# BENCHMARKING AND BUILDING LONG-CONTEXT RETRIEVAL MODELS WITH LOCO AND M2-BERT

**Jon Saad-Falcon, Daniel Y. Fu, Simran Arora, Neel Guha & Christopher Ré**
Department of Computer Science
Stanford University
Stanford, CA 94305, USA
{jonsaadfalcon,danfu,simarora,nguha,chrismre}@stanford.edu

## ABSTRACT

Retrieval pipelines are an integral component of many machine learning systems. However, they perform poorly in domains where documents are long (e.g., 10K tokens or more) and where identifying the relevant document requires synthesizing information across the entire text. Developing long-context retrieval encoders suitable for these domains raises three challenges: (1) how to evaluate long-context retrieval performance, (2) how to pretrain a base language model to represent both short contexts (corresponding to queries) and long contexts (corresponding to documents), and (3) how to finetune this model for retrieval under the batch size limitations imposed by GPU memory constraints. To address these challenges, we first introduce LoCoV1, a 12 task benchmark constructed to measure long-context retrieval where chunking is not possible or not effective. We next present the M2-BERT retrieval encoder, an 80M parameter state-space encoder model built from the Monarch Mixer architecture, capable of scaling to documents up to 32K tokens long. We describe a pretraining data mixture which allows this encoder to process both short and long context sequences, and a finetuning approach that adapts this base model to retrieval with only single-sample batches. Finally, we validate the M2-BERT retrieval encoder on LoCoV1, finding that it outperforms competitive Transformer-based models by at least 22.2 points, despite containing $90\times$ fewer parameters.

## 1 INTRODUCTION

Retrieval is an essential component of machine learning pipelines for tasks like search, question-answering, dialogue, and fact verification Chen et al. (2017); Lewis et al. (2021); Dinan et al. (2019); Petroni et al. (2021). Most retrieval systems rely on pretrained text models that are only capable of processing short input sequences (e.g., approximately 512 to 8192 tokens) Reimers & Gurevych (2019); Lassance & Clinchant (2022); Karpukhin et al. (2020); Santhanam et al. (2022). Yet from our analysis of domain-specific datasets, such as those in law and medicine (Section 4), the documents or queries may be tens of thousands of tokens long, and identifying the relevant document requires synthesizing information across a long text sequence Li et al. (2023). Examples include legal contracts, company financial documents, patient notes, screenplays, and other documents with specific contextual details and cross-document references Bai et al. (2023); Shaham et al. (2022); Dasigi et al. (2021); Xu et al. (2023). Our work explores how to benchmark and build high quality and efficient retrieval systems for long-context corpora.

Popular retrieval models are built using the Transformer architecture Vaswani et al. (2023), which scales quadratically in sequence length, making it expensive to extend existing retrieval recipes to the long-context setting. Recent work on *state-space* architectures, such as S4 Gu et al. (2022), Mamba Gu & Dao (2023), Monarch Mixer Fu et al. (2023a), and more Wang et al. (2022); Smith et al. (2023); Hasani et al. (2022); Fu et al. (2023b); Poli et al. (2023), suggests that the subquadratic scaling properties enjoyed by these models make them amenable for long contexts. However, adapting state-space models for retrieval raises three challenges:

- **Evaluation**: Existing benchmarks for retrieval contain query-document pairs where the relevant information is contained either within the first 512 tokens of the document, or within a small sequence of text Thakur et al. (2021); Muennighoff et al. (2022). As a result, naive truncation-based and chunking baselines perform nearly optimally, regardless of the document length. Validating long-context retrievers thus requires benchmarks on which identifying the relevant document requires

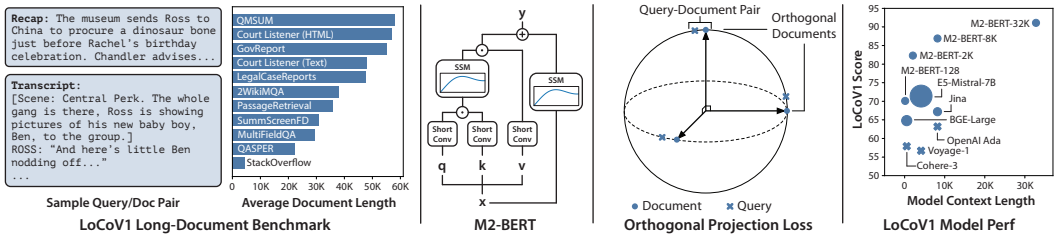


Figure 1: **Left:** The LoCoV1 long document retrieval benchmark and the average document length of its constituent datasets. **Center Left:** M2-BERT sequence mixer. **Center Right:** The orthogonal projection loss. **Right:** Performance of various retrieval models and M2-BERT at different sequence lengths on LoCoV1. Circles are open models, where circle area corresponds to model size. X marks are closed models.

reasoning across longer spans of text (e.g. medical, financial, or legal documents with many repeated textual phrases amongst in-class documents but key contextual details throughout the document).

- **Pretraining**: Retrieval encoders must be pretrained to process both short sequences (corresponding to queries) and long sequences (corresponding to documents). Prior work on state-space model pretraining, in contrast, has focused exclusively on tasks requiring pretraining on uniformly shorter textual inputs Fu et al. (2023a); Wang et al. (2022). In Section B.4.1, we show that naive pretraining strategies for model weight initialization are insufficient for preparing retrieval encoders to long input sequences.
- **Finetuning**: Retrieval encoders are usually finetuned from pretrained models using a contrastive loss function (multiple negatives ranking loss, known as MNRL). MNRL treats other in-batch positive passages as negative passages for a given query Reimers & Gurevych (2019). MNRL then pushes the embeddings of the positive pair and passage together, while pushing apart the embeddings of negative passages. With a large batch size, this loss creates an embedding geometry that aligns positive pairs together, while distributing them uniformly around the embedding hypersphere Wang & Isola (2022); Leszczynski et al. (2022); Henderson et al. (2017) Training long-context models with the requisite batch sizes is challenging due to GPU memory constraints, necessitating alternate loss functions that can create a similar geometry with smaller batch sizes (e.g., $B=1$).

Our work addreses these three challenges. First, to **evaluate** long-context retrieval performance, we construct LoCoV1 (Figure 1), a novel benchmark consisting of 12 tasks drawn from law, medicine, science, finance, corporate governance, government reports, and more. LoCoV1 tasks are drawn from real-world datasets spanning diverse domains, including Tau Scrolls, QASPER, LongBench, and the Legal Case Reports corpus Shaham et al. (2022); Dasigi et al. (2021); Bai et al. (2023); Galgani (2012). Unlike previous benchmarks, performance on LoCoV1 requires long-context reasoning, and naive truncation and chunking baselines perform poorly (Table 13 in the Appendix).

Next, we present the **M2-BERT retrieval encoder**, an 80M parameter long-context retriever based on the Monarch Mixer architecture Fu et al. (2023a) and capable of processing up to 32K-length sequences, generating embeddings substantially faster than Transformer-based encoders. To **pretrain** M2-BERT to reason over both short and long contexts, the initial model is pretrained on a mixture of short and long text sequences from C4, Wikipedia, and BookCorpus Raffel et al. (2019); Foundation (2022); Zhu et al. (2015). Building beyond prior pretraining frameworks for M2-BERT, the long-context versions of this model are also warm-started from shorter-context checkpoints to ensure convergence.

To **finetune** M2-BERT for retrieval, we explore two alternative strategies that aim to achieve the same embedding geometry as contrastive loss, but are batch-size independent. First, we explored prototype loss (PL) Li et al. (2021), but found weak performance for downstream retrieval. Instead, we turned to orthogonal projection loss (OPL) Ranasinghe et al. (2021), which allowed more degrees of freedom for aligning the embeddings of query-passage pairs. Furthermore, unlike the common MNRL, OPL optimizes the distance between a query and any relevant/irrelevant document while only requiring a batch size of $B=1$ (Figure 1). This allows for finetuning with single-sample batches that fit in memory.

**Results** Experiments comparing the M2-BERT retrieval encoder to competitive baselines illustrate both performance and efficiency advantages (Figure 1). In a dense retriever setting, the M2-BERT retrieval encoder substantially outperforms models 5x to 90x its size, beating zero-shot E5-Mistral (7.11B) by 22.2 points and fine-tuned BGE-Large (335M) by 30.2 points on average for LoCoV1 Wang et al. (2023); Xiao et al. (2023). M2-BERT also outperforms other retrieval approaches, such as ColBERTv2 San-

| Dataset | Model | Max. Seq. Length | Score | $\Delta$ vs. SOTA |
|---|---|---|---|---|
| BEIR | E5-Mistral | 4096 | 56.9 | 0.0 |
| | OpenAI Ada | 8192 | 53.3 | -3.6 |
| | BGE-Large | 512 | 54.3 | -2.6 |
| LoCo | E5-Mistral | 4096 | 71.4 | -19.7 |
| | OpenAI Ada | 2048* | 63.4 | -27.7 |
| | BGE-Large | 512 | 54.8 | -36.3 |

Table 1: BEIR vs. LoCoV1 on Truncation-Based Approaches (we truncate Ada embeddings at 2048 tokens since it is scores higher than truncating at the 8192 max length).

thanam et al. (2022), a retrieval model that trades off additional compute at inference time for higher quality, and BM25 Jones et al. (2000), a bag-of-words retrieval function that scales easily to longer contexts. With only 80 million trainable parameters, M2-BERT beats several popular API services, such as OpenAI's *text-embedding-ada-002*, Voyager's *voyage-01*, and Cohere's *embed-english-v3.0* by 31.3 points, averaged across the LoCoV1 datasets. Depending on the number of tokens to embed, M2-BERT is also 3 to 676$\times$ more efficient at embedding generation than the next state-of-the-art Transformer-based model (E5-Mistral) while also being pretrained on substantially less data. We provide model checkpoints for the 128, 2048, 8192, and 32768-maximum sequence length versions of the M2-BERT retrieval encoder.

Overall, our work makes the following contributions: (1) the long-context (LoCoV1) retrieval benchmark for evaluating and comparing approaches to long-context retrieval, (2) the M2-BERT retrieval encoder, a state-of-the-art retriever and the first retriever utilizing a state-space architecture, (3) a pretraining and fine-tuning framework for training new M2-BERT retrieval encoders, and (4) an experimental study of the M2-BERT retrieval encoder that illustrates its strengths and weaknesses on long-context tasks. [1]

## 2 LoCoV1 Retrieval Benchmark

We first motivate the need for retrieval benchmarks which *require* long-context reasoning. We find that on existing benchmark datasets, context length does not correlate with performance, and short-context models yield near state-of-the-art performance. We then describe LoCoV1, which consists of retrieval tasks with long documents. We empirically illustrate that on LoCoV1, performance is correlated with context length, suggesting that LoCoV1 better measures long-context retrieval abilities.

**Existing Benchmarks** We explore whether existing retrieval benchmark datasets adequately capture regimes in which long-context reasoning is *essential* for high performance. We examine BEIR Thakur et al. (2021) within the MTEB leaderboard Muennighoff et al. (2022), a popular retrieval benchmark consisting of 17 tasks spanning different domains, query formats, document formats, and query-to-document ratios. In Table 1, we compare performance for three high-scoring models with different sequence lengths (using truncation): *E5-Mistral* (4096 tokens), *OpenAI Ada* (8192), and *BGE-Large* (512). First, we observe that the best performing retrieval model, *E5-Mistral*, is only 2.6 accuracy points, on average, ahead of *BGE-Large-en-v1.5*, despite handling 8$\times$ longer input sequence length (e.g. 4096 vs. 512). Second, we observe that for most BEIR tasks, the longest documents are only several thousand tokens (Figure 6). Qualitatively, we note that many BEIR examples have overlap between the query and the beginning of the document (Table 14 in the Appendix). Overall, these findings suggest that existing benchmark tasks do not effectively capture real-world scenarios where long context retrieval is essential for the downstream ML pipeline (e.g. long context documentation in medicine, law, finance, and more).

