Work

Knowledge Distillation

Knowledge distillation (KD) is proposed to transfer the intrinsic knowledge from a teacher model to a student model by approximating the teacher's soft targets, such as output logits and intermediate representations. Cross-Tokenizer KD extends this traditional framework to scenarios involving different tokenizers, each with distinct vocabularies, which is crucial for LLM distillation. Various KD methods have been explored, ranging from logit-based distillation to representation-based distillation. These methods typically employ divergence measures like KL divergence (Hinton, Vinyals, and Dean 2015; Agarwal et al. 2024; Wu et al. 2024; Zhou, Xu, and McAuley 2022; Zhang et al. 2023; Liu et al. 2022), RKL (Tu et al. 2020; Gu et al. 2023b; Ko et al. 2024), and JS divergence (Wen et al. 2023; Yin et al. 2020; Fang et al. 2021). These measures compute discrepancies on each dimension, requiring a one-to-one correspondence between teacher and student logit dimensions. SinKD (Cui et al. 2024b,a) addresses the limitations of these traditional measures by using the Sinkhorn distance. However, its approach still requires dimensional correspondence in the cost matrix. In cross-tokenizer distillation, such dimensional correspondence is absent, making these methods inapplicable.

To overcome this challenge, both ULD (Boizard et al. 2024) and DSKD (Zhang et al. 2024b) propose promising solutions for token-wise alignment. ULD measures token-wise OT distance between the logits of the student and teacher models, eliminating the dependency on dimensional correspondence. DSKD attempts to transform the hidden states of one model to that of another by training projectors, but the transformed distribution often exhibits low accuracy. Comparatively, the proposed method differs in the following aspects: 1) ULD only considers local information while neglecting global distributional properties. Its padding approach, more like an ad-hoc brutal tactic, limits it to a singular cost matrix. In contrast, we stem from the token-level and sequence-level perspectives and deduce different

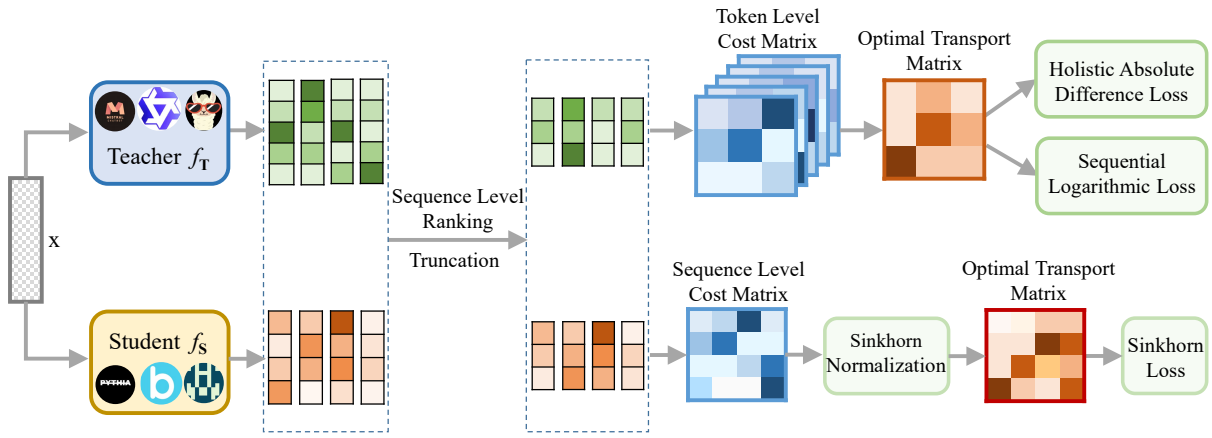

Figure 2: Illustration of our pipeline. MultiLevelOT computes sequence-aware token-level and sequence-level optimal transport distances between the output logits of the teacher and student models. This approach effectively transfers local and global information within the logits distribution, accommodating vocabulary differences and enabling cross-tokenizer distillation.

forms of cost matrices for lexical and semantic alignment. 2) While DSKD relies on traditional divergence measures, which suffer from issues like mode-averaging and mode-collapsing (Cui et al. 2024b), we employ the Sinkhorn distance to fully capture the geometric characteristics of logit distributions. In addition, we do not explicitly enforce cross-model space mapping because such dual-space projection lacks semantic interpretability and thereafter hinders sequence comprehension. 3) Both ULD and DSKD assume a rigid token-by-token correspondence, which is often impractical. Our approach uses sequence-level OT, which automatically identifies corresponding relationships between tokens, thereby eliminating the need for strict token correspondence.

Optimal Transport

Optimal transport (OT) theory offers a robust mathematical framework for comparing probability distributions by calculating the minimal cost required to transform one distribution into another. The Wasserstein distance, a pivotal concept in OT, quantifies this cost and excels in capturing the geometric structure of distributions (Villani and Villani 2009; Zhang, Liu, and Tao 2021). This metric has been instrumental across various domains, including causal discovery (Wei et al. 2022; Weilin et al. 2023), image generation (Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017; Peyré, Cuturi et al. 2019), unsupervised learning (Gu et al. 2023a; Chen et al. 2022; He et al. 2022), and reinforcement learning (Du et al. 2023; Lan et al. 2023; Zhang et al. 2024a).

