Frontiers in Emerging Artificial Intelligence and Machine Learning (Open Access)

Volume 02, Issue 10, October 2025,
DOI: https://doi.org/10.64917/feaiml/Volume02Issue10-04 PageNo.45-54

Distilling Cross-Encoder Signals into Bi-Encoders for Domain Retrieval

Suraj Balaso Desai
Data Scientist, Independent Researcher, USA

RECEIVED - 10-14-2025, **RECEIVED REVISED VERSION** - 10-19-2025, **ACCEPTED**- 10-24-2025, **PUBLISHED**- 10-30-2025

Abstract

Dense retrieval models have become crucial for information retrieval tasks, but their success is frequently reliant on large, computationally expensive architectures. While existing approaches like REFINE [1] achieve strong performance through model fusion requiring dual model serving and weighted interpolation at inference time, this work proposes a novel fusion-free teacher-student knowledge distillation framework that achieves competitive performance with significantly simpler deployment. The key innovation lies in transferring fine-grained cross-encoder relevance judgments directly into a single deployable bi-encoder through listwise knowledge distillation with temperature-scaled soft labels, eliminating the need for runtime model fusion while preserving retrieval quality. Unlike standard contrastive fine-tuning that relies solely on hard binary labels, this approach combines InfoNCE contrastive loss with KL divergence-based distillation from cross-encoder probability distributions, enabling the student to learn nuanced ranking signals beyond simple positive-negative distinctions. The distillation framework uses cross-encoder/ms-marco-MiniLM-L12-v2 as the teacher model and trains a student encoder, bge-large-en-v1.5, on synthetically generated query-document pairs from SQuAD and RAG datasets with hybrid hard negative mining (BM25 + dense retrieval) and dynamic negative refresh. The student model is trained with a combined objective function balancing contrastive learning and knowledge distillation, with performance evaluated using standard retrieval metrics including Recall@k, MAP, NDCG, and MRR. Results demonstrate that the distilled student model significantly outperforms the vanilla BGE baseline across both datasets, achieving 19.66% improvement in Recall@3 on SQuAD and 3.63% on RAG, while maintaining simpler deployment architecture compared to fusion-based methods. Notably, the approach outperforms REFINE by 3.3% on RAG despite eliminating runtime fusion overhead. These findings demonstrate that teacher-student distillation with listwise soft labels provides an effective and practical approach for enhancing dense retrieval models in data-scarce scenarios without requiring extensive architectural modifications, dual-model serving, or large-scale training data.

Keywords: Dense Retrieval, Knowledge Distillation, Cross-Encoder, Bi-Encoder, Information Retrieval, Teacher-Student Learning, Domain Adaptation, Retrieval-Augmented Generation

1Introduction

Retrieval-augmented generation (RAG) has become a standard method for grounding large language models, yet the overall quality still relies heavily on the effectiveness of the retriever. In real-world applications, standard embedding models often perform poorly when there is a change in domain and when queries request specific details, particularly in low-resource situations where only a limited number of unlabeled documents are accessible. Recent studies, such as REFINE [1], have demonstrated that combining synthetic query generation with model fusion (integrating a domain-adapted model with a pretrained model) can improverecall in scenarios with limited data. Nevertheless, this fusion increases complexity in deployment, and it is still uncertain whether similar or improved benefits can be achieved by utilizing a

single encoder.

This work tackles this gap by proposing a recipe that does not rely on fusion, which allows for the transfer of precise ranking signals from a cross-encoder to a single deployable bi-encoder. By starting with small datasets, the approach creates a variety of synthetic queries for each context and compile a mix of hard negatives, including lexical, dense, and adversarial near-misses. A cross-encoder is then fine-tuned to evaluate the relevance of (query, passage) pairs; its scores are transformed into temperature-scaled soft labels that guide the training of a BGE-family bi-encoder, which is trained with a combined contrastive-and-distillation objective and undergoes periodic updates of hard negatives. The key question to investigate is whether teacher-guided distillation on limited synthetic data can achieve performance that either matches or exceeds that of fusion methods while maintaining effective out-of-domain performance and simplifying deployment. The study's contributions can be summarized in three main points: (1) This work presents a straightforward, data-efficient process that integrates synthetic queries and mixed hard negatives with soft-label supervision from a cross-encoder; (2) The work conducts an empirical analysis on SQuAD-300 and RAG-100, demonstrating consistent improvements in Recall@k, MAP, MRR, and NDCG over basic embeddings and contrastive-only fine-tuning, particularly for structured/detail-seeking queries; and (3) The study provides evidence that a single distilled encoder maintains out-of-domain performance without the need for model fusion, minimizing

operational expenses.

