xPRT: CROSS-PASSAGE RETRIEVAL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense retrievers are a foundational component underlying modern RAG and search systems. Training these models typically requires a curated collection of annotated query-document pairs, which are prohibitively expensive to obtain outside a handful of well-resourced domains. This has motivated a line of work that leverages large-scale unsupervised retrieval pretraining from unlabelled text alone. Most such methods learn by optimizing a language modelling or contrastive learning objective over *short spans within a document*. This raises the question: *how can we extract richer learning signals from a single document for retrieval pretraining?* To address it, we introduce **xPRT** (Cross-Passage Retrieval Pretraining), an unsupervised framework that learns from the key information distributed across an entire document and the connections between its parts. We evaluate xPRT against five state-of-the-art objectives on MTEB English v2, continuing pretraining from a public unsupervised checkpoint while holding the corpus, schedule, and training procedure fixed. Across diverse tasks spanning all task types, we observe that xPRT consistently outperforms the strongest baseline by $\mathbf{3.2}\%$ on an average. Notably, in this continual pretraining setting, many baseline methods incur substantial performance regressions, whereas our method delivers the largest complementary gains over the initial pretrained checkpoint. Controlled ablations confirm that xPRT's performance gains stem from its proposed cross-passage signals. xPRT demonstrates that a single unlabelled document is a much richer source of retrieval signal than prior work assumed, and offers **complementary gains** to what current retrievers already capture.

1 INTRODUCTION

Dense retrievers are a core component of modern search, open-domain question answering and retrieval-augmented generation (RAG). A dual encoder maps queries and passages to vectors in a shared space and retrieves candidates from a prebuilt index by approximate nearest-neighbour search (Karpukhin et al., 2020; Reimers & Gurevych, 2019; Zhao et al., 2024). In RAG, a language model answers from the passages it is given (Lewis et al., 2020), so a passage the retriever misses cannot be recovered later in the pipeline. Since the encoder is the whole retriever, retrieval quality rests on how well a single pooled vector stands in for a passage. Learning that vector has traditionally relied on costly supervision: large collections of annotated query-passage pairs, mined hard negatives and cross-encoder filtering (Nguyen et al., 2016; Xiong et al., 2021; Qu et al., 2021). Such data exist for a handful of well-resourced domains and are prohibitively expensive to curate elsewhere.

This cost motivates learning from unlabelled text. Pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020) are the standard initialization for dense retrievers (Karpukhin et al., 2020; Xiong et al., 2021). However, they are pre-trained with token-level supervision, and the pooled vector never learns to represent a passage as a whole, in a manner that is suitable for retrieval (Chang et al., 2020; Reimers & Gurevych, 2019), and thus labelled data must close the gap.

Retrieval-oriented pre-training aims to close part of this gap without labels, by training the pooled vector directly on unlabelled text. **Contrastive** methods build pseudo-relevant pairs from the corpus, such as an inverse-cloze pair or two crops of the same document, and train the pooled vector to prefer its counterpart over passages from other documents (Lee et al., 2019; Izacard et al., 2022; Ram et al., 2022; Gao & Callan, 2022). **Bottleneck auto-encoding** methods instead train the pooled vector to summarize its passage (Gao & Callan, 2021; Xiao et al., 2022; Wang et al., 2023; Ma et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

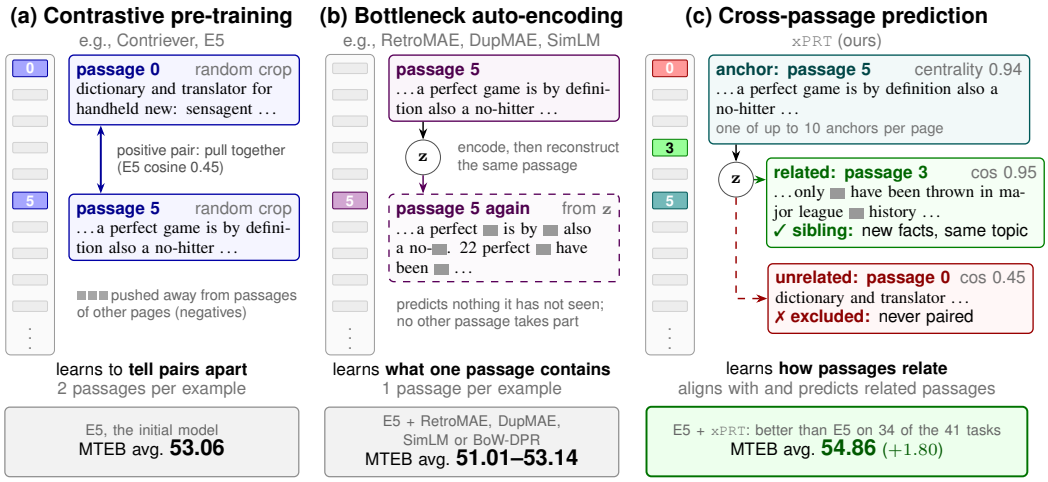


Figure 1: **How pre-training objectives use a document's passages**, on one FineWeb page. (a) Contrastive pre-training separates a passage pair from other pages'; a random pair can wrongly link an article to a dictionary widget. (b) Bottleneck auto-encoding reconstructs a passage from its own embedding $\mathbf{z}$, so it need not anticipate unseen content; (c) `xPRT` makes an anchor's embedding predict its four most related passages, adding new facts on the query–answer relation a retriever needs; unrelated passages (the widget) are neither anchors nor targets. Bottom: average over 41 MTEB(eng, v2) tasks for E5 and E5 after continued pre-training on 100K documents (Table 1).

2024; Li et al., 2023b; Shen et al., 2023). The encoder compresses a passage into the vector, and a decoder must reconstruct the passage from it, so the vector is pushed to retain the passage's content. Both families, however, make some strong assumptions for sourcing the learning signal that might contradict with what a good retriever needs. Auto-encoders reconstruct a passage from itself, so the pooled vector is never asked to anticipate content it has not read, which is exactly what a query asks of the passage that answers it. Contrastive methods look beyond the span, but pair it with a second span chosen by position or at random. They reduce the relation between the two to a single label, same document or not, applied with equal weight whether the spans share a subject or only a URL. On web text, a navigation block is as likely to be drawn as the paragraph that states the topic (Penedo et al., 2024), and nothing else in the document answers a query drawn from boilerplate. CoT-MAE (Wu et al., 2023) does look outside the passage, but supervises only a single adjacent context span. Two document-level signals thus go unused: *which* passages carry a document's subject, and *how* those passages relate to one another. This raises our central question: *how much more retrieval signal can a single unlabelled document supply?*

We propose `xPRT` (Cross-Passage Retrieval Pre-training), a self-supervised objective that exploits both signals using the document alone. Content, not position, decides what a training example contains. A passage becomes an *anchor* only if its embedding under the frozen initial encoder lies close to the centroid of its document, which keeps boilerplate out of the query side. The anchor's *siblings* are the passages of the same document most related to it under fused dense and lexical similarity (§2.1). Supervision then runs *across* passages rather than within one. A contrastive loss makes the anchor single out its most related sibling among passages of other documents (§2.2). A lightweight decoder that sees the anchor only through its embedding predicts the masked tokens of sibling passages the encoder never read (§2.3). This mirrors the relation retrieval needs between a query and the passage that answers it.

We continue pre-training the public `e5-base-unsupervised` checkpoint (Wang et al., 2022) on FineWeb (Penedo et al., 2024) under each objective. We hold the backbone, corpus, schedule, batch size, passage length, retention regularizer and seeds fixed, and evaluate every checkpoint under one protocol without fine-tuning, so that the objective is the only variable (§4). On MTEB(eng, v2) (Muennighoff et al., 2023; Enevoldsen et al., 2025), `xPRT` outperforms the strongest of five baselines across three data scales experimented with, by $\mathbf{3.2}\%$ on average over 41 tasks and by $\mathbf{6.1}\%$ nDCG@10 on the ten retrieval tasks.

Our contributions are:

- We identify two document-level signals that unsupervised retrieval objectives leave unused: which passages represent a document, and how its passages relate. We propose `xPRT`, which learns from both by making the embedding of a representative passage predict the related passages it never read (§2).
- In a controlled continual pre-training comparison against five established objectives, `xPRT` is the only one that improves its starting checkpoint substantially. It gains 1.8 to 1.9 points on MTEB(eng, v2), on every task type, at every corpus scale from 100K to 1M documents. No baseline gains more than 0.4 points at any scale, and four of the five end below the checkpoint at 100K (§4.1 and §4.2).
- Controlled ablations trace the gain to the cross-passage signal from curated related passages. Removing cross-passage prediction and pairing each anchor with a random passage of its document costs 9.3 nDCG@10. Within-passage even given as much text, or up to five times the steps, stays between 18.8 - 19.2 nDCG@10, below the strongest baseline (§4.3).

2 Cross-Passage Retrieval Pre-Training

Established retrieval pre-training objectives learn a passage's embedding from the passage itself or from heuristic pairs, such as random crops or adjacent spans. `xPRT` instead learns from *cross-passage signals*: the embedding of a passage that represents its document must single out, and predict the content of, the passages of the same document that are most related to it or complements it, e.g. answer a question posed by the first passage. The encoder f_θ, a 12-layer Transformer, is later deployed as the retriever. Documents are split into passages of 128 tokens, and each training example is a *group* $g = (a, s_1, \ldots, s_n)$: an anchor passage a and $n = 4$ sibling passages of the same document, ranked by their relatedness to the anchor (§2.1). For a passage x, e_x is the mean of the encoder's final hidden states, the embedding used for retrieval. During training, the encoder reads a light view $m_L(x)$ of every passage, with 15% of the tokens replaced by [MASK], and a one-layer decoder h_ϕ, discarded after pre-training, reads a heavy view $m_H(x)$, with 60% masked at the positions $M_H(x)$, a superset of the light mask (Appendix B). The objective combines two cross-passage terms, which carry this signal, with a self-prediction term and two standard regularizers:

$$\mathcal{L}_{\texttt{xPRT}} = \underbrace{\mathcal{L}_{\text{xp}}}_{\text{contrastive xp signal (§2.2)}} + \underbrace{\lambda_c \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{self}}}_{\text{content supervision signal(§2.3)}} + \underbrace{\mathcal{L}_{\text{MLM}} + \lambda_u \mathcal{L}_{\text{unif}}}_{\text{regularizers (§2.3)}}, \tag{1}$$

with $\lambda_c = 0.1$, $\lambda_u = 0.01$, and all other weights 1 (Table 4). Figure 2 shows how terms are computed for a group.

2.1 Selecting related passages

A sibling is useful only if it is related to its anchor, and an anchor only if it represents its document rather than, say, navigation text or a widget (Figure 1). We select groups once, before training and for both training regimes, with frozen E5 (Wang et al., 2022) as a dense scorer and inverse document frequency (IDF) as a lexical one (Spärck Jones, 1972). An anchor must be *on-topic*: the cosine similarity between its embedding and the centroid of its document's passage embeddings must exceed a threshold τ_a (Appendix C). Among the eligible passages, we keep up to ten anchors per document, preferring those that are both central and lexically informative, as combined by reciprocal rank fusion (RRF; Cormack et al., 2009). For each anchor, every other passage of the document is a candidate sibling; we rank the candidates by fusing their dense and TF-IDF similarity to the anchor, again with RRF, and keep the top four. Siblings therefore come from anywhere in the document rather than from adjacent spans (Wu et al., 2023), and every document contributes several groups: on the 100K corpus, 99% of documents yield groups (7.0 per document on average), which cover 68% of all passages.