While the Wasserstein distance may be simplified in some low-dimensional cases, it can be computationally intensive in other scenarios. To address this, the Sinkhorn distance has been proposed as an approximation, which introduces an entropy regularization term to the OT problem, making it more tractable (Cuturi 2013). This approach has demonstrated success in diverse applications such as machine translation (Li, Unanue, and Piccardi 2023), domain adaptation (Nguyen and Luu 2022; Xu et al. 2023), classification (Liu et al. 2023), and teacher model selection (Lu, Ye, and Zhan 2022; Bhardwaj, Vaidya, and Poria 2021).

Our approach employs both token-level and sequence-level OT for cross-tokenizer knowledge distillation. This dual-level OT captures global and local information, enhancing geometry information transfer and model efficacy.

Methods

Problem Statement

Given a sample $\mathbf{x}$ and its ground-truth label $\mathbf{y}$, the output logits with softmax activation σ_τ from the teacher $f_\mathbf{T}$ and the student $f_\mathbf{S}$ are $\mathbf{t} \in \mathbb{R}^{T\times m}$ and $\mathbf{s} \in \mathbb{R}^{T\times n}$, respectively:

$$\mathbf{t} = \sigma_\tau(f_\mathbf{T}(\mathbf{x})), \quad \mathbf{s} = \sigma_\tau(f_\mathbf{S}(\mathbf{x})), \tag{1}$$

where τ represents the temperature parameter, m and n denote the dimensions of the teacher and student output vocabularies, respectively, and T is the total number of tokens in the generated sequence. We denote the i-th dimension of the teacher and student logits for the t-th token as $\mathbf{t}_i(t)$ and $\mathbf{s}_i(t)$, respectively. Our objective is to minimize the optimal transport distance between the distributions of the teacher's and student's outputs for knowledge transfer. In scenarios where the ground-truth label is unavailable, we use teacher-generated text as a substitute.

Reconstructing optimal transport in ULD

ULD (Boizard et al. 2024) leverages OT to address the challenge of cross-tokenizer knowledge distillation. To ensure equal support size between the teacher and student distribution spaces, ULD pads the smaller vocabulary with zero values, matching the larger size $\max(m, n)$. The ULD loss is then computed by summing the token-wise Wasserstein distances. The OT distance for the t-th token is defined as:

$$\min_{\mathbf{P}(t)} \sum_{i=1}^{\max(m,n)} \sum_{j=1}^{\max(m,n)} \mathbf{P}_{ij}(t)\mathbf{C}_{ij}(t), \tag{2}$$

where $\mathbf{P}$ is the optimal transport matrix and $\mathbf{C}$ is the cost matrix. ULD asserts that each transport cost is equal to 1 and applies the following constraints on $\mathbf{P}$:

$$\sum_i \mathbf{P}_{ij}(t) = \mathbf{s}_j(t) \quad \forall j, t, \sum_j \mathbf{P}_{ij}(t) = \mathbf{t}_i(t) \quad \forall i, t. \quad (3)$$

However, the original formulation lacks flexibility. We propose a more adaptable reformulation by setting $\mathbf{C}_{ij} = |\mathbf{t}_i(t) - \mathbf{s}_j(t)|$ and using these constraints:

$$\sum_i \mathbf{P}_{ij}(t) = 1 \quad \forall j, t, \qquad \sum_j \mathbf{P}_{ij}(t) = 1 \quad \forall i, t. \quad (4)$$

Both formulations yield the same optimal transport distance:

$$\mathcal{L}_{\text{ULD}} = \sum_{t=1}^{T} \sum_{i=1}^{\max(m,n)} |\mathbf{t}_{\text{TR},i}(t) - \mathbf{s}_{\text{TR},i}(t)|, \quad (5)$$

where $\mathbf{s}_{\text{TR}}(t)$ and $\mathbf{t}_{\text{TR}}(t)$ are the token-wise ranked logits of the student and teacher, respectively:

$$\begin{aligned} \mathbf{s}_{\text{TR}}(t) &= \mathbf{s}\left[\text{argsort}\left(\mathbf{s}(t), \text{descending}\right)\right] \\ \mathbf{t}_{\text{TR}}(t) &= \mathbf{t}\left[\text{argsort}\left(\mathbf{t}(t), \text{descending}\right)\right]. \end{aligned} \quad (6)$$

By reconstructing equivalent optimal transport problems, we can design various cost matrices and extend token-wise optimal transport distance to multi-level optimal transport.

