9 SUPPLEMENTARY MATERIAL

9.1 COMPUTATIONAL RESOURCE USAGE

We pre-train all models on 8 NVIDIA H100 GPUs. In the slowest setting, training a biencoder for a single unsupervised epoch (235M pairs) takes approximately one day. Training our contextual archiecture for a single epoch takes approximately two days. Shorter sequence-length experiments are 10-20x faster, and can be run on a single GPU.

9.2 INITIAL EXPERIMENTS

We conducted two preliminary experiments to verify (i) the need for contextual training strategy and (ii) the need for in-batch false negative filtering when doing adversarial contrastive learning on a real dataset.

Preliminary experiment (i). We conduct a preliminary experiment to verify this issue. Starting from several trained retrieval systems we compute performance on a variety of different tasks from the BEIR dataset. Additionally we compute the IDF statistics from the datasets, and compare the divergence from the base IDF statistics of the training set. Figure 8 shows that datasets with high-divergence have very high correlation with the accuracy degradation of models when measured in comparison to BM25, which is able to measure and adapt to statistics of the test corpus.
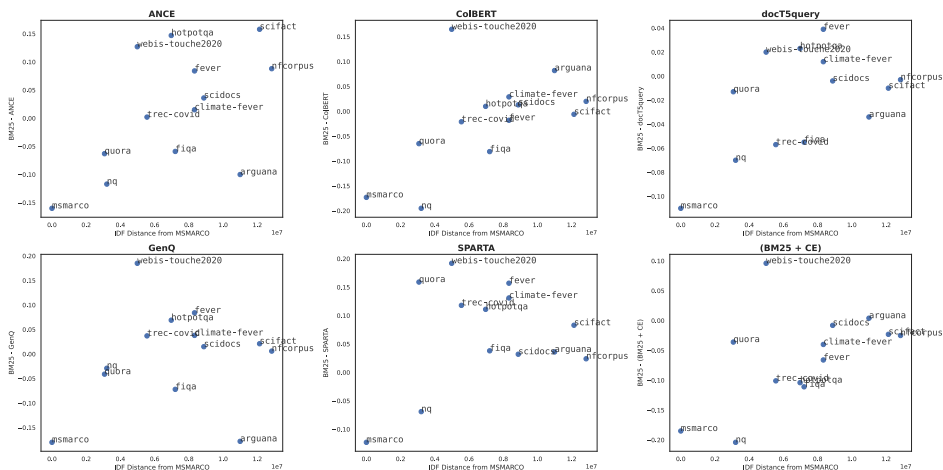

Figure 8: Analysis of domain shift for popular neural retrieval methods. Performance difference from BM25 (y-axis) correlates with the different in IDF of the test corpus $\mathcal{D}$ form the training corpus $\mathcal{D}_T$.

Preliminary experiment (ii). We select a random document from an unsupervised corpus and look at its nearest neighbors, displayed in Table 3. We observe that the nearest neighbors to a given document in a large corpus are very close; in fact, many of them could be considered valid documents for the given query as well.

This challenge motivates our embedding contextualization. In this section, we describe two complementary methods for remediation, (a) a contextual training method, (b) a contextual encoding method.

9.3 INTERACTIONS BETWEEN CONTRASTIVE LOSS AND DISTRIBUTED DATA PARALLEL

The authors note that it can be notoriously difficult to train models using both contrastive loss and the distributed data parallel (DDP) setting. In particular, when aggregating samples between GPUs, if any artifact reveals which GPU a model came from (for example, if the GPU model weights are initialized slightly differently) than the model can quickly deteriorate to a suboptimal solution, each

Table 3: Nearest-neighbors to a single query in a large unsupervised dataset.

