TALAS: Teacher-Anchored Layer Alignment with Adaptive Sharpness-Aware Minimization for Embedding Distillation

Quoc Phong Dao[1]*, Hoang Son Nguyen[1]*, Pham Khanh Chi[1]*,
Linh Ngo Van[1,†], Diep Thi-Ngoc Nguyen[2], Thien Huu Nguyen[3], Trung Le[4]

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

Abstract

Knowledge Distillation (KD) has established itself as a pivotal technique for compressing large pre-trained language models. However, existing methods that force a student to strictly mimic the teacher's sentence embeddings or internal features often incur prohibitive computational costs and yield suboptimal performance due to the inherent capacity gap. To address these challenges, we propose **TALAS** (**T**eacher-**A**nchored **L**ayer **A**lignment with **S**harpness-aware minimization), a unified framework that synergizes hierarchical (multi-layer) alignment with robust optimization. First, we introduce a Teacher-Anchored mechanism that selectively distills final sentence embeddings only into the student's upper layers, thereby reducing overhead while respecting capacity constraints. Second, we bridge the semantic gap in lower layers via Layer-Aligned Self-Distillation, which propagates knowledge top-down using internal geometric relational constraints in the embedding space. Finally, to prevent the student from memorizing point-wise teacher noise, we integrate Adaptive Sharpness-Aware Minimization (ASAM) into the training objective, guiding the model towards flat minima for enhanced generalization. Empirical results on standard sentence embedding benchmarks demonstrate that TALAS consistently outperforms strong distillation baselines while achieving superior training efficiency in terms of computational cost and memory footprint.

1 Introduction

With the rapid advancement of natural language processing, text embedding models have become a core component of Retrieval-Augmented Generation (RAG) systems, enabling effective semantic retrieval to supply relevant context for large language models (Gao et al., 2024; Zhao et al., 2025c). They play a critical role in downstream tasks such as information retrieval, question answering, and text classification by facilitating meaning-based comparison and representation of textual data (Ramesh Kashyap et al., 2024). Despite their strong performance on benchmarks such as MTEB (Muennighoff et al., 2023) state-of-the-art text embedding models are often characterized by large parameter counts and high-dimensional representations, which substantially increase computational and deployment costs.

In parallel, knowledge distillation (Hinton et al., 2015) has emerged as a fundamental technique for compressing large neural models, with recent advances primarily focusing on large language models. However, distillation methods specifically tailored to embedding models - despite their central role in many downstream tasks (Ramesh Kashyap et al., 2024) - remain comparatively less explored and less focused. A key challenge in embedding model distillation lies in the trade-off between effectiveness and efficiency. Approaches that exploit internal representations such as hidden states or attention patterns provide rich supervision and have been shown to achieve strong performance in knowledge transfer settings (Sun et al., 2019; Jiao et al., 2020; Wang et al., 2020; Passban et al., 2020; Zhang et al., 2024a; Dasgupta and Cohn, 2025; Zhao et al., 2025a). However, when the teacher is a large LLM-based embedding model, leveraging such intermediate signals typically incurs substantial computational and memory overhead, which limits their practicality in large-scale or resource-constrained scenarios. In contrast, methods that rely solely on the teacher's output representations offer a significantly more efficient training pipeline and are particularly attractive when the teacher is only accessible through inference (Hinton et al., 2015; Ko et al., 2024; Anshumann et al., 2025; Vu et al., 2026a). Nevertheless, compressing rich semantic information into a single output embedding, combined with the capacity gap between teacher

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

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32705–32723

and student models, can lead to the student failing to capture all of the important nuances present in the teacher's distribution; instead, the student may focus on more dominant or easier aspects of the teacher's outputs, as discussed in prior work on distillation objectives for generative models (Gu et al., 2024).

A recurring obstacle in knowledge distillation is the capacity gap between a high-capacity teacher and a compact student: when the student lacks sufficient expressiveness, a stronger teacher does not necessarily produce a better distilled model and can even degrade performance (Cho and Hariharan, 2019). This issue becomes particularly salient for embedding distillation with LLM-based teachers, where directly forcing the student to mimic the teacher across the network can be unstable - especially in shallow layers that primarily encode low-level linguistic features - because aligning them to highly abstract teacher representations may disrupt feature extraction and hinder optimization. Prior work has proposed using intermediate assistant models to help bridge the gap between large teachers and compact students and to stabilize the distillation process by gradually reducing representational differences through one or more auxiliary networks (Mirzadeh et al., 2019b; Son et al., 2021; Zhou and Ai, 2024). Inspired by this principle, we adopt a progressive bridging strategy: we restrict teacher-anchored supervision to the student's upper layers while using the student's own higher layers as a lightweight "assistant" signal to guide lower layers through sequential, top-down alignment. This design directly targets the capacity-gap bottleneck while remaining compatible with resource-constrained settings where accessing or storing teacher internals is often impractical. Moreover, prior studies suggest that sharp minima in the loss landscape are closely associated with larger generalization gaps, motivating us to use Adaptive Sharpness-Aware Minimization (ASAM) (Kwon et al., 2021) of loss-surface sharpness as a diagnostic and optimization signal. In the context of embedding distillation, where generalization across domains and tasks is crucial, controlling sharpness provides a promising direction for balancing efficiency and performance.

Our main contributions include:

- We propose **TALAS** (**T**eacher-**A**nchored **L**ayer **A**lignment with **S**harpness-aware minimization) - a resource-efficient unsupervised knowledge distillation framework for embedding models that aligns multiple student layers with teacher output embeddings, eliminating the need for teacher inference during training.
- We introduce a layer-aligned self-distillation mechanism that propagates relational structure across student layers without relying on teacher hidden states. Additionally, we incorporate adaptive sharpness-aware optimization to mitigate the capacity gap between large teacher models and compact students, improving generalization.
- We empirically demonstrate effective unsupervised distillation from large LLM-based embedding models to BERT-based students using only cached teacher embeddings.

2 Background

This section introduces the background concepts underlying our approach. We first review the fundamentals of knowledge distillation, followed by Adaptive Sharpness-Aware Minimization (ASAM). An extended discussion of related work is in appendix A.

Knowledge Distillation. Knowledge Distillation aims to transfer knowledge from a teacher model T to a student model S by encouraging the student to approximate the teacher's behavior under a chosen supervision signal. In general, given an input x, the teacher produces an output $z_T(x)$ and the student produces a corresponding output $z_S(x)$, where $z(\cdot)$ may represent logits, hidden representations, or embeddings depending on the task. The student is trained by minimizing a distillation loss of the form

$$\mathcal{L}_{\text{KD}} = \mathcal{D}\big(z_S(x), z_T(x)\big), \tag{1}$$

where $\mathcal{D}(\cdot,\cdot)$ denotes a discrepancy measure such as cross-entropy, mean squared error, or a similarity-based distance.

Adaptive Sharpness-Aware Minimization. Adaptive Sharpness-Aware Minimization (ASAM) is an optimization framework designed to improve generalization by encouraging convergence to flat minima. Instead of minimizing the empirical loss $\mathcal{L}(w)$ directly, ASAM solves a local min-max problem:

$$\min_{w} \max_{\|\mathbf{T}_w^{-1}\epsilon\|_2 \le \rho} \mathcal{L}(w+\epsilon), \tag{2}$$

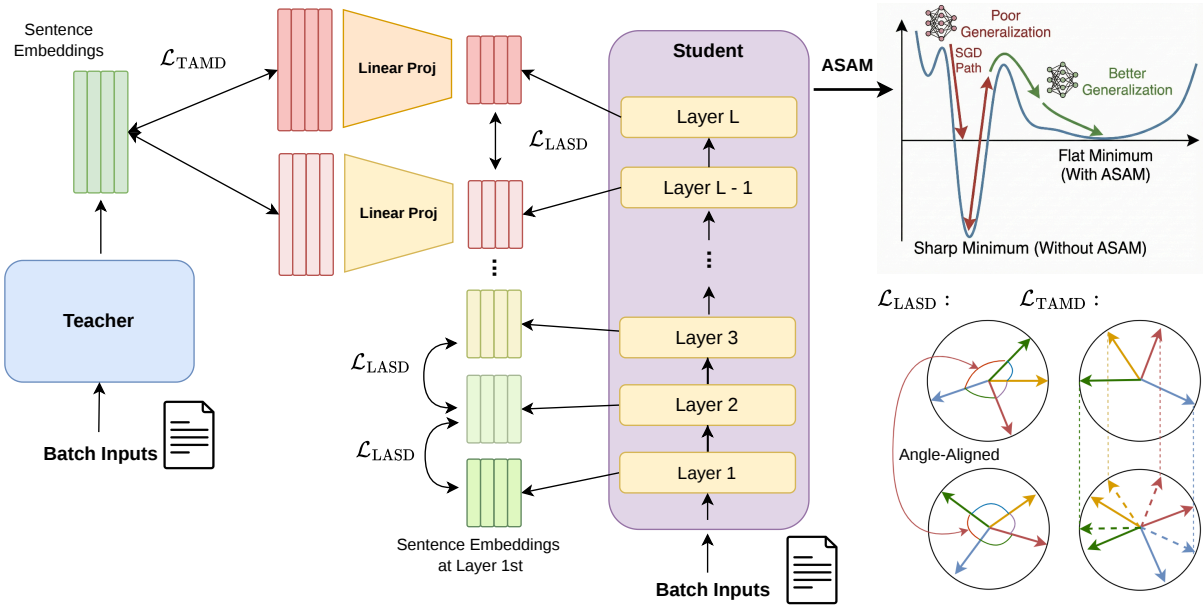

Figure 1: An illustration of the proposed TALAS framework. The training process synergizes two complementary distillation strategies: (1) **Teacher-Anchored Multi-Layer Distillation**($\mathcal{L}_{\text{TAMD}}$) aligns the student's upper layers with the teacher's stationary embedding via learnable linear projections; and (2) **Layer-Aligned Self-Distillation** ($\mathcal{L}_{\text{LASD}}$) enforces geometric structural consistency between adjacent student layers. The entire training is optimized with **Adaptive Sharpness-Aware Minimization (ASAM)** to seek flat minima for better generalization.

