TRANSPORTING LAYERWISE STRUCTURE FOR TEXT EMBEDDING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large text embedding models achieve strong performance but incur high deployment costs, motivating knowledge distillation (KD) into lightweight models. Existing text embedding distillation methods primarily focus on matching final embeddings, leaving the student's intermediate computation underconstrained. We first observe that strong text embedding teachers share a consistent *layerwise trajectory*: pairwise distances expand into a pronounced hump in the middle layers before contracting, while the positive–negative margin widens rapidly in the final third of the network. However, output-only KD does not transfer this intermediate trajectory, and existing layerwise objectives also fail to recover it when applied to text embeddings. These observations expose three challenges: (i) failure to preserve relations among input texts, (ii) inability to align teacher and student networks with different numbers of layers, and (iii) neglect of the late-layer separation trend between positive and negative samples. To address these challenges, we propose **Layerwise Structure Transport (LST)**, which (i) encodes relations among input texts in a common space, (ii) learns a soft correspondence to align teacher and student networks of different depths, and (iii) transfers ranking preferences with greater weight at depths where positive–negative separation emerges. Experiments show that LST reliably transfers the teacher's layerwise trajectory and consistently improves student performance under all tested Output KD objectives. Our code can be found in the Reproducibility Statement.

1 INTRODUCTION

Large-scale text embedding models underpin semantic search, retrieval-augmented generation, reranking, clustering, and other language services (Morris & Rush, 2025; Weller et al., 2025; Li et al., 2025b; Su et al., 2023; Li et al., 2025a). Billion-parameter encoders provide strong general-purpose representations, but online encoding latency, accelerator memory, and throughput costs limit deployment at scale (Lee et al., 2025; Muennighoff et al., 2025). Knowledge distillation therefore provides a way to transfer the capabilities of a strong teacher to a lightweight student (Gu et al., 2024; Agarwal et al., 2024; Gu et al., 2025; Shing et al., 2025; Ildiz et al., 2025; Ko et al., 2024; 2025).

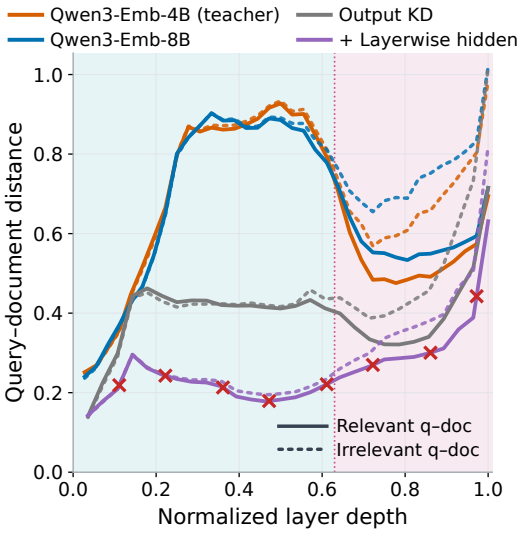


Figure 1: Layerwise query–document distances for different models. Red crosses mark the eight teacher layers omitted by the fixed layer map.

Intermediate-layer supervision is a standard distillation strategy, ranging from feature hints (Romero et al., 2015) to recent relational objectives in visual representation learning and language modeling (Shi et al., 2024; Dasgupta & Cohn, 2025). Yet text embedding experiments find that adding layerwise hidden-state or attention losses could hurt retrieval performance (Vujanic & Rückstieß, 2026). Existing text embedding

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distillation methods therefore supervise only the final embedding (Akram et al., 2026; Vujanic & Rückstieß, 2026). However, we argue that the teacher's ability to organize sample relations across layers remains valuable for distillation. As shown in Figure 1, both embedding teachers' query–document distances rise into a pronounced hump in the middle layers (0.25-0.63 layer depth), whereas the Output KD student does not reproduce this expansion. This pattern also extends to embedding models from other families. Moreover, existing hidden-state matching fails to preserve this relational trajectory and performs worse than Output-only KD (Vujanic & Rückstieß, 2026). This motivates our central question: *how can we transfer a teacher's ability to organize sample relations across layers to a lightweight student?*

To overcome these limitations, we identify three challenges. (i) **Preserving sample relations.** Layerwise hidden-state objectives use learned projectors to reconcile different token dimensions, but coordinate matching after projection does not guarantee that the student preserves relations among arbitrary input text samples in its native representation space (Csiszárik et al., 2021; Huang et al., 2022). (ii) **Aligning complete layer sequences.** Teachers and students often have different numbers of layers. Fixed, random, or manually selected pairs may match different computational stages or omit parts of the teacher trajectory (Passban et al., 2021; Chen et al., 2021). (iii) **Transferring late-emerging relevance.** Figure 1 also shows that positive and negative query–document distances separate most clearly in later layers. Supervision should capture where this separation emerges and transfer the resulting ranking preferences (van Aken et al., 2019; Liang et al., 2023; Yang et al., 2024).

Layerwise Structure Transport (LST) uses three modules to address these challenges. (i) **Relational Geometry Encoding** expresses each layer in a dimension-independent sample-relation space, enabling direct supervision of the student's native geometry without intermediate projectors. (ii) **Soft Layer Alignment** replaces predefined layer pairs with a learned soft correspondence plan over the complete teacher and student layer sequences, allowing each student layer to aggregate supervision from multiple teacher stages. (iii) **Emergence-Weighted Ranking** derives depth-dependent weights from the teacher's retrieval transition and uses them to transfer its layerwise positive–negative preferences. Our contributions are summarized as follows:

- We identify a shared layerwise trajectory with distinct stages in the tested high-capacity embedding teachers, and find that Output-only KD does not transfer it to the student.
- We formulate LST to preserve native sample relations, align complete teacher and student layer sequences, and transfer the teacher's late-emerging ranking preferences.
- To our knowledge, we provide the first evidence that a teacher's layerwise trajectory can be reliably transferred in text embedding distillation (Reimers & Gurevych, 2020; Xie et al., 2023; Mao & Nakagawa, 2023; Xu et al., 2023; Truong et al., 2025; Akram et al., 2026; Vujanic & Rückstieß, 2026). As a training-only plugin, LST consistently improves both tested Output KD objectives without adding inference-time overhead.

2 Layerwise Structure Transport

Let f_{T} and f_{S} be teacher and student text embedding models with L_{T} and L_{S} layers and token dimensions d_{T} and d_{S}; in general, $L_{\mathrm{T}} \neq L_{\mathrm{S}}$ and $d_{\mathrm{T}} \neq d_{\mathrm{S}}$. For a batch of B texts, $\boldsymbol{Z}_m^{\mathrm{T}}$ and $\boldsymbol{Z}_l^{\mathrm{S}}$ collect their last-token representations at teacher layer m and student layer l. LST augments an output-level distillation objective with structural and ranking auxiliary terms, weighted by λ_{str} and λ_{rank}, respectively. The structural term transfers layerwise sample relations through a learned soft correspondence, while the ranking term transfers the teacher's ranking preferences as positive–negative separation emerges across layers. Both terms access the teacher only during training step and leave the student's inference architecture unchanged.

Figure 2 illustrates the overall pipeline. Each layer produces a within-batch relation operator. For a labeled triplet of positive document, query, and negative document, the same layer states also produce a positive–negative rank probability. The complete operator sequences determine cross-model and within-model soft layer plans. The ranking branch reuses the teacher-to-student plan $\boldsymbol{A}^{\mathrm{ST}}$ and student self-plan $\boldsymbol{A}^{\mathrm{SS}}$ to construct aligned teacher and student probabilities at each student depth. It then applies a Bernoulli KL weighted by the transported teacher-emergence signal w_l^{S}.

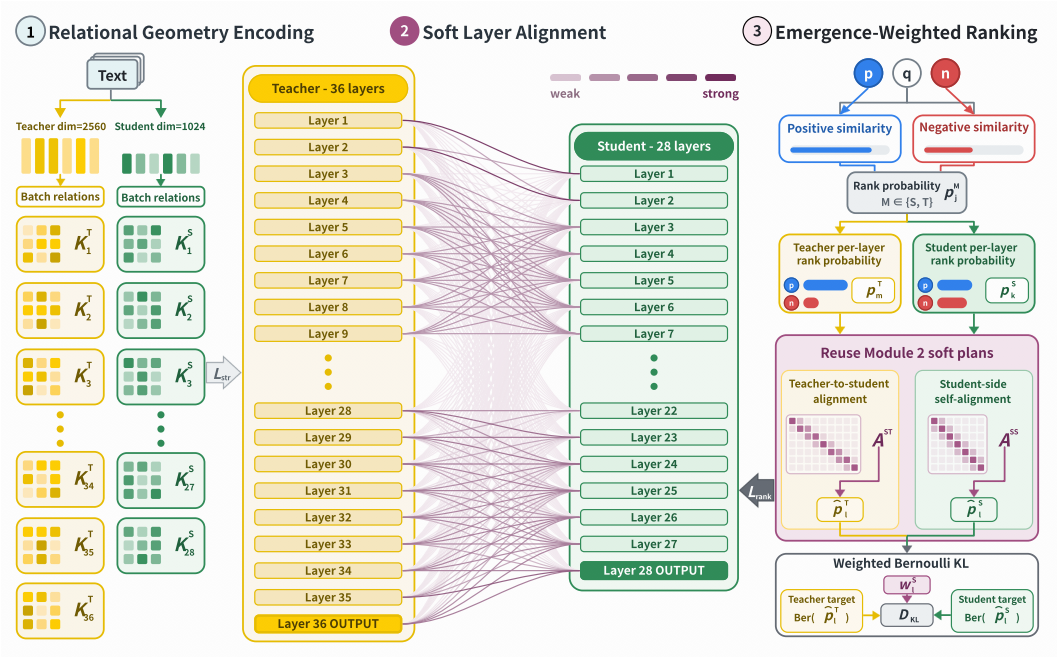


Figure 2: Overview of LST. Gold and green denote teacher and student quantities, while purple edges represent the soft alignment plans. (i) Relational Geometry converts representations with different token dimensions into batch relation operators. (ii) Soft Layer Alignment learns soft teacher–student and self-alignment plans. (iii) Emergence-Weighted Ranking forms per-layer rank probabilities, reuses the Soft Layer Alignment plans and applies an emergence-weighted KL divergence.

2.1 Relational Geometry Encoding

To preserve sample relations in the student's native representation space, we use centered Gram matrices to compare relations among the same texts independently of token dimension (Kornblith et al., 2019; Zhou et al., 2024; Dasgupta & Cohn, 2025; Huh et al., 2024). For model $M \in \{\mathrm{S}, \mathrm{T}\}$ and layer j, let $\boldsymbol{U}_j^M = \mathrm{RowNorm}(\boldsymbol{Z}_j^M)$ and $\boldsymbol{H} = \boldsymbol{I}_B - B^{-1}\mathbf{1}\mathbf{1}^\top$. We define

$$t_j^M = \mathrm{tr}\big(\boldsymbol{H}\boldsymbol{U}_j^M(\boldsymbol{U}_j^M)^\top \boldsymbol{H}\big), \qquad \boldsymbol{K}_j^M = \frac{\boldsymbol{H}\boldsymbol{U}_j^M(\boldsymbol{U}_j^M)^\top \boldsymbol{H}}{t_j^M}. \tag{1}$$

Each $\boldsymbol{K}_j^M$ is a $B \times B$ matrix, regardless of d_M, and is invariant to orthogonal changes of feature basis. Trace normalization removes total centered relation energy, so the comparison targets relative sample geometry without intermediate projectors.

2.2 Depth-Regularized Soft Layer Alignment

To overcome the inability to align teacher and student models with unequal numbers of layers, we use entropic optimal transport (Cuturi, 2013) with a structural cost and a soft depth-locality prior ($\lambda_d > 0$). Let normalized depths be $\delta_l^{\mathrm{S}} = (l-1)/(L_{\mathrm{S}}-1)$ and $\delta_m^{\mathrm{T}} = (m-1)/(L_{\mathrm{T}}-1)$. The teacher's trajectory variation sets a common structural scale:

$$\overline{\boldsymbol{K}}^{\mathrm{T}} = \frac{1}{L_{\mathrm{T}}}\sum_m \boldsymbol{K}_m^{\mathrm{T}}, \qquad s_{\mathrm{T}} = \max\left\{\frac{1}{L_{\mathrm{T}}B^2}\sum_m \|\boldsymbol{K}_m^{\mathrm{T}} - \overline{\boldsymbol{K}}^{\mathrm{T}}\|_F^2,\ 10^{-8}\right\}. \tag{2}$$

The cross-model cost is

$$C_{lm}^{\mathrm{ST}} = \frac{\|\boldsymbol{K}_l^{\mathrm{S}} - \boldsymbol{K}_m^{\mathrm{T}}\|_F^2}{B^2 s_{\mathrm{T}}} + \lambda_d(\delta_l^{\mathrm{S}} - \delta_m^{\mathrm{T}})^2. \tag{3}$$

Let $\boldsymbol{a} \in \mathbb{R}^{L_{\mathrm{S}}}$ and $\boldsymbol{b} \in \mathbb{R}^{L_{\mathrm{T}}}$ be the plan's row and column marginals, specifying how much matching mass is assigned to each student and teacher layer. We take both to be uniform, $a_l = 1/L_{\mathrm{S}}$ and

$b_m = 1/L_{\mathrm{T}}$, so that no layer is weighted a priori. The plan is then

$$\boldsymbol{\pi}^{\mathrm{ST}} = \arg\min_{\boldsymbol{\pi}\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{\pi}, \boldsymbol{C}^{\mathrm{ST}}\rangle - \varepsilon\mathcal{H}(\boldsymbol{\pi}), \qquad \mathcal{U}(\boldsymbol{a},\boldsymbol{b}) = \{\boldsymbol{\pi} \geq 0 : \boldsymbol{\pi}\mathbf{1} = \boldsymbol{a}, \boldsymbol{\pi}^\top\mathbf{1} = \boldsymbol{b}\}. \quad (4)$$

Here $\mathcal{H}$ denotes entropy. We compute finite-iteration Sinkhorn plans $\widehat{\boldsymbol{\pi}}^{XY}$ for cross-model and within-model pairs with the same teacher scale and depth penalty, and row-normalize them as $A^{XY}_{lm} = \widehat{\pi}^{XY}_{lm} / \sum_{m'} \widehat{\pi}^{XY}_{lm'}$, so that each row is a distribution over target layers. Writing $\mathcal{T}_\varepsilon(X,Y) = \langle \widehat{\boldsymbol{\pi}}^{XY}, \boldsymbol{C}^{XY}\rangle$, the structural objective is

$$\begin{aligned} \Delta_{\mathrm{str}} &= \mathcal{T}_\varepsilon(\mathrm{S},\mathrm{T}) - \tfrac{1}{2}\mathcal{T}_\varepsilon(\mathrm{S},\mathrm{S}) - \tfrac{1}{2}\mathcal{T}_\varepsilon(\mathrm{T},\mathrm{T}), \\ \mathcal{L}_{\mathrm{str}} &= \rho(\Delta_{\mathrm{str}}), \qquad \rho(x) = \frac{\sqrt{\max\{x, 10^{-12}\}}}{1 + \sqrt{\max\{x, 10^{-12}\}}}. \end{aligned} \quad (5)$$

This self-debiased, bounded cost controls the auxiliary loss scale.

