Prompt Discriminative Language Models for Domain Adaptation

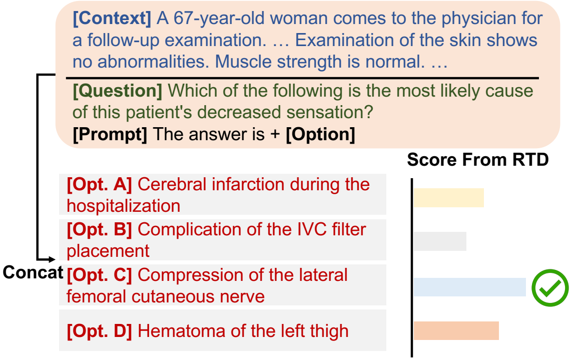
Keming Lu[1], **Peter Potash**[2], **Xihui Lin**[2], **Yuwen Sun**[4], **Zihan Qian**[3], **Zheng Yuan**[5]
Tristan Naumann[2], **Tianxi Cai**[3] and **Junwei Lu**[3]
[1]University of Southern California [2]Microsoft Research [3]Harvard University
[4]Xi'an Jiaotong-Liverpool University [5]Alibaba Damo Academy
[1]keminglu@usc.edu [2]{Peter.Potash,xihlin,tristan}@microsoft.com
[3]{zihanqian,tcai,junweilu}@hsph.harvard.edu
[4]Yuwen.Sun19@student.xjtlu.edu.cn [5]yuanzheng.yuanzhen@alibaba-inc.com

Abstract

Prompt tuning offers an efficient approach to domain adaptation for pretrained language models, which predominantly focus on masked language modeling or generative objectives. However, the potential of discriminative language models in biomedical tasks remains underexplored. To bridge this gap, we develop BIODLM, a method tailored for biomedical domain adaptation of discriminative language models that incorporates prompt-based continual pretraining and prompt tuning for downstream tasks. BIODLM aims to maximize the potential of discriminative language models in low-resource scenarios by reformulating these tasks as span-level corruption detection, thereby enhancing performance on domain-specific tasks and improving the efficiency of continual pertaining. In this way, BIODLM provides a data-efficient domain adaptation method for discriminative language models, effectively enhancing performance on discriminative tasks within the biomedical domain.

1 Introduction

Recent years witnessed the development of biomedical pretrained language models (PLMs) (Kalyan et al., 2022). These domain-specific PLMs contribute to a large number of downstream tasks in the biomedical domain, such as named entity recognition (Yuan et al., 2021; Khandelwal et al., 2022; Watanabe et al., 2022), entity linking (Zhang et al., 2022; Liu et al., 2020), relation extraction (Li et al., 2022a; Sarrouti et al., 2022), and question answering (Jin et al., 2019a; Pappas et al., 2022).



[Context] A 67-year-old woman comes to the physician for a follow-up examination. ... Examination of the skin shows no abnormalities. Muscle strength is normal. ...

[Question] Which of the following is the most likely cause of this patient's decreased sensation?

[Prompt] The answer is + [Option]

Score From RTD
Concat
[Opt. A] Cerebral infarction during the hospitalization
[Opt. B] Complication of the IVC filter placement
[Opt. C] Compression of the lateral femoral cutaneous nerve
[Opt. D] Hematoma of the left thigh

Figure 1: A case for prompting discriminative pretrained language models (DLMs) on multi-choice biomedical question answering. Each option is first concatenated with a predefined hard prompt: "The answer is". They are separately concatenated with the context and question as input. We rank the score from the head of replaced token detection (RTD) in DLMs to determine the best option.

Most existing domain-specific PLMs rely on tremendous in-domain corpus and computing resources for continual pretraining (Lee et al., 2020; Rasmy et al., 2021; Yuan et al., 2022; Alsentzer et al., 2019) or pretraining from scratch (Gu et al., 2021; Yasunaga et al., 2022), which could be infeasible with limited resources. Meanwhile, PLMs for general purposes usually fails to achieve comparable performance on biomedical tasks with finetuning compared with in-domain PLMs at the same model scale (Gu et al., 2021). To combat these issues, exploring a prompt-based domain adaptation method that better leverages existing knowledge learned in pretaining is necessary. Recent research demonstrates that prompts or instructions can activate the hidden abilities of PLMs (Liu et al., 2022; Radford et al., 2019; Brown et al., 2020), including cross-domain inference (Yeh et al., 2022; Fries et al., 2022; Yao et al., 2022b). Therefore, prompt tuning on general PLMs can be a data-efficient domain adaptation method as they are proven promising on various downstream tasks (Wang et al., 2018, 2019).

Existing explorations about prompt-based domain adaptation mainly focus on PLMs with masked language modeling (Lai et al., 2022; Sung et al., 2021) or generative objectives (Luo et al., 2022). However, we identify that discriminative pretrained language models (DLMs) also hold

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 247–258
July 14, 2023 ©2023 Association for Computational Linguistics

great potential for prompt-based domain adaptation but remains understudied. DLMs are pretrained to distinguish between alternatives and proved to be stronger few-short learners than PLMs with other training objectives (Xia et al., 2022). Therefore, DLMs are better choices for domain adaptation since many downstream tasks in the biomedical domain focus on discriminative objectives (Gu et al., 2021). However, complex model architecture and training recipes hinder DLMs from efficient adaptation to other domains.

To shed light on this topic, we develop BIODLM (Prompt-based Biomedical Domain Adaptation for Discriminative Language Models), which can efficiently take advantage of the state-of-the-art DLMs in the general domain. BIODLM is a prompt-based biomedical domain adaptation method designed explicitly for DLMs, including prompt-based continual pertaining and prompt tuning for downstream tasks. Inspired by Xia et al. (2022), we first formulate discriminative downstream tasks in the biomedical domain, such as multi-choice question answering, as span-level corruption detection.