where w denotes model parameters, ϵ is an adversarial perturbation, ρ controls the neighborhood size, and $\mathbf{T}_w$ is a parameter-dependent scaling matrix that normalizes perturbations according to parameter magnitude. By accounting for parameter scale, ASAM provides a more stable and adaptive variant of sharpness-aware optimization. This property is especially beneficial in knowledge distillation settings with strong supervision signals or large teacher-student capacity gaps.

3 Methodology

In this section, we present **TALAS**, a unified framework for robust embedding distillation under teacher–student capacity mismatch. TALAS decomposes the training objective into three complementary components. Section 3.1 introduces Teacher-Anchored Distillation, which defines *where* semantic alignment with the teacher is applied. Section 3.2 presents Layer-Aligned Self-Distillation, which governs *how* semantic knowledge is propagated across layers. Finally, Section 3.3 integrates Adaptive Sharpness-Aware Minimization (ASAM) to enable robust optimization. An overview is shown in Figure 1.

Problem Setup. Let $\mathcal{T}$ and $\mathcal{S}$ denote the teacher and student models, respectively. We define $\mathbf{e}^{\mathcal{T}}$ and $\mathbf{e}^{\mathcal{S}}$ as their corresponding sentence embeddings. For the student, let $\mathbf{e}_l^{\mathcal{S}}$ represent the pooled embedding at layer l. Depending on the backbone architecture, these representations are derived via specific pooling strategies, such as `[CLS]` (Devlin et al., 2019), `[EOS]` (Zhang et al., 2025b), or mean pooling (Lee et al., 2024).

3.1 Teacher-Anchored Multi-Layer Embedding Distillation

Unlike prior distillation methods that rely on complex internal signals such as attention maps and hidden states (e.g., EOFD (Zhao et al., 2025b), FDD (Gong et al., 2025)), our approach focuses exclusively on sentence embedding distillation, significantly reducing computational overhead. To compensate for the reduced supervision and maximize semantic transfer, we draw inspiration from the representation compression principle proposed in ESE (Li et al., 2025), which advocates for encoding salient semantics into lower network layers. Accordingly, we aim to guide both intermediate and final layers of the student to align with the teacher. However, acknowledging the inherent capacity mismatch, directly forcing alignment across all layers can be suboptimal. Therefore, we apply Teacher-Anchored distillation strictly to the upper layers, where the student possesses sufficient capacity to

approximate the teacher's embedding space.

Formally, we treat the teacher's final embedding as a stationary *anchor* to supervise the student's representations across multiple upper layers. The teacher-anchored distillation objective is formulated as follows:

$$\mathcal{L}_{\text{TAMD}}(l) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{D}_{\cos}\left(\mathbf{e}_{l,i}^{\mathcal{S}}\mathbf{W}_l, \mathbf{e}_i^{\mathcal{T}}\right)$$
$$\mathcal{L}_{\text{TAMD}} = \frac{1}{k+1}\sum_{l=L-k}^{L} \mathcal{L}_{\text{TAMD}}(l) \quad (3)$$

where L is the total number of student layers, N is the batch size, $\mathcal{D}_{\cos}(u, v) = 1 - \frac{u^\top v}{\|u\|\|v\|}$ denotes the cosine distance and $\mathbf{W}_l \in \mathbb{R}^{d_S \times d_T}$ is a learnable projection matrix at l-th layer.

In this formulation, $\mathbf{e}^{\mathcal{T}}$ serves as a shared semantic anchor that guides both intermediate and final representations. By anchoring multiple upper layers to the same teacher embedding, the student is encouraged to progressively align its semantic space with that of the teacher, rather than relying solely on supervision at the final layer. This design facilitates *early semantic injection* during the forward pass, enabling intermediate layers to acquire high-level semantic information before it is fully consolidated at the output layer. Compared to single-layer distillation, such multi-layer anchoring stabilizes training and reduces representation drift across layers.

3.2 Layer-Aligned Embedding Self-Distillation

While Section 3.1 anchors the upper layers to the teacher's semantic space, extending this supervision to lower layers is non-trivial. Prior study on layer-wise distillation (Sun et al., 2019) have shown that directly forcing shallow student layers to match highly abstract teacher representations under large capacity gaps can destabilize training and degrade feature learning. Early layers primarily encode low-level linguistic patterns and directly forcing them to approximate the teacher's highly abstract representations can distort their feature extraction behavior. At the same time, leaving the lower layers entirely unsupervised introduces a representational discontinuity across depth: the abrupt transition from unsupervised shallow layers to teacher-anchored upper layers yields structurally misaligned internal geometries, impeding the smooth propagation of semantic structure through the network.

To overcome this limitation, we propose a progressive self-distillation mechanism that propagates knowledge sequentially from top to bottom. Instead of mapping lower layers directly to the teacher, we utilize the student's own upper layer $(l+1)$ as a dynamic "guide" for the immediate lower layer (l). This chain-based approach ensures a smooth semantic transition, allowing deep layers to maintain high-level abstractions while gradually adjusting shallow layers to the target embedding space without destroying their intrinsic structural properties.

We formulate this process as a geometric structural consistency between adjacent student layers problem using Relational Knowledge Distillation (RKD). Unlike point-wise distillation, RKD focuses on preserving the structural relations between data samples within a batch. Let $\mathbf{E}_l^{\mathcal{S}} \in \mathbb{R}^{N\times d}$ denote the batch of embeddings at layer l of student model, where N is the batch size. We first apply L_2 normalization to obtain $\tilde{\mathbf{E}}_l^{\mathcal{S}}$ and compute the relational matrix $\mathbf{R}_l^{\mathcal{S}} \in \mathbb{R}^{N\times N}$, which captures the pair-wise similarity structure:

$$\mathbf{R}_l^{\mathcal{S}} = \tilde{\mathbf{E}}_l^{\mathcal{S}}(\tilde{\mathbf{E}}_l^{\mathcal{S}})^\top, \quad (4)$$

where each element $(\mathbf{R}_l^{\mathcal{S}})_{ij}$ corresponds to the cosine similarity between the i-th and j-th samples, i.e., $\cos(\mathbf{e}_i, \mathbf{e}_j)$.

Here, $\mathbf{R}_l^{\mathcal{S}}$ represents the geometric topology of the embedding space at layer l. Our objective is to minimize the structural divergence between adjacent layers. The Layer-Aligned Self-Distillation loss is defined using the Frobenius norm of the difference between relation matrices:

$$\mathcal{L}_{\text{LASD}} = \frac{1}{L-1}\sum_{l=1}^{L-1}\frac{1}{N^2}\left\|\mathbf{R}_{l+1}^{\mathcal{S}} - \mathbf{R}_l^{\mathcal{S}}\right\|_F^2, \quad (5)$$

where N is the batch size and $\|\mathbf{A}\|_F^2 = \sum_{i,j} A_{ij}^2$ denotes the squared Frobenius norm, serving as a matrix-level Mean Squared Error proxy.

By optimizing Eq. 5, the student model creates a seamless *semantic bridge*, where the teacher's knowledge (anchored at the top) is progressively distilled down to the input level through structural alignment, bridging the capacity gap effectively.

3.3 Training Objective and Optimization

In this section, we describe the overall training objective and optimization strategy for TA-

LAS. Our design combines complementary objectives that promote semantic alignment and well-structured representations, together with an optimization scheme tailored to teacher-student distillation, with the goal of improving generalization under a capacity gap.

Overall Training Objective. The objectives introduced in Sections 3.1 and 3.2 focus on aligning the student with the teacher's semantic space. However, exclusive reliance on such direct supervision carries a risk: if the teacher's representations exhibit collapse or high anisotropy within the specific training domain, the student may inherit these suboptimal geometric properties, potentially leading to a restricted embedding space (Ethayarajh, 2019; Li et al., 2020). In contrast, contrastive learning mitigates such collapse by optimizing for *uniformity* (Wang and Isola, 2020), which encourages distinct samples to separate and populate the hypersphere evenly. This prevents the representation space from degenerating into anisotropic clusters, maximizing the model's discriminative power. Therefore, we incorporate an unsupervised contrastive objective $\mathcal{L}_{\text{SimCSE}}$ based on SimCSE (Gao et al., 2022) into our framework. This term serves as a regularizer that complements the distillation process: it ensures the student benefits from the teacher's semantic guidance while simultaneously maintaining a broader, more uniform representation space, thereby enhancing generalization.

