MCW-KD: Multi-Cost Wasserstein Knowledge Distillation for Large Language Models

Hoang Tran Vuong[1], **Tue Le**[1], **Quyen Tran**[2], **Linh Ngo Van**[1], **Trung Le**[3]
[1]Hanoi University of Science and Technology, Hanoi, Vietnam
[2]Rutgers University, New Jersey, USA
[3]Monash University, Clayton, VIC 3800, Australia
hoang.tv224855@sis.hust.edu.vn, tue.ldt210909@sis.hust.edu.vn, qt60@rutgers.edu, linhnv@soict.hust.edu.vn, trunglm@monash.edu

Abstract

Knowledge distillation (KD) is widely recognized as an effective approach for compressing large language models (LLMs). However, standard KD methods often falter when confronted with architectural or tokenization heterogeneity between teacher and student models, which creates a mismatch in their representations. While Optimal Transport (OT) provides a promising solution to align these representations, most OT-based methods rely on a single cost function, which isn't enough to capture the multifaceted discrepancies between models with distinct designs. To address this limitation, we introduce **Multi-Cost Wasserstein Knowledge Distillation (MCW-KD)**, a novel framework that enhances KD by simultaneously optimizing several cost functions within a unified OT formulation. MCW-KD employs specific cost matrices to effectively align both the final hidden states and the output distributions of the models. We also provide a rigorous theoretical foundation for the proposed Multi-Cost Wasserstein Distance, ensuring both mathematical validity and computational ability. Extensive experiments on instruction-following datasets demonstrate that MCW-KD significantly improves student model performance compared to state-of-the-art KD baselines, especially when teacher and student models have different tokenizers.

1 Introduction

The rapid advancement of large language models (LLMs) has greatly advanced the field of natural language processing (NLP), leading to major progress across diverse tasks (Brown et al. 2020; Touvron et al. 2023; Team et al. 2023; He et al. 2025). These models, often comprising billions of parameters, rely on vast computational resources to achieve state-of-the-art performance. However, their practical deployment is constrained by high computational and memory requirements. Knowledge distillation (KD) has emerged as a promising approach for transferring knowledge from a large, high-capacity teacher model to smaller student models, enabling deployment on limited resources while preserving performance (Wen et al. 2023; Gu et al. 2024; Zhang et al. 2024b). Effective alignment of representations between teacher and student models is critical to the success of KD. Conventional KD approaches, such as those employing Kullback-Leibler (KL) divergence (Hinton, Vinyals, and Dean 2015; Kim and Rush 2016), Jensen-Shannon (JS) divergence (Wen et al. 2023), or Reverse KL divergence (Gu et al. 2023), primarily focus on aligning output distributions of teacher and student models. However, these methods struggle when teacher and student models differ in their architectures or tokenization strategies. Such differences lead to misaligned representational spaces, resulting in suboptimal knowledge transfer and decreased student model performance. The Wasserstein distance, grounded in optimal transport (OT) theory, provides a more robust framework for distribution alignment (Nguyen et al. 2017), particularly when teacher and student models differ in architecture or tokenization. Unlike divergence-based methods, Wasserstein distance effectively quantifies dissimilarities between distributions residing in different metric spaces by preserving their geometric structure (Villani et al. 2008; Peyré, Cuturi et al. 2019). Moreover, divergence-based methods often suffer from mode-averaging and mode-collapse issues, as observed in generative adversarial networks (Arjovsky, Chintala, and Bottou 2017). In KD, these issues manifest as student models oversmooths rich output distributions of a teacher model or become overly focused on specific modes. Wasserstein distance, as demonstrated by (Arjovsky, Chintala, and Bottou 2017), mitigates these problems, enabling more comprehensive knowledge transfer between teacher and student models.

Recent advances have adopted OT as a robust alignment framework for KD, leveraging metrics like Wasserstein distance (Boizard et al. 2024; Lv, Yang, and Li 2024) or minimizing transport costs to align distributions (Cui et al. 2025). However, most OT-based methods depend on a single cost function, which may fail to capture the diverse aspects of representational discrepancies between teacher and student models, especially when their architectures differ. For instance, ULD (Boizard et al. 2024) focuses on local token-level alignment but neglects sequence-level relationships and semantic coherence. Similarly, MultiLevelOT (Cui et al. 2025) utilizes multiple cost matrices, but it lacks a mechanism to effectively integrate these cost matrices or apply them across diverse representational spaces. Effective alignment in KD requires simultaneously addressing multiple perspectives and objectives, such as syntactic alignment at token level, semantic coherence across sequences or distributional consistency in output probabilities. Capturing these

diverse perspectives is essential to fully exploit the multi-faceted knowledge encoded in the teacher model.

One approach to achieve this is to formulate an optimization problem integrating multiple Wasserstein distances, each weighted to balance cost matrices. However, this strategy requires careful weight tuning to ensure each cost matrix contributes appropriately to the alignment process. Moreover, the influence of individual cost matrices may vary dynamically during training, fixed weight assignments can lead to suboptimal alignment. An alternative approach is represent the problem as a multi-objective optimization task (Yu et al. 2020; Liu et al. 2021; Navon et al. 2022; Ban and Ji 2024; He et al. 2024), optimizing a set of Wasserstein distances simultaneously. However, this approach necessitates computing and aggregating gradients for each distance, introducing significant computational overhead, particularly for LLMs where efficiency and scalability are critical. These limitations underscore the need for a more flexible and comprehensive alignment approach in knowledge distillation. To address these challenges, we introduce **Multi-Cost Wasserstein Knowledge Distillation (MCW-KD)**, a novel framework that enhances KD by simultaneously optimizing multiple cost matrices to achieve robust alignment between teacher and student representations. Unlike prior approaches, our method defines distinct cost matrices for specific alignment objectives, targeting both the last hidden state and output distribution spaces. By integrating these aspects within a unified OT framework, our method effectively bridges the capacity gap between teacher and student, even when they have different architectures or tokenization schemes. To the best of our knowledge, we are the first to develop the theory of multi-cost for Wasserstein distance in the context of KD, establishing a generalized OT-based framework that concurrently optimizes multiple cost matrices. We provide a rigorous theoretical foundation, ensuring the mathematical soundness and practical applicability of the method. Our key contributions are as follows:

