EMO: Embedding Model Distillation via Intra-Model Relation and Optimal Transport Alignments

Minh-Phuc Truong[1]*, Hai An Vu[1]*, Tu Vu[2]*, Diep Thi-Ngoc Nguyen[3], Linh Ngo Van[1,†], Thien Huu Nguyen[4], Trung Le[5]

[1]Hanoi University of Science and Technology, [2]ByteDance Inc,
[3]VNU University of Engineering and Technology, [4]University of Oregon, [5]Monash University

Abstract

Knowledge distillation (KD) is crucial for compressing large text embedding models, but faces challenges when teacher and student models use different tokenizers (Cross-Tokenizer KD - CTKD). Vocabulary mismatches impede the transfer of relational knowledge encoded in deep representations, such as hidden states and attention matrices, which are vital for producing high-quality embeddings. Existing CTKD methods often focus on direct output alignment, neglecting this crucial structural information. We propose a novel framework tailored for CTKD embedding model distillation. We first map tokens one-to-one via Minimum Edit Distance (MinED). Then, we distill intra-model relational knowledge by aligning attention matrix patterns using Centered Kernel Alignment, focusing on the top-m most important tokens of the directly mapped tokens. Simultaneously, we align final hidden states via Optimal Transport with Importance-Scored Mass Assignment, which emphasizes semantically important token representations, based on importance scores derived from attention weights. We evaluate distillation from state-of-the-art embedding models (e.g., LLM2Vec, BGE) to a Bert-base-uncased model on embedding-reliant tasks such as text classification, sentence pair classification, and semantic textual similarity. Our proposed framework significantly outperforms existing CTKD baselines. By preserving attention structure and prioritizing key representations, our approach yields smaller, high-fidelity embedding models despite tokenizer differences.

1 Introduction

Knowledge distillation (KD) has emerged as a highly effective technique for model compression, enabling the transfer of knowledge from large, computationally expensive teacher models to smaller, more efficient student models (Hinton et al., 2015). This is crucial for Large Language Models (LLMs), whose state-of-the-art performance often entails significant deployment challenges (Zhao et al., 2025). KD offers a promising avenue to create compact models that preserve much of the teacher's capabilities while being suitable for resource-constrained environments, as demonstrated by influential works like DistilBERT, TinyBERT, and MiniLM (Sanh et al., 2020; Jiao et al., 2019; Wang et al., 2020).

In the specific domain of representation learning, KD plays a vital role in developing efficient text embedding models, as demonstrated by various efforts to compress large embedding models while preserving their semantic representation capabilities (Sanh et al., 2020; Jiao et al., 2019; Wang et al., 2020; Zhang et al., 2025; Gao et al., 2023). State-of-the-art embedding models, benchmarked on MTEB (Muennighoff et al., 2023), tend to possess a large number of parameters and high-dimensional outputs (Lee et al., 2025; Xiao et al., 2024), posing challenges for practical deployment. Consequently, distilling these large embedding teachers into smaller students, as explored in works such as Jasper (Zhang et al., 2025) and DistillCSE (Gao et al., 2023), is an area of significant interest.

A fundamental assumption in many conventional KD frameworks, including those recently applied to embedding models like Jasper (Zhang et al., 2025) or general embedding models such as TinyBERT (Jiao et al., 2019) and DistilBERT (Sanh et al., 2020), is that the teacher and student models share the same tokenizer, vocabulary. This homogeneity simplifies the alignment process, typically involving the minimization of a distance metric (e.g., KL divergence) between the probability distributions of the two models at each token position (Sun et al., 2019; Gu et al., 2024). However, this assumption limits flexibility given diverse modern LLMs' tokenization. Distilling knowledge between

*Equal contribution
†Corresponding author: linhnv@soict.hust.edu.vn

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7606–7618
November 4-9, 2025

models with different tokenizers introduces significant challenges: divergent tokenization strategies cause sequence misalignments, and differing vocabularies result in output spaces with mismatched dimensions and semantics.

Recent approaches to Cross-Tokenizer KD (CTKD) include using Optimal Transport (OT) for distribution and sequence level alignment (Boizard et al., 2025; Cui et al., 2024), employing MinED for token mapping (Wan et al., 2024), and unifying output spaces (Zhang et al., 2024). Despite these advancements, current CTKD methods mainly align logits, overlooking rich intermediate layer knowledge such as hidden states (Sun et al., 2019) or attention patterns (Clark et al., 2019), which are crucial for effective distillation into smaller models (Jiao et al., 2019; Sun et al., 2019).

To overcome these limitations, we propose a framework **EMO** (**E**mbedding **M**odel Distillation via Intra-Model Relation and **O**ptimal Transport Alignments) for Cross-Tokenizer Knowledge Distillation, specifically designed for learning high-quality text embeddings. Our approach goes beyond simple output alignment by distilling knowledge from intermediate layers while explicitly preserve intra-model token relationships. We identify reliably one-to-one mapped tokens between student and teacher sequences using Minimum Edit Distance (MinED). Then we focus on the top-m most important tokens of these specific matched tokens to align the internal model structures by distilling their corresponding attention matrices using Centered Kernel Alignment (CKA) (Kornblith et al., 2019). This step explicitly transfers the learned token relationships and contextual dependencies captured within the teacher's self-attention mechanism, preserving vital structural information often lost in cross-tokenizer scenarios. In addition, we introduce Optimal Transport (Villani, 2008; Cuturi, 2013) with importance-based mass assignment to directly align tokens' representations across the teacher and student models. While CKA preserves structural dependencies within mapped token pairs, OT focuses on mapping semantically tokens between the models, addressing the token misalignment caused by tokenizer differences. This direct alignment of contextualized tokens' representations ensures that critical information is transferred even when sequence lengths or token boundaries differ. We summarize the contributions of our study as follows:

1. We propose **EMO**, a novel embedding model distillation framework that integrates two complementary components for improved cross-model alignment. First, Intra-Model Relational Alignment (IRA) captures structural correspondences by aligning attention patterns using Centered Kernel Alignment (CKA), focusing specifically on the top-m most salient token pairs identified via direct mapping from MinED. Second, Optimal Transport with Importance-Scored Mass Assignment (OTIS) ensures robust representation alignment across models by leveraging token-level importance scores to guide the optimal transport process. Together, these components enable EMO to distill both relational and representational knowledge effectively.

