Choose Your Words Wisely: Domain-adaptive Masking Makes Language Models Learn Faster

Vanshpreet S. Kohli
IIIT Hyderabad
vanshpreet.k@research.iiit.ac.in

Aaron Monis
IIIT Hyderabad
aaron.monis@students.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

Foundational Language Models perform significantly better on downstream tasks in specialised domains (such as law, computer science, and medical science) upon being further pre-trained on extensive domain-specific corpora, but this continual pre-training incurs heavy computational costs. Indeed, some of the most performant specialised language models such as BioBERT incur even higher computing costs during domain-specific training than the pre-training cost of the foundational models they are initialised from. In this paper, we argue that much of the extended pre-training is redundant, with models seemingly wasting valuable resources re-learning lexical and semantic patterns already well-represented in their foundational models such as BERT, T5 and GPT. Focusing on Masked Language Models, we introduce a novel domain-specific masking strategy that is designed to facilitate continual learning while minimizing the training cost. Using this approach, we train and present a BERT-based model trained on a biomedical corpus that matches or surpasses traditionally trained biomedical language models in performance across several downstream classification tasks while incurring up to 11 times lower training costs.

1 Introduction

Rapid advancements in Large Language Models (LLMs) (OpenAI, 2024, Touvron et al., 2023) have resulted in an increased focus on their capabilities in specialised domains like biology, law and computer science (Lai et al., 2024, Chen et al., 2024). The significance of foundational models like BERT, T5, GPT and LLaMa for application in such fields is particularly evident from the multitude of models based on them – such as LegalBERT (Chalkidis et al., 2020), SciFive (Phan et al., 2021), BioGPT (Luo et al., 2022) and PMC-LLaMa (Wu et al., 2023) – delivering state-of-the-art results on benchmarks in their respective domains. Given the vast amount of training data available for continual pre-training across various fields, existing literature shows that further pre-training foundational language models on a domain-specific corpus yields better model performance across downstream tasks (Gururangan et al., 2020, Rongali et al., 2021). However, pre-training language models can be resource-prohibitive, both in terms of monetary cost and time spent on training. This is especially true for large, parameter-dense language models like LLaMa (Touvron et al., 2023), which are not only significantly more expensive to train further, but also exhibit much smaller improvements in downstream performance per unit of compute spent (Chen et al., 2024).

It is therefore worthwhile to attempt to leverage the capabilities of contemporary foundational language models for tasks across such domains without expending exorbitant computing resources on diminishing returns. In this paper, we present a novel strategy for continual/mixed-domain pre-training that emphasises selecting relevant (as opposed to random) training samples to maximise compute efficiency. We test our strategy in the biomedical domain by further pre-training BERT on a corpus of PubMed abstracts, mirroring the selection of architecture and pre-training corpora of BioBERT (Lee et al., 2020), one of the most popular biomedical language models. In our testing over 8 Named Entity Recognition (NER) tasks, the resultant model significantly outperforms BioBERT-v1.0 at two-thirds of the compute cost, and performs similarly to BioBERT-v1.1 at about one-eleventh of its compute cost.

2 Related Works

2.1 Biomedical Language Models

The vast majority of domain-adapted language models employ the transformer architecture

Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025), pages 87–91
May 4, 2025 ©2025 Association for Computational Linguistics

(Vaswani, 2017), either as encoder layers (like BERT), decoder layers (like GPT) or a combination of the two (like T5). The two most popular strategies to train domain-specific language models are: **1.** pre-training from scratch on a corpus relevant to the domain, and **2.** further pre-training a foundational model on the corpus. The former approach has generally been shown to yield better results when large corpora are available for training, because such models use a vocabulary relevant to their corpus instead of inheriting the vocabulary from a general-domain model (Gu et al., 2021). However, experiments from the likes of Chalkidis et al. (2020) and Lee et al. (2020) demonstrate that the latter approach yields competitive results while requiring significantly less training due to the transfer of learning from their foundational models.