- We propose **Multi-Cost Wasserstein Knowledge Distillation (MCW-KD)**, the first OT-based framework to optimize multiple cost matrices concurrently in knowledge distillation, enabling robust alignment between models with diverse representations.
- We establish a solid theoretical foundations, incorporating ε-entropic dual formulations (Genevay et al. 2016), ensuring the mathematical validity and computational efficiency of our method.
- We validate our method through extensive experiments on instruction-following datasets, demonstrating significant performance improvements in student models over knowledge distillation baselines, particularly in scenarios with architectural heterogeneity.

2 Related Work

Knowledge Distillation for LLMs: Knowledge distillation (KD) in NLP is a widely explored technique to transfer knowledge from a teacher model to student model by aligning the teacher's soft targets, such as output logits or representations. With the rise of large language models (LLMs) across various tasks, KD techniques tailored for LLMs have been developed (Gu et al. 2023; Peng et al. 2023; Hsieh et al. 2023; Xu et al. 2024; Ko et al. 2024). Common approaches employ divergence measures like KL divergence (Kim and Rush 2016; Park, Kim, and Yang 2021; Agarwal et al. 2024), Jensen-Shannon (JS) divergence (Fang et al. 2021; Wen et al. 2023), or Reverse KL divergence (Gu et al. 2023; Wu et al. 2024; Ko et al. 2024). These methods require pointwise correspondence between student and teacher outputs, necessitating use of the same tokenizer and consistent vocabularies, which limits their applicability when tokenizers differ. To address this, solutions such as ULD (Boizard et al. 2024) aligns distributions of individual tokens using optimal transport; DSKD (Zhang et al. 2024b) aligns token representations in a shared space, and MinED (Wan et al. 2024) minimizes edit distance between tokenized sequence logits. However, these approaches often neglect sequence-level semantic dependencies, limiting robust knowledge transfer. Effective KD for LLMs requires demands capturing diverse perspectives to fully exploit the teacher model's knowledge.

Optimal Transport: Optimal transport (OT) theory offers a robust mathematical framework for comparing probability distributions by minimizing the cost of transforming one distribution into another. The Wasserstein distance, a robust OT metric that quantifies distribution dissimilarities effectively (Villani et al. 2008; Zhang, Liu, and Tao 2021). This metric has been widely applied in NLP applications such as clustering (Xu et al. 2018; Zhuang, Chen, and Yang 2022), sentence similarity (Colombo et al. 2021) or classification (Shi et al. 2023). To improve computational efficiency, Sinkhorn distance has been proposed as an approximation for Wasserstein distance by adding an entropy regularization term, making OT more tractable (Cuturi 2013).

3 Method

3.1 Problem Setting and Notions

Given a training set $D = \{(x^i, y^i)\}_{i=1}^n$, an input sentence x is tokenized by teacher model and student into a sequence $[x_1^T, x_2^T, \ldots, x_M^T]$ and $[x_1^S, x_2^S, \ldots, x_N^S]$, respectively. Due to distinct tokenization strategies, sequence lengths M and N may differ ($M \neq N$). Then, teacher f_{T} and student f_{S} predict logit vectors $z^{(T)} \in R^{\mathrm{V}_T \times M}$ and $z^{(S)} \in R^{\mathrm{V}_S \times N}$, obtain the prediction probabilities:

$$\begin{aligned} p^{(T)} &= \mathrm{softmax}(z^{(T)}/\tau^{(T)}) \\ p^{(S)} &= \mathrm{softmax}(z^{(S)}/\tau^{(S)}) \end{aligned} \tag{1}$$

We denote the layer hidden state representations as $H^S \in R^{N \times d}$ for student and $H^T \in R^{M \times D}$ for teacher, where d and D are the hidden dimensions. The i-th row vector of H^S, denoted $h_i^S \in R^d$, corresponds to the hidden state of i-th student token x_i^S. Similarly, $h_j^T \in R^D$ represents the hidden state for j-th teacher token x_j^T.

3.2 Our Motivations

In the context of knowledge distillation for LLMs, it is crucial not only to distill the generation outputs from teacher to student, but also to align their internal representations.

Specifically, we aim to match the representations of the teacher and student LLMs at the distributional level. To achieve this, we define two empirical distributions over the hidden states, denoted as P^S and P^T, corresponding to the student and teacher models, respectively. Subsequently, we aim to minimize a suitable divergence or distance metric between the distributions P^S and P^T.

Among various divergence measures such as KL divergence, JS divergence, and general f-divergences, the Wasserstein (WS) distance stands out as particularly suitable due to its theoretical advantages and empirical effectiveness (Arjovsky, Chintala, and Bottou 2017). To define a WS distance between P^S and P^T, we must first specify a cost metric between the representations of teacher and student models. Each such cost metric captures a specific aspect of the relationship between the sequences generated by the teacher and student for the same input. To achieve a comprehensive alignment, it is desirable to compare and align these sequences from multiple complementary perspectives. To address the problem of multi-criteria alignment, one can formulate in two ways: (i) a *weighted sum over the distances*, i.e., $\sum_i \lambda_i \mathcal{W}_{c_i}\left(P^T, P^S\right)$, or (ii) a *multi-objective optimization problem* (Yu et al. 2020; Liu et al. 2021; Navon et al. 2022; Ban and Ji 2024; He et al. 2024), i.e., $\min_{f^S}\left[\mathcal{W}_{c_i}\left(P^T, P^S\right)\right]_i$. While the first approach requires tuning a potentially large number of trade-off parameters λ_i, second approach is computationally prohibitive, as it involves multiple backward passes to compute gradients and costly gradient aggregation procedures, especially in the context of LLMs. It is essential to develop an efficient WS-based multi-cost alignment mechanism, along with meaningful cost metrics that capture diverse aspects of token sequences, to enable more effective alignment between teacher and student models.