Given a batch of input sentences, we feed them through the student encoder twice with different standard dropout masks z, z' to obtain two views of embeddings $\mathbf{e}_i^z$ and $\mathbf{e}_i^{z'}$. The loss is formulated as:

$$\mathcal{L}_{\text{SimCSE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\text{sim}(\mathbf{e}_i^z,\mathbf{e}_i^{z'})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{e}_i^z,\mathbf{e}_j^{z'})/\tau}}, \quad (6)$$

where N is the batch size, τ is the temperature hyperparameter, and $\text{sim}(\cdot)$ denotes cosine similarity.

The final training objective is defined as a weighted sum of three terms:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{SimCSE}} + \lambda_2\mathcal{L}_{\text{TAMD}} + \lambda_3\mathcal{L}_{\text{LASD}}, \quad (7)$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters balancing the contributions of contrastive learning, teacher-anchored distillation, and self-distillation, respectively.

Algorithm 1: TALAS Training via ASAM

Input: Training corpus $\mathcal{D}$, Teacher $f_{\mathcal{T}}$, Student $f_{\mathcal{S}}(\cdot;\mathbf{w})$, learning rate η, ASAM radius ρ

Pre-compute and cache teacher embeddings: $\mathbf{D}_e^{\mathcal{T}} \leftarrow \{f_{\mathcal{T}}(x)_{\text{anchored}}\}_{x\in\mathcal{D}}$

for *each training step* **do**

  Sample mini-batch $\mathcal{B} \subset \mathcal{D}$

  Retrieve cached targets $\mathbf{E}^{\mathcal{T}} \in \mathbf{D}_e^{\mathcal{T}}$ for $\mathcal{B}$

  `// 1. Ascent Step`

  Compute standard loss following Equation: $\mathcal{L} \leftarrow \mathcal{L}_{\text{total}}(f_{\mathcal{S}}(\mathcal{B};\mathbf{w}), \mathbf{E}^{\mathcal{T}})$

  Generate perturbed weights $\tilde{\mathbf{w}}$: $\tilde{\mathbf{w}} \leftarrow \text{ASAM\_Perturb}(\mathbf{w}, \nabla_{\mathbf{w}}\mathcal{L}, \rho)$

  `// 2. Descent Step`

  Compute gradients at the position $\tilde{\mathbf{w}}$: $\mathbf{g}_{\text{ASAM}} \leftarrow \nabla_{\tilde{\mathbf{w}}}\mathcal{L}_{\text{total}}(f_{\mathcal{S}}(\mathcal{B};\tilde{\mathbf{w}}), \mathbf{E}^{\mathcal{T}})$

  Update original parameters: $\mathbf{w} \leftarrow \mathbf{w} - \eta \cdot \mathbf{g}_{\text{ASAM}}$

end

Optimization via ASAM. While the proposed objectives effectively align student representations with the teacher, optimizing the distillation loss using standard SGD can be risky under a teacher-student capacity gap. Cosine-based embedding distillation imposes strong instance-level constraints, which may cause the student to overfit the teacher's outputs without capturing the underlying data structure. As shown by Stanton et al. (2021), such overfitting can result in high fidelity to the teacher but poor generalization.

To mitigate this issue, we optimize $\mathcal{L}_{\text{total}}$ using Adaptive Sharpness-Aware Minimization (ASAM) (Kwon et al., 2021). Following the standard ASAM formulation, we seek solutions that remain stable within a local neighborhood by solving the corresponding min-max problem in Eq. 2, which is efficiently optimized via a two-step ascent–descent procedure (Algorithm 1).

In our distillation setting, ASAM encourages convergence to flatter solutions, helping the student learn more robust and transferable representations under a capacity gap.

	Classification (F1)			Pair Classification (AP)			STS (Spearman)			Domain Avg		
	Banking77	**Tweet**	**Emotion***	**MRPC**	**SciTail**	**WiC***	**SICK**	**STS12**	**STSB***	**Avg-In**	**Avg-Out**	**Avg.**
Qwen3-Embedding 0.6B → MiniLMv2 H384 22M												
Teacher	93.21	71.33	66.53	83.37	90.00	66.85	80.64	77.11	84.60	72.66	82.61	79.29
Student base	68.29	41.07	69.90	78.39	64.56	59.58	47.60	21.95	22.08	40.91	58.45	52.60
SimCSE-unsup	86.67	69.56	54.00	82.29	73.76	65.71	66.04	59.11	62.08	60.60	72.91	68.80
CDM	86.07	70.32	53.04	81.97	72.97	68.34	65.97	61.12	65.75	62.38	73.07	69.51
DSKD	85.66	69.86	52.25	82.17	73.52	68.58	66.54	62.52	66.70	62.51	73.38	69.76
Jasper and Stella	85.94	71.25	57.99	83.36	76.17	68.25	70.57	63.81	70.11	65.45	75.18	71.94
DistillCSE	**88.03**	69.87	54.24	83.20	77.66	67.88	70.09	68.19	71.53	64.55	76.17	72.30
EMO	87.03	71.67	59.63	83.45	78.17	**68.89**	70.81	65.78	71.42	66.65	76.15	72.98
TALAS	86.69	**72.77**	**60.94**	**85.10**	**81.42**	66.65	**72.51**	**70.77**	**76.22**	**67.94**	**78.21**	**74.79**
BGE-M3 → MiniLMv2 H768 66M												
Teacher	93.52	73.85	68.56	85.81	91.87	61.47	79.18	78.73	84.87	71.63	83.83	79.76
Student base	81.84	47.70	71.67	79.67	65.06	60.89	49.46	26.76	24.12	44.24	62.41	56.35
SimCSE-unsup	87.89	71.45	55.65	83.87	76.32	67.96	66.93	59.66	63.97	62.53	74.35	70.41
CDM	89.15	70.48	58.71	84.16	75.91	**70.00**	68.14	62.74	67.78	65.50	75.10	71.90
DSKD	89.51	70.53	57.27	84.39	75.36	69.96	68.81	66.78	70.05	65.76	75.90	72.52
Jasper and Stella	88.38	74.13	62.81	85.98	80.33	68.38	74.35	66.80	74.55	68.58	78.33	75.08
DistillCSE	90.50	72.70	59.88	84.39	78.48	69.49	73.97	70.00	73.73	67.70	78.34	74.79
EMO	90.19	74.26	61.48	85.31	80.82	68.30	75.41	67.06	76.35	68.71	78.84	75.46
TALAS	**90.62**	**74.81**	**63.30**	**87.05**	**81.94**	66.28	**76.86**	**69.04**	**79.01**	**69.53**	**80.05**	**76.55**
Qwen3-Embedding 4B → Bert-base 109M												
Teacher	93.63	72.22	68.94	84.79	91.77	66.80	83.43	82.55	86.87	74.20	84.73	81.22
Student base	84.86	47.79	68.02	77.36	67.74	59.22	42.43	21.54	20.30	42.44	60.33	54.36
SimCSE-unsup	89.21	69.13	53.55	82.86	76.59	71.64	68.19	61.68	70.36	65.18	74.61	71.47
CDM	89.75	70.58	53.11	84.83	75.19	71.06	69.37	68.48	73.92	66.03	76.37	72.92
DSKD	89.66	69.93	54.51	83.13	77.95	**71.53**	68.65	63.37	71.42	65.82	75.45	72.24
Jasper and Stella	89.71	73.83	65.98	85.33	80.99	70.50	73.96	67.90	75.65	70.71	78.62	75.98
DistillCSE	90.94	70.90	60.80	82.86	78.98	68.63	75.12	72.61	77.08	68.84	78.57	75.32
EMO	90.78	73.54	67.48	84.46	81.57	69.76	75.66	68.83	77.17	71.47	79.14	76.58
TALAS	**91.43**	**74.30**	**70.71**	**86.47**	**82.66**	69.24	**78.38**	**75.40**	**80.88**	**73.61**	**81.44**	**78.83**

Table 1: Experimental results on nine datasets. Datasets marked with ⋆ are in-domain, while the remaining datasets are out-of-domain.

4 Experiments

4.1 Experimental Setup

Datasets. Our experiments span three task categories: **Classification**, **Pair Classification**, and **Semantic Textual Similarity (STS)**. For **Classification**, we report F1 scores on BANKING77 (Casanueva et al., 2020), EMOTION (Saravia et al., 2018), and TWEETEVAL-SENTIMENT (Barbieri et al., 2020). For **Pair Classification**, we evaluate Average Precision (AP) on MRPC (Dolan and Brockett, 2004), SCITAIL (Khot et al., 2018), and WIC (Pilehvar and Camacho-Collados, 2019). For **STS**, we report Spearman's rank correlation on SICK (Marelli et al., 2014), STS12 (Agirre et al., 2012), and STS-BENCHMARK (Cer et al., 2017).