2. Through extensive experiments distilling a LLM2Vec (BehnamGhader et al., 2024) or BGE (Chen et al., 2024) teacher to a Bert-base-uncased student, we demonstrate that our framework significantly outperforms existing state-of-the-art cross-tokenizer KD methods and conventional KD baselines for text embedding tasks, enabling the creation of smaller, yet highly performant, embedding models.

2 Related Work and Background

This section reviews prior work in knowledge distillation, focusing on techniques for same- and cross-tokenizer scenarios, and introduces foundational concepts like Optimal Transport and Centered Kernel Alignment.

2.1 Related Work

Text embedding models play an important role in several domains such as retrieval-augmented generation (RAG) (Nguyen et al., 2025), information extraction (Pham et al., 2025; Le et al., 2025; Anh et al., 2025), topic model (Vuong et al., 2025; Vu et al., 2025), etc. Recent advances in embedding models largely result from fine-tuning large language models on representation learning tasks (BehnamGhader et al., 2024). However, the substantial size of these architectures results in high computational overhead and significant resource costs during training and inference. Therefore, there is a need for lightweight yet effective embedding models. Knowledge distillation (KD) (Hinton et al., 2015) is a standard compression technique. With shared tokenizers, KD evolved from

logit matching (Gu et al., 2024) to distilling intermediate layer information, including hidden states (Sun et al., 2019; Liang et al., 2023; Jiao et al., 2019) and attention matrices (Clark et al., 2019; Wang et al., 2020; Jiao et al., 2019), recognizing that these capture crucial structural and relational knowledge.

However, distillation between models with different tokenizers (Cross-Tokenizer KD - CTKD) introduces challenges of sequence and vocabulary mismatches (Zhang et al., 2024). While early black-box methods relied on teacher outputs (Kim and Rush, 2016), recent white-box CTKD approaches employ techniques such as Optimal Transport for aligning output distributions like ULD (Boizard et al., 2025), MultiLevelOT (Cui et al., 2024), or dynamic programming for sequence alignment like MinED (Wan et al., 2024), or unified output spaces via projections like DSKD (Zhang et al., 2024).

Despite these advances in aligning overall outputs or logits, current CTKD methods often lack mechanisms to explicitly distill the relational information captured within the teacher's attention mechanism – a technique proven valuable in same-tokenizer KD. This gap is particularly relevant when distilling large text embedding models (Zhang et al., 2025; Muennighoff et al., 2023), where preserving contextual and structural understanding is paramount for downstream tasks. While some work exists on same-tokenizer embedding KD (Gao et al., 2023), effectively transferring relational knowledge across disparate tokenizers for embedding models remains an open challenge. Our work addresses this by specifically adapting relational distillation principles, focusing on attention structure, to the cross-tokenizer embedding scenario.

2.2 Background

2.2.1 Knowledge Distillation Fundamentals

Knowledge Distillation (KD) (Hinton et al., 2015) is a model compression technique where a smaller student model learns from a larger teacher model. Instead of only using ground-truth labels, the student matches the teacher's softened output probability distributions, typically derived from teacher logits (x_t) after temperature scaling. This knowledge transfer is achieved by minimizing a distillation loss $\mathcal{L}_{KD}(x_t, x_s)$ (e.g., KL divergence) between the student's and teacher's distributions:

$$\mathcal{L}_{KD}(x_t, x_s) \tag{1}$$

This $\mathcal{L}_{KD}$ is usually combined with the standard supervised cross-entropy loss $\mathcal{L}_{CE}$ on ground-truth labels y for the student's training:

$$\mathcal{L} = \mathcal{L}_{CE}(y, p(x_s)) + \mathcal{L}_{KD}(x_t, x_s) \tag{2}$$

2.2.2 Optimal Transport Principles

Optimal Transport (OT) (Villani, 2008) offers a robust framework for comparing probability distributions by finding the minimal cost to transform one into another, excelling where classical divergences like KL struggle with differing supports. This makes OT highly suitable for aligning outputs from language models with disparate vocabularies.

For discrete distributions $f = \sum \alpha_i \delta_{\boldsymbol{x}_i}$ and $g = \sum \beta_j \delta_{\boldsymbol{y}_j}$, OT finds a transport plan $\boldsymbol{T}$ detailing mass T_{ij} moved from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$, respecting marginal probabilities $\boldsymbol{\alpha}, \boldsymbol{\beta}$. Given a cost matrix $\boldsymbol{D}$ (e.g., based on distances $d(\boldsymbol{x}_i, \boldsymbol{y}_j)^p$), the OT (Wasserstein) distance d_W is the minimum total transport cost $\langle \boldsymbol{T}, \boldsymbol{D} \rangle$ over all valid plans $\boldsymbol{T}$:

$$d_W(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{D}) = \min_{\boldsymbol{T} \in U(\boldsymbol{\alpha}, \boldsymbol{\beta})} \langle \boldsymbol{T}, \boldsymbol{D} \rangle \tag{3}$$

This facilitates sequence-level alignment across different vocabularies in CTKD.

2.2.3 Measuring Representational Similarity: From CCA to CKA

Canonical Correlation Analysis (CCA) (Hardoon et al., 2004) has been employed to measure representational similarity by seeking linear projections that maximize the correlation between two sets of representations. However, CCA faces limitations: it is sensitive to simple transformations of the representations (such as isotropic scaling or rotation) and can be computationally intensive, especially for the high-dimensional representations common in deep learning (Kornblith et al., 2019). Centered Kernel Alignment (CKA), introduced by Kornblith et al. (2019), offers a more robust and computationally efficient alternative. A core insight of CKA is its shift from comparing individual multivariate features to comparing the *learned similarity structures* within each representation space.

