Google DeepMind

Rethinking the Role of Token Retrieval in Multi-Vector Retrieval

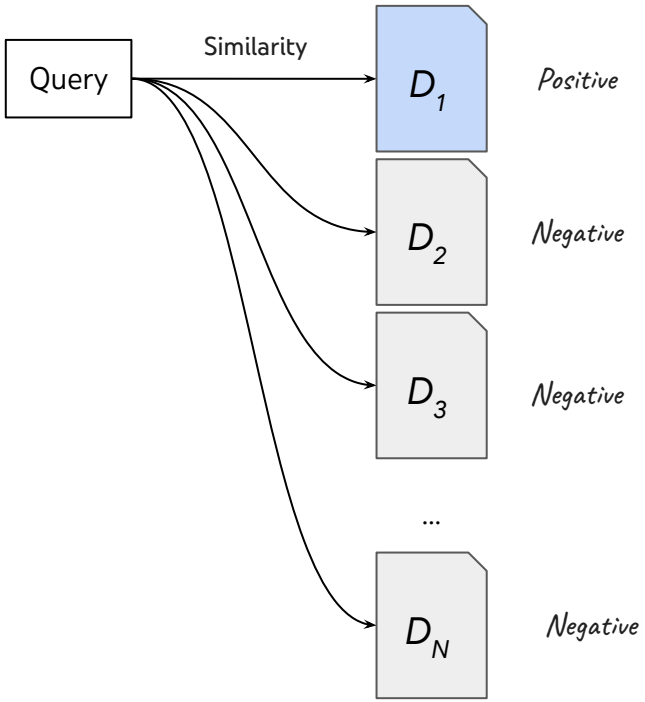
Jinhyuk Lee, Zhuyun Dai, Sai Meher Karthik Duddu, Tao Lei,
Iftekhar Naim, Ming-Wei Chang, Vincent Y. Zhao