Multi-Level Optimal Transport

Instead of considering each token independently, our method jointly optimizes all tokens within a sequence through sequence-aware multi-level OT, effectively aligning the distributions of teacher and student output logits. The primary objective is to minimize the sum of token-level and sequence-level costs using an optimal transport plan $\mathbf{P}$:

$$\min_{\mathbf{P}} \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{P}_{ij} \sum_{t=1}^{T} \mathbf{C}_{ij}^{tok}(t) + \min_{\mathbf{P}} \sum_{i=1}^{T}\sum_{j=1}^{T} \mathbf{P}_{ij}\mathbf{C}_{ij}^{seq}, \quad (7)$$

where $\mathbf{C}^{tok}$ and $\mathbf{C}^{seq}$ represent the token-level and sequence-level cost matrices, respectively. Specific mathematical formulations will be detailed in subsequent paragraphs. The optimization is subject to the constraints:

$$\sum_i \mathbf{P}_{ij} = 1 \quad \forall j, \qquad \sum_j \mathbf{P}_{ij} = 1 \quad \forall i. \quad (8)$$

We model the token-level cost using both absolute difference and logarithmic forms, while the sequence-level cost is captured through the optimal transport distance between tokens. For token-level alignment, our optimization strategy integrates both global and local information by considering the entire sequence within the optimal transport process. The full pipeline is illustrated in Figure 2.

Holistic Absolute Difference Loss We define the first token-level cost matrix $\mathbf{C}_{ij}^{tok}(t)$ using the absolute difference between logits: $\mathbf{C}_{ij}^{tok}(t) = |\mathbf{t}_i(t) - \mathbf{s}_j(t)|$, so that the Wasserstein distance can be obtained by solving this optimization problem:

$$\min_{\mathbf{P}} \sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{P}_{ij} \left|\mathbf{t}_i(t) - \mathbf{s}_j(t)\right|. \quad (9)$$

While ULD employs a separate optimal transport matrix for each token, leading to inconsistent dimensional relationship, our approach ensures robustness by performing sequence-level ranking across all logits within a sequence. This allows us to use a single optimal transport matrix for all tokens, ensuring consistent dimensional ordering within each token t. Our sequence-level ranking process is defined as follow:

$$\mathbf{t}_{\text{SR}} = \mathbf{t}\left[\text{argsort}\left(\sum_{t=1}^{T} \mathbf{t}(t), \text{descending}\right)\right], \quad \mathbf{s}_{\text{SR}} = \mathbf{Q}\mathbf{s}, \quad (10)$$

where $\mathbf{Q} = \mathbf{Q}^*$ is a permutation matrix used to match the dimensions of $\mathbf{s}$ with the corresponding dimensions of $\mathbf{t}_{\text{SR}}$ at the sequence level, satisfying:

$$\mathbf{Q}^* = \underset{\mathbf{Q}}{\text{argmin}} \sum_{t=1}^{T}\sum_{i=1}^{m} |\mathbf{t}_{\text{SR},i}(t) - [\mathbf{Q}\mathbf{s}(t)]_i|. \quad (11)$$

To ensure the consistency of the support size, allow for a logarithmic cost matrix, prevent mode-averaging, and reduce noise from unlikely words, we conduct the top-k truncation as follows:

$$\mathbf{s}_{\text{SR,Tr}}(t) = \mathbf{s}_{\text{SR}}(t)[:k], \quad \mathbf{t}_{\text{SR,Tr}}(t) = \mathbf{t}_{\text{SR}}(t)[:k], \quad (12)$$

where $[:k]$ denotes the slicing operation for choosing the top-k elements of the vector. Then the optimization problem can be reformulated as:

$$\min_{\mathbf{P}} \sum_{t=1}^{T}\sum_{i=1}^{k}\sum_{j=1}^{k} \mathbf{P}_{ij} \left|\mathbf{t}_{\text{SR,Tr},i}(t) - \mathbf{s}_{\text{SR,Tr},j}(t)\right|. \quad (13)$$

The optimal transport matrix to the above Eq. (13) is $\mathbf{P}^* = \mathbf{P}^{\text{HAD}}$, where $\mathbf{P}_{ij}^{\text{HAD}}$ is 1 only when $i = j$, and 0 otherwise. The absolute difference loss, representing the solution to this optimization problem, is then computed as:

$$\mathcal{L}_{\text{HAD}} = \sum_{t=1}^{T}\sum_{i=1}^{k} \left|\mathbf{t}_{\text{SR,Tr},i}(t) - \mathbf{s}_{\text{SR,Tr},i}(t)\right|. \quad (14)$$

In the following text, all instances of $\mathbf{t}^k$ and $\mathbf{s}^k$ refer to $\mathbf{t}_{\text{SR,Tr}}$ and $\mathbf{s}_{\text{SR,Tr}}$, respectively.

Sequential Logarithmic Loss For the token-level cost matrix, in addition to the absolute difference, we also incorporate a logarithmic form: $\mathbf{C}_{ij}^{tok}(t) = -\mathbf{t}_i(t)\log \mathbf{s}_j(t)$. We apply the previously mentioned top-k truncation, which ensures that no zero-value elements are present in the student logits, thus making this cost matrix meaningful and effective. Given that each dimension is equally important, the optimization problem for computing the Wasserstein distance can be formulated in a sequence-level ranked order:

$$\min_{\mathbf{P}} \sum_{t=1}^{T}\sum_{i=1}^{k}\sum_{j=1}^{k} -\mathbf{P}_{ij}\mathbf{t}_i^k(t)\log \mathbf{s}_i^k(t). \quad (15)$$

