STRATADISTILL: STRUCTURE-AWARE DISTILLATION FOR LONG-DOCUMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-document retrieval often depends on evidence confined to a small part of a document. Single-vector embedding models encode documents without the query, making local relevance difficult to preserve amid unrelated content. We introduce StrataDistill, a distillation method that uses document structure to guide supervised fine-tuning (SFT) of embedding models. Frozen teachers combine full-document similarity with the strongest local match to construct relevance distributions over candidate documents. A shared student learns these distributions through full-document scoring alongside supervised contrastive learning. Using source-specific teachers consolidates supervision from separately adapted models in the same retriever. The method reuses existing retrieval data and document structure without query-to-section annotations, and also supports short documents without decomposition. Across five long-document datasets and four embedding backbones, StrataDistill with source-specific teachers improves macro-average nDCG@10 by 2.85 to 4.66 points over mixed-task SFT. Further analysis shows a smaller gap between full-document and local-view scoring, consistent with partial transfer of local relevance into document representations. Inference preserves the original index size and scoring procedure, storing one vector per document.

1 INTRODUCTION

Retrieval supports document-grounded assistance and long-form generation (Feng et al., 2021; Gu et al., 2025). These applications draw on research papers and technical documentation, making effective long-document retrieval important in practice (Dasigi et al., 2021; Cohen et al., 2025). Embedding models enable efficient search with precomputed document vectors (Karpukhin et al., 2020; Xiong et al., 2021). They perform well on passage retrieval benchmarks (Wang et al., 2022b; 2024; Lee et al., 2025), yet long-document retrieval remains challenging (Zhu et al., 2024; Saad-Falcon et al., 2024). Figure 1(a) illustrates this contrast between short- and long-document retrieval.

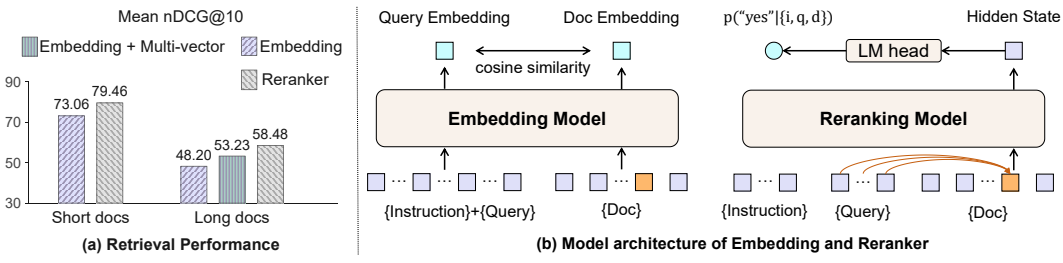


Figure 1: Motivation for structure-aware distillation. (a) The evaluated models perform better on short documents than on long ones. Local-view multi-vector scoring improves long-document embedding retrieval and narrows the gap to reranking. (b) Rerankers use query-conditioned attention over document content, whereas embedding models encode queries and documents independently.

In a long document, relevant evidence may occupy only one section amid unrelated content. Figure 1(b) contrasts how reranking and embedding models match this evidence. Rerankers jointly process queries and documents, allowing attention to identify relevant content (Ren et al., 2021;

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.

Zhang et al., 2025b). Embedding models use query-independent document encoding, so document attention cannot adapt to the query. The document vector must serve future queries about different sections (Luan et al., 2021). Longer contexts admit more text but still encode documents independently of the query (Nussbaum et al., 2025; Zhu et al., 2024).

Multi-vector retrieval partly compensates by retaining local representations for query-time matching, as in chunk-based retrieval and late interaction (Chen et al., 2024b; Günther et al., 2024; Khattab & Zaharia, 2020; Santhanam et al., 2022). The local-view gains in Figure 1(a) suggest that local representations expose relevance signals missed by full-document scoring. Our goal is to use these signals during training while retaining single-vector retrieval at inference.

Document-level labels identify relevant documents but do not specify where supporting evidence occurs. Our analysis shows that local-view scoring retains an advantage after supervised fine-tuning (SFT). Document structure defines local views through section boundaries, making their query-relevance scores available as additional supervision. Distillation can use these scores to train a single-vector retriever (Lin et al., 2020; Ren et al., 2021; Zhou et al., 2024).

We introduce **StrataDistill**, a structure-aware distillation method for supervised fine-tuning of embedding models. A frozen embedding teacher combines full-document similarity with the strongest match among local views defined by document structure. The teacher converts these scores into a relevance distribution over candidate documents. The student learns this distribution through full-document scoring, alongside the usual supervised contrastive objective.

When training spans multiple data sources, StrataDistill can use a teacher adapted to each source. A shared student consolidates their relevance distributions into one retriever. The teachers can use the same backbone as the student. Distillation reuses existing query-document pairs and structural boundaries without query-to-section annotations. Local-view computation stays in the offline training pipeline. Inference uses one vector per document and standard similarity search. For short documents without decomposition, the same objective applies using full-document teacher scores.

Our contributions are as follows:

- **Structure-derived relevance targets.** We combine full-document similarity with the strongest local match to construct relevance targets for document-level retrieval.
- **Integration with embedding fine-tuning.** Shared or source-specific teachers supervise one retriever without changing its architecture, index size, or online scoring.
- **Cross-backbone evaluation.** Across five retrieval datasets and four backbones, StrataDistill improves on shared-model SFT, with additional gains from source-specific teachers.

2 Related Work

Long-document retrieval. Long-context embedding models extend the encoder input length (Nussbaum et al., 2025; Zhu et al., 2024). Longtriever aggregates interacting document blocks into one representation (Yang et al., 2023). Late chunking constructs context-aware chunk embeddings (Günther et al., 2024). Landmark embeddings and the ColBERT family retain finer-grained representations for query matching (Luo et al., 2024; Khattab & Zaharia, 2020; Santhanam et al., 2022). LongRAG approximates relevance to a long retrieval unit using the maximum similarity over its constituent chunks (Jiang et al., 2024). StrataDistill uses local matches as training supervision while preserving the backbone architecture, index size, and single-vector scoring.

Document structure in retrieval. RAPTOR and DISRetrieval organize content into hierarchical retrieval nodes (Sarthi et al., 2024; Chen et al., 2026). For representation learning, SAILER exploits sections of legal documents (Li et al., 2023a), and CoSAEmb learns section-aware representations of scientific papers (Singh & Singh, 2024). SEAL trains query-document matching with tagged, untagged, and partially tag-masked document variants (Huang et al., 2025). StrataDistill derives query-conditioned relevance targets from existing structure and document-level training pairs, without requiring query-to-section annotations.

Distillation for dense retrieval. Cross-encoder distillation transfers relevance scores to efficient retrieval models (Hofstätter et al., 2020; Wang et al., 2022a). RocketQAv2 uses listwise distillation

to jointly train a retriever and reranker (Ren et al., 2021). Lin et al. (2020) distill ColBERT MaxSim scores into dot-product scoring for single-vector retrieval. FGD aligns multi-granular student scores with a cross-encoder teacher during training while retaining single-vector inference (Zhou et al., 2024). M3-Embedding combines dense, sparse, and multi-vector relevance scores as self-distillation targets (Chen et al., 2024a). StrataDistill uses global and local scores from frozen checkpoints of the student backbone.

3 METHOD

3.1 RETRIEVAL SETUP AND OVERVIEW

We consider a retrieval training corpus with $M \geq 1$ data sources. Each training set $\mathcal{D}_a$ contains queries, relevant documents, and sampled negatives. StrataDistill uses document structure to supervise one shared retriever with a single vector per document. Let f_θ be an encoder with shared parameters for queries and documents and ℓ_2-normalized outputs. The retrieval score is

$$s_\theta(q, d) = f_\theta(q)^\top f_\theta(d). \tag{1}$$

Queries and documents use their respective input formats. We use each full document as input, subject to the encoder input limit.

Figure 2 illustrates three stages using source-specific teachers: checkpoint preparation, structure-aware distillation, and single-vector inference. During distillation, the teacher scores both full documents and local views, while the student scores only full documents.