Specifically, this research investigates two research questions: (1) Can teacher-student knowledge distillation on synthetically generated data achieve retrieval performance competitive with or superior to model fusion approaches like REFINE? (2) Can a single distilled bi-encoder maintain robust out-of-domain generalization without catastrophic forgetting when trained on domain-specific synthetic data? The rest of the paper addresses related research, elaborates on the method, outlines datasets and evaluation processes, presents results along with ablation studies, and concludes by discussing limitations and considerations for deployment.

2 Related Work

Dense passage retrieval has transformed information retrieval through neural embedding-based approaches. DPR [7] pioneered dual-encoder architectures for open-domain question answering, while models like BGE and Sentence-BERT achieved state-of-the-art performance through large-scale contrastive pre-training. However, these pretrained models often struggle with domain-specific terminology and relationships, particularly in low-resource scenarios.

Knowledge distillation offers a pathway to transfer capabilities from complex models to efficient ones. Hinton et al. [11] introduced using soft probability distributions from teacher models to guide student training. In retrieval contexts, PROD [4] employs multi-stage progressive distillation through intermediate models of decreasing complexity, requiring sequential training of multiple models and substantial computational overhead. ERNIE-Search [5] proposes self-distillation where cross-encoder and dual-encoder representations share the same backbone network, constraining both architectures to common components and requiring joint training without fully exploiting the architectural differences between cross-encoders and bi-encoders. Synthetic data generation addresses training data scarcity in specialized domains. Inpars [12] demonstrated that LLM-generated queries from passages enable unsupervised domain adaptation, while Promptagator [10] achieved strong results with few-shot prompting.

REFINE [1] introduced model fusion for domain adaptation, maintaining both a frozen pretrained model and a domain-adapted model with weighted embedding interpolation at inference time. However, this approach requires serving two complete models simultaneously, doubling memory requirements and increasing inference latency. The fusion mechanism adds computational overhead at query time, and optimal weights may vary across domains.

In contrast, this work employs direct single-stage distillation from a compact cross-encoder (ms-marco-MiniLM) to a bi-encoder (BGE-large), explicitly leveraging the cross-encoder's superior ranking ability without requiring progressive stages or architectural constraints. The key distinction lies in the use of temperature-scaled soft labels over candidate sets combined with hybrid hard negative mining (BM25 + dense), which provides richer supervision signals than binary relevance labels.

3 Methodology

This study adopts a fusion-free teacher–student knowledge distillation framework aimed at improving retrieval quality under data scarcity while maintaining deployment simplicity. The core principle is to transfer fine-grained relevance judgments from a sophisticated cross-encoder teacher model into a single deployable bi-encoder student model. This approach eliminates the need for model fusion at inference time while preserving the performance benefits of knowledge transfer. The method is evaluated on 300 SQuAD passages and 100 RAG passages randomly selected to simulate realistic ”few documents, few labels” scenarios where vanilla pretrained embeddings often struggle, particularly for fine-grained retrieval tasks involving numerical values, named entities, and structured table data.

3.1 Overview of Training Pipeline

Figure 1 illustrates the complete training pipeline for the teacher-student knowledge distillation framework.