As shown in Fig. 1, this prompt tuning reformulation allows general-domain DLMs to be used as zero-shot or few-shot learners in biomedical tasks, which is also supported by our probing experiments in §4.2. We develop an efficient prompt-based continual pretraining method to further enhance the performance of DLMs on biomedical tasks. As Bajaj et al. (2022) revealed, the selection of corrupted tokens and the corruption methods play a vital role in pretraining DLMs and is highly related to the performance on downstream tasks. BIODLM selects domain-specific words, defined as different vocabulary between in-domain and general models, as corrupted tokens to lead the continual pretraining focusing on new domain knowledge and improve pretraining efficiency. For corruption, BIODLM employs fixed in-domain PLMs as encoders to corrupt selected tokens instead of co-training encoders and decoders in DLMs. BIODLM is a flexible domain adaptation method that can be applied to any existing DLMs.

The contributions of this work are mainly twofold. First, we explore prompt tuning general-domain DLMs on various biomedical downstream tasks, showing prompting DLMs has significant potential on these tasks under low-resource scenarios. Second, we develop a data-efficient continual pretraining method based on replaced token detection, which employs in-domain PLMs as generators to corrupt domain-specific words in the biomedical corpus. In summary, BIODLM efficiently improves low-resource performance on discriminative tasks in the biomedical domain.

2 Related Works

Discriminative PLMs. Discriminative PLMs (DLMs) incorporate replaced token detection (RTD) or other discriminative objectives during pretraining. Clark et al. (2020) first propose a discriminative pretraining method, which trains a generator to create replaced tokens and a discriminator to distinguish between real and replaced tokens. This approach increases the pretraining efficiency by reducing the computation required in the head compared with previous masked language modeling. Meng et al. (2021) further improves the RTD to corrective language modeling, which requires both RTD and language modeling for correcting the replaced tokens. Bajaj et al. (2022) proposes a more stable and efficient training recipe for DLMs. In this work, we explore domain adaptation for these methods in the biomedical domain. We use METRO-LM (Bajaj et al., 2022) in our experiments of BIODLM since it demonstrates the best performance on general benchmarks, such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019).

Prompt tuning for DLMs. Prompt tuning for DLMs is an emerging topic in general and biomedical domains. Ni and Kao (2022) presents empirical evidence showing that ELECTRA can perform well on downstream tasks without fine-tuning or additional training. Xia et al. (2022) introduces a prompt-based fine-tuning approach that leverages discriminative prompts to guide the model towards learning specific downstream tasks with only a few examples. Li et al. (2022b) proposes a few-shot learning approach with pre-trained token-replaced detection models to transform traditional classification and regression tasks into token-replaced detection problems. Yao et al. (2022a) suggests fine-tuning DLMs with prompts for task-specific downstream tasks by adding a small number of task-specific parameters as a prompt to guide the model's output. However, these works are limited to a single method ELECTRA and do not explore biomedical tasks. We follow the recipe of prompt tuning in Xia et al. (2022) and use it on biomedical discriminative tasks.

Biomedical Domain Adaptation. Biomedical domain adaptation of PLMs is a fast-developed topic summarized adequately in the survey from Kalyan et al. (2022). Therefore, we only provide a highly selected review. Alrowili and Vijay-Shanker (2021) propose a novel method for pre-training large biomedical language models that combine BERT, ALBERT, and ELECTRA architectures. Raj Kanakarajan et al. (2021) propose a biomedical domain-specific language encoder model that extends ELECTRA to obtain state-of-the-art performance on numerous biomedical natural language understanding benchmarks. Tinn et al. (2023) propose PubmedELECTRA, a domain-specific version of ELECTRA by continually pertaining ELECTRA on PubMed articles. Luo et al. (2022) propose a generative pre-trained Transformer language model on a large corpus of biomedical articles for biomedical text generation and mining. Our method, BIODLM proposes another perspective that employs prompt-based continual pretraining to adapt DLMs to the biomedical domain, which is understudied in this topic.

3 Methods

We describe preliminaries (§3.1), prompt-based continual pretraining with RTD (§3.2), and prompt tuning for discriminative PLMs (§3.3).

3.1 Preliminaries

Replaced Token Detection. BIODLM is a prompt-based method based on the RTD task. RTD is one of the core pretraining objectives of DLMs (Clark et al., 2020). During the pretaining of DLMs, the input is a sequence of tokens $\mathbf{x} = \{x_i\}_{i=1}^{n}$, where n is the length of input sequences. A random set of tokens in this sequence is selected and corrupted with a generator by masked language modeling. Predictions from the generator will be used to replace the original tokens to obtain a corrupted input $\tilde{\mathbf{x}} = \{\tilde{x}\}_{i=1}^{n}$. At the same time, token-level binary labels are constructed by $\mathbf{y} = \{I(x_i = \tilde{x}_i)\}_{i=1}^{n}$, where $I(\cdot)$ is the indicator function[1]. The discriminator of DLMs is trained with token-level classification on the corrupted input and corresponding labels to detect the replaced tokens.

Method Overview. Similar to the "pretraining-and-finetuning" workflow, BIODLM involves a prompt-based continual pretraining (§3.2) and a prompt-tuning method on downstream tasks (§3.3). As shown in Fig. 2, BIODLM first builds a domain-specific vocabulary for the prompt-based continual pertaining. Then, we corrupt the original biomedical corpus with a fixed in-domain language model as the generator. The corrupted corpus is used to train the general-domain discriminator with RTD for domain adaptation. After the continual pertaining, we explore prompt tuning with RTD to apply BIODLM to biomedical downstream tasks. We reformulate biomedical discriminative tasks into single-token or multi-token RTD, as the example in Fig. 1. BIODLM can also be further tuned on a reformulated training set with RTD objective to enhance downstream performance.

