Improving Referring Ability for Biomedical Language Models

Anonymous ACL submission

Abstract

Existing auto-regressive large language models (LLMs) are primarily trained using documents from general domains. In the biomedical domain, continual pre-training is a prevalent method for domain adaptation to inject professional knowledge into powerful LLMs that have been pre-trained in general domains. Previous studies typically conduct standard pre-training by randomly packing multiple documents into a long pre-training sequence. Recently, some existing works suggest that enhancing the relatedness of documents within the same pre-training sequence may be advantageous. However, these studies primarily focus on general domains, which cannot be readily applied in the biomedical domain where the distinction of fine-grained topics is harder. Is it possible to further improve the pre-training for biomedical language models (LMs) using exactly the same corpus? In this paper, we explore an improved approach to continual pre-training, which is a prevalent method for domain adaptation, by utilizing information from the citation network in this challenging scenario. Empirical studies demonstrate that our proposed LinkLM data improves both the intra-sample and inter-sample referring abilities of auto-regressive LMs in the biomedical domain, encouraging more profound consideration of task-specific pre-training sequence design for continual pre-training.

1 Introduction

Pre-trained language models (PLMs) benefit from large-scale, readily accessible, unsupervised texts. Particularly in the biomedical domain, numerous studies conducted pre-training on academic papers and abstracts to enhance representations and professional knowledge (Gu et al., 2021; Beltagy et al., 2019; Bolton et al., 2024). Most of them are encoder-based language models (Ho et al., 2024). With the development of auto-regressive language models (LMs), numerous studies have demonstrated their superior generalization ability and performance compared to encoder-based PLMs when the models are sufficiently large (Brown et al., 2020; Ouyang et al., 2022; Taylor et al., 2022). They can not only understand instructions or background information provided in the context, which can be considered as the *intra-sample referring ability* (as shown in Figure 1), but also adapt to new tasks by referring several provided demonstrations, which can be regarded as the *inter-sample referring ability*. Moreover, with the advent of remarkable open-sourced large language models (LLMs), such as the Llama family (Touvron et al., 2023a,b), researchers turn to explore the possibility of conducting continual pre-training to develop LLMs tailored for specific-domains (Chen et al., 2023; Huang et al., 2023; Wu et al., 2024).

Several pre-training methods have been proposed for encoder-based models, including masked

```
# PubMedQA
Abstract:
To examine patterns of knowledge and attitudes among
adults aged>65 years unvaccinated for influenza. […]
Question:
Do patterns of knowledge and attitudes exist among
unvaccinated seniors?
Answer: yes
```

```
# MedMCQA
Question:
In a 6-month-old child, thick curd like white patch
appears on the buccal mucosa. On rubbing it leaves an
erythematous patch. Most likely diagnosis is:
A. Tuberculosis
B. Lichen planus
C. Lupus erythematous
D. Candidiasis
Answer: Candidiasis
```

Figure 1: Examples of PubMedQA and MedMCQA datasets. PubMedQA requires intra-sample referring ability, whereas MedMCQA mainly measures acquired knowledge from the LM itself or needs to refer to few-shot examples (inter-sample referring).

language modeling, next sentence prediction (Devlin et al., 2019), document relation prediction (Yasunaga et al., 2022), translation language modeling (CONNEAU and Lample, 2019). These methods have effectively helped in learning specific knowledge and significantly promoted the development of encoder-based LMs. However, to the best of our knowledge, most auto-regressive LMs adhere to a conventional method for preparing input sequences for pre-training or continual pre-training, which involves first shuffling the corpora, followed by the random packing (concatenation) of documents until the concatenated sequence reaches the prescribed maximum input length (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023a; Chen et al., 2023).