Query	Document
looks like my card payment was duplicated after all. [...]	
why is there an extra €1 fee in my statement?	why is there an extra charge on my statement?
what is this fee for card payment?	why was a fee charged for my card payment?
why do i have duplicate transactions for one purchase?	why was my transaction charged twice?
i have two of the same charges on my account!	why was my transaction charged twice?
my transaction went through but i was charged a fee. why?	why was a fee charged for my transfer?
my account shows i have been charged twice for the same meal. [...]	
will i get extra charges?	why was a fee charged for my transfer?
i got charged in double and want a refund	why was my transaction charged twice?
where do i pay with my debit or credit card?	why is my card not accepted?
why did i get charged a fee for my card payment?	why was a fee charged for my card payment?
my statement shows different transaction times.	why was my transaction charged twice?

GPU learning a different final model and "cheating" to classify samples based on which GPU they came from.

This issue is made extra difficult by the fact that gradient-syncing must be disabled for large-batch contrastive learning to work efficiently. If gradient syncing becomes totally disabled, the training silently diverges as each model learns a degenerate solution. We advise practitioners to take care when controlling gradient-syncing and run many control experiments to determine performance equivalence between DDP and non-DDP scenarios.

One potential benefit of our method is that it greatly decreases the number of hard negatives required per batch, which means that negative-sharing across GPUs may not be necessary in most settings. If possible, the most sanity-preserving way to perform contrastive training could be to

9.4 REMOVING POSITIONALITY WITH ROTARY EMBEDDINGS

One detail of our model architecture is that it does not track positionality between dataset input tokens. Although disabling positionality would be trivial an a BERT-like encoder model that uses learned positional embeddings, we use a version of BERT with *rotary* positional embeddings which inject positional information at each layer of the transformer. To circumvent this step, we modify the model internals to set dataset input tokens to zero for the self-attention step only, and add a residual connection propagating the dataset input tokens past the attention phase.

9.5 ADDITIONAL RESULTS

Section 9.5 show sweeps over batch and cluster sizes under our small experimental settings when performing unsupervised pretraining with contextual architecture. We see similar trends to those observed with the biencoder architecture, however we note that performance is higher across the board and our transductive model is able to perform well even at higher cluster sizes and low batch sizes.

