Enhancing abstract screening classification in evidence-based medicine: incorporating domain knowledge into pre-trained models.

OFORI-BOATENG, R., ACEVES-MARTINS, M., WIRANTUGA, N. and MORENO-GARCIA, C.F.

2024

Enhancing Abstract Screening Classification in Evidence-Based Medicine: Incorporating Domain Knowledge into Pre-trained Models

Regina Ofori-Boateng[1 [0000−0002−0319−773X]], Magaly Aceves-Martins[2 [0000−0002−9441−142X]], Nirmalie Wirantuga[1 [0000−0003−4040−2496]], and Carlos Francisco Moreno-García[1 [0000−0001−7218−9023]]

[1] School of Computing, Robert Gordon University, Aberdeen, Scotland
[2] The Rowett Institute, University of Aberdeen, Scotland
{r.ofori-boateng, c.moreno-garcia}@rgu.ac.uk

Abstract. Evidence-based medicine (EBM) represents a cornerstone in medical research, guiding policy and decision-making. However, the robust steps involved in EBM, particularly in the abstract screening stage, present significant challenges to researchers. Numerous attempts to automate this stage with pre-trained language models (PLMs) are often hindered by domain-specificity, particularly in EBMs involving animals and humans. Thus, this research introduces a state-of-the-art (SOTA) transfer learning approach to enhance abstract screening by incorporating domain knowledge into PLMs without altering their base weights. This is achieved by integrating small neural networks, referred to as knowledge layers, within the PLM architecture. These knowledge layers are trained on key domain knowledge pertinent to EBM, PICO entities, PubmedQA, and the BioASQ 7B biomedical Q&A benchmark datasets. Furthermore, the study explores a fusion method to combine these trained knowledge layers, thereby leveraging multiple domain knowledge sources. Evaluation of the proposed method on four highly imbalanced EBM abstract screening datasets demonstrates its effectiveness in accelerating the screening process and surpassing the performance of strong baseline PLMs.

Keywords: Pre-trained Language Models · Domain Integration · Transfer Learning · Evidence-Based Medicine · Abstract Text Classification.

1 Introduction

Evidence-Based Medicine (EBM) presents the highest form of reliable evidence in shaping healthcare policies and decision-making [1]. Generally, the process involves (i) formulating a protocol, (ii) defining the research question using entity frameworks such as PICO[3] to encapsulate the inclusion and exclusion criteria,

[3] where PICO denotes Population, Intervention, Comparison, Outcome

(iii) searching, (iv) screening abstracts, (v) extracting and analysing data from pertinent articles, and (vi) interpreting and publishing the findings. This process although structured is labour-intensive, further exacerbated by the daily increase in published articles. It is reported that the typical time frame for completing an EBM is approximately 15 months [1]. Thus, most EBMs become outdated before completion, needing major revisions.

Among all the stages in EBM, *abstract screening* has been reported to be the most challenging stage [2]. For example, research indicates that an experienced researcher typically spends 30-90 sec screening a single abstract, and estimated that 5,000 publications usually require 8-125 hrs[2]. Numerous methodologies for automating this stage have been proposed [2] ranging from traditional machine learning (ML) models to advanced PLMs, where they are fully fine-tuned (FFT) on EBM abstract datasets. However, most of these approaches are hindered by domain specificity, especially in highly imbalanced studies involving humans and animals [3]. Furthermore, PLMs comprise an extensive number of parameters; thus, in FFT, the parameters of the PLMs are updated whenever a new EBM dataset is introduced, resulting in increased computational costs and memory requirements. To tackle these issues, this paper investigates a SOTA method to integrate domain knowledge into PLMs for abstract screening tasks [4].

2 Related work

Many methods have been proposed for abstract screening, from traditional ML algorithms like Support Vector Machine (SVM) and Naive Bayes (NB) to SOTA PLMs. Timsina et al. [4] proposed using ULMS as a feature extraction technique and a softMax SVM classifier for abstract classification. Almeida et al. [5] also suggested the addition of MeSH and keywords to the abstracts for training a decision tree classifier. Similarly, Kontonatsios et al. [6] presented using MesH heading to train a neural network, and [7] proposed using Latent Dirichlet Allocation (LDA). With the rise of PLMs, medical domain knowledge PLMs such as SciBERT, PubMedBERT (PMBERT), BioBERT and CBERT (CBERT) have been proposed. For example, for this task, Hasny et al. [8] proposed using variants of BERT base models such as BERT-Meduim, SciBERT, BioBERT and CBERT. Ofori-Boateng et al. [2] also presented attention mechanisms with LSTM and Bi-LSTM. Moreno et al. [9] presented a zero-shot classification method for abstract screening. Similarly, [10] also proposed using GPT. Despite their advancements, these PLMs were originally trained on unstructured corpora, lacking the structured domain knowledge essential for biomedical tasks. As such, these PLMs treat biomedical concepts as conventional tokens, limiting their effectiveness [3].