Recently, some studies demonstrate that the standard pre-training method for auto-regressive LMs can be further improved by designing appropriate pre-training sequences (Levine et al., 2021; Gu et al., 2023; Shi et al., 2023; Zhao et al., 2024), such as incorporating relevant texts into the preceding context. LinkBERT (Yasunaga et al., 2022) constructs three types of segment pairs based on a citation network to classify whether they are continuous, linked, or random, motivating models to capture the citing relationship between two text segments. Considering its success, we consider whether this methodology can be extended to auto-regressive LMs, helping them learn to capture relationships between multiple text segments and improving their referring ability. Therefore, in this paper, we explore the linking information from the citation network to construct sequences for training an auto-regressive LM, which we call it as LinkLM. Specifically, we design the pre-training sequences by organizing the documents based on their citing relationships. When optimizing the language modeling objective, auto-regressive LMs can learn to refer to possible information from the previous context. As illustrated in Figure 2, when predicting the tokens in the abstract D_1^1 (<PMID 37893869>), models can access information from its citing papers, learning from the findings about other detection tools (e.g., ENFEN Battery in D_2^1) and different aspects (e.g., neurobiology in D_2^2). Furthermore, by referring D_2^1, D_3^1, and D_4^1, we can understand Attention Deficit Hyperactivity Disorder (ADHD) with a series of related works along the science history. Therefore, training with LinkLM data encourages LMs to refer to necessary information from the previous context, and therefore enhances models' referring ability, which can be used in tasks such as open-book question answering (Mihaylov et al., 2018; Jin et al., 2019) and the In-Context Learning (ICL) setting (Dong et al., 2022).

Though the success of constructing appropriate pre-training sequences has been revealed by some previous works (Gu et al., 2023; Shi et al., 2023; Zhao et al., 2024), they primarily focus on general domains where the distinction of topics is less challenging than that in the biomedical domain. Additionally, they only trained their models from scratch. However, after pre-training with large-scale, randomly concatenated documents, LMs may tend to avoid breaking document boundaries (i.e., $[EOS]$ token) to refer to adjacent concatenated documents. Whether the conclusion still holds under the continual pre-training scenario is not clear. Since continual pre-training is a prevalent practice for developing biomedical LLMs, we focus on this setting in our experiments.

In summary, our contributions are threefold:

- We propose a novel algorithm for pre-training sequence design exploiting citation information from a citation network to improve referring ability for biomedical language models.
- Our empirical studies fill the gaps in previous research, demonstrating that constructing appropriate pre-training sequences is also promising under the continual pre-training setting, improving both intra-sample and inter-sample referring ability of auto-regressive language models.
- Our experiments on one-shot evaluation with retrieved demonstrations show that our method can further boost performance in this scenario, emphasizing the potential of designing task-specific pre-training sequences.

2 Related Work

2.1 Domain Adaptation

Among domain-specific LMs, there are three dominant architectures: encoder-only, encoder-decoder, and decoder-only Transformer (Ho et al., 2024). For encoder-only models, BioLinkBERT (Yasunaga et al., 2022) introduced a pre-training objective, document relation prediction (DRP), to identify whether a pair of segments is contiguous, linked, or random. For encoder-decoder models,