**LoCoV1** Through the LoCoV1 benchmark, we present a new set of naturalistic, domain-specific retrieval tasks that reflect real-world use cases for long-context queries and documents. LoCoV1 draws from several existing long-context benchmarks, including Tau Scrolls Shaham et al. (2022), LongBench Bai et al. (2023), and QASPER Dasigi et al. (2021), as well as several domain-specific datasets not originally intended for retrieval, like CourtListener, the Australian Legal Court Reports dataset Galgani (2012), and the StackOverflow forum. (details about each task can be found in Table 10 in the Appendix). Each dataset was selected for both **a)** the longer, more complex formatting of its queries and documents as well as **b)** its ability to gauge long-context handling by containing relevant

[1]The M2-BERT code and LoCoV1 datasets are publically available on Github and HuggingFace, respectively.

| **Model** | BGE-Large Zeroshot | BGE-Large Finetuned | E5-Mistral | BM25 | Jina Embeds. | OpenAI Ada | ColBERTv2 | M2-BERT 128 | M2-BERT 2k | M2-BERT 8k | M2-BERT 32k |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Param. Count** | 335M | 335M | 7.11B | N/A | 137M | N/A | 110M | 80M | 80M | 80M | 80M |
| **Max. Seq. Length** | 512 | 512 | 4096 | N/A | 8192 | 8192 | 512 | 128 | 2048 | 8192 | 32768 |
| **LoCoV1 Score** | 56.9 | 65.0 | 73.0 | 81.5 | 67.2 | 63.9 | 54.3 | 69.7 | 81.6 | 88.9 | **95.2** |
| **LoCoV1 Score w. Chunks** | 54.8 | 61.6 | 70.3 | N/A | 19.2 | 63.2 | N/A | N/A | N/A | N/A | N/A |

Table 2: M2-BERT Retrieval Encoder and Baseline Model Performances on LoCoV1.

information throughout its queries and documents. Violin plots depicting document lengths for each of the LoCoV1 tasks can be found in Figure 5.

In Table 1, we provide results for the same three encoders on LoCoV1. In contrast to BEIR, we find that the relative performance of each model correlates with its sequence length. Additional experiments on LoCoV1 are described in 4.

## 3 M2-BERT Retrieval Encoder Architecture

Motivated by the need for longer-sequence reasoning on LoCoV1, we describe the architecture for the M2-BERT retrieval encoder. For notational clarity, we let $\mathbf{S}$ denote maximum sequence length (for information on pretraining and fine-tuning the M2-BERT retrieval encoder, see Sections B.2 and B.3).

### 3.1 Architecture

The M2-BERT retrieval encoder relies on the Monarch Mixer (M2) architecture, a BERT-like model that utilizes Monarch matrices for language modeling. Monarch Mixer is part of a new class of architectures called *state-space models* (SSMs), which include S4, Mamba, and BiGS (Gu et al., 2022; Gu & Dao, 2023; Wang et al., 2022). Unlike regular BERT and long-context Transformer-based encoder like LongFormer (Beltagy et al., 2020), M2-BERT can handle longer input contexts by leveraging Monarch matrices as a subquadratic primitive along both input sequence length and model dimension. While new Transformer-based models capable of encoding 8k tokens have emerged (Günther et al., 2023), the M2-BERT encoders can handle up to 32k input tokens, undergo fine-tuning substantially faster than attention-based models, run inference 3 to 676x more rapidly (Table 5), and still achieve state-of-the-art on long context retrieval tasks.

## 4 Experiments

We begin by evaluating the M2-BERT retriever's performance relative to existing competitive retriever methods. We choose five of the best performing models from BEIR. These are: BGE-Large-en-v1.5 (Xiao et al., 2023), E5-Mistral (Wang et al., 2023), Jina Embeddings (Günther et al., 2023), OpenAI Ada embeddings (*text-embedding-ada-002*), and ColBERTv2 (Santhanam et al., 2022). The Appendix reports additional models that we evaluated but that have worse performance.

The baseline models have maximum sequence lengths shorter than some documents in LoCoV1. We therefore study two approaches for generating embeddings. The first approach truncates each document to the length of the model's maximum sequence length, while the second approach segments the document into chunks (each the size of the model's maximum sequence length) and computes a document embedding as the average of chunk embeddings. All M2-BERT models are evaluated with the LoCoV1 and BEIR retrieval benchmarks.

We use nDCG@10 (Wang et al., 2013) as the quality metric for LoCoV1. nDCG@10 measures the ranking quality of information retrieval systems, accounting for both the position and quality of the items in the retrieved sequence. We evaluate efficiency by calculating the time it takes to embed 32k document tokens, on average, whether that is through one single embedding or multiple chunked embeddings. Appendix B.15 provides additional information.

**LoCoV1** The baseline models have maximum sequence lengths shorter than some documents in LoCoV1. We therefore study two approaches for generating embeddings. The first approach truncates each document to the length of the model's maximum sequence length, while the second approach segments the document into chunks (each the size of the model's maximum sequence length) and computes a document embedding as the average of chunk embeddings. All M2-BERT models are evaluated with the LoCoV1 and BEIR retrieval benchmarks.

| Model | Max. Seq. Length | Param. Count | BEIR Score | Δ Params | Δ BEIR Score |
|---|---|---|---|---|---|
| SentenceBERT | 512 | 110M | 40.0 | 0% | 0 |
| M2-BERT-128 | 128 | 80M | 38.7 | -27% | -1.3 |

Table 3: M2-BERT vs. SentenceBERT on BEIR.

We use nDCG@10 Wang et al. (2013) as the quality metric for LoCoV1. nDCG@10 measures the ranking quality of information retrieval systems, accounting for both the position and quality of the items in the retrieved sequence. We evaluate efficiency by calculating the time it takes to embed 32k document tokens, on average, whether that is through one single embedding or multiple chunked embeddings. Appendix B.15 provides additional information.

Table 2 compares averaged nDCG scores for all methods on the LoCoV1 benchmark (Table 13 in the Appendix provides results by task). Performance improvements are significant — we found that M2-BERT-32k outperformed the next best baseline approach (*BM25*) by an average of 13.7 points, the next best truncation-baseline approach (*E5-Mistral*) by an average of 22.2 points, and the next best chunked-baseline approach (*E5-Mistral*) by an average of 24.9 points. On a per-task level, M2-BERT-32k outperforms all baseline methods on 7 of 12 tasks, and all Transformer-based methods on 9 of 12 tasks.

We also observe that retrieval accuracy increased as we incrementally scaled maximum sequence length of the M2-BERT retrieval encoder for each of our models. The overall performance improvement for going from a sequence length of 128 tokens to 32k tokens is approximately 35.5 points (average). In contrast, alternate retrieval strategies—like chunking—appeared to barely improve other base retrieval models, and sometimes even worsen them. Overall, our findings demonstrate that standard retrieval approaches, whether it is truncation or chunking with embedding averaging, are not sufficient for handling long-context documents in retrieval, and that M2-BERT outperforms baseline models while being substantially smaller.

**BEIR** By testing the M2-BERT architecture on the BEIR benchmark, we study **1)** whether M2-BERT retrieval encoders can match Transformer-based models when pretrained and fine-tuned on similar data and **2)** whether M2-BERT retrieval encoders sacrifice short-context performance when fine-tuned on longer contexts.

For question #1, we compare to SentenceBERT, a language model of comparable size with a similar pretraining ensemble and identical fine-tuning process for BEIR (e.g. fine-tuning on the MS MARCO retrieval dataset Bajaj et al. (2018)). Holding the data constant, we fine-tune a separate pretrained checkpoint of the M2-BERT-128 model using the same fine-tuning process and the same number of MSMARCO examples as SentenceBERT. We find that our M2-BERT-128 checkpoint approximately matches SentenceBERT performance (Table 4), averaging 1.3 nDCG@10 points lower, and performs better than SentenceBERT on some of the longer context classification datasets (e.g. AmazonPolarityClassification and AmazonReviewsClassification).

For question #2, we jointly fine-tune a separate pretrained checkpoint of the M2-BERT-128 model on MSMARCO and LoCoV1, using the same number of training examples for both datasets (e.g. 500K examples) to study whether a single model can generalize to both short and long sequences. We find that our M2-BERT-128 checkpoint performs worse on BEIR and LoCoV1 than checkpoints solely fine-tuned on each dataset individually; BEIR performance drops by 7.2 nDCG@10 points, on average, and LoCoV1 performance drops by 5.1 nDCG@10 points, on average. While M2-BERT-128 can handle both short queries and long documents for LoCoV1, our results for configuration #2 suggests there is likely negative transfer between the longer context LoCoV1 datasets and shorter context BEIR datasets. Future work should explore how to balance both short and long context handling in the next generation of retrieval encoders.

## IMPACT STATEMENT

This paper seeks to improve the robustness and utility of existing ML pipelines, particularly those with retrieval-based components. During document indexing, we imagine the M2-BERT retrieval encoder will significantly increase the maximum passage length used for chunking. We hope the extended passage length will improve the utility of retrieval-based applications for language models by improving the quality of the retrieved context, allowing researchers and practitioners to improve language model generation quality for a variety of tasks: question-answering, fact verification, dialogue generation, and more.

## REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*, 2017.

Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, and Christopher Ré. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pp. 3090–3122. PMLR, 2022.

Together Computer. Redpajama: an open dataset for training large language models, October 2023. URL `https://github.com/togethercomputer/RedPajama-Data`.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*, 2021.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations (ICLR)*, 2019.

Wikimedia Foundation. Wikimedia downloads, 2022. URL `https://dumps.wikimedia.org`.

Daniel Y. Fu, Mayee F. Chen, Michael Zhang, Kayvon Fatahalian, and Christopher Ré. The details matter: Preventing class collapse in supervised contrastive learning. 2022.

Daniel Y Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. Monarch mixer: A simple sub-quadratic gemm-based architecture. In *Advances in Neural Information Processing Systems*, 2023a.

Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry Hungry Hippos: Towards language modeling with state space models. In *International Conference on Learning Representations*, 2023b.

Filippo Galgani. Legal Case Reports. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5ZS41.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces, 2022.

Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang, and Han Xiao. Jina embeddings: A novel set of high-performance sentence embedding models, 2023.

Ramin Hasani, Mathias Lechner, Tsun-Huang Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models. *arXiv preprint arXiv:2209.12951*, 2022.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply, 2017.

K Sparck Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840, 2000.

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 *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://www.aclweb.org/anthology/2020.emnlp-main.550.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Carlos Lassance and Stéphane Clinchant. An efficiency study for splade models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 2220–2226, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531833. URL https://doi.org/10.1145/3477495.3531833.

Megan Leszczynski, Daniel Y. Fu, Mayee F. Chen, and Christopher Ré. Tabi: Type-aware bi-encoders for open-domain entity retrieval, 2022.

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, 2021.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations, 2021.

Zehua Li, Neel Guha, and Julian Nyarko. Don't use a cannon to kill a fly: An efficient cascading pipeline for long documents. 2023.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search, 2022.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi: 10.48550/ARXIV.2210.07316. URL https://arxiv.org/abs/2210.07316.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. 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. 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*, 2023.