3.2 Prompt-based Continual Pretraining

Continual pretraining on in-domain corpus significantly improve downstream performance on downstream tasks (Gu et al., 2021). However, unlike other training objectives, pretraining with RTD requires self-supervised training corpus construction with corruption. Therefore, we develop a prompt-based continual pretraining method to adapt DMLs to the biomedical domain. The continual pretraining involves a token corruption generator and an RTD discriminator. The recipe of token corruption is essential for both efficiency and effectiveness of the pretraining of DLMs (Bajaj et al., 2022). Therefore, we design a corrupted token selection recipe focusing on in-domain vocabulary and employ fixed in-domain PLMs as generators to corrupt these tokens.

Corrupted Token Selection. Corrupted token selection aims to select the tokens in the in-domain corpus that the generator will corrupt. We first build a domain-specific vocabulary by extracting different tokens from in-domain to general-domain vocabulary. The first challenge is that in-domain and general language models may have very different tokenizers. However, most of them share similar pre-tokenizers to segment context into words. Therefore, we conduct word-level corruption instead of token-level corruption in the traditional design of RTD so that in-domain and general-domain vocabulary can be aligned with each other in the corruption. The detailed selection recipe is described below:

1. We filter tokens that are in in-domain vocabulary but not in the general-domain vocabulary.
2. To conduct word-level corruption, we filter out

[1] The definition of labels may vary in different DLMs. Our introduction follows the recipe in Bajaj et al. (2022).

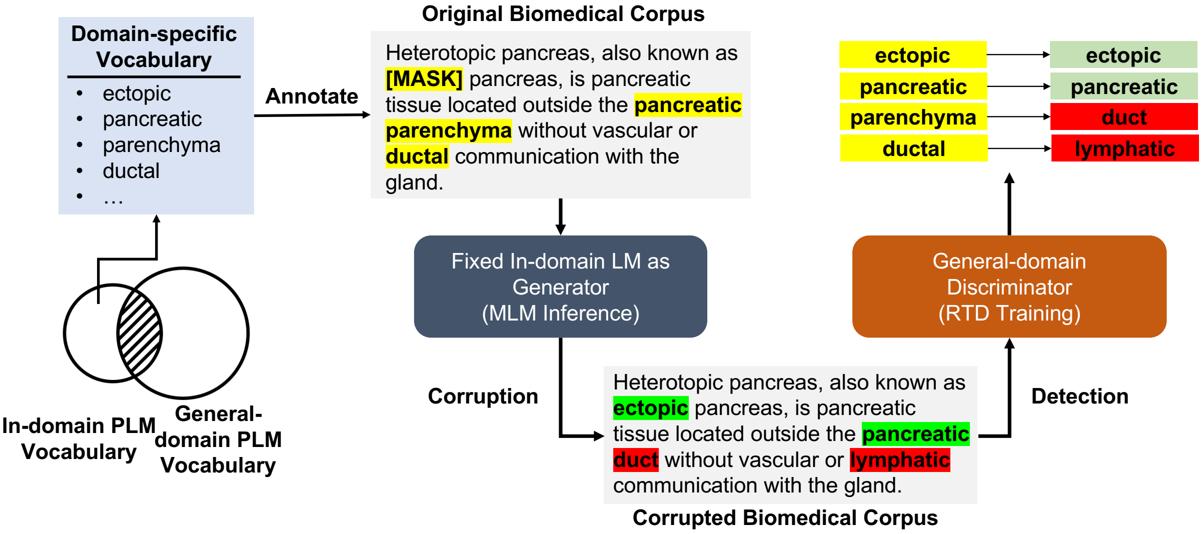

Domain-specific Vocabulary
• ectopic
• pancreatic
• parenchyma
• ductal
• …
Annotate
Original Biomedical Corpus
Heterotopic pancreas, also known as [MASK] pancreas, is pancreatic tissue located outside the pancreatic parenchyma without vascular or ductal communication with the gland.
In-domain PLM Vocabulary
General-domain PLM Vocabulary
Fixed In-domain LM as Generator (MLM Inference)
Corruption
Heterotopic pancreas, also known as ectopic pancreas, is pancreatic tissue located outside the pancreatic duct without vascular or lymphatic communication with the gland.
Corrupted Biomedical Corpus
General-domain Discriminator (RTD Training)
Detection
ectopic
pancreatic
parenchyma
ductal
ectopic
pancreatic
duct
lymphatic

Figure 2: Overview of prompt-based continual petraining in BIODLM. A vocabulary is collected by differing in-domain and general-domain PLMs vocabulary. And we annotate the in-domain corpus with this vocabulary and use this annotation as a set of words for sampling corrupted tokens. Selected tokens are corrupted with a fixed in-domain language model as the generator via masked language modeling inference. The corrupted corpus is then used to continually pretrain a general-domain discriminator with replaced token detection.

all tokens that are not a whole word in the set of tokens we collect in the previous step.

3. We tokenize the remained words in the previous step with the tokenizer of general-domain DLM and filter out any words that contain "unknown" tokens[2]. The rest are our domain-specific vocabulary $\mathcal{D}$.

We use the vocabulary of PubmedBERT (Gu et al., 2021) as our in-domain vocabulary and the vocabulary of MetroLM (Bajaj et al., 2022) as general-domain vocabulary. We eventually have 12,919 words remaining in domain-specific vocabulary $\mathcal{D}$. Most words in $\mathcal{D}$ are biomedical terms, and a sample is listed in §4.3.