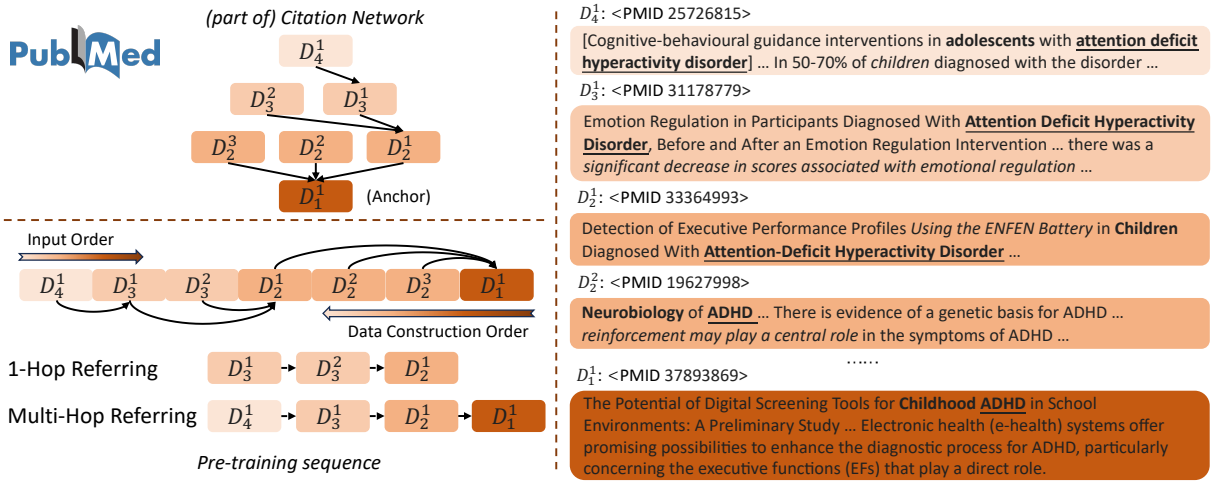


Figure 2: Example of LinkLM data construction. The detailed process is described in Algorithm 1. In this example, the pre-training sequence contains a series of works discussing Attention Deficit Hyperactivity Disorder (ADHD). Training with LinkLM data, models can not only learn to predict an anchor abstract by referring to its citing references, but also benefit from the multi-hop references, which are not linked directly.

BioT5 (Pei et al., 2023) constructed various tasks by incorporating molecule and protein representations into pure texts, learning the relation between biochemistry representations and their surrounding contexts. For decoder-only models, Galactica (Taylor et al., 2022) and Meditron (Chen et al., 2023) carefully processed input texts by inserting the title of the cited paper when the input texts contain citation annotations. This series of work shows that careful design of pre-training input sequences can indeed improve LMs beyond the standard pre-training. However, most of them require fine-grained annotations, which are expensive to collect. Although BioLinkBERT exploited the citation network, it remains unclear whether it is still available and how it can be applied to auto-regressive LMs.

2.2 Pre-training Sequence Design

Recently, in the general domain, some researchers have shown that even without fine-grained annotations, we can still construct meaningful and useful input sequences for pre-training. Levine et al. (2021) proved that by pre-pending semantically related texts based on RoBERTa (Liu et al., 2019) sentence embeddings, sentence representations and open-domain question-answering abilities of auto-regressive LMs can be improved. Gu et al. (2023) trained a task-specific classifier to identify the intrinsic tasks within the pre-training texts and clustered those whose intrinsic tasks are the same into the same context, improving the in-context learning ability of LMs. Shi et al. (2023) retrieved similar texts using Contriever (Izacard et al., 2022) and concatenated them one by one to form long input sequences. Zhao et al. (2024) showed that packing documents from a single source could be more effective than packing documents sampled randomly from the entire pre-training corpora.

In this paper, we explore a more challenging case, where all documents discuss a similar topic. Even the *standard* way can provide pre-training sequences with relevant context (belonging to the biomedical-related topics). Therefore, this leads to a research question: Is it possible to further improve the pre-training for biomedical language models using exactly the same corpus?

Additionally, existing studies primarily explore training models from scratch (Gu et al., 2023; Shi et al., 2023; Zhao et al., 2024). However, it is unclear whether this conclusion still holds in continual pre-training, which is a prevalent method in domain adaptation. Levine et al. (2021) integrated similar texts selected via K-Nearest Neighbor (KNN) into the context after several steps of warming up, which could be considered as an attempt at continual pre-training. However, the LMs they used were relatively small, containing only 345M parameters. In this paper, we focus on this continual pre-training setting to improve the referring ability of biomedical language models.

3 Preliminary Experiment

All references of a given paper can serve as background information, but their importance towards the given paper is different. Therefore, it is neces-

sary to rank them based on their significance. A natural solution is using retrievers. As one of our preliminary experiments, we realize that retrievers are not as reliable as we expect in identifying the most appropriate reference for a given abstract. Before using the retriever to select references that provide sufficient background information for the following anchor abstract, we should first understand *how well a retriever can find out the reference that provides the most information for predicting a given abstract*. We know the information that a reference provides can be measured by

$$I(ref; anchor) = P(anchor) - P(anchor|ref) \quad (1)$$

where $P(anchor)$ is the perplexity of an anchor abstract, and $P(anchor|ref)$ is the perplexity of the anchor abstract when the reference is provided in the context. For each reference, $P(anchor)$ is constant, so we can measure the information and rank references directly by $P(anchor|ref)$.