2.1 Research Questions

We explore integrating essential domain knowledge into the models to address these issues. Specifically, we focus on incorporating PICO entities along with two

[4] For reproducibility, the source code and datasets are available on Github. https://github.com/reginaofori/EBM-Domain-Integration-PLMs

biomedical Q&A datasets, PubMedQA[5] and BioASQ 7B[6]. PICO entities are fundamental in EBM, while PubMedQA and BioASQ 7B offer formats similar to the EMB abstract datasets (context/abstracts, question, and decision). To this end, we ask the following **RQs**:

1. How can the diverse domain knowledge crucial for abstract screening tasks be integrated into a base PLM without adjusting model parameters? We insert small neural networks (knowledge layers) into the layers of a base PLM, SciBERT, using the principle of adapters [11] and train on the domain knowledge. Our choice of SciBERT is from a practical viewpoint as it was trained to cover a broad biomedical domain, thus advantageous for this task.
2. What is the effect of different configurations of the knowledge layers (where they are inserted) on the downstream task? We investigate and compare three configurations of inserted networks to analyse their influence on the downstream task.
3. Can adapter-based tuning perform better than SOTA FFT PLMs proposed for EBM abstracts? We empirically compare the performance of the trained knowledge layers with FFT SciBERT. Additionally, we examine the transferability and modularity of the method by inserting the already-trained networks into variants; CBERT, PMBERT, and BioBERT, adapter-tuning them, and comparing them against their FFT versions.

3 Methodology

3.1 Class Imbalance: Back translation

EBM abstract classification struggles with class imbalance, where the number of excluded abstracts outweighs the included. Traditional methods have been proposed, such as cost-sensitive classifiers and data resampling [4]. However, this study proposes a SOTA data augmentation technique to address this issue called *Back-translation*. It involves translating the original text into another language and then back into the original language, generating a paraphrased version. Despite potential inaccuracies that may be introduced during re-translation, this method has demonstrated effectiveness in NLP tasks [8]. For this study, the Google Translate API[7] was utilised to translate English abstracts in the training dataset into seven different source languages (Spanish, French, German, Italian, Chinese (simplified), Chinese (traditional), and Irish), followed by re-translation back to English. Notably, back translation was applied only to the minority class (include). Further details on partitioning the downstream dataset for translation are provided in Section 4.

[5] https://pubmedqa.github.io/
[6] http://participants-area.bioasq.org/datasets
[7] https://translate.google.com/

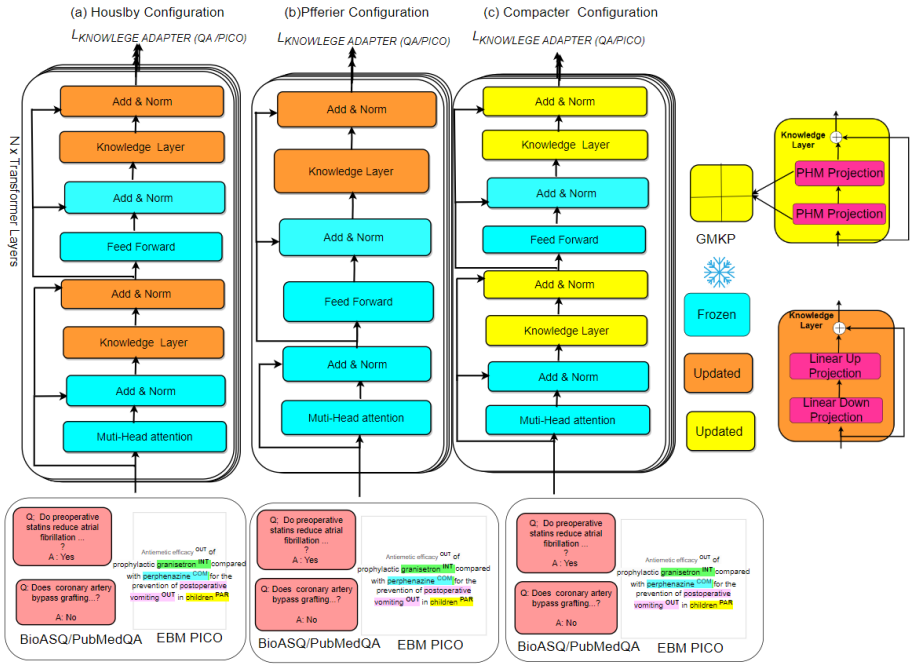

Fig. 1: Methodology for training the individual knowledge layers (PICO, BioASQ, and PubMedQA). The training involves 1) where the knowledge layer is inserted within the SciBERT PLM with frozen parameters and 2) where we investigate training with three configurations, in (a) the Houlsby (H), (b) the Pfeiffer, and in (c) the Compacter (C), a similar architecture of (a) but with a modification.