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. *arXiv e-prints*, 2019.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:201646309.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Sultan, and Christopher Potts. UDAPDR: Unsupervised domain adaptation via LLM prompting and distillation of rerankers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11265–11279, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.693. URL https://aclanthology.org/2023.emnlp-main.693.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. 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. 3715–3734, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. Scrolls: Standardized comparison over long language sequences, 2022.

Jimmy TH Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Ellen M Voorhees, Donna K Harman, et al. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

Junxiong Wang, Jing Nathan Yan, Albert Gu, and Alexander M Rush. Pretraining without attention. *arXiv preprint arXiv:2212.10544*, 2022.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere, 2022.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of ndcg type ranking measures, 2013.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. Retrieval meets long context large language models, 2023.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to augment large language models, 2023.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

# A APPENDIX

# B ACKNOWLEDGEMENTS

We thank Silas Alberti, Sabri Eyuboglu, Chris Fifty, Omar Khattab, Gautam Machiraju, Eric Nguyen, Krista Opsahl-Ong, Christopher Potts, Benjamin Spector, Alyssa Unell, Benjamin Viggiano, Michael Wornow, and Michael Zhang for their constructive feedback during the composition of the paper. We would also like to thank our collaborators at the Stanford Artificial Intelligence Laboratory (SAIL) and TogetherAI.

We gratefully acknowledge the support of NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); US DEVCOM ARL under No. W911NF-21-2-0251 (Interactive Human-AI Teaming); ONR under No. N000141712266 (Unifying Weak Supervision); ONR N00014-20-1-2480: Understanding and Applying Non-Euclidean Geometry in Machine Learning; N000142012275 (NEPTUNE); NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, Google Cloud, Salesforce, Total, the HAI-GCP Cloud Credits for Research program, the Stanford Data Science Initiative (SDSI), Stanford EDGE Fellowship, GEM Fellowship Program, and members of the Stanford DAWN project: Facebook, Google, and VMWare. Neel Guha is supported by the Stanford Interdisciplinary Graduate Fellowship and the HAI Graduate Fellowship. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of NIH, ONR, or the U.S. Government.

## B.1 RELATED WORK

We overview existing retrieval benchmarks and contrast them with LoCoV1. We also describe existing state-of-the-art approaches for retrieval models and compare them to M2-BERT.

**Retrieval Benchmarks** There are a variety of existing retrieval benchmarks for guiding embedding development, such as BEIR, TREC, NaturalQuestions (NQ), SQuAD, and LoTTE (Thakur et al., 2021; Voorhees et al., 2005; Kwiatkowski et al., 2019; Rajpurkar et al., 2018; Santhanam et al., 2022) While these datasets cover a wide breadth of domains, none of them reliably gauge long-context handling during retrieval. The Tau Scrolls datasets (Shaham et al., 2022) seek to gauge long-context handling in language models but it focuses on other knowledge-intensive tasks, such as summarization, fact verification, and natural language inference.

We explore whether existing retrieval benchmark datasets adequately capture regimes in which long-context reasoning is *essential* for high performance. We examine BEIR (Thakur et al., 2021) within the MTEB leaderboard (Muennighoff et al., 2022), a popular retrieval benchmark consisting of 17 tasks spanning different domains, query formats, document formats, and query-to-document ratios. In Table 1, we compare performance for three high-scoring models with different sequence lengths (using trunction): *E5-Mistral* (4096 tokens), *OpenAI Ada* (8192), and *BGE-Large* (512). First, we observe that the best performing retrieval model, *E5-Mistral*, is only 2.6 accuracy points, on average, ahead of *BGE-Large-en-v1.5*, despite handling $8\times$ longer input sequence length (e.g. 4096 vs. 512). Second, we observe that for most BEIR tasks, the longest documents are only several thousand tokens (Figure 6). Qualitatively, we note that many BEIR examples have overlap between the query and the beginning of the document (Table 14). Overall, these findings suggest that existing benchmark tasks do not effectively capture real-world scenarios where long context retrieval is essential for the downstream ML pipeline (e.g. long context documentation in medicine, law, finance, and more).

With the Long-Context (LoCo) Benchmark (V1), we seek to accurately gauge long-context handling in retrieval encoders. We selected datasets for which increases to a model's maximum input context will substantially improve retrieval accuracy.

**Embedding Models for Retrieval** Embedding models are frequently utilized in machine learning pipelines during retrieval. Many neural embedding models utilize an encoder-only Transformer architecture (Vaswani et al., 2023) that is fine-tuned to maximize cosine similarity between queries and their relevant passages (Reimers & Gurevych, 2019; Leszczynski et al., 2022; Chen et al., 2022). Alternative neural retrieval approaches have emerged to further boost retrieval accuracy while minimizing growing training time, inference time, and memory utilization: examples include dense passage retrieval (DPR) (Karpukhin et al., 2020), late-interaction techniques with ColBERTv2 (Santhanam et al., 2022), and sparse lexical representations with SPLADEv2 (Lassance & Clinchant,

| Length Type | C4 | Wikipedia | BookCorpus |
|---|---|---|---|
| Variable | 10% | 10% | 10% |
| Maximum | 24% | 23% | 23% |

Table 4: Pretraining dataset proportions based on text source and sequence length type.

2022). However, new embeddings models based on the generative pretrained transformer (GPT) architecture, such as SGPT, BGE, and E5-Mistral, have reached state-of-the-art accuracy on the BEIR retrieval benchmark (Thakur et al., 2021), leading to higher quality embedding representations that increase domain generalization (Muennighoff, 2022; Zhang et al., 2023; Wang et al., 2023).

In ML pipelines, researchers and practitioners have sought to avoid longer contexts by simply chunking the passages into smaller inputs and averaging the embeddings (Lewis et al., 2021). However, for long-context benchmarks, we found that the M2-BERT retrieval encoder outperforms existing models, both when they truncate the input context and when they employ chunking strategies (Table 10 and Table 13). This finding suggests that there is indeed a benefit to being able to retrieve over full documents, rather than employing chunking strategies.

### B.2 PRETRAINING

Retrieval encoders frequently rely on model backbones which have already been pretrained on corpora from the relevant language (Reimers & Gurevych, 2019; Santhanam et al., 2022; Karpukhin et al., 2020; Lassance & Clinchant, 2022). This equips the model with the capacity to understand and reason over text sequences, and enables high performance even when the retrieval-specific finetuning dataset is small (Saad-Falcon et al., 2023). The difficulty with using the M2 architecture for our encoder is that previous work has only (1) studied pretraining M2 for sequence lengths up to 128 tokens, and (2) studied pretraining in regimes where downstream tasks consisted of sequences mostly uniform in length (e.g., short GLUE tasks). In contrast, the long-context retrieval setting requires that the base model be capable of understanding both short sequences (for queries) and long sequences (for documents).

The first technical challenge is designing a pretraining dataset over which the masked language modeling (MLM) objective enables the model to learn both short and long sequences. Experimentally, we find that training with only short or only long sequences is insufficient, and that instead the pretraining data must contain a mixture of both short and long context samples (see B.4.1 for comparisons to alternative strategies). For the source of these samples, we rely on three high quality datasets routinely used for pretraining: C4 (Raffel et al., 2019), Wikipedia (Foundation, 2022), and BookCorpus (Zhu et al., 2015). For our short context examples, we include variable length passages from our three training corpora, which can range from 10 tokens to our maximum input sequence length of 128, 2048, 8192, or 32768 tokens, depending on the M2-BERT model. For our long context examples, we concatenate multiple successive training examples together to generate sequences that reach our maximum input sequence length.

The second technical challenge is ensuring pretraining convergence when the maximum sequence length is greater. We find that traditional initialization with random weights is sufficient when $\mathbf{S} \in \{128, 2\text{k}, 8\text{k}\}$. For $\mathbf{S} = 32\text{k}$ however, we find that models initialized with random weights do not converge to sufficient MLM accuracies within a reasonable amount of time. Therefore to accelerate training convergence for this model, we warm start with the weights of a pretrained 8k checkpoint, and initialize the 32k positional embeddings with the initial 8k positional embeddings by extending them through replication. Under this strategy, the 32K model converges.

### B.3 FINE-TUNING

To adapt a pretrained model for a specific retrieval task (e.g., identifying the relevant legal case given a description), it is common practice to finetune that model on a collection of representative queries and documents (Reimers & Gurevych, 2019; Santhanam et al., 2022; Karpukhin et al., 2020; Saad-Falcon et al., 2023).

**MNRL** A popular approach is to finetune the base model using a contrastive learning loss called multiple negatives ranking loss (Henderson et al., 2017), which encourages the model to learn embeddings of queries and documents for which the cosine similarity of relevant query-document pairs is high, and irrelevant query-document pairs is low. It requires a dataset of query and relevant document pairs $(\{(q_i, d_i)\}_{i=1}^{n}$. For a query $q_i$, MNRL samples $k$ random documents from $\{d_j\}_{j=1, j \neq i}^{n}$ as "negative" passages, and generates a "prediction" for $q_i$ against $d_i$ and the $k$ distractors by computing pairwise

cosine similarities (e.g. *PCS*). CrossEntropyLoss (e.g. *CE*) is applied to these predictions, treating the $k$ distractors as the negative class and $d_i$ as the positive class. For a given query $q_k$, we compute MNRL as:

$$\begin{aligned} \textit{MNRL}(\{q_k,d_i\}_{i=1}^n) &= \textit{CE}(\text{Scores},\text{Labels}) \\ \text{Scores} &= [\textit{PCS}(q_k,d_i)_{i=1}^n] \\ \text{Labels} &= [1,...,n] \end{aligned}$$

MNRL is closely related with constrastive loss, and induces an embedding geometry of *alignment* between query-document pairs, and *uniformity* of document embeddings across the hypersphere (Wang & Isola, 2022; Chen et al., 2022; Leszczynski et al., 2022; Fu et al., 2022). This loss function requires large batch sizes for quality.

In MNRL, a single query and all $k+1$ documents must fit within a single batch. In the long-context regime, GPU memory requirements thus force a tradeoff between $k$ and $\mathbf{S}$. When $\mathbf{S}$ is small (e.g., 128 tokens), $k$ can be large and still fit in GPU memory (e.g., $k=128$). When $\mathbf{S}$ is large however (e.g., 32k tokens), the memory footprint of a single document is larger, and $k$ must be considerably smaller (e.g., $k=2$). The technical challenge is that MNRL only works well for large $k$ (Henderson et al., 2017), and thus, suboptimal for long-sequences (see Appendix B.4.1).

**Prototype Loss** In our work, we seek a method to achieve the same embedding geometry as MNRL, but in a batch-independent way. One approach is *prototype loss* (Li et al., 2021), which uses a target model's embeddings to guide the contrastive learning of a student model. By leveraging the learned embeddings of a model trained with MNRL (e.g. M2-BERT-128), we may be able to rapidly fine-tune a long-context embedding model that is limited to a much smaller batch size (e.g. M2-BERT-32k). Given query $q_k$, passage $p_k$, target embedding model *TM*, and student embedding model *SM*, we calculate prototype loss (PL) as:

$$\begin{aligned} PL(\{q_k,p_k\}) &= \text{Query Loss} + \text{Passage Loss} \\ \text{Query Loss} &= \textit{PCS}(\textit{TM}(q_k),\textit{SM}(q_k)) \\ \text{Passage Loss} &= \textit{PCS}(\textit{TM}(p_k),\textit{SM}(p_k)) \end{aligned}$$

Even if our M2-BERT-32k model successfully learns the embeddings from the M2-BERT-128 model, it still requires further fine-tuning from the starting M2-BERT-128 representations to develop robust embeddings for 32k context length. After using prototype loss to fine-tune our M2-BERT-32k with the fine-tuned M2-BERT-128 model as the target embeddings (Table 8), we find that the M2-BERT-128 embeddings are not the ideal starting weights for further fine-tuning of M2-BERT-32k; the learned representations at 128 context length are substantially different than the learned representation at 32k context length (Appendix B.7).