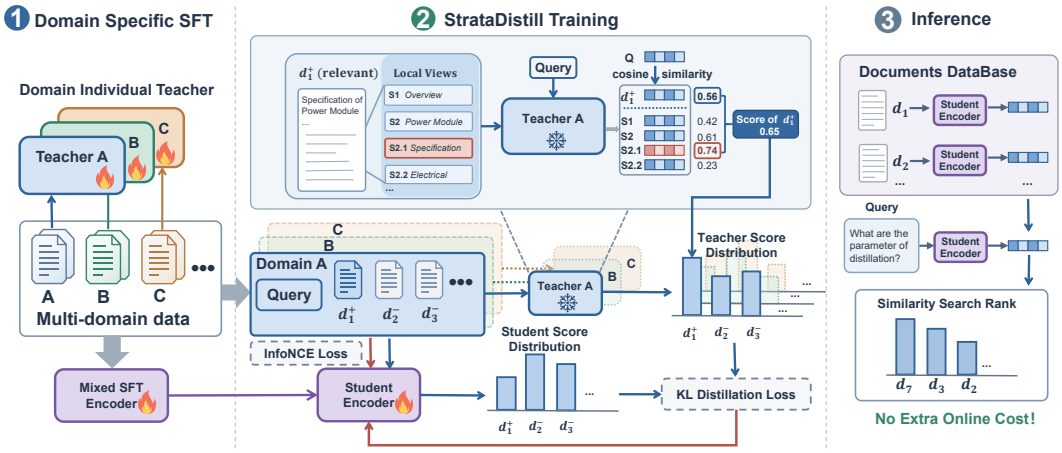


Figure 2: StrataDistill with source-specific teachers. Stage 1 prepares one SFT teacher per source and a mixed-data SFT checkpoint for student initialization. In Stage 2, each frozen source teacher combines full-document and maximum local-view scores into a document-level distribution. The shared student learns through KL distillation and InfoNCE on labeled query-document pairs. Stage 3 uses only the student for single-vector retrieval. Snowflakes and flames denote frozen and trainable encoders, respectively.

3.2 TEACHER AND STUDENT PREPARATION

Let θ_0 denote the original embedding checkpoint. In Stage 1, we fine-tune a copy of θ_0 on each $\mathcal{D}_a$ to obtain a source-specific teacher ϕ_a. We also fine-tune θ_0 on the pooled data $\bigcup_{a=1}^{M} \mathcal{D}_a$ to obtain θ_{mix}. Both paths use document-level contrastive learning with Equation 6. The shared student starts from θ_{mix}, and the source-specific teachers are frozen before distillation. This separates source-specific adaptation from consolidation in one student, as also explored in multi-teacher language-model distillation (Ma et al., 2026).

For a query q from source a, the illustrated configuration uses

$$\theta^{(0)} = \theta_{\text{mix}}, \qquad \phi(q) = \phi_a. \tag{2}$$

The same selected teacher encodes the query, every candidate document, and its local views. Batches from different sources use their corresponding teachers and update the same student.

A shared-teacher configuration sets $\phi(q) = \theta_{\text{mix}}$ for all sources and keeps the same student initialization. We evaluate these two SFT-based configurations. Initial-stage SFT can instead set $\theta^{(0)} = \phi(q) = \theta_0$ without fine-tuned teachers, but we do not evaluate this setting.

3.3 Structure-Aware Distillation

Teacher scores. For a document d, let $U(d) = \{u_1, \ldots, u_{K_d}\}$ denote a nonempty set of textual views. We use sections and subsections when structural boundaries are available, and chunks otherwise. When no decomposition is needed, we set $U(d) = \{d\}$. Views use the same input format as full documents. Given frozen teacher parameters $\phi = \phi(q)$, we compute

$$\begin{aligned} m_\phi(q, d) &= \max_{u \in U(d)} f_\phi(q)^\top f_\phi(u), \\ t_\phi(q, d) &= \frac{1}{2}\left[s_\phi(q, d) + m_\phi(q, d)\right]. \end{aligned} \tag{3}$$

Here, $s_\phi(q, d)$ is the full-document score from Equation 1, evaluated with teacher parameters. We average full-document similarity and the strongest local match for each query. In Figure 2, the full-document score of 0.56 and highest local score of 0.74 give a teacher score of 0.65. We apply this rule to every positive and negative candidate, with no query-to-section annotations. For a short document with $U(d) = \{d\}$, $t_\phi(q, d) = s_\phi(q, d)$ and the same training objective remains applicable.

Document-level distributions. Each per-device distillation batch contains queries from one source. For a query q, the candidate pool $\mathcal{C}_q^{\text{KD}}$ contains the positive documents and sampled negatives of all queries in that batch. The teacher uses $s_T(q, d) = t_{\phi(q)}(q, d)$, while the student uses only the full-document score $s_\theta(q, d)$. We normalize both sets of scores over these candidates with temperature $\tau_{\text{KD}} > 0$:

$$\begin{aligned} p_T(d \mid q) &= \frac{\exp(s_T(q, d)/\tau_{\text{KD}})}{\sum_{d' \in \mathcal{C}_q^{\text{KD}}} \exp(s_T(q, d')/\tau_{\text{KD}})}, \\ p_\theta(d \mid q) &= \frac{\exp(s_\theta(q, d)/\tau_{\text{KD}})}{\sum_{d' \in \mathcal{C}_q^{\text{KD}}} \exp(s_\theta(q, d')/\tau_{\text{KD}})}. \end{aligned} \tag{4}$$

The target distribution is over documents, with local views contributing to each teacher score. Teachers score candidates in their own embedding spaces, so distillation requires no coordinate alignment.

Student optimization. We minimize the KL divergence from the frozen teacher distribution to the student distribution:

$$\ell_{\text{KD}}(q) = D_{\text{KL}}(p_T(\cdot \mid q) \,\|\, p_\theta(\cdot \mid q)) = \sum_{d \in \mathcal{C}_q^{\text{KD}}} p_T(d \mid q) \log \frac{p_T(d \mid q)}{p_\theta(d \mid q)}. \tag{5}$$

The student also learns from labeled query-document pairs through an InfoNCE loss (van den Oord et al., 2018). For a labeled positive d^+ and a retrieval candidate pool $\mathcal{C}_q^{\text{ret}}$, it is

$$\ell_{\text{ret}}(q, d^+) = -\log \frac{\exp(s_\theta(q, d^+)/\tau_{\text{ret}})}{\sum_{d' \in \mathcal{C}_q^{\text{ret}}} \exp(s_\theta(q, d')/\tau_{\text{ret}})}, \tag{6}$$

where $\tau_{\text{ret}} > 0$ is the retrieval temperature. The retrieval pool includes cross-device negatives, while distillation uses candidates on the same device. The StrataDistill training objective is

$$\mathcal{L}(\theta) = \mathbb{E}_{a,(q,d^+)}\left[\lambda_{\text{ret}} \ell_{\text{ret}}(q, d^+) + \lambda_{\text{KD}} \ell_{\text{KD}}(q)\right], \tag{7}$$

where the expectation follows the training sampler over sources and examples, and both loss weights are positive. Both losses update query and document representations through the shared student encoder, as shown in Stage 2 of Figure 2. We precompute and fix teacher embeddings for queries, full documents, and local views. Teacher scores and candidate distributions are computed from these embeddings during training.

3.4 SINGLE-VECTOR INFERENCE

At inference, the student encodes each document once and stores one vector. Queries are matched against these vectors using Equation 1. Neither teacher selection nor local-view scoring is needed at inference. For the same encoder architecture and vector dimension, StrataDistill preserves the index size and online scoring procedure of the full-document bi-encoder. Teacher preparation and local-view computation add offline cost only.

4 EXPERIMENTS

We evaluate StrataDistill through five research questions. **RQ1:** Does it improve single-vector retrieval beyond SFT across embedding backbones? **RQ2:** Can individual teachers combine source-specific adaptation and local matching in one shared retriever? **RQ3:** Do native structural partitions improve supervision, and does distillation remain effective without them? **RQ4:** Are the long-document gains retained when training also includes short-document tasks? **RQ5:** Can structure-derived supervision match distillation from strong teachers?

4.1 EXPERIMENTAL SETUP

Datasets. We use five long-document retrieval datasets: ESG, MIRACL, MultiDoc2Dial, QASPER, and WixQA.

