ListT5: Listwise Reranking with Fusion-in-Decoder Improves Zero-shot Retrieval

ACL 2024 Bangkok, Thailand

Paper Code

Soyoung Yoon[1]*, Eunbi Choi[2], Jiyeon Kim[3], Hyeongu Yun[2], Yireun Kim[2], Seung-won Hwang[1]†

[1]Seoul National University [2]LG AI Research [3]KAIST AI

†Corresponding Author *Work done during an internship at LG AI Research

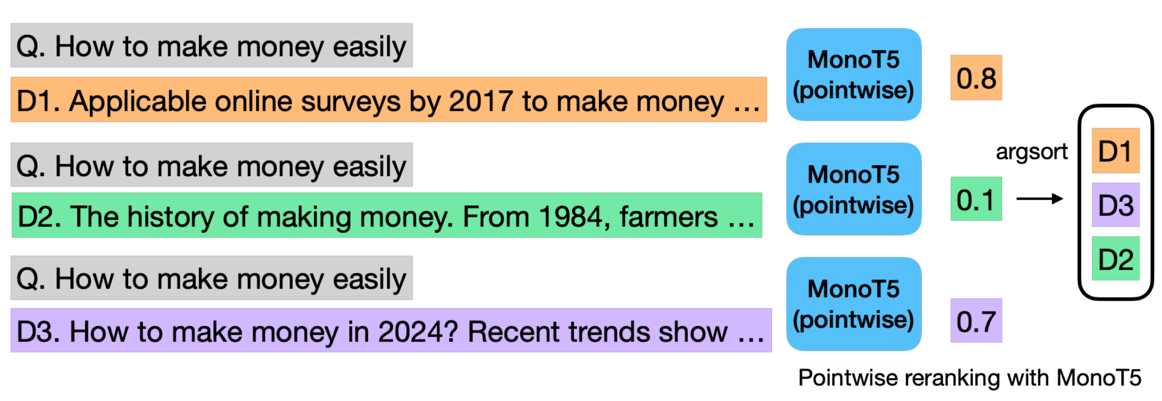
KAIST AI LG AI Research

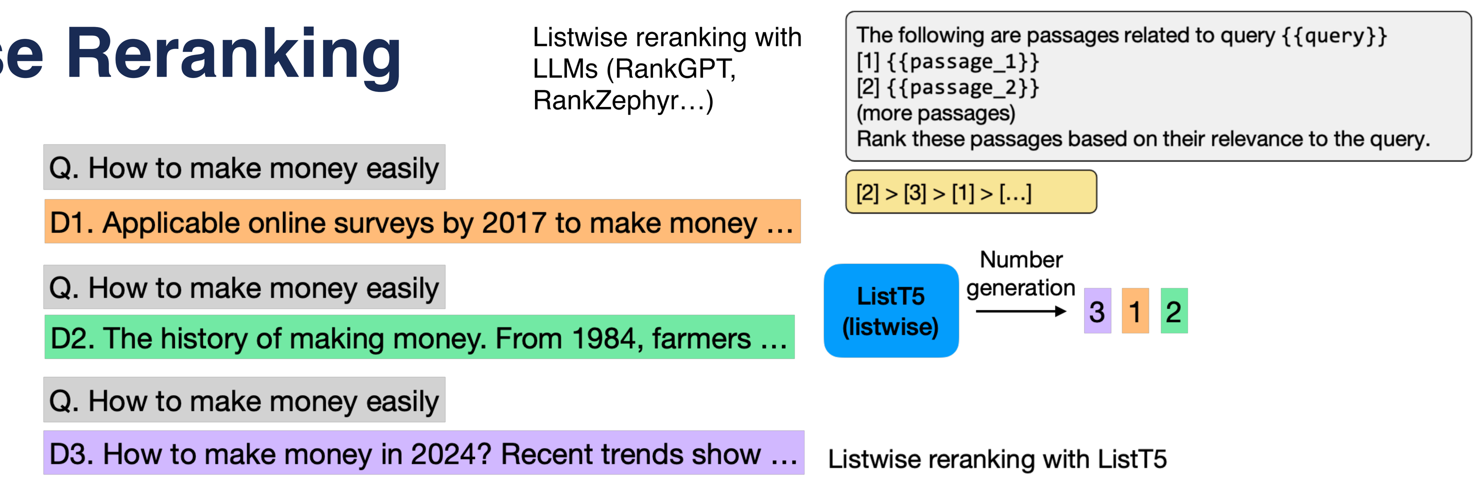
Overview

- Neural-based Information Retrieval systems still struggle on zero-shot retrieval compared with statistical retrievers (e.g., BM25)
- Listwise reranking models are shown to be effective on zero-shot retrieval, but previous listwise reranking had limitations: small-sized models only implement pairwise reranking with impractical efficiency, and large-sized models suffer from the lost-in-the middle problem due to its long input length.