3.3 Multi-Cost Wasserstein Distance

Let X and Y be two Polish spaces equipped with probability measures μ and ν, respectively. Given a set of lower semi-continuous cost functions $c_1, \ldots, c_m : X \times Y \to R$, we aim to find a coupling $\pi \in \Gamma(\mu, \nu)$ that *minimizes multiple total costs* $\int c_i d\pi, \quad i = 1, \ldots, m$ *simultaneously*:

$$\mathrm{MWS}(\mu, \nu, c) = \inf_{\pi \in \Gamma(\mu,\nu)} \left[\int c_1 \, d\pi, \ldots, \int c_m \, d\pi\right] \quad (2)$$

where $\mathrm{MWS}(\mu, \nu, c)$ specifies the *multi-cost Wasserstein distance* between two distributions μ and ν w.r.t. the cost functions $c_1, ..., c_m$. We cast the above multi-cost optimal transport to the following optimization problem:

$$\inf_{\pi \in \Gamma(\mu,\nu)} \inf_{\alpha \in \Delta_{m-1}} \int c_\alpha d\pi \quad (3)$$

where Δ_{m-1} is the $(m-1)$-simplex and $c_\alpha = \sum_{i=1}^m \alpha_i c_i$ is the corresponding weighted cost function. We can interpret the optimization problem in (3) as follows: *given a transport plan $\pi \in \Gamma(\mu, \nu)$, we aim to find $\alpha \in \Delta_{m-1}$ to minimize $\sum_{i=1}^m \alpha_i \int c_i \, d\pi$ and then find the optimal transport plan π^* that minimizes all these minimizations.*. To ensure all cost functions contribute meaningfully to the computation for a given transport plan π, we incorporate a regularization term:

$$\Omega(\alpha) = \sum_{i=1}^m \lambda_i \left(\alpha_i - \frac{1}{m}\right)^2 \quad (4)$$

where $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_m]$, with $\lambda_i > 0, \ \forall i$, encouraging balanced participation of each cost. Then, we obtain the following optimization problem:

$$\inf_{\pi \in \Gamma(\mu,\nu)} \inf_{\alpha \in \Delta_{m-1}} \left\{\int c_\alpha \, d\pi + \Omega(\alpha)\right\} \quad (5)$$

By the optimization problem in (5), we learn the weights to combine the cost functions $c_1, \ldots, c_m$. To make a tractable primal form, we add the stochastic entropic regularization term to the above primal form to reach:

$$\begin{aligned} \inf_{\pi \in \Gamma(\mu,\nu)} \Bigg\{ \inf_{\alpha \in \Delta_{m-1}} \left\{\int c_\alpha \, d\pi + \Omega(\alpha)\right\} \\ + \varepsilon \mathrm{KL}(\pi \| \mu \otimes \nu) \Bigg\} \end{aligned} \quad (6)$$

where $\varepsilon > 0$. This formulation allows us to learn an optimal transport plan that balances multiple cost perspectives while maintaining computational efficiency. The following theorem presents the dual-form for the primal problem in (6), making it trainable. The primal form in (6) has the following dual form:

$$\begin{aligned} &\inf_{\alpha \in \Delta_{m-1}} \sup_{\varphi, \psi} \left\{-\varepsilon \int e^{\frac{\varphi(x)+\psi(y)-c_\alpha(x,y)-\Omega(\alpha)}{\varepsilon}} d\mu(x) d\nu(y) \right. \\ &\qquad \left. + \int \varphi \, d\mu + \int \psi \, d\nu \right\} \\ &= \inf_{\alpha \in \Delta_{m-1}} \sup_{\varphi} \left\{\int \varphi \, d\mu + \int \varphi^c \, d\nu\right\} \end{aligned} \quad (7)$$

where we have defined

$$\varphi^c(y) = -\varepsilon \log \int e^{\frac{\varphi(x)-c_\alpha(x,y)-\Omega(\alpha)}{\varepsilon}} d\mu(x) \quad (8)$$

The detailed theoretical foundations and proofs are provided in Appendix. Our proposed multi-cost Wasserstein framework is particularly well-suited for knowledge distillation scenarios, where teacher and student models may produce sequences of different lengths or representational spaces. In the following sections, we apply this framework to align representations in two critical domains. The last hidden state space (Section 3.4) provides comprehensive alignment of hidden state sequences, capturing syntactic structure, local semantic magnitude, and global semantic direction simultaneously. The output distribution space (Section 3.5) reflects the model's predictive behavior, enabling the student to mimic the teacher's predictions while capturing precise token-level reproduction, distributional robustness, and semantic consistency. By employing tailored cost matrices, our approach bridges the capacity gap between teacher and student models, achieves robust alignment across divergent model tokenizations.

3.4 Alignment in Last Hidden State Space

Multi-Cost Wasserstein framework, introduced in Section 3.3, provides a robust mechanism for aligning sequences by simultaneously optimizing multiple cost matrices. In this section, we apply it to align the last hidden state of teacher and student LLMs. We define three cost matrices, each capturing a distinct perspective of the representational relationship between teacher and student model.

