RBC BOREALIS

Embedding-Based Context-Aware Reranker

Ye Yuan, Mohammad Amin Shabani, Siqi Liu

April 2026

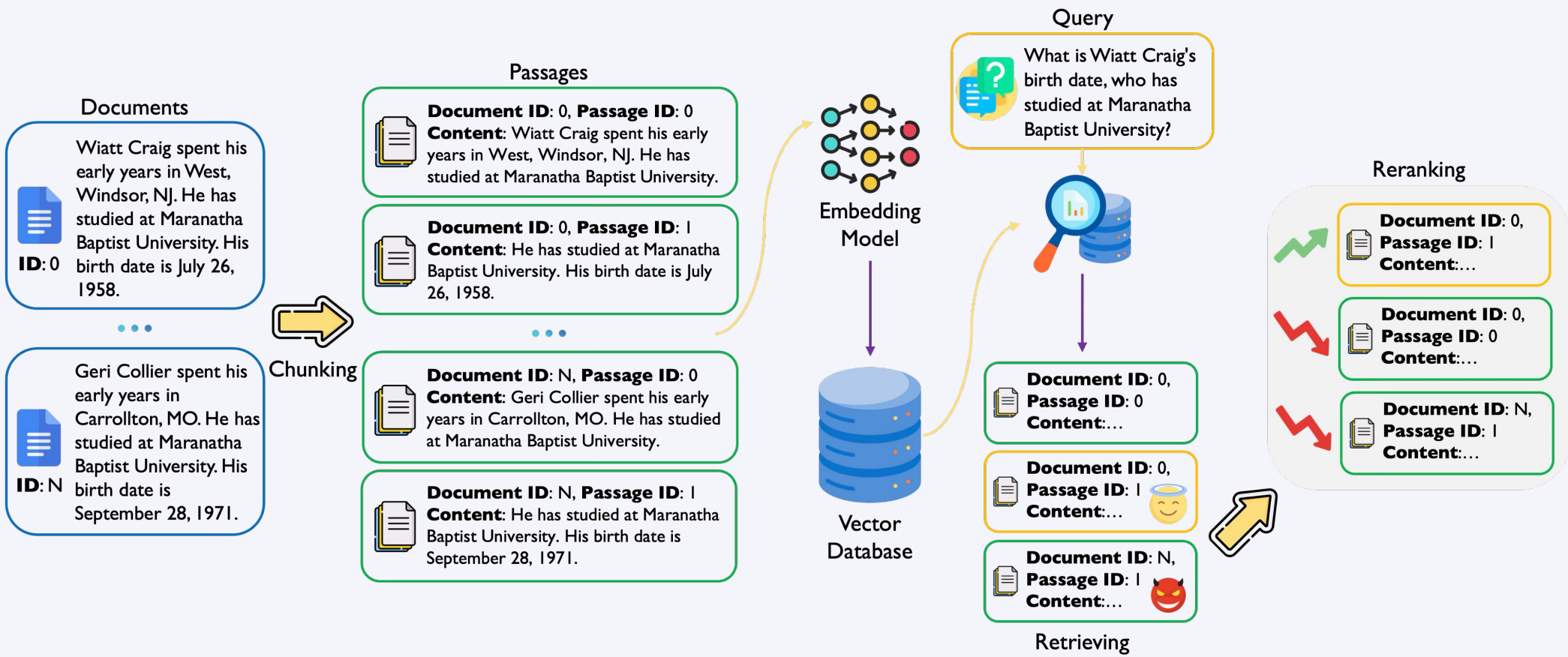
McGill UNIVERSITY

Mila

Problem Background

- Modern pipelines often rely on passage-level indexing, where a long document is split into shorter fixed-size passages for retrieval.
- Most rerankers rely on feeding the raw text of retrieved passages and the query into LLMs for scoring, incurring substantial computational cost and latency.
- Existing methods are evaluated on idealized benchmarks that assume the necessary information for answering a query is fully contained within a single passage, leaving the capabilities of current reranking strategies underexplored in scenarios that demand cross-passage inference.