- We present **ListT5** that overcomes the aforementioned limitations with the following advantages:

1. **Computational Efficiency:** Efficient than pairwise methods & listwise methods w/ LLMs, comparable to pointwise methods, applicable to small models (e.g., T5-base).
2. **Robustness to Positional Bias:** Effectively overcomes the lost-in-the middle problem, better than RankGPT-4, by the nature of Fusion-in-Decoder.
3. **Zero-shot performance:** Shows superior performance than pointwise(MonoT5, RankT5) and listwise (RankZephyr, RankVicuna, RankGPT3.5) counterparts.

Background: Pointwise v.s. Listwise Reranking



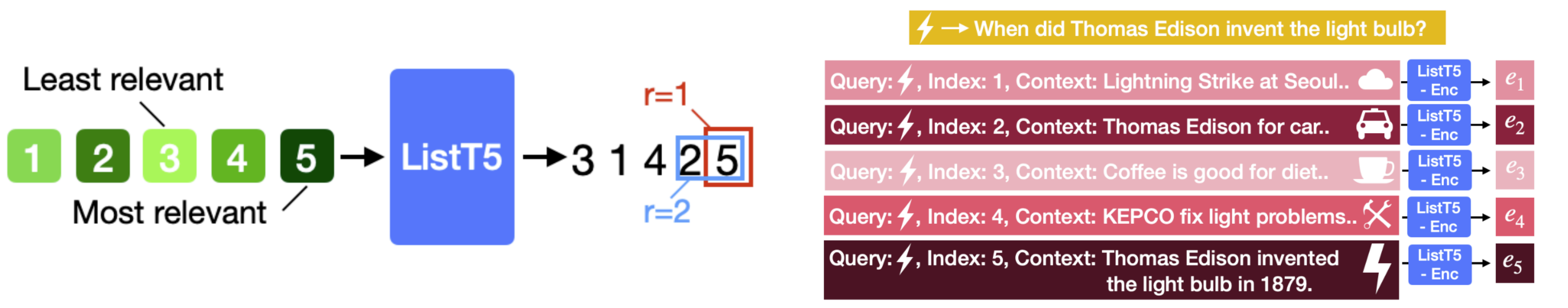


- **Pointwise** (MonoT5, RankT5): Individually assigns *definite* relevance scores for each documents
- **Listwise** (ListT5, RankGPT, RankZephyr..): Given multiple documents as input, sort documents and compute *relative* ordering between them