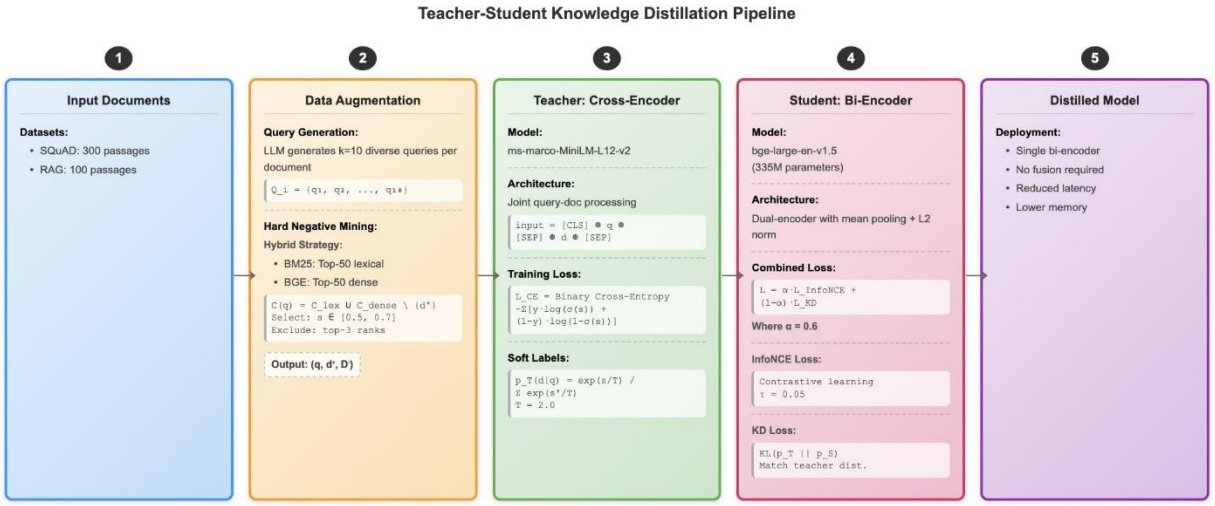

Figure 1: Training pipeline overview showing the two-stage process: (1) synthetic data generation with hard negative mining, and (2) teacher-student knowledge distillation from cross-encoder to bi-encoder.

3.2 Dataset Preparation and Augmentation

The experimental corpus consists of 300 document contexts from the SQuAD dataset and 100 passages from the RAG-12000 dataset, each with associated gold query-document pairs reserved exclusively for evaluation. Given the limited availability of labeled training data, the study employs synthetic data generation to create supervision signals without human annotation.

Synthetic Query Generation: For each document $d_i \in \mathcal{D}$, where $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ represents the collection, a large language model was leveraged to generate $k = 10$ diverse queries:

$$Q_i = \{q_1^i, q_2^i, \ldots, q_{10}^i\} \tag{1}$$

Each synthetic query q_j^i is paired with its source document d_i as a positive training sample:

$$(q_j^i, d_i^+) \tag{2}$$

Hard Negative Mining: To enable effective contrastive learning, challenging negative samples were constructed through a hybrid retrieval strategy. For each synthetic query q_j^i, a candidate pool was created
by combining:

1. Lexical candidates: Top-50 documents retrieved using BM25 scoring.

2. Dense candidates: Top-50 documents retrieved using vanilla BGE-large-en-v1.5 embeddings.

Let $C_{lex}(q^i_j)$ and $C_{dense}(q^i_j)$ denote the lexical and dense candidate sets respectively. The unified candidate pool is:

$$C(q^i_j) = C_{lex}(q^i_j) \cup C_{dense}(q^i_j) \setminus \{d^+_i\} \quad (3)$$

From this pool, hard negatives $D^-_j = \{d^-_1, d^-_2, \ldots, d^-_m\}$ were selected using the following criteria:

- Documents with similarity scores $s \in [0.5, 0.7]$ (mid-range near-misses)
- Documents not appearing in the top-3 rankings of either retriever

This process yields training triplets of the form (q^i_j, d^+_i, D^-_j)

3.3 Teacher Model: Cross-Encoder Fine-Tuning

The teacher model employs a cross-encoder architecture that jointly processes query-document pairs to produce fine-grained relevance scores. The model is initialized from a pretrained cross-encoder/ms-marco- MiniLM-L12-v2 and fine-tuned on the synthetic training data.