2.3 Emergence-Weighted Ranking

To capture the teacher's late-layer separation between positive and negative samples, we transfer the teacher's positive–negative preferences through the same correspondence. Relation matching preserves normalized sample geometry but neither captures the absolute strength of positive–negative margins nor identifies the depths at which retrieval discrimination emerges. We express each layer's positive–negative preference as a probability derived from the cosine-similarity margin. For triplet i, layer j, and model $M \in \{\mathrm{S},\mathrm{T}\}$, define

$$r^M_{i,j} = \cos(\boldsymbol{q}^M_{i,j}, \boldsymbol{p}^M_{i,j}) - \cos(\boldsymbol{q}^M_{i,j}, \boldsymbol{n}^M_{i,j}), \qquad p^M_{i,j} = \sigma(r^M_{i,j}/\tau). \quad (6)$$

With $\tau > 0$, $p^M_{i,j} > 1/2$ exactly when the positive outranks the negative. Figure 2 omits the triplet index i for simplicity. Using the row-normalized plans defined above, we form

$$\widetilde{p}^{\mathrm{T}}_{i,l} = \mathrm{clip}_\eta\left(\sum_m A^{\mathrm{ST}}_{lm} p^{\mathrm{T}}_{i,m}\right), \qquad \widetilde{p}^{\mathrm{S}}_{i,l} = \mathrm{clip}_\eta\left(\sum_k A^{\mathrm{SS}}_{lk} p^{\mathrm{S}}_{i,k}\right), \quad (7)$$

where clip_η maps its argument into $[\eta, 1-\eta]$, $\eta = 10^{-6}$.

To emphasize the retrieval transition, let $\xi_m \in [0,1]$ be the normalized, smoothed positive depth-change of the teacher's mean margin. We use $g_m = \alpha + (1-\alpha)\xi_m$, with $\alpha = 0.25$, and transport these weights as $h_l = \sum_m A^{\mathrm{ST}}_{lm} g_m$. After normalization, $w^{\mathrm{S}}_l = h_l / \sum_k h_k$; the implementation's intermediate mean normalizations cancel. The ranking objective is

$$\mathcal{L}_{\mathrm{rank}} = \rho(J), \qquad J = \sum_l w^{\mathrm{S}}_l d_l, \qquad d_l = \frac{1}{N}\sum_i D_{\mathrm{KL}}\left[\mathrm{Ber}(\widetilde{p}^{\mathrm{T}}_{i,l}) \,\|\, \mathrm{Ber}(\widetilde{p}^{\mathrm{S}}_{i,l})\right]. \quad (8)$$

2.4 Optimization

With the teacher frozen, we minimize $\mathcal{L} = \mathcal{L}_{\mathrm{out}} + \lambda_{\mathrm{str}}\mathcal{L}_{\mathrm{str}} + \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}}$, where $\mathcal{L}_{\mathrm{out}}$ is the output-level distillation loss, and λ_{str} and λ_{rank} weight structural and ranking supervision, respectively. The structural loss backpropagates through Sinkhorn iterations, whereas the ranking loss uses detached transport plans and emergence weights. Appendix A provides the corresponding theoretical analysis.

3 Experiments

3.1 Experimental Setup

Models and data. Following Akram et al. (2026), we use Qwen3-Embedding-4B as teacher (36 layers, token dimension 2560) and initialize the student from Qwen3-0.6B-Base (28 layers, token dimension 1024). We train on the 1.51M-example public E5 mixture curated by Springer et al. (2025).

Evaluation. We evaluate on 15 MTEB tasks following K et al. (2025), covering retrieval, reranking, clustering, pair classification, classification, and semantic textual similarity. Table 1 reports the primary MTEB metric for each task: nDCG@10 for retrieval, MAP for reranking, V-measure for clustering, average precision for pair classification, accuracy for classification, and Spearman correlation for semantic textual similarity (Muennighoff et al., 2023).

Table 1: MTEB results. Color hue indicates change from the corresponding Output KD baseline (green: gain; red: drop), while intensity indicates magnitude. Bold marks the best result within each host. Consistent with prior work (Vujanic & Rückstieß, 2026), hidden-state matching, attention matching, and their joint TinyBERT objective all reduce performance relative to Output KD in our experiments. We further evaluate RAIL-KD and EMD, for a total of five representative layerwise baselines. LST improves the average score under both output-level objectives.

Model / objective	Retrieval			Reranking		Clustering			Pair	Classification			STS			**Avg.**
	SciFact	ArguAna	NFCorpus	SO-Dup	SciDocs	BioRxiv	MedRxiv	20News	SprintDup	Banking77	Emotion	MASSIVE	STS17	SICK-R	STS-B	
Non-distilled																
Qwen3-Embedding-4B	79.09	74.84	41.02	59.34	89.58	44.06	40.44	59.15	98.14	86.20	68.17	76.21	91.63	87.91	93.67	72.63
Qwen3-0.6B-Base	0.22	0.03	1.07	16.06	25.07	4.52	13.14	9.93	1.58	3.01	17.21	4.08	18.78	26.76	8.37	9.99
Jina																
Output KD	71.25	60.23	35.41	51.62	81.05	33.50	32.96	48.13	**97.54**	80.29	54.81	49.40	80.50	83.17	89.49	63.29
+ Layerwise hidden	66.09	51.48	27.33	47.21	78.81	33.06	31.92	46.70	95.02	74.42	47.13	42.36	66.58	76.46	80.94	57.70
+ Layerwise attention	51.62	33.38	17.69	40.66	70.61	27.06	27.57	39.19	91.13	70.28	34.97	37.57	57.72	72.03	79.16	50.04
+ TinyBERT	42.69	32.83	13.64	37.70	68.91	24.89	25.14	35.57	84.23	61.38	35.88	32.40	55.23	71.16	75.08	46.45
+ RAIL-KD	71.52	59.51	34.20	51.61	80.70	34.01	33.21	48.23	97.08	79.76	54.88	48.55	78.77	82.95	89.28	62.95
+ EMD	69.96	58.90	33.10	49.76	78.78	33.71	32.27	44.27	96.18	78.09	53.79	46.64	75.95	82.07	88.58	61.47
+ LST	**72.88**	**61.25**	**36.15**	**51.98**	**82.18**	**34.98**	**33.52**	**49.26**	97.01	**81.22**	**56.65**	**50.95**	**81.09**	**83.88**	**89.71**	**64.18**
LEAF																
Output KD	72.23	60.76	34.93	51.85	81.20	**34.11**	32.69	48.26	**97.54**	80.89	55.35	49.55	80.20	83.14	89.54	63.48
+ Layerwise hidden	66.74	53.13	28.77	47.23	79.03	33.39	32.14	46.80	95.08	74.86	48.01	43.28	67.99	76.77	82.42	58.38
+ Layerwise attention	53.01	37.16	19.28	42.03	71.30	27.41	27.87	39.61	90.60	70.16	34.80	38.49	60.40	72.89	80.05	51.00
+ TinyBERT	43.34	34.62	15.09	40.63	68.64	24.97	25.86	37.15	88.40	62.44	35.65	33.76	58.16	72.31	77.06	47.87
+ RAIL-KD	71.88	60.15	34.49	51.75	80.49	33.63	**32.70**	48.32	97.03	79.98	54.55	48.74	79.47	83.05	89.34	63.04
+ EMD	70.27	59.78	33.77	50.05	79.03	33.67	32.59	44.44	96.37	78.82	53.35	46.98	77.46	82.48	89.23	61.89
+ LST	**72.25**	**61.65**	**36.04**	**52.02**	**82.03**	33.99	32.32	**48.33**	97.13	**81.40**	**56.85**	**50.61**	**81.22**	**84.07**	**89.63**	**63.97**

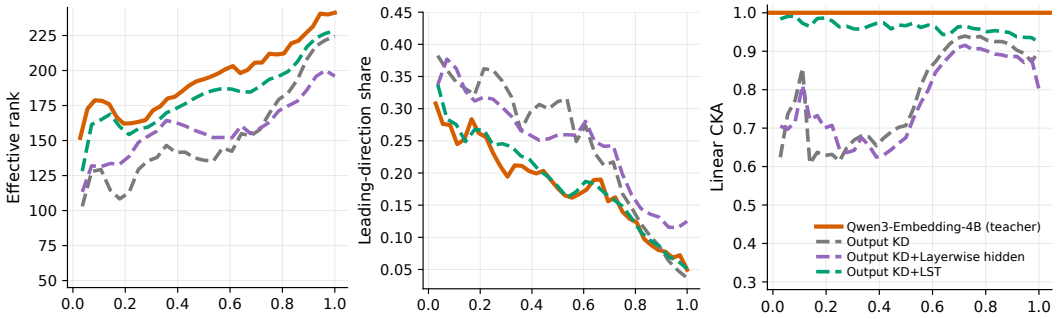
Baselines. We use two output-level objectives: Jina-style and LEAF-style matching after projecting the student final embedding to the teacher dimension (Akram et al., 2026; Vujanic & Rückstieß, 2026). On each host, we compare five intermediate-layer configurations: Layerwise hidden (projected hidden-state matching), Layerwise attention, their joint TinyBERT objective, RAIL-KD random ordered layer matching, and the hidden-state transport component of BERT-EMD (Jiao et al., 2020; Haidar et al., 2022; Li et al., 2020). All methods share the teacher, student, corpus and training budget. We run the main experiments with three different random seeds and report the mean results. Appendix B gives the baseline equations, architecture adaptations, code-level reproduction protocol, and the remaining training, evaluation, and probe details.

3.2 Overall Effectiveness

LST improves the 15-task average under both output-level objectives. We first test whether layerwise structural supervision contributes information beyond endpoint matching and whether this contribution persists across output-level objectives. We attach the same LST objectives to Jina and LEAF while holding the teacher, student, corpus, and training budget fixed, and compare against five representative intermediate-layer configurations (Table 1).

On the Jina host, the $+0.89$ gain over Output KD has a paired-task bootstrap 95% CI of $[+0.68, +1.10]$ across three seeds (Appendix B.2). Classification shows the most consistent gains, which is compatible with Relational Geometry Encoding's transfer of semantic neighborhood structure in the student's native space rather than the teacher's coordinate basis. Retrieval and reranking also improve, which is consistent with Emergence-Weighted Ranking's focus on relative candidate ordering (Tian et al., 2020; Huang et al., 2022; Tao et al., 2024). The STS gain is also consistent with preservation of graded similarity, rather than only discrete class or relevance boundaries.

These comparisons suggest that the gains depend on the form of intermediate supervision. Hidden-state matching, attention matching, and TinyBERT all underperform Output KD. These objectives



(a) Relational richness. (b) Directional concentration. (c) Same-depth alignment.

Figure 3: Layerwise recovery of the teacher trajectory. The three panels measure effective rank, directional concentration, and same-depth CKA. Output KD and Layerwise hidden remain farther from the teacher on these diagnostics, whereas LST follows the teacher's rank and concentration trajectories and reaches near-unity CKA. These patterns indicate that Relational Geometry Encoding recovers intermediate relational structure.

match hidden states or attention patterns, but do not explicitly preserve relations among text samples. RAIL-KD relaxes layer selection but retains hard layer pairs, while EMD expands layer coverage without changing the representations being matched. LST instead aligns the full sequences of native sample relations and uses the resulting correspondence to transfer ranking preferences. We report results on additional models in Appendix C.

3.3 Does Relational Geometry Encoding Recover the Teacher Trajectory?

The downstream gains in Table 1 do not by themselves show whether LST recovers the teacher's internal trajectory or merely regularizes the final embedding. We therefore examine the mechanism targeted by Relational Geometry Encoding: preserving teacher-like sample relations without requiring teacher and student token coordinates to share a dimension.

We build on the layerwise ablation of Vujanic & Rückstieß (2026), who apply direct matching of projected hidden states within text embedding distillation (Jiao et al., 2020). We compare Output KD+Layerwise hidden, Output KD, and LST on the Jina host, measuring every student in its native representation space. The same diagnostics for the remaining baselines appear in Appendix D.

We assess trajectory recovery using three complementary diagnostics (Figure 3). Effective rank measures the breadth of the relation spectrum, leading-direction share captures concentration along dominant directions, and same-depth CKA measures agreement with the corresponding teacher stage. Output KD has lower effective rank and greater directional concentration than the teacher despite matching the final embedding, showing that endpoint agreement alone does not determine geometry at intermediate layers. Layerwise hidden changes the relation spectrum only modestly and does not improve same-depth CKA. Thus, projected coordinate matching does not consistently bring the intermediate geometry closer to the teacher. In contrast, LST moves all three diagnostics toward the teacher in the student's native inference space, without relying on an intermediate projector, and reaches a same-depth CKA of 0.962. The query–document probes provide consistent evidence, with LST reducing the query–positive distance MAE on held-out MTEB triplets from 0.321 to 0.083 relative to Output KD (Appendix D.2). **LST recovers the teacher's intermediate relational trajectory in the student's native representation space.**

3.4 Does Soft Layer Alignment Improve Cross-Depth Correspondence?

Soft Layer Alignment preserves relational structure while covering the complete teacher trajectory, unlike fixed hard mappings and unconstrained full-coverage alignment. Models with different depths require an explicit correspondence between their complete layerwise trajectories. Predefined, randomly sampled, or manually selected layer pairs may connect different computational

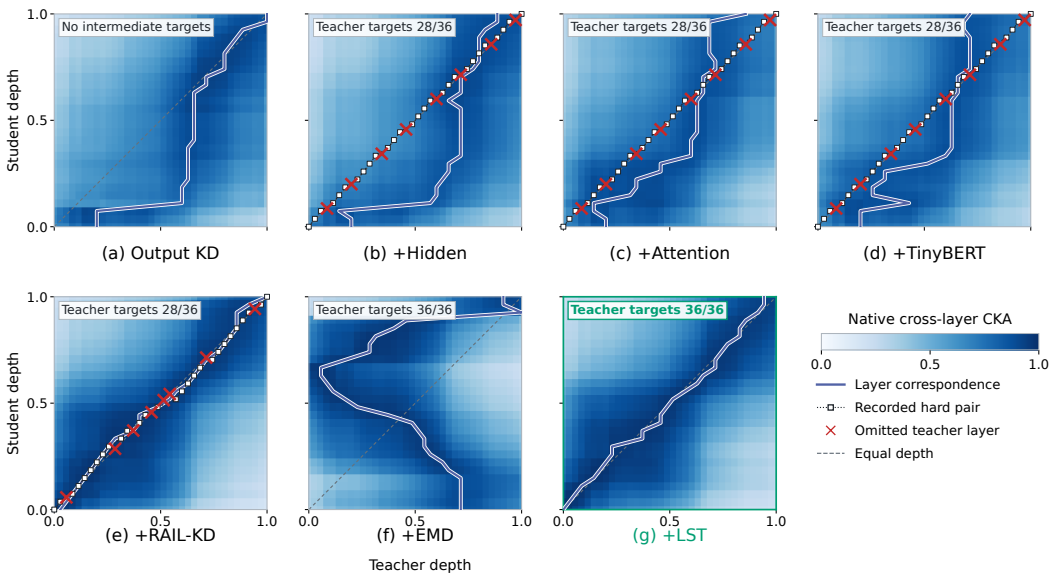


Figure 4: Teacher–student layer correspondence and hard-map coverage. Indigo curves show correspondence, squares recorded hard pairs, and red crosses omitted teacher layers. Dashed lines indicate equal depth, and insets direct teacher-layer coverage.

stages or leave parts of the teacher trajectory without direct supervision. We compare the pairs used by existing methods with post-hoc correspondences derived from the trained representations. We then test whether Soft Layer Alignment can replace these prescribed pairs with a learned correspondence over all teacher and student layers.

3.4.1 Diagnosing Teacher–Student Layer Correspondence

Existing layer assignments either miss parts of the teacher trajectory or align layers without preserving relational richness; LST avoids both failure modes. For each trained student layer, we identify the teacher layer with the highest post-hoc relation similarity. Figure 4 overlays this recovered path with the layer pairs actually used for supervision (squares) and the teacher layers they omit (red crosses).

