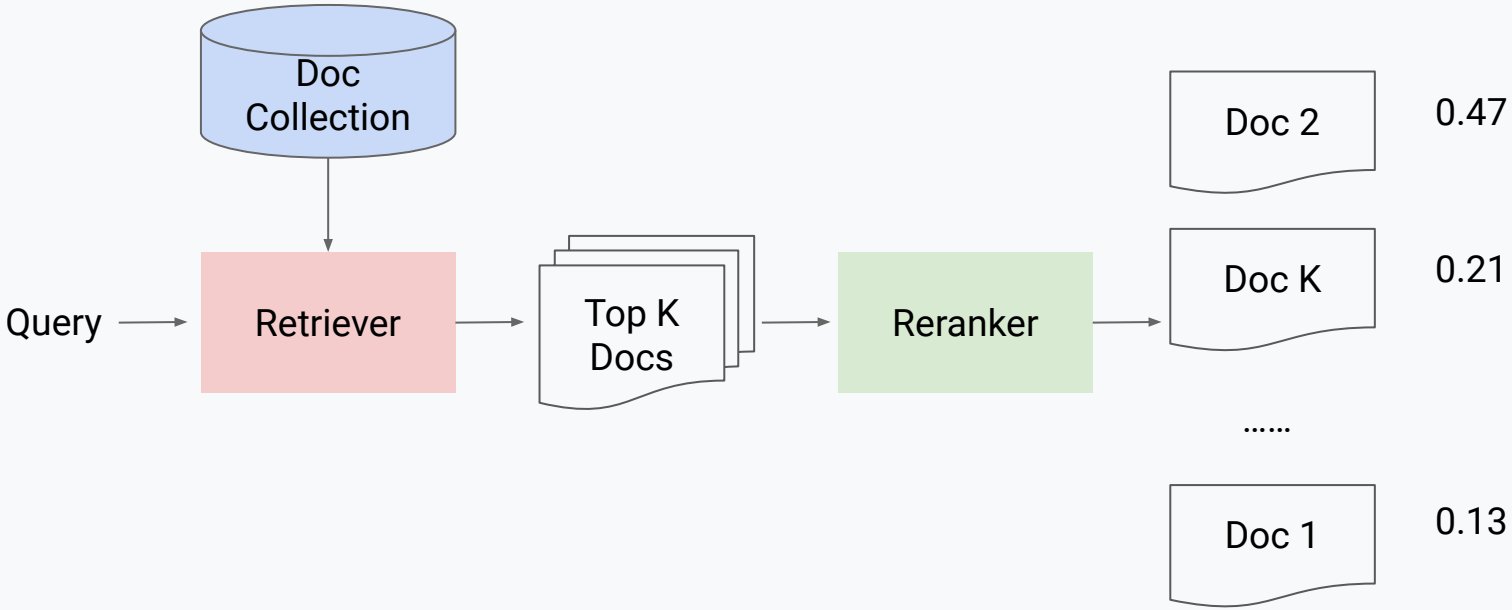
HYRR: Hybrid Infused Reranking for Passage Retrieval

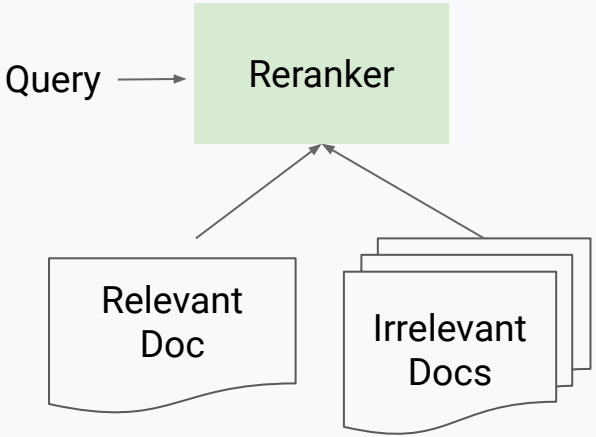
Jing Lu*, Keith Hall◇, Ji Ma*, Jianmo Ni✧
*Google Research, ◇Sizzle AI, ✧Google DeepMind