To the best of our knowledge, Meditron (Chen et al., 2023) is currently the best open-sourced biomedical LM because it is continually pre-trained with biomedical texts on the top of the powerful LLM, Llama-2 (Touvron et al., 2023b), so that it can provide a relatively accurate measurement for conditional perplexity. Therefore, we use Meditron-7B to compute the ranking of references as the ground truth. Subsequently, we use some popular models including the Contriever[1] to rank the references of a given abstract. We selected 1,000 anchor abstracts for this analysis. Results are summarized in Table 1. Kendall's Tau measures the correspondence between two rankings, while HitN@Top5 represents the proportion that one of the top-N predictions exists in top-5 references ranked by Meditron-7B.

Model	Params	Kendall's Tau	Hit1@Top5	Hit3@Top5
GPT-2	0.1B	0.087	43.4%	69.0%
GPT-2 medium	0.3B	0.665	69.5%	88.5%
GPT-2 large	0.6B	0.664	70.3%	88.3%
BioMedLM	2.7B	0.590	66.0%	86.8%
Llama-2-7B	7B	0.882	89.7%	98.5%
Contriever	0.1B	0.098	48.6%	71.4%
Meditron-7B	7B	1.000	100%	100%

Table 1: Ranking performance of models. HitN@Top5 represents the proportion that one of the top-N predictions exists in top-5 references ranked by Meditron-7B.

[1]We use `facebook/contriever-msmarco` checkpoint (supervised version) from Hugging Face.

Considering Kendall's Tau and HitN@Top5, we realize that Contriever cannot accurately provide the most appropriate reference for the given abstract, despite its widespread usage in information retrieval. Specifically, only 48.6% of the top-1 retrieved reference falls in the top-5 references ranked by Meditron-7B. And the proportion of the cases where at least one of the top-3 retrieved references falls in the top-5 references ranked by Meditron-7B is 71.4%. Compared to GPT-2 (Radford et al., 2019) which has a similar number of parameters, Contriever does not show a superior performance. However, we should point out that the dense passage retriever (DPR) is more computationally efficient than auto-regressive LMs because it decouples the encoding of a pair of texts. Nevertheless, it is still a good choice in the field of information retrieval. Therefore, as a trade-off, using DPR necessitates retrieving multiple references simultaneously to ensure that the selected references can provide sufficient information to predict the following anchor abstract.

4 Methodology

In the scenario of pre-training biomedical LMs, we usually collect abstracts or full papers as the pre-training corpus. The key of our methodology is to construct a long input sequence containing relevant information in the context. Scientific researchers typically cite pertinent papers to support their conclusions and these citing papers are often previous stages of their research. Based on this, we construct the pre-training input sequence with the help of the citation network, which is easy to obtain in the biomedical domain. Algorithm 1 shows the procedure of our methodology.

To develop biomedical LMs, we use one of the most commonly used data sources, PubMed Abstract[2]. After pre-processing the raw data, we extract both textual and citing information, forming a citation network $\mathcal{G}$. We begin with a randomly selected abstract as the anchor (e.g., D_1^1 in Figure 2). Unlike previous works (Shi et al., 2023; Zhao et al., 2024), we select multiple relevant references at the same time to increase the hit rate of selected references. This approach addresses the limitations of retrievers, which do not always retrieve the most relevant reference from the given candidates, as discussed in Section 3. To increase the diversity of our LinkLM data, we randomly sample the num-

