A Derivatives w.r.t. Similarity Scores

Sum-of-max Here, we use a cross-entropy loss $\mathcal{L}_{\text{CE}}$ with the sum-of-max operator f_{ColBERT} and analyze the derivatives with respect to the token similarity scores.

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp f(Q, D^+)}{\sum_{b=1}^{B} \exp f(Q, D_b)} = -f_{\text{ColBERT}}(Q, D^+) + \log \sum_{b=1}^{B} \exp f_{\text{ColBERT}}(Q, D_b) \tag{5}$$

$$f_{\text{ColBERT}}(Q, D) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{A}_{ij} \mathbf{P}_{ij} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{P}_{i\hat{j}} \tag{6}$$

Here, we denote $\hat{j}$ as the index of the row-wise maximum value, dependent on each i (i.e., $\mathbf{A}_{ij} = 1$). Given the cross-entropy loss with the sum-of-max operator, we compute the gradient with respect to one of the maximum token similarities $\mathbf{P}^+_{i\hat{j}}$ for a positive document $D^+ \in D_{1:B}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{P}^+_{i\hat{j}}} &= -\frac{f(Q, D^+)}{\partial \mathbf{P}^+_{i\hat{j}}} + \frac{1}{\sum_{b=1}^{B} \exp f(Q, D_b)} \frac{\partial}{\partial \mathbf{P}^+_{i\hat{j}}} \sum_{b=1}^{B} \exp f(Q, D_b) \\
&= -\frac{\partial}{\partial \mathbf{P}^+_{i\hat{j}}} \frac{1}{n} \sum_{i=1}^{n} \max_{1 \le j \le m} \mathbf{P}^+_{ij} + \frac{1}{\sum_{b=1}^{B} \exp f(Q, D_b)} \sum_{b=1}^{B} \frac{\partial}{\partial \mathbf{P}^+_{i\hat{j}}} \exp f(Q, D_b) \\
&= -\frac{1}{n} + \frac{1}{\sum_{b=1}^{B} \exp f(Q, D_b)} \sum_{b=1}^{B} \exp f(Q, D_b) \frac{\partial f(Q, D_b)}{\partial \mathbf{P}^+_{i\hat{j}}} \\
&= -\frac{1}{n} + \frac{1}{n} \frac{\exp f(Q, D^+)}{\sum_{b=1}^{B} \exp f(Q, D_b)} = -\frac{1}{n}[1 - P(D^+|Q, D_{1:B})].
\end{aligned}$$

Similarly, the gradient w.r.t. a maximum token similarity $\mathbf{P}^-_{i\hat{j}}$ for a negative document $D^- \in D_{1:B}$ is computed as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{P}^-_{i\hat{j}}} &= -\frac{f(Q, D^+)}{\partial \mathbf{P}^-_{i\hat{j}}} + \frac{1}{\sum_{b=1}^{B} \exp f(Q, D_b)} \frac{\partial}{\partial \mathbf{P}^-_{i\hat{j}}} \sum_{b=1}^{B} \exp f(Q, D_b) \\
&= \frac{1}{n} \frac{\exp f(Q, D^-)}{\sum_{b=1}^{B} \exp f(Q, D_b)} = \frac{1}{n} P(D^-|Q, D_{1:B}).
\end{aligned}$$

Hence, the positive token-level score $\mathbf{P}^+_{i\hat{j}}$ will gradually increase until $P(D^+|Q, D_{1:B}) \to 1$ and the negative token-level score $\mathbf{P}^-_{i\hat{j}}$ will decrease until $P(D^-|Q, D_{1:B}) \to 0$. This shows that the token-level scores are trained based on the document-level scores, which might stagnate the token-level scores. For instance, even if $\mathbf{P}^-_{i\hat{j}}$ is very high—later causing $\mathbf{d}^-_{\hat{j}}$ to be retrieved instead of ones from positive documents—it will not be penalized as long as $P(D^-|Q, D_{1:B})$ is low enough.