3.2 Overview of Adapters/Knowledge Layers

Adapters, originally proposed by Rebuffi et al. [13], are small trainable neural networks integrated within the layers of PLMs. An adapter consists of four main components: a FeedForward Linear Down Projection (FFD), FeedForward Linear Up Projection (FFU), a non-linear activation function (LeakyReLU), and a skip residual. The FFD and FFU reduce dimensionality, converting input from the PLM's high-dimensional space to a lower-dimensional one. For example, the FFD of the adapter maps the input data from the original high-dimensional space, d_{PLM}, to a much lower-dimensional space, h_{adapter}, where typically $h_{\text{adapter}} \ll \mathrm{d}_{PLM}$. Readers are referred to [11] for the detailed mathematical explanation. The LeakyReLU enables the adapter to handle negative inputs, ensuring a more dynamic range for the activations. Lastly, the skip residual ensures the model doesn't lose essential information during transformation.

3.3 Approach–Training the Knowledge Layers

To address **RQ1** and **RQ2**, Figure 1 illustrate the training of the three domain knowledge layers/adapters (PICO, PubMedQA, and BioASQ). We employ a comprehensive method in two phases: In Phase 1, the knowledge layers are integrated within every layer of the base SciBERT PLM to ensure a granular capture of information. In Phase 2, to examine the effect of different adapter configurations, we experiment with three distinct existing configurations in training the knowledge layers: (a) the Houlsby Configuration (H) [11], where the adapters modules are placed before the multi-head attention mechanism and the FeedForward layer of the SciBERT model as seen in Figure 1, (b) The Pfeiffer Configuration (Pf) [15], where the adapter modules are placed exclusively after the FeedForward layer and (c) The Compacter Configuration (C) similar to the Houlsby configuration, but replaces the standard linear FFD and FFU with a more intricate Parameterised Hypercomplex Multiplication (PHM) layer [14]. The PHM layer uniquely determines its weights by computing the Global Multiplier of the Kronecker Product (GMKP) between two concise matrices. Readers are referred to the work done by [14] for a detailed explanation of how the GMKP and PHM work in the compacter. During the training of the knowledge layers, the integrated layers introduce trainable parameters, denoted by Φ_n which are only updated, while the core weights of the base SciBERT, Θ, remains static. This strategy accelerates the training process.

Training the Q&A Knowledge Layers–PubMedQA and BioASQ. The main goal of training a Q&A knowledge layer is to facilitate efficient transfer learning for our downstream abstract classification task, capitalising on the capabilities of SciBERT. PubMedQA and BioASQ, the two Q&A datasets used for training, are described in Table A2 in the appendix. In refining the training quality for PubMedQA (made up of three labels; yes/no/maybe), the"maybe" labels are excluded from both training and validation sets, ensuring a focus on clear-cut include ("yes") or exclude ("no") decisions to avoid potential ambiguities during training and in real life cases. Thus given the classification task (predicting "yes" or "no"), SciBERT is initialised with a binary sequence classification head, while the adapter module explained in Section 3.3 is introduced for training, keeping the main parameters of the SciBERT model frozen. The raw text Q&A data is tokenized using SciBERT's tokenizer, combining questions with their corresponding contexts e.g., "[CLS] question [SEP] context [SEP] and the label "yes" is mapped to the label 1, while "no" is 0. Given our binary classification task, the cross-entropy loss function for optimisation is mathematically given as:

$$L_{\text{Q\&A}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{1}$$

where $L_{\text{Q\&A}}$ is the loss for the Q&A datasets, N is the total number of samples in each datasets, y_i denotes the actual label of the i-th sample, and $\hat{y}_i$ represents the predicted probability for the i-th sample being labelled as "yes".