The optimization objective is minimized by the sequential transfer between logit dimensions, making the optimal transport matrix $\mathbf{P}^{\text{SL}}$ equivalent to $\mathbf{P}^{\text{HAD}}$. Consequently, the loss function is defined as:

$$\mathcal{L}_{\text{SL}} = -\sum_{t=1}^{T}\sum_{i=1}^{k} \mathbf{t}_i^k(t)\log \mathbf{s}_i^k(t). \tag{16}$$

Sinkhorn Distance Loss We employ the optimal transport distance between tokens to measure pairwise differences between the i-th and j-th tokens in a sequence, constructing the sequence-level cost matrix $\mathbf{C} \in \mathbb{R}^{T\times T}$ with entries $\mathbf{C}_{ij}^{seq} = \sum_{l=1}^{k}\sum_{q=1}^{k} \mathbf{P}_{lq}^{\text{HAD}} \left|\mathbf{t}_l^k(i) - \mathbf{s}_q^k(j)\right|$. Following SinKD (Cui et al. 2024b,a), we use Sinkhorn distance as an efficient approximation for Wasserstein distance, retaining its benefits while significantly reducing computational costs for online distillation. The Sinkhorn distance is based on the relaxed formulation of an OT plan with entropy regularization. The OT plan $\mathbf{P}^{\lambda}$ is obtained by minimizing:

$$\mathbf{P}^{\lambda} = \underset{\mathbf{P}}{\text{argmin}} \sum_{i=1}^{T}\sum_{j=1}^{T} \mathbf{P}_{ij}\mathbf{C}_{ij} - \lambda h\left(\mathbf{P}\right), \tag{17}$$

where $h(\mathbf{P})$ is the entropy of the matrix $\mathbf{P}$, $\lambda > 0$ is the entropy regularization weight. To solve this iteratively, we construct the kernel matrix $\mathbf{K}^0 \in \mathbb{R}^{T\times T}$ by applying the Gaussian kernel to $\mathbf{C}$ with the parameter λ:

$$\mathbf{K}^0 = \exp\left(-\frac{\mathbf{C}}{\lambda}\right). \tag{18}$$

The OT plan $\mathbf{P}^{\lambda}$ is then derived through sequence-level Sinkhorn normalization, using iterative updates on $\mathbf{K}$:

$$\widehat{\mathbf{K}}^i \leftarrow \mathbf{K}^{i-1} \oslash \left(\mathbf{K}^{i-1}\mathbf{1}_b\mathbf{1}_b^{\top}\right), \mathbf{K}^i \leftarrow \widehat{\mathbf{K}}^i \oslash \left(\mathbf{1}_b\mathbf{1}_b^{\top}\widehat{\mathbf{K}}^i\right). \tag{19}$$

For simplicity, irrelevant constants are excluded from the equations. After a pre-determined number of iterations N, the OT matrix is obtained as $\mathbf{P}^{\lambda} = \mathbf{K}^N$. The sequence-level optimal transport distance loss is then computed as:

$$\mathcal{L}_{\text{SD}} = \left\langle \mathbf{P}^{\lambda}, \mathbf{C}\right\rangle = \sum_{i=1}^{T}\sum_{j=1}^{T} \mathbf{K}_{i,j}^{N}\mathbf{C}_{i,j}. \tag{20}$$

Total Loss We combine the Cross-Entropy (CE) loss with the weighted holistic absolute difference loss, sequential logarithmic loss, and Sinkhorn distance loss for distillation. For a sequence of T tokens, the total loss is defined as:

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_{\text{CE}}\left(\mathbf{y}(t), \mathbf{s}(t)\right) + \alpha(\mathcal{L}_{\text{HAD}} + \beta\mathcal{L}_{\text{SL}} + \gamma\mathcal{L}_{\text{SD}}), \tag{21}$$

where α, β and γ are weights for each loss component.

Experiments

Experimental Settings

Datasets. We evaluate our method on three representative tasks: an extractive QA task (QED) (Lamm et al. 2021), a generative QA task (FairytaleQA) (Xu et al. 2022), and a summarization task (DIALOGSum) (Chen et al. 2021). For evaluation, we use the F1 score for the QED and the Rouge-LSum (Giarelis, Mastrokostas, and Karacapilidis 2023) for others. More details are given in the appendix.

Model	Method	QED (F1)	FairytaleQA (Rouge-LSum)	DIALOGSum (Rouge-LSum)
LLaMA2-7B	Few-Shot	61.68	50.90	37.75
OPT-350M	Origin	12.46	11.16	14.02
	SFT	55.71	46.04	35.59
	SeqKD	49.61	39.19	30.71
	MinED	56.03	46.11	35.82
	ULD	56.76	45.82	36.05
	Ours	**58.97**	**46.96**	**37.61**
Pythia-410M	Origin	22.87	15.14	4.41
	SFT	59.03	47.23	36.06
	SeqKD	51.12	39.78	31.57
	MinED	59.21	47.31	35.97
	ULD	59.71	47.81	36.07
	Ours	**61.79**	**49.10**	**37.45**
Bloomz-560M	Origin	47.67	43.47	11.82
	SFT	60.48	49.07	36.52
	SeqKD	52.33	45.68	31.83
	MinED	60.52	49.10	36.39
	ULD	61.22	49.87	36.40
	Ours	**62.58**	**50.94**	**37.68**
Average	Origin	27.67	23.25	10.08
	SFT	58.41	47.45	36.05
	SeqKD	50.99	41.55	31.37
	MinED	58.58	47.47	36.06
	ULD	59.30	47.83	36.17
	Ours	**60.99**	**49.00**	**37.58**

Table 1: Performance of the students in labeled distillation. Both the teacher and ground-truth provide supervision.