2.2 Cross-Passage Contrastive Signals

The main objective, $\mathcal{L}_{\text{xp}}$, asks the embedding of an anchor to single out its most related sibling among passages of other documents. A symmetric InfoNCE loss (van den Oord et al., 2018) with

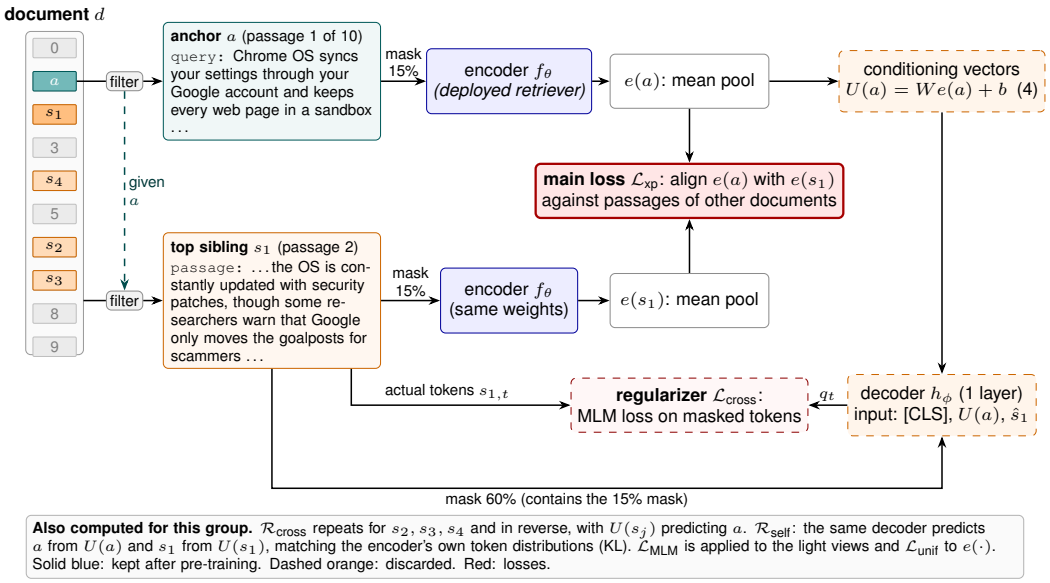


Figure 2: **xPRT training.** *Left:* From a document, first a group of anchor and n sibling crops are selected. Anchors are passages whose embedding is close to the embedding of the full document; the siblings of an anchor are the four other passages most related to it by a fused dense and TF-IDF similarity metric, so the anchor determines its siblings. *Right:* the encoder, the only model we keep, encodes the anchor/sibling passages with light (15%) masking. The cross-passage loss $\mathcal{L}_{\text{xp}}$ aligns the embeddings of the anchor and its top sibling against anchor/sibling passages of other documents. The content supervision loss $\mathcal{L}_{\text{cross}}$ asks a one-layer decoder, which receives the anchor's embedding as four conditioning vectors $U(a)$, to recover tokens from its heavily masked siblings. Excerpts are paraphrased from a real group of a 10-passage document.

cosine similarity cos and temperature $\tau = 0.05$ aligns each anchor with its top-ranked sibling s_1:

$$\mathcal{L}_{\text{xp}} = -\frac{1}{2}\left[\log \frac{\exp\left(\cos(e_a, e_{s_1})/\tau\right)}{\sum_{g'} \exp\left(\cos(e_a, e_{s'_1})/\tau\right)} + \log \frac{\exp\left(\cos(e_{s_1}, e_a)/\tau\right)}{\sum_{g'} \exp\left(\cos(e_{s_1}, e_{a'})/\tau\right)}\right], \tag{2}$$

where (a', s'_1) is the anchor and top sibling of group g', and g' ranges over the groups of the global batch except the other groups of the same document, which would be false negatives. The loss has the form of contrastive pre-training (Izacard et al., 2022); what xPRT changes is the positive. Contrastive methods pair a passage with a random crop, an adjacent span or a structural counterpart such as a title, whether or not the two texts are related. xPRT pairs a representative anchor with the passage of its document that is most related to it, so that the embedding must capture the relation between two different passages on one topic – the relation between a query and a passage that answers it.

2.3 Cross-Passage Content Supervision Signals

$\mathcal{L}_{\text{xp}}$ uses one sibling per anchor and constrains only embedding geometry. Two prediction terms supervise *content*: an anchor's embedding should encode its related passages, and every embedding its own passage. Both use the decoder h_ϕ, one freshly initialized Transformer layer sharing the encoder's token embeddings and MLM head. A linear map projects the conditioning embedding e_x to $U(x) \in \mathbb{R}^{4\times d}$, inserted after [CLS] in the heavy view $m_H(y)$ of a target passage y, and h_ϕ predicts $q_t(\cdot \mid m_H(y), U(x))$ at each masked position $t \in M_H(y)$. One layer cannot recover 60% of y from the rest, so it must draw on $U(x)$.

$\mathcal{L}_{\text{cross}}$, the prediction term that xPRT introduces, asks the embedding of one passage to predict the masked tokens of another: those of each sibling from the anchor, and those of the anchor from each

sibling. With the masked-language-modeling cross-entropy on the actual tokens y_t,

$$C(y \mid x) = -\frac{1}{|M_H(y)|} \sum_{t \in M_H(y)} \log q_t\big(y_t \mid m_H(y), U(x)\big), \tag{3}$$

and $\mathcal{L}_{\text{cross}} = \frac{1}{2n} \sum_{j=1}^{n} \big[C(s_j \mid a) + C(a \mid s_j)\big]$. To predict a passage that it has never read, the anchor's embedding must encode what the two passages share; bottleneck auto-encoders, in contrast, reconstruct a passage from its own embedding (Section §5).

2.4 Within-Passage Content Supervision Signals

Beyond its tokens, the encoder's reading of a passage, its output distribution at every masked position, describes what the passage is about. $\mathcal{L}_{\text{self}}$ distills this reading into the embedding: the encoder, which reads the light view, is the teacher, and the decoder, which reads the heavy view conditioned on the same passage's embedding, is the student: $\mathcal{L}_{\text{self}} = \frac{1}{2}\big[D(a) + D(s_1)\big]$ with

$$D(y) = \frac{1}{|M_H(y)|} \sum_{t \in M_H(y)} \mathrm{KL}\Big(\mathrm{sg}\left[p_t(\cdot \mid m_L(y))\right] \,\Big\|\, q_t\big(\cdot \mid m_H(y), U(y)\big)\Big), \tag{4}$$

where p_t is the distribution of the encoder's MLM head and $\mathrm{sg}[\cdot]$ stops gradients through the targets. Every target comes from a view that saw more of y than the student (Hinton et al., 2015). Both terms are token-weighted batch averages (Appendix D).

An MLM loss $\mathcal{L}_{\text{MLM}}$ on the light views of a and s_1 keeps the teacher's distributions grounded in the actual tokens, and the uniformity loss $\mathcal{L}_{\text{unif}}$ of Wang & Isola (2020) on the normalized anchor embeddings keeps them from collapsing. After pre-training, the projection, the decoder and the MLM head are discarded; the retriever is f_θ with mean pooling, so inference cost is unchanged. Per group, training costs five encoder passes and ten one-layer decoder passes.

3 Experimental setup

Pre-training corpora. All models are pre-trained on English web documents from FineWeb (Penedo et al., 2024). We take the first one million non-empty documents in the corpus's released order, tokenize them with the BERT uncased vocabulary shared by all our models, and split them into 128-token passages, keeping a trailing passage only if it has at least 64 tokens; documents with fewer than five passages cannot form a group and are discarded. From this pool we build three corpora. **100K** is a length-stratified sample of 100,000 documents, of which 40%, 30%, 21% and 9% have 5–8, 9–16, 17–32 and 33–64 passages. **1M** contains every document of the million with 5–64 passages, and **500K** is the subset of 1M that falls in a fixed random half of the million.

Continual pre-training from a strong retriever (main setting). We initialize every model from E5-base-unsupervised (Wang et al., 2022), a BERT-base encoder contrastively pre-trained on large-scale weakly supervised text pairs. E5 ships without an MLM head, so we initialize the head from BERT-base-uncased, which shares its vocabulary. Pre-training continues for one pass over the groups of each corpus with batch size 512 (1 epoch over all the documents) and a peak learning rate of 10^{-6} with 5% linear warm-up and linear decay; the encoder is frozen for the first 100 steps while the newly initialized components warm up (Table 4). Continual pre-training must not erase what the initial model already knows, so we apply a retention regularizer of the same form, weight and schedule to *every* method, baselines included (Appendix E lists the passages each method uses): a symmetric KL divergence between the in-batch similarity distributions of the current model and of the frozen initial model (Tung & Mori, 2019; Park et al., 2019). Its weight decays from 40 to 0 (Appendix E) as $\lambda_{\text{ret}}(t) = 40\left(1 - \sqrt{(t-1)/(T-1)}\right)$ over the T training steps, which protects the initial model early and lets the pre-training objective dominate late.

Pre-training from BERT. To test whether the signal depends on a contrastive starting point, we also pre-train from BERT-base-uncased (Devlin et al., 2019). `xPRT` uses the same configuration as in the continual setting – four siblings per anchor and every term of Eq. 1 – but without the retention regularizer, since there is no pre-trained retriever to preserve. Training uses the 100K groups for

six epochs (8,172 steps, batch 512) at learning rate 10^{-4}, without a backbone freeze (Table 4). The groups are the same as in the continual setting and were selected with E5, so this regime tests the objective, not a selection-free pipeline. Because each run is cheap, we also use this regime for the ablations of §4.3.

Baselines. We compare with five retrieval-oriented pre-training objectives, using their official implementations adapted to our data pipeline. Four of them learn within a passage: RetroMAE (Xiao et al., 2022), masked auto-encoding with an enhanced one-layer decoder; DupMAE (Liu et al., 2023), which adds a bag-of-words objective to RetroMAE; SimLM (Wang et al., 2023), replaced-language modeling through a representation bottleneck; and BoW-DPR (Ma et al., 2024), bag-of-words prediction without a decoder. Each of them trains on the anchor passage of every group – the same passages that `xPRT` uses as anchors, including the multiple anchors per document – with its native masking rates and loss weights (Table 6). The fifth, Revela (Cai et al., 2026), learns from other passages: a language model (Qwen2.5-0.5B with LoRA) predicts each passage while attending to the other passages of its group, weighted by the retriever's similarities. Each Revela group holds an anchor, its four siblings and eleven anchors of other documents. All baselines use the initialization, number of steps, batch size, learning-rate schedule and, in the continual setting, retention regularizer of `xPRT`. We also report the initial model, E5, without further training.

Evaluation. We evaluate the final checkpoint of every run, without fine-tuning, on the 41 tasks of MTEB(eng, v2) (Muennighoff et al., 2023; Enevoldsen et al., 2025). The tasks span classification (8), clustering (8), pair classification (3), reranking (2), retrieval (10), semantic textual similarity (9) and summarization (1). Embeddings are mean-pooled and L2-normalized, inputs carry E5's `query:` and `passage:` prefixes, the maximum length is 512 tokens, and inference runs in bf16. We report the average main score over the 41 tasks (*Avg.*) (Table 8), the number of tasks on which a model is better or worse than the initial E5 (W/L), and nDCG@10, MRR@10, Recall@10 and Recall@100 averaged over the ten retrieval tasks. We select hyperparameters for the final config through validation loss that is disjoint from the evaluation benchmark which is contrastive perplexity on the MS MARCO passage-ranking dev set (Nguyen et al., 2016).

Table 1: **Continual pre-training from E5.** Established objectives leave E5 essentially unchanged, whereas `xPRT` improves every task type at every corpus scale. Columns give the average main score per task type (number of tasks below; nDCG@10 for retrieval) and over all 41 tasks (Avg.); Δ is the change in Avg. with respect to E5, computed before rounding. Scores are $\times 100$, and the best score in each column and corpus is in bold. Win/loss counts and further retrieval metrics are in Table 10 and Table 8.

Corpus	Method	Class. (8)	Clust. (8)	Pair (3)	Rerank (2)	Retr. (10)	STS (9)	Summ. (1)	Avg. (41)	Δ
–	E5 (initial)	55.72	46.06	79.19	44.80	43.26	63.36	31.46	53.06	–
100K	RetroMAE	55.96	45.70	79.25	44.70	42.41	63.69	32.19	52.92	−0.14
	DupMAE	55.95	45.90	79.24	44.69	42.44	63.73	32.13	52.97	−0.09
	SimLM	56.22	39.62	**80.67**	44.29	36.93	66.32	28.10	51.01	−2.05
	BoW-DPR	56.12	43.58	80.07	44.60	43.11	65.67	29.57	53.14	+0.08
	Revela	56.10	45.86	79.16	**44.96**	42.99	63.17	32.13	53.02	−0.05
	`xPRT`	**56.74**	**46.75**	79.84	44.83	**45.78**	**66.85**	**32.87**	**54.86**	**+1.80**
500K	RetroMAE	55.99	45.81	79.31	44.69	42.38	63.90	32.05	52.99	−0.07
	DupMAE	56.00	45.93	79.32	44.66	42.32	63.90	32.16	53.00	−0.06
	SimLM	**57.00**	45.23	**80.40**	44.52	42.01	65.30	30.38	53.32	+0.26
	BoW-DPR	56.63	45.24	79.92	44.76	43.38	64.62	31.61	53.44	+0.38
	Revela	56.07	45.92	79.00	**45.06**	42.66	63.10	31.15	52.89	−0.17
	`xPRT`	56.67	**47.04**	79.91	44.87	**45.67**	**67.04**	**32.93**	**54.93**	**+1.87**
1M	RetroMAE	56.13	45.88	79.39	44.66	41.94	64.07	32.40	52.97	−0.09
	DupMAE	56.11	45.73	79.39	44.65	41.99	64.03	32.41	52.94	−0.12
	SimLM	56.50	43.82	79.52	43.85	37.93	63.07	29.96	51.36	−1.70
	BoW-DPR	**56.76**	45.13	79.62	44.39	42.36	64.07	30.95	53.02	−0.04
	Revela	56.11	46.10	79.07	**44.88**	42.47	62.82	30.22	52.80	−0.26
	`xPRT`	56.68	**47.15**	**79.89**	44.87	**45.33**	**67.13**	**33.00**	**54.89**	**+1.83**

4 RESULTS

We ask whether the relations between the passages of a document add to what a strong retriever and established objectives already capture (§4.1), where the gains appear (§4.2), and which ingredients produce them (§4.3 and §4.4).