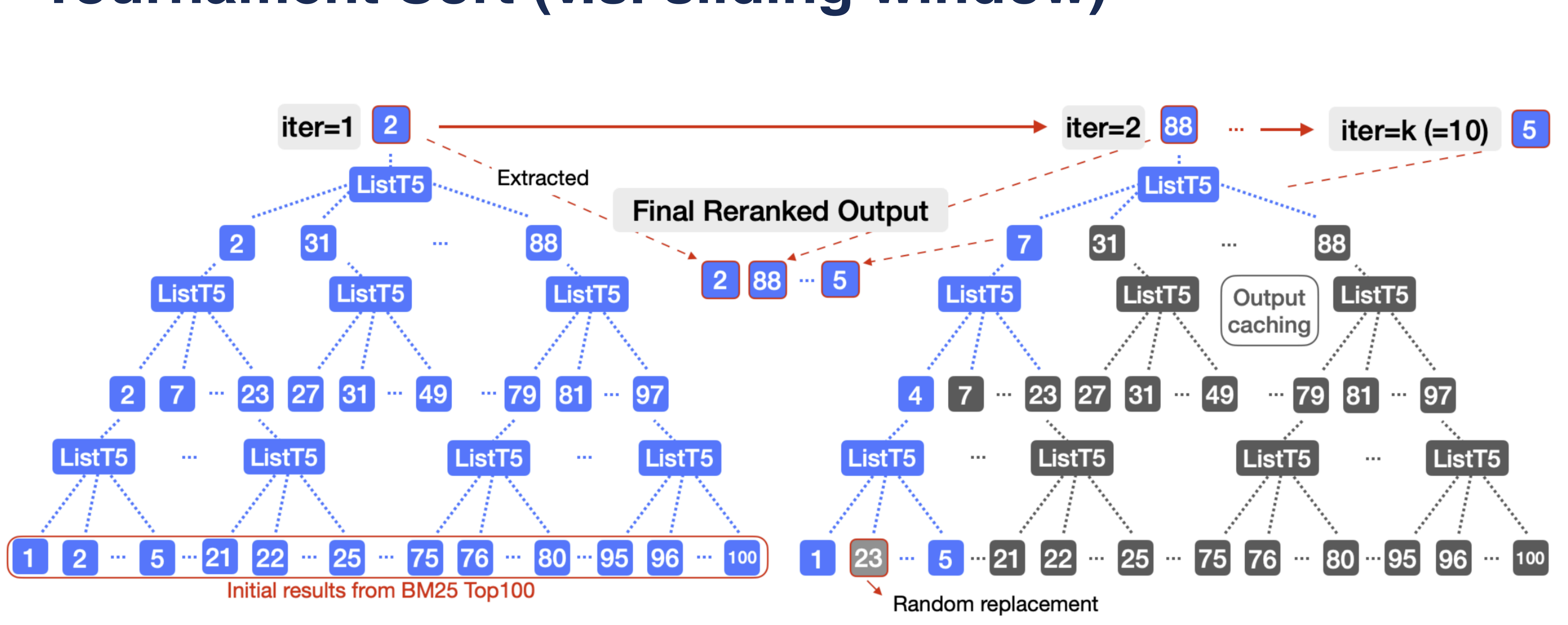
- However, they exhibit the lost in the middle problem, positionally biased to passages presented in the **first** and **last** parts of the listwise input.
- How can we train the model to efficiently see multiple passages at once, while being fairly efficient and exhibit less positional bias?
  -> **Fusion-In Decoder** with **tournament sort**!
- Each passage is processed by the encoder with identical positional encodings by FiD
  - so, ListT5 **cannot** exploit positional bias.