We follow the task selection protocol of EMO (Truong et al., 2025) but differ in the training paradigm. While EMO employs supervised fine-tuning (SFT), we adopt an *unsupervised SimCSE-style training objective* (Gao et al., 2022) to encourage the student model to learn general-purpose semantic representations without relying on task-specific labels.

For training, we uniformly sample **5,000 sentences from each of three in-domain datasets** and merge them into a single unlabeled corpus that is shared across all teacher–student configurations. Specifically, EMOTION, WIC, and STS-BENCHMARK are treated as *in-domain*, while the remaining datasets are considered *out-of-domain*. The resulting corpus is evaluated on all nine downstream benchmarks, yielding a balanced and domain-diverse training setup while strictly avoiding any form of supervised signal.

Training and Evaluation Settings. The student models considered in our experiments include BERT-base (Devlin et al., 2019), MINILMV2 H384, and MINILMV2 H768 (Gu et al., 2024). For teacher models, we use BGE-M3 (Chen et al., 2024), QWEN3-EMBEDDING-0.6B, and QWEN3-

EMBEDDING-4B (Zhang et al., 2025b), and systematically evaluate multiple teacher → student distillation configurations.

Detailed descriptions of the model architectures, data preprocessing procedures, training hyperparameters, and baseline implementations are provided in Appendix D and Appendix E.

Baselines. To rigorously evaluate the proposed framework, we compare it against representative knowledge distillation methods spanning different levels of granularity. For **token-level KD**, we include DSKD (Zhang et al., 2024c), CDM (Chen et al., 2025), and EMO (Truong et al., 2025); these approaches transfer knowledge by aligning fine-grained token-level hidden representations or relational structures between teacher and student models. For **sentence-level KD**, we benchmark against JASPER & STELLA (Zhang et al., 2025a) and DISTILLCSE (Xu et al., 2023), state-of-the-art methods that distill knowledge exclusively via global sentence embeddings without relying on token-wise supervision.

4.2 Main Results

To assess the effectiveness of *TALAS*, we report the main experimental results with a primary focus on **embedding generalization**. Specifically, we evaluate whether the proposed method improves the robustness and transferability of sentence representations across diverse downstream tasks, thereby validating the semantic quality of the distilled student models compared to strong baselines. Comprehensive ablation studies and additional experimental analyses are provided in the Appendix E

Table 1 shows that *TALAS* consistently achieves the best or second-best performance across most datasets and attains the highest **Domain Average** score, indicating superior generalization on both in-domain and out-of-domain benchmarks. In particular, *TALAS* yields notable improvements on challenging out-of-domain datasets such as *Tweet* and *STS12*, reflecting stronger robustness to domain shift. Compared with token-level distillation methods and sentence-level baselines, our approach delivers consistent gains without sacrificing efficiency, highlighting its ability to better preserve the geometric structure of the teacher embedding space while maintaining compact student representations.

$\mathcal{L}_{\text{SimCSE}}$	$\mathcal{L}_{\text{TAMD}}$	$\mathcal{L}_{\text{LASD}}$	ASAM	**Avg-In**	**Avg-Out**	**Avg-All**
Qwen3-Embedding 0.6B → MiniLMv2 H384 22M						
✓				44.24	62.41	56.35
✓	✓			65.50	75.24	71.99
✓	✓	✓		67.03	75.77	72.85
	✓		✓	67.65	77.10	73.95
	✓	✓	✓	67.27	77.78	74.28
✓	✓		✓	68.04	76.93	73.97
✓	✓	✓	✓	67.94	78.21	74.79
BGE-M3 0.6B → MiniLMv2 H768 66M						
✓				40.91	58.45	52.60
✓	✓			70.08	79.21	76.17
✓	✓	✓		69.37	79.63	76.21
	✓		✓	69.46	79.12	75.90
	✓	✓	✓	68.67	79.06	75.59
✓	✓		✓	69.83	79.27	76.13
✓	✓	✓	✓	69.53	80.09	76.57
Qwen3-Embedding 4B → Bert-base 109M						
✓				65.18	74.61	71.47
✓	✓			69.58	77.53	74.88
✓	✓	✓		71.28	77.99	75.75
	✓		✓	73.05	80.29	77.88
	✓	✓	✓	72.32	80.97	78.09
✓	✓		✓	73.32	80.44	78.06
✓	✓	✓	✓	73.61	81.44	78.83

Table 2: Ablation study on different loss combinations. Results are reported in terms of in-domain, out-of-domain, and overall average performance.

5 Analysis

Impact of each component in TALAS. $\mathcal{L}_{\text{TAMD}}$ functions as the primary bridge for cross-tokenizer alignment. As evidenced in Table 2, configurations lacking $\mathcal{L}_{\text{TAMD}}$ suffer from a catastrophic performance drop, verifying its status as the fundamental backbone upon which other objectives operate. The integration of ASAM further elevates performance, yet its impact varies across configurations. While ASAM yields marginal gains for the *BGE-M3 → MiniLMv2 H768* pair, it delivers substantial improvements for the *Qwen3-Embedding 4B → BERT-base* and *Qwen3-Embedding 0.6B → MiniLMv2 H384* settings. This disparity highlights ASAM's critical role in bridging the capacity gap: when the student is extremely compact (22M) or the teacher is vastly superior (4B), the optimization landscape becomes perilous, and ASAM's ability to seek flat minima proves essential for effective knowledge transfer. Finally, the auxiliary objectives exhibit complementary roles: $\mathcal{L}_{\text{SimCSE}}$ enhances local clustering for in-domain tasks (Avg-In) but shows signs of bias toward the training dis-

tribution. Conversely, replacing it with $\mathcal{L}_{\text{LASD}}$ consistently improves out-of-domain generalization (Avg-Out), suggesting that preserving the geometric relationships between layers is more effective for robust transfer learning.

Computational Efficiency. Table 3 analyzes the resource consumption for the Qwen3-Embedding 4B $\rightarrow$ BERT-base distillation setting, measured on an NVIDIA T4 GPU with a batch size of 32. Token-level baselines (e.g., EMO, DSKD) incur prohibitive overheads with peak memory exceeding 12 GB due to the necessity of parallel teacher inference and the impracticality of caching dense states. Among lightweight alternatives, *Jasper* shows low latency but requires a cumbersome two-stage pipeline, while *DistillCSE*, although faster, still consumes more memory and lacks structural guidance from intermediate representations. In contrast, our framework achieves the optimal synergy of efficiency and performance. TALAS (w/o ASAM) records the *second fastest* distillation speed of 195 ms/step—slightly behind DistillCSE (182 ms/step) and outperforming Jasper (204 ms/step)—while still delivering highly competitive results with a minimal memory footprint (3.98 GB). Even with ASAM enabled, TALAS remains significantly more resource-friendly than token-level methods, making it the most viable solution for high-performance distillation on consumer-grade hardware.

Method	Time (ms/step)	Mean Mem (MB)	Peak Mem (MB)	Std Mem (MB)
SimCSE-unsup	151	2662	3690	235
DistillCSE	182	3373	3674	124
Jasper and Stella	204	2932	3884	71
EMO	1223	11203	13831	492
DSKD	1002	10991	12393	336
CDM	1078	10814	12156	314
TALAS (w/o ASAM)	195	2864	3982	261
TALAS (with ASAM)	375	3317	4499	277

Table 3: Runtime and GPU memory while training comparison for Qwen3-Embedding-4B $\rightarrow$ BERT-base.

Impact of Sharpness-Aware Minimization Variants. Figure 2 presents the comparative results of SAM, ASAM, and DISAM on the Qwen3-Embedding 0.6B $\rightarrow$ MiniLMv2 H384 distillation task (refer to Appendix F for detailed breakdowns on individual datasets). As illustrated, ASAM consistently achieves the superior performance, recording the highest overall average score of **74.79** compared to DISAM (74.37) and SAM (73.90). Specifically, ASAM demonstrates remarkable robustness in Out-of-Domain generalization (78.21) while maintaining the best In-Domain stability (67.94). Consequently, we identify ASAM as the optimal optimizer for TALAS to ensure a robust balance between generalization capabilities and representation fidelity.

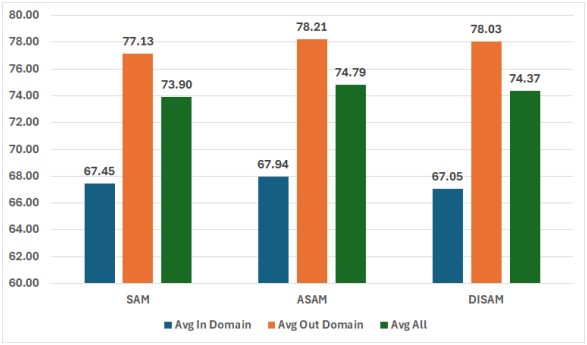

Figure 2: Comparison of SAM variants (SAM, DISAM, and ASAM) for Qwen3-Embedding 0.6B $\rightarrow$ MiniLMv2 H384.