Training EBM-PICO Knowledge Layer. The PICO framework is a fundamental structure for formulating clinical questions in EBM. In training the

Table 1: Summary of results on the LN_19 (1) and AH_19(2) datasets.[8]

DB		Methods	Precision	Recall	F1	WSS@95	AUC_PR	ROC
1	Know. Int. Baselines	KRISSBERT	0.9957	0.9957	0.9949	0.9465	0.82	0.97
		CODER-BERT	0.9948	0.9948	0.9935	0.9474	0.72	0.97
	FFT	SciBERT	0.9953	0.9957	0.9853	0.9448	0.79	0.95
	Modif. SciBERT	FPBPA(H)	**0.9957**	**0.9957**	**0.9949**	0.9465	**0.87** ↑ 0.05	**0.99**
		FPBPA(Pf)	0.9931	0.9931	0.9903	0.9491	0.85	0.87
		FPBPA (C)	0.9845	0.9922	0.9884	**0.9500**	**0.86**	**0.99**
	FFT	PMBERT	0.9940	0.9942	0.9920	0.9483	0.47	0.92
	Modif. PMBERT	FPBPA(H)	0.9948	0.9958	0.9935	0.9474	**0.81**	**0.99**
		FPBPA(Pf)	**0.9966** ↑ 0.09	**0.9970** ↑ 0.13	**0.9963** ↑ 0.14	**0.9591** ↑ 1.26	0.79	0.97
		FPBPA(C)	0.9942	0.9938	0.9935	0.9474	0.77	**0.99**
	FFT	BioBERT	0.9932	0.9934	0.9923	0.9483	0.57	0.98
	Modif. BioBERT	FPBPA(H)	0.9943	0.9940	0.9920	0.9483	0.66	**0.99**
		FPBPA(Pf)	**0.9952**	**0.9957**	**0.9949**	**0.9481**	**0.75**	**0.99**
		FPBPA(C)	0.9934	0.9931	0.9903	0.9442	0.67	0.98
	FFT	CBERT	0.9945	0.9948	0.9935	0.9474	0.62	0.98
	Modif. CBERT	FPBPA(H)	**0.9953**	**0.9957**	**0.9961**	0.9448	0.68	**0.99**
		FPBPA(Pf)	0.9939	0.9936	0.9934	0.9405	**0.77**	**0.99**
		FPBPA(C)	0.9931	0.9930	0.9903	**0.9491**	0.67	0.98
2	Know. Int. Baselines	KRISSBERT	0.9745	0.9777	0.9728	0.9388	0.46	0.82
		CODER-BERT	0.9428	0.9710	0.9567	0.9500	0.61	0.88
	FFT	SciBERT	0.9539	0.9725	0.9600	**0.9555** ↑ 0.55	0.69	0.93
	Modif. SciBERT	FPBPA(H)	0.9821	0.9821	0.9754	0.9299	**0.70** ↑ 0.24	**0.95** ↑ 0.13
		FPBPA(Pf)	**0.9832**↑ 0.87	**0.9829** ↑ 0.52	**0.9794** ↑ 0.66	0.9433	0.52	0.92
		FPBPA(C)	0.9715	0.9762	0.9713	0.9388	0.57	0.85
	FFT	PMBERT	0.9682	0.9657	0.9675	0.9433	0.50	0.96
	Modif. PMBERT	FPBPA(H)	**0.9775**	**0.9769**	**0.9694**	0.9433	**0.58**	**0.91**
		FPBPA(Pf)	0.9650	0.9739	0.9633	**0.9500**	0.41	0.82
		FPBPA(C)	0.9719	0.9754	0.9691	0.9388	0.47	0.87
	FFT	BioBERT	0.9662	0.9721	0.9680	0.9188	0.60	0.88
	Modif. BioBERT	FPBPA(H)	**0.9728**	**0.9756**	**0.9695**	0.9433	**0.69**	**0.93**
		FPBPA(Pf)	0.9650	0.9739	0.9633	**0.9500**	0.41	0.82
		FPBPA(C)	0.9663	0.9747	0.9659	0.9433	0.57	0.86
	FFT	ClincalBERT	0.9439	0.9699	0.9519	0.9478	0.33	0.76
	Modif. CBERT	FPBPA(H)	**0.9532**	**0.9717**	**0.9584**	**0.9478**	**0.52**	**0.94**
		FPBPA(Pf)	0.9427	0.9688	0.9556	0.9478	0.34	0.95
		FPBPA(C)	0.9458	0.9677	0.9567	0.9411	0.20	0.82