One confounding factor in these experiments is that since the number of contextual documents is fixed, the number of different contextual inputs seen during training decreases with higher batch size.
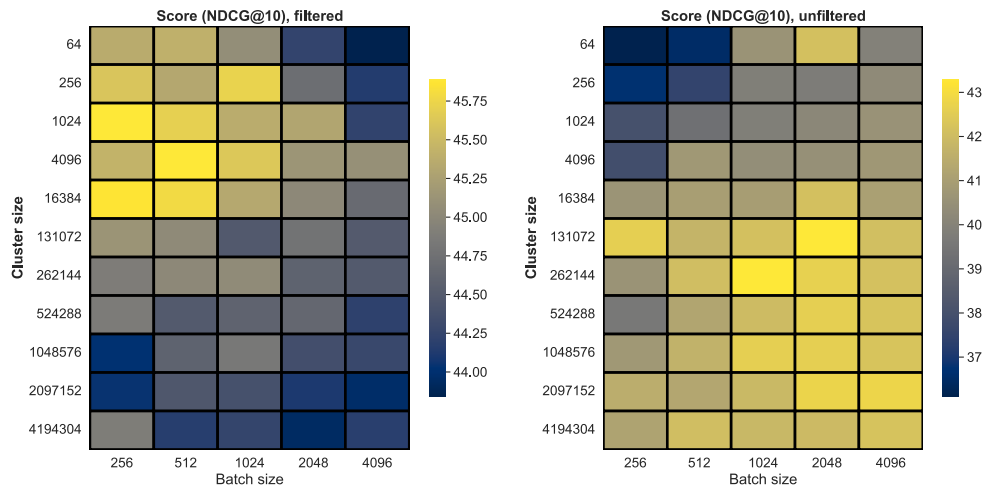
Figure 9: Contextual performance with filtering (left) and without (right) across batch and cluster sizes during unsupervised contrastive pre-training. Here, clustering with small cluster sizes clearly improves performance, and larger batch sizes do not.
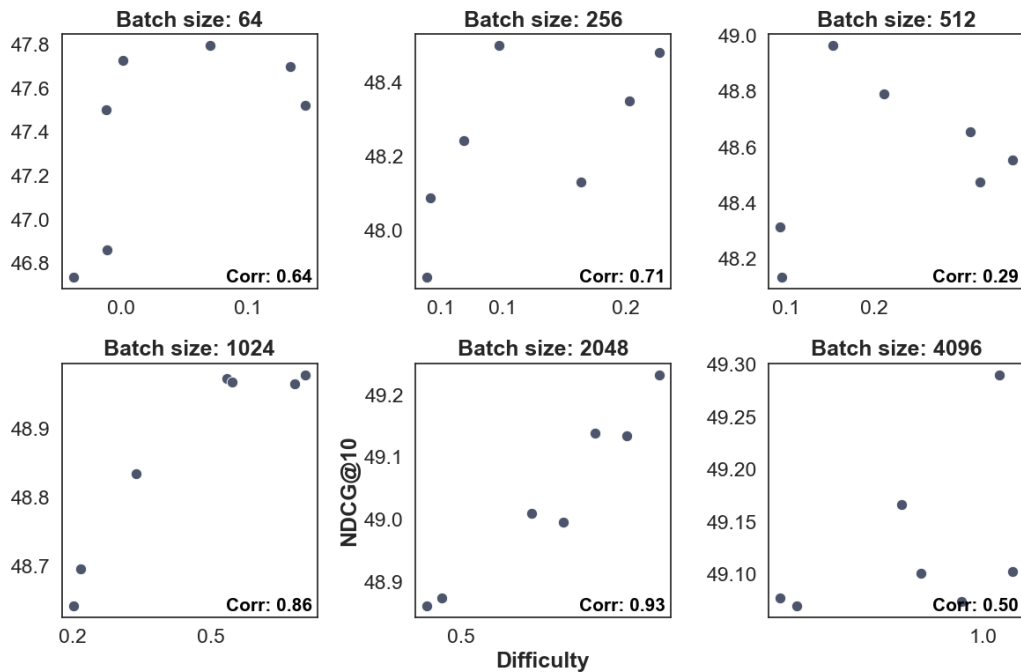

Figure 10: Correlation between batch difficulty and perforamnce after supervised training.

This might explain part of why performance stagnates with higher batch sizes; increasing the batch size decreases the total number of learning examples seen by our contextual model.

Supervised training: difficulty correlations. In Section 9.5 we plot the correlation between batch difficulty and downstream performance across cluster sizes (and within batch sizes) in the supervised setting. In this case we also see the best performance through the most difficult clusters.
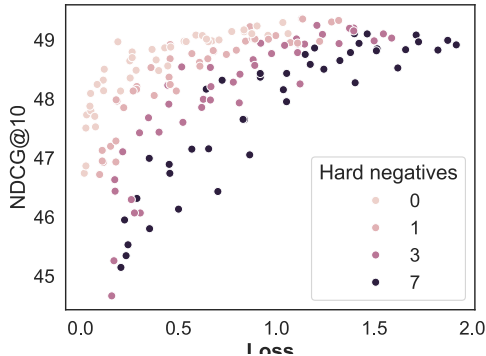
Figure 11: Performance of all supervised models, across numbers of hard negatives.
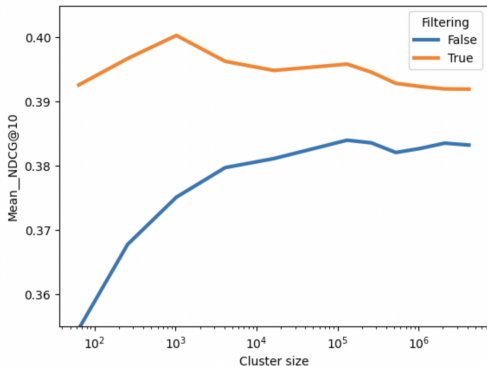