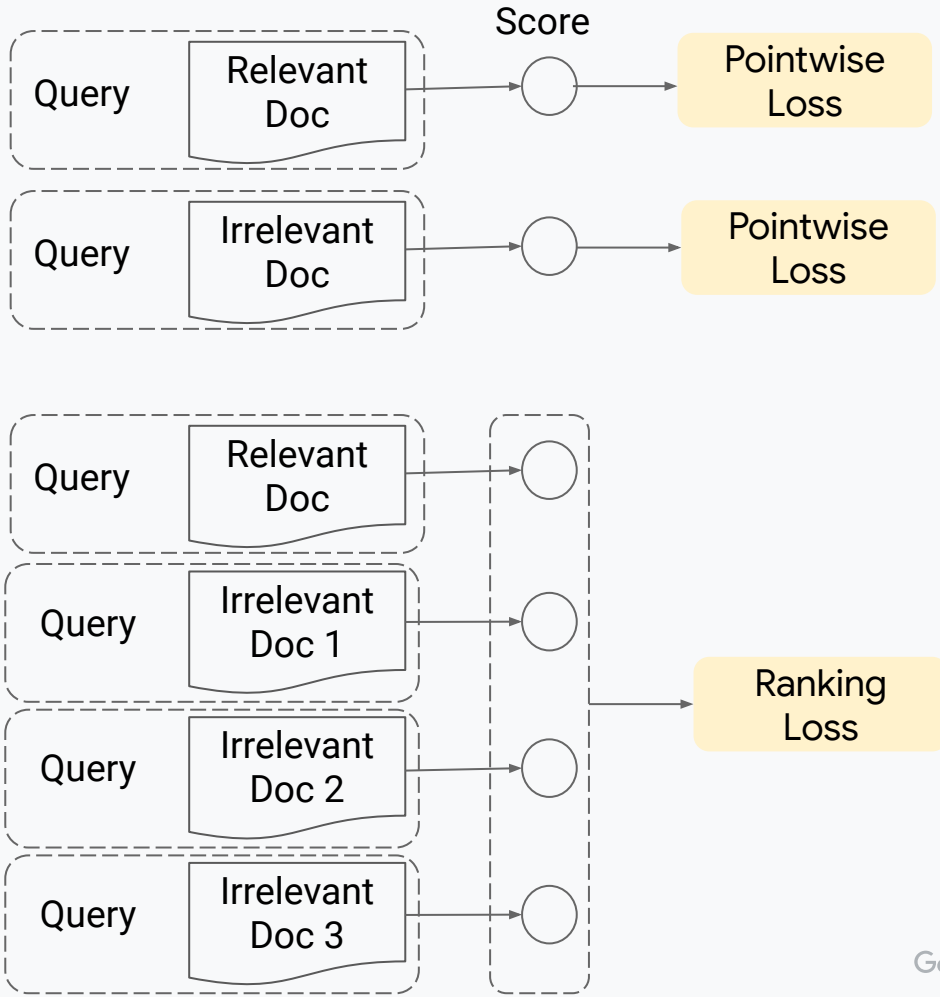
May 2024

Two-stage retrieval pipeline

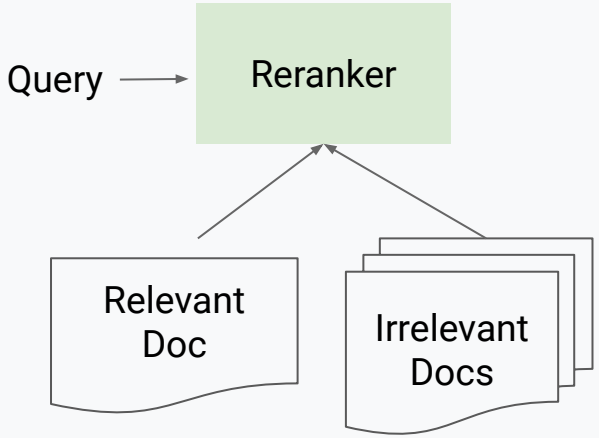


Training rerankers





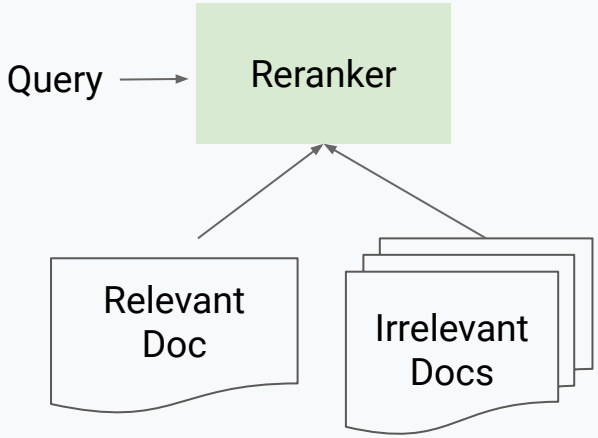
Training rerankers



Source of Irrelevant Docs:

- Labeled docs
  - Independently of the first-stage retrievers
  - E.g. monoT5 (Nogueira et al. 2020)
- Top results from retrievers