4.1 MAIN RESULTS

We conduct our experiments on E5, which was contrastively pre-trained on hundreds of millions of text pairs. We observe that well-established pre-training objectives could not improve it meaningfully (Table 1) in a continual pre-training setting. Across three corpora we observe that, the four within-passage objectives exhibit an average performance difference of about -2.05 to $+0.38$ points. Even the SimLM objective is unstable, losing 2.05 and 1.70 points at 100K and 1M data scales. Revela, which learns from other passages through a language model, ends 0.05 to 0.26 points below the E5 starting point. `xPRT`, with identical data and budget, improves E5 by 1.80, 1.87 and 1.83 points at 100K, 500K and 1M, on 34-36 of the 41 tasks (Table 8) and on every task type at every scale. Furthermore, its nDCG@10 at 100K, 45.78, is 6.2% above that of the strongest baseline, demonstrating the effectiveness of the proposed method.

One intuition that may help explain the result is that E5 as well as most of the pre-training objectives in literature lack direct supervision from the relation between different passages in a single document. E5's pairs are linked by web structure, such as a title and its body, and reconstructing a passage from its own embedding, in each of the four forms we tried, provides minimal complementary value. `xPRT` supervises the embedding with this intra-document relations directly. Our objective forces the anchor's embedding to single out its most related passage among passages of other documents, and learn sufficiently from the related passages' content for a one-layer decoder to predict them. Related passages alone do not suffice, as Revela shows. It trains on the same groups as `xPRT` – an anchor, its four siblings and eleven anchors of other documents – but the retriever's similarities only weight the attention of a language model that predicts each passage, so the retriever learns only indirectly, from how much each passage helps the language model. In `xPRT`, the relation acts on the embedding itself.

Crucially, the signal does not depend on the model used as the starting point. In a controlled ablation, we observe that pre-training BERT for six epochs on the 100K groups (Table 2), `xPRT` scores 4.4 points above the strongest baseline, SimLM, with a 41% higher retrieval nDCG@10 (29.21 against 20.68), and it is the best of the six models on 28 of the 41 tasks (Table 11). Revela, the other objective that learns from other passages, ends 0.5 points below SimLM. The gains lie mostly in retrieval, STS, pair classification and clustering; classification and summarization are the only task types on which baselines remain ahead.

4.2 GAINS ACROSS TASK TYPES AND CORPUS SCALES

The gains generalize across various types and corpora that `xPRT` is evaluated on. At 100K, gains are largest on STS (+3.5) and retrieval (+2.5 nDCG@10), which score whether two different texts mean the exact thing, and smallest on the reranking task (under 0.1), pair classification (+0.6) and clustering (+0.7). Within retrieval, the largest gains appear on tasks whose query and target are linked by an argument or a claim: Touché-2020 (+6.6), ArguAna (+5.8), FEVER (+4.1), Climate-FEVER (+3.6) and HotpotQA (+3.0) (Table 10). The two regressions, TREC-COVID (-1.4) and FiQA (-1.1), are in specialized domains far from web text, and the two reranking tasks, which reorder short candidate lists by fine distinctions between similar questions (AskUbuntu) or by a user's clicks (MIND), are unchanged. The gain is also stable across scales: from 100K to 1M documents, with clustering and STS still rising and nDCG@10 slightly lower at 1M (45.33) than at 100K (45.78), as TREC-COVID and FiQA degrade (Table 9). We attribute the plateau to the continual budget, one pass at a learning rate of 10^{-6} under a decaying retention constraint.

4.3 CURATION OF RELATED PASSAGES IS CRITICAL

Table 3 ablates `xPRT` in the BERT regime, isolating the contributions of each term in the objective and of sibling selection by removing ingredients one at a time. The decisive ingredient is the cross-

Table 2: **Pre-training from BERT** (100K groups, six epochs, no retention regularizer). Columns as in Table 1; Δ is the change in Avg. relative to SimLM, the strongest baseline. `xPRT` and its variant without $\mathcal{L}_{\text{cross}}$ (Table 3) are means over three seeds. Per-task scores are in Table 11.

Method	Class. (8)	Clust. (8)	Pair (3)	Rerank (2)	Retr. (10)	STS (9)	Summ. (1)	Avg. (41)	Δ
RetroMAE	51.43	36.96	72.28	38.60	19.65	54.37	33.28	41.96	-0.77
DupMAE	50.16	35.79	70.71	38.13	18.89	52.66	**33.77**	40.80	-1.93
SimLM	52.02	40.16	71.55	38.60	20.68	53.81	31.38	42.73	–
BoW-DPR	**52.46**	39.11	71.23	39.25	19.04	55.31	32.29	42.57	-0.16
Revela	51.25	40.28	69.88	39.43	17.45	56.14	31.35	42.24	-0.49
`xPRT`	51.33	**44.29**	**76.20**	**40.77**	**29.21**	**59.83**	26.23	**47.12**	$+\mathbf{4.39}$
without $\mathcal{L}_{\text{cross}}$ (§4.3)	50.96	44.17	75.85	40.74	28.46	59.27	25.25	46.67	$+3.94$

Table 3: **Ablations when pre-training from BERT** (100K groups, six epochs; $\times 100$): averages over 41 tasks and 10 retrieval tasks.

	Variant	MTEB (41) Avg.	MTEB (41) Δ	Retrieval (10) nDCG@10	Retrieval (10) Δ
	`xPRT` (Eq. 1, $\lambda_c = 0.1$)	**47.12**		29.21	
	without $\mathcal{L}_{\text{cross}}$	46.67	-0.45	28.46	-0.75
$\mathcal{L}_{\text{xp}}$	random passage of the same document	42.22	-4.90	19.87	-9.34
	top sibling of another document	40.80	-6.32	18.03	-11.18
	SimLM (strongest baseline)	42.73	-4.39	20.68	-8.53
objective	without $\mathcal{L}_{\text{xp}}$ (prediction only)	43.72	-3.4	19.47	-9.74
	without $\mathcal{L}_{\text{self}}$	44.40	-2.72	24.78	-4.4
	without $\mathcal{L}_{\text{MLM}}$	43.19	-3.94	22.93	-6.29
	without $\mathcal{L}_{\text{unif}}$	46.73	-0.39	28.45	-0.76
no cross-passage signal	self-prediction, all five passages	42.62	-4.50	19.20	-10.01
	self-prediction, a and s_1, $2\times$ steps	42.75	-4.37	18.80	-10.41
	RetroMAE, all five passages, $5\times$ steps	41.73	-5.39	19.21	-10.00
	siblings ranked by E5 cosine only (no TF-IDF)	44.82	-2.30	25.88	-3.33
	siblings ranked by TF-IDF only (no E5 cosine)	47.10	-0.02	**29.89**	+0.67

passage signal from curated related passages. Removing cross-passage prediction alone lowers the average by 0.45 points and nDCG@10 by 0.75, on every task type (Table 2). Removing the cross-passage signal altogether – dropping $\mathcal{L}_{\text{cross}}$ and pairing each anchor with a random passage of its document instead of its selected sibling – costs 4.9 points on average and 9.3 of nDCG@10, 32% of the retrieval score, and falls below SimLM; a sibling from another document is worse still. More text cannot substitute for this signal: three controls that drop both cross-passage terms and give self-prediction or RetroMAE as much text as `xPRT` or more, with up to five times the training steps, still plateau at 18.8–19.2 nDCG@10.

Random pairs fail because web pages mix genuine content with widgets, link lists, and contact details (Figure 1; Table 5) – noise inherited by random same-document pairing, as used in contrastive pre-training (Izacard et al., 2022; Gao & Callan, 2022). Just as filtering noisy pairs benefits weakly supervised pre-training (Wang et al., 2022), curating them is essential in the unsupervised setting. Once the pairs are curated, the remaining design choices matter far less. Sibling ranking method changes the average by at most 2.3 points (TF-IDF alone even gives a slightly higher nDCG@10 than the fused ranking, $+0.67$), whereas a random passage costs 4.9. Among the loss terms, $\mathcal{L}_{\text{xp}}$ carries most of the gain (-9.7 nDCG@10 when removed), $\mathcal{L}_{\text{MLM}}$ is also necessary (-3.9 points), since it grounds the prediction targets in the actual tokens, and likewise $\mathcal{L}_{\text{self}}$ (-2.7 points) matters, removing $\mathcal{L}_{\text{unif}}$, by contrast, changes the average by only 0.4.

4.4 WHAT THE CROSS-PASSAGE SIGNAL CHANGES

The cross-passage signal teaches the embedding which passages address one another, beyond which ones share words. ArguAna shows this most clearly: its target is the counter-argument of the query, it has one of the largest gains (+5.8 nDCG@10), and its debate pages, which collect points and counterpoints on one motion, resemble xPRT's groups. Re-encoding its 1,406 queries in full precision (Appendix H), xPRT ranks a passage from the query's own debate first for 79.2% of the queries, against 73.9% for E5. The gain does not come from lexical matching: after splitting the queries into thirds by the IDF-weighted overlap between query and counter-argument, xPRT improves over E5 in every third (+3.7 to +8.8 nDCG@10), whereas BoW-DPR, whose objective is lexical, gains only in the highest-overlap third and loses 10.3 points in the lowest; CQADupstack-Gaming shows the same pattern. For the argument that legalizing drugs would reduce crime, E5 ranks first another argument of the same debate, which does not address crime, whereas xPRT ranks first the counter-argument that most drug crimes are committed by people on drugs (Table 7 gives six such queries).

5 RELATED WORK

Unsupervised pair construction: Lacking labelled pairs (Nguyen et al., 2016), one family builds pseudo-relevant pairs from an unlabelled corpus and trains the pooled vector contrastively (van den Oord et al., 2018): inverse cloze pairs a sentence with its context (Lee et al., 2019; Chang et al., 2020), Contriever and SPIDER pair two crops of one document (Izacard et al., 2022; Ram et al., 2022), coCondenser adds a corpus-aware term over a bottleneck (Gao & Callan, 2022), and weakly supervised mining filters noisy pairs at scale (Wang et al., 2022); supervised pipelines then add hard negatives and cross-encoder filtering (Xiong et al., 2021; Qu et al., 2021). In all of them the positive is fixed by position or by chance, so a pair counts as equally relevant whether its two passages share a subject or only a URL.

Bottleneck auto-encoding: A second family reconstructs a passage through the pooled vector, weakening the decoder so that the vector is the only route to the target (Lu et al., 2021; Gao & Callan, 2021; Xiao et al., 2022; Li et al., 2023b). Later variants change what is reconstructed rather than where it comes from: a bag-of-words term (Liu et al., 2023), replaced-language modeling (Wang et al., 2023), a lexicon bottleneck (Shen et al., 2023), or the passage's own words with no decoder at all (Ma et al., 2024). Every target still comes from the passage just read.

Supervision that leaves the passage: CoT-MAE reconstructs one neighbouring span alongside the input (Wu et al., 2023), and Revela reaches the retriever through a language model that predicts each passage while attending to the others in its batch (Cai et al., 2026). Our machinery is standard: soft targets from a better-informed view (Hinton et al., 2015), distillation that transfers geometry rather than logits (Park et al., 2019; Tung & Mori, 2019), and InfoNCE read through alignment and uniformity (Wang & Isola, 2020). What xPRT adds is which passages to supervise (§2.1), which separates it from contrastive objectives whose positive is a random crop, an adjacent span or a structural counterpart, from auto-encoders whose targets never leave the encoded passage, and from Revela, where the relation between passages reaches the retriever through a language model's attention rather than through the embedding.

6 CONCLUSION

Established retrieval pre-training objectives leave unused how the passages of a document relate to one another. xPRT selects representative anchors and their most related passages, aligns their embeddings and predicts the related passages through the anchor's embedding. Under a controlled protocol, it is the only objective that improves E5 substantially, by 1.8–1.9 points on MTEB(eng, v2) and on every task type at every corpus scale, and from BERT it outperforms every baseline by 4.4 points or more; ablations trace the gain to the cross-passage signal from curated related passages.

AI USE STATEMENT

In this work, we used generative AI tools to propose and refine hypotheses, to give feedback on our experimental design, to help implement parts of the method and the evaluation harness, and to help interpret results. Additionally, we used generative AI tools to search for and summarize related work, to identify relevant literature and format references, to draft and edit parts of the manuscript for readability, and to suggest its structure.