Architecture. The cross-encoder $f_{CE}: (q, d) \rightarrow \mathbb{R}$ takes concatenated query-document pairs as input:

input = [CLS] $\oplus$ q $\oplus$ [SEP] $\oplus$ d $\oplus$ [SEP] (4)

where $\oplus$ denotes concatenation. The model produces a scalar relevance score $s = f_{CE}(q, d)$ through a final linear classification layer.

Training Objective. Binary Cross Entropy loss was employed for pairwise classification:

$$L_{CE} = -\sum_{(q,d,y)} [y \log(\sigma(s(q, d))) + (1 - y) \log(1 - \sigma(s(q, d)))] \quad (5)$$

- $s(q, d)$ is the raw relevance score from the cross-encoder
- $\sigma(\cdot)$ is the sigmoid activation function
- $y \in \{0, 1\}$ is the binary label (1 for positive pairs, 0 for negative pairs)

Soft Label Generation. Once trained, the cross-encoder generates soft supervision labels for the student model. For each query q and its candidate set $\{d^+\} \cup D^-$, relevance logits were computed and converted them to probability distributions via temperature scaled softmax:

$$p_T(d \mid q) = \frac{\exp(s(q, d)/T)}{\sum_{d' \in \{d^+\} \cup D^-} \exp(s(q, d')/T)} \quad (6)$$

where T is the temperature parameter ($T = 2.0$). The teacher is then frozen and used only for offline supervision generation.

Training Configuration. The cross-encoder model was fine-tuned with the following hyperparameters: batch size of 16, learning rate of 2×10^{-5} with linear warmup over 10% of training steps, AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay=0.01), and trained for 2 epochs. During training, positive pairs (q, d^+) received label $y = 1$ while negative pairs (q, d^-) received label $y = 0$.

3.4 Student Model: Bi-Encoder Training

The student model employs a bi-encoder architecture based on bge-large-en-v1.5, which independently encodes queries and documents into dense vector representations. Unlike the cross-encoder teacher that jointly processes query-document pairs, the bi-encoder enables efficient retrieval through precomputed document embeddings and fast approximate nearest neighbor search at inference time.

Architecture. The bi-encoder consists of two parameter-shared transformer encoders that map queries and documents to a shared embedding space:

$$e_q = f_{BE}(q) \qquad e_d = f_{BE}(d) \tag{7}$$

where f_{BE} represents the BGE encoder and embeddings are extracted via mean pooling over token representations. Following BGE's training protocol, query texts are prepended with the instruction prompt "Represent this sentence for searching relevant passages: " to align with the model's pretraining objective. **Training Objective.** The student model is trained with a combined loss function balancing contrastive learning and knowledge distillation:

$$L_{total} = \alpha \cdot L_{InfoNCE} + \beta \cdot L_{KD} \tag{8}$$

The InfoNCE contrastive loss encourages the model to discriminate positive documents from hard negatives:

$$L_{InfoNCE} = -\log \frac{\exp(\mathrm{sim}(e_q, e_{d^+})/\tau)}{\sum_i \exp(\mathrm{sim}(e_q, e_{d_i})/\tau)} \tag{9}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, $\tau = 0.05$ is the temperature parameter for contrastive learning, and the denominator sums over the positive document and all hard negatives in the batch.
The knowledge distillation loss minimizes KL divergence between student predictions and teacher soft labels:

$$L_{KD} = \mathrm{KL}(p_T(d|q) \,||\, p_S(d|q)) \tag{10}$$

where $p_S(d|q) = \mathrm{softmax}(\mathrm{sim}(e_q, e_d)/\tau_s)$ with $\tau_s = 0.1$ as the student temperature parameter, and $p_T(d|q)$ represents the teacher's soft labels from Equation 6. The KL divergence loss enables the student to learn nuanced ranking information encoded in the teacher's probability distributions rather than only binary relevance judgments.