Figure 12: Model performance vs. cluster size with and without filtering. When false negative filtering is enabled, we see more improvements in performance from clustering at small cluster sizes.
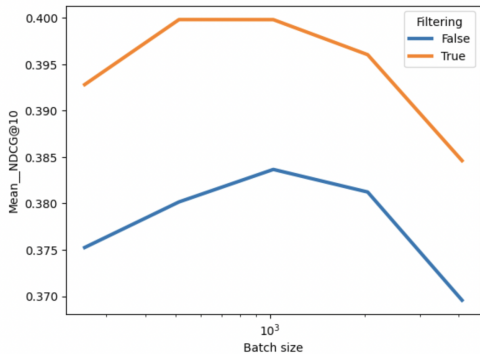
Figure 13: Model performance vs. batch size with and without filtering. With and without filtering, the optimal batch size ranges between 10^2 and 10^4; performance starts to decrease as batch size grows too large.

Supervised training: full results. We plot the full results of all supervised training experiments in Section 9.5. Our experiments in this setting (using the mined negatives from the Nomic supervised meta-datasets) generally show *decreasing* performance with additional hard negatives.

TSP Packing. We compare randomly packing clusters into batches vs. a greedy traveling salesman-style solution, similar to (Shi et al., 2024). In our scenario, we first cluster datapoints, then find the centroid embedding of each cluster. We begin packing by randomly selecting a cluster, and then choose the next cluster by finding the cluster with the closest centroid to the current one. Results are shown in Figure 14. Although these results appear slightly noisy, we see an improvement from TSP-style packing especially at smaller cluster sizes (where packing has an outsized impact). We therefore opt to use this packing procedure for our main model.

Impact of context size We consider contextual embeddings might move in space as their conditioning varies. Section 9.5 displays a few qualitative examples. We generate embeddings for randomly sampled documents from the TREC-Covid dataset and visualize their embeddings with PCA, where unique document inputs with different contextual embeddings are visualized in the same color. By changing only the conditioning we reshape the embedding space and our model produces different embedding for the same text. Note that although the embeddings are clearly moving in response to changing the contextual inputs, they still remain closer to each other than to different documents.
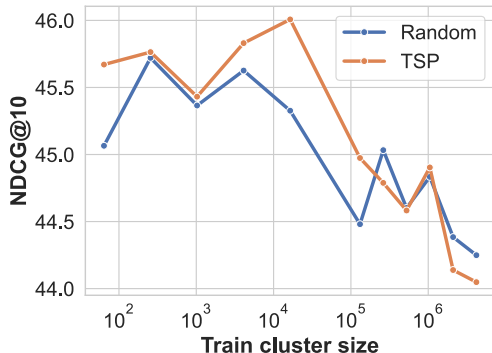
Figure 14: Pre-training with TSP vs. random batching across cluster sizes.

Figure 15: Each color indicates a single document input d. Different points represent different values $\phi(d; \mathcal{D})$ for different contexts.
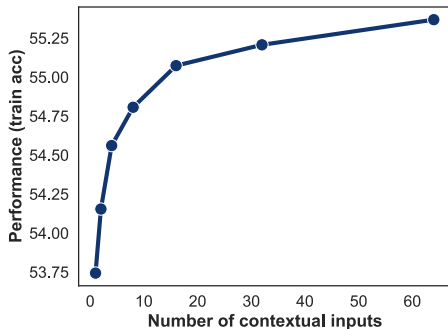

Figure 16: Performance of CDE model as the number of contextual examples increases.

We also consider how additional context is improving our model. Because the model includes an optional null token, we can supply any number of contextual inputs. We plot our model's performance across context sizes in Figure 9.5. We see that our model is able to utilize partial context window sizes, and even perform reasonably with no context (i.e. all null token inputs) but offers the best performance given a full context window size.

9.6 CLUSTER TEXT EXAMPLES

We include random examples from a cluster gathered from our supervised dataset, shown in Table 4. This particular cluster appears to be a combination of documents about county populations in the Untied States (in Kentucky, Iowa, Pennsylvania, etc.) and documents about criminal trials (mentioning hearings, depositions, and courts).