We have reviewed all AI-assisted work. Every reference was checked against primary metadata from arXiv or the publisher rather than accepted from model output, because language models are known to fabricate citations, and references we could not verify were removed. Every number reported in this paper comes from a training or evaluation run that we executed and logged; no result was produced or estimated by a model. Where a model proposed an experiment or an explanation for a result, we ran the experiment and confirmed the outcome before any claim resting on it entered the paper, and AI-assisted code was read and tested by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 2 states the objective in full, including every loss term and the procedure that selects anchors and their siblings, and Section 3 states how the three corpora are derived from FineWeb, which public checkpoints initialize each run, the optimization schedule and batch size, the retention regularizer applied identically to every method (Appendix E), the configuration of each baseline (Appendix F), and the evaluation protocol. Table 4 gives complete hyperparameter settings for the continual and the BERT regime, and Appendix I gives per-task scores for every run from which the averages in the main text are computed. Group selection is deterministic given a corpus and the frozen scorer (Appendix C), so the training examples underlying every run can be regenerated exactly. Table 2 and the first rows of Table 3 report means over three seeds. We provide the complete code base, covering group selection, all pre-training objectives and the evaluation harness, with README files for reproducing each table, as anonymous supplementary material.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, June 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Fengyu Cai, Tong Chen, Xinran Zhao, Sihao Chen, Hongming Zhang, Sherry Wu, Iryna Gurevych, and Heinz Koeppl. Revela: Dense retriever learning via language modeling. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 1372–1395, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0343104ddbfc48f35f06aaae88980e48-Paper-Conference.pdf.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkg-mA4FDr.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3829–3846, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.232. URL https://aclanthology.org/2023.emnlp-main.232/.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrø m, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, RafałPoświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Harries, Loïc Magne, Isabelle Mohr, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lu, Jordan Clive, Gayatri K, Maksimova Anna, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James V. Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 101715–101771, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fc0e3f908a2116ba529ad0a1530a3675-Paper-Conference.pdf.

Luyu Gao and Jamie Callan. Condenser: a pre-training architecture for dense retrieval. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 981–993, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.75. URL https://aclanthology.org/2021.emnlp-main.75/.

Luyu Gao and Jamie Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2843–2853, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.203. URL https://aclanthology.org/2022.acl-long.203/.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 2591–2607, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0b276510ec2d3f6613a8b60c41ff0438-Paper-Conference.pdf.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new

approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=jKN1pXi7b0.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016. URL http://arxiv.org/abs/1612.00796.

Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Yx1OzVU_SRi.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612/.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023a.

Zehan Li, Yanzhao Zhang, Dingkun Long, and Pengjun Xie. Challenging decoder helps in masked auto-encoder pre-training for dense passage retrieval, 2023b. URL https://arxiv.org/abs/2305.13197.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947, December 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2773081. URL https://doi.org/10.1109/TPAMI.2017.2773081.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zheng Liu, Shitao Xiao, Yingxia Shao, and Zhao Cao. RetroMAE-2: Duplex masked auto-encoder for pre-training retrieval-oriented language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2635–2648, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.148. URL https://aclanthology.org/2023.acl-long.148/.

Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tie-Yan Liu, and Arnold Overwijk. Less is more: Pretrain a strong Siamese encoder for dense text retrieval using a weak decoder. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2780–2791, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.220. URL https://aclanthology.org/2021.emnlp-main.220/.

Guangyuan Ma, Xing Wu, Zijia Lin, and Songlin Hu. Drop your decoder: Pre-training with bag-of-word prediction for dense passage retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 1818–1827, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657792. URL https://doi.org/10.1145/3626772.3657792.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268, 2016. URL http://arxiv.org/abs/1611.09268.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*,

volume 37, pp. 30811–30849. Curran Associates, Inc., 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/370df50ccfdf8bde18f8f9c2d9151bda-Paper-Datasets_and_Benchmarks_Track.pdf.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.466. URL https://aclanthology.org/2021.naacl-main.466/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. Learning to retrieve passages without supervision. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2687–2700, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.193. URL https://aclanthology.org/2022.naacl-main.193/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Xiaolong Huang, Binxing Jiao, Linjun Yang, and Daxin Jiang. LexMAE: Lexicon-bottlenecked pretraining for large-scale retrieval. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PfpEtB3-csK.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. doi: 10.1108/eb026526.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 1179–1206. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/tosh21a.html.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv e-prints*, art. arXiv:1807.03748, July 2018. doi: 10.48550/arXiv.1807.03748.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. SimLM: Pre-training with representation bottleneck for dense passage retrieval. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2244–2258, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.125. URL https://aclanthology.org/2023.acl-long.125/.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Xing Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. Contextual masked auto-encoder for dense passage retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4738–4746, Jun. 2023. doi: 10.1609/aaai.v37i4.25598. URL https://doi.org/10.1609/aaai.v37i4.25598.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 538–548, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.35. URL https://aclanthology.org/2022.emnlp-main.35/.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4), February 2024. ISSN 1046-8188. doi: 10.1145/3637870. URL https://doi.org/10.1145/3637870.

A THEORETICAL ANALYSIS

Our experiments establish two facts. Cross-passage prediction outperforms within-passage reconstruction (Table 3), and reconstruction-based pretrainers such as RetroMAE and DupMAE end *below* the retriever they are initialized from (Table 1). The first fact is expected. The second is not: a weaker training signal should yield a smaller gain, not a loss. The analysis explains both through a single mechanism.

1. **The text pairs dictate what any of the objectives learns.** Under squared error or KL, the optimal predictor is the average target given the input. When the target is another passage of the same document, this average removes everything except what the two passages share: their subject, which is exactly what relevance depends on (Theorem A.1).
2. **Reconstruction makes the subject compete with passage-specific content.** When the decoder also reads a masked copy of the target and the conditioning vector has limited capacity, reconstruction pays for each kind of information in proportion to how often the copy hides it. Topical words recur within a passage, so the copy reveals the subject more often than passage-specific content, and the subject loses the competition for capacity. In the limit of an unrestricted encoder, the optimum carries no information about the subject at all (Theorem A.4). This is consistent with the degradation.
3. **The conditions are made explicit.** The results assume that sibling passages share nothing beyond their subject and that queries are encoded like passages. They therefore also indicate where cross-passage training should help less: siblings that share non-topical content, and queries unlike passages. This result highlights promising directions for future work by improving how cross-passage training pairs are constructed for self-supervised retrieval model training.

The analysis does not imply that within-passage signals are useless: removing $\mathcal{L}_{\text{self}}$ from the full objective lowers performance (Table 3). Cross-passage prediction supplies the subject; within-passage prediction can complement it, but on its own it puts the subject at a disadvantage. The results concern the optima of the objectives, not training dynamics, and Theorems A.3 and A.4 hold in a model of passages rather than on web data. Section I states the setting and the results. The proofs in Section A.2 can be skipped without losing the argument.

A.1 SETTING AND RESULTS

Setting. Let $Y \in \{1, \dots, k\}$ be the *subject* of a document, that is, what it is about, with prior $\pi_y = P(Y = y) > 0$. The *subject posterior* of a text x is the vector $p(x) \in \mathbb{R}^k$ with entries $p_y(x) = P(Y = y \mid x)$. It records everything x reveals about the subject, and nothing else.

Both objectives train a predictor ψ that maps an input a to features $\varphi(t) \in \mathbb{R}^d$ of a target t. In `xPRT`, $\varphi(t)$ is the teacher's token distribution at a masked position of t, and $\psi(a)$ is the decoder's output given the conditioning vector $U(a)$. The two objectives differ only in the pair (a, t):

- *cross-passage:* a is an anchor and t is a sibling s from the same document;
- *within-passage:* a is the visible part of a passage and t is its masked part.

For squared error, and for the KL divergence from teacher to student used in D, the optimal predictor is the conditional mean (Lemma A.5):

$$\psi^*(a) \;=\; \mathbb{E}\big[\varphi(t) \mid a\big].$$

ψ^* is what an objective asks for, independently of architecture and optimization. Because one input can be followed by many targets, ψ^* averages over them. What survives this averaging depends on the target, and this is the mechanism behind every result below.

The *profile* of subject y is its typical target, $\mu_y = \mathbb{E}[\varphi(t) \mid Y = y]$. Collecting the profiles as the columns of $M = [\mu_1 \cdots \mu_k] \in \mathbb{R}^{d \times k}$, the vector $M\,p(a) = \sum_y p_y(a)\,\mu_y$ mixes the profiles in proportion to the posterior of a.

Premises.

(P1) Relevance is a shared subject. A query and its relevant passage are independent given their common subject, and an irrelevant passage is an independent draw from the corpus. Every query and passage is a possible anchor, and the subject posterior of a text does not depend on whether it is read as a query, a passage or an anchor.

(P2) Clean pairs. Selected pairs are independent given their subject: $a \perp s \mid Y$.

Under P1, the score

$$S^*(q, x) \;=\; \sum_y \frac{p_y(q)\, p_y(x)}{\pi_y}$$

orders candidates by their posterior probability of being relevant (Lemma A.6), so no retriever can rank better.

P1 defines the retrieval task, as any analysis of retrieval must; its second sentence is the transfer assumption that lets a pretraining optimum be applied to queries. P2 is the property that anchor filtering and sibling ranking are responsible for. It is the conditional-independence condition used to analyze self-supervised learning (Tosh et al., 2021; Lee et al., 2021). It does not require a passage to have a single topic, since Y may describe a mixture of topics, nor does it require a passage to reveal its subject with certainty. It excludes sharing that is not about the subject, such as a sentence split across adjacent passages or common boilerplate. Within-passage pairs violate P2 by construction, and the contrast between the two objectives rests on this.

Cross-passage prediction. The siblings that can follow an anchor differ in everything except their subject, so averaging over them removes everything else.

Theorem A.1 (Cross-passage prediction retains exactly the subject)**.** *Under P2, for every anchor a,*

$$\psi^*(a) \;=\; M\, p(a).$$

Consequently:

(i) Nothing else is retained: $\psi^*(a)$ *depends on a only through $p(a)$.*

(ii) The subject is retained up to what targets distinguish: *for every query q and passage x (under P1),*

$$\psi^*(q)^\top \psi^*(x) \;=\; \sum_{y,y'} p_y(q)\, p_{y'}(x)\, \langle \mu_y, \mu_{y'} \rangle,$$

so the dot product scores subject overlap, weighted by the similarity of the subjects' profiles. If $\mu_1, \ldots, \mu_k$ are linearly independent, a fixed reweighting G gives $\psi^*(q)^\top G\, \psi^*(x) = S^*(q, x)$, the best possible ranking.

Linear independence means that different subjects produce different targets on average. Theorem A.1 is the cross-passage instance of the observation of Lee et al. (2021) that predicting one view from another recovers the latent the views share.

Remark A.2 (The decoder's view of the target). In `xPRT` the decoder also reads a heavily masked view $\hat{s}$ of the sibling. Part (i) of Theorem A.1 persists: under P2, the optimal predictor $\mathbb{E}[\varphi(s) \mid a, \hat{s}]$ depends on a only through $p(a)$.

A model of passages. A comparison with reconstruction requires a passage to contain something besides its subject. In the *component model*, a passage is $x = (Y, \nu_x)$, where the *specific content* ν_x is independent of Y and of the content of every other passage. Each text (an anchor, a sibling, or the visible or masked part of a passage) is generated from the content of its passage, with randomness independent of other passages. Target features are additive in the two components:

$$\mathbb{E}\big[\varphi(t) \mid a,\, Y,\, \nu_t\big] \;=\; c(Y) + d(\nu_t), \qquad \text{(CM)}$$

where ν_t is the specific content of the passage that t comes from. For cross pairs, equation CM only asks the conditional mean of φ to be additive. For within pairs it is an idealization: the features of the masked part depend on the content of the passage, not on which tokens were masked. Cross pairs satisfy P2 in this model.

Within-passage prediction. When the target is the masked part of the same passage, its specific content does not average out, because the visible part predicts it.

Theorem A.3 (Within-passage prediction also retains passage-specific content)**.** *In the component model, for every pair* (a, t)*,*

$$\psi^*(a) \;=\; M\,p(a) + \Delta(a), \qquad \Delta(a) \;=\; \mathbb{E}\big[d(\nu_t) \mid a\big] - \mathbb{E}\big[d(\nu_t)\big].$$

(i) Cross-passage: $\Delta \equiv 0$*, which recovers Theorem A.1.*

(ii) Within-passage: *for visible parts* v, v' *with* $p(v) = p(v')$*,*

$$\psi^*(v) - \psi^*(v') \;=\; \mathbb{E}\big[d(\nu) \mid v\big] - \mathbb{E}\big[d(\nu) \mid v'\big],$$

where ν *is the specific content of the passage each part comes from. Passages about the same subject receive different predictions whenever their visible parts indicate different specific content.*

Adding the bottleneck and the decoder's copy. Theorem A.3 omits two features of `xPRT`: the decoder also reads a heavily masked copy of the target, and the conditioning vector has limited capacity. With both, the vector is rewarded only for what the copy leaves out.

In the component model, let Y and ν be uniform on $\mathbb{Z}_m$ with $m \geq 2$, and let a passage be its content, $x = (Y, \nu_x)$. The encoder maps its input to a random vector U with capacity $I(U; \text{input}) \leq B$, where $0 < B \leq \log m$. The decoder reads U and a masked copy $\hat{t}$ of the target t. The copy reveals Y with probability β_Y and ν_t with probability β_ν, independently of each other and of everything else, with $0 < \beta_Y, \beta_\nu < 1$. The decoder predicts t under the log loss. By Lemma A.8, the loss of the best decoder is $H(t \mid \hat{t}) - R$, where the *reward*

$$R \;=\; I\big(U;\, t \mid \hat{t}\big)$$

is the information the vector adds about the target beyond the copy. In the cross-passage objective the input is an anchor a and the target is a sibling s. In the within-passage objective the input and the target are both the passage x.