ESG is a new collection of corporate report pages on environmental, social, and governance topics. We apply page-level OCR and retain content and structure in HTML. **MIRACL** (Zhang et al., 2023) matches factual questions to Wikipedia passages. We reconstruct English articles by merging their passages.

MultiDoc2Dial (Feng et al., 2021) contains public-service documents with title hierarchies, sections, and text spans. **QASPER** (Dasigi et al., 2021) pairs questions with NLP research papers, retaining their texts and sections. Each question is a query, and its associated paper is the relevant document. **WixQA** (Cohen et al., 2025) matches questions about website building and product use to enterprise help-center articles. We evaluate on 400 expert-written and simulated queries, excluding the synthetic subset. Table 1 summarizes document lengths and section counts. Appendix A provides preprocessing and corpus details.

Table 1: Document lengths and section counts across datasets. We report the mean and 90th-percentile document lengths in tokens, along with the average number of sections per document.

Dataset	Mean	P90	Sections
ESG	747	1,089	6.97
MIRACL	1,188	2,926	10.79
MultiDoc2Dial	1,009	1,689	8.57
QASPER	5,283	7,978	14.82
WixQA	479	1,161	5.19

Training comparisons. We fix the embedding backbone and treat each dataset as one task. **Original** uses the embedding model without additional fine-tuning. **Mixed-task SFT** trains one shared model on all datasets with the document-level contrastive objective in Equation 6. **Task-specific SFT** trains and evaluates a separate model per dataset with the same objective, reporting five retrievers. **Continued SFT (CE-only)** adds a second contrastive stage to Mixed-task SFT on the same data to test whether further supervised training explains the gains.

Each indented + **StrataDistill** row freezes the preceding SFT model or models as teachers. Both variants train a shared student from the same Mixed-task SFT initialization using Equation 7. The mixed-task variant uses one teacher, while the task-specific variant routes each example to its source-dataset teacher. All systems in Table 2 use full-document single-vector retrieval.

Backbones and training. We evaluate four backbones with approximately 0.5B to 2B parameters. **KaLM-Embedding-V2.5** (Zhao et al., 2026) adapts Qwen2-0.5B with bidirectional attention and mean pooling. **Jina-embeddings-v3** (Sturua et al., 2024) is a 570M model based on XLM-RoBERTa with rotary position embeddings and task-specific LoRA adapters. **GTE-Qwen2-1.5B-instruct** (Li et al., 2023b) is a 1.5B instruction-based model built on Qwen2. **Youtu-Embedding** (Zhang et al., 2025a) is a 2B mean-pooling model for retrieval and semantic similarity.

Table 2: Recall@10 and nDCG@10 results on five long-document retrieval datasets. Avg. reports the unweighted mean across datasets. Arrows indicate score-point changes from Mixed-task SFT within each backbone block. Bold indicates the best result. † denotes one model per task.