Empirical Sharpness Analysis. To further substantiate the role of sharpness in our setting, we directly measure curvature during training by estimating the largest eigenvalue of the Hessian, denoted as $\lambda_{\max}$. This quantity serves as a widely adopted proxy for loss landscape sharpness, where larger values indicate sharper minima and smaller values correspond to flatter regions.

We track $\lambda_{\max}$ of the full training objective across epochs for both ASAM and AdamW under identical training conditions. The results are summarized in Table 4.

Epoch	ASAM	AdamW
1	14.89	30.24
3	12.94	21.48
5	11.59	19.66
10	10.85	18.66
30	9.93	17.41

Table 4: Largest Hessian eigenvalue ($\lambda_{\max}$) across training epochs. Lower values indicate flatter minima.

As shown in Table 4, ASAM consistently produces significantly smaller $\lambda_{\max}$ values compared to AdamW at all measured epochs. Notably, the gap remains substantial throughout training, indicating that ASAM systematically steers optimization toward flatter regions of the loss landscape rather than merely converging to a similar solution with reduced variance.

This behavior aligns with the theoretical motivation of sharpness-aware methods, which explicitly penalize sensitivity to parameter perturbations. By minimizing the worst-case loss within a local neighborhood, ASAM effectively avoids sharp minima characterized by high curvature. In contrast, AdamW, which lacks such regularization, tends to converge to sharper regions with higher $\lambda_{\max}$.

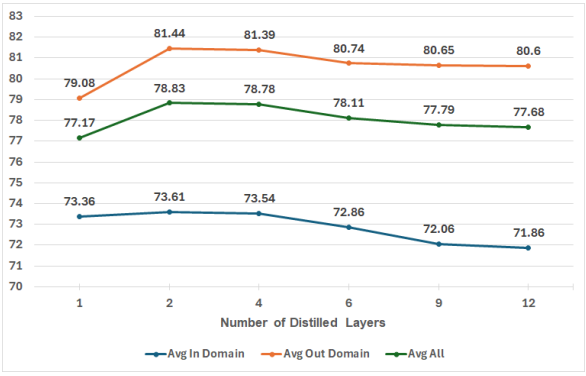

Figure 3: Effect of the number of distilled layers from teacher for Qwen3-Embedding 4B $\rightarrow$ BERT.

Effect of Distilled Layer Depth. Figure 3 illustrates the impact of the number of teacher-anchored layers (K) on performance. We observe that effectiveness peaks at $K = 2$ (Avg All **78.83**) rather than scaling linearly, with consistent degradation occurring at deeper alignments ($K > 4$). This trend suggests that given the substantial teacher-student capacity gap, enforcing constraints on too many internal layers induces over-regularization, restricting the student's intrinsic feature organization. Consequently, anchoring only the top two layers yields the optimal balance between high-level semantic transfer and the student's representational flexibility.

6 Conclusion

We presented **TALAS**, a novel distillation framework that effectively compresses LLM capabilities into compact encoders via a dual-strategy approach: aligning upper layers with the teacher ($\mathcal{L}_{\text{TAMD}}$) and enforcing internal geometric consistency ($\mathcal{L}_{\text{LASD}}$). Furthermore, we demonstrated that integrating ASAM is critical for navigating the student's loss landscape, enabling convergence to flat minima that generalize well across domains. Experimental results verify that TALAS outperforms existing baselines in both in-domain and out-of-domain scenarios. Future work will focus on scaling the training data and extending our evaluation to the MTEB to assess the model's versatility.

7 Limitations

Despite the effectiveness of TALAS, we acknowledge certain limitations in our study. First, a constrained computational budget restricted our experimental scope, precluding a comprehensive investigation into the framework's scalability with large models or large-scale training corpora. Second, our current evaluation focuses primarily on English-centric benchmarks; thus, the framework's adaptability to multilingual settings or low-resource languages remains to be fully explored. Finally, while our method significantly reduces memory footprint, the integration of ASAM necessitates two forward-backward passes per update step, effectively doubling the training duration. This trade-off between convergence stability and training speed may present a constraint for scenarios requiring rapid iteration cycles or when scaling to very large student models where temporal efficiency is paramount.

Acknowledgments

Trung Le was supported by the Air Force Office of Scientific Research under award number FA9550-23-S-0001. Thien Huu Nguyen was supported by 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.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393. Association for Computational Linguistics.

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.

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. 2025. Sparse logit sampling: Accelerating knowledge distillation in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18085–18108, Vienna, Austria. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1644–1650. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for Conversational AI*, pages 38–45. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14. Association for Computational Linguistics.

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.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Enhancing cross-tokenizer knowledge distillation with contextual dynamical mapping.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations.

Bao-Ngoc Dao, Minh Le, Quang Nguyen, Luyen Ngo Dinh, Nam Le Hai, and Linh Ngo Van. 2026. Wave++: Capturing within-task variance for continual relation extraction with adaptive prompting. *Neurocomputing*, page 132915.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

William B. Dolan and Chris Brockett. 2004. The microsoft research paraphrase corpus. Technical report, Microsoft Research.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. Hong Kong, China. Association for Computational Linguistics.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Guoqiang Gong, Jiaxing Wang, Jin Xu, Deping Xiang, Zicheng Zhang, Leqi Shen, Yifeng Zhang, JunhuaShu JunhuaShu, ZhaolongXing ZhaolongXing, Zhen Chen, et al. 2025. Beyond logits: Aligning feature dynamics for effective knowledge distillation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23067–23077.

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

Nam Le Hai, Linh Ngo Van, and Sang Dinh. 2026. Mozila: Continual event detection through the lens of multi-objective optimization and language model head preservation. *Computational Linguistics*, pages 1–44.

Cheng Han, Qifan Wang, Sohail A. Dianat, Majid Rabbani, Raghuveer M. Rao, Yi Fang, Qiang Guan, Lifu Huang, and Dongfang Liu. 2024. Amd: Automatic multi-step distillation of large-scale vision models.

Nguyen Manh Hieu, Vu Lam Anh, Hung Pham Van, Nam Le Hai, Diep Thi-Ngoc Nguyen, Linh Ngo Van, and Thien Huu Nguyen. 2025. Magix: A multi-granular adaptive graph intelligence framework for enhancing cross-lingual rag. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5202–5219.

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

Kun Huang, Xin Guo, and Meng Wang. 2023. Towards efficient pre-trained language model via feature correlation distillation. In *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. 2024. Distillm: Towards streamlined distillation for large language models.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. 2021. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks.

Anh Duc Le, Nam Le Hai, Thanh Xuan Nguyen, Linh Ngo Van, Nguyen Thi Ngoc Diep, Sang Dinh, and Thien Huu Nguyen. 2025a. 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.

Thanh-Thien Le, Viet Dao, Linh Nguyen, Thi-Nhung Nguyen, Linh Ngo, and Thien Nguyen. 2024. Sharpseq: Empowering continual event detection through sharpness-aware sequential-task learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3632–3644.

Tue Le, Hoang Tran Vuong, Quyen Tran, Linh Ngo Van, Mehrtash Harandi, and Trung Le. 2025b. Token-level self-play with importance-aware guidance for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. Online. Association for Computational Linguistics.

Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2025. Ese: Espresso sentence embeddings. In *ICLR*.

Zhenghao Lin, Yeyun Gong, Xiao Liu, Hang Zhang, Chen Lin, Anlei Dong, Jian Jiao, Jingwen Lu, Daxin Jiang, Rangan Majumder, and Nan Duan. 2023. Prod: Progressive distillation for dense retrieval.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, and Hassan Ghasemzadeh. 2019a. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. *CoRR*, abs/1902.03393.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2019b. Improved knowledge distillation via teacher assistant.

Yongyu Mu, Yuzhang Wu, Yuchun Fan, Chenglong Wang, Hengyu Li, Jiali Zeng, Qiaozhi He, Murun Yang, Fandong Meng, Jie Zhou, Tong Xiao, and Jingbo Zhu. 2024. Cross-layer attention sharing for pre-trained large language models. *arXiv preprint*, arXiv:2408.01890. ArXiv:2408.01890 [cs.CL].

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Truong Nguyen, Phi Van Dat, Ngan Nguyen, Linh Ngo Van, Trung Le, and Thanh Hong Nguyen. 2026. CTPD: cross tokenizer preference distillation. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI*, pages 37783–37790. AAAI Press.

Flavio Di Palo, Prateek Singhi, and Bilal H Fadlallah. 2024. Performance-guided LLM knowledge distillation for efficient text classification at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3687, Miami, Florida, USA. Association for Computational Linguistics.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation.

Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. 2020. Alp-kd: Attention-based layer projection for knowledge distillation.

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.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1267–1273. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. 2024. A comprehensive survey of sentence representations: From the BERT epoch to the CHATGPT era and beyond. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1738–1751, St. Julian's, Malta. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics.

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

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697. Association for Computational Linguistics.

Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. 2021. Densely guided knowledge distillation using multiple teacher assistants.