**Orthogonal Projection Loss** To overcome these challenges, we instead finetune our M2-BERT base model using *orthogonal projection loss* (OPL) (Ranasinghe et al., 2021). Unlike MNRL, OPL is compatible with single-sample batches by using Mean Squared Error (e.g. *MSE*). Unlike prototype loss, OPL does not require a teacher model for embeddings. Given a query $q_k$ and passage $p_k$, we calculate OPL as:

$$\begin{aligned} OPL(\{q_k,p_k\}) &= \textit{MSE}(\text{Score},\text{Label}) \\ \text{Score} &= \textit{PCS}(q_k,p_k) \\ \text{Label} &= 1.0 \text{ for positives, } 0.0 \text{ for negatives} \end{aligned}$$

Intuitively, OPL finetunes the model to encourage embeddings for positive query-document pairs to be aligned with each other, and for negative query-document pairs to be orthogonal to each other. Because OPL operates on a single query-document pair, it performs well on single-sample batches, and is thus ideal for our long-context setting. Similar to MNRL, we sample negative documents for query $q_i$ from $\{d_j\}_{j=1,j\neq i}^n$. Lastly, we note that while OPL proves effective for fine-tuning our M2-BERT encoder (Appendix B.4.1), OPL is just one choice of loss function; other functions with similar properties may be useful.

### B.4 Additional Experiments

Our experimental evaluations focus on three questions: (1) How efficient are the M2-BERT encoders at embedding generation compared to standard retrieval encoders? (2) How necessary are the pretraining and finetuning approaches proposed in section 3, and how do they compare to standard retriever

pretraining/finetuning methods? (3) Can the representations learned by the fine-tuned M2-BERT models be used for non-retrieval tasks, like data visualization or clustering-based classification?

**Computational Efficiency** We compare M2-BERT to baseline methods in terms of throughput, i.e., the time it takes to both tokenize and embed the entirety of an $X$ token document (Table 5) on A100. For models that cannot tokenize and embed the $X$ tokens all at once, we create separate embeddings for $Y$ token chunks, where $Y$ is the maximum sequence length of the model. We find that M2-BERT-32K provides the greatest throughput, producing an embedding $3.13\times$ more efficiently for a 512 token document and $676\times$ more efficiently for a 32768 token document relative to the next best state-of-the-art model, *E5-Mistral*.

| | | **Time to Encode $X$ Tokens** | | | |
|---|---|---|---|---|---|
| **Models** | **Max. Seq. Length** | **128** | **2048** | **8192** | **32768** |
| BGE-Large | 512 | 0.015 | 0.029 | 0.12 | 0.49 |
| E5-Mistral | 4096 | 0.029 | 0.11 | 1.2 | 4.8 |
| Jina Embeds. | 8192 | 0.0070 | 0.0070 | 0.0070 | 0.028 |
| M2-BERT-128 | 128 | 0.028 | 0.057 | 0.12 | 0.46 |
| M2-BERT-2k | 2048 | 0.0071 | 0.0071 | 0.028 | 0.057 |
| M2-BERT-8k | 8192 | 0.0072 | 0.0072 | 0.0072 | 0.028 |
| M2-BERT-32k | 32768 | 0.0071 | 0.0071 | 0.0071 | 0.0071 |
| **$\Delta$ 32k Speed vs. Mistral** | | 3.13x | 14.9x | 169x | 676x |

Table 5: M2-BERT Efficiency Comparison to Baseline Models.

**Needle-in-the-Haystack Synthetic** We perform a more detailed analysis of the M2-BERT retriever's ability to encode long contexts using a synthetic modeled off "needle-in-the-haystack" tasks that have been used in other studies of longer context tasks (Liu et al., 2023). For our version, we adapted the Natural Questions (NQ) benchmark (Kwiatkowski et al., 2019), which contains contains query-relevant passage pairs (derived from Google and Wikipedia). We use the original queries provided by the NQ benchmark, but modified the passages by adding nine "distractor" passages to each relevant passage in the answer set. These distractor passages are selected by randomly sampling other Wikipedia passages. We study how the location of the relevant passage (e.g., appearing first vs. appearing seventh amongst all the passages) impacted retrieval performance. Since we have 10 passages total, there are exactly 10 different positions to place the relevant passage within the sequence.

We compare M2-BERT-32k to the best-performing baselines: *BGE-Large-en-v1.5*, *E5-Mistral* and Jina Embeddings (Figure 2). We observe a relationship between the position of the relevant passage and the relative performance improvement of M2-BERT-32k. When the relevant passage is closer to the start of the concatenated sequence, the baseline models perform almost as well as M2-BERT-32k. However, as the relevant passage moves to the end of the concatenated sequence, the performances of the baseline models substantially drops since the models cannot see the relevant passage within the total sequence, due to their shorter maximum sequence lengths (see Table 15 for complete results).

| **Model** | **Max. Seq. Length** | **Training Selection** | **LoCoV1 Score** |
|---|---|---|---|
| M2-BERT | 2048 | Short Examples | 37.2 |
| M2-BERT | 2048 | Long Examples | 44.9 |
| M2-BERT | 2048 | Mixed Examples | **55.4** |

Table 6: M2-BERT Training Example Selection for Pretraining.

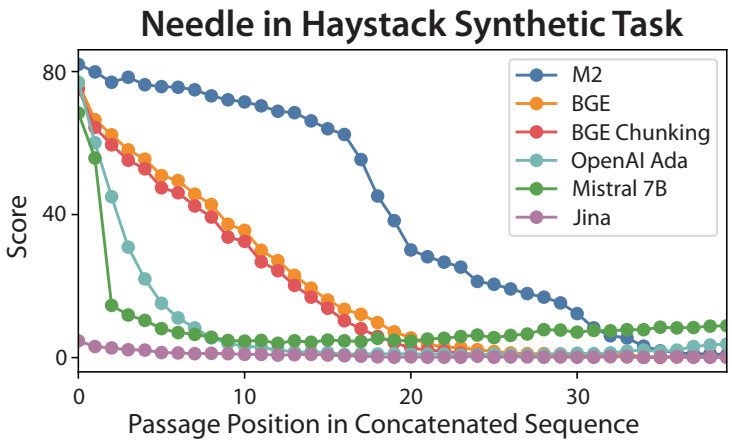


Figure 2: M2-BERT and Baseline Model Performance on Needle-in-the-Haystack Synthetic Task.

| Model | Max. Seq. Length | Checkpoint Selection | MLM Accuracy |
|---|---|---|---|
| M2-BERT-32k | 32768 | Warm-Start | **33.9** |
| M2-BERT-32k | 32768 | Cold-Start | 4.8 |

Table 7: Warm vs. Cold Start for M2-BERT-32768 Pretraining - MLM Train Accuracy after 6,000 Training Steps.

#### B.4.1 Ablation of Pretraining and Finetuning

Section 3 presents two design choices for the M2-BERT retriever—pretraining data mixture and finetuning loss objective. This subsection evaluates those choices in comparison to alternative pretraining and finetuning approaches.

**Pretraining** In Section 3, we describe selecting a pretraining mixture for the M2-BERT base model consisting of both short and long sequences. We compare this to two alternate pretraining regimes: (1) solely using short training examples, and (2) solely using long training examples (Table 6). For each regime, we pretrain the M2-BERT-2048 architecture to 5,000 training steps before further fine-tuning on the LoCoV1 dataset but with a limited number of negatives (e.g. 8 negative passages per query-positive passage pair). We observe that the model trained on the mixed short/long sequence dataset performs best, beating solely long sequence pretraining by 10.5 points on average.

We also illustrate the necessity of initializing M2-BERT-32k with the weights of a M2-BERT-8k checkpoint (Table 7 and Figure 4). Compared to random initialization, we find that the version with warm-starting converges dramatically faster, successfully completing pretraining in the same number of steps as our other M2-BERT encoders.

**Finetuning** Section 3 describes how GPU memory constraints limit the training batch size for longer context M2-BERTs, necessitating the use of OPL loss function, which can function with single-sample batch sizes. We illustrate the batch size-performance tradeoff incurred by the traditionally used MNRL loss function by comparing (1) OPL trained with batch size 1, to (2) MNRL trained with the maximum batch size possible on an A100 GPU (Table 8). For fine-tuning M2-BERT-32k, we find that OPL improved average nDCG@10 on LoCoV1 by 29.4% compared to MNRL.

#### B.4.2 Applications of M2-BERT Retrieval Encoders

Finally, we explore whether the embeddings from the M2-BERT retrieval model are useful for other embedding tasks.

**Zero-shot Clustering** We find our M2-BERT retrieval encoders can be used effectively for zero-shot clustering of textual datasets. Using our M2-BERT-32k model, we take a sample of the RedPajama-v1 dataset (Computer, 2023) and generate embeddings for datapoints from each of the constituent datasets:

| Model | Loss Function | Batch Size | LoCoV1 Score | Δ Scores |
|---|---|---|---|---|
| M2-BERT-32k | MNRL | 2 | 70.4 | 0 |
| M2-BERT-32k | PL | 2 | 63.2 | -7.2 |
| M2-BERT-32k | OPL | 1 | **91.1** | 29.4 |

Table 8: OPL vs. MNRL for Fine-tuning M2-BERT-32k.

| Model | Max. Seq. Length | MTEB Score | Δ Params | Δ MTEB Score |
|---|---|---|---|---|
| SentenceBERT | 512 | **64.5** | 0% | 0 |
| M2-BERT-128 | 128 | 63.4 | -27% | -1.1 |

Table 9: M2-BERT vs. SentenceBERT on MTEB Classification.

C4, StackExchange, BookCorpus, ArXiv, and Github. In Figure 3, we visualize the M2-BERT embeddings for the sampled datapoints from RedPajama. We find that the datapoints for Github and StackExchange tend to be grouped together, likely due to their overlapping subject terminologies. Additionally, we find limited overlap between C4 and BookCorpus due to some shared subjects between the two constituent datasets. Lastly, the ArXiv datapoints seem mostly isolated by its unique mix of technical topics.

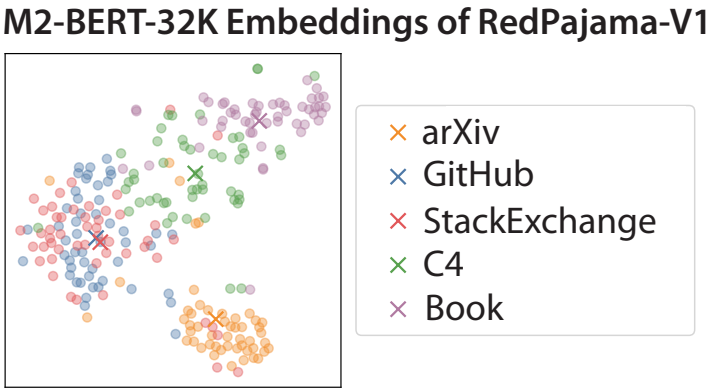


Figure 3: t-SNE Visualization of M2-BERT-32K Embeddings of RedPajama-V1 sample.