In-batch token retrieval Compared to the sum-of-max operator, our in-batch sum-of-max f_{XTR} considers the max values only when they are retrieved over other negative tokens in the mini-batch.

$$f_{\text{XTR}}(Q, D_{1:B}) = \frac{1}{Z} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{A}_{ij} \hat{\mathbf{A}}_{ij} \mathbf{P}_{ij} = \frac{1}{Z} \sum_{i=1}^{n} \mathbf{P}_{i\bar{j}}$$

Here, we denote $\bar{j}$ as the index of the row-wise maximum value that is also within the mini-batch top-k_{train} given q_i (i.e., satisfies both $\mathbf{A}_{ij} = 1$ and $\hat{\mathbf{A}}_{ij} = 1$). If there is no such $\bar{j}$, we simply use $\mathbf{P}_{i\bar{j}} = 0$. We also use a normalizer Z, which is the number of non-zero $\mathbf{P}_{i\bar{j}}$. In this analysis, we assume $Z > 0$ since if every $\mathbf{P}_{i\bar{j}}$ is zero, the gradient is undefined.

The gradient w.r.t. the maximum token similarity $\mathbf{P}^+_{i\bar{j}}$ (non-zero) for a positive document $D^+ \in D_{1:B}$ is computed as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{P}^+_{i\bar{j}}} &= -\frac{f(Q, D^+)}{\partial \mathbf{P}^+_{i\bar{j}}} + \frac{1}{\sum_{b=1}^{B} \exp f(Q, D_b)} \frac{\partial}{\partial \mathbf{P}^+_{i\bar{j}}} \sum_{b=1}^{B} \exp f(Q, D_b) \\ &= -\frac{1}{Z^+}\big[1 - \frac{\exp f(Q, D^+)}{\sum_{b=1}^{B} \exp f(Q, D_b)}\big] \\ &= -\frac{1}{Z^+}\big[1 - P(D^+|Q, D_{1:B})\big].\end{aligned}$$

This is a very similar result compared to the sum-of-max operator except that 1) the gradient is defined only when $\mathbf{P}^+_{i\bar{j}}$ is non-zero (i.e. retrieved) and 2) it is dependent on Z^+, which means that the gradient will be large whenever there is a small number of retrieved tokens from the positive document. If only a handful of tokens are retrieved for D^+, our objective function increases $\mathbf{P}^+_{i\bar{j}}$.

For negative similarity score $\mathbf{P}^-_{i\bar{j}}$, we have the following:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{P}^-_{i\bar{j}}} &= -\frac{f(Q, D^+)}{\partial \mathbf{P}^-_{i\bar{j}}} + \frac{1}{\sum_{b=1}^{B} \exp f(Q, D_b)} \frac{\partial}{\partial \mathbf{P}^-_{i\bar{j}}} \sum_{b=1}^{B} \exp f(Q, D_b) \\ &= -\frac{1}{Z^-}\big[- \frac{\exp f(Q, D^-)}{\sum_{b=1}^{B} \exp f(Q, D_b)}\big] \\ &= \frac{1}{Z^-} P(D^-|Q, D_{1:B}).\end{aligned}$$

Again, it is similar to the sum-of-max result, but it depends on Z^-. In this case, even when $P(D^-|Q, D_{1:B})$ is low, if there is a small number of retrieved tokens from D^- (i.e., small Z^-), $\mathbf{P}^-_{i\bar{j}}$ will be decreased significantly. Note that when Z^- is large, Z^+ naturally becomes smaller as they compete for in-batch token retrieval, which causes positive tokens to have higher scores.

B Inference Complexity

We compare the complexity of ColBERT and XTR during the scoring stage in terms of FLOPs. We do not measure the complexity for the online query encoding and maximum inner product search (MIPS), which have been extensively studied for both dual encoders and multi-vector retrieval [Santhanam et al., 2022a,b, Guo et al., 2020].

For the scoring stage, both ColBERT and XTR have $\mathcal{O}(nk')$ candidate documents. Here, we assume the worst case nk' where each document token comes from a unique document. For each candidate document, ColBERT loads a set of document vectors of $\bar{m}d$ floating points ($\bar{m}$ = average document length) and computes eq. (1) with the query vectors of nd floating points. Computing eq. (1) per candidate document requires $2n\bar{m}d$ FLOPs for token-level inner products, $n\bar{m}$ for finding the row-wise max, and n for the final average. In total, ColBERT requires $n^2k'(2\bar{m}d + \bar{m} + 1)$ FLOPs for the scoring stage. Note that this does not include the latency of loading the $\mathcal{O}(nk'\bar{m}d)$ floating points onto the memory, which amounts up to 450MB per query when $n = 16, k' = 1000, \bar{m} = 55, d = 128$.