CKA operationalizes this concept using the Hilbert-Schmidt Independence Criterion (HSIC), a non-parametric kernel-based measure of statistical dependence. For two sets of representations, $\boldsymbol{X} \in \mathbb{R}^{m \times S}$ and $\boldsymbol{Y} \in \mathbb{R}^{m \times T}$ (for m common inputs), CKA constructs kernel matrices $\boldsymbol{K}$ and $\boldsymbol{L}$. Here, $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(y_i, y_j)$ capture

the pairwise similarities using kernel functions k and l. HSIC then measures the dependence between these kernel matrices after centering them with a matrix $\boldsymbol{H} = I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T$ to remove mean effects:

$$\text{HSIC}(\boldsymbol{K}, \boldsymbol{L}) = \frac{1}{(m-1)^2}\text{tr}(\boldsymbol{K}\boldsymbol{H}\boldsymbol{L}\boldsymbol{H}) \quad (4)$$

A high HSIC value signifies that the structure of pairwise similarities in $\boldsymbol{X}$ (captured by $\boldsymbol{K}$) is strongly related to the structure in $\boldsymbol{Y}$ (captured by $\boldsymbol{L}$). CKA then normalizes HSIC to achieve invariance to isotropic scaling and orthogonal transformations, yielding a robust similarity score:

$$\text{CKA}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\text{HSIC}(\boldsymbol{K}, \boldsymbol{L})}{\sqrt{\text{HSIC}(\boldsymbol{K}, \boldsymbol{K})\text{HSIC}(\boldsymbol{L}, \boldsymbol{L})}} \quad (5)$$

This normalized index effectively quantifies the alignment of the overall representational geometries, making CKA a valuable tool for determining if two models or layers have learned to organize input examples in a structurally similar fashion.

3 Methodology

This section details our proposed EMO framework, which distills intra-model relational knowledge via attention matrix alignment using CKA and aligns cross-model representation of hidden states using Optimal Transport with Importance-Scored Mass Assignment (OTIS).

3.1 Intra-Model Relation Distillation via Attention Matrices

A core challenge in CTKD is the inherent misalignment between the student token sequences $\mathbf{x_s}$ and the teacher token sequences $\mathbf{x_t}$. Directly comparing hidden states or attention token-by-token is often infeasible or inaccurate. Furthermore, simply aligning final outputs neglects the rich relational information learned within the transformer layers. Our approach IRA (**I**ntra-Model **R**elation Distillation via **A**ttention Matrices) addresses this by distilling the attention patterns associated with them.

Token mapping via MinED

We use Minimum Edit Distance (MinED) to identify n reliably mapped one-to-one token pairs between the student tokenized sequence $\mathbf{x_s}$ (length S) and the teacher tokenized sequence $\mathbf{x_t}$ (length T), such that $n \leq \min(S, T)$. We denote the set of indices corresponding to these n mapped tokens within their respective sequences as $\mathcal{N}_s \subset \{1, ..., S\}$ and $\mathcal{N}_t \subset \{1, ..., T\}$ ($|\mathcal{N}_t| = |\mathcal{N}_s| = n$), and the set of mapped token pairs as $\mathcal{N} = \{(i, j) | i \in \mathcal{N}_s, j \in \mathcal{N}_t, \text{token } i \text{ maps to token } j\}$. Token mapping statistics are in Appendix A.

Identifying Important Tokens

Inspired by Li et al. (2023), we identify the most salient tokens based on the attention patterns in the final layer of the teacher model h_t (h_t and h_s are the number of hidden layers in the teacher and student models, respectively). The final layer often provides a more global perspective on token importance (Li et al., 2023). We compute the average attention matrix $\mathbf{A}_{h_t}^{\text{avg},Tea} \in \mathbb{R}^{T \times T}$ by averaging $\mathbf{A}_{h_t}^{\text{full},Tea}$ across all heads. Similarly, we define $\mathbf{A}_{h_s}^{avg,Stu} \in \mathbb{R}^{S \times S}$ for student. The importance score for the j-th teacher token is calculated as the sum of attention it receives from all other tokens:

$$\text{score}_j = \sum_{i=1}^{T} (\mathbf{A}_{h_t}^{\text{avg},Tea})_{ij} \quad (6)$$

We select the top-m indices from $\mathcal{N}_t$ based on their corresponding scores, where $m \leq n$. The resulting subset is denoted as $\mathcal{N}_t^{\text{top-}m}$. The corresponding student token indices are $\mathcal{N}_s^{\text{top-}m}$. The choice of m is discussed in Section 5. Focusing on the top-m most salient tokens (among the already mapped n) allows us to distill the most critical structural interactions efficiently. This selective focus is inspired by observations that not all tokens contribute equally to learning (Lin et al., 2025). It also reduces the dimensionality for CKA computation we demonstrate below.

Structural Alignment with CKA

For each student layer k, we map to layer l of the teacher with $l = M(k) = \left\lfloor \frac{h_t}{h_s} \right\rfloor \cdot k$. We extract the attention patterns from these top-m tokens to the entire sequence within their respective models. Let $\mathbf{A}_k^{Stu} \in \mathbb{R}^{m \times S}$ be the submatrix of $\mathbf{A}_k^{\text{avg},Stu}$ containing rows corresponding to the indices in $\mathcal{N}_s^{\text{top-}m}$. Similarly, let $\mathbf{A}_l^{Tea} \in \mathbb{R}^{m \times T}$ be the submatrix of $\mathbf{A}_l^{\text{avg},Tea}$ with rows corresponding to the indices in $\mathcal{N}_t^{\text{top-}m}$. We use a linear kernel, which offers reduced computational cost, vital for applying this algorithm to large language models. Let us define two kernel matrices:

$$\mathbf{P}_k^{Stu} = \mathbf{A}_k^{Stu}(\mathbf{A}_k^{Stu})^\top \quad (7)$$

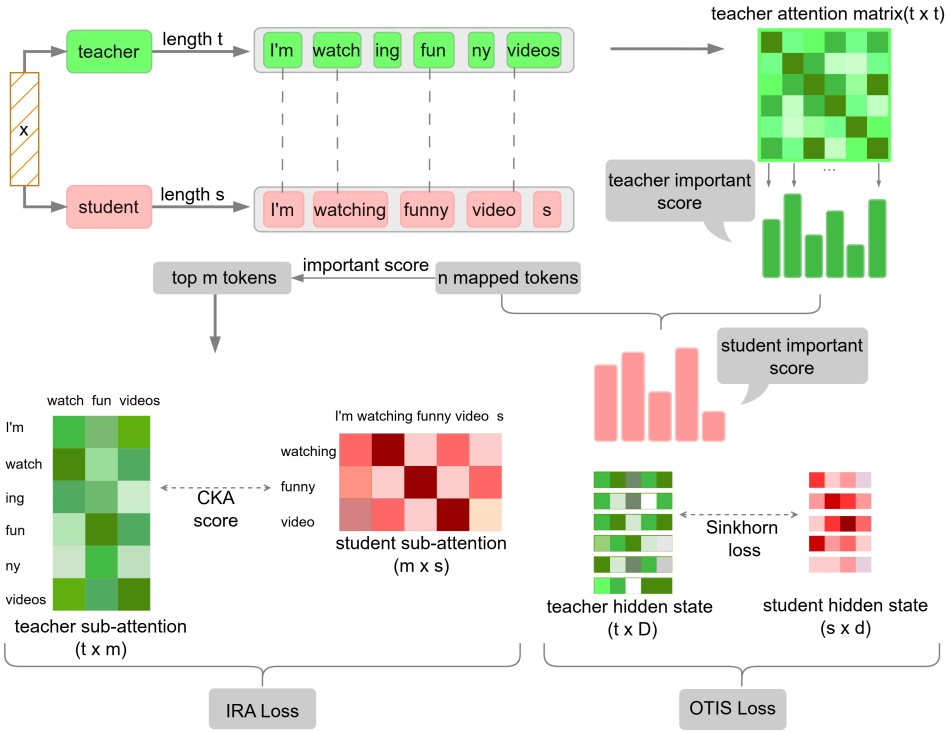

Figure 1: Overall workflow of our EMO framework. We perform Intra-Model Relational Distillation (IRA) using MinED for token mapping, followed by CKA on attention matrices and employ Optimal Transport with Importance-Scored Mass Assignment (OTIS) for cross-model representation alignment.

$$\mathbf{P}_l^{Tea} = \mathbf{A}_l^{Tea}(\mathbf{A}_l^{Tea})^\top \tag{8}$$

We define $\tilde{\mathbf{A}}_k^{Stu}$ and $\tilde{\mathbf{A}}_l^{Tea}$ are the centered matrices of $\mathbf{A}_k^{Stu}$ and $\mathbf{A}_l^{Tea}$:

$$\tilde{\mathbf{A}}_k^{Stu} = \mathbf{A}_k^{Stu}(I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T) \tag{9}$$

$$\tilde{\mathbf{A}}_l^{Tea} = \mathbf{A}_l^{Tea}(I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T) \tag{10}$$

The HSIC between the teacher and student sub-attention matrices is:

$$HSIC(\mathbf{P}_k^{Stu}, \mathbf{P}_l^{Tea}) = \left\|\text{cov}((\tilde{\mathbf{A}}_k^{Stu})^\top, (\tilde{\mathbf{A}}_l^{Tea})^\top)\right\|_F^2 \tag{11}$$

where cov() is a covariance function.

From section 2.2.3, the linear CKA between $\mathbf{A}_k^{Stu}$ and $\mathbf{A}_l^{Tea}$ is defined as:

$$\text{CKA}\left(\mathbf{A}_k^{Stu}, \mathbf{A}_l^{Tea}\right) = \frac{HSIC(\mathbf{P}_k^{Stu}, \mathbf{P}_l^{Tea})}{\sqrt{HSIC(\mathbf{P}_k^{Stu}, \mathbf{P}_k^{Stu}) \cdot HSIC(\mathbf{P}_l^{Tea}, \mathbf{P}_l^{Tea})}} \tag{12}$$

Combining Eq.11 and Eq.12, we obtain the formula for linear CKA between two sub-attention matrices:

$$\text{CKA}(X, Y) = \frac{\left\|\text{cov}\left(\tilde{X}^\top, \tilde{Y}^\top\right)\right\|_F^2}{\left\|\text{cov}\left(\tilde{X}^\top, \tilde{X}^\top\right)\right\|_F \cdot \left\|\text{cov}\left(\tilde{Y}^\top, \tilde{Y}^\top\right)\right\|_F} \tag{13}$$

where X denotes $\mathbf{A}_k^{\text{Stu}}$ and Y denotes $\mathbf{A}_l^{\text{Tea}}$. Note that CKA values lie within the interval $[0, 1]$.

We then define the Intra-Model Relation Distillation between these $m \times S$ and $m \times T$ matrices, applied to the last z layers of the student model:

$$\mathcal{L}_{\text{IRA}} = \sum_{k=h_s-z+1}^{h_s} \left(1 - \sqrt{\text{CKA}\left(\mathbf{A}_k^{Stu}, \mathbf{A}_l^{Tea}\right)}\right) \tag{14}$$

While previous work has effectively employed CKA to measure similarity between hidden state representations in language models (Dasgupta and Cohn, 2025), we apply it directly to attention matrices ($\mathbf{A}_k^{Stu}$ and $\mathbf{A}_l^{Tea}$). $\mathcal{L}_{\text{IRA}}$ leverages CKA to measure the structural similarity between the attention patterns from the top-m important tokens towards their respective full sequences. Specifically, $\mathcal{L}_{\text{IRA}}$ assesses whether the student and teacher models exhibit similar overall attention patterns (whether key tokens focus on similar parts of the sequence

in both models). By optimizing CKA loss on these attention matrices, we encourage the student to learn the teacher's high-level attentional structure regarding its top salient tokens.