**Classification with M2** To further explore the robustness of the M2-BERT embeddings, we test our M2-BERT retrieval encoders on the MTEB classification set. In Table 9, we compare the zero-shot classification results of our M2-BERT-128 retrieval encoder to the SentenceBERT baseline for MTEB classification benchmark, evaluating on only the English datasets (for the expanded results, see Table 16). We found that M2-BERT-128 performed comparably to the SentenceBERT model, scoring only 1.1 accuracy points below SentenceBERT despite substantially less pretraining data and 27% less parameters. We are interested to explore further applications of M2-BERT in both classification and clustering tasks, particularly for long-context tasks.

### B.5 LoCoV1 Overview

Table 10 provides an overview of the LoCo benchmark.

| Dataset | Source | Domain | Query Type | Answer Type | # of Train Queries | # of Train Documents | # of Test Queries | # of Test Documents | Avg. Query Length | Avg. Doc. Length |
|---|---|---|---|---|---|---|---|---|---|---|
| SummScreenFD | Tau Scrolls | Screenwriting | Summary | Dispersed | 3673 | 3673 | 338 | 338 | 590 | 30,792 |
| Gov. Report | Tau Scrolls | Government | Summary | Dispersed | 17457 | 17457 | 972 | 972 | 3,871 | 55,280 |
| QMSUM | Tau Scrolls | Corporate Management | Summary | Dispersed | 1257 | 162 | 272 | 35 | 430 | 58,129 |
| QASPER Title to Full Text | QASPER | Science | Title | Dispersed | 888 | 888 | 416 | 416 | 71 | 22,315 |
| QASPER Abstract to Full Text | QASPER | Science | Abstract | Dispersed | 888 | 888 | 416 | 416 | 931 | 22,315 |
| MultiFieldQA | LongBench | General Domain | Question | Answer Span | 120 | 120 | 30 | 30 | 62 | 29,465 |
| 2WikimQA | LongBench | General Domain | Question | Answer Span | 240 | 240 | 60 | 60 | 69 | 37,867 |
| Passage Retrieval | LongBench | General Domain | Summary | Dispersed | 240 | 240 | 60 | 60 | 840 | 35,814 |
| Court Listener Plain Text | CourtListener | Law | Summary | Dispersed | 10000 | 10000 | 2000 | 2000 | 146 | 48,190 |
| Court Listener HTML | CourtListener | Law | Summary | Dispersed | 10000 | 10000 | 2000 | 2000 | 146 | 57,028 |
| Australian Legal Case Report | Australian Legal Case Report | Law | Summary | Dispersed | 3094 | 3094 | 770 | 770 | 14,986 | 47,536 |
| StackOverflow | StackOverflow | Programming | Question | Dispersed | 1599 | 18005 | 400 | 7741 | 758 | 4,544 |

Table 10: Overview of Long-Context (LoCo) benchmark (V1) and its constituent datasets.

### B.6 LoCoV1 Query and Document Examples

| LoCoV1 Dataset | Query Example | Document Example |
|---|---|---|
| Tau Scrolls Summ ScreenFD | It's the first day of school at Degrassi Community School, and eighth-grader Ashley already has her sights set on becoming the school's newest student council president. Her seemingly sure win is soon threatened when her stepbrother, Toby, becomes frustrated by her unchallenged status and convinces his friend J.T. to run against her. Meanwhile, Emma and Manny deal with eighth-grader Spinner's bullying. Note: This episode marks the first appearances of Sarah Barrable-Tishauer, Lauren Collins, Aubrey Graham, and Shane Kippel | [The Kerwin House - Ashley's Room]<br><br>(While getting ready for school, she's talking to her friend Terri on the phone.)<br><br>Ashley: This is gonna be the best year ever! (Working on her poster for Degrassi<br><br>... |
| Tau Scrolls Gov. Report | Members of Congress and Administrations have periodically considered reorganizing the federal government's trade and development functions to advance various U.S. policy objectives. The Better Utilization of Investments Leading to Development Act of 2018 (BUILD Act), which was signed into law on October 5, 2018 (P.L. 115-254), represents a potentially major overhaul of U.S. development finance efforts... | Background<br>What is the U.S. International Development Finance Corporation (IDFC)? The IDFC is authorized by statute to be a "wholly owned Government corporation ... under the foreign policy guidance of the Secretary of State" in the executive branch. Its purpose is to "mobilize and facilitate the participation of private sector capital and skills in the economic development" of developing and... |
| Tau Scrolls QMSUM | According to the Industrial Design, there might be only a few choices for the energy source and materials from the current manufacturer, so he suggested that they had better look for another manufacturer for more alternatives. The Marketing put forward to design a user friendly interface while the User Interface came up with the idea of including the voice recognition system into the remote control... | Summarize the ideas of the individual presentations.<br>Marketing: {vocalsound} That went well , thank you .<br>Project Manager: That's great .<br>Industrial Designer: {vocalsound} 'Kay .<br>Marketing: Perfect .<br>Project Manager: Alright , let me just PowerPoint this up .<br>{vocalsound} {vocalsound} {vocalsound} Right so um this |
| QASPER Title | Knowledge Authoring and Question Answering with KALM | Introduction: Knowledge representation and reasoning (KRR) is the process of representing the domain knowledge in formal languages (e.g., SPARQL, Prolog) such that it can be used by expert systems to execute querying and reasoning services. KRR have been applied in many fields including financial regulations, medical diagnosis, laws, and so on. One major obstacle in KRR is the creation of large-scale... |
| QASPER Abstract | Knowledge representation and reasoning (KRR) is one of the key areas in artificial intelligence (AI) field. It is intended to represent the world knowledge in formal languages (e.g., Prolog, SPARQL) and then enhance the expert systems to perform querying and inference tasks. Currently, constructing large scale knowledge bases (KBs) with high quality is prohibited by the fact that the construction | Introduction: Knowledge representation and reasoning (KRR) is the process of representing the domain knowledge in formal languages (e.g., SPARQL, Prolog) such that it can be used by expert systems to execute querying and reasoning services. KRR have been applied in many fields including financial regulations, medical diagnosis, laws, and so on. One major obstacle in KRR is the creation of large-scale... |
| MultiField QA | What algorithm is engaged in the PLMS-PPIC method? | \section{Introduction}\label{S1}<br>The multiple access interferences (MAI) is the root of user limitation in CDMA systems \cite{R1,R3}. The parallel least mean square-partial parallel interference cancelation (PLMS-PPIC) method is a multiuser detector for code division multiple access (CDMA) receivers which reduces the effect of MAI in bit detection. In this method and similar to its former version |
| 2WikimQA | Where did the director of film The Brave Bulls (Film) die? | Passage 1: The Brave Archer<br>The Brave Archer, also known as Kungfu Warlord, is a 1977 Hong Kong film adapted from Louis Cha's novel The Legend of the Condor Heroes. The film was produced by the Shaw Brothers Studio and directed by Chang Cheh, starring Alexander Fu Sheng and Tien Niu in the lead roles. The film is the first part of a trilogy and was followed by The Brave Archer 2 (1978) and... |
| Passage Retrieval | During World War II, navy nurses played a crucial role in providing medical care and preventing further casualties. They were present during the initial Japanese attack on Pearl Harbor, as well as in Kaneohe Bay, the Philippines, Guam, and aboard the Solace. The nursing profession was recognized for its essential contribution and was placed under the War Manpower Commission. Despite shortages, ... | Paragraph 1: Thermometric titrimetry Thermometric titrimetry is an extraordinarily versatile technique. This is differentiated from calorimetric titrimetry by the fact that the heat of the reaction (as indicated by temperature rise or fall) is not used to determine the amount of analyte in the sample solution. Instead, the equivalence point is determined by the rate of temperature change. Because ... |
| Court Listener (HTML) | noting that "[a]s a court of limited jurisdiction, we begin, and end, with an examination of our jurisdiction" | <citances>[c]Sellar v Lasotav Pty Ltd: In the matter of Lasotav Pty Ltd [2008] FCA 1612 (27 October 2008)</citances><br></citances> Home Databases WorldLII Search Feedback<br>Federal Court of Australia<br>You are here: AustLII Databases Federal Court of Australia<br>2008 2008 FCA 1612 </citances> |
| Court Listener (Plain Text) | noting that "[a]s a court of limited jurisdiction, we begin, and end, with an examination of our jurisdiction" | [c]Sellar v Lasotav Pty Ltd: In the matter of Lasotav Pty Ltd [2008] FCA 1612 (27 October 2008)<br>Home Databases WorldLII Search Feedback Federal Court of Australia<br>You are here: AustLII Databases<br>Federal Court of Australia 2008 2008 FCA 1612 |
| Legal Case Reports | <citphrase id="cp0.0">cited from="[2006] FCA 1222">corporations law</citphrase><br><citphrase id="cp0.1">cited from="[2006] FCA 1222">whether funds held pursuant to terminated deed of company arrangement are held for the benefit of deed creditors or property of the company in liquidation</citphrase><br><citphrase id="cp0.2">cited from="[2006] FCA 1222">direction that the funds be administered a | On 14 November 2008, Ms Swee Yen Tay instituted a proceeding in this Court against the Migration Review Tribunal ("the Tribunal") and the Minister for Immigration and Citizenship.</sentence><br><br><sentence id="s1">The claim made in the proceeding is said by the applicant to be within the original jurisdiction of the Court, "being an application for a declaration as to the proper construction of s 494C |
| Stack Overflow | Multithreading Design Best Practice — Consider this problem: I have a program which should fetch (let's say) 100 records from a database, and then for each one it should get updated information from a web service. There are two ways to introduce parallelism in this scenario: | You could use an Observer pattern. A simple functional way to accomplish this: "'<br><php Plugin system listeners = array(); Create an entry point for plugins |

Table 11: LoCoV1 Examples for each Dataset

### B.7 M2-BERT PRETRAINING, FINE-TUNING, AND EVALUATION DETAILS

For pretraining the M2-BERT encoders, we use the C4, Wikipedia, and Bookcorpus datasets for training examples. For our dataset split, we sample each dataset equally (e.g. 33% each). For our example length ratio, we selected 0.3 variable length examples (e.g. short examples) and 0.7 maximum concatenated examples (e.g. long examples). We utilize the masked-language modeling (MLM) pretraining objective with an MLM probability of 0.3 to prepare the encoders for downstream language modeling. For training evaluation, we use the C4 validation set with an MLM probability of 0.15. For our scheduler, we use linear decay with warmup, where warmup is 0.06 of the total training duration. For our optimizer, we use a learning rate of $5.0e-4$ with an epsilon of $1e-06$, betas of $0.9$ and $0.98$, a weight decay of $1e-5$.

For fine-tuning the M2-BERT encoders, we use the Sentence Transformers library (Reimers & Gurevych, 2019). For all M2-BERT configurations, we use a learning rate of $5e-6$, a true batch size of 32, 1 epoch of fine-tuning, a maximum gradient norm of 1.0, and a ratio of 32 negative passages per query-positive passage pair. When using orthogonal projection loss (OPL) for fine-tuning, we use cosine similarity distance for calculating loss. When using prototype loss (PL), we first fine-tune the M2-BERT-32k model with the fine-tuned M2-BERT-128 model as the teacher model. To improve downstream retrieval accuracy, we then have a second-phase of fine-tuning in which we fine-tune with MNRL with a batch size of 2.

For evaluation, we use the BEIR library (Thakur et al., 2021) to calculate retrieval accuracy on both the LoCoV1 and BEIR benchmarks. For accuracy, we use normalized discounted cumulative gain at 10 (nDCG@10) (Wang et al., 2013).