The comparison identifies two limitations of prior alignment strategies. The fixed hidden-state, attention, and TinyBERT maps deviate from the recovered path and leave the same eight teacher layers unsupervised. RAIL-KD follows the recovered path more closely, but its hard pairs still yield substantially lower mean effective rank than both LST and the teacher. Conversely, EMD covers every teacher layer yet produces a displaced, nonmonotonic correspondence and the lowest mean effective rank in Table 2. Thus, these comparisons suggest that preserving relational structure is important in addition to selecting plausible layer pairs and covering the teacher trajectory.

Table 2: Mean and final effective rank with omitted teacher layers. The omitted sets are $\mathcal{F} = \{4, 8, 13, 17, 22, 26, 31, 35\}$ for the fixed maps and $\mathcal{R} = \{3, 11, 14, 17, 19, 20, 26, 34\}$ for RAIL-KD. LST combines complete layer coverage with teacher-like relational richness.

Objective	Mean	Final	Omitted layers
Teacher	194.7	241.3	–
Output KD	153.2	224.6	1–36
+ Hidden	158.4	195.9	$\mathcal{F}$
+ Attention	149.2	190.5	$\mathcal{F}$
+ TinyBERT	148.7	155.7	$\mathcal{F}$
+ RAIL-KD	117.8	212.0	$\mathcal{R}$
+ EMD	112.9	207.0	None
+ LST	**182.5**	**228.2**	**None**

LST uses a soft correspondence that covers the teacher sequence and remains close to the recovered path. It also improves the average score under both output-level objectives. The near-diagonal path is not an artifact of the depth prior: removing the depth term still leaves 99.93% of the transport mass near-diagonal, whereas depth alone gives 71.20% (Appendix D.6).

3.4.2 Effect of Correspondence Softness

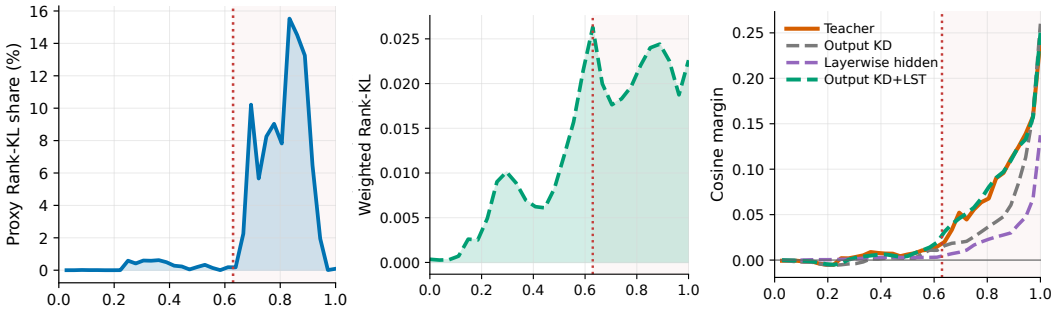
Table 3: Soft correspondence and row-hard assignment. Δ Overall is relative to Row-hard; effective teacher layers measure mean exponentiated row entropy at fixed matching cost.

Mapping / ε	Overall	Δ Overall vs. Row-hard	Effective teacher layers
Row-hard	63.85	–	1.00
0.01	63.96	+0.11	2.19
0.02	**64.18**	**+0.33**	**2.76**
0.05	64.16	+0.31	4.16
0.10	64.14	+0.30	5.94
0.20	64.07	+0.22	8.66

Moderately soft correspondence performs best; broadening support further does not improve Overall. Soft Layer Alignment learns a distributed correspondence rather than assigning each student layer a single teacher index, with entropy controlling the breadth of supervision. To examine this mechanism, we retrain the Jina-host recipe across positive entropy values and compare it with a row-hard rule selecting the minimum structure-plus-depth cost independently in each row. The positive-entropy runs keep the remaining objective fixed; downstream scores come from separate training runs, while effective teacher support is recomputed on one fixed final-checkpoint cost matrix.

Table 3 shows how entropy controls the breadth of the learned correspondence and, in turn, downstream performance. Relative to the row-hard score of 63.85, soft alignment reaches 64.18 Overall at $\varepsilon = 0.02$, where each student layer receives support from 2.76 effective teacher layers. All tested soft configurations outperform the row-hard assignment, while Overall varies only modestly across positive entropy values and decreases slightly as the correspondence becomes overly diffuse.

3.5 Does Emergence-Weighted Ranking Track the Ranking Transition?



(a) Diagnostic proxy. (b) Training-objective loss. (c) Retrieval-margin recovery.

Figure 5: Emergence-Weighted Ranking diagnostics: (a) offline Teacher–Output KD Rank-KL by depth, (b) transported ranking loss at the final LST checkpoint, and (c) teacher and student retrieval margins across depth. The red dotted line marks normalized depth 0.63, where the implemented loss peaks. LST places ranking supervision at the layers where the teacher's retrieval preferences emerge and recovers their margin strength.

Emergence-Weighted Ranking concentrates supervision where retrieval preferences appear and improves the student's margin trajectory in that region. The component study in Appendix E shows that Emergence-Weighted Ranking provides clear downstream gains. Here, we test whether its layerwise emphasis follows the teacher's retrieval transition observed in Figure 1.

Before applying the ranking objective, we use an offline diagnostic to identify where ranking preferences differ between the teacher and the Output KD student. Figure 5(a) plots the same-depth Rank-KL mismatch against normalized layer depth, with the vertical axis reporting each depth's share of the total proxy mismatch. The mismatch is concentrated in the later layers: 95.2% of the total lies at or beyond depth 0.63, where the teacher's retrieval margin begins to increase. **This pattern motivates focusing Emergence-Weighted Ranking on these late layers to transfer the teacher's emerging positive–negative separation to the student.**

Figure 5(b) uses normalized student depth on the horizontal axis and the per-layer transported, emergence-weighted Rank-KL on the vertical axis. The loss peaks at student layer 18 of 28 (depth 0.63), with 67.0% of its mass at that layer or later. Thus, LST places ranking supervision in the late-depth region identified by the offline diagnostic rather than imposing an arbitrary fixed cutoff. Figure 5(c) uses the same depth axis and plots the mean cosine retrieval margin, $\cos(q, p) - \cos(q, n)$, for the teacher and each student. Across depths 0.55–0.80, LST reduces the mean teacher-relative margin error by 73.6% compared with Output KD. **This shows that Emergence-Weighted Ranking module learns where the teacher's retrieval preferences emerge and how strongly positives should be separated from negatives.**

4 Related Work

Relational representation distillation. Since matching the teacher's hidden activations coordinate-by-coordinate is sensitive to width and basis (Romero et al., 2015; Sun et al., 2019; Jiao et al., 2020; Sun et al., 2020), relational distillation transfers relations among samples instead: pairwise distances and similarities (Park et al., 2019; Tung & Mori, 2019), probabilistic structure (Passalis & Tefas, 2018), contrastive information (Tian et al., 2020; Nguyen et al., 2024), attention relations (Wang et al., 2020; Truong et al., 2025; Vu et al., 2026; Li et al., 2026), and, most closely related to us, similarity between hidden states of different widths (Dasgupta & Cohn, 2025; Dao et al., 2026). These methods show that sample relations can be transferred, but most operate at a single layer or reduce relations across depth to one summary statistic. They therefore leave the evolution of relational structure unspecified. LST retains the complete depth-indexed sequence of native sample relations and learns a cross-depth correspondence that determines which teacher stages supervise each student layer.

Cross-depth alignment and optimal transport. Prior work fixes a uniform layer map, randomizes an ordered teacher subset each epoch to reduce this bias, or replaces hard assignment with many-to-many layer transport driven by hidden-state distances (Jiao et al., 2020; Haidar et al., 2022; Li et al., 2020). Optimal transport underlies related objectives elsewhere in distillation and model merging, including component fusion (Imfeld et al., 2024), selective transfer from an inverse-transport view (Shi et al., 2025), and output-distribution alignment within a batch or across tokenizers (Cui et al., 2024; 2025). The recurring limitation is that alignment operates on width-dependent coordinates and optimizes a map in isolation, so a well-aligned assignment need not carry the teacher's relational structure. LST instead computes transport over a width-independent relation cost with a soft depth prior, and then reuses the same correspondence to route ranking supervision.

Layerwise representation emergence. Studies of language models have examined how hidden representations change across depth, including the emergence of high-dimensional abstractions and the evolution of features through the residual stream (Skean et al., 2025; Cheng et al., 2025; Liu et al., 2025; Lawson et al., 2025; Yang et al., 2025). These studies show that intermediate computation depends on depth, but they focus on representation formation within a model rather than trajectory transfer between embedding teachers and students. Text embedding distillation, by contrast, primarily matches final embedding geometry and does not directly supervise these layerwise dynamics (Kim et al., 2023; Vera et al., 2025; Zhang et al., 2024; Feng & Hong, 2025). LST targets this unsupervised component by transferring depth-dependent positive–negative retrieval preferences.

5 Conclusion

We introduce LST, a training-only plugin for transferring a teacher's layerwise trajectory during text embedding distillation. LST compares width-mismatched representations through their sample relations and learns a soft alignment between teacher and student layers. It reuses this alignment to transfer ranking preferences, emphasizing the depths at which the teacher separates relevant from irrelevant documents. LST improves the 15-task average while moving the student's intermediate geometry and retrieval-margin trajectory toward the teacher, without changing the inference architecture. To our knowledge, these results provide the first evidence that layerwise trajectories can be effectively transferred in text embedding distillation.

AI USE STATEMENT

Generative AI tools were used for limited language polishing of author-written text and retrieval related work, with all suggested edits reviewed and revised by the authors. The proposed framework, experimental design, and analysis are the original contributions of the authors.

REPRODUCIBILITY STATEMENT

Section 3.1 summarizes the models, data, baselines, and evaluation. Appendix B provides the complete training settings, evaluation protocols, diagnostic probes, and baseline implementations. The accompanying artifact contains the code, experiment configurations, and scripts needed to reproduce the reported results. Our anonymized code repository is available at https://anonymous.4open.science/r/LST-8C30/.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Mohammad Kalim Akram, Saba Sturua, Nastia Havriushenko, Quentin Herreros, Michael Günther, Maximilian Werk, and Han Xiao. jina-embeddings-v5-text: Compact and robust text embedding models using task-targeted distillation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4454–4458. ACM, 2026. doi: 10.1145/3805712.3808455. URL https://doi.org/10.1145/3805712.3808455.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7028–7036, 2021. doi: 10.1609/aaai.v35i8.16865. URL https://ojs.aaai.org/index.php/AAAI/article/view/16865.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *International Conference on Learning Representations*, volume 2025, pp. 35389–35414, 2025.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5656–5668, 2021. URL https://arxiv.org/abs/2110.14633.

Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. Sinkd: Sinkhorn distance minimization for knowledge distillation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):11887–11901, 2024.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. In *Proceedings of AAAI*, pp. 23724–23732, 2025. doi: 10.1609/aaai.v39i22.34543.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Linh Ngo Van, Diep Thi-Ngoc Nguyen, Thien Huu Nguyen, and Trung Le. TALAS: Teacher-anchored layer alignment with adaptive sharpness-aware minimization for embedding distillation. *arXiv preprint arXiv:2606.21851*, 2026. doi: 10.48550/arXiv.2606.21851. URL https://arxiv.org/abs/2606.21851.

Sayantan Dasgupta and Trevor Cohn. Improving language model distillation through hidden state matching. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IcVSKhVpKu.

Shuya Feng and Yuan Hong. DPED: Multi-layer noise distillation for privacy-preserving text embeddings. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25237–25245. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1282. URL https://aclanthology.org/2025.emnlp-main.1282/.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://iclr.cc/virtual/2024/poster/19420.

Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. MiniPLM: Knowledge distillation for pre-training language models. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/28062.

Md Akmal Haidar, Nithin Anchuri, Mehdi Rezagholizadeh, Abbas Ghaddar, Philippe Langlais, and Pascal Poupart. RAIL-KD: Random intermediate layer mapping for knowledge distillation. In *Findings of NAACL*, 2022. URL https://aclanthology.org/2022.findings-naacl.103.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/da669dfd3c36c93905a17ddba01eef06-Abstract-Conference.html.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Muhammed Ildiz, Alperen Gozeten, Ege Taga, Marco Mondelli, and Samet Oymak. High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling laws. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/31172.

Moritz Imfeld, Jacopo Graldi, Marco Giordano, Thomas Hofmann, Sotiris Anagnostidis, and Sidak Pal Singh. Transformer fusion with optimal transport. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=LjeqMvQpen.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of EMNLP*, 2020.

Thennal D K, Tim Fischer, and Chris Biemann. Large language models are overparameterized text encoders. In *Proceedings of the 10th Workshop on Representation Learning for NLP*, pp. 170–184, 2025. URL https://aclanthology.org/2025.repl4nlp-1.13/.

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. EmbedDistill: A geometric knowledge distillation for information retrieval. *arXiv preprint arXiv:2301.12005*, 2023. URL https://arxiv.org/abs/2301.12005.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=lsHZNNoC7r.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=rc65N9xIrY.

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

Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual stream analysis with multi-layer SAEs. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/29317.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lgsyLSsDRe.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Defu Lian, Yingxia Shao, and Zheng Liu. Making text embedders few-shot learners. In *International Conference on Learning Representations*, volume 2025, pp. 38107–38124, 2025a.

Jianquan Li, Xiaokang Liu, Honghong Zhao, Ruifeng Xu, Min Yang, and Yaohong Jin. BERT-EMD: Many-to-many layer mapping for BERT compression with earth mover's distance. In *Proceedings of EMNLP*, 2020. URL https://aclanthology.org/2020.emnlp-main.242/.

Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. ESE: Espresso sentence embeddings. In *International Conference on Learning Representations*, 2025b. URL https://iclr.cc/virtual/2025/poster/28282.

Zhen Li, Xin Huang, Liang Wang, Nan Yang, Ting Song, Yan Xia, Xun Wu, Shaohan Huang, Huishuai Zhang, Furu Wei, and Dongyan Zhao. BitNet Text Embeddings. *arXiv preprint arXiv:2606.25674*, 2026. doi: 10.48550/arXiv.2606.25674. URL https://arxiv.org/abs/2606.25674.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20852–20867. PMLR, 2023. URL https://proceedings.mlr.press/v202/liang23j.html.

Ziyin Liu, Isaac Chuang, Tomer Galanti, and Tomaso Poggio. Formation of representations in neural networks. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/29863.

Zhuoyuan Mao and Tetsuji Nakagawa. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1886–1894. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.138. URL https://aclanthology.org/2023.eacl-main.138/.

John X. Morris and Alexander Rush. Contextual document embeddings. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/29341.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BC4lIvfSzv.

Cong-Duy Nguyen, Thong Nguyen, Xiaobao Wu, and Anh Tuan Luu. KDMCSE: Knowledge distillation multimodal sentence embeddings with adaptive angular margin contrastive 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)*, pp. 733–749. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.naacl-long.42/.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of CVPR*, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of ECCV*, 2018.

Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. ALP-KD: Attention-based layer projection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13657–13665, 2021. doi: 10.1609/aaai.v35i15.17610. URL https://ojs.aaai.org/index.php/AAAI/article/view/17610.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4512–4525. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.365. URL https://aclanthology.org/2020.emnlp-main.365/.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Bowen Shi, Xiaopeng Zhang, Yaoming Wang, Li Jin, Wenrui Dai, Junni Zou, Hongkai Xiong, and Qi Tian. Hybrid distillation: Connecting masked autoencoders with contrastive learners. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5598cf1b2905a26ddb863e6705588327-Abstract-Conference.html.