Model	Method	QED (F1)	FairytaleQA (Rouge-LSum)	DIALOGSum (Rouge-LSum)
LLaMA2-7B	Few-Shot	61.68	50.90	37.75
OPT-350M	Origin	12.46	11.16	14.02
	Raw Text	49.61	39.19	30.71
	ULD	50.71	39.86	32.03
	Ours	**51.96**	**40.68**	**36.88**
Pythia-410M	Origin	22.87	15.14	4.41
	Raw Text	51.12	39.78	31.57
	ULD	52.09	40.69	34.15
	Ours	**53.56**	**41.28**	**36.52**
Bloomz-560M	Origin	47.67	43.47	11.82
	Raw Text	52.33	45.68	31.83
	ULD	53.02	46.72	34.21
	Ours	**54.15**	**47.88**	**37.10**
Average	Origin	27.67	23.25	10.08
	Raw Text	50.99	41.55	31.37
	ULD	51.94	42.42	33.46
	Ours	**53.22**	**43.28**	**36.83**

Table 2: Performance of the students in unlabeled distillation. The ground-truth is unavailable for supervision.

Implementation details. We use four advanced teacher models: LLaMA2 7B Chat (Touvron et al. 2023b), Mistral3 7B Instruct (Jiang et al. 2023), Qwen 7B Chat (Bai et al. 2023) and LLaMA3 8B Instruct (Meta 2024). These models are chosen for their proficiency in few-shot learning and their unique vocabulary coverage (Brown et al. 2020). For student models, we use a range of LLMs from various families and architectures, including OPT 350M (Zhang et al. 2022), Pythia 160M, Pythia 410M, Pythia 1B (Biderman et al. 2023), Bloomz 560M (Muennighoff et al. 2023), and mT0 300M (Muennighoff et al. 2023), initializing them with their pretrained weights. Following ULD (Boizard et al. 2024), we set the learning rate $lr = 1e-6$, $\alpha = 0.15$, $\beta = 0.1$. Additionally, we empirically set $\gamma = 0.1$, $\tau_{\text{SL}} = 1$, $\tau_{\text{SD}} = 2$, $\lambda = 0.1$, $N = 20$ and $k = 50$. Discussions on the effects of key factors N, and k are presented later. Although further tuning may enhance performance, we maintain a consistent set of hyper-parameters across all tasks to underscore the robustness of our approach.

Baselines. Our experiments involve two settings: labeled distillation and unlabeled distillation. Labeled distillation, commonly used in most distillation studies, involves supervision with ground-truth labels. In contrast, unlabeled distillation relies solely on the generated texts from the teacher as pseudo-targets (Boizard et al. 2024). For labeled distillation, we compare our approach against the following baselines: Supervised Fine-Tuning (SFT), Sequence-level KD (SeqKD) (Kim and Rush 2016), MinED (Wan et al. 2024), and ULD (Boizard et al. 2024). SeqKD can be interpreted as a form of supervised fine-tuning using the teacher's outputs, deriving knowledge exclusively from the teacher model. MinED, which can align the logits using dynamic programming , is also included in our comparison. For unlabeled distillation, we follow the ULD to adopt the same baselines. In both settings, we use the official code and default hyper-parameters for each baseline to ensure a fair comparison. We exclude DSKD (Zhang et al. 2024b) from our comparison as it introduces additional modules whose increased learnable parameters may cause unfair comparison.

Results and Discussions

Comparison with SOTA. Results on labeled distillation and unlabeled distillation are presented in Table 1 and Table 2, respectively. MultiLevelOT consistently outperforms all baseline methods across all datasets and student models.

CE	AD	TR	SR	Tr	SL	SD	OPT	Pythia	Bloomz
✓							55.71	59.03	60.48
✓	✓	✓					56.76	59.71	61.22
✓	✓		✓				58.02	60.18	61.56
✓	✓		✓	✓			58.01	60.22	61.58
✓	✓		✓	✓	✓		58.17	61.10	61.87
✓	✓		✓	✓		✓	58.15	61.20	61.90
✓			✓	✓	✓	✓	58.23	61.17	61.80
✓	✓		✓	✓	✓	✓	**58.97**	**61.79**	**62.58**

Table 3: Ablation Study on QED across three students.

	OPT	Pythia	Bloomz
w/o SD loss	58.17	61.10	61.87
w token-level SD loss	58.32	61.22	61.95
w sequence-level SD loss	**58.97**	**61.79**	**62.58**

Table 4: Comparison of token-level and sequence-level Sinkhorn distance loss on QED across three students.