Method	ESG		MIRACL		MultiDoc2Dial		QASPER		WixQA		Avg.	
	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10
KaLM-Embedding-V2.5 (496M parameters)												
Original	61.31	47.96	77.79	65.72	57.37	42.88	32.70	24.14	80.75	58.30	61.98	47.80
Mixed-task SFT	73.60	59.76	80.88	66.57	60.00	44.40	34.38	25.43	84.83	63.99	66.74	52.03
+ StrataDistill	75.76 $_{\uparrow 2.16}$	62.02 $_{\uparrow 2.26}$	81.28 $_{\uparrow 0.40}$	68.57 $_{\uparrow 2.00}$	60.52 $_{\uparrow 0.52}$	44.78 $_{\uparrow 0.38}$	38.09 $_{\uparrow 3.71}$	28.27 $_{\uparrow 2.84}$	84.67 $_{\downarrow 0.16}$	64.59 $_{\uparrow 0.60}$	68.06 $_{\uparrow 1.32}$	53.64 $_{\uparrow 1.61}$
Continued SFT (CE-only)	72.87 $_{\downarrow 0.73}$	59.09 $_{\downarrow 0.67}$	80.18 $_{\downarrow 0.70}$	65.45 $_{\downarrow 1.12}$	59.79 $_{\downarrow 0.21}$	43.88 $_{\downarrow 0.52}$	33.95 $_{\downarrow 0.43}$	25.18 $_{\downarrow 0.25}$	84.25 $_{\downarrow 0.58}$	64.47 $_{\uparrow 0.48}$	66.21 $_{\downarrow 0.53}$	51.61 $_{\downarrow 0.42}$
Task-specific SFT†	74.38 $_{\uparrow 0.78}$	60.70 $_{\uparrow 0.94}$	80.94 $_{\uparrow 0.06}$	68.12 $_{\uparrow 1.55}$	60.58 $_{\uparrow 0.58}$	44.76 $_{\uparrow 0.36}$	35.87 $_{\uparrow 1.49}$	26.82 $_{\uparrow 1.39}$	86.96 $_{\uparrow 2.13}$	66.13 $_{\uparrow 2.14}$	67.75 $_{\uparrow 1.01}$	53.31 $_{\uparrow 1.28}$
+ StrataDistill	**76.07** $_{\uparrow 2.47}$	**62.28** $_{\uparrow 2.52}$	**81.34** $_{\uparrow 0.46}$	**70.04** $_{\uparrow 3.47}$	**61.35** $_{\uparrow 1.35}$	**45.37** $_{\uparrow 0.97}$	**38.27** $_{\uparrow 3.89}$	**28.83** $_{\uparrow 3.40}$	**88.38** $_{\uparrow 3.55}$	**67.89** $_{\uparrow 3.90}$	**69.08** $_{\uparrow 2.34}$	**54.88** $_{\uparrow 2.85}$
Jina-embeddings-v3 (570M parameters)												
Original	49.49	36.49	67.49	52.52	53.19	39.40	27.81	20.05	82.00	61.64	56.00	42.02
Mixed-task SFT	66.81	51.57	78.32	64.50	55.57	40.60	30.78	22.28	81.21	55.21	62.54	46.83
+ StrataDistill	67.99 $_{\uparrow 1.18}$	52.93 $_{\uparrow 1.36}$	79.52 $_{\uparrow 1.20}$	67.55 $_{\uparrow 3.05}$	55.56 $_{\downarrow 0.01}$	40.68 $_{\uparrow 0.08}$	31.61 $_{\uparrow 0.83}$	23.00 $_{\uparrow 0.72}$	82.42 $_{\uparrow 1.21}$	57.11 $_{\uparrow 1.90}$	63.42 $_{\uparrow 0.88}$	48.25 $_{\uparrow 1.42}$
Continued SFT (CE-only)	66.51 $_{\downarrow 0.30}$	51.22 $_{\downarrow 0.35}$	78.21 $_{\downarrow 0.11}$	64.75 $_{\uparrow 0.25}$	55.04 $_{\downarrow 0.53}$	40.18 $_{\downarrow 0.42}$	30.98 $_{\uparrow 0.20}$	22.06 $_{\downarrow 0.22}$	81.75 $_{\uparrow 0.54}$	56.40 $_{\uparrow 1.19}$	62.50 $_{\downarrow 0.04}$	46.92 $_{\uparrow 0.09}$
Task-specific SFT†	**68.69** $_{\uparrow 1.88}$	**53.98** $_{\uparrow 2.41}$	77.13 $_{\downarrow 1.19}$	63.26 $_{\downarrow 1.24}$	56.28 $_{\uparrow 0.71}$	40.69 $_{\uparrow 0.09}$	30.94 $_{\uparrow 0.16}$	22.14 $_{\downarrow 0.14}$	**85.00** $_{\uparrow 3.79}$	**65.33** $_{\uparrow 10.12}$	63.61 $_{\uparrow 1.07}$	49.08 $_{\uparrow 2.25}$
+ StrataDistill	68.05 $_{\uparrow 1.24}$	53.15 $_{\uparrow 1.58}$	**79.56** $_{\uparrow 1.24}$	**68.74** $_{\uparrow 4.24}$	**56.40** $_{\uparrow 0.83}$	**41.50** $_{\uparrow 0.90}$	**32.20** $_{\uparrow 1.42}$	**23.49** $_{\uparrow 1.21}$	84.71 $_{\uparrow 3.50}$	61.63 $_{\uparrow 6.42}$	**64.19** $_{\uparrow 1.65}$	**49.70** $_{\uparrow 2.87}$
GTE-Qwen2-1.5B-instruct (1.5B parameters)												
Original	54.02	40.56	77.93	65.40	58.07	43.59	33.08	24.81	81.33	57.49	60.89	46.37
Mixed-task SFT	74.06	59.48	79.18	64.57	57.93	42.97	34.54	25.72	79.92	57.71	65.12	50.09
+ StrataDistill	**75.76** $_{\uparrow 1.70}$	61.50 $_{\uparrow 2.02}$	79.89 $_{\uparrow 0.71}$	66.84 $_{\uparrow 2.27}$	58.54 $_{\uparrow 0.61}$	43.35 $_{\uparrow 0.38}$	37.77 $_{\uparrow 3.23}$	28.82 $_{\uparrow 3.10}$	82.29 $_{\uparrow 2.37}$	59.60 $_{\uparrow 1.89}$	66.85 $_{\uparrow 1.73}$	52.02 $_{\uparrow 1.93}$
Continued SFT (CE-only)	74.72 $_{\uparrow 0.66}$	60.57 $_{\uparrow 1.09}$	**80.06** $_{\uparrow 0.88}$	66.57 $_{\uparrow 2.00}$	55.88 $_{\downarrow 2.05}$	40.90 $_{\downarrow 2.07}$	35.57 $_{\uparrow 1.03}$	26.35 $_{\uparrow 0.63}$	83.38 $_{\uparrow 3.46}$	60.13 $_{\uparrow 2.42}$	65.92 $_{\uparrow 0.80}$	50.91 $_{\uparrow 0.82}$
Task-specific SFT†	74.37 $_{\uparrow 0.31}$	60.48 $_{\uparrow 1.00}$	79.45 $_{\uparrow 0.27}$	65.24 $_{\uparrow 0.67}$	57.42 $_{\downarrow 0.51}$	41.62 $_{\downarrow 1.35}$	38.80 $_{\uparrow 4.26}$	28.94 $_{\uparrow 3.22}$	84.04 $_{\uparrow 4.12}$	62.55 $_{\uparrow 4.84}$	66.82 $_{\uparrow 1.70}$	51.77 $_{\uparrow 1.68}$
+ StrataDistill	**75.76** $_{\uparrow 1.70}$	**61.79** $_{\uparrow 2.31}$	79.98 $_{\uparrow 0.80}$	**67.16** $_{\uparrow 2.59}$	**59.19** $_{\uparrow 1.26}$	**43.60** $_{\uparrow 0.63}$	**39.22** $_{\uparrow 4.68}$	**29.99** $_{\uparrow 4.27}$	**85.83** $_{\uparrow 5.91}$	**64.28** $_{\uparrow 6.57}$	**68.00** $_{\uparrow 2.88}$	**53.36** $_{\uparrow 3.27}$
Youtu-Embedding (2B parameters)												
Original	63.72	50.22	82.91	73.30	56.63	42.22	31.53	23.37	74.29	55.02	61.82	48.83
Mixed-task SFT	69.66	54.80	83.18	69.86	58.60	43.25	31.31	22.61	75.13	53.72	63.58	48.85
+ StrataDistill	73.15 $_{\uparrow 3.49}$	59.15 $_{\uparrow 4.35}$	**84.81** $_{\uparrow 1.63}$	73.15 $_{\uparrow 3.29}$	60.15 $_{\uparrow 1.55}$	44.77 $_{\uparrow 1.52}$	35.10 $_{\uparrow 3.79}$	25.95 $_{\uparrow 3.34}$	79.29 $_{\uparrow 4.16}$	57.05 $_{\uparrow 3.33}$	66.50 $_{\uparrow 2.92}$	52.01 $_{\uparrow 3.16}$
Continued SFT (CE-only)	68.85 $_{\downarrow 0.81}$	54.29 $_{\downarrow 0.51}$	83.56 $_{\uparrow 0.38}$	69.89 $_{\uparrow 0.03}$	58.00 $_{\downarrow 0.60}$	42.07 $_{\downarrow 1.18}$	32.80 $_{\uparrow 1.49}$	23.93 $_{\uparrow 1.32}$	77.83 $_{\uparrow 2.70}$	56.43 $_{\uparrow 2.71}$	64.21 $_{\uparrow 0.63}$	49.32 $_{\uparrow 0.47}$
Task-specific SFT†	69.28 $_{\uparrow 0.38}$	54.45 $_{\uparrow 0.35}$	83.67 $_{\uparrow 0.49}$	71.47 $_{\uparrow 1.61}$	57.89 $_{\downarrow 0.71}$	42.53 $_{\downarrow 0.72}$	32.09 $_{\uparrow 0.78}$	23.18 $_{\uparrow 0.57}$	79.17 $_{\uparrow 4.04}$	57.03 $_{\uparrow 3.31}$	64.42 $_{\uparrow 0.84}$	49.73 $_{\uparrow 0.88}$
+ StrataDistill	**74.28** $_{\uparrow 4.62}$	**60.40** $_{\uparrow 5.60}$	84.73 $_{\uparrow 1.55}$	**73.50** $_{\uparrow 3.64}$	**61.30** $_{\uparrow 2.70}$	**45.78** $_{\uparrow 2.53}$	**36.86** $_{\uparrow 5.55}$	**27.36** $_{\uparrow 4.75}$	**81.83** $_{\uparrow 6.70}$	**60.52** $_{\uparrow 6.80}$	**67.80** $_{\uparrow 4.22}$	**53.51** $_{\uparrow 4.66}$

We mine negatives with Qwen3-Embedding-0.6B (Zhang et al., 2025b). For each query, we randomly sample seven non-positive documents from its top-50 results. All stages, methods, and backbones reuse these examples and negatives. InfoNCE shares negatives across GPUs, while KL distillation uses candidates on the same GPU. We set $\lambda_{\text{ret}} = 0.3$ and $\lambda_{\text{KD}} = 0.7$ and use an 8k-token training input limit, covering most documents in full. We report nDCG@10 and Recall@10 on a 0–100 scale with dataset-equal macro averages. Appendix B provides implementation details.

4.2 RQ1: Retrieval Effectiveness across Backbones

Gains across backbones. With task-specific teachers, StrataDistill improves both metrics over Mixed-task SFT in all 20 backbone–dataset combinations in Table 2. Average nDCG@10 increases by 2.85, 2.87, 3.27, and 4.66 points for KaLM, Jina, GTE, and Youtu, respectively. Average Recall@10 increases by 1.65 to 4.22 points. The gains span encoder and decoder models up to 2B.

Additional supervision from the same checkpoint. The shared-teacher variant improves nDCG@10 in all 20 combinations, with macro gains of 1.42 to 3.16 points. Teacher and student start from the same checkpoint. Local scoring can distinguish candidates differently even though the checkpoint and labeled pairs are unchanged. Local views add relevance supervision, while deployment retains the same index size and full-document scoring.

Distillation versus additional SFT. Continued SFT changes average nDCG@10 by -0.42 to $+0.82$ points, below the gains from either StrataDistill configuration. The shared-teacher variant exceeds this control by 1.11 to 2.69 points on the same data.

For Youtu, Mixed-task SFT barely changes average nDCG@10, from 48.83 to 48.85, and Continued SFT reaches 49.32. StrataDistill reaches 52.01 with the shared teacher and 53.51 with task-specific teachers. Together, these comparisons identify structure-derived targets as useful supervision beyond another pass over document-level labels.

4.3 RQ2: Individual Teachers for a Shared Retriever

Individual adaptation produces better supervision. Task-specific SFT exceeds Mixed-task SFT in both macro metrics for every backbone, gaining 0.88 to 2.25 nDCG@10 points. Its nDCG@10 advantage holds in 15 of the 20 backbone–dataset combinations. Replacing the shared teacher with individual teachers improves student nDCG@10 in all 20 combinations. Macro gains are 1.24 to 1.50 nDCG@10 points and 0.77 to 1.30 Recall@10 points. WixQA shows the largest nDCG@10 benefit from both individual SFT and individual-teacher distillation for every backbone. The latter gains range from 3.30 to 4.68 points over the shared teacher. Source-specific adaptation thus improves supervision for a shared retriever with student initialization, data, and objective unchanged.