Edit Distance Cost Matrix. The cost matrix is designed to capture syntactic structure and token-level correspondence between student and teacher sequences. This is particularly crucial when models utilize different tokenizers, leading to sequences of potentially different lengths. We define a matrix from student to teacher $C^{S\to T} \in R^{N\times M}$, each element $C^{S\to T}_{i,j}$ represents the cost of aligning student's token x^S_i with teacher's token x^T_j:

$$C^{S\to T}_{i,j} = \begin{cases} \text{edit}(x^S_i, x^T_j) & \text{if } x^S_i \text{ is aligned with } x^T_j \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where $\text{edit}(x^S_i, x^T_j)$ is the Levenshtein distance. Similarly, we define cost matrix $C^{T\to S} \in R^{M\times N}$ for alignments from teacher to student sequences. The final cost matrix is constructed as:

$$\mathrm{C}_{\mathrm{MED}} = \frac{C^{S\to T} + (C^{T\to S})^T}{2} \quad (10)$$

Contextual Cost Matrix. Beyond syntactic alignment, effective KD requires aligning semantic meaning. To capture local semantic dissimilarity, we define cost matrix C_{MC} based on contextualized representations. We first project teacher's hidden states $h^T \in R^{M\times D}$ into student's d-dimensional space using a trainable linear projection $P \in R^{D\times d}$, and compute representations for both student and projected teacher by averaging them within a local window. Cost matrix C_{MC} is defined to quantify the discrepancy between these representations, encouraging student's contextualized semantics to closely match those of the teacher:

$$\mathrm{C}_{\mathrm{MC}}[i,j] = \frac{1}{2K+1} \left\| \sum_{l=-K}^{K} \left(h^S_{i+l} - P(h^T_{j+l})\right) \right\|_2 \quad (11)$$

where K is window radius. Complementing the magnitude-based C_{MC}, C_{SC} targets directional similarity in contextualized representations. To achieve this, we normalize contextualized embeddings of student and teacher at last hidden states, yielding $\bar{h}^S$ and $\bar{h}^T$. By minimizing C_{SC}, OT plan prioritizes alignments between tokens whose contextualized representations exhibit similar semantic directions:

$$\mathrm{C}_{\mathrm{SC}}[i,j] = 1 - \langle \bar{h}^S_j,\ \bar{h}^T_i \rangle \quad (12)$$

Through the joint integration of these cost matrices, MCW-KD provides a comprehensive alignment of the last hidden state sequences—capturing syntactic structure, local semantic magnitude, and global semantic direction.

3.5 Alignment in Output Distribution Space

To transfer knowledge effectively from teacher to student model, aligning their output distributions is essential. Building on the framework in Section 3.3, we propose an OT-based alignment objective that integrates multiple cost matrices, each capturing a distinct perspective of the distributional discrepancy between teacher and student models. Rather than directly comparing outputs, we construct a time-dependent interpolated distribution. This design provide a smoother alignment target, enhancing training stability and mitigating representational disparities due to the capacity gap. Formally, we normalize and sort logit vectors in descending order, truncating them to retain only top-k highest values, obtaining $z^{(T)}_{\text{top-k}}$ and $z^{(S)}_{\text{top-k}}$. The interpolated logit for teacher at position i is defined as:

$$\tilde{z}^{(T)}_i = (1-t)\cdot z^{(S)}_{i,\text{top-k}} + t \cdot z^{(T)}_{i,\text{top-k}} \quad (13)$$

where t is linear interpolation factor increasing over training. Applying softmax with temperature $\tau^{(T)}$, we obtain interpolated teacher distributions $\tilde{p}^{(T)} = \text{softmax}(\tilde{z}^{(T)}/\tau^{(T)})$. This interpolation strategy allows student to gradually shift from relying on its own predictions toward effectively learning from the teacher as training progresses.

Euclidean Distance Cost Matrix. The cost matrix quantifies the overall Euclidean discrepancy between the interpolated teacher distribution $\tilde{p}^{(T)}$ and the truncated student distribution $\tilde{p}^{(S)} = \text{softmax}(z^{(S)}_{\text{top-k}}/\tau^{(S)})$ at each token position. This cost is motivated by the need for a computationally efficient measure to ensure that student's predictions approximate teacher's output distribution:

$$\mathrm{C}_{\mathrm{ED}}[i,j] = \left\| \tilde{p}^{(T)}_i - \tilde{p}^{(S)}_j \right\|_2 \quad (14)$$

Divergence Cost Matrix. Based on the observation that in NLPs, tokens with high-probability predictions often carry critical semantic information, cost matrix C_{KL} uses KL divergence to measure discrepancy between $\tilde{p}^{(S)}_j$ and $\tilde{p}^{(T)}_i$, prioritizing tokens with high predictive significance. C_{KL} penalize mismatches in these critical tokens, thereby enabling student model to capture the teacher's most confident predictions:

$$\mathrm{C}_{\mathrm{KL}}[i,j] = \sum_{\ell=1}^{k} \tilde{p}^{(T)}_{i,\ell} \cdot \log\left(\frac{\tilde{p}^{(T)}_{i,\ell}}{\tilde{p}^{(S)}_{j,\ell} + \varepsilon}\right) \quad (15)$$

By prioritizing alignment in high-confidence regions of distribution, C_{KL} enhances global perspective of C_{ED}, helping student better match teacher's key predictive features.

Salience Score Cost Matrix. Cost matrix C_{SS} accounts for token importance in a sequence based on its hidden state representation. It is motivated by the fact that not all tokens contribute equally to sequence's meaning, particularly in long or complex inputs. We quantify token significance by computing salience scores for student and teacher model, and define the cost as their absolute difference. The scores are derived from hidden states via learned linear projections:

$$\mathrm{C}_{\mathrm{SS}}[i,j] = \left|\sigma(W_s \cdot h_j^{(S)} + b_s) - \sigma(W_t \cdot h_i^{(T)} + b_t)\right| \quad (16)$$

Large salience score differences at position (i, j) will prompt OT plan to prioritize alignments between student and teacher tokens with similar importance. By integrating these cost matrices, MCW-KD achieves a comprehensive alignment of output distributions, enabling student to effectively capture teacher's predictive behavior.