Token Corruption. With domain-specific vocabulary $\mathcal{D}$, we employ fixed in-domain LM as a generator to corrupt the in-domain corpus with the inference of masked language modeling. Given an input of the in-domain corpus, such as a PubMed abstract[3], we sample a fixed proportion of words in the input to corrupt. We follow Clark et al. (2020) to set the percentage to 30%. We first pre-tokenize it into words $\mathbf{x} = \{x_i\}_{i=1}^n$, where the length of word sequence is n. Then, we identify any domain-specific words in $\mathcal{D}$, denoting them as a bag of words $\mathcal{C}$. The words for corruption are sampled with a strategy that favors domain-specific words:

- $|\mathcal{C}| > \lfloor 0.3n \rfloor$: We randomly select $0.3n$ words from $\mathcal{C}$ as candidates for corruption.
- $|\mathcal{C}| \leqslant \lfloor 0.3n \rfloor$: We randomly select $\lfloor 0.3n \rfloor - |\mathcal{C}|$ words from the rest of the input to meet the requirement of the proportion of corrupted words.

This strategy ensures domain-specific words will be corrupted first, which leads the pretraining to focus on domain knowledge and enhances pretraining efficiency. After identifying the candidates, each word in the candidates will be replaced with a mask token, such as "[MASK]" in the PubmedBERT, and conduct inference of whole-word masked language modeling with the in-domain PLM. The predictions from the in-domain PLM then replace the words in the original inputs to obtain the corrupted in-domain training corpus.

Training. We use the corrupted biomedical corpus for continual pretraining general-domain discriminators with RTD. We conduct word-level corruption—all tokens in corrupted words are labeled with "replaced" and the rest are "original". Otherwise, continual pretraining is the same as §3.1.

3.3 Prompt Tuning with RTD

We explore prompt tuning with RTD on biomedical downstream tasks in BIODLM. Prompt tuning enables DLMs to conduct low-resource inference and

[2] These "unknown" tokens refer to out-of-vocabulary tokens in the general-domain tokenizer, such as the "[UNK]" token in the MetroLM (Bajaj et al., 2022).

[3] PubMed Official Site: `https://pubmed.gov`

helps DLMs better leverage pretraining knowledge in the general domain. Here, we introduce how to reformulate inputs of biomedical discriminative tasks to conduct low-resource inference with RTD.

Input Reformulation. We follow the recipe from Xia et al. (2022) to prompt DLMs on biomedical downstream tasks. We denote the context as $\mathbf{C}$ and labels as $\mathbf{y} = \{y_i\}_{i=1}^{c}$ of discriminative tasks, where c is the number of labels. We first verbalize labels with predefined words or templates and denote the verbalized templates as $T(\mathbf{y}) = \{t(y_i)\}_{i=1}^{c}$, where $t(\cdot)$ is a manually designed verbalizer for each label. For example, labels from a binary classification task are verbalized as "yes" and "no". As for multi-choice question answering, the labels are already phrases so no verbalization will be applied. Each verbalized label is concatenated with context and a predefined prompt as inputs, denoting as $x = \{C \oplus t(y_i)\}_{i=1}^{c}$, where $\oplus$ is the text concatenation operation. The inputs are fed into the DLMs, and we collect scores from the RTD head within the spans of labels as outputs. The RTD head classifies tokens in labels into "replaced" or "original", where "original" suggests the correct answer to the discriminative problem. The classification scores from the RTD head reveal the semantic correlation between the context and verbalized labels. When verbalized labels are tokenized into more than one token, we use the average RTD scores as the score of these labels. However, the RTD head aims to identify token-level corruption, so averaging multiple tokens do not align well with the pretraining objective and potentially hinders the performance of prompt inference. Therefore, we separately analyze **single-token** and **multi-token** labels in this work. This reformulation allows us to conduct zero-shot prompt inference with DLMs on biomedical discriminative tasks.

Fig. 1 shows a case that we apply prompt inference for DLMs on a multi-choice biomedical question answering dataset. The context is made of a description of the patient background marked in blue and a question marked in green. Then, it is concatenated with four options individually, with a predefined prompt, "The answer is". We consider the average RTD score in each option span as the classification score. And we select the option with the highest average RTD score as the prediction.

Training. In addition to the zero-shot inference, we also conduct prompt tuning on downstream tasks. With the input reformulation described before, discriminative tasks can be reformulated as multi-label binary classification tasks. We further tune the parameters of DLMs in this way to conduct few-shot and fully supervised inference.

4 Experiments

This section introduces an experimental evaluation of prompting discriminative PLMs for biomedical domain adaptation. We describe the experimental setup (§4.1), main results (§4.2), and ablation study (§4.3) on incorporated techniques.

4.1 Experimental Setup

Training corpus. The biomedical corpus for the continual pretraining in this work is the PubMed abstracts in the PubMed Central (PMC) Open Access (OA) Subset[4] (Gamble, 2017; Bethesda, 2003). We process this dump with the open-source tool *pubmed_parser*[5] (Achakulvisut et al., 2020) to extract abstracts of articles. We then follow the preprocessing recipe of Bajaj et al. (2022) and segment the corpus into paragraphs. The original PMC OA Subset contains 21 million paragraphs from biomedical journal articles. We only randomly select three million paragraphs for continual pretraining due to the limitation of computation resources.