PICO adapter, the objective was to capture the inherent relationships embedded in PICO tags shown in Table A2 in the appendix. Thus, we implemented a token classification methodology, on the EBM PICO data tags. To mitigate training bias, the zero entity class was strategically excluded from the EBM-PICO dataset, creating an effective learning environment for the remaining relevant classes. Each relevant token is encoded with SciBERT into which the adapter module is integrated for training the EBM-PICO. After this encoding process, the token is directed to a classification layer to be classified into one of the three distinct PICO tags. The training objective for this is optimised using the cross-entropy loss function:

$$L_{\text{PICO}} = -\sum_{i=1}^{N} \left[y_i^{\text{PAR}} \log(\hat{y}_i^{\text{PAR}}) + y_i^{\text{INT}} \log(\hat{y}_i^{\text{INT}}) + y_i^{\text{OUT}} \log(\hat{y}_i^{\text{OUT}}) \right] \quad (2)$$

where $y_i^{\text{PAR}}, y_i^{\text{INT}}, y_i^{\text{OUT}}$, $\hat{y}_i^{\text{PAR}}, \hat{y}_i^{\text{INT}}, \hat{y}_i^{\text{OUT}}$ represent the ground-truth labels and the corresponding predicted probability distributions for the i i-th token, respectively, within the categories of Participants, Intervention/Comparison, and Outcome, and N encapsulates the cumulative count of tokens within the dataset.

3.4 Fusing the trained Knowledge Layers/Adapters

To address **RQ3**, we integrate and tune the trained adapters in SciBERT, PMBERT, BioBERT and CBERT to show transferability. The individually trained

adapters for PICO, PubMedQA, and BioASQ encapsulate different facets of information, each with its relevance to the downstream abstract task. Thus, to leverage the variability in the information stored by each trained adapter, we employ AdapterFusion [15]. AdapterFusion functions analogously to the attention layer in a standard transformer model, where the primary output from the PLM operates as the query. In contrast, the outputs from the various adapters act as keys and values. Readers are referred to the work done by [15] for further details. For clarity in this work, the combination of the trained PICO, PubMedQA and BioASQ is referred to as FPBPA.

4 Experimental Setup

Downstream SR Datasets for evaluation. The proposed model was evaluated on four complex highly imbalanced EBM abstract datasets. One of these datasets, the Aceves-Martins_2022 dataset (AM_22) [16], is private focusing on oral health in children and nutritional disparities among prisoners. The remaining datasets; Appenzeller-Herzog_2019 (AH_19), Van-Dis_2019 (VD_20) and Leenars_2019 (LN_19) are publicly available on Github[9]. Each study's research question and inclusion/exclusion criteria were combined to form the "question" and the abstract was the "context". A summary of the datasets is provided in Table A3.

Implementation and Hyperparameters. The AdapterHub[10], HuggingFace library[11], and the PyTorch framework were employed for training the knowledge adapters evaluation. Our experimental setup was done with Nvidia 2080Ti GPUs. To ensure uniform input dimensions during the training of the knowledge layers, we truncated/pad sequences to a consistent length of 512 tokens. We split each of the datasets in Table A2 into 90% train and 10% validation split, to find the optimal hyperparameters. In training the PICO adapter, we deployed the following hyper-parameters; warmup step: $[0, \mathbf{500}, 1000]$, epochs: $[3, \mathbf{5}, 10, 20]$, batch size: $[8, \mathbf{16}, 64, 256]$, weight decay: $[\mathbf{0.0}, 0.1, 0.01, 0.001]$ and learning rate: $[\mathbf{1}e^{-4}, 3e^{-5}, 1e^{-5}]$ with the AdamW as the optimizer. Similarly, the same hyper-parameters were in training in PubMedQA and BioASQ adapters. However, the best-performing batch size and epochs for the PubMedQA were **64** and **3**, whereas the best-performing learning rate for the BioASQ was $3e^{-5}$. We modulated three random seeds (42, 10 and 50) and reported on the aggregated results over the iterations to ensure robustness.

[9] https://github.com/asreview/synergy-dataset

[10] https://adapterhub.ml/

[11] https://huggingface.co/docs/transformers/index

[12] Similar to [9], but here ↑ denotes the % increment of the best results compared to the strongest baseline (CODER-BERT).

[12] The **Bold** values represent scenarios where the FPBPA method outperforms the FFT PLMs baselines within its category (SciBERT, BioBERT, PMBERT, CBERT) for the dataset. **Bold** also denotes the overall best value for each metric e.g precision, recall, and F1 in each dataset (LN_19, AH_19). The ↑ denotes the % increment of the best results compared to the strongest baseline (KRISSBERT).