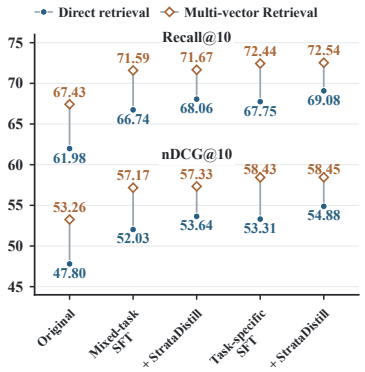


Figure 3: Direct and multi-vector retrieval with KaLM.

Consolidation improves on the individual retrievers. The student exceeds its corresponding task-specific SFT retriever in both metrics in 18 of the 20 combinations, except Jina on ESG and WixQA. Both macro metrics improve for every backbone, with nDCG@10 gains of 0.62 to 3.78 points. The individual-teacher student also outperforms the shared-teacher student in all five cases where individual SFT initially trails mixed training.

On Jina MIRACL, individual SFT scores 63.26 nDCG@10 versus 64.50 for Mixed-task SFT, yet the individual-teacher student reaches 68.74. The shared-teacher student reaches 67.55. Teacher accuracy in direct retrieval does not fully determine distillation value. The student combines structure-aware targets with pooled supervision.

Distillation narrows the gap left by ordinary SFT. We use KaLM for Figure 3 and the remaining analyses. Section-aware evaluation reranks the top 1,000 full-document candidates with a full-document score floor. Local scoring raises nDCG@10 by 5.46 points for the original checkpoint and by 5.14 and 5.12 points after mixed and individual SFT. SFT leaves local relevance signals that are not fully reflected in direct retrieval.

After StrataDistill, these within-model gaps fall to 3.69 and 3.57 points for the two students. Relative to the respective teachers, student direct scores rise by 1.61 and 1.57 points, while section-aware scores change by only 0.16 and 0.02 points. The gains concentrate in direct retrieval, consistent with partial transfer of the local-matching advantage into full-document representations.

Individual teachers also raise the section-aware reference by 1.26 points over the shared teacher, accompanied by a 1.24-point gain in student direct retrieval. Teacher specialization and local-view scoring serve different roles. Specialization adapts the teacher to each source, while local scoring exposes relevance within documents. The shared student benefits from both without retaining separate source-specific retrievers. Its gains over individual SFT show that consolidation can improve retrieval rather than merely reproduce the separate models.

4.4 RQ3: Native Sections versus Fixed-Token Chunks

We replace native views with non-overlapping chunks of 128, 256, 512, or 1024 KaLM tokens. Native views follow document structure or dataset-provided passages, as detailed in Appendix A. Teacher checkpoints, student initialization, training examples, loss weights, and schedule remain fixed. Both variants use a full-document score floor and single-vector student inference.

Native structure refines the supervision. Native sections outperform all four fixed lengths in both macro metrics in Figure 4. They reach 54.88 nDCG@10 and 69.08 Recall@10, exceeding the best fixed length of 128 tokens by 0.29 and 0.24 points. The nDCG@10 margin increases to 0.62 points against 1024-token chunks. Native sections also lead the section-aware teacher reference, scoring 58.43 versus 57.91 for the best fixed length.

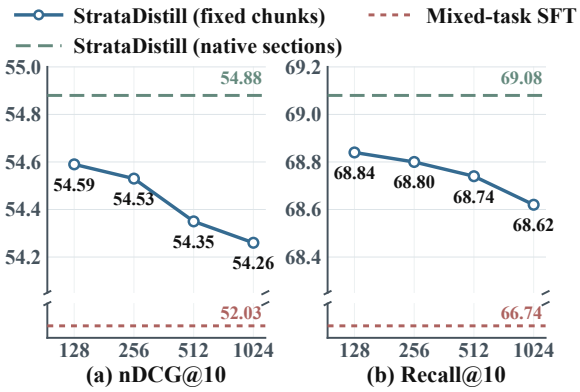


Figure 4: Student scores by chunk length. Dashed lines mark native sections and SFT.

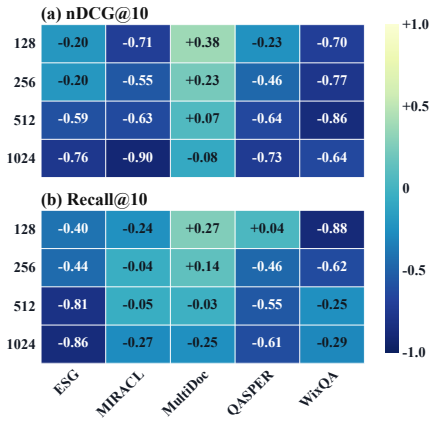


Figure 5: Fixed-chunk gains relative to native document sections.

In Figure 5, native sections exceed every fixed length in nDCG@10 on ESG, MIRACL, QASPER, and WixQA. MultiDoc2Dial is the exception, with gains of up to 0.38 points from short fixed chunks. Both macro metrics decrease as fixed chunks grow longer. Only the teacher views change, so this effect reaches retrieval through supervision while student inputs remain full documents. Native partitions adapt view boundaries, lengths, and counts to the document instead of imposing one token length on every local unit.

Structural annotations are useful but not required. Fixed chunks improve both metrics over Mixed-task SFT in all 20 length–dataset combinations. Macro nDCG@10 gains range from 2.23 to 2.56 points, and Recall@10 gains range from 1.88 to 2.10 points. With 1024-token chunks, the nDCG@10 gain remains 2.23 points. The best fixed length preserves 89.8% of the native-section nDCG@10 gain. Decomposition therefore supplies useful supervision without chunk-level search, while native structure further improves it.

4.5 RQ4: Extension to Short-Document Tasks

We jointly train on the five long-document datasets and four short-document tasks: DuReader, FEVER, NQ, and SQuAD v2. Table 3 compares matched nine-dataset runs using full-document single-vector retrieval. Appendix C.3 reports per-dataset scores and averaging conventions.

Table 3: Joint training on long and short documents. R/N@10 denotes Recall/nDCG@10. † denotes one model per task.

Method	Long (5)		Short (4)		All (9)	
	R@10	N@10	R@10	N@10	R@10	N@10
Original	61.98	47.80	93.35	87.50	75.92	65.45
Mixed-task SFT	65.98	52.05	94.45	89.24	78.63	68.58
+ StrataDistill	67.97	54.22	94.47	89.31	79.75	69.82
Task-specific SFT†	67.63	53.28	94.36	89.40	79.51	69.33
+ StrataDistill	**69.11**	**54.95**	**94.57**	**89.58**	**80.43**	**70.34**

The individual-teacher student improves nDCG@10 over Mixed-task SFT on all nine datasets. Both metrics improve on every long-document task. Long-document averages rise from 52.05 to 54.95 in nDCG@10 and from 65.98 to 69.11 in Recall@10. Short-document nDCG@10 gains range from 0.23 to 0.40 points across the four tasks, raising the macro score from 89.24 to 89.58. Macro Recall@10 increases from 94.45 to 94.57.

Individual teachers add 0.73 nDCG@10 points on long documents and 0.27 on short documents over the shared teacher. Their student exceeds the nine task-specific SFT retrievers in all six group-level metrics in Table 3. Long-document gains therefore persist alongside short-document improvements in one shared retriever.

For an undecomposed document, $U(d) = \{d\}$ makes the teacher score equal to full-document similarity in Equation 3. The contrastive and distillation losses both remain active. Native sections, fixed chunks, and whole documents can therefore share one training objective.

4.6 RQ5: Structure-Derived Supervision versus Strong Teachers

Table 4 compares StrataDistill with distillation from Qwen3-Embedding-8B and Qwen3-Reranker-8B (Zhang et al., 2025b). Embedding-8B KD uses query-to-full-document similarity, while StrataDistill combines global and local scores from source-specific KaLM teachers. Both embedding-teacher settings share KL candidates within each device batch. Qwen3-Reranker-8B instead scores only the query-specific one positive and seven negatives for KL distillation. All configurations train the same KaLM student with identical Mixed-task SFT initialization and training schedule. Appendix B gives the full distillation protocol.