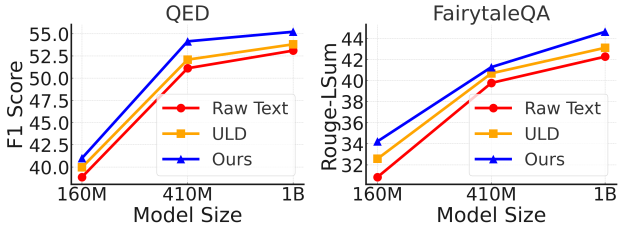

Figure 3: Performance at different student scales (Pythia 160M, 410M, and 1B) on QED and FairytaleQA.

Notably, compared with ULD (Boizard et al. 2024), MultiLevelOT reduces the performance gap between the student and the teacher by over 71% in the QED task on labeled distillation. This improvement highlights the effectiveness of MultiLevelOT in bridging the performance gap by transferring sequence-level and sequence-aware token-level knowledge from the teacher to the student. The superior performance of our approach is also attributed to the well-rounded design of the cost matrix. By employing diverse cost matrices, we facilitate effective geometry distribution information extraction and enhance the knowledge transfer process.

Each components plays its role in MultiLevelOT. The ablation study on the QED task, as shown in Table 3, demonstrates the critical role of each component in the MultiLevelOT framework. The baseline model, utilizing only cross-entropy (CE) loss, corresponds to the standard SFT. Adding the absolute difference (AD) and token-wise ranking (TR), as in ULD, provides a reference for improvement. However, the key advancements come from our proposed components. Integrating sequence-level ranking (SR) and truncation (Tr) with AD results in the **Holistic Absolute Difference Loss**, which shows significant gains by capturing both global and local geometrical information. Incorporating the **Sequential Logarithmic Loss (SL)** further boosts performance, highlighting the value of various cost matrices in capturing different aspects of the distribution. Finally, in-

Method	**QED (F1)**	**FairtaleQA (Rouge-LSUM)**	**DIALOGSum (Rouge-LSUM)**
Raw Labels	34.96	29.73	28.88
ULD	37.25	31.52	30.04
Ours	**41.37**	**34.01**	**33.01**

Table 5: Generalizability of MultiLevelOT in student architecture. Teacher: LLaMA, student: mT0-300M.

Method	LLaMA2	Mistral3	Qwen	LLaMA3
Teacher	61.68	64.03	62.16	65.96
Raw Text	49.61	51.24	51.21	51.91
ULD	50.71	52.08	52.89	52.81
Ours	**51.96**	**52.96**	**53.99**	**54.38**

Table 6: Generalizability of MultiLevelOT across different teacher models on QED. Student : OPT-350M.

Number of Iterations N	5	10	20	50	100
OPT-350M	58.26	58.52	58.97	59.02	58.99
Pythia-410M	60.56	61.24	61.79	61.76	61.78

Table 7: Effect of N on QED.

tegrating the **Sinkhorn Distance Loss (SD)** results in the best performance, underlining the necessity of sequence-level knowledge for effective knowledge transfer.

Sequence-level Sinkhorn distance excels token-level Sinkhorn distance. Table 4 demonstrates that sequence-level Sinkhorn distance outperforms token-level distance across all student models. The sequence-level approach captures the geometric properties of logit distributions more comprehensively, providing a robust framework for understanding global contextual relationships among tokens. In contrast, while token-level distance, akin to a Holistic Absolute Difference Loss with an added entropy term, enhances robustness and mitigates sparsity, it fails to fully encapsulate the overarching patterns of entire sentences.

MultiLevelOT generalizes well on student LLMs across scales. We evaluate the impact of student LLMs' sizes on the efficacy of MultiLevelOT through a detailed analysis in an unlabeled distillation context. Using two diverse tasks, QED (Lamm et al. 2021) and FairytaleQA (Xu et al. 2022), as illustrated in Figure 3, we observe that MultiLevelOT consistently enhances the performance of student models across various scales. This improvement substantiates MultiLevelOT's advanced capability to effectively utilize optimal transport for knowledge distillation, clearly outperforming the ULD method (Boizard et al. 2024).

Generalization of MultiLevelOT across student architectures. Since MultiLevelOT relies solely on logits in the distillation process, it can be applied to any architecture. In addition to decoder-only models, we also test it on the encoder-decoder model mT0 (Muennighoff et al. 2023). Results in Table 5 reveal significant performance enhancements, underscoring MultiLevelOT's flexibility and effectiveness across various architectural frameworks.

Truncation Threshold k	5	20	50	100	1000
OPT-350M	58.54	58.84	58.97	58.78	58.42
Pythia-410M	61.42	61.50	61.79	61.40	61.32

Table 8: Effect of k on QED.

Generalization of MultiLevelOT across teacher LLMs. An extensive evaluation of MultiLevelOT's performance with varying teacher LLMs is conducted, employing models including LLaMA2 7B Chat (Touvron et al. 2023a), Mistral 7B Instruct (Jiang et al. 2023), Qwen 7B (Bai et al. 2023), and LLaMA3 8B Chat (Meta 2024). As shown in Table 6, MultiLevelOT consistently outshines its counterparts. This highlights MultiLevelOT's robust capacity to leverage the distinct advantages of various teacher models.