Liangliang Shi, Zhengyan Shi, and Junchi Yan. SelKD: Selective knowledge distillation via optimal transport perspective. In *International Conference on Learning Representations*, 2025.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/29025.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of ICML*, 2025. URL https://openreview.net/forum?id=WGXb7UdvTX.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition improves language model embeddings. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ahlrf2HGJR.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1102–1121. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-acl.71/.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of EMNLP-IJCNLP*, 2019.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. MobileBERT: A compact task-agnostic BERT for resource-limited devices. In *Proceedings of ACL*, 2020.

Chongyang Tao, Chang Liu, Tao Shen, Can Xu, Xiubo Geng, Binxing Jiao, and Daxin Jiang. ADAM: Dense retrieval distillation with adaptive dark examples. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11639–11651. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.692. URL https://aclanthology.org/2024.findings-acl.692/.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgpBJrtvS.

Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. 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*, pp. 7594–7606, 2025. doi: 10.18653/v1/2025.emnlp-main.385. URL https://aclanthology.org/2025.emnlp-main.385/.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of ICCV*, 2019.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. How does BERT answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1823–1832, 2019. doi: 10.1145/3357384.3358028. URL https://doi.org/10.1145/3357384.3358028.

Henrique Schechter Vera, Sahil Dua, et al. EmbeddingGemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025. URL https://arxiv.org/abs/2509.20354.

Hai An Vu, Minh-Phuc Truong, Tu Vu, and Linh Ngo Van. Mol: Mixture of layers in cross-tokenizer embedding model distillation. *Know.-Based Syst.*, 343(C), July 2026. ISSN 0950-7051. doi: 10.1016/j.knosys.2026.116001. URL https://doi.org/10.1016/j.knosys.2026.116001.

Robin Vujanic and Thomas Rückstieß. LEAF: Knowledge distillation of text embedding models with teacher-aligned representations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43362–43383. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2008. URL https://aclanthology.org/2026.acl-long.2008/.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2020.

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. Promptriever: Instruction-trained retrievers can be prompted like language models. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/28345.

Jian Xie, Xin He, Jiyang Wang, Zimeng Qiu, Ali Kebarighotbi, and Farhad Ghassemi. SimTDE: Simple transformer distillation for sentence embeddings. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2389–2393. ACM, 2023. doi: 10.1145/3539618.3592063. URL https://doi.org/10.1145/3539618.3592063.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. DistillCSE: Distilled contrastive learning for sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8153–8165. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.547. URL https://aclanthology.org/2023.findings-emnlp.547/.

Eugene Yang, Dawn Lawrie, James Mayfield, Douglas W. Oard, and Scott Miller. Translate-distill: Learning cross-language dense retrieval by translation and distillation. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval (ECIR)*, pp. 50–65. Springer, 2024.

Yukang Yang, Declan Campbell, Kaixuan Huang, Mengdi Wang, Jonathan Cohen, and Taylor Webb. Emergent symbolic mechanisms support abstract reasoning in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=y1SnRPDWx4.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of SOTA embedding models. *arXiv preprint arXiv:2412.19048*, 2024. URL https://arxiv.org/abs/2412.19048.

Zikai Zhou, Yunhang Shen, Shitong Shao, Linrui Gong, and Shaohui Lin. Rethinking centered kernel alignment in knowledge distillation. In *Proceedings of IJCAI*, pp. 5680–5688, 2024. doi: 10.24963/ijcai.2024/628. URL https://www.ijcai.org/proceedings/2024/628.

Technical Appendix

TABLE OF CONTENTS

A Theoretical Analysis of LST **16**
A.1 Limits of Output Supervision . . . 16
A.2 Preliminaries . . . 16
A.3 Native Geometry and Retrieval Margins . . . 16
A.4 Depth-Local Trajectory Alignment . . . 17
A.5 Weighted Preference Transfer . . . 17
A.6 Connection to the Training Losses . . . 18

B Experimental Protocols **18**
B.1 Training Configuration . . . 18
B.2 Result Stability . . . 19
B.3 Evaluation and Diagnostic Probes . . . 19
B.4 Baseline Implementations . . . 20

C Student Transfer and Objective Comparisons **22**
C.1 Qwen3.5-0.8B Student . . . 22
C.2 Negative-Aware Training Objectives . . . 23

D Trajectory Diagnostics **23**
D.1 Teacher Trajectories Across Model Families . . . 23
D.2 Semantic Geometry and Retrieval Separation . . . 24
D.3 Structural Similarity Across Depth . . . 24
D.4 Baseline Spectra and Alignment . . . 25
D.5 Distance Geometry Across Baselines . . . 28
D.6 Layer Correspondence from Relation Costs . . . 28

E Component and Sensitivity Studies **28**
E.1 Relation Operators . . . 29
E.2 Layer Assignment . . . 29
E.3 Entropy Regularization . . . 30
E.4 Auxiliary Ranking Targets . . . 31

F Limitations **32**

A Theoretical Analysis of LST

A.1 Limits of Output Supervision

The layerwise mismatch in Figure 1 can persist even under exact output matching. The following counterexample isolates what endpoint supervision cannot determine.

Proposition A.1 (equal outputs can hide opposite intermediate rankings). There exist two-layer linear teacher and student encoders with identical outputs for every input but opposite intermediate cosine rankings on the same triplet.

Proof. Let the teacher apply two identity maps and the student apply $M = \mathrm{diag}(1, 3)$ followed by M^{-1}. Both return x for every row-vector input x. For $q = (2, 1)$, $p = (1, 0)$, and $n = (0, 1)$, their intermediate cosine margins are

$$\begin{aligned} r^T &= \cos(q, p) - \cos(q, n) = \frac{1}{\sqrt{5}} > 0, \\ r^S &= \cos(qM, pM) - \cos(qM, nM) = -\frac{1}{\sqrt{13}} < 0. \end{aligned} \tag{A.1}$$

The second student map cancels the intermediate distortion exactly. □

Implication and scope. Output equality does not identify intermediate relations. The same construction also distinguishes projected from native matching: if xM is the deployed student representation and M^{-1} is a training-only projector, projected outputs agree while native rankings differ. Neither observation implies that every projector is harmful or that intermediate matching is necessary for accurate retrieval; identical *native* final outputs already give identical final rankings. They establish only that output supervision alone does not enforce the internal geometry. The remainder of this section characterizes the additional constraints supplied by LST; whether these constraints help a capacity-limited student trained on finite data is an empirical question.

A.2 Preliminaries

We next give three batch-level results: relation matching controls normalized retrieval margins, depth-regularized alignment controls deviation from the teacher trajectory, and weighted preference matching controls ranking errors under explicit conditions. These are deterministic statements about the achieved representations and plans, not guarantees of optimization or generalization.

Use the notation of Section 2, writing S, T for student and teacher and omitting boldface. Both models encode the same ordered batch of $B \geq 3$ texts at $L_S, L_T \geq 2$ layers. Write $G = UU^\top$ and use t, K from Eq. (1). Assume feature norms exceed 10^{-12} and centered traces are at least 10^{-8}, so the normalized rows are unit vectors and the trace floor is inactive. The $N \geq 1$ triplets refer to distinct query, positive, and negative row indices within each triplet. Let A^{ST}, A^{SS} be the row-normalized finite plans, with original row masses at least 10^{-12}; thus both are nonnegative and row-stochastic. The teacher scale $s_T > 0$ retains its floor in Eq. (2). No exact transport optimality or marginal convergence is assumed.

A.3 Native Geometry and Retrieval Margins

Proposition A.2 (native relations control retrieval margins). For a triplet (q, p, n) with cosine margin r, define

$$\mathcal{M}_{qpn}(K) = K_{qp} - K_{qn} + \tfrac{1}{2}(K_{nn} - K_{pp}). \tag{A.2}$$

Then $r/t = \mathcal{M}_{qpn}(K)$, and any student–teacher layer pair satisfies

$$\left| \frac{r^S}{t^S} - \frac{r^T}{t^T} \right| \leq \sqrt{\frac{3}{2}} \left\| K^S - K^T \right\|_F . \tag{A.3}$$

Proof. Centering leaves row differences unchanged, so $\|u_a - u_b\|^2 = t(K_{aa} + K_{bb} - 2K_{ab})$. Unit row norms give $2r = \|u_q - u_n\|^2 - \|u_q - u_p\|^2$, proving the identity. On symmetric matrices, $\mathcal{M}_{qpn}$ is a Frobenius inner product with four off-diagonal and two diagonal coefficients of magnitude $1/2$. Its coefficient norm is $\sqrt{3/2}$, so Cauchy–Schwarz proves the bound. □

Ordering versus strength. Exact relation matching implies $r^S = (t^S/t^T)r^T$ and therefore preserves every within-batch cosine ordering in the native student space. It need not be attainable when the teacher operator's rank exceeds the student width. Nor does it fix absolute margin strength: for unit vectors v_a with a nonzero triplet margin, consider $u_a^{(b)} = (\sqrt{b}\, v_a, \sqrt{1-b})$, $0 < b \le 1$. Their Gram matrix is $bVV^\top + (1-b)\mathbf{1}\mathbf{1}^\top$. Centering removes the shared component and trace normalization cancels b, giving the same K but margins scaled by b whenever the trace floor is inactive. Thus ranking targets add preference-strength information, not an ordering signal entirely absent from relational geometry.

A.4 Depth-Local Trajectory Alignment

Proposition A.3 (structure and depth control trajectory error). Let $K_T^\dagger$ linearly interpolate the observed teacher operators on $[0,1]$, and set $L_K = \max_{m<L_T} \|K_{m+1}^T - K_m^T\|_F/(\delta_{m+1}^T - \delta_m^T)$. Define the actual row cost $C_l = \sum_m A_{lm}^{ST} C_{lm}^{ST}$, its average $C_A = L_S^{-1}\sum_l C_l$, and $c_K = B^2 s_T + L_K^2/\lambda_d$, where $\lambda_d > 0$. Then

$$\begin{aligned} \|K_l^S - K_T^\dagger(\delta_l^S)\|_F^2 &\le c_K C_l, \\ \frac{1}{L_S}\sum_l \|K_l^S - K_T^\dagger(\delta_l^S)\|_F^2 &\le c_K C_A. \end{aligned} \tag{A.4}$$

Proof. The interpolant is L_K-Lipschitz, including across breakpoints. Put $F_l = \sum_m A_{lm}^{ST}\|K_l^S - K_m^T\|_F^2$ and $D_l = \sum_m A_{lm}^{ST}(\delta_l^S - \delta_m^T)^2$. The triangle inequality and weighted Cauchy–Schwarz give

$$\begin{aligned} \|K_l^S - K_T^\dagger(\delta_l^S)\|_F &\le \sum_m A_{lm}^{ST}\big(\|K_l^S - K_m^T\|_F + L_K|\delta_l^S - \delta_m^T|\big) \\ &\le \sqrt{F_l} + L_K\sqrt{D_l} \le \sqrt{c_K\big(F_l/(B^2 s_T) + \lambda_d D_l\big)} = \sqrt{c_K C_l}. \end{aligned}$$

The last inequality is Cauchy–Schwarz in two dimensions. Squaring and averaging proves both claims. □

Consequence for the final embedding. At $l_* = L_S$, the reference is exactly $K_{L_T}^T$ because $\delta_{l_*}^S = 1$. Writing $\epsilon_* = \sqrt{c_K C_{l_*}}$, Proposition A.2 yields the sufficient condition

$$\frac{r_{i,L_T}^T}{t_{L_T}^T} > \sqrt{\frac{3}{2}}\,\epsilon_* \quad\Longrightarrow\quad r_{i,l_*}^S > 0. \tag{A.5}$$

Hence only triplets whose final teacher normalized margin is below or at this threshold can have nonpositive student margins. The bound concerns locality on the normalized-depth grid; it does not assert that equal depths are semantically equivalent stages.

A.5 Weighted Preference Transfer

Write $P_{il} = \widetilde{p}_{i,l}^T$ and $Q_{il} = \widetilde{p}_{i,l}^S$ for the clipped mixtures in Eq. (7), and use natural-log Bernoulli KL, denoted kl. Let $d_l = N^{-1}\sum_i \mathrm{kl}(P_{il}\|Q_{il})$ and $J = \sum_l w_l^S d_l$ as in Eq. (8). The result below uses only $0 < \alpha \le 1$ and $0 \le \xi_m \le 1$; the specific emergence signal is detailed in Appendix B.1.

Proposition A.4 (preference error and native ranking). For $h_l = \sum_m A_{lm}^{ST}[\alpha + (1-\alpha)\xi_m]$ and $w_l^S = h_l/\sum_k h_k$,

$$\frac{1}{NL_S}\sum_{i,l}|P_{il} - Q_{il}|^2 \le \frac{J}{2\alpha}, \qquad d_l \le \frac{J}{w_l^S}. \tag{A.6}$$

To relate the mixtures to a native layer, define $V_{il} = \sum_k A_{lk}^{SS}|p_{ik}^S - p_{il}^S|$. For $0 \le \beta < \gamma \le 1/2 - \eta$, let b_l be the fraction of triplets satisfying $P_{il} < 1/2 + \gamma$ or $V_{il} > \beta$. Then

$$\mathrm{Err}_l := \frac{1}{N}\sum_i \mathbf{1}\{r_{i,l}^S \le 0\} \le b_l + \frac{d_l}{2(\gamma-\beta)^2} \le b_l + \frac{J}{2w_l^S(\gamma-\beta)^2}. \tag{A.7}$$

Proof. Stochastic averaging gives $\alpha \le h_l \le 1$, hence $w_l^S \ge \alpha/L_S$. Bernoulli KL satisfies $\mathrm{kl}(a\|b) \ge 2(a-b)^2$: as a function of a, it has curvature $1/[a(1-a)] \ge 4$ and value and slope zero at $a = b$. Summing over triplets and layers gives Eq. (A.6), using $J \ge w_l^S d_l$ for its second inequality. If $r_{i,l}^S \le 0$, then $p_{il}^S \le 1/2$. Outside the fraction b_l, the student self-mixture is at most $1/2 + \beta$. Clipping preserves this upper bound, while $P_{il} \ge 1/2 + \gamma$, so every such error incurs KL at least $2(\gamma - \beta)^2$. Summing these contributions proves Eq. (A.7). □

Role of emergence and self-mixing. The weight ratio is at most $1/\alpha$, or four for the default $\alpha = 0.25$; when $\xi = 0$, weights are uniform. Emergence therefore reallocates the error budget without excluding any layer, but the bound does not establish that this particular weighting rule is optimal. Equation (A.7) applies also to $l = L_S$ and explicitly requires confident transported teacher preferences and limited self-mixing; the latter is observable, with $V_{il} \le 1 - A_{ll}^{SS}$. Small mixed KL alone cannot guarantee native ordering or equality of all unclipped margins.

A.6 Connection to the Training Losses

The preceding bounds use the actual row costs and pre-transform ranking loss. To connect them to the optimized scalars, let $\mathcal{R}_{XY} = \langle \widehat{\pi}^{XY}, C^{XY} \rangle$, $\widehat{a}_l = \sum_m \widehat{\pi}_{lm}^{ST}$, and $\chi = L_S \min_l \widehat{a}_l > 0$. For $\psi(y) = (y/(1-y))^2$, $0 \le y < 1$, Eq. (5) gives

$$C_A \le \frac{\mathcal{R}_{ST}}{\chi} \le \frac{\psi(\mathcal{L}_{\mathrm{str}}) + (\mathcal{R}_{SS} + \mathcal{R}_{TT})/2}{\chi}, \qquad J \le \psi(\mathcal{L}_{\mathrm{rank}}). \tag{A.8}$$

Indeed, $\mathcal{R}_{ST} = \sum_l \widehat{a}_l C_l \ge \chi C_A$ by nonnegativity of row costs, and $\psi(\rho(x)) = \max\{x, 10^{-12}\} \ge x$ proves the remaining inequalities. Thus a small structural scalar alone does not certify small trajectory error: the self-costs and row coverage must also be accounted for.