Benchmarks. We evaluate BioDLM on five public biomedical datasets: (1) **PubmedQA** (Jin et al., 2019b) contains 1k expert-labeled question-answer pairs based on PubMed abstracts with yes/no/maybe multiple-choice answers. (2) **BioASQ** (Tsatsaronis et al., 2012) is a large question-answering dataset containing biological questions and answers, and related biomedical papers and abstracts. (3) **MedQA(USMLE)** (Jin et al., 2021) is a question-answering dataset containing multiple-choice questions and related answer options in US Medical License Exam (USMLE) format, which were obtained with a choice of 4 or 5 possible answers from the National Medical Board Examination in the United States. (4) **MMLU (Professional Medicine)** (Hendrycks et al., 2020) involves difficult exam questions consisting of four multiple-choice questions with corresponding answers in the biomedical domain. (5) **MedMCQA** (Pal et al., 2022) is a new large-scale

[4] PMC OA Subset:https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[5] Github repository of *pubmed_parser*: https://github.com/titipata/pubmed_parser

	Dataset	Size	**Random**	**Zero-shot (RTD Prompt)**			**Fully Supervised (CLS Finetuning)**			
				MetroLM	Electra	BioElectra	MetroLM	Electra	BioElectra	PubmedBERT
Single	PubmedQA	500	33.3	**64.0**	58.0	48.0	63.8	57.0	62.2	55.8
	BioASQ	140	50.0	**74.3**	73.6	67.1	94.3	73.6	75.7	87.6
Multi	MedQA(USMLE)	1273	25.0	**25.3**	22.1	19.5	28.1	27.4	40.3	39.3
	MMLU	272	25.0	**25.7**	19.9	20.6	27.6	25.8	44.1	29.1
	MedMCQA*	4183	25.0	26.6	**26.8**	20.7	35.5	34.8	40.8	41.2
	Macro Avg.		31.7	**43.2**	40.1	35.2	49.8	43.7	52.6	50.6

Table 1: Probing experiment results display the zero-shot performance with the RTD prompt of various DLMs on the test sets of our benchmark. We also report the CLS-based finetuning performance in the full training setting and involve an in-domain PLM, PubmedBERT, for comparison. We report accuracy on each data split and the macro average accuracy on our benchmark. The best zero-shot performance on each dataset is marked in **bold**. * We report performance on the development set of MedMCQA since we have not received official scores on the test set.

	Dataset	**Random**	**0% (Zero-shot)**			**10% (Few-shot)**			**100% (Full)**		
			CLS	Prompt	BioDLM	CLS	Prompt	BioDLM	CLS	Prompt	BioDLM
Single	PubmedQA	33.3	31.1	**64.0**	57.0	56.0	**62.8**	58.0	63.8	**69.9**	66.8
	BioASQ	50.0	35.2	74.3	**77.1**	77.9	77.9	**80.0**	**94.3**	85.3	89.8
Multi	MedQA(USMLE)	25.0	9.8	25.3	**27.7**	26.5	25.7	**29.1**	28.1	27.0	**29.6**
	MMLU	25.0	11.0	25.7	**26.8**	21.7	30.8	**32.7**	27.6	31.2	**31.6**
	MedMCQA*	25.0	6.4	26.6	**27.4**	**30.7**	22.9	30.1	**35.5**	27.2	33.2
	Macro Avg.	33.7	18.7	43.1	**43.4**	42.6	44.1	**49.9**	50.0	48.1	**50.2**

Table 2: Results of BioDLM in the zero-shot, few-shot, and full settings compared with finetuning CLS representations on the test sets of our benchmark. We use MetroLM as the backbone in BioDLM for results in this table. The prompt baseline is MetroLM with RTD prompt tuning without continual pretraining in BioDLM. We report accuracy on each data split and the macro average accuracy on our benchmark. Finetuning CLS requires the training of a classification head, so we conduct zero-shot inference of CLS representations by semantic matching between context and options. The best accuracy on each dataset in each setting is marked in **bold**. * We report performance on the development set of MedMCQA since we have not received official scores on the test set.

Multiple-Choice Question Answering dataset containing about 194k 4-option multiple-choice questions from Indian medical entrance exams (AIIMS/NEET). In our benchmark, PubmedQA and BioASQ are single-token datasets as their labels are short as “yes/no/maybe”. However, other multi-token datasets, such as MedQA(USMLE), are more challenging since they have longer options and at least four options. We report accuracy scores on the test sets. And we only report the performance on the development set on MedMCQA since we have not received official feedback for the test scores.

Baselines. In the probing experiments, we consider Electra and BioElectra as baselines for MetroLM. We also include PubmedBERT for reference. **Electra** (Clark et al., 2020) is a PLM that uses replaced token detection as a self-supervised task for language representation learning. The central concept of Electra is to train a text encoder to identify input tokens from high-quality negative samples generated by a small generator network, resulting in superior performance on downstream tasks compared to conventional masked language modeling. **BioELECTRA** (Raj Kanakarajan et al., 2021) is a biomedical PLM adapted from the ELECTRA model for the biomedical domain. It is pretrained from scratch on the biomedical domain-specific text and achieves state-of-the-art performance on various biomedical NLP tasks, demonstrating that pretraining from scratch with biomedical domain text enhances the model’s capacity. **PubMedBERT** (Gu et al., 2021) is a biomedical PLM that has been pretrained on PubMed abstracts. It achieves state-of-the-art results in several benchmark datasets, making it a strong baseline model for biomedical language understanding tasks. We also include a random baseline, which is the accuracy based on a random guess.

Configurations. We develop BioDLM based on a strong discriminative pretrained language model *MetroLM-base* (Bajaj et al., 2022). This DLM demonstrates the best zero-shot performance in our probing experiments described in §4.2. We run continual pretraining on 8 NVIDIA V100 GPUs for 10 hours and evaluation on each dataset in our benchmark on 1 NVIDIA V100 GPU for less than 1 hour. The hyper-parameters are determined with the grid search based on the accuracy of the development set. Detailed hyper-parameters are shown in Appx. §A.

4.2 Results

We first show the results of a probing experiment to demonstrate DLMs are zero-shot learners on biomedical tasks. Then we present our main results to show the effectiveness of BIODLM on both full- and low-resource scenarios on our benchmark.