Table 4: KaLM student retrieval with different teachers. Embedding-8B KD and Reranker-8B KD use full-document scores from the corresponding Qwen3 teachers. StrataDistill combines global and local scores from source-specific KaLM SFT teachers. R/N@10 denotes Recall/nDCG@10, and Avg. is the dataset-equal mean. Best results per row and metric are bold.

Dataset	Mixed-task SFT		Embedding-8B KD		Reranker-8B KD		StrataDistill KaLM teachers	
	R@10	N@10	R@10	N@10	R@10	N@10	R@10	N@10
ESG	73.60	59.76	69.73	55.74	73.37	59.77	**76.07**	**62.28**
MIRACL	80.88	66.57	79.51	68.96	80.08	70.03	**81.34**	**70.04**
MultiDoc2Dial	60.00	44.40	61.25	**46.06**	61.12	45.36	**61.35**	45.37
QASPER	34.38	25.43	**40.17**	**30.59**	32.82	24.62	38.27	28.83
WixQA	84.83	63.99	86.50	63.40	86.88	63.29	**88.38**	**67.89**
Avg.	66.74	52.03	67.43	52.95	66.85	52.62	**69.08**	**54.88**

Structure provides competitive supervision. StrataDistill reaches the highest macro scores at 54.88 nDCG@10 and 69.08 Recall@10. This student exceeds Embedding-8B KD and Reranker-8B KD by 1.93 and 2.26 nDCG@10 points, respectively. Recall@10 gains are 1.65 and 2.23 points.

The Embedding-8B KD student achieves the best QASPER results and MultiDoc2Dial nDCG@10, whereas StrataDistill leads both metrics on ESG, MIRACL, and WixQA. Only StrataDistill improves both metrics over Mixed-task SFT on all five datasets.

These results support document structure as an effective source of supervision for long-document retrieval. StrataDistill combines source adaptation with explicit local matching, using teachers of the student backbone. This combination yields higher average student scores than either strong-teacher baseline across the evaluated long-document tasks. The local-scoring gap after SFT in Section 4.3 likewise points to useful evidence beyond document-level labels. Local views expose relevance cues within long documents during training, while inference retains one vector per document and the original similarity scoring procedure.

5 Conclusion

StrataDistill uses document structure to enrich supervision for embedding fine-tuning. Frozen teachers combine global and local relevance scores, which a shared student learns through full-document scoring. Across five datasets and four backbones, the method improves on shared-model SFT, with further gains from source-specific teachers. The reduced local-scoring gap suggests that part of this additional relevance signal transfers into single-vector retrieval. Native sections give the best average results, while fixed chunks remain effective when structural annotations are unavailable. The same objective also supports joint training on short documents. These gains preserve the original index size and online scoring procedure. Future work could use structural relations to combine evidence across sections when constructing teacher targets.

AI USE STATEMENT

AI tools assisted language editing and data processing, and Qwen3-Max generated synthetic data. The authors reviewed the AI-assisted content and take full responsibility for this work.

ETHICS STATEMENT

This work studies a retrieval training method using publicly available datasets, pretrained models, and an ESG corpus derived from publicly accessible corporate reports. It involves no recruitment of human participants or collection of new personal data. We use the source resources in accordance with their applicable licenses and terms of use. Synthetic data may contain biases, factual errors, or queries that do not reflect the source documents. These limitations should inform data reuse and interpretation of the results.

REPRODUCIBILITY STATEMENT

Section 3 specifies the method, and Section 4.1 describes the datasets, backbones, training, and evaluation. Appendices A and B detail data preparation and implementation. Appendix C reports additional analyses. We will release the training, evaluation, data synthesis, and filtering code, together with the training and test datasets.

REFERENCES

Huiyao Chen, Yi Yang, Yinghui Li, Meishan Zhang, Baotian Hu, and Min Zhang. Beyond chunking: Discourse-aware hierarchical retrieval for long document question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18176–18198. Association for Computational Linguistics, 2026.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024a.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense x retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15159–15177, 2024b.

Dvir Cohen, Lin Burg, Sviatoslav Pykhnivskyi, Hagit Gur, Stanislav Kovynov, Olga Atzmon, and Gilad Barkan. WixQA: A multi-dataset benchmark for enterprise retrieval-augmented generation. *arXiv preprint arXiv:2505.08643*, 2025. URL https://arxiv.org/abs/2505.08643.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, 2021. URL https://aclanthology.org/2021.naacl-main.365/.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6162–6176, 2021. URL https://aclanthology.org/2021.emnlp-main.498/.

Hongchao Gu, Dexun Li, Kuicai Dong, Hao Zhang, Hang Lv, Hao Wang, Defu Lian, Yong Liu, and Enhong Chen. RAPID: Efficient retrieval-augmented long text generation with writing planning and information discovery. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16742–16763. Association for Computational Linguistics, 2025.

Michael Günther, Isabelle Mohr, Daniel J Williams, Bo Wang, and Han Xiao. Late chunking: Contextual chunk embeddings using long-context embedding models. *arXiv preprint arXiv:2409.04701*, 2024.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. DuReader: a chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pp. 37–46, 2018. doi: 10.18653/v1/W18-2605. URL https://aclanthology.org/W18-2605/.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*, 2020.

Xinhao Huang, Zhibo Ren, Yipeng Yu, Ying Zhou, Zulong Chen, and Zeyi Wen. SEAL: Structure and element aware learning to improve long structured document retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8526–8536, 2025. doi: 10.18653/v1/2025.emnlp-main.429. URL https://aclanthology.org/2025.emnlp-main.429/.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*, 2024.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

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*, volume 2025, pp. 79310–79333, 2025.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. Sailer: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1035–1044, 2023a.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023b.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386*, 2020.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 2021.

Kun Luo, Zheng Liu, Shitao Xiao, and Kang Liu. BGE landmark embedding: A chunking-free embedding method for retrieval augmented long-context large language models. *arXiv preprint arXiv:2402.11573*, 2024.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Zach Nussbaum, John Xavier Morris, Andriy Mulyar, and Brandon Duderstadt. Nomic embed: Training a reproducible long context text embedder. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=IPmzyQSiQE. Reproducibility Certification.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2825–2835, 2021. doi: 10.18653/v1/2021.emnlp-main.224.

Jon Saad-Falcon, Daniel Y Fu, Simran Arora, Neel Guha, and Christopher Ré. Benchmarking and building long-context retrieval models with loco and m2-bert. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 42918–42946, 2024.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, 2022.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 32628–32649, 2024.

Shruti Singh and Mayank Singh. Cosaemb: Contrastive section-aware aspect embeddings for scientific articles. In *Proceedings of the fourth workshop on scholarly document processing (sdp 2024)*, pp. 283–292, 2024.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task LoRA. *arXiv preprint arXiv:2409.10173*, 2024. URL https://arxiv.org/abs/2409.10173.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2345–2360, 2022a. doi: 10.18653/v1/2022.naacl-main.168.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022b.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, 2024.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021.

Junhan Yang, Zheng Liu, Chaozhuo Li, Guangzhong Sun, and Xing Xie. Longtriever: a pre-trained long text encoder for dense document retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3655–3665, 2023.

Bowen Zhang, Zixin Song, Chunquan Chen, Qian-Wen Zhang, Di Yin, and Xing Sun. CoDiEmb: A collaborative yet distinct framework for unified representation learning in information retrieval and semantic textual similarity. *arXiv preprint arXiv:2508.11442*, 2025a. URL https://arxiv.org/abs/2508.11442.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. URL https://aclanthology.org/2023.tacl-1.63/.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025b. URL https://arxiv.org/abs/2506.05176.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. KaLM-Embedding-V2: Superior training techniques and data inspire a versatile embedding model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y7qzhvWhcz.

Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Jianbing Shen, Guodong Long, Can Xu, and Daxin Jiang. Fine-grained distillation for long document retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19732–19740, 2024.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Longembed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 802–816, 2024.