3.6 Overall

Our MCW-KD framework is grounded in the novel Multi-Cost Wasserstein Distance (Section 3.3), enabling robust alignment between teacher and student across tokenization disparities by simultaneous optimizing multiple cost matrices. We apply this mechanism to two critical domains: last hidden states (P_h^T, P_h^S) and output distributions (P_o^T, P_o^S), using tailored cost sets $c_{1:3}^h = [\mathrm{C}_{\mathrm{MED}}, \mathrm{C}_{\mathrm{MC}}, \mathrm{C}_{\mathrm{SC}}]$ and $c_{1:3}^o = [\mathrm{C}_{\mathrm{ED}}, \mathrm{C}_{\mathrm{KL}}, \mathrm{C}_{\mathrm{SS}}]$, respectively:

$$\begin{aligned} \mathcal{L}_{\mathrm{Hidden}} &= \mathrm{MWS}(P_h^T, P_h^S, c_{1:3}^h) \\ \mathcal{L}_{\mathrm{Output}} &= \mathrm{MWS}(P_o^T, P_o^S, c_{1:3}^o). \end{aligned} \quad (17)$$

These multi-perspective alignments are key to bridging capacity gaps and enabling student performance. Motivated by the effectiveness of Dual-Space Knowledge Distillation (Zhang et al. 2024b) in unifying output distribution spaces via a shared language model head, we incorporate its loss as a supplementary term to further improve knowledge transfer. The final distillation objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{Hidden}} + \mathcal{L}_{\mathrm{Output}} + \gamma \mathcal{L}_{\mathrm{DSKD}} \quad (18)$$

where $\mathcal{L}_{\mathrm{CE}}$ is the standard cross-entropy loss between student predictions and ground truth, $\mathcal{L}_{\mathrm{DSKD}}$ is DSKD's distillation loss with $\gamma > 0$. By jointly optimizing this multi-component loss, MCW-KD effectively captures the teacher's knowledge across multiple perspectives for enhanced performance on student model, despite tokenization disparities. Pseudo-code for our approach are provided in Appendix.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on several instruction-following datasets. Following (Gu et al. 2024), we use DATABRICKS-DOLLY-15K dataset to conduct KD process. For evaluation, in addition to this primary dataset, we include four instruction-tuning benchmarks as additional test sets for out-of-distribution evaluation: SUPER-NATURAL-INSTRUCTIONS (**S-NI**) (Wang et al. 2022), VICUNA-EVALUATION (**VICUNAEVAL**) (Chiang et al. 2023), DIALOGSUM (**DIALOG**) (Chen et al. 2021) and SELF-INSTRUCT (**SELFINST**) (Wang et al. 2023). This diverse evaluation allows us to assess model's generalization capabilities across a wide range of instruction domains.

Models. We focus on scenarios between teacher and student models with distinct tokenizers and vocabularies.

Methods	Dolly	S-NI	SelfInst	Dialog	Vicuna	Avg.
Qwen1.5-1.8B → GPT2-120M						
Teacher	28.23	34.36	19.58	14.18	19.59	23.19
SFT	23.78	7.81	6.78	8.29	**17.04**	12.74
ULD	23.77	14.04	9.30	8.63	14.33	14.01
MinED	24.21	16.40	10.02	9.79	14.96	15.08
MultiLevelOT	23.02	12.26	8.41	8.79	13.79	13.25
DSKD	24.26	17.15	10.07	10.03	15.25	15.35
MCW-KD	**25.03**	**18.71**	**11.41**	**10.22**	15.91	**16.26**
Qwen2.5-7B-Instruct → GPT2-1.5B						
Teacher	28.49	39.87	24.67	16.86	20.48	26.07
SFT	21.83	21.66	13.62	10.91	15.95	16.79
ULD	24.52	26.18	15.11	11.72	15.94	18.69
MinED	25.52	26.25	15.39	11.79	16.15	19.02
MultiLevelOT	24.40	23.94	14.53	10.84	15.97	17.94
DSKD	25.38	25.82	16.10	12.19	16.84	19.27
MCW-KD	**27.94**	**28.70**	**18.35**	**14.12**	**19.19**	**21.66**

Table 1: Rouge-L scores (%) averaged over 5 random seeds on several benchmarks for two teacher–student model pairs. Best scores are highlighted in bold.

For student models, we select a range of LLMs across diverse architectures: GPT2 (120M, 340M, 1.5B) (Radford et al. 2019), TinyLLaMA-1.1B (Zhang et al. 2024a) and OPT2.7B (Zhang et al. 2022). We paired Qwen1.5-1.8B (Bai et al. 2023) as the teacher for GPT-2-120M and GPT-2-340M. Mistral-7B (Jiang 2024) served as the teacher for TinyLlama-1.1B, while Qwen2.5-7B-Instruct (Yang et al. 2024) was chosen for GPT-2-1.5B and OPT-2.7B.

Baselines. We compare our method against a range of approaches, from supervised fine-tuning (SFT) to several state-of-the-art knowledge distillation techniques designed for addressing discrepancies in tokenizers and vocabularies, including: **ULD** (Boizard et al. 2024), **MinED** (Wan et al. 2024), **DSKD** (Zhang et al. 2024b) and **MultiLevelOT** (Cui et al. 2025). These baselines provide a comprehensive evaluation framework for assessing our method's ability to handle tokenizer and vocabulary disparities. Additional details on model configurations, along with the training and evaluation setup, are provided in Appendix.