Probing Experiments. Tab. 1 shows the probing experiment results about the zero-shot performance of three DLMs with RTD prompt. We also report finetuning results based on the CLS representations of these DLMs, along with the zero-shot prompt tuning performance. First, we notice that models with zero-shot RTD prompt tuning even outperform their finetuning counterparts on several datasets, marked with the underline in Tab. 1. For example, the accuracy of MetroLM with zero-shot RTD prompt tuning in PubmedQA is 64.0, 0.2 absolute percentage higher than its fully supervised finetuning counterpart. Similar cases are also witnessed in other DLMs, such as Electra on the test split of PubmedQA and BioElectra on the development set of BioASQ. These cases show that prompt tuning of general-domain DLMs has great potential as zero-shot learners on biomedical tasks. And these results also provide evidence that reformulating biomedical discriminative tasks as replaced token detection contributes to leveraging general-domain knowledge in pertaining, which is proposed in §3.3. Furthermore, MetroLM significantly outperforms other DLMs on most datasets, achieving 43.2 macro average accuracy. Therefore, we choose MetroLM as the backbone to conduct the following experiments and analyses of BIODLM. Tab. 8 in Appx. §B is an extended version of Tab. 1 containing results on both development and test sets.

Main Results. Tab. 2 shows the main results of BIODLM in the zero-shot, few-shot, and fully supervised settings on the test sets of our benchmark. In the zero-shot setting, BIODLM outperforms MetroLM with only prompt tuning on most datasets, improving macro average accuracy by 0.3 percent. We conduct zero-shot inference with CLS representations by semantic matching between context and options based on CLS representations. However, it can not perform well in the zero-shot setting since the context and options are significantly different. In the few-shot setting, the macro average accuracy of BIODLM is higher than CLS and prompt methods by 7.3% and 5.8%, respectively. These results prove that BIODLM enables general-domain DLMs to conduct inference on biomedical downstream tasks under low-resource scenarios. Furthermore, even though the traditional finetuning method outperforms prompt tuning in the fully supervised setting by 1.9% in accuracy, we notice BIODLM still slightly outperforms the finetuning method by 0.2% on macro average accuracy. This observation suggests that BIODLM benefits from the prompt-based continual pertaining. And we summarize that BIODLM is a better choice under low-resource scenarios, but both traditional CLS finetuning and BIODLM perform well with adequate supervision.

4.3 Study

We provide the following analyses to evaluate further the core components of BIODLM, including corruption methods, prompt templates, and domain-specific vocabulary.

Corruption Methods. In this analysis, we conduct ablation study experiments to demonstrate the effectiveness and data efficiency of the corruption method proposed in BIODLM. We design a random strategy that randomly selects 30% words in the input as the baseline of the domain-specific token selection strategy for corruption. As for the generator, we use the general-domain pretrained language model BERT as the baseline of the in-domain pretrained language model PubmedBERT. We conduct continual pretraining on different combinations of corrupted token selection and generators with 1 million to 3 million samples.

Tab. 3 shows the results of the ablation study on corruption models of BIODLM. Comparing the random and domain-specific token selection, we notice the macro average accuracy on the benchmark of the domain-specific strategy is consistently higher than that of the random strategy. Within each corrupted token selection strategy, PubmedBERT, as the generator, outperforms BERT in most cases, showing that fixed in-domain PLMs with more precise corruption benefit the continual pretraining in BIODLM. Furthermore, we notice the domain-specific token selection strategy with Pub-

Token Selection	Generator	Pretraining Samples		
		1M	2M	3M
Random	BERT	49.3	49.6	50.0
	PubmedBERT	49.3	49.5	49.9
Domain-specific	BERT	49.7	50.4	51.8
	PubmedBERT	50.2	50.9	52.1

Table 3: Ablation study on corruption methods. We compare two token selection recipes based on Random and In-domain vocabulary and two generators BERT (general-domain) and PubmedBERT (in-domain), with pertaining samples from 1 million to 3 million. We report macro average accuracy scores on our benchmark.

medBERT as the generator used in BIODLM with only 1 million training samples can outperform the random token strategy with 3 million training samples. This result provides valuable insight that corruption methods in BIODLM can significantly improve data efficiency in continual pretraining.

Prompt Templates. Prompt templates play a vital role in prompt tuning. We adopt manually designed prompt templates in BIODLM to verbalize labels and reformulate inputs of discriminative tasks. To better evaluate the influence of manual template design, we construct three prompt templates for two biomedical question answering datasets:

- **Template A**: "[Context]. [Question]? The answer is [prompt label]."
- **Template B**: "[Context] [Question]? The answer is [prompt label]."
- **Template C**: "Context: [Context]. Question: [Question]? The answer is [prompt label]."

There are only minor differences among these templates. Using each prompt template, we then run zero-shot inference with MetroLM and RTD prompt on two datasets. Tab. 5 shows that the design of prompt templates may influence zero-shot performance, which could be related to the specific dataset. It is worth noticing that prompt template B only slightly differs from prompt template A but performance on the test set of BioASQ dropped by half, suggesting an obvious spurious correlation on the punctuation in prompt templates.

We also conduct additional prompt ablation studies on the multi-token prompt datasets. We have manually designed two prompts for multi-choice question-answering datasets in our benchmark:

- **Template D**: "[Context]. [Question]? The answer is [Option]."

Prompt	MedQA(USMLE)		MMLU		MedMCQA
	dev	test	dev	test	dev
D	27.6	25.3	38.7	25.7	26.6
E	25.9	25.1	31.2	23.9	24.0

Table 4: Zero-shot accuracy of MetroLM with RTD prompting on multi-token prompt datasets with two manually designed prompt templates.