Of the architectures discussed above, the versatility of encoder representations in downstream tasks makes BERT-like models vastly popular for domain adaptation. This is particularly true in the biomedical domain which is littered with models like SciBERT (Beltagy et al., 2019), BioBERT, BioLinkBERT (Yasunaga et al., 2022), DistilBioBERT (Rohanian et al., 2022), PubMedBERT and so on. Indeed, among the 8 tasks we test our model on, the current State-Of-The-Art (SOTA) results[1] are claimed by a non–BERT-style model only for two of the tasks.

2.2 Curriculum Learning

In a curriculum learning setting (Bengio et al., 2009), training samples are presented to a model not arbitrarily, but in an "easy-to-difficult" order, where the method for ranking the difficulty of samples depends on the model and task involved. This framework is designed to better simulate human cognition, wherein humans learn complex concepts more easily after having learnt basic ones. Recent studies indicate that employing this approach demonstrably accelerates convergence compared to random presentation of samples in many settings (Roy et al., 2024, Jarca et al., 2024, Tang et al., 2024).

We approach domain-adaptation of a foundational language model as an analogous task to curriculum learning. We posit that since the model has already been trained on a general ("easy") corpus and must now be trained in a specific ("difficult") domain, we can apply the same curriculum learning principle of curating training samples such that they specifically facilitate domain-specific learning. To the best of our knowledge, this approach has not thus far been tested or reported on.

3 Methodology

We begin by creating a biomedical corpus consisting of PubMed abstracts publicly available at https://pubmed.ncbi.nlm.nih.gov/download/, amounting to about 9.4GB of text. Leveraging the linguistic difficulty criterion and subsequent curriculum generation approach introduced by Lee et al. (2022), who claim that frequently occuring words that have many connections in a large knowledge graph are easier to learn, we build a set S of "basic" concepts – i.e. the n concepts with the most connections in a large-scale knowledge graph that occur in the corpus above a threshold frequency f. Iterating through all the elements s_i of S, we add s_i and every concept in ConceptNet within k hops of s_i to a new set C, which acts as a "curriculum" consisting of relevant concepts. Following manual assessment of the curriculum generated, we settled on using $f = 200,000$, $n = 5,000$ and $k = 5$.

For our purposes, despite the availability of biomedical knowledge graphs like BIKG (Geleta et al., 2021) and BIOS (Yu et al., 2022), we chose to use the general-domain ConceptNet (Speer et al., 2017) as the knowledge graph. We did not assess the overlap between these specific knowledge graphs and our corpus, and could not be certain that their usage would not be counterproductive given that BERT's vocabulary itself is not tailored towards biomedical terms. Moreover, this makes our approach easier to generalise for other domains without pre-existing large knowledge graphs. Nonetheless, we recognise that the use of domain-specific knowledge graphs for concept extraction, wherever available, is worth investigating in future studies.

Since this curriculum includes general-domain concepts already represented well in BERT, we iterate once over BERT's corpus (Wikipedia + BooksCorpus), identify concepts occuring more than $f/3$ times and remove them from C. Note that this cutoff frequency has been scaled down with respect to the threshold frequency used for the PubMed corpus above to account for the difference in sizes between the corpora. We then iterate through C and remove any concepts that do not occur at all in the PubMed corpus, ensuring that

[1] sourced from https://paperswithcode.com/sota

Dataset	Entity type	No. of annotations
NCBI Disease	Disease	6,881
BC5CDR	Disease	12,694
BC5CDR	Drug/Chem.	15,411
BC4CHEMD	Drug/Chem.	79,842
BC2GM	Gene/Protein	20,703
JNLPBA	Gene/Protein	35,460
LINNAEUS	Species	4,077
Species-800	Species	3,708

Table 1: Statistics of the biomedical NER datasets.

the concepts now contained in C are relevant to the biomedical domain.