Theorem A.4 (Under a bottleneck, reconstruction disfavors the subject)**.** *In this model:*

(i) Cross-passage. *For every encoder,* $R_{\text{cross}} \;=\; (1 - \beta_Y)\, I(U; Y)$*. Every maximizer has* $I(U; Y) = B$*: it spends its whole capacity on the subject.*

(ii) Within-passage: competition. *Suppose* U *describes the subject and the specific content separately:* $U = (U_Y, U_\nu)$*, where* U_Y *is a randomized function of* Y *and* U_ν *of* ν_x*, each with its own randomness. Let* $b_Y = I(U; Y)$ *and* $b_\nu = I(U; \nu_x)$*, with* $b_Y + b_\nu \leq B$*. Then*

$$R_{\text{within}} \;=\; (1 - \beta_Y)\, b_Y + (1 - \beta_\nu)\, b_\nu,$$

so each component is paid in proportion to how often the copy hides it. If $\beta_Y > \beta_\nu$*, every maximizer has* $b_Y = 0$ *and* $b_\nu = B$*.*

(iii) Within-passage: unrestricted encoders. *Every maximizer of* R_{within} *has* $I(U; x) = B$ *and* $I(U; Y) = I(U; \nu_x) = 0$*, for every* $\beta_Y, \beta_\nu \in (0, 1)$*.*

(iv) Retrieval. *Let a query be a passage* $q = (Y, \nu_q)$ *with fresh specific content, ranked against* N *candidates: one relevant passage with the same subject and* $N - 1$ *independent draws from the corpus. Let the retrieval embedding be any function of* U*, possibly randomized, and let the score be arbitrary, with ties broken at random. For every maximizer in (iii), the rank of the relevant passage is uniform on* $\{1, \ldots, N\}$*. For every cross-passage maximizer with* $B = \log m$*, the subject is a function of* U*, and ranking by agreement of subjects is the best possible ranking.*

All maxima are attained.

Reading Theorem A.4. Part (ii) is the general picture. Reconstruction pays for each kind of information in proportion to how often the decoder's copy hides it. Topical words recur within a passage, so the copy reveals the subject more often than a specific fact ($\beta_Y > \beta_\nu$), and the subject loses the competition for capacity. Part (iii) is the limit: an encoder that may also combine the two components drops the subject entirely, using codes such as $(Y + \nu_x) \bmod m$ that are informative only together with the copy. Cross-passage prediction involves no such competition. The sibling's specific content cannot be predicted from the anchor, so the subject is the only information worth encoding. The comparison with Theorem A.3 isolates the cause: without the copy, the within-passage optimum contains the subject term $M\,p(v)$; with it, the subject is at a disadvantage.

Scope. The results describe the optima of the objectives, not training dynamics. Theorems A.3 and A.4 hold in the component model, and Theorem A.4 abstracts the decoder as predicting the whole target rather than a token distribution. The analysis concerns the predictive pathway, meaning the conditioning vector and the decoder. Retrieval in `xPRT` uses the mean-pooled e, which $\mathcal{L}_{\text{xp}}$ shapes directly. P1 treats passages that share a subject as equally relevant, so the analysis does not model the value of passage-specific content for ranking among them. We attribute the contribution of $\mathcal{L}_{\text{self}}$ to this, but the theory does not establish it.

A.2 Proofs

Throughout, masks, dropout and the teacher use fresh randomness for every text. Texts take countably many values, φ is bounded, and in Theorem A.4 all variables take finitely many values.

The optimal predictor and the optimal ranking

Lemma A.5 (The optimal predictor is the conditional mean)**.** *Let $\psi^*(a) = \mathbb{E}[\varphi(t) \mid a]$. For every predictor ψ:*

(i) $\mathbb{E}\|\varphi(t) - \psi(a)\|^2 = \mathbb{E}\|\varphi(t) - \psi^*(a)\|^2 + \mathbb{E}\|\psi^*(a) - \psi(a)\|^2$;

(ii) *if $\varphi(t)$ and $\psi(a)$ are distributions over the vocabulary,* $\mathbb{E}\,\mathrm{KL}\big(\varphi(t) \,\|\, \psi(a)\big) = \mathbb{E}\,\mathrm{KL}\big(\varphi(t) \,\|\, \psi^*(a)\big) + \mathbb{E}\,\mathrm{KL}\big(\psi^*(a) \,\|\, \psi(a)\big)$.

In both cases the first term does not depend on ψ, and the second vanishes only at $\psi = \psi^*$.

Proof. Part (i).

$$\begin{aligned}\mathbb{E}\|\varphi(t) - \psi(a)\|^2 &\overset{(a)}{=} \mathbb{E}\|\varphi(t) - \psi^*(a)\|^2 + \mathbb{E}\|\psi^*(a) - \psi(a)\|^2 + 2\,\mathbb{E}\big[(\varphi(t) - \psi^*(a))^\top(\psi^*(a) - \psi(a))\big] \\ &\overset{(b)}{=} \mathbb{E}\|\varphi(t) - \psi^*(a)\|^2 + \mathbb{E}\|\psi^*(a) - \psi(a)\|^2.\end{aligned}$$

(a) Add and subtract $\psi^*(a)$, then expand the square.
(b) Tower property: the second factor is a function of a, and $\mathbb{E}[\varphi(t) - \psi^*(a) \mid a] = 0$.

Part (ii).

$$\begin{aligned}\mathbb{E}\,\mathrm{KL}\big(\varphi(t) \,\|\, \psi(a)\big) &\overset{(a)}{=} \mathbb{E}\sum_v \varphi_v(t) \log \frac{\varphi_v(t)}{\psi^*_v(a)} + \mathbb{E}\sum_v \varphi_v(t) \log \frac{\psi^*_v(a)}{\psi_v(a)} \\ &\overset{(b)}{=} \mathbb{E}\,\mathrm{KL}\big(\varphi(t) \,\|\, \psi^*(a)\big) + \mathbb{E}\sum_v \psi^*_v(a) \log \frac{\psi^*_v(a)}{\psi_v(a)}.\end{aligned}$$

(a) Add and subtract $\log \psi^*_v(a)$.
(b) Tower property: the log-ratio is a function of a, and $\mathbb{E}[\varphi_v(t) \mid a] = \psi^*_v(a)$.

The last sum is $\mathrm{KL}(\psi^*(a) \,\|\, \psi(a))$. □

Lemma A.6 (The optimal ranking)**.** *Assume P1. Fix a query q and N candidates $x_1, \dots, x_N$: one relevant passage at a uniformly random index J, and $N-1$ independent draws from the corpus. Then*

$$P\big(J = j \mid q, x_1, \dots, x_N\big) \;=\; \frac{S^*(q, x_j)}{\sum_i S^*(q, x_i)}.$$

Ranking by S^* therefore orders candidates by their posterior probability of being relevant. In particular, placing the maximizer of S^* first maximizes the probability that the relevant passage is ranked first.

Proof. Given $J = j$, the candidates are independent, with $x_j \sim P(\cdot \mid q, \mathrm{rel})$ and $x_i \sim P(\cdot)$ for $i \neq j$. Since J is uniform,

$$P\big(J = j \mid q, x_1, \dots, x_N\big) \;\propto\; P(x_j \mid q, \mathrm{rel}) \prod_{i \neq j} P(x_i) \;=\; \Big(\prod_i P(x_i)\Big) \frac{P(x_j \mid q, \mathrm{rel})}{P(x_j)}.$$

The prefactor does not depend on j. For the ratio,

$$\frac{P(x \mid q, \mathrm{rel})}{P(x)} \overset{\text{(a)}}{=} \sum_y p_y(q) \, \frac{P(x \mid Y = y)}{P(x)} \overset{\text{(b)}}{=} \sum_y \frac{p_y(q)\, p_y(x)}{\pi_y} \;=\; S^*(q, x).$$

(a) P1: given its subject, the relevant passage is independent of the query, so $P(x \mid q, \mathrm{rel}) = \sum_y P(Y = y \mid q)\, P(x \mid Y = y)$.
(b) Bayes' rule: $P(x \mid Y = y)/P(x) = p_y(x)/\pi_y$.

Normalizing over j gives the formula. A ranking that places index j first is correct with probability $P(J = j \mid q, x_1, \dots, x_N)$, which is largest at the maximizer of S^*. □

CROSS-PASSAGE AND WITHIN-PASSAGE PREDICTION

Proof of Theorem A.1.

$$\begin{aligned}
\psi^*(a) \;=\; \mathbb{E}\big[\varphi(s) \mid a\big] &\overset{\text{(a)}}{=} \sum_y P(Y = y \mid a)\, \mathbb{E}\big[\varphi(s) \mid a,\, Y = y\big] \\
&\overset{\text{(b)}}{=} \sum_y p_y(a)\, \mathbb{E}\big[\varphi(s) \mid Y = y\big] \\
&\overset{\text{(c)}}{=} \sum_y p_y(a)\, \mu_y \;=\; M\, p(a).
\end{aligned}$$

(a) Law of total expectation over Y.
(b) P2: given Y, the sibling is independent of the anchor.
(c) Definitions of μ_y and M.

Part (i) is immediate.

$$\psi^*(q)^\top \psi^*(x) \;=\; \sum_{y, y'} p_y(q)\, p_{y'}(x)\, \langle \mu_y, \mu_{y'} \rangle$$

For part (ii), linear independence of the columns of M makes $M^\top M$ invertible, so $M^+ = (M^\top M)^{-1} M^\top$ satisfies $M^+ M = I_k$ and $p(x) = M^+ \psi^*(x)$. By P1, this holds for queries and passages as well as anchors. Let $\Pi = \mathrm{diag}(\pi)$ and $G = (M^+)^\top \Pi^{-1} M^+$. Then

$$\psi^*(q)^\top G\, \psi^*(x) \;=\; \big(M^+ \psi^*(q)\big)^\top \Pi^{-1} \big(M^+ \psi^*(x)\big) \;=\; p(q)^\top \Pi^{-1} p(x) \;=\; S^*(q, x),$$

and Lemma A.6 completes the proof. □

Proof of Remark A.2.

$$\mathbb{E}\big[\varphi(s) \mid a, \hat{s}\big] \overset{(a)}{=} \sum_y P(Y = y \mid a, \hat{s})\, \mathbb{E}\big[\varphi(s) \mid a, \hat{s}, Y = y\big]$$
$$\overset{(b)}{=} \sum_y P(Y = y \mid a, \hat{s})\, \mathbb{E}\big[\varphi(s) \mid \hat{s}, Y = y\big],$$
$$P(Y = y \mid a, \hat{s}) \overset{(c)}{=} \frac{p_y(a)\, P(\hat{s} \mid Y = y)}{\sum_{y'} p_{y'}(a)\, P(\hat{s} \mid Y = y')}.$$

(a) Law of total expectation over Y.
(b) $\hat{s}$ is generated from s with fresh randomness, so P2 gives $(s, \hat{s}) \perp a \mid Y$.
(c) Bayes' rule, using $\hat{s} \perp a \mid Y$.

Both factors depend on a only through $p(a)$. □

Proof of Theorem A.3.

$$\psi^*(a) \;=\; \mathbb{E}\big[\varphi(t) \mid a\big] \overset{(a)}{=} \mathbb{E}\Big[\, \mathbb{E}\big[\varphi(t) \mid a, Y, \nu_t\big] \;\Big|\; a\Big]$$
$$\overset{(b)}{=} \mathbb{E}\big[c(Y) + d(\nu_t) \mid a\big]$$
$$\overset{(c)}{=} \sum_y p_y(a)\, c(y) + \mathbb{E}\big[d(\nu_t) \mid a\big]$$
$$\overset{(d)}{=} \sum_y p_y(a)\, \mu_y - \mathbb{E}\big[d(\nu_t)\big] + \mathbb{E}\big[d(\nu_t) \mid a\big] \;=\; M\, p(a) + \Delta(a).$$

(a) Tower property.
(b) equation CM.
(c) Linearity, then total expectation over Y.
(d) $\mu_y = c(y) + \mathbb{E}[d(\nu_t) \mid Y = y] = c(y) + \mathbb{E}[d(\nu_t)]$, because $\nu_t \perp Y$; also $\sum_y p_y(a) = 1$.

Part (i). The anchor is generated from (Y, ν_a) and randomness of its own passage, all of which are independent of ν_s. Hence $\nu_s \perp a$, so $\mathbb{E}[d(\nu_s) \mid a] = \mathbb{E}[d(\nu_s)]$ and $\Delta \equiv 0$. The same independence gives $a \perp s \mid Y$, so P2 holds in the model.

