Ferdinand Schlatt Maik Fröbe Harrisen Scells Shengyao Zhuang Bevan Koopman
Guido Zuccon Benno Stein Martin Potthast Matthias Hagen

Rank-DistiLLM: Closing the Effectiveness Gap Between Cross-Encoders and LLMs for Passage Re-Ranking

Cross-Encoders vs LLM Re-Rankers

Cross-Encoder (MS MARCO) < Cross-Encoder (Distilled from LLM) < LLM Re-Ranker

Rank-DistiLLM

➜ We propose a new dataset, Rank-DistiLLM, to close the effectiveness gap between cross-encoders and LLM re-rankers

Insights from MS MARCO Fine-Tuning

Effective fine-tuning using MS MARCO uses three strategies:

(1) Hard-negatives (2) Deep sampling (3) Listwise loss

Previous LLM distillation datasets and studies have not considered these strategies. We apply (1) and (2) to Rank-DistiLLM and propose a new listwise loss for (3)

Experimental Setup

Rank-DistiLLM Dataset

- Re-rank top $k = 100$ passages from BM25 and ColBERTv2 for 10k queries using RankZephyr
  ➜ BM25 (Easy-negatives) vs ColBERTv2 (Hard-Negatives)
- Sub-sample for $k \in \{10, 25, 50\}$
  ➜ Evaluate different sampling depths

Model

- Distill RankZephyr ranking into pointwise monoELECTRA

Listwise Loss

LLM distillation usually uses the pairwise RankNet loss

$$\mathcal{L}_{\text{RankNet}} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \log(1 + \exp(s_{d_j} - s_{d_i}))$$

We propose a listwise Approximate Discounted Rank MSE loss

$$\mathcal{L}_{\text{ADR-MSE}} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\log_2(i+1)} (i - \hat{\pi}(d_i))^2$$

s_{d_i} is the relevance score and $\hat{\pi}(d_i)$ is the approximated rank of passage d_i

Main Insights

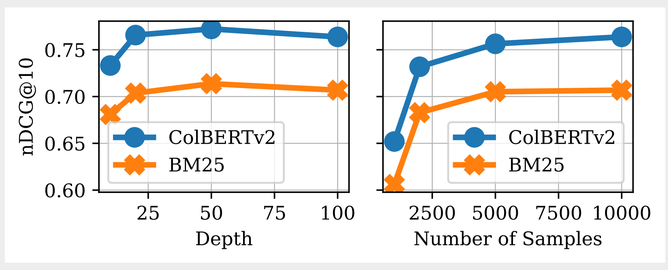
Hard-Negatives

nDCG@10 on TREC DL '19 / '20

Negatives	BM25		ColBERTv2	
	DL 19	DL 20	DL 19	DL 20
BM25	0.644	0.622	0.674	0.654
CBv2	0.709	0.704	0.774	0.754

➜ Hard-negatives are crucial

Deep Sampling



➜ 10k queries & 50 passages/query suffice

Listwise Loss

nDCG@10 on TREC DL '19 / '20

Loss	BM25		ColBERTv2	
	DL 19	DL 20	DL 19	DL 20
RankNet	0.720	0.711	0.768	0.770
ADR-MSE	0.716	0.709	0.770	0.765

➜ Listwise loss unnecessary

Comparing LLM Distillation Datasets

1. Rank-DistiLLM (RDL) outperforms existing datasets [1,2]
2. Hard-negative mining with manual labels is still competitive

nDCG@10 on TREC DL '19 / '20; † denotes $p < 0.05$ compared to last row

Dataset	BM25		ColBERTv2	
	DL 19	DL 20	DL 19	DL 20
MS MARCO	0.687	0.698	0.739	0.760
LLM-Distillation – Single-Stage Fine-Tuning				
RankGPT [1]	0.696	0.666†	0.690†	0.662†
TWOLAR [2]	0.693	0.669†	0.754	0.730
RDL (Ours)	0.709	0.704	**0.774**	0.754
LLM-Distillation – Two-Stage Fine-Tuning				
RankGPT [1]	0.664†	0.634†	0.477†	0.472†
TWOLAR [2]	0.715	0.706	0.763	0.760
RDL (Ours)	**0.720**	**0.711**	0.768	**0.770**

[1] Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents, Sun et al. (EMNLP 2023)
[2] TWOLAR: A TWO-Step LLM-Augmented Distillation Method for Passage Reranking, Baldelli et al. (ECIR 2024)

Comparing with SOTA Re-Rankers

1. Our distilled models are competitive with SOTA models
2. They are also orders of magnitude more efficient

nDCG@10 on TREC DL '19 / '20 and geometric mean on TIREx

Model	BM25		ColBERTv2		TIREx
	DL 19	DL 20	DL 19	DL 20	
First Stage	0.480	0.494	0.732	0.724	0.290
RankGPT-4	0.713	0.713	0.766	0.793	–
RankZephyr	0.719	0.720	0.749	0.798	0.321
monoT5$_{3B}$	0.705	0.715	0.745	0.757	0.305
RankT5$_{3B}$	0.710	0.711	0.752	0.772	**0.323**
mELECTRA$_{Base}$	0.720	0.711	**0.768**	0.770	0.312
mELECTRA$_{Large}$	**0.733**	**0.727**	0.765	**0.799**	0.320

➜ We are able to close the effectiveness gap between cross-encoders and LLM re-rankers

Resources

zenodo.org/records/12528410 github.com/webis-de/rank-distillm webis.de/publications.html#schlatt_2025b