Prompt	PubmedQA		BioASQ	
	dev	test	dev	test
A	50.2	64.0	72.0	74.3
B	50.2	64.0	62.9	38.7
C	48.6	68.0	71.7	72.3

Table 5: Zero-shot accuracy of MetroLM with RTD prompting on BiomedQA and BioASQ with three manually designed prompt templates.

- **Template E**: "[Context] [Question]? The answer [Option] is [right/wrong]."

The underlined spans include tokens for the RTD. Template D is used in our main results, while template E reformulates multi-token prompts into single-token prompts by simply judging whether the option is right or wrong. Tab. 4 shows the results of these two templates. Template D consistently outperforms template E, suggesting direct RTD on the option spans works better in our multi-token prompt datasets. Therefore, prompt templates need to be carefully designed to achieve the best performance on each dataset.

Domain-specific Vocabulary. We present a brief case study of vocabulary differences between in-domain and general-domain PLMs to justify our design in the corrupted token selection. §4.3 shows cases in the domain-specific vocabulary and their corresponding categories. Most words in this vocabulary fall into categories such as Gene, Protein, Disease, Chemical, and Drug. These categories contain rich biomedical terms frequently used in the downstream tasks. Therefore, continual pretraining on the domain-specific vocabulary helps DLMs focus on biomedical knowledge and improves data efficiency of domain adaptation.

5 Conclusion

We study an efficient way to adapt general-domain DLMs to the biomedical domain and propose BIODLM. BIODLM consists of data-efficient continual pretraining that focuses on domain-specific vocabulary and leverages domain knowledge in the

Categories	Words
Gene & Protein	TGFβ1, IGF1R, phosphatases,Synaptophysin
Disease	Adenomatous,malarial, atherosclerotic, cholangiocarcinoma
Chemical & Drug	Phosphatidylcholine, cycloheximide,azithromycin, minocycline,hygromycin, Methylprednisolone

Table 6: A case study of domain-specific vocabulary used for continual pretraining. We present randomly-selected words and their categories in this vocabulary.

in-domain PLMs by employing them as RTD generators. We also conduct experiments on a biomedical benchmark with six biomedical datasets, verifying that prompt tuning is an effective way to adapt DLMs on biomedical discriminative tasks directly. Future works include extending BioDLM to more DLMs, such as ELECTRA (Clark et al., 2020) and COCO-LM (Meng et al., 2021), and experimenting with BioDLM on other discriminative tasks in the biomedical domain.

Limitations

BioDLM adopts DLMs as backbone models. Compared to PLMs with other training objectives, DLMs may miss language modeling benefits and squeeze representation space. Besides, our benchmarks can be extended to more biomedical discriminative tasks, such as relation extraction, document classification, and entity disambiguation. We consider extending our exploration to more DLMs and biomedical tasks as valuable future works.

Ethics Statement

All datasets in our benchmark and continual pretraining are obtained according to each dataset's respective data usage policy.

References

Titipat Achakulvisut, Daniel Acuna, and Konrad Kording. 2020. Pubmed parser: A python parser for pubmed open-access xml subset and medline xml dataset xml dataset. *Journal of Open Source Software*, 5(46):1979.

Sultan Alrowili and K Vijay-Shanker. 2021. Biomtransformers: building large biomedical language models with bert, albert and electra. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Payal Bajaj, Chenyan Xiong, Guolin Ke, Xiaodong Liu, Di He, Saurabh Tiwary, Tie-Yan Liu, Paul Bennett, Xia Song, and Jianfeng Gao. 2022. Metro: Efficient denoising pretraining of large scale autoencoding language models with model generated signals. *arXiv preprint arXiv:2204.06644*.

Bethesda. 2003. PMC Open Access Subset. *National Library of Medicine*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, et al. 2022. Bigbio: A framework for data-centric biomedical natural language processing. *Advances in Neural Information Processing Systems*, 35:25792–25806.

Alyson Gamble. 2017. Pubmed central (pmc). *The Charleston Advisor*, 19(2):48–54.

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 (HEALTH)*, 3(1):1–23.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019a. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019b. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Anshita Khandelwal, Alok Kar, Veera Raghavendra Chikka, and Kamalakar Karlapalem. 2022. Biomedical NER using novel schema and distant supervision. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 155–160, Dublin, Ireland. Association for Computational Linguistics.

Zhaohong Lai, Biao Fu, Shangfei Wei, and Xiaodong Shi. 2022. Continuous prompt enhanced biomedical entity normalization. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part II*, pages 61–72. Springer.

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.

Lishuang Li, Ruiyuan Lian, Hongbin Lu, and Jingyao Tang. 2022a. Document-level biomedical relation extraction based on multi-dimensional fusion information and multi-granularity logical reasoning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2098–2107, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zicheng Li, Shoushan Li, and Guodong Zhou. 2022b. Pre-trained token-replaced detection model as few-shot learner. *arXiv preprint arXiv:2203.03235*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.

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).

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34:23102–23114.

Shiwen Ni and Hung-Yu Kao. 2022. Electra is a zero-shot learner, too. *arXiv preprint arXiv:2207.08141*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Dimitris Pappas, Prodromos Malakasiotis, and Ion Androutsopoulos. 2022. Data augmentation for biomedical factoid question answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 63–81, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kamal Raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Mourad Sarrouti, Carson Tao, and Yoann Mamy Randriamihaja. 2022. Comparing encoder-only and encoder-decoder transformers for relation extraction from biomedical texts: An empirical study on ten benchmark datasets. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 376–382, Dublin, Ireland. Association for Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4).