  - First-stage retrievers
    - Similar to the distribution at inference time
    - E.g. RocketQA (Ren et al., 2021)
  - Other retrievers

Training rerankers

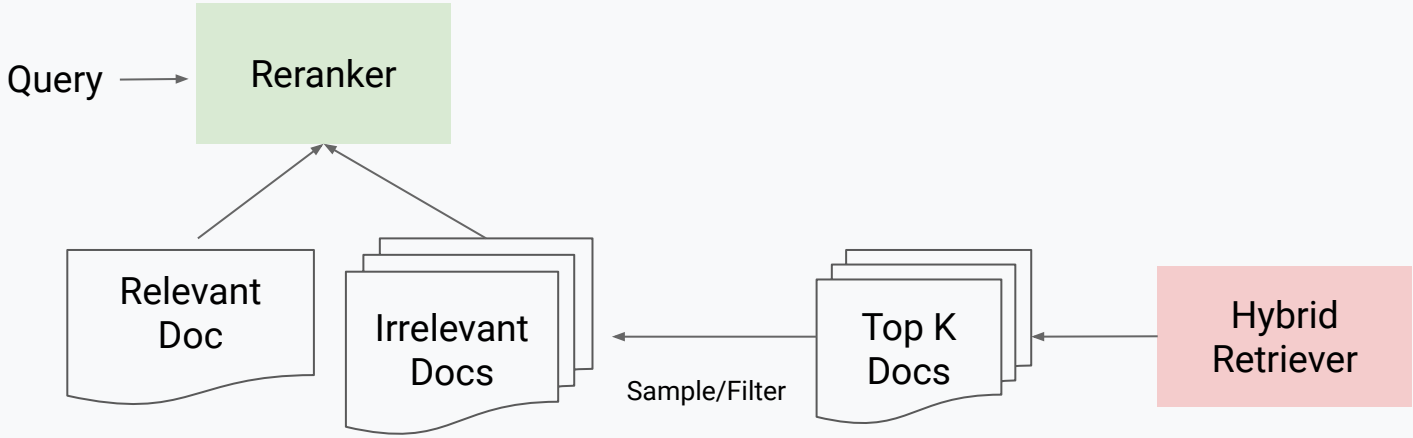


- When the retriever used to generate irrelevant docs is weaker than the first-stage retriever, performance drops severely
  - Gao et al. 2021