On the other hand, XTR first imputes the missing similarity, which is simply done by caching the k'-th token retrieval score for each query token. Then, each of nk' candidate documents requires $n\bar{r}$ FLOPs for finding row-wise max and n for the average where $\bar{r}$ is the average number of retrieved tokens per each candidate document. In total, we have $n^2k'(\bar{r} + 1)$ FLOPs. Table 1 shows the estimated FLOPs of the two models. XTR reduces the FLOPs at the scoring stage by 4000× making multi-vector retrieval more efficient and practical.

C Implementation Details

XTR uses k_{train} for retrieving in-batch document tokens. Since we retrieve over mini-batches, the size of mini-batch affects the performance for different k_{train}, which is shown in §5.3. In our experiments, we tried $k_{\text{train}} = \{32, 64, 128, 256, 320\}$ for each batch size and choose the best model based on their performance on the MS MARCO development set. For inference, XTR uses k' for the token retrieval. We use $k' = 40,000$, which is possible due to the efficient scoring stage of XTR.[6] We analyze the effect of using different k''s as well as its relationship to k_{train} in §5.3. We initialize XTR from the base and xxl versions of the T5 encoder [Raffel et al., 2020] and provide XTR_{base} and XTR_{xxl}. For multilingual XTR, we initialize XTR from mT5 [Xue et al., 2021]. We fine-tune XTR for 50,000 iterations with the learning rate to 1e-3. Up to 256 chips of TPU v3 accelerator were used depending on the size of the model. We use ScaNN [Guo et al., 2020] for the MIPS during the token retrieval stage. For BEIR, we use 13 datasets (AR: ArguAna. TO: Touché-2020. FE: Fever. CF: Climate-Fever. SF: Scifact. CV: TREC-COVID. NF: NFCorpus. NQ: Natural Questions. HQ: HotpotQA. FQ: FiQA-2018. SD: SCIDOCS. DB: DBPedia. QU: Quora).

Baselines There are two main paradigms on training retriever models for the out-of-domain evaluation. The first group trains a single retriever for each dataset (or domain) by generating queries for each out-of-domain corpus. Typically, this approach generates N datasets to train N independent models for N different domains. For this *one-retriever-per-domain* approaches, we include GenQ [Thakur et al., 2021], GPL [Wang et al., 2022], and Promptagator [Dai et al., 2022]. The second group builds a single retriever—typically trained on a large-scale IR dataset such as MS MARCO—and directly applies it on the out-of-domain corpora and queries. For this *one-retriever-for-all* approaches, we present results of state-of-the-art retrievers including $\text{Splade}_{\text{v2}}$ [Formal et al., 2021], $\text{ColBERT}_{\text{v2}}$ [Santhanam et al., 2022b], and GTR_{xxl} [Ni et al., 2021]. We also show the results of $\text{T5-ColBERT}_{\text{xxl}}$ [Qian et al., 2022], which is a T5-initialized ColBERT model and shares the same backbone LM and training dataset with XTR. Note that T5-ColBERT uses the heavy scoring stage based on the original sum-of-max. All of our one-retriever-for-all baselines, as well as XTR, are trained on English MS MARCO, unless otherwise stated.

[6] In fact, XTR with $k' = 40,000$ has still two-to-three orders of magnitude cheaper scoring stage than ColBERT with $k' = 1,000$ and T5-ColBERT with $k' = 4,000$.

D Additional Results

In Table D.1, we show Recall@100 on BEIR.

	MS	AR	TO	FE	CF	SF	CV	NF	NQ	HQ	FQ	SD	DB	QU	Avg.
One Retriever per Domain															
GenQ	88.4	97.8	45.1	92.8	45.0	89.3	45.6	28.0	86.2	67.3	61.8	33.2	43.1	98.9	64.2
$\text{PTR}_{\text{retriever}}$	-	98.9	47.5	94.1	53.1	91.8	55.9	30.6	89.8	74.6	76.5	41.6	46.3	99.6	69.2
One Retriever for All															
BM25	65.8	94.2	53.8	93.1	43.6	90.8	49.8	25.0	76.0	74.0	53.9	35.6	39.8	97.3	63.6
ColBERT	86.5	91.4	43.9	93.4	44.4	87.8	46.4	25.4	91.2	74.8	60.3	34.4	46.1	98.9	64.5
GTR_{base}	89.8	97.4	44.3	92.3	52.2	87.2	41.1	27.5	89.3	67.6	67.0	34.0	41.8	99.6	64.7
$\text{T5-ColBERT}_{\text{base}}$	91.8	76.0	49.9	90.4	46.2	91.3	55.4	27.6	90.5	78.3	63.0	34.2	50.5	97.9	65.5
XTR_{base}	91.0	92.1	50.8	92.5	51.6	90.5	57.3	28.0	91.6	80.7	63.5	34.8	52.0	98.9	68.0
GTR_{xxl}	91.6	98.3	46.6	94.7	55.6	90.0	40.7	30.0	94.6	75.2	78.0	36.6	49.4	99.7	68.4
$\text{T5-ColBERT}_{\text{xxl}}$	93.3	81.4	50.1	91.7	49.8	94.6	60.3	29.0	95.5	81.6	72.5	38.5	54.6	99.1	69.1
XTR_{xxl}	93.0	95.6	52.7	93.7	56.2	95.0	62.1	30.7	95.8	82.2	73.0	39.4	54.5	99.3	**71.6**

