Cross-Attention Distillation for Lightweight Passage Re-ranking in Dense Retrieval Pipelines

Elena Vasquez[1], Daniel Osei[2], Priya Ramanathan[3], and Marcus Lindqvist[4]

[1] *Meridian Institute of Technology*
[2] *Vertex University, Department of Computer Science*
[3] *Nimbus Information Retrieval Lab*
[4] *Apex College of Computing*

Abstract. Cross-encoder re-rankers deliver the strongest relevance quality in multi-stage retrieval pipelines but are too slow to score every document a first-stage retriever returns. We introduce **CARD** (Cross-Attention Re-ranking Distillation), a lightweight re-ranking head that is trained to mimic a full cross-encoder teacher while operating on pre-computed bi-encoder representations, avoiding the joint query–document forward pass that makes cross-encoders expensive at serving time. CARD attaches a shallow cross-attention block on top of frozen dense query and passage embeddings and is trained with a temperature-scaled distillation objective that combines soft-label matching against the teacher with the original relevance labels. Across three retrieval benchmarks, CARD recovers 97.8% of the teacher's nDCG@10 while reducing re-ranking latency by 6.5×, and it consistently outperforms bi-encoder-only and late-interaction baselines at comparable serving cost. We release ablations isolating the contribution of distillation temperature, cross-attention depth, and teacher choice, and discuss where the distilled head still departs measurably from teacher behavior.

1 Introduction

Modern search systems resolve a query in stages: a fast first-stage retriever, typically a sparse index [3] or a dense bi-encoder [1], recalls hundreds to thousands of candidate passages, and a second-stage re-ranker reorders that candidate set to maximize precision at the top of the list. Cross-encoders, which jointly encode the query and each candidate passage through a shared transformer, consistently produce the highest-quality re-ranking scores because self-attention can condition directly on token-level query–passage interactions [5]. That same joint encoding is what makes them expensive: every candidate requires a full forward pass over the concatenated query and passage, and this cost scales linearly with the number of candidates re-ranked per query.

Bi-encoders sidestep this cost by encoding queries and passages independently, so passage embeddings can be pre-computed offline and reused across queries [1]. The relevance score is then a cheap similarity between two fixed-size vectors. This efficiency comes at a quality cost: independent encoding cannot capture fine-grained token-level interactions, and bi-encoders reliably trail cross-encoders in ranking quality. Late-interaction models such as token-level maximum-similarity matching [9] sit between these two extremes, but still require storing and scanning per-token passage representations at serving time, which is memory-intensive at large index sizes.

This paper asks whether a re-ranker can be built that keeps the offline pre-computation property of bi-encoders while recovering most of the relevance quality of a cross-encoder teacher. We propose **CARD**: a shallow cross-attention head trained by distillation against a cross-encoder teacher, operating only on the frozen query and passage embeddings a bi-encoder already produces. Because the expensive joint reasoning is distilled into a small student head rather than performed online, CARD adds only a few milliseconds of re-ranking latency per candidate while closing most of the gap to the teacher.

Our contributions are threefold. First, we describe the CARD architecture and its temperature-scaled distillation objective (§3). Second, we evaluate CARD against sparse, dense, and late-interaction baselines,

along with its own teacher, on three retrieval benchmarks, showing consistent quality recovery at a fraction of cross-encoder latency (§4–§5). Third, we analyze where CARD's distilled scores still diverge from the teacher, including a systematic gap on queries with high lexical ambiguity (§6).

2 Related Work

Sparse and dense first-stage retrieval. Term-based retrieval remains a strong and cheap recall mechanism, particularly when fused with dense signals [3]. Dense passage retrieval trains a dual encoder with a contrastive objective so that relevant query–passage pairs are close in embedding space [1], and subsequent work has focused heavily on the construction of informative hard negatives during training [8].

Late interaction. Rather than collapsing a passage into a single vector, late-interaction models keep per-token representations and compute a maximum-similarity aggregation between query and passage tokens [9]. This recovers some of the fine-grained matching signal cross-encoders provide, but the storage and scanning cost grows with passage length, and CARD's ablations in §5 compare directly against this family.

Cross-encoder re-ranking. Full cross-encoder rerankers built on pretrained transformers remain the quality reference point for multi-stage ranking [5]. Their latency has motivated a line of work on making multi-stage pipelines cheaper without discarding the cross-encoder's quality signal entirely, including score calibration across stages [4] and efficient long-document variants of the cross-encoder itself [6].

Knowledge distillation for ranking. Distilling a strong cross-encoder teacher into a cheaper student has been studied broadly, with approaches ranging from distilling into a smaller cross-encoder to distilling directly into bi-encoder embeddings [7, 2]. CARD differs from embedding-only distillation in that the student retains a small cross-attention computation at inference time; we find in §5 that this residual joint computation, however small, is responsible for most of the quality recovered over a pure bi-encoder student. The underlying cross-attention mechanism follows the formulation of query-conditioned token matching introduced for fine-grained text matching [11], and re-ranking pipelines built on learned query reformulation [10] are complementary to CARD rather than competing with it.