9.7 TASK PREFIXES

Prefixes are hand-written for each dataset in both meta-training sets. We follow the same prefix selection procedure as Nussbaum et al. (2024), inspired by Reimers et al. (2023):

- `search_query`
- `search_document`
- `classification`

19

query	document
population of breckenridge mi	breckenridge, michigan. breckenridge is a village in gratiot county in the u. s. state of michigan. the population was 1, 328 at the 2010 census. the village is located in wheeler township.
can a deposition be used in a criminal case	depositions are commonly used in civil litigation (suits for money damages or equitable relief) [...]
what cases require strict scrutiny	the strict scrutiny standard is one of three employed by the courts in reviewing laws and government policies. the rational basis [...]
function of state supreme courts	it has also initiated several programs designed to improve the effectiveness of the court system. a primary function of the supreme court is to ensure [...]
what is the population in idaho	idaho ' s population grows to nearly 1. 7 million. idaho ' s population grew by 1. 2 percent between mid - 2014 and mid - 2015, the 12th strongest increase among the states and four - tenths of a percentage point ahead of the national growth rate.
what is the population of manson, ia	manson, iowa. manson is a city in calhoun county, iowa, united states. the population was 1, 690 at the 2010 census.
what happens after a sentencing hearing	find answers. sentencing. after a criminal defendant is convicted or pleads guilty, a judge will decide [...]
flathead county population	flathead county, montana. flathead county is a county located in the u. s. state of montana. as of the 2010 census, the population was 90, 928, making it [...]
whiting, ks population	the city of whiting had a population of 177 as of july 1, 2017. whiting ranks in the lower quartile for population density and diversity index when compared to the other cities, towns [...]
what is the population of lewiston id	lewiston, id population and races. as of 2010 - 2014, the total population of lewiston is 32, 178, which is 4. 12% more than it was in 2000. [...]
what happens if you don't show up for jury	what happens if you don't show up for jury duty in california? a : according to california courts, judicial branch of california, if a citizen fails to show up for jury duty, the juror can accrue fines up to $1,500. if service presents an undue hardship, a juror can request a postponement or to be excused. otherwise, citizens are not exempt from jury duty.
population of clearfield county pa	clearfield is a borough and the county seat of clearfield county, pennsylvania, united states. the population was 6, 215 at the 2010 census, and the borough is part of the dubois, pa micropolitan statistical area, as well as the larger state college - dubois, pa combined statistical area.
how long can it take for a trial	the preliminary hearing phase of the trial usually takes place 5 - 6 days after an arraignment. in the case of a misdemeanor [...]
population clinton ky	clinton county is a county located in the u. s. state of kentucky. as of the 2010 census, the population was 10, 272. its county seat is albany. the county was formed in 1835 and named for dewitt clinton, the seventh governor of new york. it is a prohibition or dry county.
population of iosco county michigan	with 25, 420 people, iosco county is the 55th most populated county in the state of michigan out of 83 counties. but watch out, iosco county, because gladwin county with 25, 411 people and manistee county with 24, 420 people are right behind you.

Table 4: Sixteen samples from a cluster our algorithm finds in the supervised training data. The full cluster size is 256 points out of a dataset of $1.5M$.

Table 5: Distribution of pretraining datasets curated in Nussbaum et al. (2024).