Problem Background



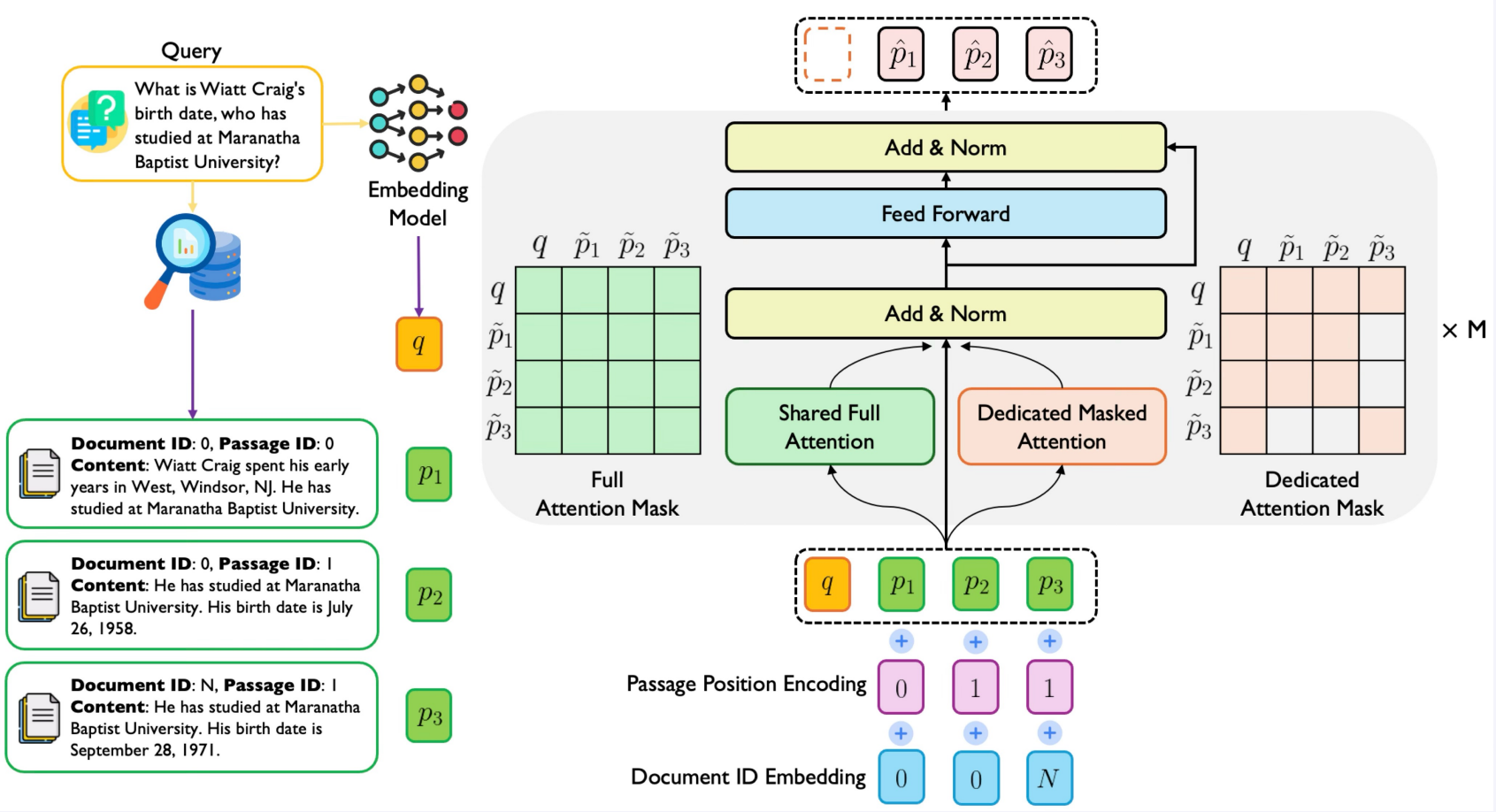
Methodology

- EBCAR operates directly on the dense embeddings of queries and retrieved passages.
- Passage embeddings are readily available, as they are stored in a vector database during indexing. The query embedding is obtained at inference time.
- To facilitate cross-passage inference, EBCAR incorporates structural signals, such as document IDs and the positions of passages within their original documents.
- These encodings are processed by a transformer encoder equipped with a hybrid attention mechanism.

Hybrid Attention Mechanism

- The shared full attention allows the query and all passage embeddings to attend to one another, capturing global interactions across retrieved passages.
- In contrast, the dedicated masked attention restricts attention scopes to the passages originating from the same documents, as determined by their document IDs.
- This hybrid attention design enables the model to infer over both inter-document context and fine-grained intra-document dependencies, promoting more accurate evidence aggregation and entity disambiguation.

Hybrid Attention Mechanism



Experiments

- We evaluate EBCAR on ConTEB, a challenging benchmark originally designed to test retrieval models' use of document-wide context but used in our work to assess rerankers.