[2]https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

Algorithm 1 LinkLM Sequence Construction

Require: $\mathcal{G} = (\mathcal{D}, \mathcal{L})$: Citation network
Require: $\mathcal{R}(d)$: Return the citing references
Require: $Retriever$
$P \leftarrow [], Q \leftarrow []$
while $|\mathcal{D}| > 0$ **do**
  Randomly select d_i from $\mathcal{D}$
  $Q.append(d_i)$
  $\mathcal{D}.remove(d_i)$
  while $\mathcal{R}(d_i) \cap \mathcal{D} \neq \emptyset$ **do**
    $K \leftarrow Poisson(3)$
    $\bar{\mathcal{D}} \leftarrow TopK(\mathcal{R}(d_i) \cap \mathcal{D}, Retriever, K)$
    $d_j \leftarrow \arg\max_{d \in \bar{\mathcal{D}}} indegree(d)$
    $Q.extend(\bar{\mathcal{D}} \backslash d_j)$
    $Q.append(d_j)$
    $\mathcal{D}.remove(\bar{\mathcal{D}})$
    $d_i \leftarrow d_j$
  end while
  $P.append(Q[:: -1])$
  $Q \leftarrow []$
end while
Shuffle P
return P: List of abstracts

ber of selected references, K, following a Poisson distribution with an expected value of three. With the help of a given retriever, we select the top-K relevant references (e.g., D_2^1, D_2^2, and D_2^3 in Figure 2) from all references. To increase the possibility of constructing longer sequences, we select the reference with the largest in-degree among these K selected references. Assuming that D_2^1 has the largest in-degree, we continue the construction with D_2^1 until none of the references have any citing papers (e.g., D_4^1 in Figure 2 has no citing papers). After the construction, we reverse the constructed sequence so that the later documents are supported by the earlier ones.

At the beginning of the data construction, we easily obtain multi-hop long sequences. However, since we delete nodes once they are visited to prevent duplication of pre-training samples, the original citation graph becomes sparse gradually. Many sequences will be composed by a single document at the end of the process. Therefore, after constructing the sequences, we perform sequence-wise shuffling so that the sequences comprising a single document will be distributed uniformly alongside other longer sequences. In this way, each batch contains linked long sequences, making full use of the constructed LinkLM data.

5 Experiments

5.1 Datasets

In the continual pre-training stage, we download the raw data from the PubMed 2024 Annual baseline[3] updated until December 14, 2023. We use PubMed parser (Achakulvisut et al., 2020) to extract necessary information including the title, abstract, and citations. We exclude isolated data points that are not cited by any paper and their citations are missing. We also exclude data points without any title or abstract. After preprocessing, we obtain approximately 25 million samples as the source for pre-training.

For evaluation, we use four widely used biomedical multi-choice question-answering (MCQA) datasets, as listed below.

- **MedMCQA** (Pal et al., 2022) is a large-scale MCQA dataset collected from the AIIMS & NEET PG entrance exam, containing more than 194k QA pairs. In the default evaluation setting, LMs can only access the question and four candidate options. Therefore, it is usually used to assess the biomedical knowledge memorized by models. Since the testing set does not provide the ground-truth answers, we use its validation set for evaluation.

- **MMLU-medical** is a subset derived from MMLU (Hendrycks et al., 2020), containing 57 tasks across various fields. We select the QA pairs if they belong to one of the following topics: *high school biology, college biology, college medicine, professional medicine, medical genetics, virology, clinical knowledge, nutrition, and anatomy*. MMLU-medical is also a four-choice MCQA task and it is mainly designed to measure knowledge acquired during pre-training. We adhere to the official setting using development set for few-shot learning.

- **USMLE-QA** (Zhang et al., 2018) is an MCQA task based on United States Medical License Exams (USMLE), which requires a certain piece of knowledge or an answer based on a patient's condition description. We use the English four-choice version subset for evaluation.

- **PubMedQA** (Jin et al., 2019) is a three-choice MCQA task (yes/no/maybe). For each ques-

[3] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

	Train	Evaluation	#Choice	#Token/Sample Aver	#Token/Sample Max	w/ Context
MedMCQA	182,822	4,183	4	61.5	573	✗
MMLU-Medical	45	1,871	4	124.1	1,192	✗
USMLE-QA	10,178	1,273	4	251.8	1,152	✗
PubMedQA	211,269	1,000	3	437.1	1,909	✓

Table 2: Statistics of four biomedical MCQA datasets. Different from the other three MCQA datasets, an extra abstract is provided for each question in the PubMedQA dataset.

tion, a related abstract from PubMed is provided, making it suitable for evaluating the intra-sample referring ability of LMs.