NeurIPS 2023

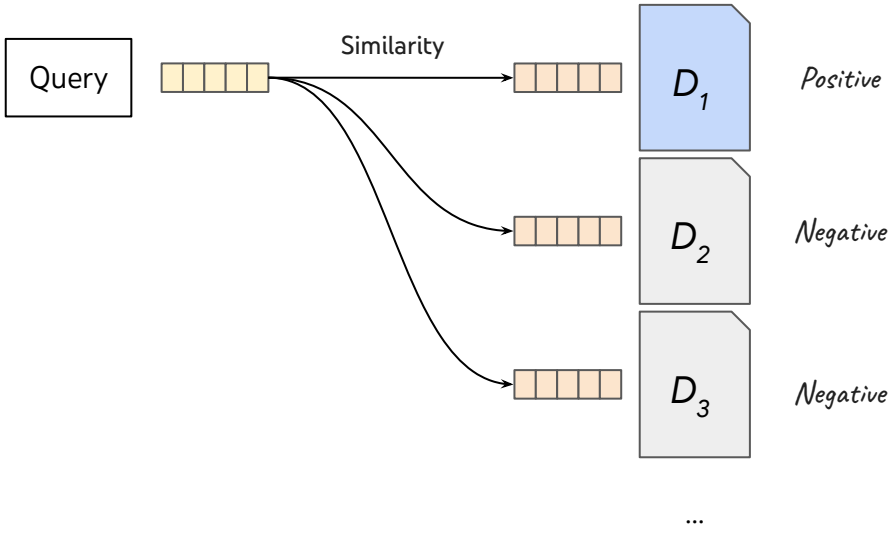
jinhyuklee@google.com

Last Updated: 11/13/2023

(Document) Retrieval

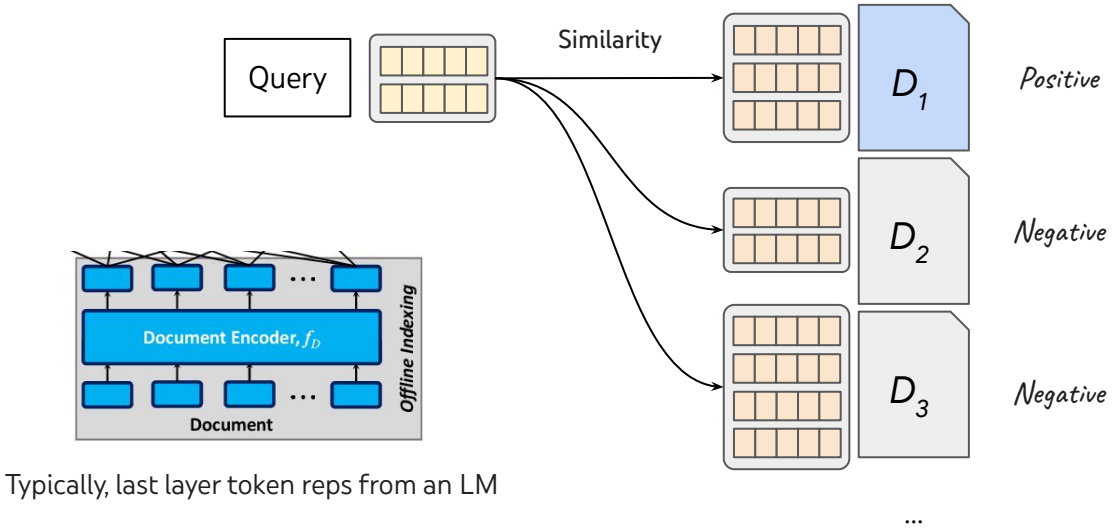


Dual Encoders



Similarity = Dot product

Multi-vector Retrievers

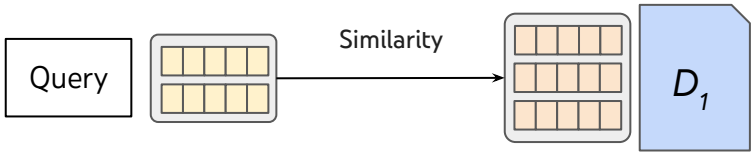


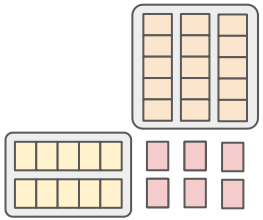
Typically, last layer token reps from an LM

Better representational capacity (= better performance)

Similarity = Dot product?

ColBERT [Khattab and Zaharia., 2020]





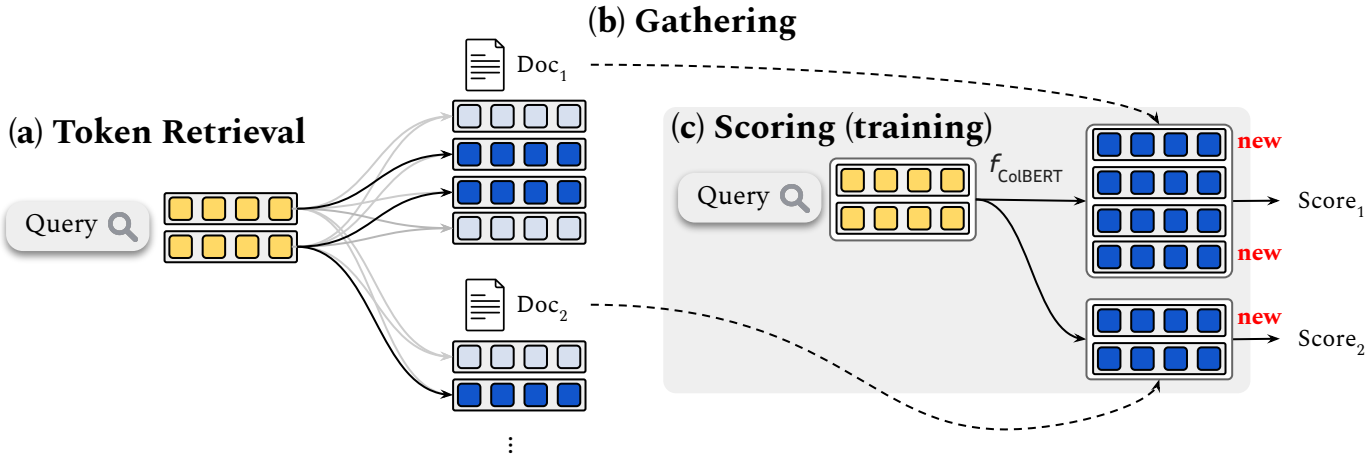
Sum-of-max $\mathbf{A}_{ij} = \mathbb{1}_{[j=\text{argmax}_{j'}(\mathbf{P}_{ij'})]}$ $1 \le j' \le m$

$$f_{\text{ColBERT}}(Q, D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{A}_{ij}\mathbf{P}_{ij} = \frac{1}{n}\sum_{i=1}^{n}\max_{1\le j\le m}\mathbf{q}_i^\top \mathbf{d}_j$$ = avg(,)

Non-linear operator between Q and D

=> Cannot be applied in large-scale MIPS (e.g. Faiss, ScaNN)

Three-stage Inference of ColBERT



Problems with Three-stage Inference

Training-testing mismatch
We do not have token retrieval during training.

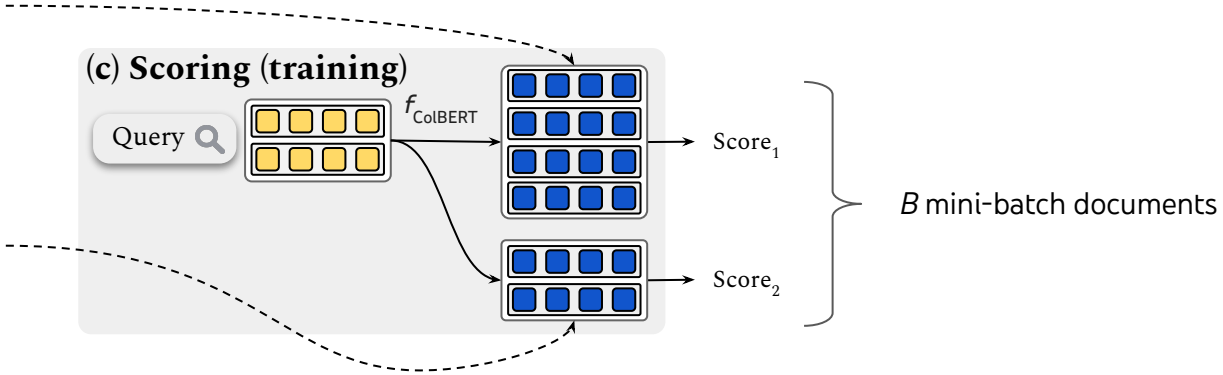
Complicated & expensive scoring
We need to load token vectors and recompute.