Part (ii). If $p(v) = p(v')$, then $M\, p(v) = M\, p(v')$, and subtracting the two expressions for ψ^* leaves the difference of the Δ terms. □

THE BOTTLENECK AND THE DECODER'S COPY

Lemma A.7 (Information in a masked view)**.** *Let $w = (w_1, w_2)$, and let $\hat{w}$ reveal w_1 with probability β_1 and w_2 with probability β_2, through indicators $E = (E_1, E_2)$ that are independent of each other and of (U, w). Then*

$$I(U; \hat{w}) = \beta_1\beta_2\, I(U; w) + \beta_1(1 - \beta_2)\, I(U; w_1) + (1 - \beta_1)\beta_2\, I(U; w_2), \tag{5}$$
$$I(U; w \mid \hat{w}) = I(U; w) - I(U; \hat{w}). \tag{6}$$

Proof.

$$I(U; \hat{w}) \overset{(a)}{=} I(U; \hat{w}, E) \overset{(b)}{=} I(U; E) + I(U; \hat{w} \mid E) \overset{(c)}{=} \sum_e P(E = e)\, I(U; \hat{w} \mid E = e).$$

(a) E can be read off $\hat{w}$, since the mask symbol differs from every value.
(b) Chain rule.
(c) $I(U; E) = 0$, because $E \perp U$.

Conditioning on $E = e$ leaves the law of (U, w) unchanged. The four values of e reveal w, w_1, w_2 and nothing, respectively, which gives equation 5. For equation 6, the chain rule gives $I(U; w) + I(U; \hat{w} \mid w) = I(U; w, \hat{w}) = I(U; \hat{w}) + I(U; w \mid \hat{w})$. Moreover $I(U; \hat{w} \mid w) = 0$, because $\hat{w}$ is a function of (w, E) and $E \perp (U, w)$. □

Lemma A.8 (The loss falls by exactly the reward)**.** *For a decoder that reads $(\hat{t}, U)$ and outputs $q(\cdot \mid \hat{t}, U)$,*

$$\mathbb{E}\big[-\log q(t \mid \hat{t}, U)\big] \;=\; H(t \mid \hat{t}) - I(U; t \mid \hat{t}) + \mathbb{E}\,\mathrm{KL}\big(P(\cdot \mid \hat{t}, U) \,\|\, q(\cdot \mid \hat{t}, U)\big).$$

$H(t \mid \hat{t})$ does not depend on the encoder, and the KL term vanishes for the best decoder.

Proof. The expected log loss equals $H(t \mid \hat{t}, U)$ plus the expected KL term, and $H(t \mid \hat{t}, U) = H(t \mid \hat{t}) - I(U; t \mid \hat{t})$ by the definition of conditional mutual information. □

Proof of Theorem A.4. Let $\mathrm{Er}(S)$ denote S passed through an erasure channel with erasure probability $1 - B/\log m$. For S uniform on $\mathbb{Z}_m$, $I(\mathrm{Er}(S); S) = (B/\log m)\, H(S) = B$.

Part (i). Let $U = U(a)$.

$$\begin{aligned} R_{\mathrm{cross}} \;=\; I(U; s \mid \hat{s}) &\overset{(a)}{=} I(U; s) - I(U; \hat{s}) \\ &\overset{(b)}{=} I(U; s) - \beta_Y \beta_\nu\, I(U; s) - \beta_Y(1 - \beta_\nu)\, I(U; Y) - (1 - \beta_Y)\beta_\nu\, I(U; \nu_s) \\ &\overset{(c)}{=} (1 - \beta_Y)\, I(U; Y). \end{aligned}$$

(a) Equation equation 6.
(b) Equation equation 5 with $w = (Y, \nu_s)$.
(c) $\nu_s \perp (U, Y)$, so $I(U; s) = I(U; Y) + I(U; \nu_s \mid Y) = I(U; Y)$ and $I(U; \nu_s) = 0$.

Y is a function of a, so $I(U; Y) \le I(U; a) \le B$. Hence $R_{\mathrm{cross}} \le (1 - \beta_Y)B$, with equality if and only if $I(U; Y) = B$. The encoder $U = \mathrm{Er}(Y)$ attains the bound, with $I(U; a) = I(U; Y) = B$. So the maximum is $(1 - \beta_Y)B$, and every maximizer has $I(U; Y) = B$.

Parts (ii) and (iii). Let $U = U(x)$.

$$R_{\mathrm{within}} \;=\; I(U; x \mid \hat{x}) \overset{(a)}{=} (1 - \beta_Y\beta_\nu)\, I(U; x) - \beta_Y(1 - \beta_\nu)\, I(U; Y) - (1 - \beta_Y)\beta_\nu\, I(U; \nu_x).$$

(a) Equations equation 6 and equation 5 with $w = (Y, \nu_x)$.

Separable codes (part ii). The pairs (U_Y, Y) and (U_ν, ν_x) are independent, so $I(U; x) = b_Y + b_\nu$, $I(U; Y) = b_Y$ and $I(U; \nu_x) = b_\nu$. Substituting,

$$\begin{aligned} R_{\mathrm{within}} &\overset{(b)}{=} \big[1 - \beta_Y\beta_\nu - \beta_Y(1 - \beta_\nu)\big]\, b_Y + \big[1 - \beta_Y\beta_\nu - (1 - \beta_Y)\beta_\nu\big]\, b_\nu \\ &\overset{(c)}{=} (1 - \beta_Y)\, b_Y + (1 - \beta_\nu)\, b_\nu \\ &\overset{(d)}{=} (1 - \beta_\nu)(b_Y + b_\nu) - (\beta_Y - \beta_\nu)\, b_Y \;\overset{(e)}{\le}\; (1 - \beta_\nu)\, B. \end{aligned}$$

(b) Collect the terms in b_Y and b_ν.
(c) Simplify each bracket.
(d) Rearrange.
(e) $b_Y + b_\nu \le B$, and $(\beta_Y - \beta_\nu)\, b_Y \ge 0$ when $\beta_Y > \beta_\nu$.

Equality in (e) holds if and only if $b_Y = 0$ and $b_\nu = B$. The code with constant U_Y and $U_\nu = \mathrm{Er}(\nu_x)$ attains it.

Unrestricted encoders (part iii). In (a), all three coefficients are positive, $I(U; x) \le B$, and mutual information is non-negative. Hence $R_{\mathrm{within}} \le (1 - \beta_Y\beta_\nu)B$, with equality if and only if $I(U; x) = B$ and $I(U; Y) = I(U; \nu_x) = 0$. Let $S = (Y + \nu_x) \bmod m$ and $U = \mathrm{Er}(S)$. For every value of

Y, S is uniform, so $S \perp Y$; likewise $S \perp \nu_x$. Since U depends on x only through S, $I(U;x) = I(U;S) = B$, while $I(U;Y) \le I(S;Y) = 0$ and $I(U;\nu_x) \le I(S;\nu_x) = 0$. The bound is therefore attained, and every maximizer satisfies the equality conditions.

Part (iv). Write $z = g(U)$ for the retrieval embedding, with fresh randomness for each text, and x^+ for the relevant passage.

Maximizer in part (iii). Since x^+ is distributed as a corpus passage, part (iii) gives $I(z(x^+);Y) \le I(U(x^+);Y) = 0$, so $z(x^+) \perp Y$. The query's specific content ν_q is independent of $(z(x^+), Y)$, hence $z(x^+) \perp (Y, \nu_q) = q$. The negatives are independent of (q, x^+) and have the same law as x^+. Given q, the N candidate embeddings are therefore i.i.d., their scores are exchangeable, and with random tie-breaking the rank of x^+ is uniform on $\{1, \dots, N\}$.

Cross-passage maximizer with $B = \log m$. Part (i) gives $I(U;Y) = \log m = H(Y)$, so $Y = f(U)$ almost surely for some function f. Every passage reveals its subject, so $p(x)$ is the point mass at the subject of x and $S^*(q,x) = m \cdot \mathbf{1}[Y_q = Y_x]$. Ranking by $\mathbf{1}[f(U(q)) = f(U(x))]$ is thus ranking by S^*, which is optimal by Lemma A.6. P1 holds in the model, because the query and the relevant passage are independent given Y. □

B Implementation Details

Table 4: **Hyperparameters of `xPRT`** in the two regimes.

	Continual from E5 (default)	From BERT
Initialization	E5-base-unsupervised; MLM head from BERT-base-uncased	BERT-base-uncased
Corpora	100K, 500K, 1M (one pass)	100K (six epochs)
Passage length / min. tail	128 / 64 tokens	128 / 64 tokens
Siblings per anchor used	4	4
Decoder	fresh 1-layer Transformer h_ϕ	same
Conditioning vectors; source	4; linear map of the mean-pooled embedding	same
Masks (light / heavy)	15% / 60%, nested	same
Prediction losses	cross: masked-token cross-entropy; self: full-vocabulary KL, $\tau = 1$	same
Terms (weights)	$\mathcal{L}_{\text{xp}}$ 1, $\mathcal{L}_{\text{cross}}$ 0.1, $\mathcal{L}_{\text{self}}$ 1, $\mathcal{L}_{\text{MLM}}$ 1, $\mathcal{L}_{\text{unif}}$ 0.01	same
$\mathcal{L}_{\text{xp}}$ negatives	511 in global batch, same-document masked	same
Retention regularizer	Eq. 7, weight $40 \to 0$, $\tau_r = 0.075$	none
Optimizer	AdamW (wd 0.01, $\epsilon = 10^{-4}$); Adam for biases/LayerNorm	same
Peak LR; warm-up; decay	10^{-6}; 5% of steps; linear	10^{-4}; 68 steps; linear
Batch	512	512
Backbone freeze	first 100 steps	none
Number of epochs	1	6
Precision; clipping; seed	bf16; 1.0; 42, 43, 44	bf16; 1.0; 42, 43, 44
Hardware	32 × AMD MI300X	8 × AMD MI300X

Inputs are `[CLS] query:` a `[SEP]` for anchors and `[CLS] passage:` s `[SEP]` for siblings. The special tokens and the prefix tokens are never masked. Masked tokens are always replaced by [MASK] (no random or unchanged replacement). One pass visits every group exactly once in a random order that is shared across GPUs. The ablations of Table 3 keep the global batch of 512, without gradient accumulation, on 8, 16 or 32 GPUs.

C Group Selection

Scores. Let $\hat{e}_i$ be the normalized E5 embedding of passage i, computed on `passage:` followed by the passage, with mean pooling.

- **Centrality.** Passage i of document d has centrality $s_i = \hat{e}_i^\top \bar{e}_d / \|\bar{e}_d\|$, where $\bar{e}_d$ is the mean of $\hat{e}_j$ over the passages of d, including i.
- **Informativeness.** ι_i is the mean of $\mathrm{idf}(t) = \ln \frac{N+1}{\mathrm{df}(t)+1} + 1$ over the unique tokens of i. Here $N = 5{,}150{,}388$ is the number of passages in the first million documents and $\mathrm{df}(t)$ counts the passages that contain token t.
- **Rank fusion.** Ranks are computed within a document, in descending order, and tied scores share the average rank. The fused score of a candidate is

$$\mathrm{RRF} = \frac{2}{10 + r_{\text{dense}}} + \frac{1}{10 + r_{\text{lex}}}.$$

Selection.

- **Anchors** are the passages with $s_i > 0.825$. Up to ten are kept per document, ranked by the fusion of s_i and ι_i; ties are broken by s_i, then ι_i, then position.
- **Siblings.** For an anchor a, each other passage c is ranked by the fusion of $\hat{e}_a^\top \hat{e}_c$ and the cosine between the TF-IDF vectors (raw counts × idf) of a and c. The top four are kept.

Example. The group in Figure 2 comes from a 10-passage blog post about Chromebooks.

- The anchor is passage 1, on account synchronization and sandboxing; its centrality is 0.842.
- The siblings, in rank order, are passages 2 (operating-system updates and security, continuing the anchor; E5 cosine 0.777), 6 (losing connectivity on a train; 0.750), 7 (cloud-centric machines for businesses; 0.750) and 4 (an audio problem; 0.646).

What the filters remove, and what they miss. The passage least related to a representative anchor is often page furniture. On the page of Figure 1 it is a dictionary widget. On an article about the Ford F-Series it is a list of music-video links, on a policy report on carbon capture a list of press contacts, and on an essay about digital magazines a stock ticker. These passages have an E5 cosine of 0.38–0.45 to the anchor, whereas the most related passage of the same page has 0.85–0.95. They also have low centrality (0.46–0.51) and never become anchors. The anchor filter, however, measures how representative a passage is of its page, not its quality. On a site-policy page, the passages of the legal disclaimer are central (0.95) and pass the filter; the passage least related to them is a list of recent visitors (cosine 0.48). Conversely, the introduction of an interview with a band fails the filter (0.79), presumably because the questions and answers that fill the rest of the page dominate its centroid. Table 5 shows these passages.

D Full Objective

The expressions for $\mathcal{L}_{\text{cross}}$ and $\mathcal{L}_{\text{self}}$ in Sections 2.3 and 2.4 are written for a single group. Over a batch of B groups $g_b = (a_b, s_{b1}, \ldots, s_{bn})$, each is averaged over masked tokens rather than over groups. Let $S(y \mid x) = |M_H(y)|\, C(y \mid x)$ be the cross-entropy of Eq. 3 summed over the masked positions of y, and $S_D(y) = |M_H(y)|\, D(y)$ the summed KL divergence of Eq. 4.