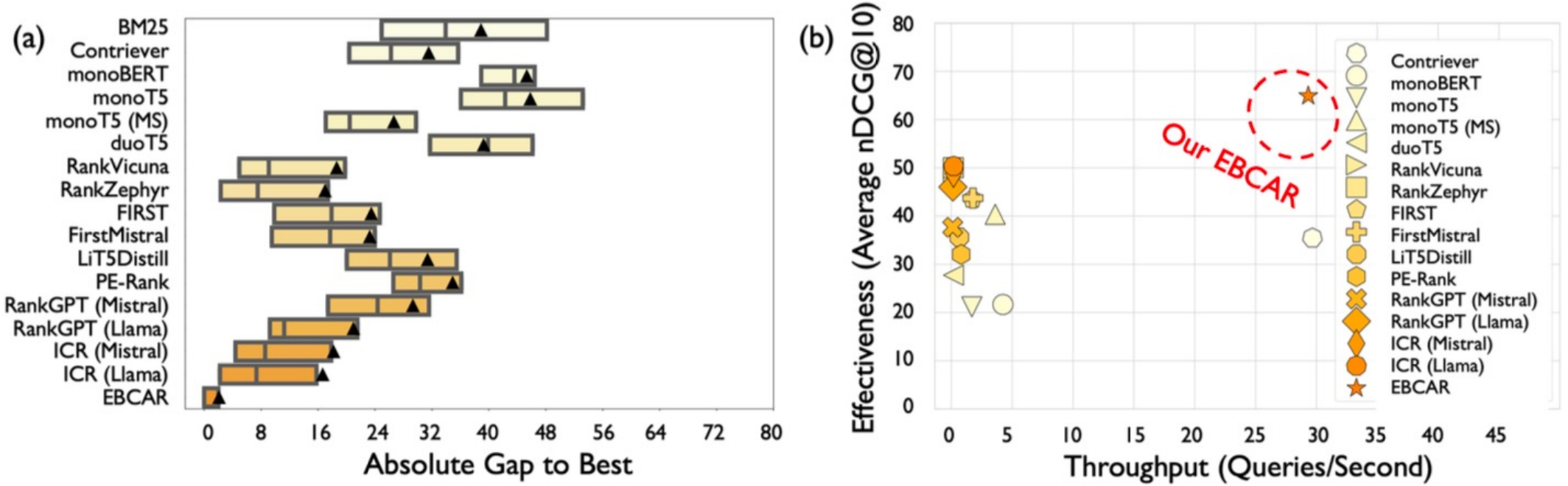
- Our experiments include both in-distribution evaluations and out-of-distribution zero-shot tests on entirely unseen domains. The results indicate that EBCAR offers a highly favorable integration of competitive ranking quality as well as fast and scalable inference.

Experiments

Table 1: nDCG@10 across datasets. Orange denotes in-distribution tests, and Green denotes out-of-distribution tests. Throughput is the number of queries processed per second on a single A100 GPU. The best and second best results are **bolded** and underlined, respectively.

Method	Training	Size	MLDR	SQuAD	NarrQA	COVID	ESG	Football	Geog	Insur	Avg	Throughput
BM25	-	-	65.42	45.07	36.20	39.79	10.10	4.69	23.66	0.00	28.12	263.16
Contriever	-	-	60.23	54.63	66.97	31.02	15.69	5.95	46.39	2.75	35.45	29.67
monoBERT (base)	ConTEB Train	110M	40.44	27.94	36.63	20.52	11.67	4.01	29.23	2.92	21.67	4.24
monoT5 (base)	ConTEB Train	223M	26.76	19.75	41.21	24.56	10.37	5.69	36.70	4.40	21.18	1.69
monoT5 (base)	MS-MARCO	223M	70.47	67.00	58.89	41.51	18.61	9.68	54.47	2.20	40.37	3.61
duoT5 (base)	MS-MARCO	223M	42.34	68.63	21.46	18.19	12.69	8.84	47.18	2.57	27.74	0.18
RankVicuna	MS-MARCO	7B	79.30	66.59	79.48	51.87	22.97	11.46	71.08	4.27	48.38	0.17
RankZephyr	MS-MARCO	7B	82.34	69.06	**81.18**	53.15	26.42	11.63	72.91	3.51	50.03	0.17
FIRST	MS-MARCO	7B	74.78	61.55	78.62	41.17	20.06	7.88	60.92	3.21	43.52	1.73
FirstMistral	MS-MARCO	7B	74.41	62.19	78.69	41.35	20.26	7.78	61.79	3.40	43.73	1.78
LiT5Distil	MS-MARCO	248M	60.33	54.87	66.94	31.22	16.07	5.98	46.66	2.78	35.61	0.66
PE-Rank	MS-MARCO	8B	52.58	45.00	51.58	33.40	19.92	6.29	46.25	1.37	32.05	0.81
RankGPT (Mistral)	-	7B	65.75	56.61	68.29	33.23	15.04	7.67	51.75	3.05	37.67	0.12
RankGPT (Llama)	-	8B	76.80	61.16	74.74	47.80	20.49	11.10	71.37	4.76	46.03	0.13
ICR (Mistral)	-	7B	79.49	67.41	79.50	52.54	25.65	10.57	71.44	4.00	48.83	0.18
ICR (Llama)	-	8B	**83.93**	69.09	79.96	53.32	28.36	10.91	73.10	4.16	50.35	0.19
EBCAR (ours)	ConTEB Train	126M	75.26	**71.62**	73.21	**59.80**	**37.20**	**80.19**	**81.30**	**40.74**	**64.92**	29.33

Experiments



Thank You!

Please forward your questions to ye.yuan3@mail.mcgill.ca.

Check Our Paper