Table 2: Summary of results on the VD_20 (1) and AM_22 (2) datasets.[12]

DB		Methods	Precision	Recall	F1	WSS@95	AUC_PR	ROC
1	Know. Int. Baselines	KRISSBERT	0.9717	0.9785	0.9741	0.9417	0.29	0.91
		CODER-BERT	0.9792	0.9829	0.9743	0.9428	0.36	0.87
	FFT	SciBERT	0.9718	0.9735	0.9783	0.9456	0.26	0.62
	Modif. SciBERT	FPBPA(H)	0.9760	0.9813	0.9750	**0.9464**	**0.31**	0.8
		FPBPA(Pf)	**0.9787**	**0.9818**	**0.9786**	0.9302	**0.31**	0.85
		FPBPA(C)	0.9725	0.9702	0.9714	0.9467	**0.31**	**0.89**
	FFT	PMBERT	0.9670	0.9685	0.9677	0.9329	**0.35**	**0.95** ↑ 0.08
	Modif. PMBERT	FPBPA(H)	**0.9735**	**0.9779**	**0.9753**	0.9379	0.30	0.87
		FPBPA(Pf)	0.9716	0.9791	0.9739	**0.9434**	0.27	0.89
		FPBPA(C)	0.9695	0.9768	0.9725	0.9412	0.18	0.75
	FFT	BioBERT	0.9675	0.9618	0.9657	0.9461	0.35	0.76
	Modif. BioBERT	FPBPA(H)	0.9741	0.9807	0.9743	0.9461	0.24	0.76
		FPBPA(Pf)	**0.9809** ↑ 0.17	**0.9873**↑ 0.44	**0.9772** ↑ 0.29	**0.9472** ↑ 0.44	0.25	0.77
		FPBPA(C)	0.9742	0.9807	0.9735	0.9461	**0.38** ↑ 0.02	**0.89**
	FFT	CBERT	0.9663	0.9624	0.9774	0.9445	0.29	0.71
	Modif. CBERT	FPBPA(H)	0.9771	0.9796	0.9782	0.9417	0.29	**0.87**
		FPBPA(Pf)	**0.9774**	**0.9818**	**0.9781**	**0.9447**	**0.30**	0.79
		FPBPA(C)	0.9707	0.9791	0.9732	0.9351	0.25	0.83
2	Know. Int. Baselines	KRISSBERT	0.9935	0.9939	0.9936	0.9393	0.75	0.98
		CODER-BERT	0.9953	0.9954	0.9953	0.9377	0.73	0.94
	FFT	SciBERT	0.9925	0.9931	0.9925	0.9210	0.64	0.89
	Modif. SciBERT	FPBPA(H)	0.9944	0.9946	0.9945	0.9370	0.77	0.97
		FPBPA(Pf)	**0.9956** ↑ 0.03	**0.9959** ↑ 0.05	**0.9959** ↑ 0.06	**0.9485** ↑ 1.08	**0.87** ↑ 1.14	**0.99** ↑ 0.01
		FPBPA(C)	0.9906	0.9916	0.9903	0.9385	0.73	0.92
	FFT	PMBERT	0.9905	0.9904	0.9877	**0.9466**	**0.86**	**0.99**
	Modif. PMBERT	FPBPA(H)	**0.9938**	**0.9935**	**0.9936**	0.9358	**0.85**	**0.99**
		FPBPA(Pf)	0.9928	0.9927	0.9928	0.9366	0.73	0.98
		FPBPA(C)	0.9920	0.9923	0.9910	0.9439	0.73	0.98
	FFT	BioBERT	0.9920	0.9927	0.9919	0.9420	0.80	**0.99**
	Modif. BioBERT	FPBPA(H)	0.9930	0.9935	0.9931	0.9397	0.64	0.92
		FPBPA(Pf)	**0.9944**	**0.9946**	**0.9945**	**0.9431**	0.72	0.96
		FPBPA (C)	0.9933	0.9931	0.9932	0.9362	**0.84**	**0.99**
	FFT	ClincalBERT	0.9941	0.9927	0.9932	0.9328	0.83	**0.99**
	Modif. CBERT	FPBPA(H)	0.9921	0.9925	0.9923	0.9397	0.82	**0.99**
		FPBPA(Pf)	**0.9948**	**0.9950**	**0.9948**	**0.9404**	**0.84**	0.98
		FPBPA (C)	0.9939	0.9943	0.9938	0.9389	0.78	0.93

Evaluation Metrics and Baselines. We report on the weighted average: precision and recall, AUC Precision-recall, AUC ROC and Work saved oversampling (WSS@95%) [5] which measures how much human burden the model can reduce. During the evaluation, we split the downstream dataset into a 60/40 train test set. We applied the back translation augmentation technique described in Section 3.1 only to the minority (include) in the train set. Further, we partitioned the final augmented and initial train sent into a 10% dev set whilst we reported the average runs on the unaugmented test set. To compare the performance of our method, we explore existing FFT proposed for abstract screening tasks. As such, **FFT-PMBERT**, **FFT-SciBERT**, **FFT-BioBERT** and **FFT-CBERT**. To further validate the performance of our model, we compare with two SOTA knowledge integrated PLMs **CODER-BERT** [13], a UMLS triples embedding integration via contrastive learning and **KRISSBERT** [14], a PMBERT that utilises self-supervised learning for entity linking.