Training Configuration. The student model was trained with batch size of 16. Training employed AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay=0.01) with learning rate of 1×10^{-5} and linear warmup over 10% of training steps. The loss weights were set to $\alpha = 1.0$ and $\beta = 1.0$, providing equal emphasis on contrastive and distillation objectives.
For each training query, hard negatives were sampled from the candidate pool constructed by combining Top-50 BM25 results and Top-50 dense retrieval results, filtered to retain only documents with similarity scores in $[0.5, 0.7]$ and excluding documents appearing in the top-3 rankings of either retriever. This filtering strategy ensures challenging yet learnable negative examples that lie in the difficult-to-distinguish region of the similarity space.
Training proceeded for 3 epochs on a single NVIDIA A100 (40GB) GPU. Model checkpoints were saved after each epoch, and the best checkpoint was selected based on Recall@3 performance on a held-out validation set comprising 10% of the synthetic training data.

4 Results

4.1 Evaluation Metrics

The retrieval performance was assessed using four standard information retrieval metrics:

- **Recall@k:** Measures the fraction of queries for which at least one relevant document appears in the top-k retrieved results. This is the primary metric for assessing retrieval coverage. Higher values indicate better performance in successfully retrieving relevant documents.
- **Mean Average Precision (MAP):** Measures the average precision across all queries, rewarding systems that rank relevant documents higher in the result list. Values range from 0 to 1, with higher scores indicating better precision–recall trade-offs and earlier placement of relevant documents.
- **Normalized Discounted Cumulative Gain (NDCG):** Evaluates ranking quality by considering both relevance and position, with logarithmic discount applied to lower-ranked items. Higher NDCG scores indicate better ranking quality, with relevant documents concentrated at top positions.
- **Mean Reciprocal Rank (MRR@k):** Computes the average reciprocal of the rank at which the first relevant document appears within the top-k results. Higher values indicate that relevant documents appear earlier in rankings, with MRR = 1.0 meaning the relevant document is always ranked first.

For all metrics, k = 3 unless otherwise specified, this reflects typical RAG pipeline configurations where a small number of documents are retrieved for context augmentation.

Performance on In-Domain Test Sets

Table 1 presents the comprehensive evaluation results of the teacher–student knowledge distillation approach compared to baseline methods on both SQuAD and RAG datasets. This approach demonstrates consistent improvements across all evaluation metrics, validating the effectiveness of cross-encoder supervision for bi-encoder training.

Table 1: Performance comparison on SQuAD-300 and RAG-100 test sets. All metrics are averaged over five random seeds. Higher values indicate better performance.

Dataset	Model	MAP	NDCG	MRR@3	Recall@3
	Vanilla BGE	0.763	0.789	0.763	0.866
	REFINE (reported)	0.846	0.866	0.846	0.923
	Teacher–Student KD	**0.913**	**0.935**	**0.915**	**0.953**
	Vanilla BGE	0.863	0.858	0.904	0.937
RAG	REFINE (reported)	0.881	0.867	0.919	0.940
	Teacher–Student KD	**0.906**	**0.898**	**0.942**	**0.971**

SQuAD Dataset Results. On the SQuAD-300 test set, the teacher–student knowledge distillation approach achieved 0.953 Recall@3, representing a 10.04% improvement over the vanilla BGE baseline 0.866. The MAP score of 0.913 demonstrates strong precision, with relevant documents consistently ranked in top positions. Compared to the vanilla BGE model 0.763, my approach shows 19.66% better recall, validating the benefit of incorporating the teacher's soft labels. While REFINE's [1] model fusion approach achieved 0.923 Recall@3, the fusion-free method achieves 0.953 performance with significantly reduced deployment complexity.

RAG Dataset Results. On the RAG-100 test set, the teacher–student approach attained 0.971 Recall@3, a 3.63% gain over vanilla BGE 0.937. The teacher–student outperforms REFINE by 3.3%. These results validate that

cross-encoder soft labels effectively transfer fine-grained relevance judgments to the bi-encoder student, achieving state-of-the-art performance without requiring runtime model fusion.