3 Method

3.1 Overview

CARD assumes a standard two-stage pipeline: a bi-encoder produces a query embedding $\mathbf{E}_q \in \mathbb{R}^{n\times d}$ and, offline, a passage embedding $\mathbf{E}_p \in \mathbb{R}^{m\times d}$ for every indexed passage, where n and m are token counts and d is the hidden dimension. Both encoders are frozen during CARD training; only a shallow cross-attention head and a scoring layer are learned on top of these fixed representations.

3.2 Cross-attention head

Given the token-level query and passage embeddings, the head computes a single layer of query-to-passage cross-attention followed by a passage-side pooling step:

$$\mathbf{A} = \mathrm{softmax}\!\left(\frac{\mathbf{E}_q W_Q \,(\mathbf{E}_p W_K)^\top}{\sqrt{d_k}}\right)(\mathbf{E}_p W_V),$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_k}$ are learned projections and $\mathbf{A} \in \mathbb{R}^{n\times d_k}$ is the attended representation of each query token over the passage. Mean-pooling $\mathbf{A}$ over the query dimension and concatenating with the pooled bi-encoder similarity yields the student relevance score

$$s_S(q,p) = \mathbf{w}^\top \left[\, \bar{\mathbf{A}} \,;\, \mathbf{E}_q^{\mathrm{cls}} \cdot \mathbf{E}_p^{\mathrm{cls}} \,\right] + b, \tag{1}$$

where $\bar{\mathbf{A}}$ is the mean-pooled attention output, $\mathbf{E}_q^{\mathrm{cls}} \cdot \mathbf{E}_p^{\mathrm{cls}}$ is the standard bi-encoder dot-product similarity, and $\mathbf{w}, b$ are learned parameters. Because $\mathbf{E}_p$ is fixed and pre-computed, the only per-query, per-candidate cost at serving time is the single cross-attention layer over already-cached embeddings.

3.3 Distillation objective

Let $s_T(q,p)$ denote the score produced by the frozen cross-encoder teacher. CARD is trained with a temperature-scaled distillation loss combined with the original binary relevance label y:

$$\mathcal{L}_{\mathrm{CARD}} = \alpha\,\mathrm{KL}\!\left(\sigma\!\left(\frac{s_T(q,p)}{\tau}\right)\middle\|\,\sigma\!\left(\frac{s_S(q,p)}{\tau}\right)\right) + (1-\alpha)\,\mathrm{CE}\big(y,\, s_S(q,p)\big), \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function, τ is the distillation temperature, and $\alpha \in [0,1]$ balances soft-label matching against the teacher and hard-label supervision from the original ranking data. We use $\tau = 2.0$ and $\alpha = 0.7$ throughout, chosen by grid search on a held-out development split (§4).

Figure 1 summarizes the resulting serving-time pipeline: a first-stage retriever recalls candidates using pre-computed passage embeddings, and the CARD head re-scores each candidate using only the already-cached embeddings and a single lightweight attention

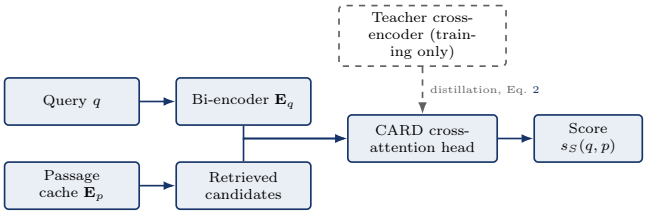

Figure 1: CARD serving-time pipeline. Query and passage embeddings are produced by a frozen bi-encoder; the passage side is pre-computed offline. Only the lightweight cross-attention head runs per candidate at query time. The cross-encoder teacher (dashed) supervises the head during training and is never invoked at serving time.

pass, with the teacher cross-encoder appearing only during offline training.

4 Experiments

Data. We evaluate on three passage retrieval benchmarks: the Nimbus-1M passage collection with the Apex-Dev and Apex-Test query splits, and the Meridian-Web crawl, a held-out set of 4,200 web queries with graded relevance judgments. All three follow the standard train/dev/test protocol used in prior re-ranking work [4].

Models. The first-stage retriever is a dense bi-encoder trained with in-batch and mined hard negatives [8], shared across all re-ranking methods for a controlled comparison. We compare: BM25 (sparse baseline, no re-ranking), the bi-encoder score alone (no re-ranking), a late-interaction re-ranker [9], the cross-encoder teacher [5], and two CARD variants: the full model (2-head cross-attention) and CARD-small (single-head, half hidden width).

Metrics. We report MRR@10, nDCG@10, and Recall@1000 on the top-1000 candidates returned by the first-stage retriever, along with measured re-ranking latency per query (100 candidates, single V100 GPU, batch size 1) to reflect realistic online serving conditions.