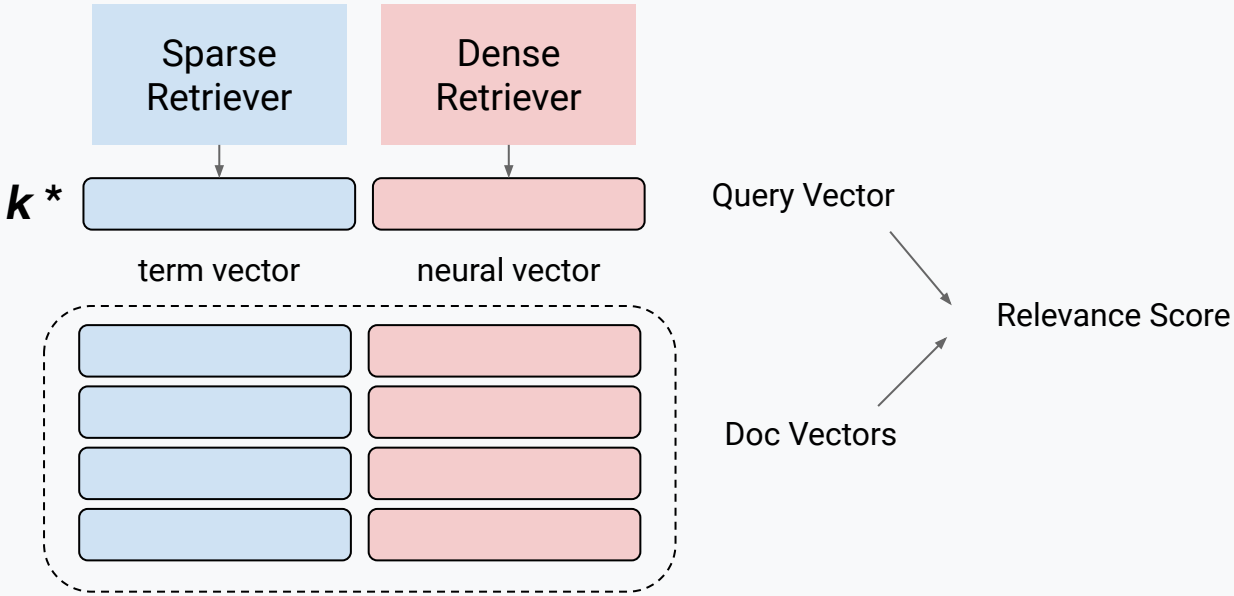
Goal

- Design a training paradigm for reranking models which are
  - Robust to the first-stage retrievers
  - Generalized in both supervised and zero-shot settings