3.2 Cross-Model Representation Alignment via Optimal Transport with Importance Scored Mass Assignment

While the intra-model relation distillation discussed earlier focuses on aligning attention patterns within each model's token sequence, this section addresses the direct alignment of token representations between the student and teacher models. We align the last hidden states of the student and teacher models via Optimal Transport (OT) (Nguyen, 2025). Specifically, our approach introduces a novel mass assignment strategy leveraging the insights from Section 3.1.

Importance-Based Mass Assignment

To assign masses to tokens, we inherit the teacher's importance scores computed in the previous section using Eq.6. These scores are normalized to form a probability distribution over the teacher's tokens: $\mu_j = \frac{\text{score}_j}{\sum_{j'=1}^{T} \text{score}_{j'}}, \quad j = 1, \ldots, T$. Thus, the empirical distribution for the teacher's last hidden state $\mathbf{h}_{h_t}^T \in \mathbb{R}^{T \times d_t}$ is:

$$\boldsymbol{\mu} = \sum_{j=1}^{T} \mu_j \delta_{\mathbf{h}_{h_t,j}^T} \tag{15}$$

where $\mathbf{h}_{h_t,j}^T$ is the hidden state of the j-th teacher token, and δ denotes the Dirac delta function.

For the student, we map the teacher's importance scores to the student's tokens using the mapped pairs of tokens $\mathcal{N} = \{(i,j) | i \in \mathcal{N}_s, j \in \mathcal{N}_t, \text{token } i \text{ maps to token } j\}$. For each mapped student token $i \in \mathcal{N}_s$, we assign the mass of its corresponding teacher token $j \in \mathcal{N}_t$: $\nu_i = \mu_j, \quad \text{for } (i,j) \in \mathcal{N}$. For student tokens $i \notin \mathcal{N}_s$ (i.e., unmapped tokens), we assign a minimal mass equal to the smallest teacher token mass: $\nu_i = \min_{j=1,\ldots,T} \mu_j$. The masses for all student tokens are then normalized to sum to 1: $\nu_i = \frac{\nu_i}{\sum_{i'=1}^{S} \nu_{i'}}, \quad i = 1, \ldots, S$. The empirical distribution for the student's last hidden state $\mathbf{h}_{h_s}^S \in \mathbb{R}^{S \times d_s}$ is:

$$\boldsymbol{\nu} = \sum_{i=1}^{S} \nu_i \delta_{\mathbf{h}_{h_s,i}^S} \tag{16}$$

where $\mathbf{h}_{h_s,i}^S$ is the hidden state of the i-th student token, and δ denotes the Dirac delta function.

Cost Matrix Computation

To align the student and teacher hidden states, we compute a cost matrix $\mathbf{C} \in \mathbb{R}^{S \times T}$ that quantifies the dissimilarity between token representations. We project the teacher's hidden states into the student's space using a learnable mapping matrix $\mathbf{P} \in \mathbb{R}^{d_t \times d_s}$. The similarity matrix is computed as:

$$\mathbf{S} = \frac{\mathbf{h}_{h_s}^S (\mathbf{h}_{h_t}^T \mathbf{P})^\top}{\sqrt{d_s}} \tag{17}$$

where $\mathbf{h}_{h_s}^S \in \mathbb{R}^{S \times d_s}$, $\mathbf{h}_{h_t}^T \in \mathbb{R}^{T \times d_t}$, and the scaling factor $\sqrt{d_s}$ ensures numerical stability. The similarity matrix is normalized row-wise using the softmax function $\mathbf{S}_{\text{norm}} = \text{softmax}(\mathbf{S})$, ensuring each row sums to 1. The cost matrix is then derived as:

$$\mathbf{C} = \mathbf{1} - \mathbf{S}_{\text{norm}} \tag{18}$$

Optimal Transport Alignment Loss with Importance-Based Masses

We compute the optimal transport plan $\mathbf{T}^*$ by solving the entropy-regularized OT problem (Cuturi, 2013):

$$\mathbf{T}^* = \arg \min_{\mathbf{T} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{T}, \mathbf{C} \rangle - \frac{1}{\lambda} H(\mathbf{T}) \tag{19}$$

where $H(\mathbf{T}) = -\sum_{i,j} T_{ij} \log T_{ij}$ is the entropy regularization term, and $\lambda > 0$ controls regularization strength. The OT-based loss is:

$$\mathcal{L}_{\text{OT}} = \langle \mathbf{T}^*, \mathbf{C} \rangle \tag{20}$$

We only apply OTIS at the final layer due to the substantial computational cost of Optimal Transport, as shown in Appendix C. Specifically, the cross-model alignment loss via Optimal Transport with Importance-Scored Mass Assignment (OTIS), applied to the last hidden states capturing refined representations, is formulated as:

$$\mathcal{L}_{\text{OTIS}} = \mathcal{L}_{\text{OT}}(\mathbf{h}_{h_s}^S, \mathbf{h}_{h_t}^T) \tag{21}$$

This loss aligns the teacher and student representations in the last hidden states, respecting the importance-based mass distributions. By restricting OTIS to the last layer, we strike a balance between effective representation alignment and feasible training efficiency.

3.3 Overall Distillation Loss

The final objective function for training the student model within our EMO framework is a combination of the standard task-specific loss and our proposed distillation losses. This overall loss $\mathcal{L}_{\text{EMO}}$ is defined as:

$$\mathcal{L}_{\text{EMO}} = \alpha \cdot \mathcal{L}_{\text{CE}} + (1-\alpha)(\mathcal{L}_{\text{IRA}} + \mathcal{L}_{\text{OTIS}}) \quad (22)$$

where $\alpha \in [0, 1]$ controls the relative influence of the standard cross-entropy loss L_{CE}, the Intra-relation via attention matrices loss L_{IRA}, and the proposed OT-based loss L_{OTIS}.

4 Experiments

4.1 Experimental Setup

We evaluate the effectiveness of our proposed framework through extensive experiments on tasks where high-quality text embeddings play a crucial role. We select the following three tasks for evaluation:

Text Classification: Requires models to capture the overall semantics of a single text input. We use Patent (Sharma et al., 2019), Imdb and Banking77 (both from MTEB (Muennighoff et al., 2023)).