The implementation uses finite Sinkhorn iterations, floored kernels, and cost-only self-debiasing. The analysis applies to the resulting row-stochastic plans, with the stated normalization assumptions, and concerns only the batch and triplets on which its quantities are measured.

B Experimental Protocols

B.1 Training Configuration

The main runs use Qwen3-Embedding-4B as teacher and Qwen3-0.6B-Base as student initialization, with approximately 1.51 million examples from the public E5 training mixture (Springer et al., 2025). A fixed exclusion list removes synonymous-prompt probes and held-out retrieval cases before training. Queries, positives, and available negatives enter a common text batch: relation operators use all encoded texts, while the ranking branch uses explicit triplets. Unless specified otherwise, training uses one epoch, eight GPUs, per-device record batch size 8, accumulation 1, global record batch size 64, maximum sequence length 512, and bfloat16. The initial learning rate is 2×10^{-4}, with 500 warmup steps, cosine decay, and minimum learning-rate ratio 0.1. LoRA rank, scaling, and dropout are 16, 32, and 0.05.

The final Jina model uses $\lambda_{\mathrm{str}} = \lambda_{\mathrm{rank}} = 0.1$, $\lambda_d = 1$, $\varepsilon = 0.02$, and 20 Sinkhorn iterations. The LEAF-host result uses the same loss weights and solver budget but $\varepsilon = 0.05$. The ranking temperature is 0.05 and the emergence floor is 0.25. Appendix C.1 records the Qwen3.5 training settings.

Emergence signal. Let $\bar{r}_m = N^{-1} \sum_i r_{i,m}^T$ and let S be a moving average with replicated boundaries and the largest odd window at most $\min(5, L_T)$. The implementation computes

$$\begin{aligned} v_1 &= 0, \qquad v_m = (L_T - 1)[(\mathsf{S}\bar{r})_m - (\mathsf{S}\bar{r})_{m-1}]_+, \quad m \ge 2, \\ z &= \mathsf{S}v, \qquad \xi_m = \begin{cases} z_m / \max_j z_j, & \max_j z_j > 10^{-8}, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{B.1}$$

Thus smoothing precedes and follows the positive depth difference. The peak-normalized ξ_m enters Section 2.3; absent a positive signal above the threshold, the weights are uniform.

Training cost. Separate eight-A800-80GB runs with 23,576 updates recorded a 4.36% wall-clock increase for LST. All auxiliary computations are training-only and do not alter the deployed student.

B.2 Result Stability

Table 4 reports the task-wise standard deviations for the main Jina-host LST configuration.

Table 4: Standard deviations in score points for the main Jina-host LST configuration.

Task	SciFact	ArguAna	NFCorpus	SO-Dup	SciDocs	BioRxiv	MedRxiv	20News	SprintDup	Banking77	Emotion	MASSIVE	STS17	SICK-R	STS-B	Overall
Std	0.43	0.10	0.49	0.06	0.10	0.14	0.17	0.46	0.07	0.34	0.77	0.25	0.13	0.09	0.09	**0.07**

The Overall standard deviation is 0.07 score points, indicating low variation across repeated training and evaluation. The 95% Student-t confidence interval (CI) for the mean Overall score is $[64.01, 64.35]$. Relative to the Jina Output KD host, LST improves Overall by $+0.89$ score points on average, with a paired-task bootstrap 95% CI of $[+0.68, +1.10]$. Together with the mean improvements over the output-level objectives, this low seed variation indicates that the gains are stable across repeated training and evaluation. Because text-embedding benchmarks aggregate over many examples and diverse tasks, small changes in the aggregate score reflect broad rather than task-specific shifts. The gains from layerwise geometry, alignment, and ranking accumulate at the level of a single shared embedding model across tasks.

B.3 Evaluation and Diagnostic Probes

Downstream evaluation. We use MTEB 1.14.12 and the 15-task selection of K et al. (2025), with left padding, final-token pooling, and L2-normalized embeddings. Asymmetric retrieval uses instructed queries and plain documents; symmetric tasks apply the task instruction on both sides. Overall is the equal-weight mean over 15 tasks, not over six category means. Retrieval, reranking, clustering, pair classification, classification, and STS contain 3, 2, 3, 1, 3, and 3 tasks. Table abbreviations follow Table 1.

Representation probes. The fixed layerwise probe contains 56 synonymous-query groups and 96 E5 triplets from MS MARCO, NQ, HotpotQA, FEVER, AllNLI, and Quora. Diagnostics use native representations without intermediate projectors. Effective-rank and CKA summaries in Figure 3 and Appendix D use the full depth range; the Gram-entropy comparison explicitly uses depths 0.25–0.75. Geometry and spectrum curves retain the recorded depth j/L; the transport grid uses $(j-1)/(L-1)$ as in Section 2.2. Same-depth CKA selects the nearest student depth for each teacher layer. Effective rank is the exponentiated entropy of normalized singular values of centered features; leading-direction share is the largest centered Gram eigenvalue divided by their sum. CKA compares centered sample relations, so its agreement need not imply equal spectral richness.

Figure 4 uses native cross-layer CKA and recorded training pairs, not a shared OT reinterpretation of the baselines. The omitted-layer sets in Table 2 identify teacher layers without direct intermediate supervision; EMD and LST cover all teacher layers through positive teacher-column constraints. For Table 3, effective teacher layers are the mean exponentiated entropy of row-normalized plans recomputed on one fixed final-LST cost matrix. The separate common-OT probe in Figure 15 uses last-token relation operators, $\varepsilon = 0.02$, depth penalty 1, and 20 iterations for every checkpoint. Its near-diagonal mass is the fraction within absolute teacher–student depth difference 0.10, not a measurement of the baselines' native training assignments.

Ranking and distance probes. Figure 5(a) constructs an offline same-depth proxy from mean teacher and Output KD margins. Panel (b) recomputes the implemented per-layer transported, weighted Rank-KL on the final LST checkpoint using 96 fixed E5 triplets and detached cross/self plans. The total bounded ranking loss on this probe is 0.1009. Panel (c) measures mean cosine-margin recovery. The separate MTEB distance probe used in Figure 8 and in the baseline distance comparisons consists of 192 triplets: 64 each from SciFact, ArguAna, and NFCorpus. Positives

follow test qrels and negatives are sampled from non-relevant documents. All eight models encode the same texts, with documents truncated to 1,200 characters, maximum sequence length 512, batch size 8, bfloat16, and SDPA attention. Each model reuses its own final RMSNorm at intermediate layers, followed by last-token pooling and L2 normalization. The three TinyBERT variants use their recorded teacher-shared tokenizer; other students retain native tokenizers. The artifact records checkpoint paths, tokenization settings, and a SHA-256 fingerprint of ordered input texts.

B.4 Baseline Implementations

We describe the baseline objectives and their implementation for decoder-only text embeddings. Let $\mathbf{h}^M_{l,i,t}$ be a token state and $\mathbf{r}^M_{l,i}$ its final-RMSNormed last-token counterpart for model $M \in \{\mathrm{S}, \mathrm{T}\}$. Write $\widehat{\mathbf{z}} = \mathbf{z}/\|\mathbf{z}\|_2$. The output projector $P_{\text{out}} : \mathbb{R}^{d_S} \to \mathbb{R}^{d_T}$ is affine; the separate intermediate projector P_{hid} is bias-free and shared across student layers. The teacher is frozen.

B.4.1 Output-Level objectives

Jina's first distillation stage projects the student embedding into teacher space and minimizes cosine distance (Akram et al., 2026). LEAF minimizes the Euclidean norm of the approximation error (Vujanic & Rückstieß, 2026). For B texts, our hosts are

$$\mathcal{L}_{\text{Jina}} = \frac{1}{B}\sum_i \left[1 - \widehat{P_{\text{out}}(\mathbf{r}^S_{L_S,i})}^{\top} \widehat{\mathbf{r}^T_{L_T,i}}\right], \tag{B.2}$$

$$\mathcal{L}_{\text{LEAF}} = \frac{1}{B}\sum_i \left\| \widehat{P_{\text{out}}(\mathbf{r}^S_{L_S,i})} - \widehat{\mathbf{r}^T_{L_T,i}} \right\|_2. \tag{B.3}$$

The LEAF norm is not squared, matching its Equation (3). Both losses apply to queries, positives, and available negatives. We evaluate native 1024-dimensional student embeddings after discarding P_{out}.

B.4.2 TinyBERT

Following the text-embedding adaptation in LEAF Appendix B.2, we replace task-logit supervision with the output host and retain hidden-state and raw-attention matching (Jiao et al., 2020; Vujanic & Rückstieß, 2026). Student layer l maps to $g(l) = \lfloor L_T l/L_S \rfloor$, $l = 1, \ldots, L_S$. For the 28-layer student and 36-layer teacher, this includes the final teacher layer. With padding mask $m_{i,t}$ and $N_{\text{tok}} = \sum_{i,t} m_{i,t}$, the hidden loss is

$$\mathcal{L}_{\text{Hidn}} = \frac{1}{L_S d_T N_{\text{tok}}} \sum_{l,i,t} m_{i,t} \left\| P_{\text{hid}}(\mathbf{h}^S_{l,i,t}) - \mathbf{h}^T_{g(l),i,t} \right\|_2^2. \tag{B.4}$$

For attention, the adaptation captures raw query and key projections before RoPE and computes

$$\mathbf{A}^M_{l,a} = \mathbf{Q}^M_{l,a}(\mathbf{K}^M_{l,a})^{\top}/\sqrt{d_h}, \qquad d_h = 128, \tag{B.5}$$

after grouped-query key expansion. The teacher's 32 query heads are averaged in adjacent pairs to match the student's 16: $\overline{\mathbf{A}}^T_{l,a} = (\mathbf{A}^T_{l,2a-1} + \mathbf{A}^T_{l,2a})/2$. For the non-padding causal query–key positions $\mathcal{V}$, we use

$$\mathcal{L}_{\text{Attn}} = \frac{1}{L_S H_S |\mathcal{V}|} \sum_{l,a} \sum_{(i,t,u)\in\mathcal{V}} \left(A^S_{l,a,i,t,u} - \overline{A}^T_{g(l),a,i,t,u} \right)^2. \tag{B.6}$$

Hidden-only and attention-only use the respective terms; joint TinyBERT uses their sum. All three variants use the teacher tokenizer to give identical token IDs and masks to both models for positionwise comparison.

Baseline degradation. The degradation of intermediate-layer baselines in Table 1 is qualitatively consistent with the text-embedding distillation results of Vujanic & Rückstieß (2026). In LEAF Appendix B, adding MiniLM, TinyBERT, or DistilBERT losses reduces NanoMSMARCO nDCG@10

from 60.7 to 54.9, 53.7, and 55.3, respectively (their Table 4), despite similar final embedding-approximation errors (their Figure 10). These findings show that adding supervision from teacher hidden states or attention patterns can reduce downstream effectiveness, even when output embeddings remain similarly aligned. Our baseline results follow the same qualitative pattern and provide a comparison for assessing whether gains arise from the specific form of intermediate supervision used by LST.

B.4.3 RAIL-KD

RAIL-KD samples a random ordered teacher subset each epoch (Haidar et al., 2022). The layerwise loss is

$$\mathcal{A}_e = \{a_{e,1} < \cdots < a_{e,L_S}\} \subseteq \{1, \ldots, L_T\}, \qquad \mathcal{L}_{\text{RAIL}} = \frac{1}{BL_S d_T} \sum_{i,l} \left\| P_{\text{hid}}(\mathbf{r}^S_{l,i}) - \mathbf{r}^T_{a_{e,l},i} \right\|_2^2 . \tag{B.7}$$

The subset is drawn without replacement and resampled each epoch. Both models retain native tokenizers because the loss compares sentence vectors, not aligned token positions.

B.4.4 BERT-EMD

BERT-EMD introduces many-to-many layer matching via transport (Li et al., 2020). Under the shared embedding setup, the transport loss is

$$C_{lm} = \frac{1}{Bd_T} \sum_i \left\| P_{\text{hid}}(\mathbf{r}^S_{l,i}) - \mathbf{r}^T_{m,i} \right\|_2^2 , \qquad \mathcal{L}_{\text{EMD}} = \min_{F \in \mathcal{U}(\mathbf{a},\mathbf{b})} \langle F, C \rangle, \tag{B.8}$$

with uniform layer marginals and the polytope in Eq. (4). At each step, SciPy HiGHS solves the linear program on detached costs up to numerical tolerance; gradients then propagate through $\langle F, C \rangle$ with fixed flow.

B.4.5 Comparison Setup and Interpretation

Baseline inclusion criterion. The empirical baselines in Table 1 are selected to test the paper's central question of layerwise distillation: they either supervise intermediate representations or define correspondences between teacher and student layers. This criterion includes TinyBERT-style hidden-state and attention matching, RAIL-KD, and BERT-EMD, but does not aim to benchmark every embedding-distillation method that uses relational information or optimal transport.

EMO combines token-level relational transfer with optimal transport for cross-tokenizer embedding distillation (Truong et al., 2025). Its transport addresses correspondence among tokens under tokenizer mismatch, whereas LST learns correspondence across the complete teacher and student layer trajectories and reuses that layer plan to transfer depth-dependent ranking preferences. Because EMO and LST apply transport to different objects and target different alignment problems, EMO is discussed as closely related embedding-distillation work rather than included as a direct empirical baseline in our controlled layerwise comparison.

The teacher, student initialization, corpus, LoRA settings, and optimizer schedule are shared. Non-attention runs use per-device batch 8 with accumulation 1 and complete 23,576 updates. Attention-only and joint TinyBERT use batch 4 with accumulation 2 and complete 23,582 updates because physical batching changes epoch rounding. The global effective record batch is 64. Attention computes four mapped layer pairs per batched QK multiplication and accumulates masked reductions in FP32.

Interpreting trajectory transfer. Throughout this paper, the teacher's layerwise trajectory describes the progression from intermediate relational organization to retrieval discrimination. Its separation component is measured by the query–positive versus query–negative cosine margins in Eq. (6), tracked across depth; the layerwise nDCG@10 curve in Figure 1(c) provides a complementary view of retrieval performance. We assess transfer through agreement with the teacher's native-space geometry and margin curves, alongside downstream embedding quality.

Prior work already transfers relational and intermediate-layer knowledge. Park et al. (2019) transfer distances and angles among embeddings, while Tung & Mori (2019) preserve sample similarities at

selected layer pairs and can supervise multiple intermediate layers. Dasgupta & Cohn (2025) extend hidden-state matching to different model widths using CKA, and BERT-EMD learns cross-layer assignments from hidden-state and attention discrepancies (Li et al., 2020). These methods establish useful forms of representation transfer, but their results do not by themselves establish recovery of the teacher's depth-dependent retrieval separation in a text embedding student.

For the baselines reproduced above, we therefore evaluate both downstream performance and the intermediate trajectory: greater representational similarity alone need not imply recovery of the teacher's relation spectrum, depth correspondence, or preference strength. Figures 3 and 4 assess native geometry and correspondence, while Figure 5(c) and Appendix D assess margin and query–document trajectory recovery. The evidence claimed here is that LST brings these complementary aspects of the student's trajectory closer to the teacher while improving the output-level objectives in Table 1. The novelty statement concerns this empirical recovery in text embedding distillation, beyond the established use of relational losses or intermediate-layer supervision.