N **as the number of Sinkhorn iterations.** We analyze the impact of varying the number of Sinkhorn iterations (N) on model performance, as summarized in Table 7. Increasing N to 20 led to substantial improvements in F1 scores for both OPT-350M (58.97) and Pythia (61.79), underscoring the importance of adequate iterations for achieving convergence. Beyond this point, however, raising N to 50 yields negligible performance gains, indicating a saturation threshold where additional iterations do not contribute further. This suggests that while sufficient iterations are necessary for convergence, excessive iterations offer diminishing returns and unnecessarily increase computational costs.

k **as the number of truncation threshold.** Table 8 illustrates the effect of the truncation threshold (k) on knowledge distillation for two student models, OPT-350M and Pythia-410M. Our findings demonstrate that $k = 50$ is optimal for both models on the QED dataset. A smaller k insufficiently captures the full sentence structure, weakening the Sinkhorn distance's ability to model high-dimensional geometric information, and thus limiting the student model's capacity to mimic the teacher's logit distribution. Conversely, a larger k introduces too many near-zero logit elements, adding noise and causing mode-averaging, which impairs the student's ability to distinguish critical information.

Conclusion

We propose MultiLevelOT for cross-tokenizer knowledge distillation that leverages both sequence-aware token-level and sequence-level optimal transport. Our method incorporates diverse cost matrices, using joint token optimization and Sinkhorn distance to provide a robust and comprehensive framework for KD. Extensive experiments demonstrate that MultiLevelOT consistently outperforms state-of-the-art cross-tokenizer KD methods across various NLP tasks. Moreover, our approach proves robust across different student model families, architectures, sizes, and teacher models, showcasing its versatility and broad applicability.

Broader Impact It is prospective to use our method for multi-teacher knowledge transfer, integrating knowledge from multiple teachers to enhance model performance. Additionally, MultiLevelOT may be suitable for cross-language and multi-modal knowledge transfer, enabling robust alignment across different languages and data modalities.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Contract 62021001, and the Youth Innovation Promotion Association CAS. It was also supported by GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of the USTC.

References

Agarwal, R.; Vieillard, N.; Zhou, Y.; Stanczyk, P.; Garea, S. R.; Geist, M.; and Bachem, O. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *ICML*, 214–223.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bhardwaj, R.; Vaidya, T.; and Poria, S. 2021. KNOT: Knowledge distillation using optimal transport for solving NLP tasks. *arXiv preprint arXiv:2110.02432*.

Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, 2397–2430.

Boizard, N.; El-Haddad, K.; Hudelot, C.; and Colombo, P. 2024. Towards Cross-Tokenizer Distillation: the Universal Logit Distillation Loss for LLMs. *arXiv preprint arXiv:2402.12030*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *NeurIPS*, 1877–1901.

Chen, P.; Zhao, R.; He, T.; Wei, K.; and Yang, Q. 2022. Unsupervised domain adaptation of bearing fault diagnosis based on Join Sliced Wasserstein Distance. *ISA transactions*, 129: 504–519.

Chen, Y.; Liu, Y.; Chen, L.; and Zhang, Y. 2021. DialogSum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Cui, X.; Qin, Y.; Gao, Y.; Zhang, E.; Xu, Z.; Wu, T.; Li, K.; Sun, X.; Zhou, W.; and Li, H. 2024a. SinKD: Sinkhorn Distance Minimization for Knowledge Distillation. *TNNLS*.

Cui, X.; Qin, Y.; Gao, Y.; Zhang, E.; Xu, Z.; Wu, T.; Li, K.; Sun, X.; Zhou, W.; and Li, H. 2024b. Sinkhorn Distance Minimization for Knowledge Distillation. In *COLING*, 14846–14858.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26.

Du, B.; Xie, W.; Li, Y.; Yang, Q.; Zhang, W.; Negenborn, R. R.; Pang, Y.; and Chen, H. 2023. Safe Adaptive Policy Transfer Reinforcement Learning for Distributed Multiagent Control. *TNNLS*.

Fang, G.; Bao, Y.; Song, J.; Wang, X.; Xie, D.; Shen, C.; and Song, M. 2021. Mosaicking to distill: Knowledge distillation from out-of-domain data. In *NeurIPS*, 11920–11932.

Giarelis, N.; Mastrokostas, C.; and Karacapilidis, N. 2023. Abstractive vs. extractive summarization: An experimental review. *Applied Sciences*, 13(13): 7620.

Gu, J.; Qian, X.; Zhang, Q.; Zhang, H.; and Wu, F. 2023a. Unsupervised domain adaptation for Covid-19 classification based on balanced slice Wasserstein distance. *COMPUT BIOL MED*, 107207.

Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2023b. Knowledge Distillation of Large Language Models. *arXiv preprint arXiv:2306.08543*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. *NeurIPS*, 30.

He, S.; Jiang, Y.; Zhang, H.; Shao, J.; and Ji, X. 2022. Wasserstein unsupervised reinforcement learning. In *AAAI*, volume 36, 6884–6892.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kim, Y.; and Rush, A. M. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Ko, J.; Kim, S.; Chen, T.; and Yun, S.-Y. 2024. DistiLLM: Towards Streamlined Distillation for Large Language Models. *arXiv preprint arXiv:2402.03898*.