We then initialize our model from the publicly available BERT-base checkpoint, and train it for Masked Language Modeling (MLM) over our corpus. Past studies indicate that the difference between using cased and uncased models to warm-start biomedical language models is minimal with no clear advantage for either (Lee et al., 2020, Gu et al., 2021, and it is beyond the scope of our current experiment to test and compare the two. For our purposes, we use the uncased version of the model.

Differing from the likes of BERT and BioBERT that randomly mask 15% of the tokens in each batch, we mask only the tokens that form a concept within the previously curated curriculum C while ensuring that no more than 20% of the tokens in any batch are masked. As concepts can span multiple tokens, we follow Lee et al. (2022)'s Whole Concept Masking (WCM) strategy such that all the tokens comprising a single concept are simultaneously masked. As is the standard, we replace 80% of the masked concepts with a mask token, replace another 10% with a random token and do not replace the remaining 10%. Since existing literature shows minimal gains from calculating the Next Sentence Prediction (NSP) loss (Liu et al., 2019), we chose to omit it; MLM was our sole pre-training objective.

4 Experimental Setup

4.1 Pre-training

We trained the model for 200K steps on four NVIDIA RTX 6000 GPUs, using PyTorch's DistributedDataParallel to share the load across the GPUs. The batch size was fixed at 256 and the maximum sequence length was set to 256, resulting in 65,536 tokens per training iteration. This equates to 33% lower compute compared to BioBERT-v1.0 trained on the same corpus (98,304 tokens per iteration and 200K iterations), and 91% lower compute than BioBERT-v1.1, which was trained on the same corpus[2] for 1.2M training steps and additionally trained on full-length PubMed Central articles ($\sim$3 times the corpus size of PubMed Abstracts) for 270K steps. Note that the computational overhead caused by curriculum generation is minimal compared to model training, as it only requires iterating over two corpora and a section of one knowledge graph.

4.2 Fine-tuning

With NER being a fundamental task for text mining, we focus our limited testing on commonly used NER benchmarks. We fine-tune and evaluate our model on 8 tasks: BC2GM (Smith et al., 2008), BC4CHEMD (Krallinger et al., 2015), BC5CDR-Chemical (Li et al., 2016), BC5CDR-Disease (Li et al., 2016), JNLPBA (Collier et al., 2004), LINNAEUS (Gerner et al., 2010), NCBI Disease (Doğan et al., 2014) and Species-800 (Pafilis et al., 2013). We use pre-processed versions of the respective datasets released by Rohanian et al. (2022). Some specifications for the datasets are listed in Table 1. Following the setup described in the BioBERT paper, we use a learning rate of 5e-5 and train for 25 epochs per dataset. We leave testing this approach in other tasks – such as Question Answering, Relation Extraction as well as other NER tasks – for future studies.

5 Results

The results obtained by our model relative to BERT, BioBERT-v1.0, BioBERT-v1.1 and the current SOTA[3] are shown in Table 2. We consider these to be the most apt comparisons to showcase because BERT is the baseline we train upon, and BioBERT most closely reflects what our model's performance would be if it had been trained using regular MLM. The models delivering the SOTA results are, for the most part, more resource-intensive to train or are tailored towards Biomedical NER tasks as opposed to being general-purpose biomedical transformers. Nevertheless, we consider their performance to be relevant benchmarks and include them in this comparison.

[2] our corpus is collected from the same source but is larger as a virtue of being more up-to-date

[3] to the best of our knowledge

Task/Model	SOTA	BERT	BioBERT-v1.0	BioBERT-v1.1	Ours
BC2GM	*86.97*	81.79	82.54	84.72	**85.43**
BC4CHEMD	*94.39*	90.04	91.26	**92.36**	90.23
BC5CDR-Chemistry	*94.88*	91.16	92.64	**93.47**	93.25
BC5CDR-Disease	*88.50*	82.41	86.2	**87.15**	85.49
JNLPBA	*82.0*	74.94	76.65	77.49	**79.29**
LINNAEUS	*92.7*	87.6	88.13	88.24	**89.23**
NCBI Disease	*89.71*	85.63	87.38	**89.71**	87.92
Species-800	*82.44*	71.63	73.08	74.06	**75.20**

Table 2: Performance comparison across different models (F1 scores). The best result other than the SOTA (*italicised*) is in **bold**, and the second-best is underlined.