XTR: Conte**X**tualized **T**oken **R**etriever

Let's retrieve important document tokens

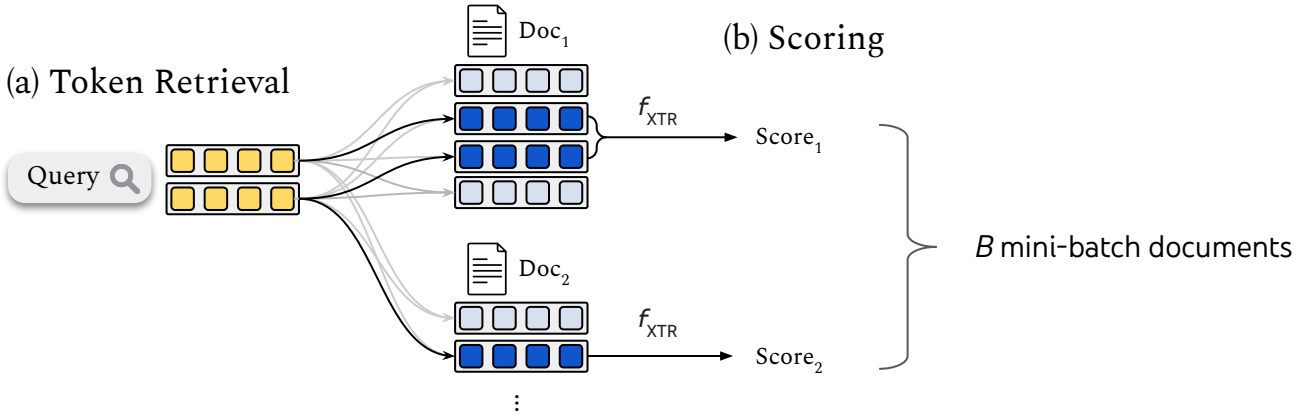
So that we can have simpler scoring with retrieved tokens

ColBERT$_{(v2)}$ Training



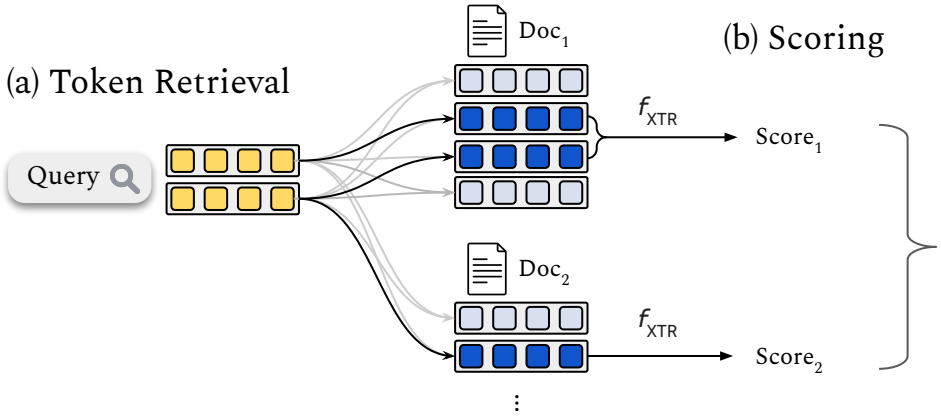
Minimize $\mathcal{L}_{\text{CE}} = -\log \frac{\exp f(Q, D^+)}{\sum_{b=1}^{B} \exp f(Q, D_b)}$

XTR Training



Important/salient tokens from positive documents should be **retrieved first** to be scored!

XTR Training

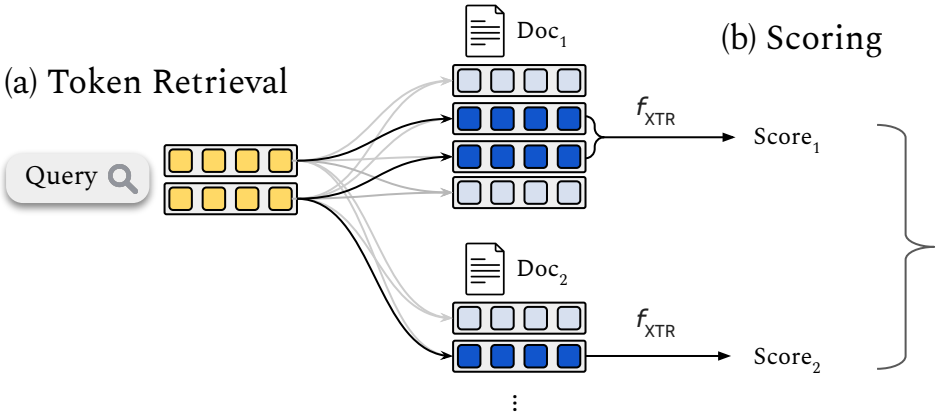


$$f_{\text{XTR}}(Q, D) = \frac{1}{Z} \sum_{i=1}^{n} \max_{1 \le j \le m} \hat{\mathbf{A}}_{ij} \mathbf{q}_i^\top \mathbf{d}_j$$

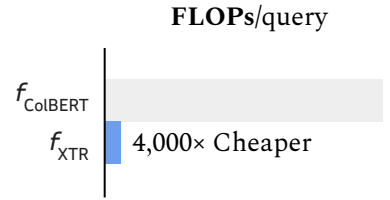
$$\hat{\mathbf{A}}_{ij} = \mathbb{1}_{[j \in \text{top-}k_{j'}(\mathbf{P}_{ij'})]}$$