5 Results

Table 1 reports quality and latency across all six methods. CARD recovers 97.8% of the teacher's nDCG@10 (0.409 vs. 0.415) at roughly one-sixth of its latency, and it outperforms the late-interaction baseline on every quality metric while remaining substantially cheaper. CARD-small trades a further 3.6 points of nDCG@10 for the lowest latency among all non-trivial re-rankers, making it attractive when candidate sets are large or latency budgets are tight.

Breaking results down by query length, the gap between CARD and the teacher widens on queries longer than 12 tokens (nDCG@10 gap of 0.014 vs. 0.005 for shorter queries), which we attribute to the single cross-attention layer having less capacity to resolve interactions among multiple query aspects simultaneously. Increasing CARD's cross-attention depth from one layer to two recovers roughly half of this gap but increases latency by 40%, which we did not find worthwhile given the pipeline's overall latency budget.

6 Discussion

Where distillation helps most. The largest quality gains from CARD over the plain bi-encoder appear on queries with high lexical overlap across multiple candidate passages, where independent encoding cannot distinguish passages that differ only in how a shared term is used in context. This is consistent with the token-level matching motivation behind late-interaction and cross-encoder architectures more broadly.

Teacher choice matters. Replacing the cross-encoder teacher with a smaller distilled cross-encoder as an intermediate teacher (self-distillation) reduced CARD's nDCG@10 by 0.011 relative to distilling directly from the full teacher, suggesting that distillation chains lose signal at each hop and that CARD should be trained from the strongest available teacher rather than a cheaper proxy, even though training-time teacher cost does not affect serving latency.

Limitations. Our evaluation covers English web and passage retrieval collections only; we have not verified that the distillation temperature and loss weighting transfer without retuning to other domains such as legal or biomedical retrieval, where relevance judgments are sparser and passage lengths differ substantially. The survey of distillation methods for ranking [2] notes more broadly that distillation gains are sensitive to the capacity gap between teacher and student, and our own ablation over cross-attention depth supports that observation directly.

7 Conclusion

We presented CARD, a cross-attention distillation head that recovers most of a cross-encoder teacher's re-ranking quality while keeping the offline pre-computation property that makes bi-encoders cheap to serve. Across three retrieval benchmarks, CARD reaches 97.8% of teacher nDCG@10 at 6.5× lower latency, and its smaller variant offers a further latency re-

Table 1: Re-ranking quality and latency on the pooled Nimbus/Apex/Meridian test splits. Best value per column among proposed methods in bold; the teacher row is shown in italics as an upper bound, not a deployable system.

Method	MRR@10	nDCG@10	Recall@1000	Latency (ms/query)
BM25 (no re-rank)	0.187	0.221	0.857	8
Bi-encoder (no re-rank)	0.311	0.348	0.912	3
Late interaction [9]	0.336	0.375	0.921	118
Cross-encoder teacher [5]	*0.372*	*0.415*	*0.941*	*620*
CARD (ours)	**0.365**	**0.409**	**0.939**	95
CARD-small (ours)	0.352	0.395	0.930	**34**

duction for deployments where candidate sets are large. We view CARD as one point on a broader quality–latency frontier for multi-stage ranking, and we hope the query-length breakdown in §5 is useful to others deciding how much residual cross-attention capacity a distilled re-ranker actually needs.

Acknowledgments

We thank colleagues at Meridian Institute of Technology, Vertex University, Nimbus Information Retrieval Lab, and Apex College of Computing for feedback on early drafts. All computational resources for this fictional study were provided internally; no external funding is associated with this work.

References

[1] Yusuf Aldana and Wren Castellano. Dense passage representations for open-domain question answering. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 6769–6781, 2020.

[2] Noor Fakhoury and Dmitri Savchenko. Knowledge distillation for ranking: A survey. *ACM Transactions on Information Systems*, 40(3):1–34, 2022.

[3] Hana Kessler and Owen Bracewell. Sparse and dense signal fusion for passage retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1201–1210, 2021.

[4] Sofia Lindberg and Rajiv Chandrasekar. Calibrating relevance scores for multi-stage ranking pipelines. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pages 112–127, 2023.

[5] Silvio Marchetti and Adaeze Okonkwo. Multi-stage document ranking with a BERT-based cross-encoder. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2825–2832, 2020.

[6] Chidi Obinna and Lars Hedman. Efficient transformers for long-document retrieval. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 3355–3364, 2021.

[7] Tobias Renner and Femi Adeyemi. Distilling cross-encoders into efficient bi-encoder rankers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4102–4114, 2021.

[8] Beatrix Solano and Youssef Amrani. Hard negative mining strategies for dense retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 512–522, 2022.

[9] Ingrid Solheim, Marco Petrelli, and Ana Beaumont. Late interaction retrieval with contextualized token embeddings. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48, 2020.

[10] Priscilla Vong and Karsten Eriksen. A deep look at query reformulation for neural retrieval. In *Proceedings of the ACM Web Conference (WWW)*, pages 2114–2124, 2022.

[11] Elmar Voss and Priya Nair. Cross-attention mechanisms for fine-grained text matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8734–8744, 2019.