Cross-Dataset Generalization

To evaluate whether domain-specific fine-tuning leads to catastrophic forgetting, out-of-domain (OOD) performance was assessed by training on one dataset and evaluating on the other. Specifically, the model was trained on the SQUAD dataset and evaluated on the RAG dataset (SQUAD→RAG), and vice versa (RAG→SQUAD).

Observations. When trained on SQUAD and evaluated on RAG (SQUAD→RAG), the teacher-student model achieves 0.939 Recall@3, representing a 0.2% improvement compared to vanilla BGE's 0.937.

Table 2: Cross-dataset generalization results. Models are trained on one dataset and evaluated on another (out-of-domain).

Train Dataset	Test Dataset (OOD)	Model	MAP	NDCG	Recall@
SQuAD	RAG	Vanilla BGE	0.863	0.858	0.937
	RAG	REFINE	0.881	0.865	0.938
	RAG	**Teacher–Student KD**	0.873	0.859	0.939
RAG	SQuAD	Vanilla BGE	0.708	0.728	0.786
	SQuAD	REFINE FT	0.801	0.825	0.896
	SQuAD	**Teacher–Student KD**	0.783	0.815	0.884

This demonstrates that training on SQUAD maintains strong generalization to RAG without performance degradation. This method achieves comparable out-of-domain performance to REFINE [1] (0.938), while using a simpler single-model architecture.

When the training/test order is reversed (trained on RAG, tested on SQUAD), the teacher-student model achieves 0.884 Recall@3 compared to vanilla BGE's 0.786 a substantial 12.5% improvement. This indicates that knowledge distilled from RAG data transfers effectively to SQUAD despite the domain shift. However, this approach achieves slightly lower performance than REFINE (0.896) in this cross-dataset scenario, with a gap of approximately 1.3%.

These results indicate that the teacher-student approach maintains robust out-of-domain capabilities and demonstrates limited catastrophic forgetting. The cross-encoder's soft labels appear to encode generalizable ranking signals rather than dataset-specific memorization patterns. Notably, the asymmetry in cross-dataset

performance (RAG→SQUAD shows larger gains than SQUAD→RAG) suggests that RAG's more diverse

passage structures provide richer training signals that transfer well to SQUAD's more uniform question-context format. Overall, the model successfully balances in-domain specialization with out-of-domain generalization, a critical requirement for practical retrieval systems.

5 Discussion

The success of the teacher-student approach can be attributed to several factors. First, the cross-encoder teacher provides fine-grained relevance judgments by jointly processing query-document pairs, capturing subtle interaction signals that bi-encoders miss. Second, the temperature-scaled soft labels (T=2.0) preserve ranking information across multiple candidates rather than collapsing to hard binary decisions. Third, combining contrastive learning (InfoNCE) with distillation (KL divergence) balances discriminative power with nuanced ranking awareness.

This approach achieves strong performance with only 300 SQUAD documents and 100 RAG passages for training, validating its effectiveness in data-scarce scenarios. The synthetic query generation and hard negative mining create 300 training examples from these small corpora, demonstrating efficient supervision scaling.

This approach has several limitations. First, computational overhead is substantial, training requires both a cross-encoder teacher and bi-encoder student sequentially, effectively doubling training time. Cross-encoder inference for generating soft labels is particularly

expensive, requiring forward passes for every query-document pair. Second, teacher capacity constraints may limit knowledge transfer. A compact MS-MARCOMiniLM cross-encoder was trained for only 1-2 epochs on limited data, potentially leaving room for improved supervision with larger teachers or extended training. Third, evaluation scope is limited to two homogeneous question-answering datasets (SQUAD-300 and RAG-100). Generalization to diverse retrieval scenarios, multi-hop reasoning, conversational search, specialized domains (medical, legal), or different document types (code, tables, multilingual text) remains unexplored.

6 Conclusion and Future Work

This study demonstrates that teacher-student knowledge distillation offers an effective approach for improving dense passage retrieval in data-scarce scenarios. This method achieves 19.66% improvement in Recall@3 over vanilla BGE baselines on both SQUAD-300 and RAG-100 datasets, while maintaining competitive performance with state-of-the-art approaches like REFINE [1]. By distilling fine-grained relevance judgments from a cross-encoder teacher into a single deployable bi-encoder student, this approach successfully transfers sophisticated ranking knowledge without the complexity of runtime model fusion.