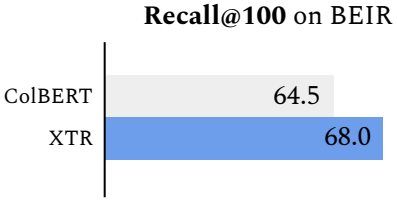
$$1 \le j' \le mB$$

XTR Inference



N total documents





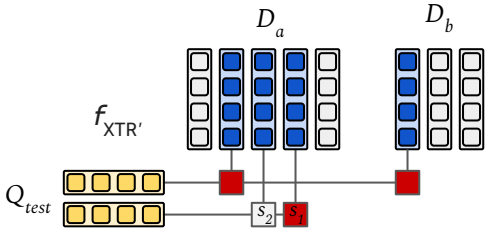
Better alignment of training and testing

Problem

# of docs are different during Training/Inference

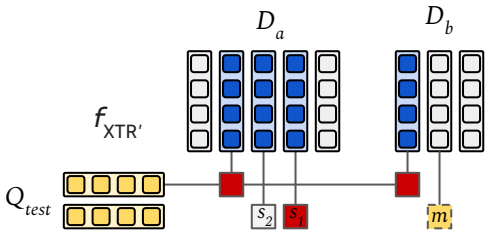
B mini-batch docs $O(10^2)$ **vs N** entire documents $O(10^6)$

Missing Similarity Imputation



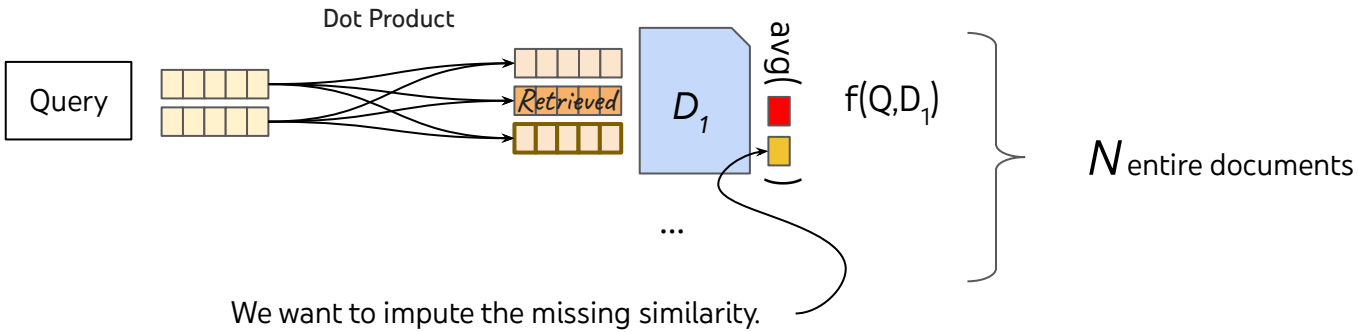
What if important tokens are not retrieved for D_b?

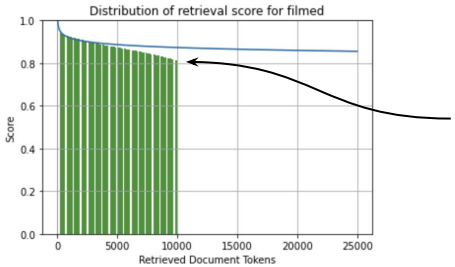
We know that... $m \leq s_2 \leq s_1$



So we impute the missing similarity!