Our model outperforms BioBERT-v1.0 in three-fourths of the tasks and, despite significantly less training on a much smaller corpus, outperforms BioBERT-v1.1 in half of the tasks, demonstrating the effectiveness of our training-sample-curation strategy.

Limitations and Future Work

We acknowledge that being a short extended abstract, this paper does not present a full comprehensive study detailing the impact of our strategy. Our aim in presenting our preliminary experiment and findings is to incite further research into this idea from the broader NLP community, encouraging exploration of this approach with different parameters, domains, corpora, model sizes, training steps, model architectures and so on.

Ethics statement

The authors have no competing interests to declare that are relevant to the contents of this article. All the datasets and models accessed as part of this study were sourced from publicly available archives and checkpoints.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina Kuttichi Keloth, Kalpana Raja, Jiming Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, Zhiyong Lu, and Hua Xu. 2024. A systematic evaluation of large language models for biomedical natural language processing: benchmarks, baselines, and recommendations. *Preprint*, arXiv:2305.16326.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

David Geleta, Andriy Nikolov, Gavin Edwards, Anna Gogleva, Richard Jackson, Erik Jansson, Andrej Lamov, Sebastian Nilsson, Marina Pettersson, Vladimir Poroshin, Benedek Rozemberczki, Timothy Scrivener, Michaël Ughetto, and Eliseo Papa. 2021. Biological insights knowledge graph: an integrated knowledge graph to support drug development. *bioRxiv*.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11:1–17.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

Andrei Jarca, Florinel-Alin Croitoru, and Radu Tudor Ionescu. 2024. Cbm: Curriculum by masking. *Preprint*, arXiv:2407.05193.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7:1–17.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2024. Large language models in law: A survey. *AI Open*, 5:181–196.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mingyu Lee, Jun-Hyung Park, Junho Kim, Kang-Min Kim, and SangKeun Lee. 2022. Efficient pre-training of masked language model via concept-based curriculum masking. *arXiv preprint arXiv:2212.07617*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

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

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Evangelos Pafilis, Sune P Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *Preprint*, arXiv:2106.03598.

Omid Rohanian, Mohammadmahdi Nouriborji, Samaneh Kouchaki, and David A. Clifton. 2022. On the effectiveness of compact biomedical transformers. *Preprint*, arXiv:2209.03182.

Subendhu Rongali, Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2021. Continual domain-tuning for pretrained language models. *Preprint*, arXiv:2004.02288.

Soumyadeep Roy, Shamik Sural, and Niloy Ganguly. 2024. *Unlocking Efficiency: Adaptive Masking for Gene Transformer Models*. IOS Press.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9:1–19.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Yao Tang, Zhihui Xie, Zichuan Lin, Deheng Ye, and Shuai Li. 2024. Learning versatile skills with curriculum masking. *Preprint*, arXiv:2410.17744.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine. *Preprint*, arXiv:2304.14454.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *Preprint*, arXiv:2203.15827.

Sheng Yu, Zheng Yuan, Jun Xia, Shengxuan Luo, Huaiyuan Ying, Sihang Zeng, Jingyi Ren, Hongyi Yuan, Zhengyun Zhao, Yucong Lin, Keming Lu, Jing Wang, Yutao Xie, and Heung-Yeung Shum. 2022. Bios: An algorithmically generated biomedical knowledge graph. *Preprint*, arXiv:2203.09975.