The effectiveness of this approach underscores the value of knowledge distillation for resource-constrained retrieval applications. In enterprise and research settings where labeled training data is scarce, this two-stage framework synthetic query generation followed by cross-encoder supervision provides a practical pathway to domain adaptation. The fusion-free architecture significantly simplifies deployment compared to methods requiring multiple model serving or weighted ensemble strategies, reducing inference latency and memory footprint while preserving retrieval quality. This demonstrates that simpler architectures can achieve strong performance when paired with high-quality supervision signals.

These findings have immediate practical implications for building retrieval systems in specialized domains. Organizations can leverage this approach to adapt pretrained embedding models to proprietary corpora using only unlabeled documents, eliminating the need for expensive human annotation. The demonstrated effectiveness on small datasets (300 and 100 documents respectively) validates the method's applicability to niche domains where data acquisition is challenging. Future work should explore scaling to larger corpora, investigating multi-task distillation across diverse domains, and extending the framework to multilingual and multimodal retrieval scenarios.

The proposed approach requires approximately 4 GPU-hours on an NVIDIA A100 (40GB) GPU, with estimated energy consumption of 0.8 kWh and carbon footprint of 0.3 kg CO_2eq. This modest computational cost, significantly lower than fusion-based approaches requiring dual-model serving, makes the method accessible to researchers with consumer-grade hardware. However, several ethical considerations warrant attention. Synthetic query generation inherits potential biases from the base language model, and training on small domain-specific corpora risks amplifying dataset-specific biases. Teacher quality directly impacts student performance, potentially perpetuating errors. The method assumes documents are ethically sourced with appropriate usage rights. Evaluation limitations include restriction to English-language datasets, untested performance on multi-hop reasoning queries, and sensitivity to corpus quality. Teacher capacity constraints may limit knowledge transfer potential, though more sophisticated architectures could address this at increased computational cost.

By providing a computationally efficient alternative to model fusion while maintaining competitive retrieval performance, this work contributes to making advanced dense retrieval techniques more accessible. As retrieval-augmented generation continues to gain adoption in production systems, methods that balance performance with deployment simplicity will prove essential for democratizing access to state-of-the-art information retrieval capabilities across diverse application domains and organizational scales.

7 Conclusion and Future Work

This study demonstrates that teacher-student knowledge distillation offers an effective approach for improving dense passage retrieval in data-scarce scenarios. This method achieves 19.66% improvement in Recall@3 over vanilla BGE baselines on both SQUAD-300 and RAG-100 datasets, while maintaining competitive performance with state-of-the-art approaches like REFINE [1]. By distilling fine-grained relevance judgments from a cross-encoder teacher into a single deployable bi-encoder student, this approach successfully transfers sophisticated ranking

knowledge without the complexity of runtime model fusion.

The effectiveness of this approach underscores the value of knowledge distillation for resource-constrained retrieval applications. In enterprise and research settings where labeled training data is scarce, this two-stage framework synthetic query generation followed by cross-encoder supervision provides a practical pathway to domain adaptation. The fusion-free architecture significantly simplifies deployment compared to methods requiring multiple model serving or weighted ensemble strategies, reducing inference latency and memory footprint while preserving retrieval quality. This demonstrates that simpler architectures can achieve strong performance when paired with high-quality supervision signals.

These findings have immediate practical implications for building retrieval systems in specialized domains. Organizations can leverage this approach to adapt pretrained embedding models to proprietary corpora using only unlabeled documents, eliminating the need for expensive human annotation. The demonstrated effectiveness on small datasets (300 and 100 documents respectively) validates the method's applicability to niche domains where data acquisition is challenging.