Lamm, M.; Palomaki, J.; Alberti, C.; Andor, D.; Choi, E.; Soares, L. B.; and Collins, M. 2021. Qed: A framework and dataset for explanations in question answering. *TACL*, 9: 790–806.

Lan, Y.; Xu, X.; Fang, Q.; and Hao, J. 2023. Sample efficient deep reinforcement learning with online state abstraction and causal transformer model prediction. *TNNLS*.

Li, S.; Unanue, I. J.; and Piccardi, M. 2023. Improving Machine Translation and Summarization with the Sinkhorn Divergence. In *PAKDD*, 149–161. Springer.

Liu, C.; Tao, C.; Feng, J.; and Zhao, D. 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)*, 1001–1011.

Liu, Y.; Zhu, L.; Wang, X.; Yamada, M.; and Yang, Y. 2023. Bilaterally normalized scale-consistent sinkhorn distance for few-shot image classification. *TNNLS*.

Lu, S.; Ye, H.-J.; and Zhan, D.-C. 2022. Faculty Distillation with Optimal Transport. *arXiv preprint arXiv:2204.11526*.

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

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Le Scao, T.; Bari, M. S.; Shen, S.; Yong, Z. X.;

Schoelkopf, H.; et al. 2023. Crosslingual Generalization through Multitask Finetuning. In *ACL*, 15991–16111.

Nguyen, T. T.; and Luu, A. T. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *AAAI*, volume 36, 11103–11111.

Park, G.; Kim, G.; and Yang, E. 2021. Distilling Linguistic Context for Language Model Compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 364–378.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *FTML*, 11(5-6): 355–607.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu, L.; Pang, R. Y.; Wiseman, S.; and Gimpel, K. 2020. ENGINE: Energy-based inference networks for non-autoregressive machine translation. *arXiv preprint arXiv:2005.00850*.

Villani, C.; and Villani, C. 2009. The wasserstein distances. *Optimal Transport: Old and New*, 93–111.

Wan, F.; Huang, X.; Cai, D.; Quan, X.; Bi, W.; and Shi, S. 2024. Knowledge fusion of large language models. *arXiv preprint arXiv:2401.10491*.

Wei, Y.; Li, X.; Lin, L.; Zhu, D.; and Li, Q. 2022. Causal discovery on discrete data via weighted normalized Wasserstein distance. *TNNLS*.

Weilin, C.; Jie, Q.; Ruichu, C.; and Zhifeng, H. 2023. On the Role of Entropy-Based Loss for Learning Causal Structure With Continuous Optimization. *TNNLS*.

Wen, Y.; Li, Z.; Du, W.; and Mou, L. 2023. f-Divergence Minimization for Sequence-Level Knowledge Distillation. In *ACL*, 10817–10834.

Wu, T.; Tao, C.; Wang, J.; Zhao, Z.; and Wong, N. 2024. Rethinking Kullback-Leibler Divergence in Knowledge Distillation for Large Language Models. *arXiv preprint arXiv:2404.02657*.

Xu, J.; Li, C.; Huang, F.; Li, Z.; Xie, X.; and Philip, S. Y. 2023. Sinkhorn distance minimization for adaptive semi-supervised social network alignment. *TNNLS*.

Xu, Y.; Wang, D.; Yu, M.; Ritchie, D.; Yao, B.; Wu, T.; Zhang, Z.; Li, T. J.-J.; Bradford, N.; Sun, B.; et al. 2022. Fantastic Questions and Where to Find Them: FairytaleQA–An Authentic Dataset for Narrative Comprehension. *arXiv preprint arXiv:2203.13947*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yin, H.; Molchanov, P.; Alvarez, J. M.; Li, Z.; Mallya, A.; Hoiem, D.; Jha, N. K.; and Kautz, J. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. *TPAMI*, 8715–8724.

Zhang, J.; Liu, T.; and Tao, D. 2021. An optimal transport analysis on generalization in deep learning. *TNNLS*, 34(6): 2842–2853.

Zhang, J.; Muhamed, A.; Anantharaman, A.; Wang, G.; Chen, C.; Zhong, K.; Cui, Q.; Xu, Y.; Zeng, B.; Chilimbi, T.; et al. 2023. ReAugKD: Retrieval-augmented knowledge distillation for pre-trained language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1128–1136.

Zhang, Q.; Leng, S.; Ma, X.; Liu, Q.; Wang, X.; Liang, B.; Liu, Y.; and Yang, J. 2024a. CVaR-Constrained Policy Optimization for Safe Reinforcement Learning. *TNNLS*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, S.; Zhang, X.; Sun, Z.; Chen, Y.; and Xu, J. 2024b. Dual-Space Knowledge Distillation for Large Language Models. *arXiv preprint arXiv:2406.17328*.

Zhou, W.; Xu, C.; and McAuley, J. 2022. BERT Learns to Teach: Knowledge Distillation with Meta Learning. In *ACL*, 7037–7049.