Sentence Pair Classification: Demands understanding the relationship or similarity between two text inputs, heavily relying on the quality of their respective embeddings. We evaluate on SciTail (Khot et al., 2018), ConTRoL-NLI (Liu et al., 2021), and Anli_r2 (Nie et al., 2020).

Semantic Textual Similarity (STS): Directly measures the ability of embeddings to capture fine-grained semantic similarity. We use STSB, STS12 (both from MTEB (Muennighoff et al., 2023)) and SICK (from (Marelli et al., 2014)).

Further details on the models used, as well as the training and evaluation setup, can be found in Appendix B.

4.2 Baselines

We compare our **EMO** framework against several state-of-the-art CTKD methods:

- **ULD** (Universal Logit Distillation) (Boizard et al., 2025): Employs Optimal Transport to align output logit distributions across different vocabularies.
- **MinED** (Wan et al., 2024): Uses Minimum Edit Distance based on dynamic programming to align token sequences before distillation.
- **DSKD** (Dual-Space Knowledge Distillation) (Zhang et al., 2024): Unifies output spaces using projections and cross-model attention to enable KD between different tokenizers.
- **MultilevelOT** (Cui et al., 2024): Extends Optimal Transport for CTKD by incorporating multi-level alignment strategies.

In our experiment, when logit-based methods such as MinED or ULD are applied to an embedding model for a classification task, they essentially reduce to a conventional KL-based knowledge distillation approach.

Moreover, because STS is a regression task predicting a continuous similarity score, the model outputs a single scalar rather than a logit vector. Thus, we only compare our method with DSKD, as other baselines rely on output logit alignment. This also highlights the advantage of our method, which does not depend on logits, can be applied across diverse tasks.

4.3 Results

We present the evaluation results across the three task categories in Table 1, and Table 2. Across all three task categories and constituent datasets, the results consistently demonstrate the superiority of our proposed framework compared to the state-of-the-art CTKD baselines (ULD, MinED, DSKD, MultilevelOT). Our framework achieves the highest scores among the distillation methods on nearly all metrics and datasets, significantly closing the gap between the student (Bert SFT) and the teacher (LLM2Vec Mistral 7B SFT).

5 Analysis

This section investigates the individual contributions of our framework's components and the impact of top-m salient token selection. Moreover, the ablation study about its robustness to different teacher models, and the impact of the distilled layer depth are in Appendix C.

Impact of Framework Components

An ablation study (Table 3) isolates the contributions of Intra-Model Relational Distillation (IRA) and Optimal Transport with Importance-Scored Mass Assignment (OTIS). Configurations using only IRA ($\textbf{EMO}_{\text{w/o OTIS}}$) or only OTIS ($\textbf{EMO}_{\text{w/o IRA}}$) both consistently outperform the **Bert SFT** baseline, demonstrating their individual benefits. The full framework **EMO** combining

Table 1: Model Performance on Classification and SentencePair Classification Tasks. "EMO" denotes our proposed framework.

Method	Classification task				SentencePair Classification task			
	Dataset	Accuracy	Precision	Recall	Dataset	Accuracy	Precision	Recall
LLM2Vec Mistral 7B SFT (Teacher)	Patent	70.0	67.7	66.1	SciTail	96.1	96.0	95.8
Bert SFT (Student)		63.1	58.7	54.4		88.1	87.7	88.8
ULD (Boizard et al., 2025)		64.8	61.4	60.9		87.0	86.4	87.8
DSKD (Zhang et al., 2024)		64.0	60.0	58.8		88.0	87.3	88.8
MinED (Wan et al., 2024)		65.0	61.6	60.8		86.9	86.1	87.5
MultilevelOT (Cui et al., 2024)		64.6	60.4	59.0		88.2	88.0	89.1
EMO		**66.5**	**63.3**	**62.4**		**90.9**	**90.1**	**91.2**
LLM2Vec Mistral 7B SFT (Teacher)	Imdb	96.6	96.6	96.6	ConTRoL-ni	63.6	62.7	62.6
Bert SFT (Student)		91.3	91.4	91.3		42.1	38.6	37.5
ULD (Boizard et al., 2025)		92.5	92.6	92.5		45.4	45.3	45.3
DSKD (Zhang et al., 2024)		93.4	93.5	93.4		42.2	41.2	39.7
MinED (Wan et al., 2024)		92.5	92.5	92.5		47.1	47.0	47.2
MultilevelOT (Cui et al., 2024)		93.3	93.4	93.3		42.5	41.4	40.1
EMO		**94.2**	**94.3**	**94.2**		**48.6**	**48.2**	**48.1**
LLM2Vec Mistral 7B SFT (Teacher)	Banking 77	93.3	93.5	93.3	Anli_r2	67.1	67.8	67.0
Bert SFT (Student)		85.7	86.4	85.7		42.7	42.6	42.6
ULD (Boizard et al., 2025)		91.4	91.9	91.4		44.8	44.7	44.7
DSKD (Zhang et al., 2024)		91.2	91.7	91.2		43.1	43.4	43.0
MinED (Wan et al., 2024)		90.0	91.2	90.0		46.4	46.6	46.4
MultilevelOT (Cui et al., 2024)		89.4	90.4	89.4		44.1	44.1	43.9
EMO		**92.3**	**92.7**	**92.3**		**47.6**	**47.8**	**47.5**

Table 2: Model Performance on Semantic Textual Similarity (STS) tasks. Metric is Spearman Correlation Coefficient (ρ). "EMO" denotes our proposed framework.

Dataset	Method	Spearman Corr (ρ)
STS-B	LLM2Vec Mistral 7B SFT (Teacher)	90.8
	Bert SFT (Student)	75.1
	DSKD (Zhang et al., 2024)	78.3
	EMO	**81.3**
STS12	LLM2Vec Mistral 7B SFT (Teacher)	80.42
	Bert SFT (Student)	49.7
	DSKD (Zhang et al., 2024)	65.3
	EMO	**75.3**
SICK	LLM2Vec Mistral 7B SFT (Teacher)	88.9
	Bert SFT (Student)	61.1
	DSKD (Zhang et al., 2024)	78.7
	EMO	**80.1**

Table 3: Ablation study results showing the impact of Intra-Model Relation Distillation (IRA) and Optimal Transport with Importance-Scored Mass Assignment (OTIS). “w/o” denotes “without”.