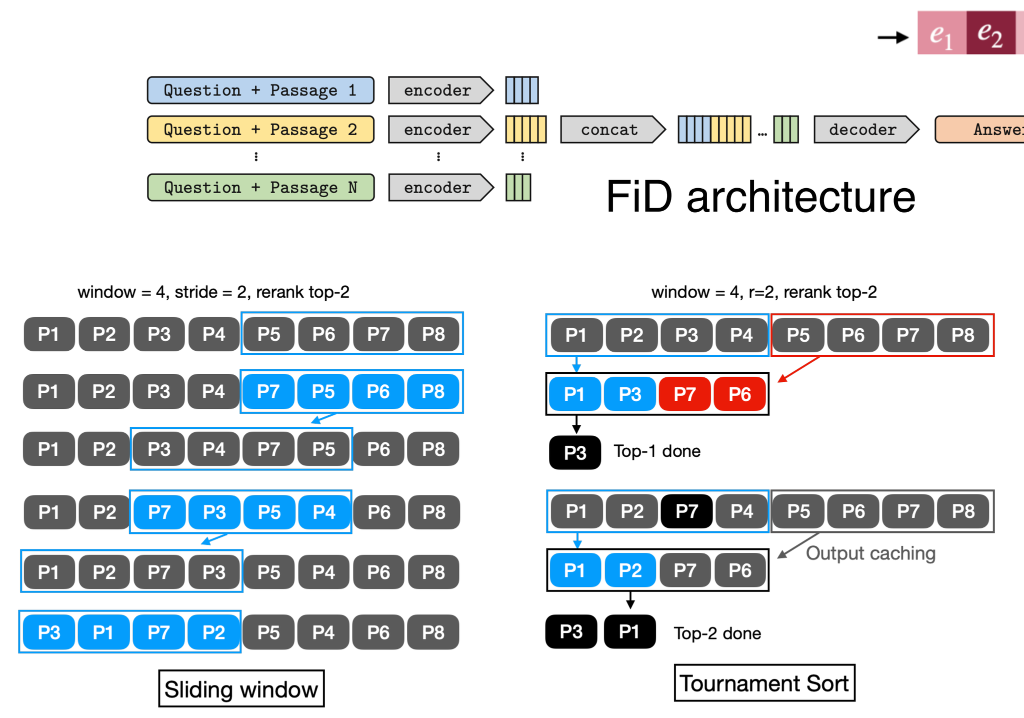
Proposed Method: ListT5 architecture

- Fusion-in-Decoder that given k (=5) contexts, output sorted index, with relevant index coming at the **last.**
- Training: MS MARCO train set, label negatives by Bi-encoder (COCO-DR/GTR)



Tournament Sort (v.s. sliding window)





Method Name	Reranking Method	Complexity
MonoT5 (Nogueira et al., 2020), RankT5 (Zhuang et al., 2022)	Pointwise	$\mathcal{O}(n)$
DuoT5 (Pradeep et al., 2021)	Pairwise	$\mathcal{O}(n^2)$
ListT5 (Ours)	Listwise	$\mathcal{O}(n + k \log n)$

- Sliding window: since window of size m can only "cache" up to m passages, full reranking top - k becomes inaccurate when k > m, and we need to run the whole iteration multiple times.
- Tournament sort: once the tree is constructed, additional iteration only requires computing a **single** path from leaf to root - most nodes can be cached & re-used for k iterations.

Zero-shot performance

- ListT5-base and ListT5-3B was superior than pointwise, pairwise, and listwise counterparts on the average NDCG@10 on full BEIR benchmark.

	TREC-DL19	TREC-DL20	TREC-COVID	NFC-orpus	Signal-1M (RT)	TREC-NEWS	Robu-st 04	Touche-2020	DBP-edia	Sci-Fact	Avg (In-domain)	Avg (BeIR)
DuoT5-base	71.4	67.4	**80.1**	35.0	31.4	**49.1**	49.6	31.8	**43.9**	69.6	69.4	48.8
ListT5-base ($r = 2$)	**71.8**	**68.1**	78.3	**35.6**	**33.5**	48.5	**52.1**	**33.4**	43.7	**74.1**	**70.0**	**49.9**
RankGPT (GPT3.5)	65.8	62.9	76.7	35.6	32.1	48.9	50.6	**36.2**	44.5	70.4	64.4	49.4
RankVicuna-7b	68.9	66.1	80.5	33.2	**34.2**	46.9	48.9	33.0	44.4	70.8	67.5	49.0
RankZephyr-7b	**73.9**	**70.9**	84.0	36.7	31.8	52.6	54.3	33.8	44.6	74.9	**72.4**	51.6
ListT5-3B ($r = 2$)	71.8	69.1	**84.7**	**37.7**	33.8	**53.2**	**57.8**	33.6	**46.2**	**77.0**	70.5	**53.0**

- Comparison with listwise LLMs & pairwise rerankers (DuoT5)