Table D.1: Recall@100 on MS-MARCO and BEIR. The last column shows the average over 13 BEIR benchmarks. Compared to GTR, T5-ColBERT only marginally improves the recall. On the other hand, XTR greatly improves the recall showing the importance of having a better token retrieval.

In Table D.2, we show nDCG@10 and Recall@100 on BEIR with different k'.

k'	MS	AR	TO	FE	CF	SF	CV	NF	NQ	HQ	FQ	SD	DB	QU	Avg.
nDCG@10															
40,000	**45.0**	40.7	**31.3**	**73.7**	**20.7**	71.0	**73.6**	34.0	**53.0**	**64.7**	**34.7**	14.5	**40.9**	86.1	**49.1**
1,000	43.2	**44.6**	29.0	72.1	20.4	**71.7**	67.5	**34.2**	49.8	61.3	33.0	**15.9**	37.0	**86.3**	47.9
Recall@100															
40,000	**91.0**	92.1	**50.8**	92.5	51.6	90.5	**57.3**	28.0	**91.6**	**80.7**	**63.5**	34.8	**52.0**	98.9	**68.0**
1,000	88.8	**96.4**	48.0	**92.5**	**53.3**	**93.1**	48.1	**28.6**	88.8	78.3	62.5	**37.0**	47.0	**99.1**	67.1

Table D.2: nDCG@10 and Recall@100 of XTR_{base} on MS-MARCO and BEIR with different k'. The last column shows the average over 13 BEIR benchmarks.

E Qualitative Analysis

In Table 6-E.5, we show token retrieval results from T5-ColBERT and XTR.

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

Rank	Token	Context of Token	Relevance
1	**la**	**la**ura 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? **la**ura branigan died on august 26, 2004 at the age of 47.	No
5	**la**	laika death in space. **la**ika 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 **la**ura 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**	**la**uren 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 **la**uren london will turn 34 on 05 december, 2018.	Yes
2	**la**	**la**uren 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, **la**uren 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, **la**uren 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 **la**ura became a housewife, but soon returned to politics to help her father-in-law, george h.w. bush's presidential campaign in 1980.	No

Table E.1: Token retrieval example from MS MARCO for the token *"la"* in the query *"lauren london age"*. Among the top 100 retrieved tokens, 100% of T5-ColBERT tokens are lexically identical as the query token `la` and 100% of XTR tokens are also lexically identical. However, top retrieved results from XTR contain the correct entity (`Lauren London`) while those from T5-ColBERT are about wrong entities (`Laura Bush`, `Laura Branigan`, etc.).

T5-ColBERT token retrieval for "***temple*** *university student population*?"			
Rank	Token	Context of Token	Relevance
1	**temple**	about **temple** university tuition, cost, financial aid, scholarships, and admission rates	No
2	**temple**	overview the application fee at **temple** university is $55. it is selective, with an acceptance rate of 61.7 percent and an early acceptance rate of 78 percent.	No
5	**temple**	the application fee at **temple** university is $55. it is selective, with an acceptance rate of 61.7 percent and an early acceptance rate of 78 percent.	No
50	**temple**	**temple** university staff accountants earn $52,000 annually, or $25 per hour, which is 14% higher than the national average for all staff accountants at $45,000 annually and 16% lower than the national salary average for all working americans	No
100	**temple**	browse expedia's selection and check out the best hotels close to **temple** university for the world-class spas and restaurants, or snatch up one of the cheap hotel deals near temple university	No