C Student Transfer and Objective Comparisons

C.1 Qwen3.5-0.8B Student

Table 5 evaluates transfer to a student initialized from Qwen3.5-0.8B-Base, with the Qwen3-Embedding-4B teacher held fixed. The Output KD control uses the Jina host; LST adds Relational Geometry Encoding, Soft Layer Alignment, and Emergence-Weighted Ranking with $\lambda_{\text{str}} = \lambda_{\text{rank}} = 0.1$, $\varepsilon = 0.02$, $\lambda_d = 1$, and 20 Sinkhorn iterations. Both configurations use one epoch, LoRA rank 16, learning rate 2×10^{-4}, maximum sequence length 512, and global effective record batch size 64. The shared optimizer and evaluation protocols are given in Appendix B.

Table 5: Qwen3.5-0.8B student with the 4B teacher. Avg. averages 15 tasks; deltas use unrounded scores. Green/red marks gains/drops from Output KD; bold marks the best score in each column.

Model / objective	Retrieval			Reranking		Clustering			Pair	Classification			STS			**Avg.**
	SciFact	ArguAna	NFCorpus	SO-Dup	SciDocs	BioRxiv	MedRxiv	20News	SprinDup	Banking77	Emotion	MASSIVE	STS17	SICK-R	STS-B	
Output KD	70.62	**60.31**	34.86	51.21	78.85	34.01	**33.05**	47.08	96.77	81.31	55.61	49.85	81.53	82.97	89.14	63.14
+ LST	**72.28**	60.18	**36.01**	**52.74**	**81.48**	**35.10**	32.95	**47.50**	**96.79**	**81.61**	**57.03**	**51.16**	**82.53**	**83.92**	**89.75**	**64.07**
Δ	+1.66	−0.12	+1.15	+1.53	+2.63	+1.09	−0.10	+0.42	+0.02	+0.30	+1.42	+1.31	+0.99	+0.95	+0.61	+0.92

LST improves 13 of 15 tasks and raises Overall by 0.92 points. The largest gains are SciDocs (+2.63), SciFact (+1.66), SO-Dup (+1.53), and Emotion (+1.42); ArguAna and MedRxiv decrease slightly. All six category means improve, with reranking (+2.08), classification (+1.01), retrieval (+0.90), and STS (+0.85) contributing the larger gains.

Model choice and practical relevance. Our main teacher–student configuration follows *jina-embeddings-v5-text: Compact and Robust Text Embedding Models using Task-Targeted Distillation* (Akram et al., 2026), which distills Qwen3-Embedding-4B into a student initialized from Qwen3-0.6B-Base. This setting follows a recent industrial embedding system reported to be competitive among models of comparable size (Akram et al., 2026). Improving this established distillation setting directly addresses a practical question: how much useful teacher information remains available beyond the final embedding? Prior text-embedding experiments found that adding hidden-state or attention supervision could reduce downstream performance (Vujanic & Rückstieß, 2026). LST shows that layerwise information can improve downstream scores when transferred through native relational geometry, soft layer correspondence, and emergence-weighted ranking.

Jina's full pipeline further specializes the distilled model with four task-specific LoRA adapters for retrieval, semantic similarity, clustering, and classification (Akram et al., 2026). LST adds two auxiliary losses during distillation to improve a single student across task categories. Under the shared protocol in Table 1, it raises Overall by 0.89 points on the Jina host and 0.49 on the LEAF host; the additional student above gains 0.92 points and improves on 13 of 15 tasks. Under this

setup, recovering intermediate teacher information through a unified training objective improves a single student checkpoint across all evaluated tasks, without task-specific adapter selection or added inference-time computation.

C.2 Negative-Aware Training Objectives

Table 6: Full MTEB-15 negative-aware controls. Bold indicates the best result in each column.

Configuration / objective	Retrieval			Reranking		Clustering			Pair	Classification			STS			**Avg.**
	SciFact	ArguAna	NFCorpus	SO-Dup	SciDocs	BioRxiv	MedRxiv	20News	SprintDup	Banking77	Emotion	MASSIVE	STS17	SICK-R	STS-B	
Hard-label control																
InfoNCE	73.68	51.77	32.55	48.37	80.13	33.42	29.76	45.75	96.06	79.09	49.92	52.34	77.40	81.18	83.44	60.99
InfoNCE + Output KD	73.33	54.97	34.62	49.93	80.25	34.29	31.78	47.60	96.74	79.78	52.30	51.20	79.05	82.27	85.96	62.27
Contextual teacher-aware objective																
KaLM-style	**75.59**	58.25	**36.24**	51.18	82.12	34.89	31.71	47.34	97.11	80.41	52.13	**52.95**	80.33	83.31	88.24	63.45
Ours																
LST Modules 1+2	72.67	**61.96**	34.95	51.56	81.87	34.91	33.00	49.01	**97.47**	**81.24**	55.61	50.37	80.09	83.33	89.58	63.84
Final LST	72.88	61.25	36.15	**51.98**	**82.18**	**34.98**	**33.52**	**49.26**	97.01	81.22	**56.65**	50.95	**81.09**	**83.88**	**89.71**	**64.18**

Table 6 compares negative-aware training objectives on MTEB-15. The rows form two matched comparisons plus two contextual references. InfoNCE and InfoNCE + Output KD share the student, candidates, labels, temperature, and budget, so their difference isolates the addition of final-embedding teacher supervision. Relational Geometry Encoding + Soft Layer Alignment and its + Rank-KL variant share the Jina host and $\varepsilon = 0.05$ structural recipe, so their difference isolates the transported ranking target within that recipe. KaLM-style changes both target and loss weights, while Final LST uses the selected $\varepsilon = 0.02$ configuration; these two rows provide context only.

Adding Output KD to InfoNCE improves 13 of 15 tasks and raises Overall by 1.28. The largest gains are ArguAna (+3.21), STS-B (+2.52), and Emotion (+2.38); MASSIVE (-1.14) and SciFact (-0.35) decrease. Adding Rank-KL to Relational Geometry Encoding + Soft Layer Alignment improves 10 tasks and raises Overall by 0.34, led by NFCorpus (+1.17), STS17 (+1.03), and 20News (+0.94). Final LST achieves higher Overall than the tested hard-label objectives.

D Trajectory Diagnostics

This section supplements the analysis of Relational Geometry Encoding, Soft Layer Alignment, and Emergence-Weighted Ranking with cross-family teacher trajectories, meaning-preserving inputs, query–document distances, and the baseline curves omitted from Figure 3. All student comparisons use the Jina host and fixed probes; statistics average the complete depth range unless a narrower interval is stated. Appendix B.3 specifies the probes and depth conventions.

D.1 Teacher Trajectories Across Model Families

As a supplementary check, we probe `microsoft/harrier-oss-v1-27b`, a 27B-parameter decoder-only embedding model from a different model family. Per MTEB model metadata, it is adapted from `google/gemma-3-27b-pt`, and differs from the Qwen3-Embedding teachers in base architecture, training data, and contrastive objective. Figure 6 applies the same three trajectory probes to Harrier-27B alongside Qwen3-Embedding-4B and Qwen3-Embedding-8B.

Harrier-27B follows the same qualitative pattern as the Qwen3-based teachers across all three subfigures: synonymous-query distance peaks at intermediate layers (Fig. 6 a), query–positive and query–negative distances diverge over the same depth range (Fig. 6 b), and full-corpus retrieval performance remains low until the final third of the network (Fig. 6 c). These observations suggest that the layerwise trajectory is not unique to Qwen3, although evidence from a single additional model is insufficient to establish broad generality across architectures or training regimes.

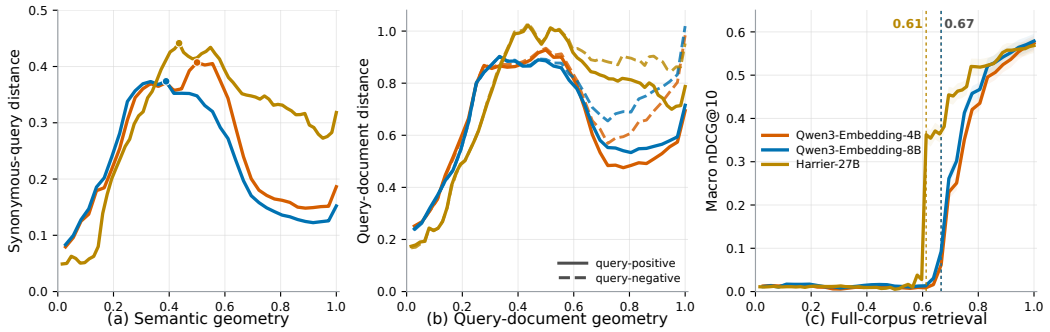


Figure 6: Teacher trajectories for Qwen3-Embedding-4B, Qwen3-Embedding-8B, and Harrier-27B. (a) Synonymous-query distance. (b) Query–document distance: solid positive, dashed negative. (c) Full-corpus retrieval nDCG@10. Horizontal axes show normalized depth.

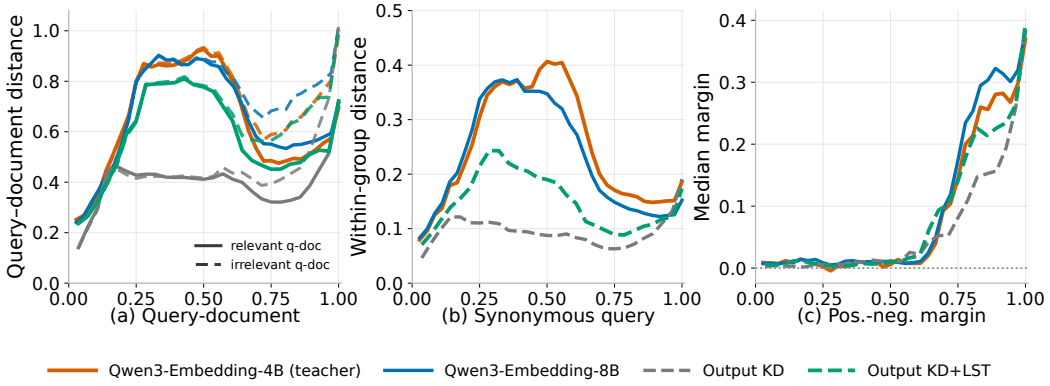


Figure 7: Supplementary geometry and ordering probes. Green denotes the final LST student. (a) Synonymous-query distance. (b) Query–document cosine similarity, with dotted negatives and solid teacher/dashed student positives. (c) Pair-order accuracy on MTEB triplets. (d) Median angular-distance margin. Horizontal axes show layer depth.

D.2 Semantic Geometry and Retrieval Separation

Figure 7 adds the final LST student to the probes in Figure 1. LST recovers part of the teacher's middle-layer expansion, but its peak remains smaller and earlier. This diagnostic measures sensitivity to meaning-preserving wording. The query–document, pair-order accuracy, and median-margin panels distinguish geometry from ordering and preference strength. On real MTEB triplets, final pair-order accuracy is similar across models, but the margin curves differ. Thus similar final pair-order accuracy can coexist with a weaker or delayed separation of relevant and irrelevant documents.

Figure 8 examines distances on 192 held-out MTEB triplets, 64 each from SciFact, ArguAna, and NFCorpus. LST reduces the full-depth query–positive and query–negative distance MAEs from 0.321 and 0.337 to 0.083 and 0.082. Each MAE compares the student's mean distance curve with the teacher curve linearly interpolated to the same depths.

D.3 Structural Similarity Across Depth

Figure 9 asks whether any student layer matches each teacher layer, whether that match occurs at the same depth, and how the relation spectrum changes. It covers all six intermediate-layer objectives together with Output KD and the teacher reference on the shared probe. LST has mean best-over-layer CKA 0.966. Panel (c) measures the excess cost of same-depth matching, $D_{\text{same}}(m) - \min_l D(l, m)$. Across all teacher layers, the ratio between mean same-depth and mean best-layer cost is 7.96 for EMD and 1.19 for LST. The ratio characterizes depth mismatch relative to each model's best available

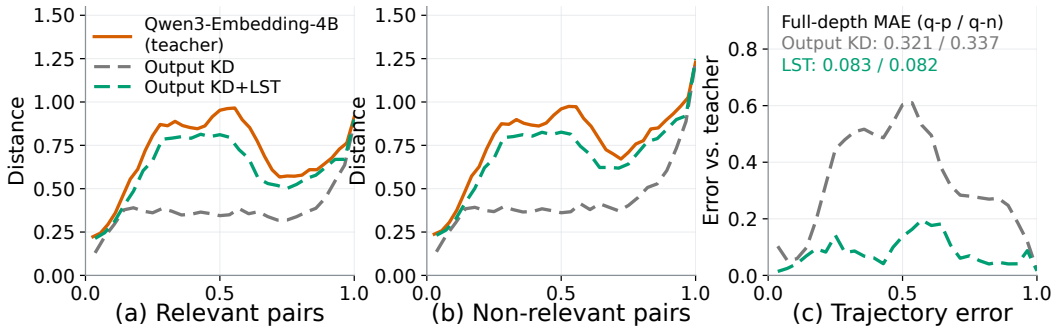


Figure 8: Query–document geometry on held-out MTEB triplets. (a) Relevant pairs. (b) Non-relevant pairs. (c) Mean absolute trajectory error. Horizontal axes show normalized depth j/L; (c) averages errors in (a–b) relative to the interpolated teacher curves.

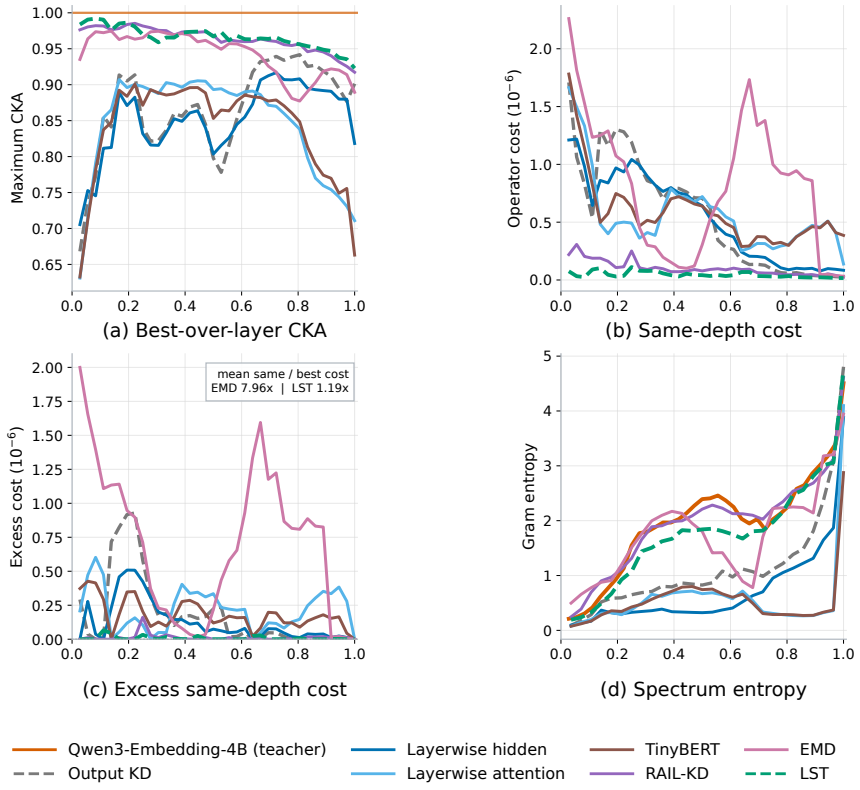


Figure 9: Native-space diagnostics across objectives. (a) Best-over-layer CKA. (b) Same-depth operator cost. (c) Excess cost, $D_{\text{same}}(m) - \min_l D(l, m)$. (d) Gram entropy. Costs are raw operator distances in units of 10^{-6}, without teacher-scale normalization. Horizontal axes show layer depth.

layer. Panel (d) is a separate spectral check: over depths 0.25–0.75, teacher-relative Gram-entropy MAE decreases from 1.167 for Output KD to 0.361 for LST. RAIL-KD has an even smaller entropy MAE of 0.123, despite its low effective rank. Gram entropy uses the cosine Gram spectrum and is not the logarithm of the centered-feature effective rank in Figure 3.