	COCO-DR Large (Init.)	MonoT5	RankT5	ListT5 (r=1)	ListT5 (r=2)
MSMARCO Top-1000 (in-domain)	41.9	43.1	46.2	46.1	**46.3**
TREC-COVID	80.8	**83.5**	**83.5**	83.2	**83.5**
NFCorpus	35.5	35.6	35.5	**36.2**	**36.2**
NQ	54.3	57.9	59.6	59.7	**60.0**
HotpotQA	63.3	68.7	**71.1**	70.3	70.9
FiQA-2018	32.3	41.2	41.3	**41.7**	**41.7**
Arguana	46.9	33.0	34.8	49.0	**49.3**
Touche-2020	21.6	25.7	**35.7**	29.1	29.6
CQADupStack	37.3	40.5	38.7	40.7	**40.9**
Quora	**87.3**	84.0	83.0	86.2	86.3
DBPedia	40.7	44.4	**46.1**	45.6	45.4
SCIDOCS	17.3	17.5	17.5	17.7	**18.3**
FEVER	74.9	78.9	79.7	79.8	**81.4**
Climate-FEVER	23.1	24.2	22.9	23.9	**24.9**
SciFact	71.9	73.5	73.6	**74.4**	74.3
Avg. BEIR	49.1	50.6	51.6	52.7	**53.1**

	BM25 Top-100							BM25 Top-1000		
	Initial	MonoT5 -base	RankT5 -base	ListT5 -base (r=2)	MonoT5 -3B	RankT5 -3B	ListT5 -3B (r=2)	MonoT5 -base	RankT5 -base	ListT5 -base (r=2)
TREC-COVID	59.5	**78.3**	77.7	**78.3**	79.8	81.7	**84.7**	78.3	79.1	**82.1**
NFCorpus	32.2	**35.7**	35.1	35.6	37.3	**37.4**	37.7	**36.1**	35.3	**36.1**
BioASQ	52.2	55.3	**58.2**	56.4	57.5	**58.3**	**58.3**	52.6	**57.6**	55.0
NQ	30.5	52.1	**53.2**	53.1	56.4	**57.8**	56.2	55.9	**57.6**	57.5
HotpotQA	63.3	71.2	**72.8**	72.6	74.3	74.8	**75.6**	70.9	**73.8**	73.6
FiQA-2018	23.6	39.2	39.2	**39.6**	**46.0**	45.2	45.1	41.2	41.1	**41.8**
Signal-1M (RT)	33.0	32.0	30.8	**33.5**	32.2	31.9	**33.8**	29.3	28.6	**30.9**
TREC-NEWS	39.5	48.0	45.4	48.5	48.3	49.5	**53.2**	47.8	45.9	**50.9**
Robust04	40.7	53.4	**54.3**	52.1	**58.5**	58.3	57.8	55.4	**57.2**	54.7
Arguana	40.8	34.4	35.5	**48.9**	46.8	37.4	**50.6**	24.2	26.6	**46.9**
Touche-2020	44.2	29.6	**37.1**	33.4	32.5	**38.8**	33.6	26.4	**37.0**	31.5
CQADupStack	30.0	38.6	37.0	**38.8**	41.3	40.3	**42.1**	40.1	38.1	**40.5**
Quora	78.9	84.6	83.3	**86.4**	84.0	83.6	**86.9**	84.2	82.9	**86.4**
DBPedia	31.8	42.8	43.7	43.7	44.8	**45.0**	46.2	43.1	**45.1**	44.9
SCIDOCS	14.9	16.7	16.8	**17.6**	19.0	18.9	**19.5**	17.0	17.1	**18.0**
FEVER	65.2	78.4	77.6	**79.8**	80.0	79.8	**82.0**	77.9	77.8	**81.0**
Climate-FEVER	16.5	23.1	21.2	**24.0**	**26.2**	24.5	24.8	23.3	20.6	**24.9**
SciFact	67.9	73.1	73.5	**74.1**	76.3	**77.1**	77.0	73.3	73.6	**74.9**
Average	42.5	49.3	49.6	**50.9**	52.3	52.2	**53.6**	48.7	49.7	**51.8**