### B.8 M2-BERT ON BEIR - EXPANDED RESULTS

| **Model** | SentenceBERT | M2-BERT |
|---|---|---|
| **Max Seq. Length** | 512 | 128 |
| **Param. Count** | 110M | 80M |
| MSMARCO | 65.0 | 59.8 |
| TREC COVID | 35.8 | 43.3 |
| NFCorpus | 23.1 | 24.6 |
| NQ | 34.5 | 30.6 |
| HotPot QA | 44.7 | 39.8 |
| FIQA | 22.0 | 22.7 |
| Arguana | 42.1 | 42.0 |
| Webis Touche 2020 | 11.0 | 19.1 |
| Quora | 84.6 | 84.2 |
| DBpedia Entity | 30.1 | 28.5 |
| SciDocs | 14.5 | 10.9 |
| Climate Fever | 55.4 | 57.8 |
| SciFact | 56.9 | 39.9 |
| **BEIR Score Average** | **64.5** | 63.4 |

Table 12: Expanded Results for M2-BERT Retrieval Encoder vs. SentenceBERT on BEIR.

### B.9 M2-BERT ON LOCOV1 - EXPANDED RESULTS

| | | | LoCoV1 Datasets | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Param. Count** | **Max.Seq. Length** | **Scrolls Summ ScreenFD** | **Scrolls Gov. Report** | **Scrolls QMSUM** | **QASPER Title** | **QASPER Abstract** | **Multi Field QA** | **2Wikim QA** | **Passage Retrieval** | **C.L. (HTML)** | **C.L. (Plain Text)** | **Legal Case Reports** | **S.O.** | **LoCo Avg.** |
| BGE-Large Zero-Shot | 335M | 512 | 65.8 | 92.6 | 51.3 | 87.1 | 94.5 | 89.3 | 69.4 | 20.1 | 10 | 10.1 | 18.9 | 74.1 | 56.9 |
| BGE-Large Fine-tuned | 335M | 512 | 84.8 | 96.0 | 70.2 | 93.5 | 97.8 | 92.1 | 71.1 | 22.5 | 22.0 | 22.8 | 42.6 | 76.5 | 65.0 |
| BGE-Large Fine-tuned w. Chunks | 335M | 512 | 80.7 | 95.5 | 60.3 | 89.3 | 96.6 | 88.4 | 66.4 | 20.3 | 21.0 | 21.8 | 39.9 | 76.7 | 61.8 |
| E5-Mistral | 7.11B | 4096 | 95.9 | 98.3 | 65.9 | **98.4** | **99.8** | 93.5 | 88.3 | 35.3 | 33.9 | 34.6 | 49.5 | 82.7 | 73.0 |
| E5-Mistral w. Chunks | 7.11B | 4096 | 95.6 | 98.4 | 61.4 | 96.8 | 99.7 | 90.5 | 84.8 | 32.9 | 32.8 | 32.7 | 49.2 | 83.1 | 71.5 |
| Jina Embeds. | 137M | 8192 | 93.3 | 98.6 | 40.5 | 95.1 | 99.4 | 86.4 | 81.6 | 60.7 | 27.0 | 26.1 | 30.7 | 69.0 | 67.2 |
| Jina Embeds. w. Chunks | 137M | 8192 | 6.1 | 25.2 | 4.2 | 32.5 | 54.3 | 43.8 | 21.6 | 10.4 | 0.9 | 0.5 | 1.8 | 28.9 | 19.2 |
| OpenAI Ada | N/A | 8192 | 86.2 | 97.1 | 57.3 | 93.8 | 98.9 | 90.1 | 78.9 | 31.2 | 16.3 | 16.8 | 28.2 | 72.3 | 63.9 |
| OpenAI Ada w. Chunks | N/A | 8192 | 86.2 | 97.1 | 52.1 | 93.8 | 98.9 | 90.1 | 78.9 | 31.2 | 16.3 | 16.8 | 30.7 | 72.3 | 63.7 |
| BM25 | N/A | N/A | 97.4 | **98.7** | 78.7 | 94.0 | 99.4 | 92.8 | **99.4** | 97.6 | 81.3 | 81.6 | 27.4 | 29.8 | 81.5 |
| ColBERTv2 | 110M | 512 | 66.5 | 88.0 | 56.0 | 85.5 | 94.5 | 85.0 | 71.7 | 21.5 | 14.7 | 17.6 | 17.2 | 44.5 | 54.3 |
| LongColBERT | 110M | N/A | 72.6 | 82.9 | 57.6 | 92.0 | 90.7 | 90.1 | 84.7 | 54.4 | 65.4 | 66.3 | 40.9 | 18.0 | 68.0 |
| Voyage-001 | N/A | 4096 | 76.7 | 92.4 | 52.9 | 88.4 | 91.7 | 88.7 | 57.0 | 17.7 | 13.0 | 12.8 | 14.0 | 74.9 | 56.9 |
| Cohere Embed-Eng. v3.0 | N/A | 512 | 75.3 | 92.2 | 50.8 | 89.8 | 93.1 | 88.9 | 68.2 | 22.1 | 13.3 | 14.3 | 24.3 | 75.3 | 59.0 |
| M2-BERT | 80M | 128 | 64.6 | 85.4 | 53.8 | 77.5 | 83.0 | 93.0 | 76.2 | 40.5 | 82.3 | 84.2 | 26.7 | 68.9 | 69.7 |
| M2-BERT | 80M | 2048 | 72.3 | 94.4 | 82.5 | 87.4 | 96.0 | 90.1 | 85.4 | 78.8 | 86.1 | 87.4 | 40.3 | 78.7 | 81.6 |
| M2-BERT | 80M | 8192 | 86.5 | 96.9 | 86.4 | 92.7 | 97.7 | **97.5** | 89.8 | 92.3 | 90.8 | 91.2 | 54.9 | 90.6 | 88.9 |
| M2-BERT | 80M | 32768 | **98.7** | 97.0 | **93.7** | 97.0 | 98.3 | 95.9 | 92.5 | **98.8** | **95.4** | **95.4** | **83.3** | **96.2** | **95.2** |

Table 13: M2-BERT and Baseline Model Performances on LoCoV1 benchmark - Complete Results.

### B.10 BEIR DATASET EXAMPLES

| BEIR Dataset | Query Example | Document Example |
|---|---|---|
| SciFact | 1/2000 in UK have abnormal PrP positivity. | OBJECTIVES To carry out a further survey of archived appendix samples to understand better the differences between existing estimates of the prevalence of subclinical infection with prions after the bovine spongiform encephalopathy epizootic and to see whether a broader birth cohort was affected, and to understand better the implications for the management of blood and blood products and for the handling of surgical instruments. DESIGN Irreversibly unlinked and anonymised large scale survey of archived appendix samples. SETTING Archived appendix samples from the pathology departments of 41 UK hospitals participating in the earlier survey, and additional hospitals in regions with lower levels of participation in that survey. SAMPLE 32,441 archived appendix samples fixed in formalin and embedded in paraffin and tested for the presence of abnormal prion protein (PrP). RESULTS Of the 32,441 appendix samples 16 were positive for abnormal PrP, indicating an overall prevalence of 493 per million population (95% confidence interval 282 to 801 per million). The prevalence in those born in 1941-60 (733 per million, 269 to 1596 per million) did not differ significantly from those born between 1961 and 1985 (412 per million, 198 to 758 per million) and was similar in both sexes and across the three broad geographical areas sampled. Genetic testing of the positive specimens for the genotype at PRNP codon 129 revealed a high proportion that were valine homozygous compared with the frequency in the normal population, and in stark contrast with confirmed clinical cases of vCJD, all of which were methionine homozygous at PRNP codon 129. CONCLUSIONS This study corroborates previous studies and suggests a high prevalence of infection with abnormal PrP, indicating vCJD carrier status in the population. |
| Quora | How do Russian politics and geostrategy affect Australia and New Zealand? | How does Russian politics affect Australia and New Zealand? |
| NQ | where does junior want to go to find hope | Throughout the novel, Junior shares his dreams with the readers. In the first chapter, he dreams of becoming a cartoon artist in order to get rich and escape the cycles of poverty and abuse on the reservation The idea that hope exists off the rez is echoed in later chapters, where Junior finds himself caught between home on the reservation and pursuing his dreams in the outside world. Junior asks his parents, "Who has the most hope?" to which they answer "White people".[h] The rez is characterized by lack of opportunity and poor education, the solution to which appears to lie in the Western world. |
| MSMARCO | cost of interior concrete flooring | For a 4 inch concrete floor, 1 yard of concrete will cover 80 square feet. The cost would be very close either way for a 4 inch concrete floor. If the floor is thicker than 4 inches, then the surface hardener is less money to use. |
| TREC-COVID | how does the coronavirus respond to changes in the weather | Abstract In this study, we aimed at analyzing the associations between transmission of and deaths caused by SARS-CoV-2 and meteorological variables, such as average temperature, minimum temperature, maximum temperature, and precipitation. Two outcome measures were considered, with the first aiming to study SARS-CoV-2 infections and the second aiming to study COVID-19 mortality. Daily data as well as data on SARS-CoV-2 infections and COVID-19 mortality obtained between December 1, 2019 and March 28, 2020 were collected from weather stations around the world. The country's population density and time of exposure to the disease were used as control variables. Finally, a month dummy variable was added. Daily data by country were analyzed using the panel data model. An increase in the average daily temperature by one degree Fahrenheit reduced the number of cases by approximately 6.4 cases/day. There was a negative correlation between the average temperature per country and the number of cases of SARS-CoV-2 infections. This association remained strong even with the incorporation of additional variables and controls (maximum temperature, average temperature, minimum temperature, and precipitation) and fixed country effects. There was a positive correlation between precipitation and SARS-CoV-2 transmission. Countries with higher rainfall measurements showed an increase in disease transmission. For each average inch/day, there was an increase of 56.01 cases/day. COVID-19 mortality showed no significant association with temperature. |