D.4 Baseline Spectra and Alignment

Figures 10–13 extend Figure 3 to TinyBERT, attention-only, RAIL-KD, and EMD supervision without repeating the layerwise hidden comparison. They retain the same teacher, Output KD, and

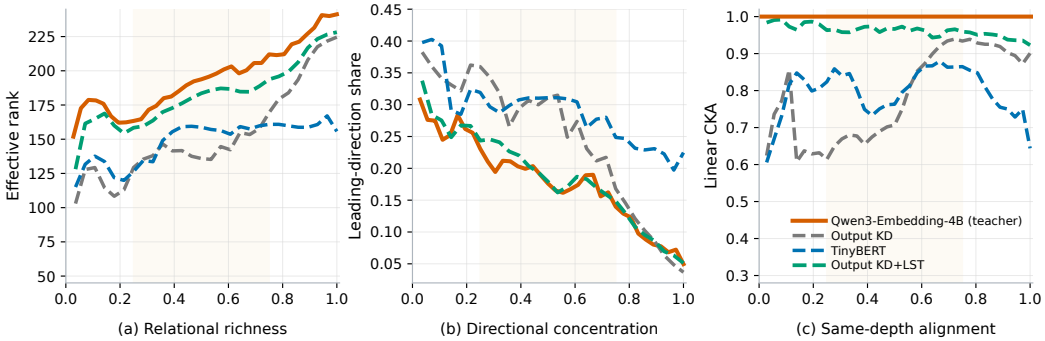


Figure 10: TinyBERT matching: effective rank, leading-direction share, and same-depth CKA. Depth is j/L; shading marks 0.25–0.60. All diagnostics use native representations.

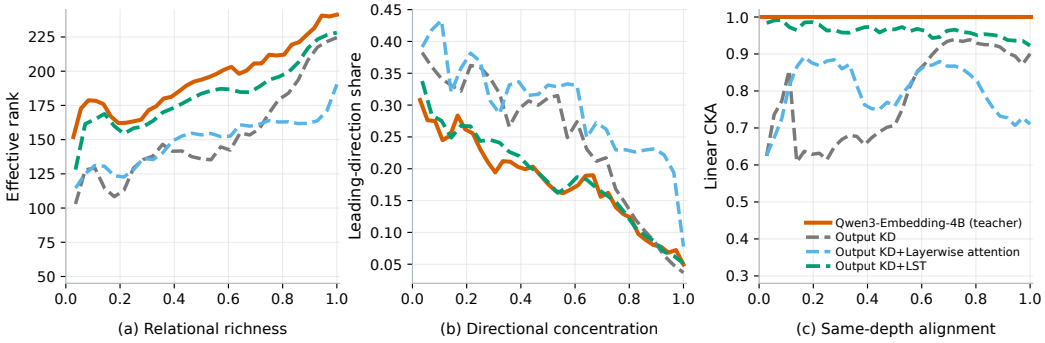


Figure 11: Layerwise attention matching; settings match Figure 10.

LST references and shared axis limits. The three columns measure effective rank, leading-direction share, and same-depth CKA in each model's native representation space. These quantities are complementary: high CKA can coexist with a concentrated relation spectrum.

TinyBERT. Figure 10 yields a full-depth mean effective rank of 148.7 and leading-direction share of 0.290, compared with Output KD's 153.2 and 0.245. Its mean same-depth CKA increases only slightly from 0.787 to 0.795, while its query–positive distance MAE remains 0.431 on the MTEB probe, compared with 0.083 for LST. Thus joint hidden-state and attention matching improves none of the three diagnostics toward the teacher: effective rank stays below Output KD, directional concentration rises, and the query–positive distance MAE remains 0.431 versus 0.083 for LST.

Attention matching. In Figure 11, mean CKA reaches 0.804, but effective rank is 149.2 and leading-direction share rises to 0.297. Compared with Output KD, stronger agreement in dominant directions coexists with a more concentrated relation spectrum. The query–positive distance MAE remains 0.426; joint TinyBERT gives 0.431 on the same probe. Appendix B.4 specifies the baseline implementations.

RAIL-KD. Figure 12 separates plausible correspondence from complete structure recovery. Its full-depth mean CKA of 0.958 is close to LST's 0.962, while effective rank remains 117.8 versus 182.5. The rank deficit persists through much of the network, even where the dominant sample relations agree with the teacher. The coverage comparison in Table 2 provides the corresponding training-map context.

EMD. Figure 13 has mean effective rank 112.9 and CKA 0.659. Its strong interior CKA peak does not extend consistently across depth. Together with the excess same-depth cost in Figure 9, this

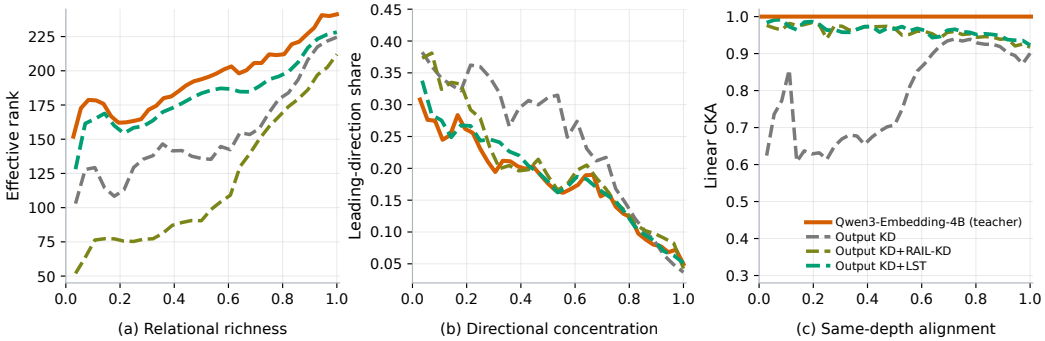


Figure 12: RAIL-KD native-space diagnostics. High same-depth CKA coexists with low effective rank; axes and references match Figure 10.

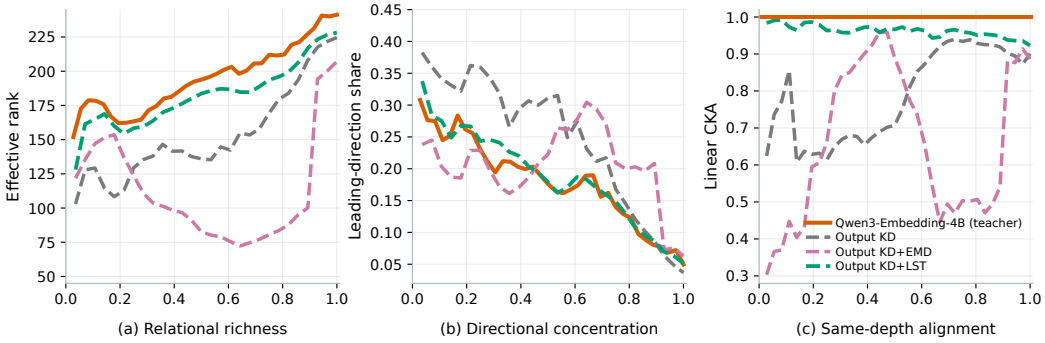


Figure 13: EMD native-space diagnostics. The displaced CKA trajectory accompanies low rank; axes and references match Figure 10.

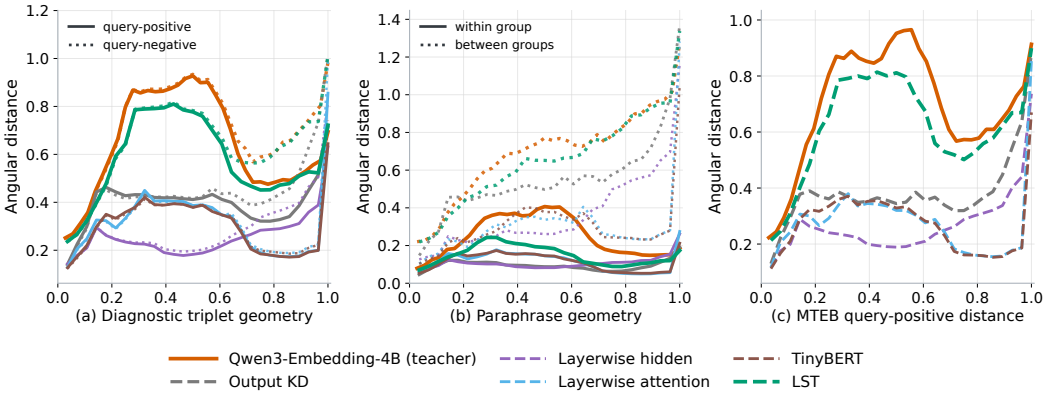


Figure 14: Results on four geometry probes: (a) diagnostic triplets, (b) within/between paraphrase groups, (c) MTEB query–positive distance, and (d) MTEB negative/positive distance ratio. Dotted curves in (a–b) denote negative or between-group pairs; other curves are solid for the teacher and dashed for students. Horizontal axes show layer depth.

indicates that many-to-many projected-hidden-state matching does not by itself recover a depth-local native trajectory. The full-depth rank values are summarized alongside teacher coverage in Table 2.

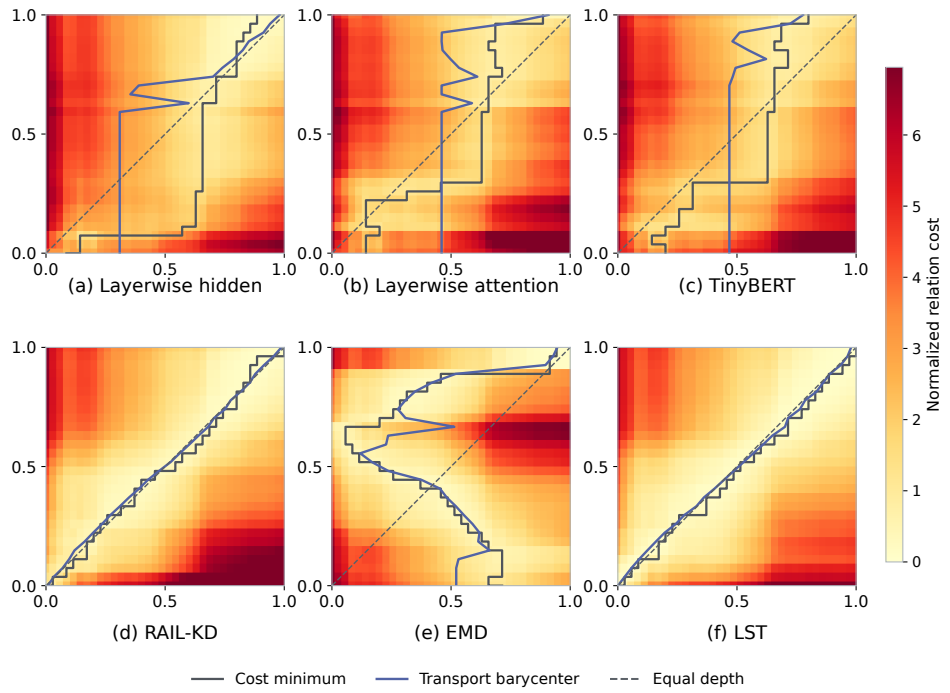


Figure 15: Relation costs and a common post-hoc OT probe. Heatmaps exclude the depth penalty; magenta traces row-wise cost minima, orange the $\varepsilon = 0.02$ transport barycenter, and gray equal depth. All panels share one cost scale.

D.5 Distance Geometry Across Baselines

Figure 14 shows layerwise distance and separation patterns across four geometry probes, with consistent model colors throughout. Every model shown appears in all four panels; the MTEB panels use the same 192 triplets and each checkpoint's training-matched tokenizer. LST follows the teacher's middle-layer query–positive expansion more closely than Output KD and captures the late-depth increase in the MTEB negative/positive distance ratio. Its final diagnostic angular-distance margin is 0.2890, compared with 0.2896 for the teacher.

D.6 Layer Correspondence from Relation Costs

Figure 15 supplements Figure 4 with relation costs and a separate common OT probe for all six intermediate objectives. Its heatmap cost excludes the depth penalty, so a near-diagonal low-cost region is not inserted by that penalty during visualization. EMD's displaced minima contrast with the more diagonal minima of RAIL-KD and LST. RAIL-KD's ordered random teacher subset already encourages depth order: its recorded layer-pair mean depth deviation is 0.0187, and every pair is within 0.10. Thus local correspondence is compatible with its low rank in Figure 12.

Magenta cost minima and orange transport barycenters need not agree because the common probe also uses depth regularization, target marginals, entropy, kernel flooring, and finite numerical iterations. On the fixed LST cost matrix, removing the depth term from this probe gives 99.93% near-diagonal mass, whereas depth cost alone gives 71.20%. The relation costs therefore contain local structure. The diagnostic plans are not the baselines' training assignments.

E Component and Sensitivity Studies

The following studies examine the relation operator, layer mapping, and auxiliary target for Relational Geometry Encoding, Soft Layer Alignment, and Emergence-Weighted Ranking, respectively. Opera-

tor and target studies use $\varepsilon = 0.05$; the mapping comparison uses the final Jina setting, $\varepsilon = 0.02$. All deltas are computed before rounding.

Development and evaluation. The MTEB-15 test tasks were not repeatedly used to guide method development. Relational Geometry Encoding, Soft Layer Alignment, and Emergence-Weighted Ranking address the representation, depth-correspondence, and ranking-transfer requirements identified in Section 2; the comparisons below evaluate their effects on embedding quality. We explored the entropy parameter ε and adopted 0.02 for the final Jina configuration. The full sweep in Appendix E.3 shows strong performance across a range of soft-alignment widths.

E.1 Relation Operators

Table 7 compares relation operators in a structure-only study that adds $0.1\mathcal{L}_{\text{str}}$ to the Jina host, without Emergence-Weighted Ranking. The reference averages last-token and masked-mean relation operators, with centering, trace normalization, self-debiasing, depth weight 1, and 20 Sinkhorn iterations. Last-token-only and mean-only change the pooling view used for intermediate supervision, not the final embedding pooling. The remaining variants respectively omit double-centering, the two self-cost corrections, or the depth penalty while retaining the other reference choices. The host-only study control is an independent rerun (63.44), not the 63.29 checkpoint in Table 1.

Table 7: Full MTEB-15 structure-only operator study at $\varepsilon = 0.05$. Color indicates task-wise change from Last + mean; bold marks the best result in this table. Output KD is the study-specific host rerun.

Configuration / objective	Retrieval			Reranking		Clustering			Pair	Classification			STS			Avg.
	SciFact	ArguAna	NFCorpus	SO-Dup	SciDocs	BioRxiv	MedRxiv	20News	SprintDup	Banking77	Emotion	MASSIVE	STS17	SICK-R	STS-B	
Last + mean (reference)	72.48	60.91	35.44	**52.14**	81.83	34.10	33.08	49.74	97.36	80.81	**56.73**	49.46	80.29	83.20	89.57	63.81
Last token only	**72.72**	61.77	35.41	52.07	81.62	34.63	33.28	49.35	**97.54**	**81.32**	55.68	**50.21**	80.25	83.34	89.39	**63.91**
Mean only	72.01	60.64	**35.71**	51.72	81.20	34.05	33.32	47.71	97.19	80.12	56.57	49.28	80.10	83.10	89.28	63.47
No centering	72.45	**62.73**	35.14	51.63	**82.37**	34.36	**33.40**	46.99	97.42	80.42	55.66	49.47	**80.62**	**83.47**	**89.69**	63.72
No self-debiasing	72.31	61.13	35.60	51.98	81.85	**34.89**	32.75	**49.78**	97.44	81.06	55.40	49.69	80.45	83.18	89.45	63.80
No depth prior	72.27	61.05	35.66	51.99	82.26	34.25	32.87	49.60	97.15	80.94	55.55	49.63	80.19	83.00	89.28	63.71
Output KD (study control)	71.79	60.02	35.58	51.80	81.07	34.29	32.71	48.93	97.34	80.94	54.41	49.62	80.50	83.25	89.41	63.44

Last-token supervision exceeds mean pooling by 0.44 Overall points and improves 8 of 15 tasks over the dual-view reference. Its largest gains are on ArguAna (+0.86), MASSIVE (+0.75), and BioRxiv (+0.53), while Emotion decreases by 1.06. Mean-only pooling improves only NFCorpus and MedRxiv and loses 0.34 Overall, supporting the last-token choice in Relational Geometry Encoding. Removing centering or depth regularization changes Overall by -0.09 or -0.10; their largest drops are on 20News (-2.76) and Emotion (-1.19), respectively. The self-debiasing difference is 0.01 Overall, with mixed task-level changes.