Imputing Missing Similarity

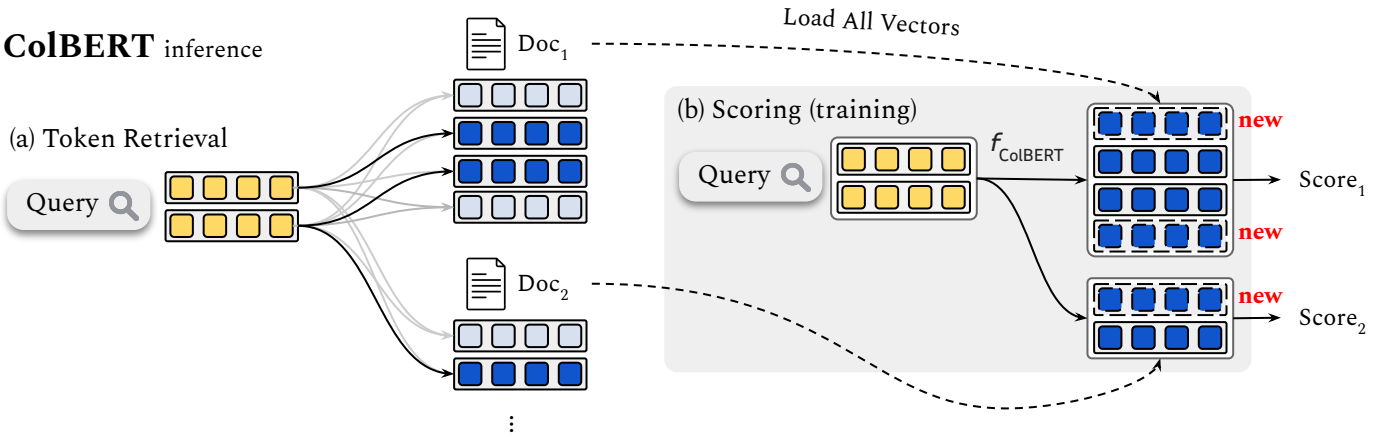


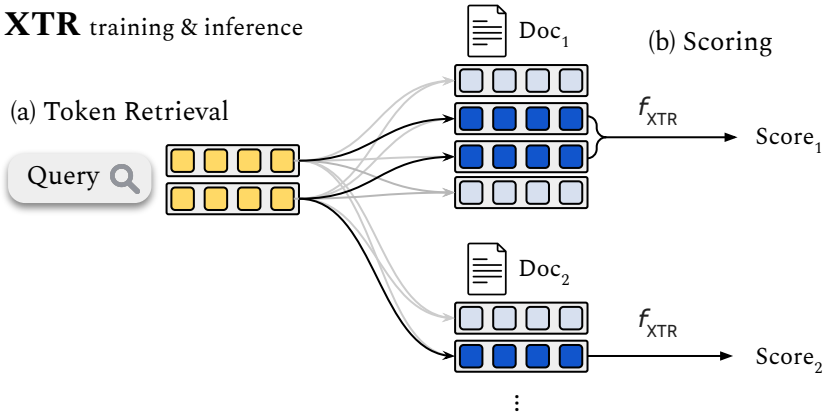


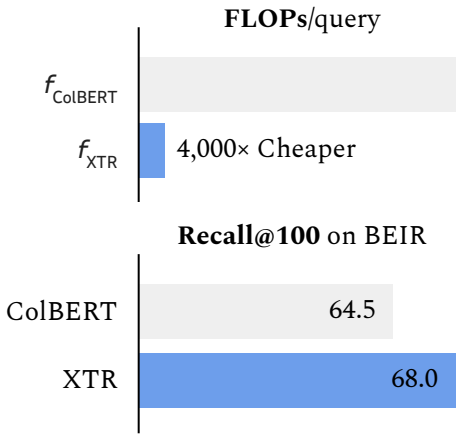
$$f_{\text{XTR}'}(Q, \hat{D}) = \frac{1}{n} \sum_{i=1}^{n} \max_{1 \le j \le m} \left[\hat{\mathbf{A}}_{ij} \mathbf{q}_i^\top \mathbf{d}_j + (1 - \hat{\mathbf{A}}_{ij}) m_i \right]$$

From top-k retrieved tokens, we know that the missing similarity is **upper-bounded by the last top-k score**.

We simply take the k-th score for our imputed value.


ColBERT inference
(a) Token Retrieval
Query
Doc1
Doc2
Load All Vectors
(b) Scoring (training)
Query
fColBERT
new
new
new
Score1
Score2


XTR training & inference
(a) Token Retrieval
Query
Doc1
Doc2
(b) Scoring
fXTR
fXTR
Score1
Score2


FLOPs/query
fColBERT
fXTR
4,000× Cheaper
Recall@100 on BEIR
ColBERT
64.5
XTR
68.0

FLOPs Comparison

	Retrieval	Refinement	Estimated FLOPs/query	Setting
GTR	$2d \log L$	0	10.3×10^3 (tok)	$L = 5 \times 10^6, d = 768$
ColBERT	$2nd \log M$	$n^2 k'(2m_{\text{avg}} d + m_{\text{avg}} + 1)$	38.8×10^3 (tok) + 0.36×10^9 (ref)	$M = 3 \times 10^9, n = 16, d = 128,$
XTR (ours)	$2nd \log M$	$n^2 k'(r_{\text{avg}} + 1)$	38.8×10^3 (tok) + 0.09×10^6 (ref)	$k' = 100, m_{\text{avg}} = 55, r_{\text{avg}} = 2.5$

4,000x smaller FLOPs!

Results on Zero-shot Document Retrieval

	MS	**BEIR**
GenQ	40.8	43.1
GPL	-	48.6
$\text{PTR}_{\text{retriever}}$	-	49.4
BM25	22.8	44.0
ColBERT	40.1	45.1
GTR_{base}	42.0	45.2
$\text{T5-ColBERT}_{\text{base}}$	45.6	46.8
XTR_{base}	45.0	49.1
$\text{Splade}_{\text{v2}}$ ♣♦	43.3	49.9
$\text{ColBERT}_{\text{v2}}$ ♣♦	-	49.9
GTR_{xxl}	44.2	49.1
$\text{T5-ColBERT}_{\text{xxl}}$	47.3	50.8
XTR_{xxl}	46.6	**52.7**

+3.9

+3.6

♣: cross-encoder distillation

nDCG@10

	MS	**BEIR**
GenQ	88.4	64.2
$\text{PTR}_{\text{retriever}}$	-	69.2
BM25	65.8	63.6
ColBERT	86.5	64.5
GTR_{base}	89.8	64.7
$\text{T5-ColBERT}_{\text{base}}$	91.8	65.5
XTR_{base}	91.0	68.0
GTR_{xxl}	91.6	68.4
$\text{T5-ColBERT}_{\text{xxl}}$	93.3	69.1
XTR_{xxl}	93.0	**71.6**

+3.3

+3.2

Recall@100

Results on EntityQuestions & OpenQA

Dense retrieval significantly falls behind sparse retrieval on this dataset.