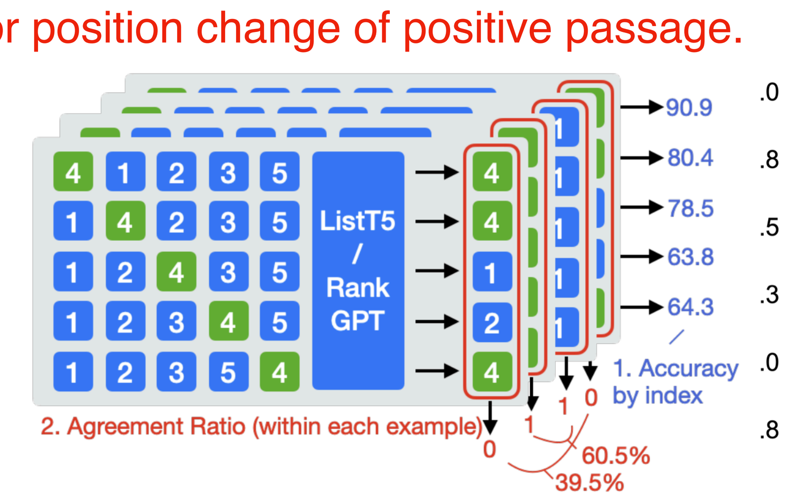
- Comparison with pointwise rerankers with BM25 / COCO-DR as first-stage retrievers

Positional Invariance

Initial ordering	DL19	DL20	TREC-COVID	TREC-NEWS	Touche -2020	Avg.
ListT5-base (tournament sort, r=2)						
No shuffle	71.8	68.1	78.3	48.5	33.4	60.0
Shuffle	71.2	68.1	77.2	48.9	32.8	59.6
Perf. drop						**-0.4**
ListT5-base (sliding windows, stride=3, iter=4)						
No shuffle	71.8	67.7	77.5	50.0	33.1	60.0
Shuffle	69.5	65.5	77.7	49.2	32.1	58.8
Perf. drop						**-1.2**
RankVicuna-7b (sliding windows)						
No shuffle	68.9	66.1	80.5	46.9	33.0	59.1
Shuffle	67.1	64.6	79.2	45.3	30.8	57.4
Perf. drop						**-1.7**
RankGPT-3.5 (sliding windows)						
No shuffle	68.4	64.9	72.6	46.5	38.2	58.1
Shuffle	62.5	57.0	66.1	38.3	22.8	49.3
Perf. drop						**-8.8**

- NDCG@10 drop before & after shuffling the initial top-100 ordering of BM25

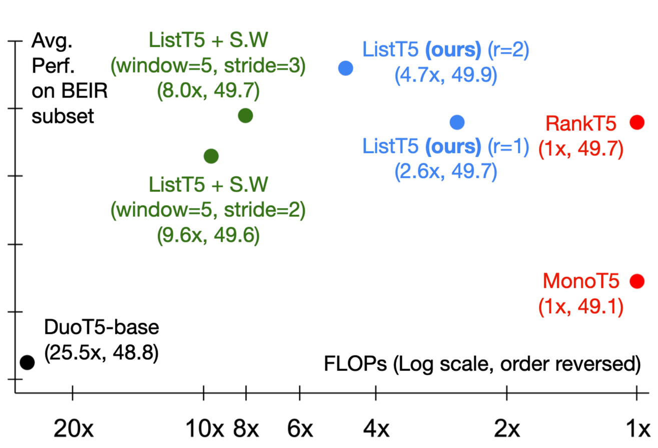
- ListT5 was more robust to initial ordering change or position change of positive passage.