Table 14: BEIR Benchmark Examples

### B.11 NEEDLE-IN-THE-HAYSTACK SYNTHETIC TASK - EXPANDED RESULTS

| **Model** | BGE-Large Zero-shot | E5-Mistral | Jina Embeds. | OpenAI Ada Embeds. | M2-BERT |
|---|---|---|---|---|---|
| **Max. Seq. Length** | 512 | 4096 | 8192 | 8192 | 32768 |
| **Param. Count** | 335M | 7.11B | 137M | N/A | 80M |
| **Answer Position in Concat. Passage** | | | | | |
| 0 | 76.7 | 68.4 | 4.7 | 77.0 | 82.0 |
| 1 | 66.5 | 60.2 | 3.1 | 60.1 | 79.9 |
| 2 | 62.3 | 42.1 | 2.7 | 45.0 | 77.0 |
| 3 | 58.1 | 23.9 | 2.2 | 30.9 | 78.4 |
| 4 | 55.5 | 10.4 | 2.1 | 22.0 | 76.3 |
| 5 | 50.9 | 8.1 | 1.4 | 15.2 | 75.8 |
| 6 | 49.5 | 6.7 | 1.3 | 11.1 | 75.6 |
| 7 | 45.7 | 5.4 | 1.1 | 8.3 | 74.9 |
| 8 | 42.8 | 4.8 | 1.2 | 5.5 | 73.2 |
| 9 | 37.3 | 4.5 | 1.1 | 4.0 | 72.1 |
| 10 | 35.6 | 3.9 | 0.9 | 3.2 | 71.5 |
| 11 | 30.0 | 3.2 | 0.9 | 2.9 | 70.4 |
| 12 | 27.1 | 3.1 | 0.7 | 2.1 | 68.9 |
| 13 | 23.0 | 2.5 | 0.8 | 1.6 | 68.5 |
| 14 | 19.4 | 1.8 | 0.9 | 1.6 | 66.2 |
| 15 | 16.1 | 2.7 | 0.7 | 1.4 | 64.0 |
| 16 | 13.5 | 2.2 | 0.5 | 0.9 | 62.4 |
| 17 | 12.1 | 1.9 | 0.4 | 1.1 | 55.4 |
| 18 | 9.8 | 1.4 | 0.2 | 1.0 | 45.2 |
| 19 | 7.2 | 0.9 | 0.1 | 1.1 | 38.3 |
| 20 | 5.5 | 1.1 | 0.2 | 1.0 | 30.1 |
| 21 | 3.8 | 1.4 | 0.1 | 1.0 | 28.2 |
| 22 | 2.9 | 0.8 | 0.1 | 1.0 | 26.7 |
| 23 | 2.8 | 1.2 | 0.1 | 0.7 | 25.3 |
| 24 | 2.2 | 1.3 | 0.4 | 0.8 | 21.3 |
| 25 | 1.7 | 1.0 | 0.1 | 1.0 | 20.5 |
| 26 | 1.3 | 1.0 | 0.2 | 0.8 | 19.2 |
| 27 | 1.1 | 1.3 | 0.2 | 0.7 | 17.9 |
| 28 | 1.1 | 0.8 | 0.1 | 0.8 | 16.9 |
| 29 | 1.0 | 1.3 | 0.2 | 1.0 | 15.3 |
| 30 | 1.0 | 1.2 | 0.1 | 1.2 | 12.3 |
| 31 | 0.7 | 0.8 | 0.1 | 1.1 | 8.3 |
| 32 | 0.8 | 0.4 | 0.3 | 1.3 | 6.1 |
| 33 | 1.0 | 0.5 | 0.3 | 1.8 | 5.5 |
| 34 | 0.1 | 0.6 | 0.1 | 2.1 | 3.2 |
| 35 | 0.1 | 0.3 | 0.2 | 1.7 | 1.9 |
| 36 | 0.2 | 0.5 | 0.3 | 2.1 | 1.4 |
| 37 | 0.2 | 0.4 | 0.1 | 3.1 | 1.0 |
| 38 | 0.2 | 0.4 | 0.1 | 3.5 | 1.1 |
| 39 | 0.1 | 0.3 | 0.1 | 3.8 | 0.8 |
| **Synth. Task Avg.** | 19.2 | 6.9 | 0.8 | 8.2 | **41.0** |

Table 15: M2-BERT and Baseline Performances on Needle-in-the-Haystack Synthetic Task - Complete Results.

### B.12 MTEB CLASSIFICATION RESULTS

| **Model** | SentenceBERT | M2-BERT |
|---|---|---|
| **Max. Seq. Length** | 512 | 128 |
| **Param. Count** | 110M | 80M |
| AmazonCounterfactualClassification | 66.0 | 66.7 |
| AmazonPolarityClassification | 63.8 | 73.4 |
| AmazonReviewsClassification | 32.5 | 37.5 |
| Banking77Classification | 81.2 | 78.2 |
| EmotionClassification | 44.3 | 42.8 |
| ImdbClassification | 59.7 | 60.4 |
| MassiveIntentClassification | 68.4 | 63.5 |
| MassiveScenarioClassification | 73.1 | 71.6 |
| MTOPDomainClassification | 91.4 | 85.1 |
| MTOPIntentClassification | 71.9 | 59.2 |
| ToxicConversationsClassification | 66.9 | 65.0 |
| TweetSentimentExtractionClassification | 54.8 | 57.6 |
| **MTEB Average Accuracy** | **64.5** | 63.4 |

Table 16: M2-BERT-128 and SentenceBERT Performance on MTEB Classification - Complete Results.

### B.13 M2-BERT PRETRAINING STRATEGIES

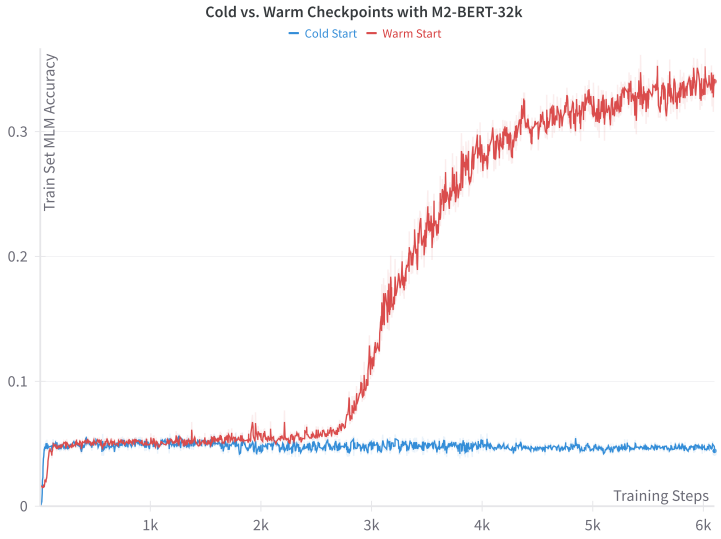


Figure 4: Cold vs. Warm Start for M2-BERT-32k Pretraining Checkpoints.

### B.14 Baseline Model Selection

- BGE-Large-en-v1.5: https://huggingface.co/BAAI/bge-large-en-v1.5
- E5-Mistral: https://huggingface.co/intfloat/e5-mistral-7b-instruct
- Jina Embeddings: https://huggingface.co/jinaai/jina-embeddings-v2-base-en
- OpenAI Ada Embeddings: https://platform.openai.com/docs/guides/embeddings
- VoyagerAI Voyage-001 Embeddings: https://docs.voyageai.com/embeddings/
- Cohere Embed-English v3.0: https://cohere.com/models/embed

### B.15 M2-BERT EFFICIENCY EXPERIMENTS

For all our efficiency experiments, we run each of the models on a single A100 GPU with 80GB of memory, running CUDA 11.7, Python 3.10, and PyTorch 1.13.1 (Paszke et al., 2019). We pre-tokenize all input sequences before measuring the time it takes to tokenize the entirety of the sequence, which can involve embedding separate chunks of the sequence if the model's maximum sequence length is less than the total sequence length.

### B.16 LoCoV0 Performance

| | | | LoCoV0 Dataset | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Param. Count.** | **Max. Seq. Length** | Summ ScreenFD | Gov. Report | QMSUM | QASPER Title | QASPER Abstract | **Average Score** |
| E5-Mistral | 7.11B | 4096 | 95.9 | 98.3 | 46.8 | 98.4 | 99.8 | 87.8 |
| BGE-Large Fine-tuned | 335M | 512 | 70.8 | 93.5 | 66.0 | 96.3 | 98.4 | 85.0 |
| Jina Embeds. | 137M | 8192 | 93.3 | 98.6 | 40.5 | 95.1 | 99.4 | 85.4 |
| OpenAI Ada | N/A | 8192 | 86.2 | 97.1 | 48.8 | 93.8 | 98.9 | 85.0 |
| Cohere Embed English v3.0 | N/A | 512 | 75.3 | 92.2 | 38.1 | 89.8 | 93.1 | 77.7 |
| Voyage voyage-01 | N/A | 4096 | 76.7 | 92.4 | 52.9 | 88.4 | 91.7 | 80.4 |
| M2-BERT-2k | 80M | 2048 | 81.8 | 94.7 | 58.5 | 87.3 | 95.5 | 83.6 |
| M2-BERT-8k | 80M | 8192 | 94.7 | 96.5 | 64.1 | 86.8 | 97.5 | 85.9 |
| M2-BERT-32k | 80M | 32768 | 98.6 | 98.5 | 69.5 | 97.4 | 98.7 | 92.5 |

Table 17: M2-BERT and Baseline Model Performances on LoCoV0

### B.17 LoCoV1 and BEIR Document Length Distributions

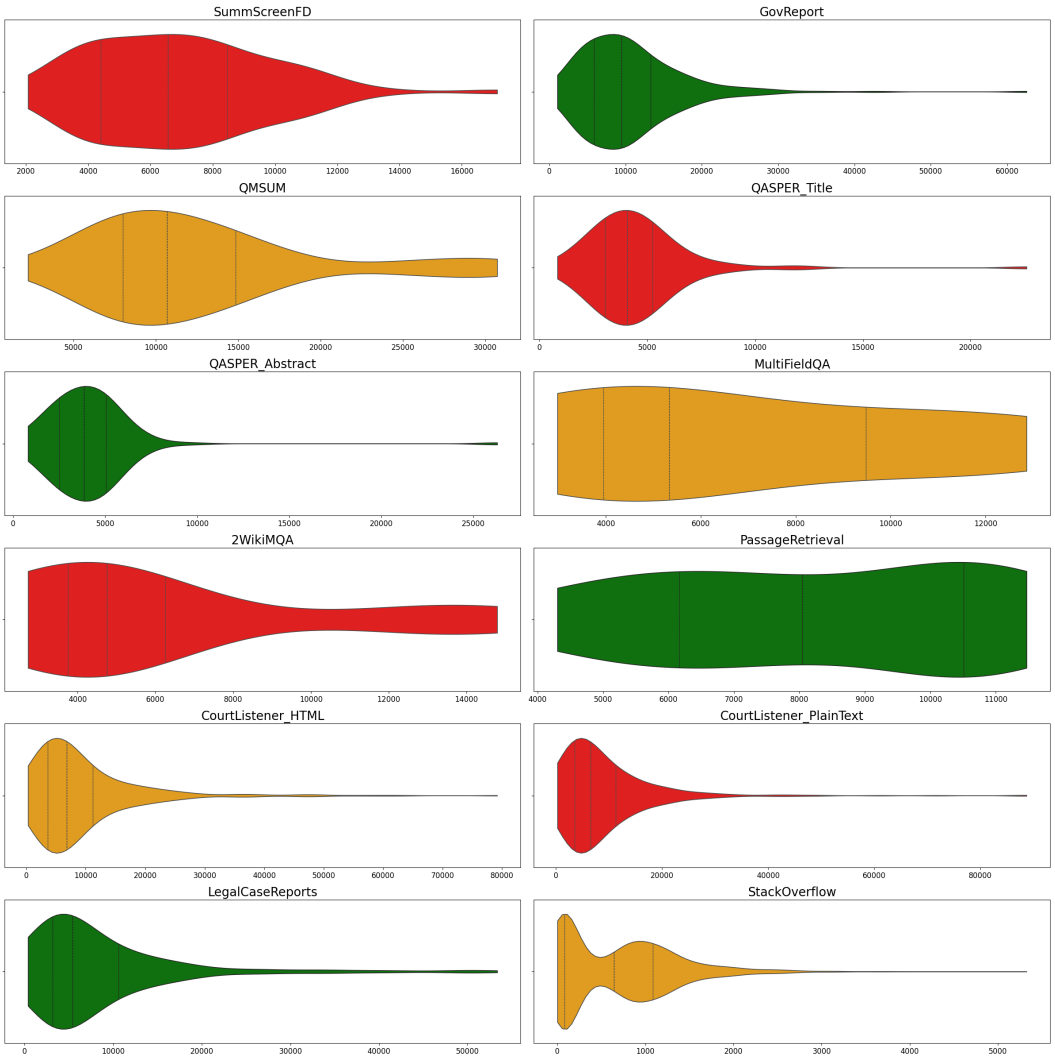


Figure 5: LoCoV1 Document Token Count Distributions.