	EQ		NQ		TQA		SQD	
	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100
Sparse Component								
BM25	71.4	80.0	62.9	78.3	76.4	83.2	71.1	81.8
DPR_{multi} + BM25	73.3	82.6	*82.6	*88.6	*82.6	*86.5	*75.1	*84.4
Retrieval Pre-training (+ Fine-tuning♣)								
CoCondenser	1.4	8.7	46.8	63.5	13.8	24.3	16.5	28.8
Contriever	63.0	75.1	67.9	80.6	73.9	82.9	63.4	78.2
Spider	66.3	77.4	68.3	81.2	75.8	83.5	61.0	76.0
MSS-DPR♣	60.6	73.7	*81.4	*88.1	*81.9	*86.6	*73.1	*84.5
$ART_{MS\ MARCO}$♣	75.3	81.9	-	-	78.0	84.1	68.4	80.4
GTR_{base}♣	73.3	80.6	49.2	60.8	76.2	83.4	65.9	77.6
GTR_{xxl}♣	75.3	82.5	83.5	89.8	81.7	86.6	70.4	80.6
+4.1 *Retrieval Fine-tuning Only*								
DPR_{multi}	56.7	70.0	*79.5	*86.1	*78.9	*84.8	*52.0	*67.7
ColBERT	-	-	79.1	-	80.3	-	76.5	-
XTR_{base}	79.0	85.2	79.3	88.1	80.3	85.5	78.2	85.9
XTR_{xxl}	**79.4**	**85.9**	**84.9**	**90.5**	**83.3**	**87.1**	**81.1**	**87.6**

Results on LOTTE

Zero-shot Retrieval

	LoTTE Search						LoTTE Forum					
	Writing	Rec.	Sci.	Tech.	Life.	Pooled	Writing	Rec.	Sci.	Tech.	Life.	Pooled
BM25	60.3	56.5	32.7	41.8	63.8	48.3	64.0	55.4	37.1	39.4	60.6	47.2
ColBERT	74.7	68.5	53.6	61.9	80.2	67.3	71.0	65.6	41.8	48.5	73.0	58.2
GTR_{base}	74.1	65.7	49.8	58.1	82.0	65.0	69.2	62.0	33.7	47.6	72.2	54.9
XTR_{base}	77.0	69.4	54.9	63.2	82.1	69.0	73.9	68.7	42.2	51.9	74.4	60.1
$Splade_{v2}$ ♣♦	77.1	69.0	55.4	62.4	82.3	68.9	73.0	67.1	43.7	50.8	74.0	60.1
$ColBERT_{v2}$ ♣♦	80.1	72.3	56.7	66.1	84.7	71.6	76.3	70.8	46.1	53.6	76.9	63.4
GTR_{xxl}	**83.9**	78.0	60.0	69.5	87.4	76.0	79.5	73.5	43.1	62.6	81.9	66.9
XTR_{xxl}	83.3	**79.3**	**60.8**	**73.7**	**89.1**	**77.3**	**83.4**	**78.4**	**51.8**	**64.5**	**83.9**	**71.2**

♣: cross-encoder distillation ♦: model-based hard negatives

Results on Multilingual Retrieval

	ar	bn	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	zh	de	yo	Avg.
BM25	48.1	50.8	35.1	31.9	33.3	55.1	18.3	45.8	44.9	36.9	41.9	33.4	38.3	49.4	48.4	18.0	-	-	-
mDPR	49.9	44.3	39.4	47.8	48.0	47.2	43.5	38.3	27.2	43.9	41.9	40.7	29.9	35.6	35.8	51.2	-	-	-
BM25 + mDPR	67.3	65.4	54.9	**64.1**	**59.4**	67.2	52.3	61.6	44.3	57.6	60.9	53.2	44.6	60.2	59.9	52.6	-	-	-
Trained on English MS MARCO																			
mContriever (en)	55.3	54.2	37.9	34.1	42.6	51.2	31.5	40.6	36.8	38.3	46.2	39.9	44.4	48.7	52.4	27.4	32.9	32.9	41.5
mXTR$_{base}$ (en)	66.1	64.7	49.4	40.5	47.9	62.2	37.5	51.4	46.9	56.8	64.0	49.8	43.0	67.7	69.2	47.2	34.5	40.6	52.2
mXTR$_{xxl}$ (en)	74.1	75.5	**56.0**	52.4	56.1	75.1	51.4	**61.8**	52.0	68.7	67.4	61.3	69.7	76.0	76.9	56.9	51.7	60.3	63.5
Trained on English MS MARCO + MIRACL (16 languages)																			
mContriever	64.6	66.4	41.2	40.3	46.3	61.9	42.9	41.9	44.6	55.6	55.4	48.1	65.3	77.6	69.3	45.9	39.6	41.9	52.7
mXTR$_{base}$	73.0	73.9	46.1	42.6	51.0	70.5	39.3	51.3	54.2	62.3	67.7	54.5	69.7	80.7	76.1	51.4	36.1	46.8	58.2
mXTR$_{xxl}$	**77.8**	**78.4**	52.5	48.9	56.0	**76.0**	**52.9**	61.5	**54.9**	**73.4**	**68.5**	**66.2**	**79.4**	**84.3**	**80.7**	**58.9**	**52.8**	**62.4**	**65.9**

Table 4: nDCG@10 on 18 multilingual retrieval tasks from MIRACL. Each row shows the performance of a single multilingual retrieval model. The last two surprise languages (de and yo) are not included in the training dataset of MIRACL. The last column shows the average over 18 languages.