Dataset	Datapoints	% Dataset
Reddit[a]	64,978,944	0.28
PAQ Lewis et al. (2021)	52,953,088	0.23
Amazon Reviews Ni et al. (2019)	38,682,624	0.16
S2ORC Title Abstract Lo et al. (2020)	35438592	0.15
WikiAnswers Fader et al. (2014)	9,912,320	0.04
S2ORC Citation Titles Lo et al. (2020)	7,585,792	0.03
S2ORC Abstract Citation Lo et al. (2020)	7,503,872	0.03
S2ORC Abstract Body Lo et al. (2020)	6,389,760	0.03
Wikipedia Title Body Foundation (2024)	6,078,464	0.03
Gooaq Khashabi et al. (2021)	1,245,184	0.01
Codesearch Husain et al. (2019)	835,584	<.01
AGNews ?	409,600	<.01
CCNews Hamborg et al. (2017)	344,064	<.01
NPR[b]	344,064	<.01
CNN See et al. (2017)	278,528	<.01
Yahoo Title-Answer[c]	262,144	<.01
AmazonQA Gupta et al. (2019)	212,992	<.01
Yahoo Title-Question[d]	196,608	<.01
Sentence Compression Filippova & Altun (2013)	163,840	<.01
YahooQA[e]	131,072	<.01
ELI5 Fan et al. (2019)	98,304	<.01
Altlex Hidey & McKeown (2016)	98,304	<.01
Wikihow Koupaee & Wang (2018)	81,920	<.01
SimpleWiki Coster & Kauchak (2011)	81,920	<.01
StackExchange Duplicate Questions[f]	65,536	<.01
StackExchange Title Body[g]	65,536	<.01
StackExchange Body Body[h]	65,536	<.01
Quora Duplicate Questions[i]	32,768	<.01
SQuAD Rajpurkar et al. (2016)	16,384	<.01
Total	234,553,344	1

[a] https://huggingface.co/datasets/sentence-transformers/reddit-title-body
[b] https://files.pushshift.io/news/
[c] https://www.kaggle.com/soumikrakshit/yahoo-answers-dataset
[d] https://www.kaggle.com/soumikrakshit/yahoo-answers-dataset
[e] https://www.kaggle.com/soumikrakshit/yahoo-answers-dataset
[f] https://data.stackexchange.com/apple/query/fork/1456963
[g] https://data.stackexchange.com/apple/query/fork/1456963
[h] https://data.stackexchange.com/apple/query/fork/1456963
[i] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

- clustering

9.8 UNSUPERVISED TRAINING DATASETS

We train on $234M$ weakly supervised query-document pairs collected for training text embedding models in Nussbaum et al. (2024). The full distribution of 29 datasets is shown in Table 5. Reddit alone makes up over 25% of the data distribution, with 19 of the datasets comprising under 1% of the total data.

Table 6: Distribution of BEIR evaluation datasets used, ordered by corpus size.

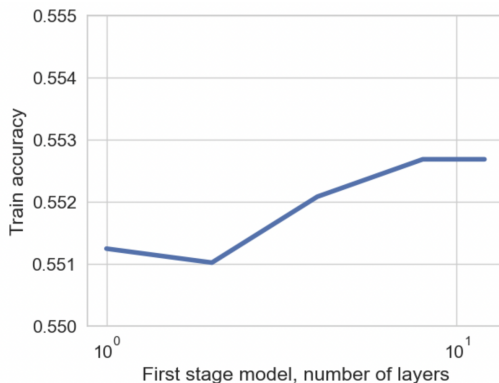| Dataset	Queries	Documents
NFCorpus	323	3,633
SciFact	300	5,183
ArguAna	1,406	8,674
SciDocs	1,000	25,657
TREC-COVID	50	171,332
Quora	5,000	522,931
Natural Questions	3,452	2,681,468
MS MARCO	6,980	8,841,823

Figure 17: System performance (training accuracy) as we scale the size of the first-stage model encoder only.

9.9 BEIR EVALUATION DATASETS

Our initial experiments involve evaluating on nine datasets from the BEIR benchmark. Datasets are detailed in Table 6. To enable fast evaluation at this stage, we obtain the top 1024 relevant documents to each document with GTR (Ni et al., 2021) and rerank only these documents at evaluation time.

9.10 ADDITIONAL MODELING ABLATIONS

First-stage model size. One consideration is whether we can improve our system without affecting search inference time by scaling the number of parameters in the backbone model only. We study this affect by scaling the number of layers in the transformer backbone of the first-stage model from 1 to the full 12. Resulting performance is shown in Section 9.10.

Our results show that scaling the first-stage model has a small positive influence on model performance. However, since the total improvement from a 12x increase in first-stage model size is less than one percent, we conclude that the second-stage model size has a much larger impact on performance.