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*. Arlington, VA: Citeseer.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Taiki Watanabe, Tomoya Ichikawa, Akihiro Tamura, Tomoya Iwakura, Chunpeng Ma, and Tsuneo Kato. 2022. Auxiliary learning for named entity recognition with multiple auxiliary biomedical training data. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 130–139, Dublin, Ireland. Association for Computational Linguistics.

Mengzhou Xia, Mikel Artetxe, Jingfei Du, Danqi Chen, and Ves Stoyanov. 2022. Prompting electra: Few-shot learning with discriminative pre-trained models. *arXiv preprint arXiv:2205.15223*.

Yuan Yao, Bowen Dong, Ao Zhang, Zhengyan Zhang, Ruobing Xie, Zhiyuan Liu, Leyu Lin, Maosong Sun, and Jianyong Wang. 2022a. Prompt tuning for discriminative pre-trained language models. *arXiv preprint arXiv:2205.11166*.

Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. 2022b. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. *arXiv preprint arXiv:2211.10265*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Hui-Syuan Yeh, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Decorate the examples: A simple method of prompt design for biomedical relation extraction. *arXiv preprint arXiv:2204.10360*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, Online. Association for Computational Linguistics.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880.

A Hyper-parameters

Tab. 7 shows details of hyper-parameters in the experiments of continual pretraining and prompt tuning. Hyper-parameters are determined by grid search.

B Comprehensive Results

We demonstrate extensive results, including performance on development sets in Tab. 8 and Tab. 9.

Parameters	Evaluation				Continual Training
	PubmedQA	BioASQ	MedQA(USMLE)	MedMCQA	PubmedQA
Batch Size	8	8	32	32	8
Learning Rate	2e-5	2e-5	5e-5	5e-5	2e-5
Warmup Steps	500	500	1000	1000	100
Epochs	20	20	10	10	1
Max Sequence Length	512	512	512	512	512

Table 7: Hyper-parameters used for BioDLM evaluation and continual training on PubmedQA, BioASQ, MedQA(USMLE), and MedMCQA.

	Dataset	Split	Size	**Random**	**Zero-shot (RTD Prompt)**			**Fully Supervised (CLS Finetuning)**			
					MetroLM	Electra	BioElectra	MetroLM	Electra	BioElectra	PubmedBERT
Single	PubmedQA	dev	50	33.3	**50.2**	46.9	46.4	62.0	56.0	54.0	52.3
		test	500	33.3	**64.0**	58.0	48.0	63.8	57.0	62.2	55.8
	BioASQ	dev	75	50.0	72.0	78.6	**82.7**	93.3	85.3	81.3	89.3
		test	140	50.0	**74.3**	73.6	67.1	94.3	73.6	75.7	87.6
Multi	MedQA(USMLE)	dev	1272	25.0	**27.6**	24.8	18.2	28.5	27.8	43.5	36.8
		test	1273	25.0	**25.3**	22.1	19.5	28.1	27.4	40.3	39.3
	MMLU	dev	31	25.0	**38.7**	29.0	16.1	25.8	29.7	45.2	32.2
		test	272	25.0	**25.7**	19.9	20.6	27.6	25.8	44.1	29.1
	MedMCQA	dev	4183	25.0	26.6	**26.8**	20.7	35.5	34.8	40.8	41.2
	Macro Avg.			32.4	**44.9**	42.2	37.7	51.0	46.4	54.1	51.5

Table 8: Probing experiment results display the zero-shot performance with the RTD prompt of various DLMs on our benchmark. We also report the CLS-based finetuning performance of these DLMs in the full training setting and involve an in-domain PLM, PubmedBERT, for comparison. We report accuracy on each data split and the macro average accuracy on our benchmark. The best zero-shot performance on each dataset is marked in **bold**.

	Dataset	Split	**Random**	**0% (Zero-shot)**			**10% (Few-shot)**			**100% (Full)**		
				CLS	Prompt	BioDLM	CLS	Prompt	BioDLM	CLS	Prompt	BioDLM
Single	PubmedQA	dev	33.3	28.7	50.2	**52.4**	48.4	62.0	**66.0**	62.0	58.7	**66.0**
		test	33.3	31.1	**64.0**	57.0	56.0	**62.8**	58.0	63.8	**69.9**	66.8
	BioASQ	dev	50.0	34.0	72.0	**75.0**	**89.3**	87.9	88.0	**93.3**	90.6	90.7
		test	50.0	35.2	74.3	**77.1**	77.9	77.9	**80.0**	**94.3**	85.3	89.8
Multi	MedQA(USMLE)	dev	25.0	10.4	**27.6**	26.4	25.6	**28.3**	27.9	28.5	25.4	**29.7**
		test	25.0	9.8	25.3	**27.7**	26.5	25.7	**29.1**	28.1	27.0	**29.6**
	MMLU	dev	25.0	4.2	38.7	**39.4**	29.0	29.3	**32.9**	25.8	30.9	**35.4**
		test	25.0	11.0	25.7	**26.8**	21.7	30.8	**32.7**	27.6	31.2	**31.6**
	MedMCQA	dev	25.0	6.4	26.6	**27.4**	**30.7**	22.9	30.1	**35.5**	27.2	33.2
	Macro Avg.		32.4	19.0	44.9	**45.5**	45.0	47.5	**49.4**	51.0	49.6	**52.5**

Table 9: Results of BioDLM in the zero-shot, few-shot, and full settings compared with finetuning CLS representations. We use MetroLM as the backbone in BioDLM for results in this table. The prompt baseline is MetroLM with RTD prompt tuning but without continual pretraining in BioDLM. We report accuracy on each data split and the macro average accuracy on our benchmark. Finetuning CLS requires the training of a classification head, so it is infeasible in the zero-shot setting. The best accuracy on each dataset in each setting is marked in **bold**.