E.2 Layer Assignment

Relational Geometry Encoding and Soft Layer Alignment jointly define the structural loss: Soft Layer Alignment computes its matching cost from the relation operators produced by Relational Geometry Encoding, and its plans are also used to construct the transported ranking targets in Emergence-Weighted Ranking. Consequently, Soft Layer Alignment has no standalone loss that can be removed while leaving the other components unchanged. Setting $\lambda_{\text{str}} = 0$ removes structural supervision, but Emergence-Weighted Ranking still requires the alignment plans; removing the alignment computation itself requires specifying a replacement correspondence. We therefore evaluate Soft Layer Alignment by replacing soft alignment with the fixed and row-hard mappings below.

Table 8 extends Table 3 with representation-independent and structure-only assignments. All rows retain the Jina host, last-token relation operators, ranking target, and nominal training budget. The soft reference uses uniform layer marginals, $\varepsilon = 0.02$, $\lambda_d = 1$, and 20 iterations. For student layer l, the hard rules select one teacher layer as follows:

Table 8: Full MTEB-15 layer-mapping study. Color indicates task-wise change from soft transport; bold marks the best result in this table. Hard rules do not enforce teacher-column coverage.

Configuration / objective	Retrieval			Reranking		Clustering			Pair	Classification			STS			Avg.
	SciFact	ArguAna	NFCorpus	SO-Dup	SciDocs	BioRxiv	MedRxiv	20News	SprintDup	Banking77	Emotion	MASSIVE	STS17	SICK-R	STS-B	
Soft transport ($\varepsilon = 0.02$)	72.88	61.25	**36.15**	51.98	82.18	**34.98**	**33.52**	**49.26**	97.01	**81.22**	56.65	**50.95**	81.09	83.88	89.71	**64.18**
Row-hard structure + depth	72.11	61.80	34.98	52.19	**82.77**	34.75	33.02	49.15	97.32	79.64	56.92	48.57	81.12	**83.90**	89.48	63.85
Fixed nearest depth	73.09	62.04	35.65	**52.34**	82.47	34.61	33.16	47.86	97.37	80.30	56.40	49.79	80.95	83.75	**89.82**	63.97
Fixed floor map	**73.20**	**62.34**	35.41	52.29	82.29	34.74	32.90	47.45	97.13	80.53	**57.15**	49.74	**81.18**	83.69	89.61	63.98
Row-hard structure only	72.98	60.39	34.75	51.62	81.31	33.74	32.28	46.16	**97.55**	79.18	55.58	47.83	80.13	82.96	89.46	63.06

Table 9: Full MTEB-15 entropy sensitivity. Color indicates task-wise change from Row-hard; bold marks the best result in this table.

Configuration / objective	Retrieval			Reranking		Clustering			Pair	Classification			STS			Avg.
	SciFact	ArguAna	NFCorpus	SO-Dup	SciDocs	BioRxiv	MedRxiv	20News	SprintDup	Banking77	Emotion	MASSIVE	STS17	SICK-R	STS-B	
Row-hard	72.11	61.80	34.98	52.19	**82.77**	34.75	33.02	49.15	**97.32**	79.64	**56.92**	48.57	**81.12**	83.90	89.48	63.85
$\varepsilon = 0.01$	72.61	**62.31**	35.89	**52.47**	82.35	33.84	32.50	48.64	97.26	80.90	56.01	50.03	80.93	83.79	**89.83**	63.96
$\varepsilon = 0.02$	72.88	61.25	36.15	51.98	82.18	**34.98**	**33.52**	**49.26**	97.01	81.22	56.65	50.95	81.09	83.88	89.71	**64.18**
$\varepsilon = 0.05$	72.85	61.91	**36.52**	52.20	82.28	34.77	32.55	48.70	97.21	**81.41**	56.59	50.79	81.11	**83.93**	89.59	64.16
$\varepsilon = 0.10$	**73.22**	61.71	36.18	52.34	82.62	34.65	32.66	48.66	97.04	81.28	56.45	50.93	81.07	83.84	89.54	64.14
$\varepsilon = 0.20$	72.60	61.32	36.19	52.08	82.59	34.63	33.17	48.24	**97.32**	81.27	56.46	**51.00**	80.93	83.85	89.39	64.07

- **Structure plus depth:** minimize $D_{lm} + \lambda_d(\delta_l^S - \delta_m^T)^2$ independently in each row, where D_{lm} is the structural part of Eq. (3).
- **Nearest depth:** minimize $|\delta_l^S - \delta_m^T|$, independently of the representations.
- **Floor map:** use $m(l) = \lfloor L_T l / L_S \rfloor$, clipped to the valid one-indexed range. This borrows only the map, not TinyBERT's objective.
- **Structure only:** minimize D_{lm} without a depth penalty.

Each rule is also applied to the self-plans. Hard plans assign mass $1/L_S$ per selected pair and permit repeated teacher choices. Their structural readout uses detached plans and uniformly averaged transported operator mismatch. The soft branch instead differentiates through its finite Sinkhorn computation and includes the depth cost.

Soft alignment exceeds the fixed rules by 0.20–0.21 points. Compared with soft transport, row-hard structure plus depth wins 7 tasks but loses 0.33 Overall: its gains on SciDocs (+0.59) and ArguAna (+0.55) are outweighed by losses on MASSIVE (-2.37), Banking77 (-1.58), and NFCorpus (-1.17). Fixed nearest and floor mappings likewise win only 6 and 7 tasks, with their largest deficits on 20News (-1.40 and -1.81). The structure-only hard rule scores 63.06 and improves only SciFact and SprintDup over soft transport; its largest losses are MASSIVE (-3.11) and 20News (-3.10).

E.3 Entropy Regularization

Table 3 combines two measurements: each downstream row is a separately trained model, whereas effective teacher support is recomputed on the fixed final-LST cost matrix. Table 9 reports task-wise sensitivity. The positive-entropy runs retain the same last-token operators, ranking target, loss weights, and training budget. The row-hard result provides a reference with discrete layer assignments.

Across all five tested settings, $\varepsilon \in \{0.01, 0.02, 0.05, 0.10, 0.20\}$, Overall ranges from 63.96 to 64.18, while effective teacher support on the fixed-cost probe spans 2.19–8.66 layers (Table 3). Every setting exceeds the row-hard reference by 0.11–0.33 points and the Jina Output KD score in Table 1 by 0.67–0.89 points. Within $\varepsilon = 0.02$–0.20, Overall varies by only 0.11 points and STS by 0.18. In particular, $\varepsilon = 0.05$ and 0.10 reach 64.16 and 64.14 Overall, within 0.02 and 0.04 points of the final $\varepsilon = 0.02$ configuration (64.18). Thus, strong results persist across the tested range of effective layer support, with limited sensitivity to the precise entropy setting.

Table 10: Full MTEB-15 auxiliary-target comparison. Color indicates task-wise change from LST Modules 1+2 ; bold marks the best result in this table. Figure 16 summarizes the Overall and retrieval changes.

Configuration / objective	Retrieval			Reranking		Clustering			Pair	Classification			STS			Avg.
	SciFact	ArguAna	NFCorpus	SO-Dup	SciDocs	BioRxiv	MedRxiv	20News	SprintDup	Banking77	Emotion	MASSIVE	STS17	SICK-R	STS-B	
LST Modules 1+2	72.67	61.96	34.95	51.56	81.87	**34.91**	33.00	49.01	97.47	81.24	55.61	50.37	80.09	83.33	89.58	63.84
Uniform extra structure	73.28	61.49	35.56	51.24	81.65	34.90	32.85	49.78	97.44	81.22	55.99	50.49	80.47	83.38	89.29	63.93
Emergence-focused structure	73.01	61.61	35.62	51.62	81.86	34.90	**33.32**	49.20	97.35	81.00	**56.27**	50.26	80.77	83.51	89.27	63.97
Trajectory derivative	72.87	61.74	34.85	**52.16**	**82.39**	34.57	32.78	48.27	**97.60**	81.02	55.27	49.45	80.24	82.96	89.47	63.71
Soft stage boundary	72.51	62.01	35.63	51.78	82.08	33.90	32.61	49.12	97.39	80.94	55.36	50.28	80.16	82.95	89.25	63.73
Curriculum focus	72.72	61.91	35.74	51.75	81.95	34.46	32.74	**50.05**	96.97	81.29	55.67	**50.75**	80.58	83.40	89.33	63.95
Transported Rank-KL	73.02	61.78	**36.12**	52.11	82.27	34.51	33.00	49.95	97.22	**81.52**	56.11	50.53	**81.12**	**84.06**	89.40	**64.18**
Combined four objectives	**73.42**	**62.48**	35.65	52.14	82.29	34.66	32.90	47.28	97.53	80.61	55.53	49.77	80.85	83.32	**89.64**	63.87

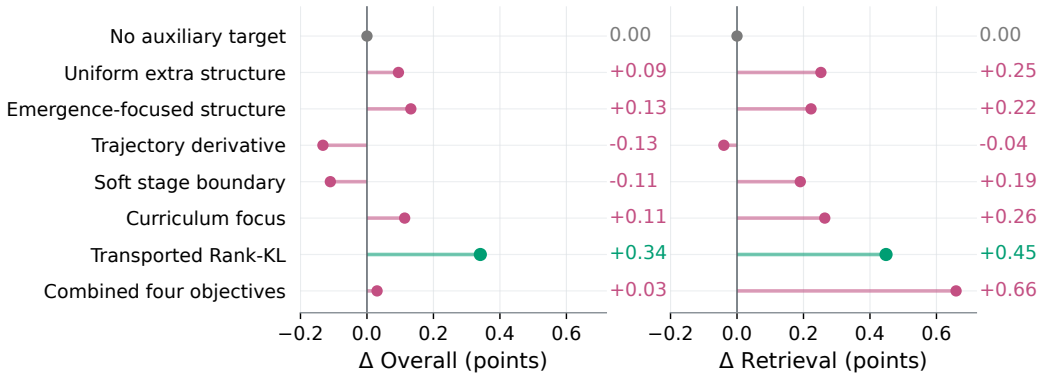


Figure 16: Auxiliary-target comparison on a shared Jina structural baseline ($\varepsilon = 0.05$). Overall (left) and retrieval (right) changes are relative to no auxiliary target (63.84 and 56.52, respectively). Green highlights transported Rank-KL.

Task-specific optima vary: $\varepsilon = 0.05$ reaches 57.10 retrieval versus 56.76 at 0.02, and 0.10 reaches 67.48 reranking versus 67.08. The final setting benefits clustering (39.26 versus 38.67 at 0.05). No single setting dominates the task columns: $\varepsilon = 0.02$ is best on only 3 of 15 tasks, while the largest sweep ranges occur on MASSIVE (2.43), Banking77 (1.77), and NFCorpus (1.55).

E.4 Auxiliary Ranking Targets

Figure 16 and Table 10 test whether explicit preferences add information beyond structural matching. Every row uses the Jina host, last-token centered operators, self-debiasing, structural weight 0.1, and $\varepsilon = 0.05$. Each nonempty auxiliary branch has weight 0.1; "None" sets it to zero. The figure shows changes from the no-auxiliary control, while the table reports absolute scores.

The "None" and "Transported Rank-KL" rows provide a direct on/off ablation of Emergence-Weighted Ranking, with the output host, Relational Geometry Encoding, and Soft Layer Alignment held fixed. They differ only in whether the ranking branch is enabled: $\lambda_{\text{rank}} = 0$ versus 0.1. Adding Emergence-Weighted Ranking raises Overall from 63.84 to 64.18 (+0.34) and improves 10 of 15 tasks. The definitions below distinguish structural reweighting from preference matching.

Structural controls. Uniform extra structure adds a transported, self-debiased operator mismatch with detached plans and equal layer weights. Emergence-focused structure uses the same mismatch with weights derived from the positive depth derivative of the smoothed teacher mean margin, a floor of 0.25, and interpolation to student depth. The extra detached-plan objective is not algebraically identical to doubling the base structural loss.

Trajectory controls. Trajectory derivative matches adjacent-layer differences of smoothed query–positive cosine-distance and mean-margin curves using teacher-scale-normalized smooth-L1 losses. Soft stage boundary matches differentiable transition depths and pre/post-transition membership, with temperatures 0.10 and 0.08, respectively. Curriculum focus shifts structural emphasis from a Gaussian band around the teacher's query–positive distance maximum (width 0.15) to the margin-opening band over the first 80% of training; it does not change data order.

Preference target. Transported Rank-KL is the objective used by Emergence-Weighted Ranking in Section 2.3, with temperature 0.05 and emergence floor 0.25. Combined four objectives averages curriculum focus, trajectory derivative, soft stage boundary, and Rank-KL before applying the outer coefficient 0.1; it excludes the two structural controls.

Table 10 reports the task-level changes summarized in Figure 16. Rank-KL gives the largest Overall gain among the tested auxiliary targets. Its largest gains are NFCorpus (+1.17), STS17 (+1.03), and 20News (+0.94); the largest decreases are BioRxiv (-0.41), SprintDup (-0.24), and ArguAna (-0.18). Uniform and emergence-focused extra structure give only +0.09 and +0.13 Overall. The combined target has the highest retrieval mean but improves only 8 tasks and loses 1.73 on 20News, leaving it 0.31 Overall below Rank-KL.

The auxiliary-target study and final checkpoints both score 64.18 Overall but differ in retrieval (56.97 versus 56.76); only the latter appears in Table 1.

F LIMITATIONS

Hyperparameter exploration. We did not conduct a systematic search over all hyperparameters: we directly set $\lambda_{\text{str}} = \lambda_{\text{rank}} = 0.1$ and the emergence floor to $\alpha = 0.25$. Our hyperparameter sweep focused on the entropy regularization parameter ε, primarily to examine the effect of Soft Layer Alignment and how different values of ε translate into different effective numbers of teacher layers per student layer. The results support the stability and effectiveness of LST under the tested settings, but do not establish that the fixed hyperparameters are optimal. Although the outer loss coefficients are fixed, the loss magnitudes and relative contributions across layers adapt to the representations and learned layer correspondence; Figure 5 illustrates the resulting layerwise ranking-loss distribution.

Model and architecture coverage. Our main experiments follow the practical distillation configuration of *jina-embeddings-v5-text* (Akram et al., 2026), using Qwen3-Embedding-4B as the teacher and Qwen3-0.6B-Base as the student initialization. This configuration provides a strong embedding teacher and a substantial compression ratio, making it sufficient for testing whether layerwise structure contributes beyond output-level distillation; Appendix C further verifies the method with a different student initialization. Smaller or weaker teachers would reduce the capability gap that makes distillation informative, while training with larger teachers exceeded the available GPU memory.