Analysis

XTR token retrieval for *"what is the **usual** pay for stock associates at michael?"*

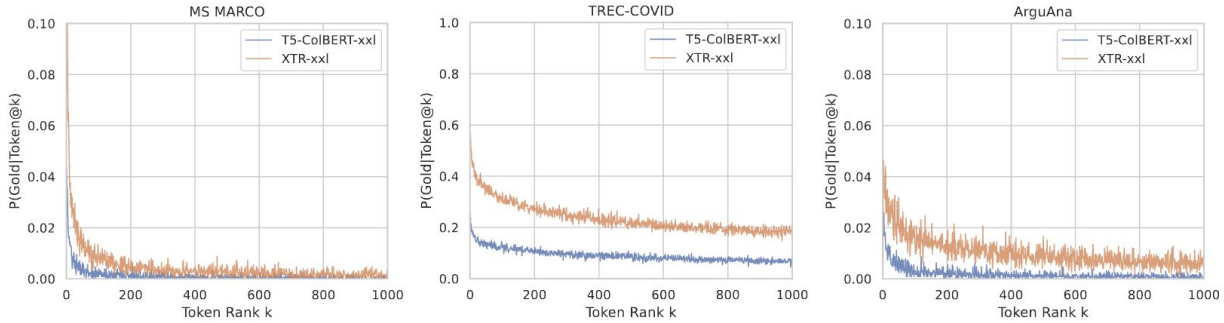
Rank	Token	Context of Token	Relevance
1	**usual**	store manager. 1 salary: the **usual** salary a store manager receives can be anywhere around $52,000 to $115,000 annually.	No
2	**usual**	1 salary: the **usual** salary a store manager receives can be anywhere around $52,000 to $115,000 annually. 2 bonuses: publix provide bonuses that could reach up to $40,000.	No
5	**average**	**average** salaries for michaels stores stock associate: $9. michaels stores hourly pay trends based on salaries posted anonymously by michaels stores employees.	**Yes**
50	**v**	i think the avg starting pay is closer to 30k for asst mgr trainees. it is an hourly position until you are fully trained (40 hours per week).	No
100	**average**	**average** macys salaries. the average salary for macys jobs is $32,000. average macys salaries can vary greatly due to company, location, industry, experience and benefits.	No

2. Given each retrieved document token, how often does it **lexically match** the query token?

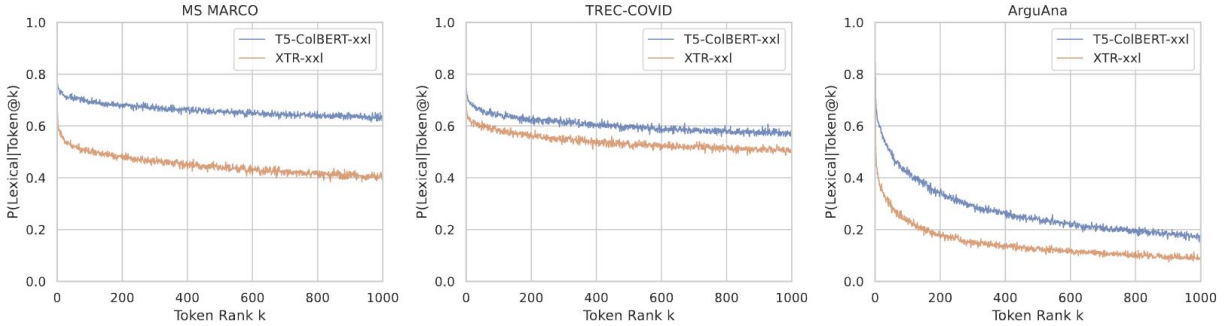
1. Given each retrieved document tokens, how often does it come from the **gold document**?

Analysis

From **gold document**?



Lexically match?



Analysis

Wrong context,
Lexically identical tokens

T5-ColBERT token retrieval for *"what is the **usual** pay for stock associates at michael?"*

Rank	Token	Context of Token	Relevance
1	**usual**	routine passport services: the **usual** waiting time in logan to get your passport is four (4) to eight (8) weeks for routine applications.	No
2	**usual**	the **usual** pay days are the 1st and 16th of each month. for annual educational paraprofessionals there is no payroll lag.	No
5	**usual**	the **usual** part xiii tax rate is 25% (unless a tax treaty between canada and your home country reduces the rate).	No
50	**usual**	this is where one can challenge the judgment debtor's claim. one option creditors have is to try and make a deal with the debtor to take less than 25% (the **usual** amount of a wage levy).	No
100	**usual**	the **usual** maximum inventory is 1 talisman, 26 elemental runes, and 26 pure essence. the ingredients must be brought to an opposing altar ... from the runes being crafted.	No

Right context,
Contextualized tokens

XTR token retrieval for *"what is the **usual** pay for stock associates at michael?"*

Rank	Token	Context of Token	Relevance
1	**usual**	store manager. 1 salary: the **usual** salary a store manager receives can be anywhere around $52,000 to $115,000 annually.	No
2	**usual**	1 salary: the **usual** salary a store manager receives can be anywhere around $52,000 to $115,000 annually. 2 bonuses: publix provide bonuses that could reach up to $40,000.	No
5	**average**	**average** salaries for michaels stores stock associate: $9. michaels stores hourly pay trends based on salaries posted anonymously by michaels stores employees.	**Yes**
50	**v**	i think the a**v**g starting pay is closer to 30k for asst mgr trainees. it is an hourly position until you are fully trained (40 hours per week).	No
100	**average**	**average** macys salaries. the average salary for macys jobs is $32,000. average macys salaries can vary greatly due to company, location, industry, experience and benefits.	No

Analysis

T5-ColBERT token retrieval for "*lauren london age*?"