Dataset	Method	Accuracy	Precision	Recall	Spearman
Patent	Bert SFT	63.1	58.7	54.4	-
	EMO$_{\text{w/o OTIS}}$	65.3	61.8	60.9	-
	EMO$_{\text{w/o IRA}}$	64.7	59.8	59.2	-
	EMO	**66.5**	**63.3**	**62.4**	-
SciTail	Bert SFT	88.1	87.7	88.8	-
	EMO$_{\text{w/o OTIS}}$	88.8	88.1	89.5	-
	EMO$_{\text{w/o IRA}}$	87.2	86.9	88.3	-
	EMO	**90.9**	**90.1**	**91.2**	-
STSB	Bert SFT	-	-	-	75.1
	EMO$_{\text{w/o OTIS}}$	-	-	-	80.9
	EMO$_{\text{w/o IRA}}$	-	-	-	78.5
	EMO	-	-	-	**81.3**

both yields the best results, indicating their complementary roles: IRA preserves internal relational structure, which OTIS then effectively aligns across models for superior knowledge transfer.

Impact of top-m Important Token Selection We analyze the effect of top-m, the number of top salient tokens selected from the n one-to-one MinED-aligned tokens in IRA. We test $m \in \{n, \lfloor n/2 \rfloor, \lfloor n/3 \rfloor\}$ on STS tasks. Table 4 shows that selecting a suitable subset of top salient tokens ($m = \lfloor n/3 \rfloor$) can achieve the best performance, suggesting that focusing CKA on the attention patterns of the most critical tokens is effective and potentially reduces noise from less important ones. Across all tested values of m, our method consistently outperforms the DSKD baseline, underscoring the benefit of structural attention distillation.

Robustness to Teacher Model Choice To evaluate framework robustness beyond the LLM2Vec-Mistral 7B teacher, we repeated experiments on SciTail using BGE-M3 (Chen et al., 2024) as the

Table 4: Impact of top-m salient tokens for CKA attention distillation on STS tasks (Pearson ρ), n: total one-to-one aligned tokens.

Dataset	$m = n$	$m = \lfloor n/2 \rfloor$	$m = \lfloor n/3 \rfloor$	DSKD (Baseline)
STSB	80.2	80.6	81.3	78.3
SICK	76.9	79.4	80.1	78.7

teacher. Table 5 shows our framework **EMO** maintains significant performance gains over both the Bert SFT baseline and all compared cross-tokenizer distillation methods (ULD, MinED, DSKD, MultilevelOT), even with this different teacher. This result indicates our framework's general applicability and effectiveness in transferring knowledge from various high-performing embedding models.

Table 5: Results on SciTail using BGE-M3 as the teacher model.

Method	Accuracy	Precision	Recall
BGE SFT (Teacher)	94.3	94.1	93.9
Bert SFT (Student)	88.1	87.7	88.8
ULD (Boizard et al., 2025)	91.2	90.6	91.9
DSKD (Zhang et al., 2024)	91.5	91.2	91.0
MinED (Wan et al., 2024)	89.5	88.9	90.4
MultilevelOT (Cui et al., 2024)	91.4	91.2	90.9
EMO	**92.7**	**91.8**	**92.3**

6 Conclusion

We address the challenge of CTKD for text embedding models by proposing a new method **EMO**. Our framework distills intra-model relational knowledge via CKA on attention matrices of aligned tokens and aligns final hidden states using Optimal Transport with Importance-Scored Mass Assignment. Experiments show our method significantly outperforms existing CTKD baselines on diverse embedding tasks. By preserving attention structure and prioritizing key representations, our approach yields smaller, high-fidelity embedding models despite tokenizer differences, offering a more comprehensive solution for effective CTKD.

7 Limitations

While our EMO framework demonstrates significant efficacy, its current focus on 1-to-1 mapped tokens via MinED and subsequent top-m selection for intra-model attention distillation inherently means some tokens are excluded from this specific structural alignment stage. This could potentially lead to a loss of nuanced relational information associated with these unaligned or less salient tokens, although the final OTIS stage considers full representations. Evaluating the precise impact of this exclusion and developing strategies to softly incorporate these tokens into attention distillation are avenues for future exploration. Future research will investigate more dynamic and context-aware methods for selecting the top-m important tokens, moving beyond static final-layer attention scores.

Acknowledgements

Trung Le was partly supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4044. Thien Huu Nguyen has been supported by the NSF grant # 2239570. He is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation therein.

References

Nguyen Hoang Anh, Quyen Tran, Thanh Xuan Nguyen, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. 2025. Mutual-pairing data augmentation for fewshot continual relation extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4057–4075.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. 2025. Towards cross-tokenizer distillation: the universal logit distillation loss for llms.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at? an analysis of bert's attention.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. 2024. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. *arXiv preprint arXiv:2412.14528*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transportation distances.

Sayantan Dasgupta and Trevor Cohn. 2025. Improving language model distillation through hidden state matching. In *The Thirteenth International Conference on Learning Representations*.

Chaochen Gao, Xing Wu, Peng Wang, Jue Wang, Liangjun Zang, Zhongyuan Wang, and Songlin Hu. 2023. Distilcse: Effective knowledge distillation for contrastive sentence embeddings.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling BERT for natural language understanding. *CoRR*, abs/1909.10351.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited.

Anh Duc Le, Nam Le Hai, Thanh Xuan Nguyen, Linh Ngo Van, Nguyen Thi Ngoc Diep, Sang Dinh, and Thien Huu Nguyen. 2025. Enhancing discriminative representation in similar relation clusters for few-shot continual relation extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2450–2467.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models.