5 Results and Discussion

Tables 1 and 2 show the results obtained from evaluating the adapter-based tuning against the FFT biomedical variants PLMs and existing strong knowledge PLM integrated baselines (CODER-BERT and KRISSBERT). Generally, the

[13] https://huggingface.co/GanjinZero/UMLSBert$_ENG$

[14] https://huggingface.co/microsoft/BiomedNLP-KRISSBERT-PubMed-UMLS-EL

tables demonstrate a consistent trend across various PLMs: tuning FPBPA (H, Pf, or C) within the PLMs leads to notable metric improvements compared to the baseline. This finding addresses **RQ3** indicating the effectiveness of FPBPA for the EBM abstract screening task. Further discussion is as follows:
Can adapter-based tuning perform better than SOTA FFT PLMs? Discussing Table 1 for the highly imbalanced ratio (IR) dataset LN_19 (IR 1:341), FPBPA(Pf) consistently achieves high precision, recall, WSS@95, and F1 score compared to the baselines, particularly in PMBERT. Additionally, FPBPA(H) and FPBPA(C) also show competitive performance, especially in terms of precision and AUC_PR. Similarly, for AH_19 (IR 1:98), SciBERT-FPBPA(Pf) consistently outperforms the strongest baseline and FFT PLMs.
In Table 2, for the VD_20 (1: 126) dataset, BioBERT FPBPA(Pf) achieves higher precision, recall, WSS@95 and F1 score compared to the strongest baseline and FFT PLMs. Similarly, for AM_22 (1:188), SciBERT-FPBPA (Pf) outperforms the FFT and the strongest knowledge-integrated baseline across all metrics.
What is the effect of the different configurations of Knowledge Layers? The different FPBPA configurations (H, Pf, C) exhibit variable impacts on different datasets seen in Tables 1 and 2. To summarise the analysis, the FPBPA(Pf) shows strength in the extremely imbalanced datasets compared to the H and C. Thus, in practicality, the use of FPBPA(Pf) may be useful in situations where the EBM to be done is broad and may lead to broad search strings, hence encompassing lots of irrelevant literature compared to the number of relevant as in the case of **LN_19** and **AM_22** dataset.

6 Conclusion and Future Works

This research explores a SOTA transfer learning method that infuses domain-specific insights into PLMs using adapters. Utilizing the PICO framework alongside resources like PubMedQA and BioASQ Q&A, our technique improves PLM capabilities for EBM abstract screening, which is critical for enhancing clinical decisions and policies. Through detailed experiments, we demonstrate that our method delivers promising outcomes across various metrics, including precision, recall, F1 score, and WSS@95. Looking ahead, we plan to incorporate additional domain-specific resources such as UMLS, DisGeNET, and the UNIPROT knowledge database to broaden our approach's relevance. Currently, our research centres on the BERT model, but future investigations will include other SOTA PLMs like GPT and LLaMA. Furthermore, a future work will be to conduct a comparative analysis of our method against baseline models such as SVM and NB +/- UMLS, employing keyword search techniques like cTAKES or a MetaMap-based model using TF-IDF or n-gram analysis.

References

1. P. B. Burns, R. J. Rohrich, and K. C. Chung, "The Levels of Evidence and Their Role in Evidence-Based Medicine," textitPlastic and Reconstructive Surgery, vol. 128, no. 1, pp. 305–310, Jul. 2011.