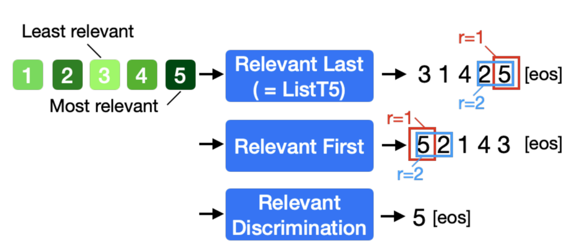
	TREC-COVID							FiQA						
	Accuracy when positive passage is at index # :						Aggrement ratio (↑)	Accuracy when positive passage is at index #:						Aggrement ratio (↑)
	1	2	3	4	5	Std. (↓)		1	2	3	4	5	Std. (↓)	
GPT-3.5	81.6	63.3	75.5	67.3	61.2	7.7	55.1	88.3	68.1	78.7	65.9	75.8	8.0	62.1
GPT-4	95.9	83.7	73.5	77.6	71.4	8.8	69.4	94.6	90.5	84.4	86.8	84.8	3.9	82.8
DuoT5	91.3	76.0	-	-	-	7.6	79.6	89.9	76.9	-	-	-	6.5	78.1
ListT5	93.9	87.8	83.7	85.7	81.6	**4.2**	**83.7**	85.3	85.6	82.2	83.3	82.6	**1.4**	**90.4**

- Agreement ratio & Std w.r.t. position of positive passage index

Efficiency



Ablations



- Relevant last was the most effective.

Idx	Base Model	Sorting method	Name	FLOPs to rerank: Top-1	Top-10
0	T5-base	pointwise	MonoT5	1x	1x
1	T5-base	tournament	ListT5(r=1)	1.3x	2.6x
2	T5-base	tournament	ListT5(r=2)	1.8x	4.7x
3	T5-base	sliding w.(s=2)	T5(FiD)	2.5x	9.8x
4	T5-base	sliding w.(s=3)	T5(FiD)	1.7x	12.3x
5	T5-3b	tournament	ListT5(r=1)	17.6x	36.3x
6	T5-3b	tournament	ListT5(r=2)	24.6x	66.0x
7	T5-3b	sliding w.(s=2)	T5(FiD)	38.5x	154x
8	T5-3b	sliding w.(s=2)	T5(no FiD)	53.8x	215.1x
9	T5-3b	sliding w.(s=3)	T5(FiD)	25.6x	128x
10	T5-3b	sliding w.(s=3)	T5(no FiD)	35.1x	175.6x

- Tournament sort was more efficient than sliding window.

Dataset	Relevant Discrimi-nation	Relevant First		Relevant Last (ListT5)	
		($r = 1$)	($r = 2$)	($r = 1$)	($r = 2$)
In-domain					
MS MARCO	40.3	40.8	**40.9**	40.7	40.7
TREC-DL19	**72.5**	69.6	70.8	71.2	71.8
TREC-DL20	67.3	67.0	66.8	67.3	**68.1**
Avg (in-domain)	60.0	59.1	59.5	59.7	**60.2**
Out-domain (BEIR)					
TREC-COVID	74.0	74.9	75.9	76.7	**78.3**
NFCorpus	34.8	35.5	35.6	35.5	**35.6**
BioASQ	55.8	56.6	56.6	**57.2**	56.4
NQ	51.1	52.7	52.9	52.0	**53.1**
HotpotQA	70.9	72.5	72.6	72.1	**72.6**
FiQA-2018	38.1	39.3	39.0	39.5	**39.6**
Signal-1M (RT)	32.9	31.8	31.7	33.3	**33.5**
TREC-NEWS	43.9	46.6	47.3	47.9	**48.5**
Robust04	49.8	**52.3**	**52.3**	52.0	52.1
Arguana	26.1	32.8	34.6	**49.7**	48.9
Touche-2020	**34.2**	31.5	31.3	**34.2**	33.4
CQADupStack	**38.8**	38.3	38.4	38.4	**38.8**
Quora	81.9	84.4	84.8	86.1	**86.4**
DBPedia	42.4	43.4	43.6	**43.9**	43.7
SCIDOCS	16.3	17.3	17.3	17.2	**17.6**
FEVER	77.6	77.4	77.7	77.8	**79.8**
Climate-FEVER	20.7	22.8	23.0	22.8	**24.0**
SciFact	73.0	74.1	**74.2**	74.1	74.1
Avg (BEIR)	47.9	49.1	49.4	50.6	**50.9**