4.2 Main Results

Table 1 and 2 summarizes the performance of our proposed MCW-KD in comparison with SFT and recent state-of-the-art knowledge distillation baselines. Evaluations were conducted across five benchmark datasets, with teacher-student pairs featuring distinct tokenization schemes. MCW-KD consistently achieves the highest average ROUGE-L scores in all settings, significantly outperforming existing methods. Notably, MCW-KD delivers substantial improvements on individual tasks across diverse teacher-student pairs. On S-NI dataset, MCW-KD surpasses the best baselines by **+9.33%** with Qwen2.5 → GPT2-1.5B and **+9.10%** with Qwen1.5 → GPT2-120M. On SelfInst, it achieves a gain of **+13.98%** over DSKD with Qwen2.5 → GPT2-1.5B,

Methods	Dolly	S-NI	SelfInst	Dialog	Vicuna	Avg.
Qwen1.5-1.8B → GPT-2 340M						
Teacher	28.23	34.36	19.58	14.18	19.59	23.19
SFT	23.11	13.03	9.09	8.00	14.89	13.62
ULD	23.90	16.26	9.96	8.76	15.04	14.78
MinED	24.48	15.69	11.21	8.98	15.56	15.18
MultiLevelOT	23.95	15.87	10.21	8.99	14.80	14.76
DSKD	25.43	17.18	11.29	8.90	15.08	15.57
MCW-KD	**26.12**	**19.98**	**11.97**	**9.67**	**16.15**	**16.78**
Mistral-7B → TinyLLaMA-1.1B						
Teacher	32.15	36.88	25.44	14.67	20.43	25.91
SFT	23.20	28.43	15.70	10.77	15.70	18.76
ULD	25.48	32.54	17.72	11.75	17.31	20.96
MinED	25.54	31.42	18.23	11.77	17.02	20.80
MultiLevelOT	24.56	27.91	15.61	12.04	16.84	19.40
DSKD	26.28	31.93	17.19	12.53	**18.74**	21.33
MCW-KD	**26.51**	**37.00**	**18.71**	**13.32**	18.40	**22.79**
Qwen2.5-7B-Instruct → OPT-2.7B						
Teacher	28.49	39.87	24.67	16.86	20.48	26.07
SFT	27.10	24.90	13.90	10.62	16.60	18.62
ULD	26.65	25.44	15.37	12.15	16.97	19.32
MinED	26.89	25.94	14.98	11.78	17.04	19.33
MultiLevelOT	26.76	24.84	15.51	11.43	16.56	19.02
DSKD	26.93	27.33	16.22	**12.43**	17.86	20.15
MCW-KD	**28.57**	**29.75**	**17.29**	**12.43**	**18.68**	**21.34**

Table 2: Rouge-L scores (%) averaged over 5 random seeds on several benchmarks for three teacher–student model pairs. Best scores are bolded.

and on Dialogsum, a margin of **+6.30%** (over DSKD, the best baseline) with Mistral → TinyLLaMA. Furthermore, MCW-KD attains the best ROUGE-L scores on Vicuna and Dolly across all teacher–student configurations, including both Qwen2.5 → GPT2-1.5B and Qwen2.5 → OPT-2.7B.

To further validate these improvements, we use API of `GPT-4o-mini` as the judge to perform pairwise comparisons between responses generated by MCW-KD and each baseline. Complementing the ROUGE-L results, GPT evaluations show that MCW-KD consistently achieves higher win rates across all comparisons, as illustrated in Figure 1. These results underscore the robustness and generalization capability of MCW-KD. By leveraging multiple cost matrices to capture different aspects of the teacher's knowledge, the student model achieves strong performance even on out-of-domain datasets without prior exposure.

4.3 Ablation Study

Impact of Each Component We conduct an ablation study to better understand contribution of each component in MCW-KD framework, focusing on multi-cost strategies (Sections 3.4, 3.5) and dynamic weighting mechanism (Section 3.3). We evaluate effects of incrementally adding $\mathcal{L}_{\text{Hidden}}$ and $\mathcal{L}_{\text{Output}}$, and compare dynamic strategy against a fixed-weight, where weights are set equally to $\frac{1}{3}$ for all

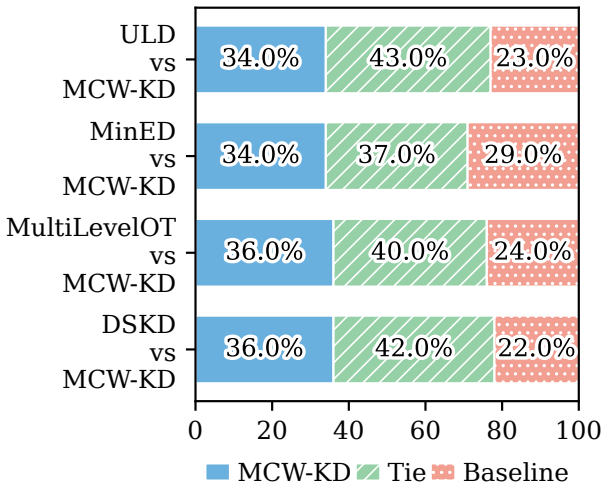

Figure 1: Win rates (%) for distilling Qwen2.5 to OPT-2.7B, evaluated by `GPT-4o-mini` on response quality.

cost matrices. Results in Table 3 indicate that incorporating multi-cost alignment in individual spaces enhances performance, while combining both provides substantially larger gains. However, using both alignment spaces without dynamic weighting results in suboptimal performance, even compared to single-space configurations. MCW-KD framework, which integrates both alignment strategies with dynamic weights, achieves highest performance, highlighting the effect of these components in robust knowledge transfer.

Effect of γ We conducted experiments to assess impact of $\mathcal{L}_{\text{DSKD}}$ component by varying hyperparameter γ in the set $[0.2, 1, 2, 5, 10]$. The results in Figure 2 show that changes in γ moderately affect the final performance. Performance slightly decreases with extremely small or excessively large values of γ, yet consistently surpasses baseline methods reported in Table 1. These findings underscore the robustness and effectiveness of our MCW-KD, demonstrating that even suboptimal choices of γ can yield competitive performance.