XTR token retrieval for "***temple*** *university student population*?"			
Rank	Token	Context of Token	Relevance
1	**temple**	by gender, the school has 18,009 male and 19,476 female students. by race/ethnicity, 20,664 white, 4,466 black, and 3,819 asian students are attending at **temple** university.	Yes
2	**temple**	below tables and charts represent the enrollment statistics including school degree, gender, race/ethnicity, and tranfer-in students at the school. at **temple** university, 37,485 students are enrolled	Yes
5	**temple**	temple university the big picture: how many students were on campus in fall 2015? of the 28,886 new freshman applicants, 56% were admitted and 31% of the admitted students enrolled at **temple** university in fall 2015.	Yes
50	**temple**	**temple** university was founded in 1884 by russell conwell, a yale-educated boston lawyer, orator, and ordained baptist minister	No
100	**temple**	kaiser said **temple**'s endowment fund is low because the university is late to the idea of fundraising.	No

Table E.2: Token retrieval example from MS MARCO for the token *"temple"* in the query *"temple university student population?"*. Among the top 100 retrieved tokens, 100% of T5-ColBERT tokens are lexically identical as the query token `temple` and 100% of XTR tokens are also lexically identical. However, top retrieved results from XTR are of the correct context (`student population`) while those from T5-ColBERT are off-topic (e.g., `tuition`, `salary`, etc.).

T5-ColBERT token retrieval for *"**aire** is expressed in some skin tumors"*			
Rank	Token	Context of Token	Relevance
1	**aire**	acids: structures, properties, and functions (university science books, sausalito, ca, 2000). humans expressing a defective form of the transcription factor **aire** (autoimmune regulator) develop multiorgan autoimmune disease.	No
2	**aire**	the primary biochemical defect in apeced is unknown. we have isolated a novel gene, **aire**, encoding for a putative nuclear protein featuring two phd-type zinc-finger motifs, suggesting its involvement in transcriptional regulation.	No
5	**aire**	control of central and peripheral tolerance by **aire**. the negative selection of self-reactive thymocytes depends on the expression of tissue-specific antigens by medullary thymic epithelial cells.	No
50	**aire**	we found that a human patient and mice with defects in aire develop similar lung pathology, demonstrating that the **aire**-deficient model of autoimmunity is a suitable translational system in which to unravel fundamental mechanisms of ild pathogenesis.	No
100	**air**	cool **air** initiates just downstream of the major sense transcript poly(a) site and terminates either early or extends into the flc promoter region.	No

XTR token retrieval for *"**aire** is expressed in some skin tumors"*			
Rank	Token	Context of Token	Relevance
1	**aire**	keratin-dependent regulation of **aire** and gene expression in skin tumor keratinocytes expression of the intermediate filament protein keratin 17 (k17) is robustly upregulated in inflammatory skin diseases and in many tumors....	Yes
2	**aire**	the thymic transcription factor autoimmune regulator (aire) prevents autoimmunity in part by promoting expression of tissue-specific self-antigens, which include many cancer antigens. for example, **aire**-deficient patients are predisposed to vitiligo, an autoimmune disease of melanocytes that is often triggered by efficacious immunotherapies against melanoma.	Yes
5	**aire**	aire regulates negative selection of organ-specific t cells autoimmune polyendocrinopathy syndrome type 1 is a recessive mendelian disorder resulting from mutations in a novel gene, **aire**, and is characterized by a spectrum of organ-specific autoimmune diseases.	No
50	**aire**	here we demonstrate a novel role for a cd4+3- inducer cell population, previously linked to development of organized secondary lymphoid structures and maintenance of t cell memory in the functional regulation of **aire**-mediated promiscuous gene expression in the thymus.	No
100	**air**	this localization is dependent on the presence of sperm in the spermatheca. after fertilization, **air**-2 remains associated with chromosomes during each meiotic division.	No

Table E.3: Token retrieval example from MS MARCO for the token *"aire"* in the query *"aire is expressed in some skin tumors"*. Among the top 100 retrieved tokens, 77% of T5-ColBERT tokens are lexically identical as the query token `aire` and 77% of XTR tokens are also lexically identical. Top retrieved results from XTR are relevant to the query (about `cancer`, `tumor`, `skin`, and `melanocyte`), while those from T5-ColBERT are off-topic.