Junyan Li, Li Lyna Zhang, Jiahang Xu, Yujing Wang, Shaoguang Yan, Yunqing Xia, Yuqing Yang, Ting Cao, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2023. Constraint-aware and ranking-distilled token pruning for efficient transformer inference. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1280–1290. ACM.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2023. Less is more: Task-aware layer-wise distillation for language model compression.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. 2025. Rho-1: Not all tokens are what you need.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context - investigating contextual reasoning over long texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13388–13396.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark.

Khai Nguyen. 2025. An introduction to sliced optimal transport. *arXiv preprint arXiv:2508.12519*.

Toan Ngoc Nguyen, Nam Le Hai, Nguyen Doan Hieu, Dai An Nguyen, Linh Ngo Van, Thien Huu Nguyen, and Sang Dinh. 2025. Improving vietnamese-english cross-lingual retrieval for legal and general domains. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 142–153.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Thanh Duc Pham, Nam Le Hai, Linh Ngo Van, Nguyen Thi Ngoc Diep, Sang Dinh, and Thien Huu Nguyen. 2025. Mitigating non-representative prototypes and representation bias in few-shot continual relation extraction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10791–10809.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression.

Cédric Villani. 2008. *Optimal transport – Old and new*, volume 338, pages xxii+973.

Tu Vu, Manh Do, Tung Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. Topic modeling for short texts via optimal transport-based clustering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7666–7680.

Hoang Tran Vuong, Tue Le, Tu Vu, Tung Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. Hicot: Improving neural topic models via optimal transport and contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13894–13920.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge fusion of large language models.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. 2024. Dual-space knowledge distillation for large language models.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models.

Appendix

A Addressing Potential Token Loss

A consequence of focusing solely on these n tokens derived from one-to-one mappings is that the remain tokens are initially excluded from the relational distillation process described in section 3.1. It is important to note that while these one-to-one mapped tokens are excluded from this specific attention distillation stage, the subsequent stage involving Optimal Transport for representation alignment (Section 3.2 operates on the full sequences or representations derived thereof, ensuring that information from all original tokens contributes to the overall distillation objective. To quantify the extent of this one-to-one mapping achievable in practice, we conducted experiments on the SciTail dataset (Khot et al., 2018), Control dataset (Liu et al., 2021) and Anli_r2 dataset (Nie et al., 2020) . We applied the MinED mapping procedure between the tokenizations produced by our student (Bert-base-uncased) and teacher (LLM2Vec-Mistral 7B) models across the train, development, and test splits. The percentage of tokens participating in the resulting 1-to-1 mapping is presented in Table 6.

Table 6: one-to-one token mapping rates (%) found by MinED between student (Bert-base-uncased) and teacher (LLM2Vec-Mistral) tokenizations.

Dataset	Teacher Mapping (%)	Student mapping (%)
SciTail	71.03	80.22
ConTRoL-nli	68.86	75.48
Anli_r2	66.37	74.47

As observed in Table 6, the proportion of tokens that can be directly mapped one-to-one between the two distinct tokenizers is substantial, generally ranging from 66-71 % for the teacher sequence and 74-80 % for the student sequence across different data splits. An average mapping rate of approximately 75% (considering both models) indicates that a significant majority of tokens have a direct counterpart found by MinED.

B Experimental Details

Table 7: Detailed training configurations

LLM2Vec Mistral 7B → Bert-base-uncased	
Epoch	5
LR	1×10^{-5}
Batch Size	4
LR Scheduler	cosine
Finetune method	LoRA
LoRA rank	256
LoRA alpha	32
LoRA dropout	0.1

Models Our student model is the standard Bert-base-uncased (110M parameters). The teacher model is LLM2Vec Mistral 7B (BehnamGhader et al., 2024), a state-of-the-art text embedding model from the MTEB leaderboard. The detail of each models training configurations in KD in Table 7.

Training and Evaluation For distillation, the student model (Bert-base-uncased) is fully finetuned. The teacher model (LLM2Vec Mistral 7B) is fine-tuned using LoRa. For Text Classification and Sentence Pair Classification tasks, we report standard metrics: Accuracy, Precision, Recall, and F1-Score. For

STS tasks, we report the Spearman Correlation Coefficient between model predictions and ground-truth similarity scores.

Table 8: The best-searched hyperparameters α for different configurations.

Method	Patent	Imdb	Banking77	Scitail	ConTRoL-nli	Anli_r2	STSB	STS12	SICK
OURS	0.5	0.1	0.1	0.5	0.5	0.5	0.5	0.5	0.5

Detailed Dataset Statistics Table 9 presents the sample counts for the training, validation, and test splits across each domain-specific dataset.

Table 9: Dataset Statistics

Dataset	**Train**	**Validation**	**Test**
Patent	25000	5000	5000
Imdb	25000	-	25000
Banking77	10000	-	3080
SciTail	23100	1300	2130
ConTRoL-nli	6720	799	805
Anli_r2	45500	1000	1000
STSB	5750	1500	1380
STS12	2230	-	3110
SICK	4500	500	4823

Hyperparameter We explored the hyperparameter α over the set $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.9$, and the optimal value for each experimental setting is reported in Table 8.

C Ablation study about the computational overhead analysis

The computational overhead analysis

Table 10: Training time per batch for each method on the SciTail dataset.

Method	ULD	DSKD	MinED	MultiOT	IRA (for 2 layers)	OTIS	EMO
Time (s)	0.18	0.31	0.67	0.72	0.41	0.56	0.97

As shown in Table 10, our full EMO framework (apply IRA for 2 layers and OTIS for 1 layer) requires approximately 0.97 seconds per batch in our experimental setup. This overhead arises mainly during training and is not counted in the inference or deployment stage, making it a reasonable trade-off for the substantial gains in the student model's performance and representation fidelity. It is also worth noting that OTIS alone incurs a relatively high training time of 0.56 seconds per batch, even though it is applied only to the final hidden layer. This observation highlights the significant computational burden of Optimal Transport, and thus justifies our design choice of restricting OTIS to the last layer to balance effectiveness with efficiency.