LiT and Lean: Distilling Listwise Rerankers into Encoder-Decoder Models

Manveer Singh Tamber Ronak Pradeep Jimmy Lin

University of Waterloo

Efficient Listwise Reranking with LiT5

Listwise reranking with LLMs has shown strong effectiveness but often uses models with billions of parameters. We propose LiT5, an encoder-decoder listwise reranker trained via distillation from RankZephyr. LiT5 maintains strong effectiveness with models between 220M–3B parameters and rerankers up to 100 passages in a single shot.

The Fusion-in-Decoder Architecture

Following FiD, LiT5 uses T5 to encode query-passage pairs independently before ranking all passages in the decoder. The decoder processes the concatenated representations to output passage identifiers ordered by relevance. FiD allows LiT5 to rerank up to 100 passages at once, with computation scaling linearly with the number of passages.

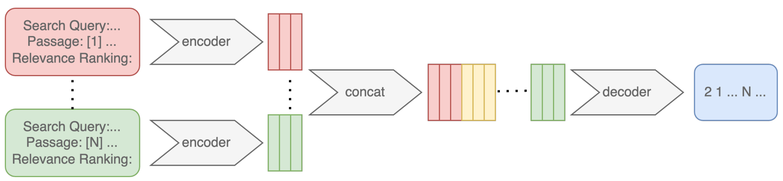

Figure 1. LiT5 architecture. Each query–passage pair is encoded separately; the decoder then generates a relevance ranking, e.g., "2 1 ...N".

RankGPT and RankZephyr

We build on RankZephyr that distills from RankGPT as a teacher model. RankZephyr bridges the effectiveness gap with RankGPT4 and in some cases even surpasses the proprietary teacher. In LiT5, we use RankZephyr as a teacher model for distillation, leveraging a cheaper, more transparent, open-source model.

BEIR Reranking Effectiveness

LiT5 generalizes well across BEIR datasets. Larger model variants (220M → 3B) scale well, achieving stronger scores

	BM25	LiT5			RankZephyr
Dataset	-	220M	770M	3B	7B
TREC-COVID	59.5	79.5	82.1	81.8	85.6
BioASQ	52.3	55.2	57.4	58.2	55.6
NFCorpus	32.2	34.2	34.9	36.1	32.2
NQ	30.6	52.9	56.1	57.7	56.9
HotpotQA	63.3	68.8	72.0	73.8	72.1
FiQA	23.6	36.5	40.0	41.7	38.7
Signal-1M	33.0	31.5	32.2	32.0	31.5
TREC-NEWS	39.5	48.0	49.9	49.4	52.2
Robust04	40.7	52.7	56.5	55.4	54.7
Arguana	39.7	29.7	35.2	39.2	42.7
Touche-2020	44.2	32.8	34.1	34.4	32.9
Quora	78.9	80.7	84.7	85.4	80.6
DBPedia	31.8	40.7	43.6	44.7	44.6
SCIDOCS	14.9	16.4	18.8	19.3	19.3
FEVER	65.1	77.6	78.1	81.6	77.1
Climate-FEVER	16.5	22.0	21.9	22.9	23.5
SciFact	67.9	72.4	74.1	74.9	76.0
Average	43.2	48.9	51.3	52.3	51.5

Table 1. Average nDCG@10 score for reranking the top 100 passages returned by BM25 on BEIR datasets.

Parameter- and Computational-Efficiency

LiT5 offers a favorable trade-off using fewer parameters and lower computational costs compared to RankZephyr, while maintaining competitive effectiveness.

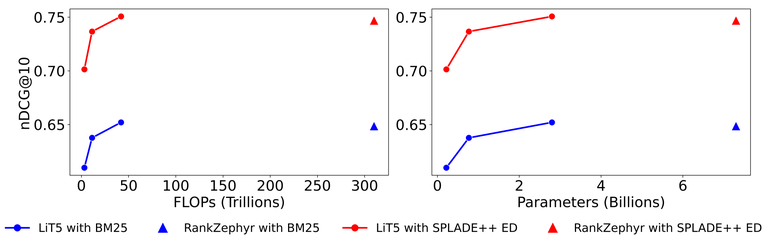

Figure 2. Average nDCG@10 across DL collections for the LiT5 model variants vs. RankZephyr, comparing reranking of the top 100 passages by average FLOPs per query (left) and model parameters (right).

MSMARCO Reranking Effectiveness

LiT5 outperforms pointwise T5-based rerankers and is competitive with RankZephyr and RankGPT4 despite using significantly fewer parameters and building on the older T5 model.

	Model Params	Source Prev.	MSv1 DL19	MSv1 DL20	MSv2 DL21	MSv2 DL22
BM25	-	None	0.506	0.480	0.446	0.269
SPLADE++ED	110M	None	0.731	0.720	0.684	0.570
MonoT5	220M	BM25	0.715	0.670	-	-
MonoT5	3B	BM25	0.718	0.689	-	-
RankT5	3B	BM25	0.712	0.695	-	-
RankVicuna	7B	BM25	0.668	0.655	0.624	0.430
RankVicuna	7B	SPLADE++ED	0.746	0.747	0.701	0.582
RankZephyr	7B	BM25	0.742	0.709	0.703	0.515
RankZephyr	7B	SPLADE++ED	0.782	**0.816**	0.760	0.669
RankGPT3.5	?	BM25	0.686	0.620	0.605	0.418
RankGPT4	?	BM25	0.750	0.704	0.707	0.508
RankGPT4	?	SPLADE++ED	0.746	0.708	**0.772**	**0.718**
$LiT5_{base}$	220M	BM25	0.717	0.667	0.645	0.484
$LiT5_{base}$	220M	SPLADE++ED	0.783	0.751	0.693	0.626
$LiT5_{large}$	770M	BM25	0.733	0.698	0.679	0.512
$LiT5_{large}$	770M	SPLADE++ED	**0.800**	0.766	0.728	0.686
$LiT5_{XL}$	3B	BM25	0.730	0.737	0.703	0.512
$LiT5_{XL}$	3B	SPLADE++ED	0.785	0.804	0.747	0.696

Table 2. nDCG@10 on DL19–DL22 reranking top-100 BM25 or SPLADE++ED passages. Best scores are in bold, and the best scores with BM25 are underlined.

Try it Out!

We have integrated LiT5 into our RankLLM repository and made the models available on HuggingFace! Check out our paper for more details.

ECIR 2025, Lucca, Italy

{mtamber, rpradeep, jimmylin}@uwaterloo.ca