- **Cross-passage supervision.**

$$\mathcal{L}_{\text{cross}} = \frac{1}{2}\left[\frac{\sum_b \sum_{j=1}^{n} S(a_b \mid s_{bj})}{n \sum_b |M_H(a_b)|} + \frac{\sum_b \sum_{j=1}^{n} S(s_{bj} \mid a_b)}{\sum_b \sum_{j=1}^{n} |M_H(s_{bj})|}\right].$$

  Each passage has a single heavy view, so the n predictions of an anchor share its masked positions; hence the factor n in the first denominator.

- **Within-passage supervision.**

$$\mathcal{L}_{\text{self}} = \frac{1}{2}\left[\frac{\sum_b S_D(a_b)}{\sum_b |M_H(a_b)|} + \frac{\sum_b S_D(s_{b1})}{\sum_b |M_H(s_{b1})|}\right].$$

- **Relation to Section 2.3.** If every heavy view masks the same number of tokens, both expressions reduce to the batch mean of the per-group terms in Section 2.3. Passages have 64 to 128 tokens (Appendix B), so the two differ only through shorter trailing passages.
- **MLM.** $\mathcal{L}_{\text{MLM}}$ is the cross-entropy of the MLM head on the lightly masked positions of a and s_1.
- **Uniformity.** $\mathcal{L}_{\text{unif}}$ is computed over the anchors on each device.

All prediction losses are computed in fp32 from bf16 logits.

E Retention Regularizer

In the continual setting, every method adds

$$\mathcal{L}_{\text{ret}} = \tfrac{1}{2}\Big[\mathrm{KL}\big(\sigma(R_1 R_2^\top/\tau_r) \,\big\|\, \sigma(E_1 E_2^\top/\tau_r)\big) + \mathrm{KL}\big(\sigma(R_2 R_1^\top/\tau_r) \,\big\|\, \sigma(E_2 E_1^\top/\tau_r)\big)\Big], \qquad (7)$$

where E_1, E_2 are normalized current embeddings of two sets of passages in the batch, R_1, R_2 are the corresponding embeddings of the frozen initial model, σ is a row-wise softmax, and $\tau_r = 0.075$. Its weight decays as $\lambda_{\text{ret}}(t) = 40\left(1 - \sqrt{(t-1)/(T-1)}\right)$ over the T training steps, which protects the initial model early and lets the pre-training objective dominate late. Equation 7 is applied to every E5-initialized run and to no BERT-initialized run.

Table 5: **Passages that pass and fail the anchor filter** on seven FineWeb pages (s: centrality; $\tau_a = 0.825$; number of passages in parentheses). On the first five pages, the failing passage is also the one least related to the passing passage (E5 cosine 0.38–0.45). The last two pages show what the filter misses: a legal disclaimer passes, and an informative introduction fails. Passages are abridged (ellipses) and shown as the lowercased model input; contact names and phone numbers are elided.

Page	Passes the filter	Fails the filter
Baseball no-hitter article (47)	($s = 0.94$) … a game in which a team allows no runners to reach base at all is a perfect game, as well as being a shutout victory. a perfect game is by definition also a no-hitter. 22 perfect games have been recorded in mlb history …	($s = 0.47$) dictionary and translator for handheld new: sensagent is now available on your handheld … with a sensagentbox, visitors to your site can access reliable information on over 5 million pages provided by sensagent.com …
Home-energy guide (25)	($s = 0.93$) … energy efficiency starts with learning how to recognize a problem, making the right choices, and sometimes changing old habits. this book will help you become more aware of energy use in your home and in the world …	($s = 0.50$) … click for more about paul scheckel's important new book deus ex machina by paul freundlich click for more about paul scheckel's important new book … copyright © 2003-2012 by the public press: the page, all rights reserved
Essay on digital magazines (34)	($s = 0.92$) … for the past year i've been pushing the theory that the age of tablets will give print media one last bite at the apple – and publishing companies that are able to make the transition could one day thrive again …	($s = 0.46$) ford motor co | 15.08 | 0.44 | 3.01% | general electric co | 23.46 | 0.19 | 0.82% | glass employees speak openly on public concerns more …
Policy report on coal and carbon capture (25)	($s = 0.94$) … although an emission performance standard that requires ccs systems for all new coal plants would pose a daunting technological and economic challenge, it will ultimately assure coal a secure and important role in the future u.s. energy mix …	($s = 0.51$) … print: [name] (women's issues, poverty, legal progress) [phone] or email@example.com print: [name] (spanish language and ethnic media) [phone] … tv: … radio: … web: …
Article on the Ford F-Series (50)	($s = 0.91$) … the f-series is a series of full-size pickup trucks from ford motor company which has been sold continuously for over six decades. the most popular variant of the f-series is the f-150 …	($s = 0.46$) … available to download on itunes including remixes of: tiesto, hard rock sofa & deniz koyu http://smarturl.it/… macklemore & ryan lewis present the official music video for can't hold us …
What the filter misses		
Site policy (23)	($s = 0.95$) … this site is controlled and operated by thesunsetchart.com from its offices within the state of california, united states of america. thesunsetchart.com makes no representation that materials in the site are appropriate or available for use in other locations …	($s = 0.57$) some of our recent visitors… bellevue, waunited states mountain view, caunited states seattle, waunited states … copyright 2013 intelligenesis, all rights reserved …
Interview with a band (22)	($s = 0.91$) … one of the things i admire most about cocore is the spirit of experimentation. can you describe your relationship to experimental music? who or what inspires you to push the limits? …	($s = 0.79$) cocore is one of south korea's first and greatest indie rock bands. in the beginning of their career, they made a name for themselves by covering nirvana and penning their own grunge anthems …

- **`xPRT`.** E_1, E_2 are the mean-pooled embeddings of the lightly masked anchor and top sibling. R_1, R_2 are computed online by a frozen copy of E5 on the unmasked, prefixed passages. Similarity matrices are gathered across GPUs (512×512).

- **Single-passage baselines.** The reference embeddings of the anchors are precomputed with the same frozen model on the same prefixed passages and stored in fp16. Both sets are the anchors of the batch.

- **Schedule.** For all methods, the weight halves after a quarter of training and reaches 0 at the last step.

F BASELINES

Table 6: **Baseline configurations.** Every baseline uses its official implementation, the same groups (the anchor passage; for Revela, the anchor with its siblings and anchors of other documents) and the same steps, batch size, learning-rate schedule and, in the continual setting, retention regularizer as `xPRT`.

Method	Native objective and settings	Passages per group
RetroMAE	encoder MLM at 30% masking; enhanced one-layer decoder reconstructs at 50%	1 (anchor)
DupMAE	RetroMAE plus bag-of-words prediction from token states (weight 0.1)	1 (anchor)
SimLM	replaced-language modeling with a frozen ELECTRA-base generator; masks 30% / 50%; two-layer decoder; generator MLM weight 0.2	1 (anchor)
BoW-DPR	encoder MLM at 30%; bag-of-words prediction (coefficient 1.0, cosine-decayed to 0.1); head mask 15%	1 (anchor)
Revela	next-token prediction by Qwen2.5-0.5B (LoRA, rank 256, $\alpha = 64$) with in-batch attention weighted by the retriever's similarities; 128-token passages	16 (anchor, 4 siblings, 11 anchors of other documents)
`xPRT`	Eq. 1	5 (anchor + 4 siblings)

G EXTENDED RELATED WORK

Contrastive pre-training for retrieval. Without labels, positive pairs have been built with the inverse cloze task (Lee et al., 2019), body–first-section and wiki-link pairs (Chang et al., 2020), recurring spans (Ram et al., 2022) and random cropping (Izacard et al., 2022), and combined with corpus-aware pre-training on spans of the same document (Gao & Callan, 2022); large-scale weakly supervised pairs, filtered with a consistency check, underlie today's general-purpose embedders (Wang et al., 2022; Li et al., 2023a). `xPRT` shares this contrastive form and changes the pair. Spider pairs passages of a document that share a recurring span, a lexical notion of relatedness; `xPRT` pairs an anchor that must represent its document with the passage most related to it by fused dense and lexical similarity, and every document contributes several such pairs. Our ablations show that this selection matters (§4.3), and `xPRT` adds a signal on top of a model pre-trained contrastively on weakly supervised pairs rather than replacing such pre-training (§4.1).

Bottleneck auto-encoding. Weak decoders force information into a single vector in SEED (Lu et al., 2021), Condenser (Gao & Callan, 2021), RetroMAE (Xiao et al., 2022) and its duplex extension DupMAE (Liu et al., 2023), SimLM with replaced-token detection (Wang et al., 2023), and LexMAE with a lexicon bottleneck (Shen et al., 2023); BoW-DPR drops the decoder in favor of bag-of-words prediction (Ma et al., 2024). The closest work is CoT-MAE (Wu et al., 2023), which reconstructs a span with the help of the embedding of its adjacent span. `xPRT` uses prediction differently: as content supervision alongside a contrastive objective, with targets drawn from related passages anywhere in the document.

Learning from related passages. Revela (Cai et al., 2026) trains a retriever through language modeling: a language model predicts each passage of a batch while attending to the other passages, weighted by the retriever's similarities, so the retriever learns which passages help to predict a given one. `xPRT` shares the premise that related passages carry a training signal, but it selects the related passages within a document, makes the retriever's own embedding align with them and predict them, and needs no language model. Under our protocol, Revela does not improve E5 (§4.1) and, from BERT, stays below SimLM (§4).

Predictive self-distillation. Predicting a model's own targets from a privileged view underlies self-supervised learning in vision and speech (Grill et al., 2020; Caron et al., 2021; Baevski et al., 2022; Assran et al., 2023). `xPRT` uses such targets in its within-passage term $\mathcal{L}_{\text{self}}$, where the prediction must pass through the passage's own embedding; its cross-passage term $\mathcal{L}_{\text{cross}}$ instead predicts the actual masked tokens of a related passage.

Conditioning on continuous vectors. Prefix and prompt tuning condition a frozen or shared model on a few continuous vectors (Li & Liang, 2021; Lester et al., 2021). Gist tokens, Auto-Compressor and ICAE compress contexts into such vectors for later generation (Mu et al., 2023; Chevalier et al., 2023; Ge et al., 2024). `xPRT` uses the same interface as a training-time bottleneck and asks the vectors to support prediction of other text rather than reconstruction of the compressed one.

Continual pre-training and forgetting. Continued pre-training adapts models to new data (Gururangan et al., 2020) but can erase prior capabilities (Kirkpatrick et al., 2016; Li & Hoiem, 2018). Our retention regularizer follows similarity-preserving and relational distillation (Tung & Mori, 2019; Park et al., 2019): it preserves the initial model's in-batch similarity structure and is annealed away during training.

H Additional Analysis

Examples. Table 7 shows six queries, each from a different debate, among the 16 for which `xPRT` ranks the counter-argument first and both E5 and BoW-DPR rank it fifth or lower. In each, E5's first passage is about the subject of the query but does not answer it: it is another argument of the same debate (drugs, tourism in Tunisia, the International Criminal Court), the counter-argument to another point (assisted suicide), or an argument from a neighboring debate (EU foreign policy, aid for accepting migrants). `xPRT`'s first passage responds to the query's point directly, down to its example of an embassy in Djibouti or its tourist numbers for 2011. This is the kind of relation that training on related passages of a page rewards: what a passage is about, and which passages address it.

I Additional Results

Table 8 complements Table 1 with win/loss counts and retrieval metrics beyond nDCG@10, and Tables 9–11 give per-task scores.

Table 7: **ArguAna queries for which `xPRT` ranks the counter-argument first and E5 does not.** Six of the 16 queries for which `xPRT` ranks the counter-argument first and both E5 and BoW-DPR rank it fifth or lower, each from a different debate (topic in italics). Passages are abridged from the corpus (ellipses); ranks come from the full-precision re-encoding described in the text.