T5-ColBERT for *"women with a higher birth weight are more likely to develop breast cancer **later** in life"*

Rank	Token	Context of Token	Relevance
1	**later**	context exposure to cardiovascular risk factors during childhood and adolescence may be associated with the development of atherosclerosis **later** in life.	No
2	**later**	n despite the high incidence of febrile seizures, their contribution to the development of epilepsy **later** in life has remained controversial.	No
5	**later**	prospectively collected data from two intervention studies in adults with severe malaria were analysed focusing on laboratory features on presentation and their association with a **later** requirement for rrt.	No
50	**later**	they did have a limited amount of proteolytic activity and were able to kill s. aureus. with time, the nuclear envelope ruptured, and dna filled the cytoplasm presumably for **later** lytic net production	No
100	**late**	finally, we address the need for a careful consideration of potential benefits of bisphosphonate therapy and the risk for osteonecrosis of the jaw, a recently recognized **late**-toxicity of their use.	No

XTR for *"women with a higher birth weight are more likely to develop breast cancer **later** in life."*

Rank	Token	Context of Token	Relevance
1	**later**	life course breast cancer risk factors and adult breast density (united kingdom) objective to determine whether risk factors in childhood and early adulthood affect **later** mammographic breast density.	Yes
2	**later**	exposure to cardiovascular risk factors during childhood and adolescence may be associated with the development of atherosclerosis **later** in life.	No
5	**subsequent**	emerging evidence suggests an association between female prenatal experience and her **subsequent** risk of developing breast cancer.	Yes
50	**later**	our nested case–control study of eh progression included 138 cases, who were diagnosed with eh and then with carcinoma (1970–2003) at least 1 year (median, 6.5 years) **later**, and 241 controls....	No
100	**during**	obesity and being overweight **during** adulthood have been consistently linked to increased risk for development of dementia later in life, especially alzheimer's disease.	No

Table E.4: Token retrieval example from Scifact for the token *"later"* in the query *"women with a higher birth weight are more likely to develop breast cancer later in life"*. Among the top 100 retrieved tokens, 72% of T5-ColBERT tokens are lexically identical as the query token `later` while only 33% of XTR tokens are lexically identical. Top retrieved results from XTR can retrieves synonyms (`sebsequent`) from relevant context, while those from T5-ColBERT are off-topic.

T5-ColBERT for *"venules have a **thinner** or absent smooth layer compared to arterioles."*

Rank	Token	Context of Token	Relevance
1	**thinner**	platelet cd40l is associated with smaller plaques and **thinner** caps, while p-selectin is associated with smaller core size. conclusions: blood cell activation is significantly associated with atherosclerotic changes of the carotid wall.	No
2	**thin**	the periosteum is a **thin**, cellular and fibrous tissue that tightly adheres to the outer surface of all but the articulated surface of bone and appears to play a pivotal role in driving fracture pain.	No
5	**thin**	immunohistological scoring showed significantly (p<0.0001) higher median 5hmc levels in bcn and dcn than in **thin** ssm, thick ssm, and cmd.	No
50	**weak**	subarachnoid haemorrhage (1·43 [1·25-1·63]), and stable angina (1·41 [1·36-1·46]), and **weak**est for abdominal aortic aneurysm (1·08 [1·00-1·17]).	No
100	**slight**	the ucp-2 gene expression was widely detected in the whole body with substantial levels in the wat and with **slight** levels in the skeletal muscle and bat.	No

XTR for *"venules have a **thinner** or absent smooth layer compared to arterioles."*

Rank	Token	Context of Token	Relevance
1	**thinner**	platelet cd40l is associated with smaller plaques and **thinner** caps, while p-selectin is associated with smaller core size. conclusions: blood cell activation is significantly associated with atherosclerotic changes of the carotid wall.	No
2	**thin**	the periosteum is a **thin**, cellular and fibrous tissue that tightly adheres to the outer surface of all but the articulated surface of bone and appears to play a pivotal role in driving fracture pain.	No
5	**thick**	in dense fibrotic zones, **thick**ening of the arterial and venous wall with severe luminal narrowing was present in each patient.	No
50	**small**	we assessed vasomotor function of the adipose microvasculature using videomicroscopy of **small** arterioles isolated from different fat compartments.	No
100	**particle**	context circulating concentration of lipoprotein(a) (lp[a]), a large glycoprotein attached to a low-density lipoprotein-like **particle**, may be associated with risk of coronary heart disease (chd) and stroke.	No

Table E.5: Token retrieval example from Scifact for the token *"thinner"* in the query *"vanules have a thinner or absent smooth later compared to arterioles"*. Among the top 100 retrieved tokens, only 1% of T5-ColBERT tokens are lexically identical as the query token `thinner` and only 1% of XTR tokens are also lexically identical.