Effect of λ To evaluate the role of λ in multi-cost optimization mechanism of MCW-KD, we conduct experiments by assigning uniform values of $\lambda \in [0.2, 0.7, 1, 2, 5]$ to all cost functions. The results, also shown in Figure 2, indicate that student model performance remains stable across all evaluation datasets regardless of λ. This observation demonstrates the generalization ability and robustness of MCW-KD, which is not overly sensitive to the choice of λ.

Results with Shared Vocabulary We also evaluate performance of MCW-KD under a shared vocabulary setting, assessing it against two distinct of KD methods: **(i)** distance-based techniques, which are designed specifically for scenarios with shared tokenizers, including: **SeqKD** (Kim and Rush 2016), reverse KL (**RKL**) and Jensen-Shannon (**JS**) (Wen et al. 2023), skewed KL and skewed reverse KL (**SKL**, **SRKL**) (Ko et al. 2024), adaptive KL (**AKL**) (Wu et al.

Methods	Dolly	S-NI	SelfInst	Dialog	Vicuna	Avg.
Mistral-7B → TinyLLaMA-1.1B						
DSKD	26.28	31.93	17.19	12.53	**18.74**	21.33
+ $\mathcal{L}_{\text{Hidden}}$	25.79	33.40	17.39	13.26	18.39	21.65
+ $\mathcal{L}_{\text{Output}}$	26.06	**37.08**	17.04	13.02	18.22	22.28
+ w/o update	25.93	33.01	17.45	13.29	18.33	21.60
MCW-KD	**26.51**	37.00	**18.71**	**13.32**	18.40	**22.79**
Qwen2.5-7B-Instruct → GPT2-1.5B						
DSKD	25.38	25.82	16.10	12.19	16.84	19.27
+ $\mathcal{L}_{\text{Hidden}}$	26.26	29.51	16.05	12.90	17.65	20.47
+ $\mathcal{L}_{\text{Output}}$	26.77	27.87	16.96	12.48	18.02	20.42
+ w/o update	26.96	29.61	16.77	13.47	17.69	20.90
MCW-KD	**27.94**	**28.70**	**18.35**	**14.12**	**19.19**	**21.66**

Table 3: Rouge-L scores (%) averaged over 5 random seeds. We report performance of DSKD with incorporating multi-cost alignment and dynamic weight strategy.

Methods	Dolly	S-NI	SelfInst	Vicuna	Avg.
Teacher	27.19	27.55	14.64	16.30	21.42
SeqKD	23.68	16.36	10.03	14.41	16.12
RKL	24.38	17.31	10.73	**15.71**	17.03
JS	23.86	16.20	10.20	15.50	16.44
SKL	24.03	17.99	10.66	14.70	16.85
SRKL	24.48	16.53	10.35	14.88	16.56
AKL	**24.75**	17.48	10.46	15.37	17.02
ULD	23.53	15.43	10.47	14.89	16.08
MinED	23.69	15.84	10.43	15.17	16.28
MultiLevelOT	23.81	14.91	10.70	14.91	16.08
DSKD	23.93	16.81	10.66	15.00	16.60
MCW-KD	24.29	**18.17**	**11.47**	15.31	**17.31**

Table 4: Mean Rouge-L scores (%) over 5 random seeds for GPT2-1.5B → GPT2-120M distillation across four datasets.

2024); **(ii)** approaches designed to mitigate tokenizer discrepancies. Our results in Table 4, demonstrate superior performance of MCW-KD compared to methods in category **(ii)**. Notably, it also exhibits competitive performance in category **(i)**, with highest overall average, despite being primarily designed to address tokenizer discrepancies. The enhanced performance of MCW-KD over methods in category **(ii)** under a shared vocabulary setting can be attributed to its novel approach of simultaneously optimizing multiple cost functions, which effectively capture diverse representational discrepancies between teacher and student models. The flexibility of MCW-KD framework suggests potential for integration with distance-based methods in **(i)**, which could enhance their robustness in knowledge distillation scenarios.

Generalizability To evaluate the generalizability of MCW-KD, we assessed its integration with other established KD baselines beyond DSKD. As our framework is designed as a versatile loss term, it can be easily incorporated into various KD objectives. Results in Table 5 demonstrate that applying MCW-KD module enhances the performance of these other baselines. This validates that MCW-KD functions as a broadly applicable and scalable enhancement, not an extension limited to a single baseline.

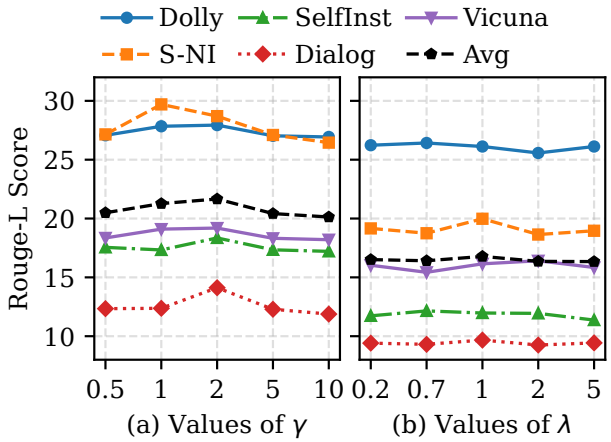

Figure 2: Ablation results of MCW-KD on two key hyperparameters: (a) Effect of γ with Qwen2.5-7B-Instruct → GPT-1.5B; (b) Effect of λ with Qwen1.5-1.8B → GPT2-340M.

Methods	Dolly	S-NI	SelfInst	Dialog	Vicuna	Avg.
ULD	23.77	14.04	9.30	8.63	14.33	14.01
+ MCW-KD	**24.24**	**16.26**	**9.87**	**9.81**	**15.61**	**15.16**
MultiLevelOT	**23.02**	12.26	8.41	8.79	13.79	13.25
+ MCW-KD	22.46	**13.51**	**9.58**	**9.44**	**15.08**	**14.01**
MinED	24.21	16.40	10.02	9.79	14.96	15.08
+ MCW-KD	**24.56**	**16.83**	**10.87**	**10.38**	**15.95**	**15.72**

Table 5: Generalizability of MCW-KD applied to existing baselines for Qwen1.5-1.8B → GPT2-120M.