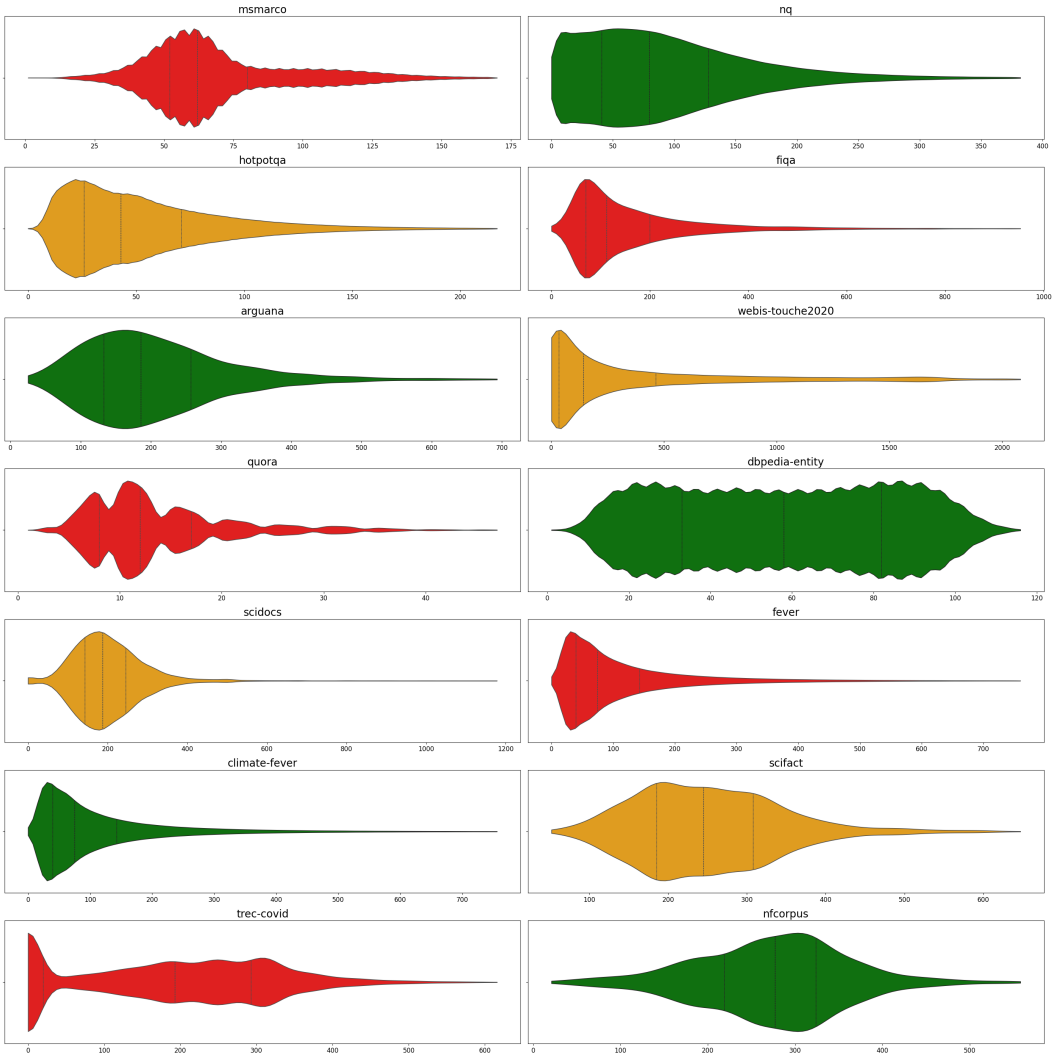


Figure 6: BEIR Document Token Count Distributions.

### B.18 LoCoV1 Performance Breakdown

| | **nDCG@10 for Document Subset** | | | |
|---|---|---|---|---|
| **Model** | <2k | >2k <8k | >8k <32k | >32k |
| BGE-Large Zero-Shot | 34.2 | 39.2 | 32.6 | 13.3 |
| Mistral | 54.5 | 60.7 | 47.9 | 24.8 |
| M2-BERT-128 | 70.8 | 60.2 | 34.3 | 15.4 |
| M2-BERT-2k | 63.9 | 68.1 | 47.9 | 24.1 |
| M2-BERT-8k | 88.5 | 90.6 | 89.9 | 81.1 |
| M2-BERT-32k | 90.4 | 93.1 | 94.4 | 86.1 |

Table 18: M2-BERT Encoder and Baseline Performances by Document Length. Queries are filtered by whether the token length of their answer documents are in the token range.

| Model | Query | Passage |
|---|---|---|
| BGE-Large | This report discusses runaway and homeless youth, and the federal response to support this population. There is no single definition of the terms "runaway youth" or "homeless youth." However, both groups of youth share the risk of not having adequate shelter and other provisions, and may engage in harmful behaviors while away from a permanent home. | Running away from home is not a recent phenomenon. Folkloric heroes Huckleberry Finn and Davy Crockett fled their abusive fathers to find adventure and employment. Although some youth today also leave home due to abuse and neglect, they often endure far more negative outcomes than their romanticized counterparts from an earlier era. Without adequate and safe shelter, runaway and homeless youth are vulnerable to engaging in high-risk behaviors and further victimization. Youth who live away from home for extended periods may become removed from... |
| BGE-Large | The professor thought it was possible to reduce the effects of reverberation by removing the low-energy segments. He thought a VAD-like approach would work. This would make it so that the model was more likely to keep an echo than throw out speech. | Professor B: I think for two years we were two months , uh , away from being done .<br>PhD A: And what was that , Morgan ? What project ?<br>Professor B: Uh , the , uh , TORRENT chip .<br>PhD A: Oh .<br>Professor B: Yeah . We were two {disfmarker} we were {disfmarker}<br>PhD C: Yeah .<br>Professor B: Uh , uh , we went through it... |
| BGE-Large | "[i]n deciding cases . . . [j]urors are not expected to lay aside matters of common knowledge or their own observations and experiences, but rather, to apply them to the facts as presented to arrive at an intelligent and correct conclusion" (internal quotation marks omitted) | The "officially released" date that appears near the beginning of each opinion is the date the opinion will be published in the Connecticut Law Journal or the date it was released as a slip opinion. The operative date for the beginning of all time periods for filing postopinion motions and petitions for certification is the "officially released" date appearing in the opinion. All opinions are subject to modification and technical correction prior to official publication in the Connecticut Reports and Connecticut Appellate Reports. In the event of discrepancies between... |
| E5-Mistral | In this paper, we describe a new national language technology programme for Icelandic. The programme, which spans a period of five years, aims at making Icelandic usable in communication and interactions in the digital world, by developing accessible, open-source language resources and software. The research and development work within the programme is carried out by a consortium of universities, institutions, and private companies, with a strong emphasis on cooperation between academia and industries. Five core projects will be the main content of the programme: language resources, speech recognition, speech synthesis, machine translation, and spell and grammar checking. We also describe other national language technology programmes and give an overview over the history of language technology in Iceland. | During the last decade, we have witnessed enormous advances in language technology (LT). Applications that allow users to interact with technology via spoken or written natural language are emerging in all areas, and access to language resources and open-source software libraries enables faster development for new domains and languages. However, LT is highly language dependent and it takes considerable resources to develop LT for new languages. The recent LT development has focused on languages that have both a large number of speakers and huge amounts of digitized language resources, like English, German, Spanish, Japanese, etc. Other languages, that have few speakers and/or lack digitized language resources, run the risk of being left behind. Icelandic is an example of a language with almost a negligible number of speakers, in terms of... |
| E5-Mistral | Who was Brooksley Elizabeth's first husband? | Brooksley Elizabeth Born (born August 27, 1940) is an American attorney and former public official who, from August 26, 1996, to June 1, 1999, was chair of the Commodity Futures Trading Commission (CFTC), the federal agency which oversees the U.S. futures and commodity options markets. During her tenure on the CFTC, Born lobbied Congress and the President to give the CFTC oversight of off-exchange markets for derivatives, in addition to its role with respect to exchange-traded derivatives, but her warnings were ignored or dismissed, and her calls for reform resisted by other regulators.Goodman, Peter S. The Reckoning - ... |
| E5-Mistral | Niles is scanning the society page when he sees a picture of Maris with another man. He plans to take an heiress on a date at a society event, the Snow Ball. He then realizes that he cannot dance but Daphne then offers to teach him. His date cancels, prompting Daphne to suggest that she go with him to the Ball. At the ball, Niles and Daphne dance, to show everyone there that he is not mourning his divorce. As they dance a tango, Niles declares that he adores Daphne, and she reciprocates. When the dance is over, Niles realizes that Daphne thought that he was just acting to try to impress everyone in the room. | ACT ONE Scene One - KACL Frasier\'s on air at KACL and he\'s running out of time. But Roz still hands him over to his next caller.\nFrasier: Well, we\'ve got about thirty seconds. I think we\'ve got time for one quick call. [presses button] Hello, Marlene, I\'m listening.\nMarlene: [v.o.] Oh my God, I\'m really on?\nFrasier: Yes, your problem, please...\nMarlene: [dog barking] Lucky, Lucky, get down. George, get the dog... |
| M2-BERT-32k | Which country Albertine, Baroness Staël Von Holstein's father is from? | Passage 1:\nAlbertine, baroness Staël von Holstein\nHedvig Gustava Albertina, Baroness de Staël-Holstein or simply Albertine (1797–1838), was the daughter of Erik Magnus Staël von Holstein and Madame de Staël, the granddaughter of... |
| M2-BERT-32k | The text is about Calvin Zabo, a biochemist who becomes obsessed with the idea of transforming into a superhuman form similar to the character Mr. Hyde in Stevenson's novel. He robs his employers to fund his experiments and seeks revenge on Donald Blake, a doctor who refuses to give him a job. Zabo successfully creates a formula that transforms him into a Hulk-like creature called Mister Hyde. Hyde attempts to kill Blake, but Blake transforms into Thor and survives. | Paragraph 1: With very few feature films made in Canada at all prior to the 1960s, in some years no Film of the Year winner was named at all, with the awards for Best Short Film or Best Amateur Film instead constituting the highest honour given to a film that year. Even the award for Film of the Year, when presented at all, often also went to a short film. The awards were also almost totally dominated by the National Film Board, to the point that independent filmmakers sometimes alleged a systemic bias which was itself a contributing factor to the difficulty of building a sustainable |
| M2-BERT-32k | "[T]he rules of criminal procedure require the appointment of counsel in PCRA proceedings." | J-S79022-17\n\n\nNON-PRECEDENTIAL DECISION - SEE SUPERIOR COURT I.O.P. 65.37\n\nCOMMONWEALTH OF PENNSYLVANIA : IN THE SUPERIOR COURT OF\n : PENNSYLVANIA\n :\n v. VERNELL MORRIS Appellant : No. 3731 EDA 2016\n\n Appeal from the PCRA Order November 3, 2016\n In the Court of Common Pleas of Philadelphia County Criminal Division at\n No(s): CP-51-CR-1113151-1992\n\n\nBEFORE: GANTMAN, P.J.,... |

Table 19: **LoCoV1 Performance Analysis by Model**: Passages that aren't highlighted were retrieved successfully while passages highlighted in red were not successfully retrieved. Retrieval success is defined as whether it was retrieved in the first 10 passages.