A Dataset Details

A.1 Long-Document Datasets

Table 5 summarizes the five processed retrieval datasets used in the main experiments. Training counts refer to sampled entries, and evaluation counts refer to queries. Document counts refer to indexed retrieval units: pages for ESG, reconstructed articles for MIRACL, and whole documents for the other datasets. We measure token lengths before applying the 8k-token training limit.

Table 5: Dataset sizes and document lengths. Train counts sampled training entries, and Test counts evaluation queries. Mean and P90 are document token lengths.

Dataset	Train	Test	Documents	Mean	P90
ESG	5,000	15,000	59,771	747	1,089
MIRACL	5,000	3,662	100,000	1,188	2,926
MultiDoc2Dial	1,869	32,266	488	1,009	1,689
QASPER	5,000	5,049	1,585	5,283	7,978
WixQA	5,000	400	6,221	479	1,161

Synthetic training queries. For all five datasets, we generate candidate training queries with Qwen3-Max. After removing duplicates, we embed the remaining training queries and the test queries with Qwen3-Embedding-0.6B. We compute their pairwise cosine similarities and discard any training query whose similarity to a test query exceeds 0.85. We then randomly sample the required number of retained queries for each training set. We will release the generated data and query-generation prompts for each task.

ESG. We downloaded approximately 3,000 publicly available environmental, social, and governance (ESG) reports in PDF format from the web. We applied OCR to the reports and reconstructed the hierarchy of each page from the OCR output. Each page is treated as one retrieval document, with its reconstructed hierarchy serving as the native structure. The resulting page-level corpus contains 59,771 documents. The ESG test examples are also generated with Qwen3-Max.

MIRACL. MIRACL provides short passages split from Wikipedia articles. To evaluate long-document retrieval, we reconstruct each article by concatenating its passages in order. The article containing a relevant passage becomes the relevant document for that query. The English source collection contains 5,758,285 articles, which is too large for our evaluation setting. We therefore construct a 100,000-article corpus from the original English training and development queries. For each query, we include the top 20 articles retrieved by BM25 and all relevant articles. We then randomly sample additional articles to reach 100,000. Because both source query splits are small, we combine them into one evaluation set of 3,662 queries.

MultiDoc2Dial. MultiDoc2Dial contains dialogue turns grounded in 488 public-service documents. Span references in the dialogue turns are mapped to their parent documents for document-level retrieval. Local views use the document title hierarchies, sections, and text spans.

QASPER. QASPER contains questions about full-text research papers. Each paper is treated as a retrieval unit, with its associated questions used as queries. The source data preserve the hierarchy of each paper, from sections and subsections to paragraphs. We use this structure to construct hierarchical section views for local matching.

WixQA. WixQA pairs questions with a corpus of 6,221 help-center items. It provides 200 expert-written queries, 200 simulated queries, and 6,221 synthetic queries. The synthetic subset is much larger and easier, producing high scores across retrievers that can obscure their differences. We therefore evaluate only the expert-written and simulated subsets, which together contain 400 queries and 488 query–article relevance pairs. We recover article structure from the HTML markup.

A.2 Short-Document Datasets

The nine-dataset extension adds four short-document retrieval datasets: DuReader, FEVER, Natural Questions (NQ), and SQuAD v2. For FEVER, NQ, and SQuAD v2, we sample 5,000 examples from each dataset's original training split. **DuReader** contains Chinese questions from Baidu Search and Baidu Zhidao (He et al., 2018). We use the DuReader corpus and test queries provided by CMTEB. To construct the training set, we retain queries from the original DuReader training split whose positive documents are included in this corpus, then randomly sample 5,000 of these queries. **FEVER** (Thorne et al., 2018) contains claims constructed by modifying Wikipedia sentences and annotated as supported, refuted, or lacking sufficient evidence. In our retrieval setting, claims serve as queries, and the annotated evidence identifies relevant documents for both supported and refuted claims. **NQ** (Kwiatkowski et al., 2019) uses Google search questions as queries and the provided annotations to identify relevant Wikipedia texts. Its original annotations include both long-answer regions and short answers, which can be text spans or yes/no responses. **SQuAD v2** (Rajpurkar et al., 2018) uses questions as queries and Wikipedia passages as retrieval units, with answer annotations identifying relevant passages for answerable questions. The source dataset also includes adversarially written questions that cannot be answered from their associated passages. We downsample the FEVER and NQ retrieval corpora to 100,000 documents to reduce evaluation cost. Table 6 summarizes dataset sizes and document lengths. Across these four corpora, mean document lengths range from 120 to 241 tokens, and the 90th-percentile length is below 400 tokens in every corpus.

The documents are substantially shorter than those in the five long-document corpora, whose mean lengths range from 479 to 5,283 tokens.

Table 6: Dataset sizes and document lengths for the four short-document datasets. Train is the number of sampled training examples and Test is the number of evaluation queries. Mean and P90 are the mean and 90th-percentile document lengths in tokens.

Dataset	Train	Test	Documents	Mean	P90
DuReader	5,000	2,000	100,000	241	373
FEVER	5,000	6,666	100,000	148	326
NQ	5,000	3,452	100,000	120	223
SQuAD v2	5,000	5,928	20,081	165	254

B Implementation Details

We use full-parameter fine-tuning for all four backbones, with each training stage running for one epoch. During distillation, the teacher remains frozen while all student parameters are updated. We use AdamW with a learning rate of 10^{-5} and a linear learning-rate schedule with a warmup ratio of 0.1. The InfoNCE and distillation temperatures are $\tau_{\text{ret}} = 0.01$ and $\tau_{\text{KD}} = 0.05$, respectively.

Training runs on eight NVIDIA A800 GPUs with four queries per GPU. We accumulate gradients over four steps, resulting in an effective batch size of 128 queries. For training on pooled datasets, a dataset-aware sampler draws all examples in each per-GPU batch from a single dataset. InfoNCE shares negatives within each group and across GPUs. The distillation loss uses only negatives shared within the same GPU.

Teacher comparison. All distillation configurations in Table 4 initialize the KaLM student from the same Mixed-task SFT checkpoint. They use the same training data, sampled negatives, loss weights, and schedule. Embedding-8B KD uses query-to-full-document cosine similarity from the frozen Qwen3-Embedding-8B teacher. Its teacher score is $s_\phi(q, d)$ from Equation 1, with no local-view scoring. StrataDistill uses source-specific KaLM teachers and the combined global and local score in Equation 3. Both embedding-teacher configurations construct KL distributions over the candidates shared within each per-device batch.

For Qwen3-Reranker-8B, let $z_{\text{yes}}(q, d)$ and $z_{\text{no}}(q, d)$ denote the yes/no token logits for a query and full document. A sigmoid maps their difference to a relevance score:

$$r(q, d) = \sigma(z_{\text{yes}}(q, d) - z_{\text{no}}(q, d)) \,. \tag{8}$$

For each query, the reranker KL candidate set is $\mathcal{C}_q^{\text{RR}} = \{d^+, d_1^-, \ldots, d_7^-\}$. It contains the positive and seven sampled negatives for that query, without sharing candidates across queries. We then normalize the sigmoid scores across these eight documents:

$$p_T^{\text{RR}}(d \mid q) = \frac{\exp(r(q, d)/\tau_{\text{KD}})}{\sum_{d' \in \mathcal{C}_q^{\text{RR}}} \exp(r(q, d')/\tau_{\text{KD}})}, \qquad \tau_{\text{KD}} = 0.05. \tag{9}$$

The student distribution normalizes full-document similarities over the same eight candidates with the same temperature. We minimize the teacher-to-student KL divergence as in Equation 5. The InfoNCE term retains cross-device negatives. Every distillation configuration uses $\lambda_{\text{ret}} = 0.3$ and $\lambda_{\text{KD}} = 0.7$ in the joint loss.

C Additional Ablation Results

The additional analyses use KaLM to examine loss weighting, per-dataset retrieval gaps, and joint training on long and short documents.

C.1 Effect of the Distillation Weight

We vary the distillation weight in Equation 7 over $\lambda_{\text{KD}} \in \{0, 0.3, 0.5, 0.7, 1\}$, with $\lambda_{\text{ret}} = 1 - \lambda_{\text{KD}}$. Weights of 0 and 1 correspond to InfoNCE-only and KL-only training, respectively, in the additional training stage. Figure 6 reports average scores of nDCG@10 and Recall@10 over the five long-document datasets.