5 Conclusion

We introduce MCW-KD, a novel Multi-Cost Wasserstein Knowledge Distillation framework addressing representation alignment between LLMs with divergent tokenization strategies. By integrating multiple cost functions within a unified Wasserstein formulation, it facilitates robust, efficient knowledge transfer from teacher to student models. Our results demonstrate performance improvements across diverse datasets, highlighting the framework's generalizability. This work establishes a principled approach to multi-perspective representation alignment in knowledge distillation, enabling efficient compression of LLMs. However, predefined cost matrices may not fully capture all representational aspects, and computing optimal transport plan can be demanding for large-scale matrices. Future work will investigate novel cost matrices tailored to diverse representation spaces and explore improving computational efficiency.

Acknowledgements

Trung Le was supported by the Air Force Office of Scientific Research under award number FA9550-23-S-0001. Linh Ngo Van is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2025.16.

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 *The Twelfth International Conference on Learning Representations*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. arXiv:1701.07875.

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*.

Ban, H.; and Ji, K. 2024. Fair resource allocation in multitask learning. *arXiv preprint arXiv:2402.15638*.

Boizard, N.; Haddad, K. E.; 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. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, Y.; Liu, Y.; Chen, L.; and Zhang, Y. 2021. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. arXiv:2105.06762.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. *URL https://lmsys.org/blog/2023-03-30-vicuna*, 3(5).

Colombo, P.; Staerman, G.; Clavel, C.; and Piantanida, P. 2021. Automatic text evaluation through the lens of Wasserstein barycenters. *arXiv preprint arXiv:2108.12463*.

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

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

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. *Advances in Neural Information Processing Systems*, 34: 11920–11932.

Genevay, A.; Cuturi, M.; Peyré, G.; and Bach, F. 2016. Stochastic optimization for large-scale optimal transport. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 3440–3448. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2023. MiniLLM: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2024. MiniLLM: Knowledge Distillation of Large Language Models. arXiv:2306.08543.

He, K.; Mao, R.; Lin, Q.; Ruan, Y.; Lan, X.; Feng, M.; and Cambria, E. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, 102963.

He, Y.; Zhou, S.; Zhang, G.; Yun, H.; Xu, Y.; Zeng, B.; Chilimbi, T.; and Zhao, H. 2024. Robust multi-task learning with excess risks. *arXiv preprint arXiv:2402.02009*.

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

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. *arXiv preprint arXiv:2305.02301*.

Jiang, F. 2024. *Identifying and mitigating vulnerabilities in llm-integrated applications*. Master's thesis, University of Washington.

Kim, Y.; and Rush, A. M. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 1317–1327.

Ko, J.; Kim, S.; Chen, T.; and Yun, S.-Y. 2024. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*.

Liu, B.; Liu, X.; Jin, X.; Stone, P.; and Liu, Q. 2021. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34: 18878–18890.

Lv, J.; Yang, H.; and Li, P. 2024. Wasserstein distance rivals kullback-leibler divergence for knowledge distillation. *Advances in Neural Information Processing Systems*, 37: 65445–65475.

Navon, A.; Shamsian, A.; Achituve, I.; Maron, H.; Kawaguchi, K.; Chechik, G.; and Fetaya, E. 2022. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*.

Nguyen, T. D.; Le, T.; Vu, H.; and Phung, D. 2017. Dual Discriminator Generative Adversarial Nets. arXiv:1709.03831.

Park, G.; Kim, G.; and Yang, E. 2021. Distilling linguistic context for language model compression. *arXiv preprint arXiv:2109.08359*.

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

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Shi, L.; Zhen, H.; Zhang, G.; and Yan, J. 2023. Relative entropic optimal transport: a (prior-aware) matching perspective to (unbalanced) classification. *Advances in Neural Information Processing Systems*, 36: 22085–22098.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Villani, C.; et al. 2008. *Optimal transport: old and new*, volume 338. Springer.

Wan, F.; Huang, X.; Cai, D.; Quan, X.; Bi, W.; and Shi, S. 2024. Knowledge Fusion of Large Language Models. arXiv:2401.10491.

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. arXiv:2212.10560.

Wang, Y.; Mishra, S.; Alipoormolabashi, P.; Kordi, Y.; Mirzaei, A.; Arunkumar, A.; Ashok, A.; Dhanasekaran, A. S.; Naik, A.; Stap, D.; et al. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2: 2.

Wen, Y.; Li, Z.; Du, W.; and Mou, L. 2023. F-divergence minimization for sequence-level knowledge distillation. *arXiv preprint arXiv:2307.15190*.

Wu, T.; Tao, C.; Wang, J.; Yang, R.; Zhao, Z.; and Wong, N. 2024. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv preprint arXiv:2404.02657*.

Xu, H.; Wang, W.; Liu, W.; and Carin, L. 2018. Distilled wasserstein learning for word embedding and topic modeling. *Advances in Neural Information Processing Systems*, 31.

Xu, X.; Li, M.; Tao, C.; Shen, T.; Cheng, R.; Li, J.; Xu, C.; Tao, D.; and Zhou, T. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

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

Zhang, P.; Zeng, G.; Wang, T.; and Lu, W. 2024a. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; Mihaylov, T.; Ott, M.; Shleifer, S.; Shuster, K.; Simig, D.; Koura, P. S.; Sridhar, A.; Wang, T.; and Zettlemoyer, L. 2022. OPT: Open Pre-trained Transformer Language Models. 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*.

Zhuang, Y.; Chen, X.; and Yang, Y. 2022. Wasserstein K-means for clustering probability distributions. *Advances in Neural Information Processing Systems*, 35: 11382–11395.