Legalization of drugs. Rank of the counter-argument: E5 9, BoW-DPR 15, `xPRT` 1.	
Query	Legalisation reduces crime. The illegality of drugs fuels a huge amount of crime that could be eliminated if drugs were legalised. Price controls would mean that addicts would no longer have to steal to fund their habits …
E5, first	More people will take drugs if they are legal. Considering that drug use has so many negative consequences, it would be disastrous to have it increase …
`xPRT`, first	Drugs affect how people think, and they take away their ability to control their actions rationally, and so people on drugs are more likely to commit crimes … "Most drug crimes aren't committed by people trying to pay for drugs; they're committed by people on drugs."
UK influence and EU membership. Rank of the counter-argument: E5 10, BoW-DPR 19, `xPRT` 1.	
Query	The EU is a force multiplier. The UK gets more bang for the buck as a result of being a member of the EU. It has representation in more countries … For example the EU have representation in Djibouti …
E5, first	*(from a debate on a common EU foreign policy)* The position of High Representative will be, and has been, largely powerless …
`xPRT`, first	It is undeniable that in some areas the EU is a force multiplier. But many of the issues it uses this leverage on are not areas of concern to a UK that has left the EU … could the UK have an embassy in Djibouti? Certainly if it wished …
Whether Tunisia should rely on tourism. Rank of the counter-argument: E5 8, BoW-DPR 7, `xPRT` 1.	
Query	Vulnerable to unrest. Relying on tourism ensures that the economy is at the mercy of unrest. The violence and break down in law and order following the Tunisian revolution resulted in a notable decrease in tourists … from 6,487,000 in 2010 to 4,456,000 in 2011 …
E5, first	Investment. Tourism should be relied upon for economic growth as it attracts significant foreign investment. Tourism is the largest form of foreign currency income …
`xPRT`, first	The long term affects that instability have on industries such as tourism is overstated. … tourism has recovered from the low point of 2011. In the first ten months of 2013 Tunisia attracted 5.5 million tourists …
An enforcement arm for the ICC. Rank of the counter-argument: E5 5, BoW-DPR 12, `xPRT` 1.	
Query	ICC enforcement would create resentment. … it would be a foreign force, that may or may not be seen as legitimate, attempting to arrest a native of that country. The result would be resentment in the community at the intrusion …
E5, first	An ICC Enforcement arm would bring in a higher proportion of defendants in to trial. Eight out of the thirty people indicted by the ICC … are still alive and avoiding justice …
`xPRT`, first	… Often a problem with arresting wanted international criminals is their support in the community … The national government's enforcement arm may be as unwelcome as the ICC …
Legalization of assisted suicide. Rank of the counter-argument: E5 7, BoW-DPR 11, `xPRT` 1.	
Query	It is vital that a doctor's role not be confused. The guiding principle of medical ethics is to do no harm … assisted suicide also places an unreasonable burden on doctors …
E5, first	*(the counter-argument to another point)* However, the idea that we should not kill is not absolute … Concerning the notion that legalised voluntary euthanasia might lead to involuntary euthanasia being carried out, there is no evidence to suggest this …
`xPRT`, first	At the moment, doctors are often put into an impossible position. A good doctor will form close bonds with their patients … when a patient … expresses a strong desire to die, they are legally unable to help …
An absolute right to asylum. Rank of the counter-argument: E5 107, BoW-DPR 9, `xPRT` 1.	
Query	The Whole System is broken. It is not clear that the system works at all. The majority of those who apply for asylum are working-age males, which implies that there is a strong economic angle …
E5, first	*(from a debate on offering aid to developing countries that accept migrants)* Migrants will simply return to the countries they have been sent from …
`xPRT`, first	Much of the fear of the asylum system being used by economic migrants is simply media hysteria and xenophobia. The vast majority of asylum claims (in the UK around 75%) are still rejected, which shows the system works …

Table 8: **Average score, win/loss counts and retrieval metrics** (continual from E5; $\times 100$). Δ is the change in Avg. with respect to E5, computed before rounding, and W/L counts the tasks on which a model is better/worse than E5. Retrieval metrics are averaged over the ten retrieval tasks.

		MTEB(eng, v2), 41 tasks			Retrieval, 10 tasks			
Corpus	Method	Avg.	Δ	W/L	nDCG@10	MRR@10	R@10	R@100
–	E5 (initial)	53.06	–	–	43.26	51.25	43.31	65.25
100K	RetroMAE	52.92	−0.14	17/24	42.41	50.34	42.82	65.05
	DupMAE	52.97	−0.09	20/21	42.44	50.34	42.82	65.06
	SimLM	51.01	−2.05	18/23	36.93	46.65	36.26	56.56
	BoW-DPR	53.14	+0.08	23/18	43.11	51.95	42.92	64.16
	Revela	53.02	−0.05	23/18	42.99	50.63	43.07	65.28
	xPRT	**54.86**	**+1.80**	**34/7**	**45.78**	**53.88**	**46.05**	**67.50**
500K	RetroMAE	52.99	−0.07	22/19	42.38	50.14	42.85	65.03
	DupMAE	53.00	−0.06	21/20	42.32	50.22	42.78	65.06
	SimLM	53.32	+0.26	26/15	42.01	49.54	41.26	63.95
	BoW-DPR	53.44	+0.38	27/14	43.38	51.09	42.91	65.14
	Revela	52.89	−0.17	14/27	42.66	50.16	42.83	65.17
	xPRT	**54.93**	**+1.87**	**36/5**	**45.67**	**53.78**	**46.00**	**67.42**
1M	RetroMAE	52.97	−0.09	22/19	41.94	49.66	42.73	64.94
	DupMAE	52.94	−0.12	22/19	41.99	49.82	42.68	64.94
	SimLM	51.36	−1.70	11/30	37.93	46.33	37.49	59.19
	BoW-DPR	53.02	−0.04	23/18	42.36	50.30	41.81	64.39
	Revela	52.80	−0.26	16/25	42.47	49.51	42.89	65.31
	xPRT	**54.89**	**+1.83**	**35/6**	**45.33**	**53.53**	**45.81**	**67.56**

Table 9: **Retrieval (nDCG@10,** $\times 100$**) per task across scales** (continual from E5).

Task	E5 (initial)	xPRT 100K	xPRT 500K	xPRT 1M
ArguAna	40.36	46.12	45.77	46.04
CQADupstack-Gaming	53.45	56.06	56.12	56.28
CQADupstack-Unix	35.26	37.00	37.00	36.50
Climate-FEVER-HN	16.48	20.06	20.47	20.94
FEVER-HN	65.58	69.65	69.83	69.78
FiQA-2018	40.00	38.87	38.30	37.71
HotpotQA-HN	56.18	59.19	59.19	58.87
SCIDOCS	21.35	21.84	21.77	21.80
TREC-COVID	64.39	62.96	62.52	60.90
Touché-2020 v3	39.51	46.08	45.72	44.51
Average	43.26	45.78	45.67	45.33

Table 10: **Per-task main scores at 100K** (continual from E5; $\times 100$).

Type	Task	E5	RetroMAE	DupMAE	SimLM	BoW-DPR	Revela	xPRT
Class.	AmazonCounterfactual	64.22	63.82	63.85	67.08	65.76	64.54	66.34
	Banking77	75.81	75.71	75.70	76.64	76.48	75.83	77.07
	Imdb	66.15	68.22	68.08	60.17	61.01	67.24	64.66
	MTOPDomain	62.54	62.57	62.64	63.37	63.41	62.55	63.24
	MassiveIntent	27.35	27.33	27.31	29.62	28.89	27.49	28.83
	MassiveScenario	34.38	34.35	34.32	36.46	35.91	34.44	35.62
	ToxicConversations	63.18	63.32	63.35	64.25	65.00	64.03	65.35
	TweetSentimentExtraction	52.11	52.35	52.37	52.17	52.55	52.68	52.82
Clust.	ArXivHierarchicalP2P	61.97	61.64	61.37	55.14	59.79	61.38	61.61
	ArXivHierarchicalS2S	58.83	58.52	58.52	57.33	58.34	58.94	60.19
	BiorxivP2P.v2	41.58	41.49	41.22	28.21	38.68	41.89	41.26
	MedrxivP2P.v2	37.13	37.12	36.87	29.69	33.90	36.91	37.57
	MedrxivS2S.v2	35.67	35.03	35.18	33.34	34.42	35.68	35.68
	StackExchange.v2	55.04	55.10	55.39	47.32	50.76	55.23	57.11
	StackExchangeP2P.v2	39.89	39.71	40.02	34.36	38.87	39.54	42.46
	TwentyNewsgroups.v2	38.37	36.96	38.59	31.53	33.86	37.35	38.07
Pair	SprintDuplicateQuestions	90.77	90.95	90.95	93.94	92.58	90.91	91.25
	TwitterSemEval2015	62.76	62.78	62.78	63.55	63.24	62.81	63.77
	TwitterURLCorpus	84.03	84.00	84.01	84.51	84.40	83.77	84.48
Rerank	AskUbuntuDupQuestions	57.99	57.68	57.66	57.42	57.62	57.96	57.97
	MindSmallReranking	31.62	31.73	31.72	31.16	31.58	31.97	31.70
Retr.	ArguAna	40.36	39.48	39.56	22.50	38.89	39.44	46.12
	CQADupstack-Gaming	53.45	53.36	53.45	50.76	51.75	53.28	56.06
	CQADupstack-Unix	35.26	34.99	35.12	33.62	34.74	35.55	37.00
	Climate-FEVER-HN	16.48	15.72	15.68	15.51	16.90	14.55	20.06
	FEVER-HN	65.58	63.79	63.65	49.52	63.11	64.92	69.65
	FiQA-2018	40.00	39.80	39.57	27.88	39.46	39.50	38.87
	HotpotQA-HN	56.18	55.80	55.84	60.79	59.52	57.02	59.19
	SCIDOCS	21.35	21.38	21.37	19.49	21.13	21.48	21.84
	TREC-COVID	64.39	60.95	61.16	57.39	65.78	64.72	62.96
	Touché-2020 v3	39.51	38.80	38.97	31.79	39.87	39.42	46.08
STS	BIOSSES	71.97	73.21	73.33	82.63	79.14	71.49	81.35
	SICK-R	68.68	68.86	68.89	67.97	69.02	68.46	70.04
	STS12	57.84	57.80	57.84	57.54	58.04	57.93	59.50
	STS13	73.46	73.92	73.90	74.97	74.37	73.17	73.71
	STS14	64.04	64.28	64.30	65.68	65.08	63.73	65.99
	STS15	75.44	76.02	76.04	79.98	78.06	74.87	78.43
	STS17	47.57	47.62	47.61	48.47	48.09	47.71	50.50
	STS22.v2	40.74	40.61	40.67	43.84	45.42	40.15	47.58
	STSBenchmark	70.51	70.95	70.96	75.81	73.78	70.99	74.53
Summ.	SummEval.v2	31.46	32.19	32.13	28.10	29.57	32.13	32.87

Table 11: **Per-task main scores when pre-training from BERT** (100K groups, six epochs; $\times 100$; xPRT: mean of three seeds).

Type	Task	RetroMAE	DupMAE	SimLM	BoW-DPR	Revela	xPRT
Class.	AmazonCounterfactual	69.80	69.45	67.92	69.77	63.41	59.99
	Banking77	56.93	53.69	60.24	59.73	60.37	65.69
	Imdb	60.84	59.13	65.00	62.73	63.69	61.56
	MTOPDomain	54.67	53.28	52.57	53.88	52.19	54.30
	MassiveIntent	24.01	22.75	21.32	22.57	21.98	24.11
	MassiveScenario	29.21	28.18	26.82	27.80	27.73	29.64
	ToxicConversations	66.22	66.48	68.94	70.36	67.49	65.44
	TweetSentimentExtraction	49.75	48.34	53.37	52.87	53.15	49.94
Clust.	ArXivHierarchicalP2P	54.52	54.09	56.92	56.07	55.96	57.44
	ArXivHierarchicalS2S	52.67	51.42	54.36	53.62	52.62	54.02
	BiorxivP2P.v2	34.43	32.26	35.00	33.28	33.22	36.86
	MedrxivP2P.v2	31.66	30.52	32.63	31.36	31.84	34.87
	MedrxivS2S.v2	27.42	26.89	28.33	28.10	28.64	32.18
	StackExchange.v2	42.62	42.10	50.35	48.42	53.63	57.91
	StackExchangeP2P.v2	33.09	32.44	34.97	33.70	34.19	39.05
	TwentyNewsgroups.v2	19.23	16.59	28.76	28.30	32.16	41.97
Pair	SprintDuplicateQuestions	76.48	73.13	74.55	72.23	73.26	90.27
	TwitterSemEval2015	60.04	60.00	60.69	60.91	58.77	57.88
	TwitterURLCorpus	80.31	79.00	79.41	80.56	77.60	80.43
Rerank	AskUbuntuDupQuestions	49.66	49.71	48.87	50.31	49.54	51.50
	MindSmallReranking	27.54	26.56	28.33	28.19	29.31	30.04
Retr.	ArguAna	33.49	31.61	41.29	35.58	28.49	40.69
	CQADupstack-Gaming	25.45	25.14	29.51	28.39	25.75	36.05
	CQADupstack-Unix	14.84	13.99	16.13	14.43	13.58	20.32
	Climate-FEVER-HN	8.90	8.44	8.97	8.99	7.17	14.54
	FEVER-HN	20.59	18.00	19.30	18.08	14.17	45.00
	FiQA-2018	7.21	5.90	7.64	6.62	6.44	16.19
	HotpotQA-HN	28.66	27.95	28.84	23.96	30.52	43.03
	SCIDOCS	8.81	8.31	8.82	7.65	6.96	13.27
	TREC-COVID	35.11	35.15	29.24	30.57	27.11	31.76
	Touché-2020 v3	13.42	14.42	17.10	16.10	14.30	31.29
STS	BIOSSES	64.31	64.82	67.90	61.37	65.10	74.24
	SICK-R	61.51	61.21	61.48	64.55	66.08	65.05
	STS12	46.87	43.42	46.41	47.46	48.59	53.37
	STS13	63.88	61.13	64.38	68.72	70.58	72.99
	STS14	58.22	55.69	56.54	60.26	56.04	63.46
	STS15	72.71	70.46	68.91	74.46	70.29	74.43
	STS17	35.15	33.93	27.34	32.40	34.85	35.76
	STS22.v2	27.02	26.17	29.64	24.05	28.95	33.46
	STSBenchmark	59.70	57.12	61.72	64.51	64.83	65.73
Summ.	SummEval.v2	33.28	33.77	31.38	32.29	31.35	26.23