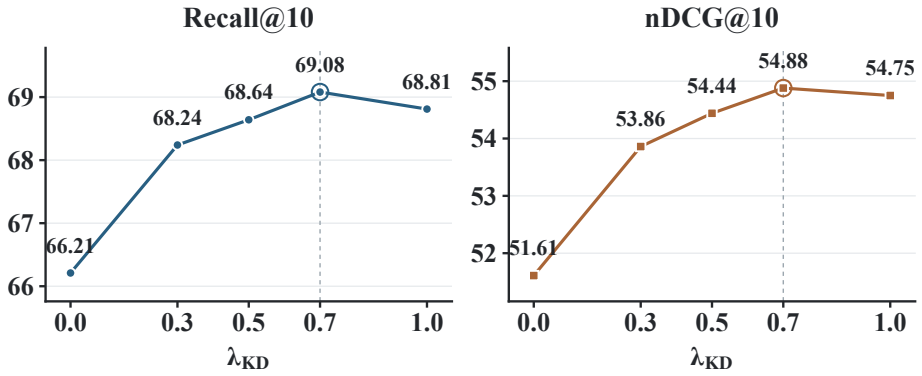


Figure 6: Effect of the KL distillation weight on KaLM retrieval scores averaged over the five long-document datasets. The InfoNCE weight is $\lambda_{\text{ret}} = 1 - \lambda_{\text{KD}}$. The dashed lines mark the default weight of 0.7.

All tested settings with a nonzero KL weight outperform InfoNCE-only training in both metrics. The default weight of 0.7 gives the highest scores among the tested settings on every dataset, raising nDCG@10 from 51.61 to 54.88 and Recall@10 from 66.21 to 69.08. KL-only training reaches 54.75 nDCG@10 and 68.81 Recall@10, retaining most of this gain. These results suggest that structure-aware KL supervision is the main source of improvement beyond SFT. Retaining a small InfoNCE term provides an anchor to the labeled query-document pairs and yields a modest improvement over relying on teacher supervision alone.

C.2 Per-Dataset Retrieval Gaps

Figure 7 shows the per-dataset results corresponding to the macro comparison in Figure 3. Each pair of markers compares direct single-vector retrieval with structure-aware multi-vector retrieval using the same checkpoint. StrataDistill uses the preceding SFT checkpoints as frozen teachers.

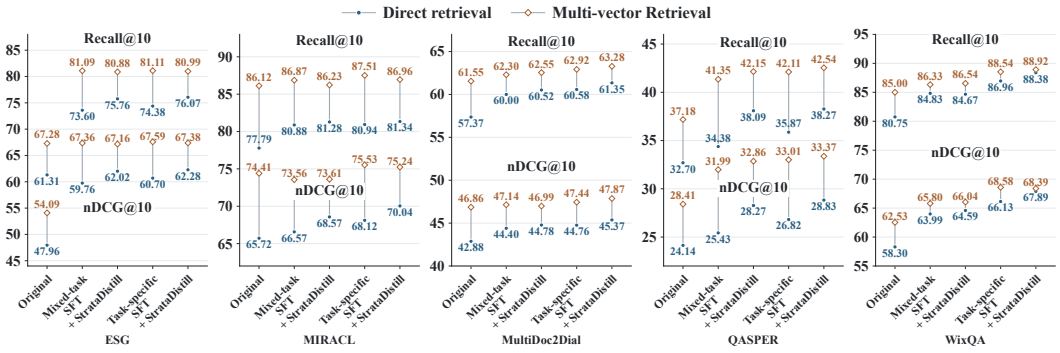


Figure 7: Direct and multi-vector retrieval with KaLM on the five datasets.

The task-specific-teacher student has a smaller gap between direct and multi-vector retrieval than task-specific SFT in both metrics on all five datasets. WixQA shows the clearest reduction, with the nDCG@10 gap decreasing from 2.45 to 0.50 points and the Recall@10 gap from 1.58 to 0.54 points. The student scores 67.89 nDCG@10 and 88.38 Recall@10 in direct retrieval, approaching

the multi-vector teacher scores of 68.58 and 88.54. The remaining differences are 0.69 and 0.16 points. The shared student thus recovers much of the local-scoring benefit on WixQA.

The improvements also extend across the other datasets. The nDCG@10 gap narrows by 1.79 points on ESG, 2.21 points on MIRACL, and 1.65 points on QASPER. The task-specific-teacher student also exceeds the direct-retrieval scores of its teacher in both metrics on all five datasets. These results support transferring local-matching benefits into single-vector retrieval.

C.3 Per-Dataset Results for the Nine-Dataset Extension

Table 7 expands the group averages reported for the nine-dataset experiment in Table 3. The training mixture combines the five long-document datasets with DuReader, FEVER, NQ, and SQuAD v2. Appendix A.2 describes preparation of the short-document datasets. Each StrataDistill column uses the preceding SFT model or models as teachers. Both students start from the nine-dataset Mixed-task SFT checkpoint. The results are from separate nine-dataset runs with matched SFT baselines. Group averages give equal weight to each dataset. They use unrounded scores for trained models and displayed scores for Original.

Table 7: Per-dataset results for joint training on long and short documents. Each cell reports nDCG@10 / Recall@10 on a scale from 0 to 100. Each + column uses the preceding SFT model or models as teachers. † denotes one retriever per task.

Dataset	Original	Mixed-task SFT	+ StrataDistill	Task-specific SFT†	+ StrataDistill
ESG	47.96 / 61.31	58.62 / 72.15	61.62 / 75.45	60.71 / 74.43	62.21 / 75.72
MIRACL	65.72 / 77.79	69.83 / 81.03	71.51 / 81.95	68.05 / 80.88	71.51 / 81.96
MultiDoc2Dial	42.88 / 57.37	44.60 / 60.40	45.39 / 61.25	44.66 / 60.45	45.18 / 61.22
QASPER	24.14 / 32.70	23.87 / 32.74	27.27 / 36.50	26.78 / 35.89	28.15 / 37.55
WixQA	58.30 / 80.75	63.33 / 83.58	65.30 / 84.71	66.19 / 86.50	67.68 / 89.12
Long avg.	47.80 / 61.98	52.05 / 65.98	54.22 / 67.97	53.28 / 67.63	**54.95 / 69.11**
DuReader	85.33 / 86.37	85.63 / 86.07	85.48 / 86.02	85.75 / 86.23	86.03 / 86.67
FEVER	93.88 / 96.66	94.54 / 97.66	94.77 / 97.71	95.03 / 97.49	94.93 / 97.42
NQ	88.03 / 96.80	89.58 / 97.92	89.67 / 97.96	89.82 / 97.86	89.91 / 97.81
SQuAD v2	82.77 / 93.56	87.21 / 96.14	87.32 / 96.20	87.01 / 95.85	87.44 / 96.37
Short avg.	87.50 / 93.35	89.24 / 94.45	89.31 / 94.47	89.40 / 94.36	**89.58 / 94.57**

Gains across long-document tasks. With task-specific teachers, StrataDistill improves both metrics over Mixed-task SFT on all five long-document datasets. The nDCG@10 gains range from 0.58 points on MultiDoc2Dial to 4.35 points on WixQA. QASPER gains 4.28 nDCG@10 points and 4.81 Recall@10 points. The long-document macro scores increase by 2.90 nDCG@10 points and 3.13 Recall@10 points. Compared with the shared-teacher student, task-specific teachers further improve the macro scores by 0.73 and 1.14 points, respectively.

Performance on short-document tasks. The task-specific-teacher student improves nDCG@10 over Mixed-task SFT on all four short-document datasets, with gains of 0.23 to 0.40 points. Its macro Recall@10 also increases from 94.45 to 94.57, although NQ decreases by 0.11 points. The shared-teacher student gives smaller macro gains and reduces DuReader nDCG@10 by 0.15 points. Relative to the task-specific SFT retrievers, the student improves both metrics on DuReader and SQuAD v2. Both metrics decrease slightly on FEVER, as does Recall@10 on NQ. The group-level gains therefore coexist with small losses on individual tasks and metrics.