Proposed training paradigm

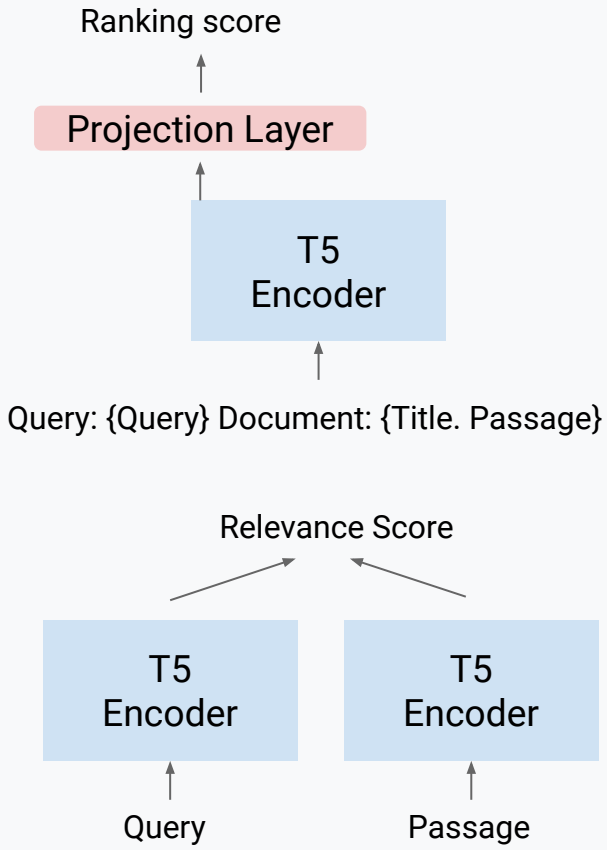


Hybrid Retriever



Evaluation

- Model
  - Reranker
    - Zhuang et al. 2023: RankT5 Encoder-only
  - Sparse retriever
    - BM25
  - Dense retriever
    - Ni et al. 2022: GTR dual encoder



Evaluation

- Fix the first-stage retriever and compare the reranking performance
  - BM25 Anserini
- Two settings
  - Supervised retrieval
    - MS MARCO passage ranking
    - Metrics: MRR@10
  - Zero-shot retrieval
    - BEIR benchmark: 18 evaluation datasets across 9 domains
    - Metrics: nDCG@10

Results - MS MARCO

	Model size	MRR@10
BM25 Anserini		0.1874
HLATR	$\text{RoBERTa}_{\text{Large}}$	0.3680
MiniLM	Distilled BERT	0.3901
monoT5	T5_{3B}	0.3980
RankT5-EncDec	T5_{Large}	0.3986
DERR_{MS}	T5_{Large} 1.1	0.4222
HYRR_{MS}	T5_{Large} 1.1	**0.4235**

Results - BEIR

	Retriever	Reranker			
	BM25	MiniLM	HYRR_{MS}	HYRR	HYRR
Model Size	Anserini	22M	400M	125M	400M
NQ	0.329	0.533	**0.569**	0.532	0.555
MS MARCO	0.228	0.413‡	**0.435**‡	0.307	0.309
Trec-Covid	0.656	0.757	0.798	0.796	**0.820**
BioASQ	0.465	0.523	**0.554**	0.551	0.549
NFCorpus	0.325	0.350	0.371	0.379	**0.382**
HotpotQA	0.603	0.707	**0.717**	0.706	0.707
FiQA-2018	0.236	0.347	0.411	0.408	**0.437**
Signal-1M	0.330	**0.338**	0.264	0.307	0.318
Trec-News	0.398	0.431	0.452	0.437	**0.453**
Robust04	0.407	0.475	0.505	0.501	**0.544**
ArguAna	**0.414**	0.311	0.351	0.344	0.342
Touché-2020	0.367	0.271	**0.467**	0.368	0.384
Quora	0.789	0.825	0.637	0.861	**0.867**
DBPedia-entity	0.313	**0.409**	0.402	0.385	0.403
SCIDOCS	0.158	0.166	0.184	0.183	**0.187**
Fever	0.753	0.819	0.825	**0.868**	0.861
Climate-Fever	0.213	0.253	0.262	0.272	**0.294**
SciFact	0.665	0.688	0.745	0.734	**0.754**
CQADupStack	0.299	0.370	0.368	0.398	**0.416**
Average	0.418	0.473	0.490	0.491	0.504
Average w/o NQ	0.423	0.470	0.486	0.489	0.501
Avg. improvement on BM25		4.63%	6.26%	6.58%	7.81%

Ablation

- Robust to different first-stage retrievers

Retriever ↓	No Reranker	BM25RR	DERR	HYRR
		MS MARCO		
BM25	0.187	0.375	0.422	**0.424**
DE	0.378	0.350	**0.440**	**0.440**
Hybrid	0.390	0.351	0.438	**0.440**
		SciFact		
BM25	0.677	0.750	0.742	**0.752**
DE	0.597	**0.755**	0.745	0.752
Hybrid	0.706	0.753	0.744	**0.759**

Conclusion

- We proposed a generic training framework for rerankers
  - Reranker is trained using negative examples retrieved from a hybrid retriever
  - Practical and generalizable

- Our rerankers are robust and outperform several strong baselines on MS MARCO and BEIR benchmark