Wonchul Son, Jaemin Na, and Wonjun Hwang. 2020. Densely guided knowledge distillation using multiple teacher assistants. *CoRR*, abs/2009.08825.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. 2021. Does knowledge distillation really work? *Advances in neural information processing systems*, 34:6906–6919.

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

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices.

Quyen Tran, Nguyen Xuan Thanh, Nguyen Hoang Anh, Nam Le Hai, Trung Le, Linh Van Ngo, and Thien Huu Nguyen. 2024. Preserving generalization of language models in few-shot continual relation extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13771–13784.

Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. 2025. EMO: Embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7605–7617. Association for Computational Linguistics.

Duc Trung Vu, Pham Khanh Chi, Dat Phi Van, Linh Ngo Van, Dinh Viet Sang, and Trung Le. 2026a. Dwa-kd: Dual-space weighting and time-warped alignment for cross-tokenizer knowledge distillation. In *Findings of the Association for Computational Linguistics: EACL*, pages 3513–3527.

Hai An Vu, Minh-Phuc Truong, Tu Vu, and Linh Ngo. 2026b. Mol: Mixture of layers in cross-tokenizer embedding model distillation. *Knowledge-Based Systems*, 343:116001.

Hoang Tran Vuong, Tue Le, Quyen Tran, Linh Ngo Van, and Trung Le. 2026. MCW-KD: multi-cost wasserstein knowledge distillation for large language models. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI*, pages 33332–33340. AAAI Press.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR.

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.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. Distilled contrastive learning for sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

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

Hang Zhang, Seyyed Hasan Mozafari, James J. Clark, Brett H. Meyer, and Warren J. Gross. 2024a. Intermediate layer distillation with the reused teacher classifier: A study on the importance of the classifier of attention-based models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7203–7212, Miami, Florida, USA. Association for Computational Linguistics.

Ruipeng Zhang, Ziqing Fan, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. 2024b. Domain-inspired sharpness-aware minimization under domain shifts. *arXiv preprint arXiv:2405.18861*.

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models.

Tianyang Zhao, Kunwar Yashraj Singh, Srikar Appalaraju, Peng Tang, Ying Nian Wu, and Li Erran Li. 2025a. On the analysis and distillation of emergent outlier properties in pre-trained language models. 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 8475–8507, Albuquerque, New Mexico. Association for Computational Linguistics.

Tianyang Zhao, Kunwar Yashraj Singh, Srikar Appalaraju, Peng Tang, Ying Nian Wu, and Li Erran Li. 2025b. On the analysis and distillation of emergent outlier properties in pre-trained language models. 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 8475–8507.

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. 2025c. A survey of large language models.

Yuhang Zhou and Wei Ai. 2024. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 265–282, Bangkok, Thailand. Association for Computational Linguistics.

Honglei Zhuang, Zhen Qin, Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2021. Ensemble distillation for BERT-based ranking models. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21)*, New York, NY, USA. ACM.

A Extended Related Work

Knowledge Distillation (KD) aims to transfer knowledge from a high-capacity teacher model to a compact student model, enabling efficient deployment while preserving performance. Since its introduction, KD has been extensively studied across a wide range of NLP tasks, including classification, text generation, and representation learning. Despite its success, effective distillation remains challenging when compressing large-scale models into small students, particularly under limited computational and memory budgets.

Output Distillation Output distillation is the most widely adopted form of knowledge distillation, where the student is trained to match the teacher's output distributions or representations. Early work focuses on distilling softened logits or output embeddings, encouraging the student to imitate the teacher's predictive behavior at the output level (Hinton et al., 2015; Sanh et al., 2020; Jiao et al., 2020; Vu et al., 2026a). In representation learning, output-level distillation is especially attractive due to its simplicity and efficiency: it does not require access to the teacher's internal states and naturally supports offline or black-box distillation scenarios (Palo et al., 2024; Agarwal et al., 2024; Gu et al., 2024). However, relying solely on output supervision can be insufficient when the capacity gap between the teacher and student is large, often leading to suboptimal generalization or overfitting to teacher-specific artifacts (Cho and Hariharan, 2019; Ko et al., 2024). These limitations motivate more structured forms of knowledge transfer beyond point-wise imitation.

Relational Distillation To address the shortcomings of output-level supervision, relational distillation methods aim to transfer structural information among samples rather than matching individual representations. Relational Knowledge Distillation (RKD) preserves pairwise distances or angular relationships in the embedding space, demonstrating improved robustness and generalization, particularly in metric learning settings (Park et al., 2019). In NLP, relational signals have been leveraged to compress large Transformer models by distilling attention patterns or hidden-state correlations, showing that preserving semantic geometry can be more effective than strict feature-level alignment (Jiao et al., 2020; Sun et al., 2020; Huang et al., 2023; Truong et al., 2025; Vuong et al., 2026). While

effective, these approaches often incur substantial memory and computational overhead due to pairwise comparisons or intermediate-state supervision, which limits scalability - especially when teacher inference must be performed online during training.

Teacher Assistant Distillation Another line of work addresses the difficulty of distilling across large capacity gaps by introducing intermediate models between the teacher and the final student. Teacher Assistant (TA) distillation progressively transfers knowledge through one or more assistant models with intermediate capacities, enabling smoother knowledge transfer (Mirzadeh et al., 2019a). This strategy has been shown to consistently outperform direct teacher-to-student distillation, particularly when the teacher is significantly larger than the student (Mirzadeh et al., 2019a; Son et al., 2020). Recent studies further explore staged or curriculum-based distillation strategies in large-scale or embedding-focused scenarios (Lin et al., 2023; Han et al., 2024). However, TA-based methods substantially increase training complexity and computational cost due to the need to train and maintain additional intermediate models, which limits their practicality in resource-constrained environments.

Knowledge Distillation for Embedding Models Text embedding models serve as a foundational component across a wide range of NLP tasks, including semantic retrieval and retrieval-augmented generation (Hieu et al., 2025; Nguyen et al., 2025), event detection (Hai et al., 2026; Le et al., 2024), and relation extraction (Dao et al., 2026; Anh et al., 2025; Tran et al., 2024; Le et al., 2025a; Pham et al., 2025). However, while knowledge distillation has recently been widely adopted to transfer knowledge across diverse settings - from reasoning capability transfer in large language models (Gu et al., 2024; Ko et al., 2024) to preference alignment (Nguyen et al., 2026; Le et al., 2025b) - its application to text embedding models, despite their central role in these downstream tasks, remains substantially underexplored.

However, while recently knowledge distillation has been widely adopted to transfer knowledge across a range of tasks - from reasoning capability transfer in large language models (Gu et al., 2024; Ko et al., 2024) to preference alignment (Nguyen et al., 2026; Le et al., 2025b) - yet its application to text embedding models remains substantially underexplored. Distilling embedding models poses challenges that differ substantially from those encountered in standard classification or text generation tasks. High-performing embedding teachers are often large language models with architectures, tokenizers, or vocabularies that differ from those of the student, making token-level alignment either computationally expensive or infeasible. As a result, relatively few studies explicitly focus on embedding-specific distillation. Representative early works include Zhuang et al. (2021), which directly mimics output logits over the vocabulary, and Jiao et al. (2020), which distills knowledge from both hidden representations and attention matrices. DistillCSE (Xu et al., 2023) takes an important step toward embedding-level distillation by transferring sentence representations through contrastive learning, thereby avoiding token-level supervision and enabling more efficient knowledge transfer. More recently, Jasper and Stella (Zhang et al., 2025a) propose a multi-stage distillation pipeline to compress state-of-the-art embedding models while maintaining strong retrieval performance. EMO (Truong et al., 2025) further explores embedding distillation from a relational perspective, aligning inter-sample structures via optimal transport. More recently, Vu et al. (2026b) proposes a Mixture-of-Layers approach for cross-tokenizer embedding distillation, adaptively routing distillation signals across multiple student layers to bridge the representational gap between heterogeneous teacher–student pairs.

Despite their effectiveness, many existing methods still rely on token-level supervision, intermediate-state matching, or repeated online teacher inference, which introduces significant memory and computational overhead during student training.

In contrast, our work focuses on *resource-efficient embedding distillation*. We distill knowledge using only cached sentence-level embeddings from the teacher, eliminating the need for teacher inference during student training. This design significantly reduces memory consumption and training latency while remaining effective in large-teacher–small-student settings, making it well suited for practical deployment scenarios.

B Sharpness-Aware Optimization

Sharpness-Aware Minimization (SAM) (Foret et al., 2020). Standard empirical risk minimiza-

tion tends to converge to sharp minima, which are sensitive to parameter perturbations and often generalize poorly. SAM addresses this by simultaneously minimizing the loss value and the loss landscape geometry (sharpness). Formally, it seeks parameters $\boldsymbol{\theta}$ that minimize the worst-case loss within a local Euclidean neighborhood of radius ρ:

$$\min_{\boldsymbol{\theta}} \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}). \tag{8}$$

To solve the inner maximization efficiently, SAM applies a first-order Taylor approximation, deriving the optimal perturbation $\hat{\boldsymbol{\epsilon}}$ as the gradient direction scaled by ρ:

$$\hat{\boldsymbol{\epsilon}} = \rho \frac{\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})}{\|\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})\|_2}. \tag{9}$$

The model then updates its weights based on the gradient computed at the perturbed position $\boldsymbol{\theta} + \hat{\boldsymbol{\epsilon}}$, effectively steering the optimization trajectory towards flat minima that improve generalization.