Rank	Token	Context of Token	Relevance
1	la	laura bush laura lane welch bush (born november 4, 1946) is the wife of the 43rd president of the united states, george w. bush.	No
2	la	is laura branigan dead? laura branigan died on august 26, 2004 at the age of 47.	No
5	la	laika death in space. laika died within hours from overheating. her body temperature got way too hot for her to survive. the heat in her spacecraft had risen to 40 degrees celsius (104 degrees fahrenheit).	No
50	la	singer laura branigan dies at 47 singer laura branigan dies at 47. laura branigan, a grammy-nominated pop singer best known for her 1982 platinum hit gloria, has died.	No
100	la	lauren bacall lauren bacall (born betty joan perske; september 16, 1924 august)	No

XTR token retrieval for "*lauren london age*?"

Rank	Token	Context of Token	Relevance
1	la	lauren london birthday, age, family & biography 33 years, 1 month, 23 days old age lauren london will turn 34 on 05 december, 2018.	Yes
2	la	lauren london current age 33 years old. lauren london height 5 feet 7 inches (1.5 m/ 157 cm) and her weight 119 lbs (54 kg).	Yes
5	la	until now, lauren taylor's age is 28 year old and have gemini constellation. count down 363 days will come next birthday of lauren taylor!	No
50	la	if dwayne johnson, 43, and his longtime girlfriend, lauren hashian, 31, have a baby, would they have a pebble? the furious 7 star and his bae are reportedly expecting their first child together.	No
100	la	laura bush biography after his defeat, bush returned to is oil business and laura became a housewife, but soon returned to politics to help her father-in-law, george h.w. bush's presidential campaign in 1980.	No

Analysis

Model	Imputation	MRR@10	R@1000
T5-ColBERT$_{\text{base}}$	None	0.0	0.0
	top-k' score	27.7	91.8
XTR$_{\text{base}}$	None	22.6	88.7
	$m_i = 0$	36.2	97.3
	$m_i = 0.2$	36.4	97.3
	power-law	**37.7**	**98.1**
	top-k' score	37.4	98.0

Table 5: Impact of training objectives and imputation methods comparing T5-ColBERT and XTR. For both models, we apply $f_{\text{XTR}'}$ during inference. We report MRR@10 and Recall@1000 on the MS MARCO development set. power-law: we fit a power-law curve using top-k' retrieved token scores and estimate the missing similarity based on the curve fitted value at $100k'$. For all imputation variants, we use $k' = 4,000$.

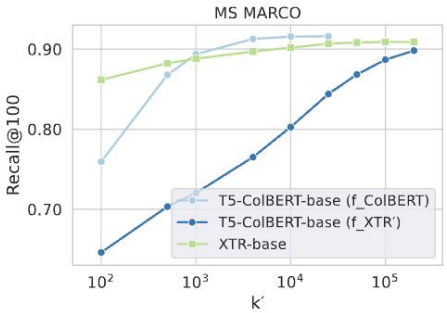


Figure 6: Recall@100 of XTR and T5-ColBERT with different k'. For T5-ColBERT, we either use $f_{\text{XTR}'}$ with the top-k' score imputation or f_{ColBERT}.

Ablation on MS-MARCO

	Training	Refinement	m_i	k'	nDCG@10	R@1000	
XTR_{base}	f_{ret}	f_{approx}	$\mathbf{q}_i^\top \mathbf{d}_{(k')}$	4,000	43.9	98.0	← Retrieved SoM + Imputation
	f_{ret}	f_{approx}	$\mathbf{q}_i^\top \mathbf{d}_{(k')}$	40,000	44.7	**98.4**	
	f_{ret}	f_{approx}	0	4,000	26.4	88.7	← w/o Imputation
	f_{ret}	f_{som}	N/A	4,000	**44.8**	**98.4**	
	f_{som}	f_{approx}	$\mathbf{q}_i^\top \mathbf{d}_{(k')}$	4,000	32.8	91.8	← w/o Retrieved SoM
	f_{som}	f_{approx}	$\mathbf{q}_i^\top \mathbf{d}_{(k')}$	200,000	43.8	**97.9**	
	f_{som}	f_{approx}	0	4,000	0.0	0.0	
	f_{som}	f_{som}	N/A	4,000	**45.9**	97.6	← Full SoM
ColBERT	f_{som}	f_{som}	N/A	1,000	40.1	96.8	
$\text{AligneR}_{\text{base}}$	f_{som}	f_{som}	N/A	4,000	45.6	97.8	

In general, f_{ret} gives better recall.
We need both **Retrieved SoM + Imputation** to make XTR working.

XTR does not need secondary “contrastive” pre-training

(e.g., GTR, Contriever)

Operates directly on contextualized token representations
=> no secondary pre-training is required (e.g. GTR, Contriever)
=> scales better with larger LMs

Could be applied to different sets of vectors in different modalities (images, video, speech, etc).

Thank you!

Rethinking the Role of Token Retrieval in Multi-Vector Retrieval (NeurIPS 2023)
Paper: https://arxiv.org/abs/2304.01982 (to be updated)
Code: Open-source coming soon!

Contact: jinhyuklee@google.com