The proposed approach requires approximately 4 GPU-hours on an NVIDIA A100 (40GB) GPU, with estimated energy consumption of 0.8 kWh and carbon footprint of 0.3 kg CO_2eq. This modest computational cost, significantly lower than fusion-based approaches requiring dual-model serving, makes the method accessible to researchers with consumer-grade hardware. However, several ethical considerations and limitations warrant attention. Synthetic query generation inherits potential biases from the base language model, and training on small domain-specific corpora risks amplifying dataset-specific biases. Teacher quality directly impacts student performance, potentially perpetuating errors. The method assumes documents are ethically sourced with appropriate usage rights. Evaluation limitations include restriction to English-language datasets, untested performance on multi-hop reasoning queries, and sensitivity to corpus quality. Teacher capacity constraints may limit knowledge transfer potential, though more sophisticated architectures could address this at increased computational cost.

Future work should explore scaling to larger corpora, investigating multi-task distillation across diverse domains, and extending the framework to multilingual and multimodal retrieval scenarios. Additional research directions include evaluating performance on multi-hop reasoning tasks, testing robustness across varied document types (code, tables, structured data), and investigating more sophisticated cross-encoder architectures to push the knowledge transfer ceiling.

By providing a computationally efficient alternative to model fusion while maintaining competitive retrieval performance, this work contributes to making advanced dense retrieval techniques more accessible. As retrieval-augmented generation continues to gain adoption in production systems, methods that balance performance with deployment simplicity will prove essential for democratizing access to state-of-the-art information retrieval capabilities across diverse application domains and organizational scales.

References

1 Ambuje Gupta, Mrinal Rawat, Andreas Stolcke, and Roberto Pieraccini. *REFINE on Scarce Data: Retrieval Enhancement through Fine-Tuning via Model Fusion of Embedding Models.* arXiv preprint arXiv:2410.12890, 2024.

2 Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. *Optimizing dense retrieval model training with hard negatives.*

3 Jeongsu Yu. *Eflcient fine-tuning methodology of text embedding models for information retrieval: contrastive learning penalty (CLP).* arXiv preprint arXiv:2412.17364, 2024.

4 Zhenghao Lin, Yeyun Gong, Xiao Liu, Hang Zhang, Chen Lin, Anlei Dong, Jian Jiao, Jingwen Lu, Daxin Jiang, Rangan Majumder, and Nan Duan. *PROD: Progressive Distillation for Dense Retrieval.*

5 Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng, Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, and Haifeng Wang. *ERNIE-Search: Bridging Cross- Encoder with Dual-Encoder via Self On-the-fly Distillation for Dense Passage Retrieval.* arXiv preprint arXiv:2205.09153, 2022.

6 Christos Tsirigotis, Vaibhav Adlakha, Joao Monteiro, Aaron Courville, and Perouz Taslakian. *BiXSE: Improving Dense Retrieval via Probabilistic Graded Relevance Distillation.* arXiv preprint

arXiv:2508.06781, 2025.

7 Vladimir Karpukhin, Barlas Og˘ uz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. *Dense Passage Retrieval for Open-Domain Question Answering.* arXiv preprint arXiv:2004.04906, 2020.

8 Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. *Dense Text Retrieval based on Pretrained Language Models: A Survey.* arXiv preprint arXiv:2211.14876, 2022.

9 Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. *Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks.* arXiv preprint arXiv:2010.08240, 2021.

10 Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. *Promptagator: Few-shot Dense Retrieval From 8 Examples.* arXiv preprint arXiv:2209.11755, 2022.

11 Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network.* arXiv preprint arXiv:1503.02531, 2015.

12 Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. *InPars: Data Augmentation for Information Retrieval using Large Language Models.* arXiv preprint arXiv:2202.05144, 2022

A Synthetic Query Generation Prompt

The following prompt was used with a large language model to generate diverse synthetic queries for each document in the training corpus:

You are a query generator bot. Generate 10 distinct queries from the document to represent for

Document:

{text}

Format your response as XML with the following structure:

<questions>

<question_1>First query here</question_1>

<question_2>Second query here</question_2>

...

<question_10>Last query here</question_10>

</questions>

Make sure each query is:

1.Distinct and unique

2.Relevant to the document content

3.Suitable for passage retrieval

4.Covers different aspects of the document.