Domain-Inspired Sharpness-Aware Minimization (DISAM) (Zhang et al., 2024b). Standard SAM operates under the crucial assumption that training samples are independent and identically distributed (i.i.d.). However, in multi-domain scenarios, this assumption often fails as data varies significantly in complexity and distribution across domains. Consequently, SAM suffers from *inconsistent convergence*, where it generates perturbations biased towards domains with larger gradients (typically those that are under-converged), potentially destabilizing the optimization for other domains.

To rectify this, DISAM integrates domain-level statistics into the sharpness estimation by imposing a variational constraint. The objective is to maximize the perturbed loss while simultaneously minimizing the divergence between domain losses. Formally, the maximization step in DISAM is defined as:

$$\mathcal{L}_{\text{DISAM}}(\mathbf{w}) = \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \Bigg[\sum_{i=1}^{M} \alpha_i \mathcal{L}_i(\mathbf{w} + \boldsymbol{\epsilon}) - \lambda \operatorname{Var}\{\mathcal{L}_i(\mathbf{w} + \boldsymbol{\epsilon})\}_{i=1}^{M} \Bigg], \tag{10}$$

where λ is a hyperparameter balancing the sharpness and the consistency constraints. Explicitly, the inter-domain variance term is defined as

$$V(\mathbf{w} + \boldsymbol{\epsilon}) = \operatorname{Var}\{\mathcal{L}_i(\mathbf{w} + \boldsymbol{\epsilon})\}_{i=1}^{M}. \tag{11}$$

It is calculated as the mean squared difference between domain losses:

$$V(\mathbf{w} + \boldsymbol{\epsilon}) = \frac{1}{2M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \Big(\mathcal{L}_i(\mathbf{w} + \boldsymbol{\epsilon}) - \mathcal{L}_j(\mathbf{w} + \boldsymbol{\epsilon}) \Big)^2. \tag{12}$$

Intuitively, this variance term acts as an adaptive regularizer. If a specific domain's loss deviates significantly from the others (violating the balance), the variance gradient counteracts the standard SAM perturbation. This mechanism essentially suppresses noise for under-converged domains (to maintain stability) while amplifying it for well-converged domains (to escape sharp minima), thereby synchronizing convergence across diverse data distributions. Our work integrates adaptive sharpness-aware optimization directly into the embedding distillation pipeline. By optimizing the distillation objective under ASAM, the student is encouraged to capture the teacher's broad semantic structure rather than overfitting to specific embedding targets. This combination of embedding-level distillation and sharpness-aware optimization is particularly effective for large teacher–small student scenarios, leading to improved robustness and out-of-domain generalization.

C Offline Teacher Inference and Cached Distillation Targets

The teacher model is not invoked during student training. Instead, it is executed **only once** in a preprocessing stage to extract and cache sentence-level representations for all training samples.

Specifically, we pass the training data through the teacher model using a dedicated dataloader and store the resulting last CLS embeddings associated with each sample. This procedure ensures full consistency between the teacher and student inputs, while avoiding discrepancies caused by differences in preprocessing or data ordering.

During student training, the cached distillation targets are retrieved alongside the input data through a *lazy loading* mechanism. For each mini-batch, only the corresponding precomputed sentence embeddings are loaded from memory or

disk, eliminating the need to perform teacher inference within the training loop. This design completely removes the computational overhead of the teacher during training and substantially reduces both training latency and GPU memory consumption compared to conventional online distillation approaches.

Rationale.

Consider a BERT-based teacher model, where each input sample produces hidden states of shape $[256, 768]$. This corresponds to

$$256 \times 768 = 196{,}608 \text{ elements per sample.}$$

When stored in `float32` format, this requires approximately $786{,}432$ bytes (≈ 768 KB) per sample. For a dataset of $100{,}000$ samples, the total storage requirement exceeds 76 GB, which is impractical for most training environments. Approaches that rely on caching intermediate-layer hidden states further amplify this cost, rendering offline storage largely infeasible in practice.

In contrast, by caching only the *sentence embedding*, each sample requires a single vector of shape $[768]$. When stored in `float32` format, this amounts to $3{,}072$ bytes (≈ 3 KB) per sample. Consequently, for $100{,}000$ samples, the total storage footprint is reduced to approximately 300 MB, making offline distillation targets both practical and scalable.

D Dataset Creation

To construct the training data for contrastive sentence representation learning, we sample sentences from multiple task categories to ensure diversity across domains and objectives. Specifically, we select 5,000 sentences from classification datasets, 2,500 sentence pairs from semantic textual similarity (STS) datasets, and 2,500 sentence pairs from pair classification datasets. All sentence pairs are then flattened into individual sentences, resulting in a total of 15,000 unique sentences. This unified sentence collection is used to form the training data for SimCSE-style contrastive learning, enabling the model to learn robust sentence embeddings from heterogeneous sources while maintaining a consistent training format.

E Experimental Details

Training Configuration We adopt a unified training protocol across the three teacher–student pairs considered in our experiments. The detailed training configurations for all methods are summarized in Table 5. All knowledge distillation (KD) experiments are conducted on the same dataset constructed as described in Section D, ensuring a fair comparison across methods. Evaluation is performed on multiple benchmark datasets spanning different domains and task types, using standard metrics including F1 score for classification, Average Precision (AP) for pair classification, and Spearman's rank correlation for semantic textual similarity.

Losses Our models were trained with $\lambda_1 = 0.001, \lambda_2 = 0.75 \; and \; \lambda_3 = 1$

Evaluation We evaluate the quality of the learned sentence embeddings across three categories of downstream tasks. (1) For **classification tasks**, we follow standard evaluation settings (Conneau and Kiela, 2018) by freezing the sentence embeddings and training a Logistic Regression classifier on top. (2) For **pair classification**, we compute the cosine similarity between sentence representations and determine the prediction based on an optimal threshold, reporting the Average Precision (AP). (3) Finally, for **Semantic Textual Similarity (STS)** tasks, we measure the alignment between the cosine similarity of the embeddings and human-annotated gold scores using Spearman correlation.

Baseline Implementation Details To ensure a rigorous evaluation, we adapt state-of-the-art generative distillation methods to the discriminative sentence embedding setting.

Contextual Dynamic Mapping (CDM). We apply CDM to align hidden representations $\mathbf{H} \in \mathbb{R}^{B \times L \times D}$. The original formulation relies on the teacher's predictive entropy to modulate the Dynamic Time Warping (DTW) alignment cost. However, since standard sentence encoders do not readily output token-level vocabulary distributions, we modify the cost function to utilize *token edit distance* (Levenshtein distance). Specifically, we compute a pairwise distance matrix between the student and teacher token sequences and employ DTW to identify the optimal monotonic alignment path that minimizes the cumulative cost. Based on this path, teacher hidden states are aggregated (via averaging) to match the student's sequence length, establishing a synchronized one-to-one correspondence for the final Mean Squared Error (MSE) loss.

Dual Space Knowledge Distillation (DSKD). We extend DSKD to enforce geometric consistency

Settings	DSKD	CDM	EMO	Stella & Jasper	DistillCSE	TALAS
Epoch	5	5	5	2 + 5 (two-stage)	5	5
LR	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Batch Size	32	32	32	32	32	32
LR Scheduler	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine

Table 5: Training configurations for different distillation and sentence embedding methods.

		Classification (F1)		Pair Classification (AP)		STS (Spearman)		Domain Avg
Model	**Training Data**	**Banking77**	**Tweet**	**MRPC**	**SciTail**	**SICK**	**STS12**	**Avg-Out**
Unsup-SimCSE	∼1M sentences	90.74	69.85	84.85	83.04	72.23	68.40	78.19
Sup-SimCSE	∼1M pairs (NLI)	*90.77*	74.10	86.76	**85.40**	**79.52**	*77.76*	**82.39**
SBERT-NLI (mean)	∼1M pairs (NLI)	88.84	74.00	86.97	76.67	69.71	70.15	77.72
SBERT-NLI+STS (mean)	∼1M pairs (NLI) + 8.6k (STS-B)	89.84	74.11	**89.28**	*83.76*	77.15	**79.38**	*82.25*
SBERT-NLI (CLS)	∼1M pairs (NLI)	89.11	**75.90**	*87.80*	77.63	72.21	69.41	78.68
TALAS	∼15K sentences	**91.43**	*74.30*	86.47	82.66	*78.38*	75.40	81.44

Table 6: Out-of-domain performance comparison with existing fine-tuned Bert-base model. Best results are shown in **bold**, and second-best results are shown in *italics*.

through a symmetric, bidirectional process. Unlike standard unidirectional approaches, DSKD projects representations into each model's respective latent space. We implement this at two granularities: (1) *Sentence-level:* Global representations (e.g., `[CLS]` tokens) are distilled via learnable linear projection heads to align the global semantic spaces. (2) *Token-level:* To address the misalignment caused by differing vocabularies and sequence lengths, we incorporate Cross-Layer Attention (CLA) (Mu et al., 2024). CLA computes an attention mechanism between the student's and teacher's intermediate layers, allowing the student to reconstruct fine-grained teacher features (and vice versa) without requiring rigid token-to-token matching.