2. R. Ofori-Boateng, M. Aceves-Martins, C. Jayne, N. Wiratunga, and C. F. Moreno-Garcia, "Evaluation of Attention-Based LSTM and Bi-LSTM Networks For Abstract Text Classification in Systematic Literature Review Automation," textitProcedia Computer Science, vol. 222, pp. 114–126, 2023.
3. Q. Xie, J. A. Bishop, P. Tiwari, and S. Ananiadou, "Pre-trained language models with domain knowledge for biomedical extractive summarization," textitKnowledge-Based Systems, vol. 252, p. 109460, Sep 2022.
4. P. Timsina, J. Liu, and O. El-Gayar, "Advanced analytics for the automation of medical systematic reviews," textitInformation Systems Frontiers, vol. 18, no. 2, pp. 237–252, Aug 2015.
5. H. Almeida, M.-J. Meurs, L. Kosseim, and A. Tsang, "Data sampling and supervised learning for HIV literature screening," textitIEEE Transactions on Nanobioscience, vol. 15, no. 4, pp. 354–361, 2016.
6. G. Kontonatsios, S. Spencer, P. Matthew, and I. Korkontzelos, "Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews," textitExpert Systems with Applications: X, vol. 6, p. 100030, Jul 2020.
7. A. Natukunda and L. K. Muchene, "Unsupervised title and abstract screening for systematic review: a retrospective case-study using topic modelling methodology," textitSystematic Reviews, vol. 12, no. 1, Jan 2023.
8. M. Hasny, A.-P. Vasile, M. Gianni, A. Bannach-Brown, M. Nasser, M. Mackay, D. Donovan, J. Šorli, I. Domocos, M. Dulloo, N. Patel, O. Drayson, N. M. Elango, J. Vacquie, A. P. Ayala, and A. Fogtman, "BERT for Complex Systematic Review Screening to Support the Future of Medical Research," in textitArtificial Intelligence in Medicine, 2023.
9. C. F. Moreno-Garcia, C. Jayne, E. Elyan, and M. Aceves-Martins, "A novel application of machine learning and zero-shot classification methods for automated abstract screening in systematic reviews," textitDecision Analytics Journal, vol. 6, p. 100162, 2023.
10. E. Guo, M. Gupta, J. Deng, Y.-J. Park, M. Paget, and C. Naugler, "Automated Paper Screening for Clinical Reviews Using Large Language Models," 2023.
11. N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-Efficient Transfer Learning for NLP," 2019.
12. M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," in textitICML-2003 Workshop on Learning from Imbalanced Data Sets II, 2003.
13. S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in textitAdvances in Neural Information Processing Systems, vol. 30, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds., 2017.
14. R. K. Mahabadi, J. Henderson, and S. Ruder, "Compacter: Efficient Low-Rank Hypercomplex Adapter Layers," 2021.
15. J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "AdapterFusion: Non-Destructive Task Composition for Transfer Learning," in textitProceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021.
16. M. Aceves-Martins, L. López-Cruz, M. García-Botello, Y. Y. Gutierrez-Gómez, and C. F. Moreno-García, "Interventions to Treat Obesity in Mexican Children and Adolescents: Systematic Review and Meta-Analysis," textitNutrition Reviews, vol. 80, no. 3, pp. 544–560, Mar. 2022.

7 Appendix

Table A1: Format of the EBM abstract screening dataset

Abstract (Abs)	**Research Question (RQ)**	**Decision**
1. Glycosylated haemoglobins and weights were recorded for 200 consecutive diabetic...	What is the prevalence of overweight and obesity among imprisoned populations worldwide?	Exclude
2. Childhood dental caries and obesity are prevalent health problems. Results from previous studies of the caries-obesity...	Is there an association between obesity or overweight and poor oral health among Mexican children and adolescents?	Include

Table A2: Statistics of datasets used to train the knowledge layers/adapters

Dataset	**Adapter**	**Format**	**Size**
EBM-PICO[13]	PICO	(I-INT, I-OUT, I-PAR)[14]	5000
PubMedQA	P-QA	(Context/Question/labels(yes/no/maybe))	211.3K
BioASQ	B-ASQ	(Context/Question/labels(yes/no))	6676

Table A3: Summary of the datasets ranging from human to animal study, where IR = Imbalance Ratio, the variables used for each EBM dataset are in Table A1

Name_of_dataset	**Subject**	**Total_papers**	**Relevant**	**Irrelevant**	**IR**	**Abs Len (Avg)**
Aceves-Martins_2022(AM_22)	Nutritional status of prisoners	13022	69	12953	1:188	1765.37
Appenzeller-Herzog_2019(AH_19)	Therapy for Wilson Disease	2873	29	2844	1:98	1282.35
Leenars_2019(LS_19)	Animal to human translation	5812	17	5795	1:341	1458.40
Van_Dis_2020(VD_20)	Cognitive Behavioral Therapy	9128	72	9056	1:126	1473.08

[13] https://github.com/bepnye/EBM-NLP

[14] where Participants is (I-PAR), Outcome (I-OUT), and a combination of Intervention/Comparison as (I-INT)