9.11 HOW MANY TOKENS PER DOCUMENT?

We consider the question of how many tokens per document is ideal while keeping the total number of document tokens fixed. Results per the nine evaluation datasets of BEIR are shown in Section 9.11.

9.12 MTEB RETRIEVAL PERFORMANCE

Table 7 shows our model performance on all datasets in the MTEB retrieval category. We see largest improvements over the baseline on the ArguAna and TREC-Covid datasets.
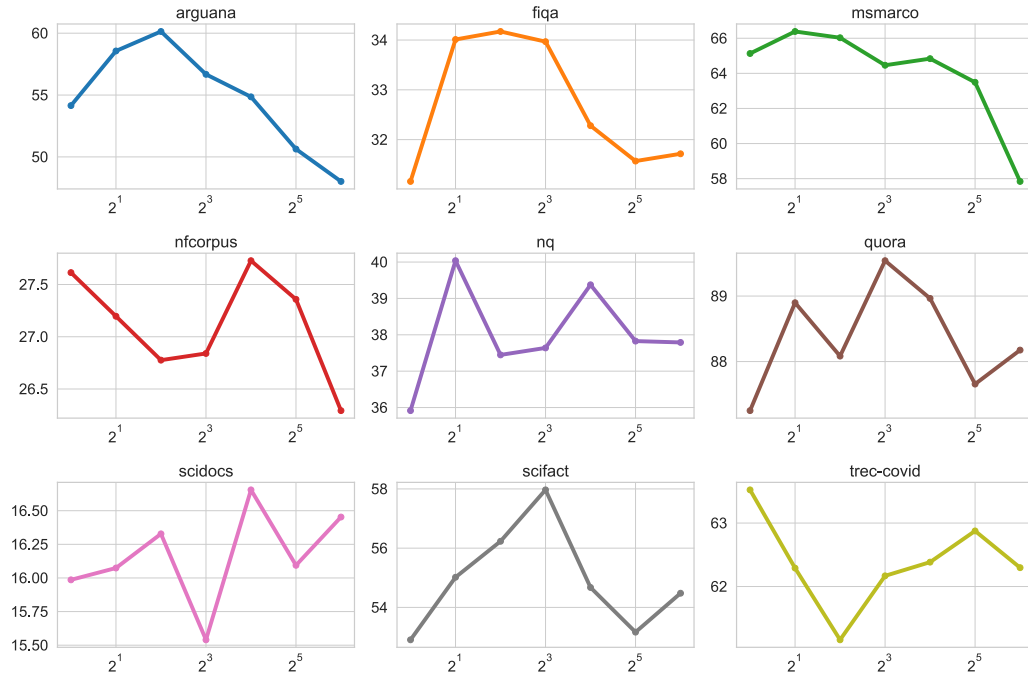
Figure 18: Performance per-dataset as we scale tokens-per-document, while keeping the total number of contextual tokens fixed. Different domains prefer a different number of tokens per document.

Method	Arg	CQA	CFEVER	DBP	FEVER	FiQA	HPQA	MSMRC	NFC	NQ	QUORA	SCID	SCIF	TREC	TOUCHE	Mean
Unsupervised																
Baseline	54.8	41.4	24.7	40.2	74.4	39.9	63.8	35.0	35.7	48.6	88.2	20.2	72.0	62.2	19.2	48.0
Contextual	54.9	43.1	24.4	40.7	79.6	42.1	68.8	38.9	36.5	57.8	88.9	21.1	72.8	77.1	21.9	51.2
Supervised																
Baseline	49.3	40.5	38.3	45.0	85.0	38.4	73.6	43.1	35.0	59.4	87.7	18.3	70.5	79.9	28.2	52.8
Contextual	53.8	41.2	38.8	43.3	89.2	40.1	73.9	42.2	35.9	61.6	87.1	20.1	72.7	82.6	27.8	54.0

Table 7: Results (NDCG@10) on the retrieval setting of the MTEB benchmark.