F Additional Experiment Results

Comparison with Existing Pretrained-Model. Table 6 highlights that our method achieves strong out-of-domain generalization despite being trained on a substantially smaller dataset. While most competing baselines rely on large-scale supervision with approximately 1M sentence pairs or sentences, our approach is trained on only 15K sentences yet attains an Avg-Out score of 81.44, remaining competitive with state-of-the-art supervised models such as Sup-SimCSE (82.39) (Gao et al., 2022) and SBERT NLI+STS (82.25) (Reimers and Gurevych, 2019).

Across individual tasks, our model demonstrates consistent robustness under domain shift, achieving competitive results in both classification and semantic similarity benchmarks. Notably, our method performs strongly on STS tasks, indicating that effective semantic alignment can be learned even under limited-data conditions. This suggests that the proposed training strategy leverages teacher supervision efficiently, enabling high-quality sentence representations without dependence on large annotated corpora.

Overall, the table highlight the strong data efficiency and practical applicability of our method. Despite being trained on substantially fewer samples, our approach achieves competitive out-of-domain performance while incurring lower training time and memory costs compared to existing knowledge distillation methods. This makes our method a compelling alternative to fine-tuned BERT-base models that rely on large-scale supervised datasets.

Ablation on Teacher Distillation. In table 7, for Qwen3-Embedding 0.6B → MiniLMv2 H384, increasing the number of distilled layers from 1 to 2 improves both Avg-Out and overall performance, with the peak Avg-Out achieved at 2 layers (78.21). Further increasing the number of distilled layers leads to gradual performance degradation, particularly on STS and out-of-domain metrics, indicating that distilling too many upper layers from the teacher can overly constrain the student representations. A similar trend is observed for Qwen3-Embedding 4B → BERT-base. The best results are obtained with 2–4 distilled layers, where Avg-Out reaches 81.44 and 81.39, respectively. Beyond this range, deeper distillation (6–12 layers) consistently reduces out-of-domain performance, despite relatively stable in-domain scores. This suggests diminishing returns when extending teacher supervision

	Classification (F1)			Pair Classification (AP)			STS (Spearman)			Domain Avg		
Num Layers	**Banking77**	**Tweet**	**Emotion***	**MRPC**	**SciTail**	**WiC***	**SICK**	**STS12**	**STSB***	**Avg-In**	**Avg-Out**	**Avg.**
Qwen3-Embedding 0.6B → MiniLMv2 H384												
1	86.28	72.33	60.46	84.33	80.73	66.57	72.15	68.14	75.43	67.49	77.33	74.05
2	86.69	72.77	60.94	85.10	81.42	66.65	72.51	70.77	76.22	67.94	78.21	74.79
3	86.03	72.35	61.01	85.10	80.55	66.86	73.01	68.60	75.76	67.88	77.61	74.36
6	85.95	71.72	60.88	85.33	79.47	67.12	71.85	67.33	74.13	67.38	76.94	73.75
Qwen3-Embedding 4B → Bert-base												
1	91.10	72.88	70.73	85.36	82.30	71.27	77.40	65.41	78.09	73.36	79.08	77.17
2	91.43	74.30	70.71	86.47	82.66	69.24	78.38	75.40	80.88	73.61	81.44	78.83
4	91.35	73.99	70.94	86.55	82.80	69.11	77.80	75.86	80.58	73.54	81.39	78.78
6	91.24	74.09	69.59	86.45	82.48	69.26	77.25	72.91	79.72	72.86	80.74	78.11
9	90.05	74.84	69.10	86.73	82.35	67.99	76.76	73.17	79.09	72.06	80.65	77.79
12	90.28	74.68	68.47	86.74	82.99	68.15	76.45	72.43	78.96	71.86	80.60	77.68

Table 7: Ablation on $\mathcal{L}_{\text{TAMD}}$

	Classification (F1)			Pair Classification (AP)			STS (Spearman)			Domain Avg		
Num Layers	**Banking77**	**Tweet**	**Emotion***	**MRPC**	**SciTail**	**WiC***	**SICK**	**STS12**	**STSB***	**Avg-In**	**Avg-Out**	**Avg.**
Qwen3-Embedding 0.6B → MiniLMv2 H384												
2	87.93	72.30	61.49	83.39	80.27	68.76	72.39	64.90	73.35	67.87	76.86	73.86
3	87.34	72.57	60.66	85.23	81.30	67.43	71.77	69.61	75.66	67.92	77.97	74.62
6	86.69	72.77	60.94	85.10	81.42	66.65	72.51	70.77	76.22	67.94	78.21	74.79
Qwen3-Embedding 4B → Bert-base												
2	91.36	73.56	68.21	84.97	81.29	70.06	77.33	70.60	78.28	72.18	79.85	77.30
4	91.59	74.50	70.96	86.30	81.42	69.63	77.76	74.04	79.40	73.33	80.94	78.40
6	91.66	74.56	70.92	86.26	82.41	70.01	78.08	74.81	80.35	73.76	81.30	78.78
9	90.63	74.69	69.41	86.52	82.40	68.48	78.27	74.89	80.79	72.89	81.23	78.45
12	91.43	74.30	70.71	86.47	82.66	69.24	78.38	75.40	80.88	73.61	81.44	78.83

Table 8: Ablation on $\mathcal{L}_{\text{LASD}}$

too deep into the student network. These results indicate that $\mathcal{L}_{\text{TAMD}}$ is most effective when applied to a small number of top layers, where it provides strong semantic guidance from the teacher while preserving sufficient flexibility in lower layers. Excessive top-layer distillation, however, appears to limit the student's ability to generalize, leading to reduced out-of-domain performance.

Ablation on Self Distillation. Table 8 shows that increasing the number of self-distillation layers consistently improves performance across both teacher–student settings. For Qwen3-Embedding 0.6B → MiniLMv2 H384, out-of-domain performance rises steadily as more layers are included, reaching the best Avg-Out score at 6 layers. This indicates that deeper self-distillation strengthens representation robustness for smaller students. The effect is more pronounced for Qwen3-Embedding 4B → BERT-base, where performance improves monotonically from 2 to 12 layers, achieving the highest Avg-Out (81.44) and overall average (78.83) at the deepest configuration. Unlike $\mathcal{L}_{\text{TAMD}}$, no degradation is observed as depth increases, suggesting that larger students benefit from deeper self-distillation. Overall, these results show that $\mathcal{L}_{\text{LASD}}$ scales well with depth and consistently enhances generalization.

Performance of SAM Variants. Table 9 presents a comprehensive performance breakdown of SAM, DISAM, and ASAM across diverse benchmarks, including single-sentence classification (e.g., Banking77, Emotion), pair classification (e.g., MRPC, WiC), and semantic textual similarity (e.g., STS-B, SICK). Validated across two distinct teacher-student setups (Qwen3-Embedding 0.6B → MiniLMv2 H384 and Qwen3-Embedding 4B → Bert-base), the results demonstrate that ASAM consistently achieves the highest aggregated performance, recording average scores of **74.79** and **78.83**, respectively. While DISAM remains highly competitive—marginally outperforming ASAM in specific tasks such as Tweet sentiment analysis—ASAM exhibits superior overall robustness and stability across the full spectrum of evaluation

	Classification (F1)			Pair Classification (AP)			STS (Spearman)			Domain Avg		
Method	**Banking77**	**Tweet**	**Emotion***	**MRPC**	**SciTail**	**WiC***	**SICK**	**STS12**	**STSB***	**Avg-In**	**Avg-Out**	**Avg.**
Qwen3-Embedding 0.6B → MiniLMv2 H384												
SAM	87.57	72.77	60.64	83.54	80.49	68.04	72.30	66.08	73.67	67.45	77.13	73.90
ASAM	86.69	72.77	60.94	85.10	81.42	66.65	72.51	70.77	76.22	67.94	78.21	74.79
DISAM	86.84	73.11	60.52	84.99	81.11	65.20	72.30	69.80	75.43	67.05	78.03	74.37
Qwen3-Embedding 4B → Bert-base												
SAM	91.16	73.97	68.39	85.10	80.84	70.38	77.48	78.77	75.65	71.47	81.22	77.97
ASAM	91.43	74.30	70.71	86.47	82.66	69.24	78.38	75.40	80.88	73.61	81.44	78.83
DISAM	91.27	74.60	71.14	86.51	81.92	68.53	78.57	76.00	80.53	73.40	81.48	78.79

Table 9: Comparison of SAM variants: SAM, DISAM, and ASAM

metrics, justifying its adoption as the primary optimizer for the TALAS framework.