Table 2 summarizes their statistics. We compute the probability of generating each option and select the one with the lowest perplexity as the final prediction. We report model accuracy and calculate micro-average accuracy since different datasets have different numbers of testing samples.

5.2 Experimental Settings

Due to the limitation of our computation resources, we chose TinyLlama-1.1B[4] as our experimental subject, which was pre-trained sufficiently using 3T tokens (Zhang et al., 2024). After tokenization, we obtained approximately 8B tokens for continual pre-training. We followed most of the original hyperparameters of pre-training TinyLlama with a context length of 2048 tokens. Further details are provided in Appendix B.1. In the following comparisons, 'Vanilla' denotes the original TinyLlama. 'Standard' and 'LinkLM' represent the continually pre-trained TinyLlama with randomly packed documents and LinkLM data, respectively.

5.3 Intra-Sample Referring Ability

As discussed in Section 5.1, among these four medical MCQA tasks, PubMedQA requires LMs to answer questions by referring to the given related abstract. Therefore, we perform a zero-shot evaluation on PubMedQA to evaluate the intra-sample referring ability of LMs. We observe fluctuations across different checkpoints. To better visualize their differences, we smooth the average accuracy with windows of size three. Figure 3 illustrates the zero-shot performance on PubMedQA. We find that after training approximately 3B tokens, the LM pre-trained with LinkLM data consistently and significantly outperforms standard pre-training, indicating the effectiveness of our proposed method.

[4]We used `TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T` checkpoint from Hugging Face.

Additionally, Table 3 shows the quantitative performances of four biomedical MCQA datasets. Compared to the vanilla TinyLlama, continual pre-training enriches the biomedical knowledge of LMs, leading to a 10.3% relative improvement (from 29.59 to 32.63) from vanilla TinyLlama to continual pre-trained TinyLlama. However, with our designed LinkLM data, though it can also achieve a 9.4% relative improvement compared to the vanilla TinyLlama, performances on some datasets (e.g., MedMCQA) slightly drop compared to standard pre-training. This observation indicates that while using LinkLM data encourages LMs to refer to previous contexts, it may also weaken memorization during pre-training.

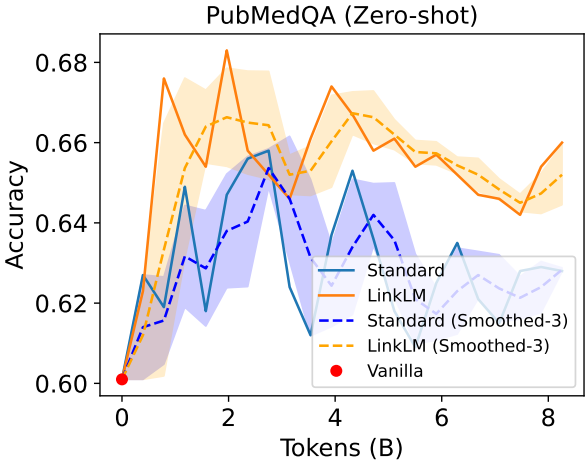


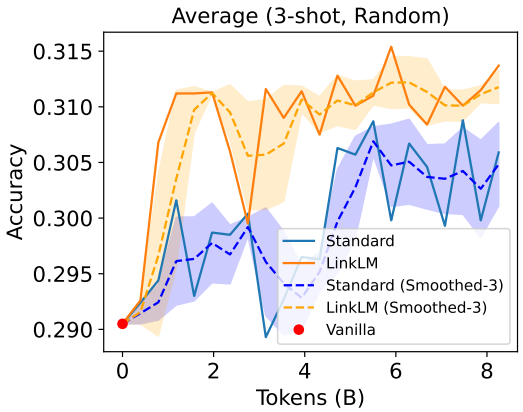
Figure 3: Comparison between different pre-training strategies on PubMedQA (Smoothing window size=3). The full and dotted lines represent the exact and smoothed values of performances, respectively. The colored area represents the standard deviation within a smoothing window.

5.4 Inter-sample Referring Ability

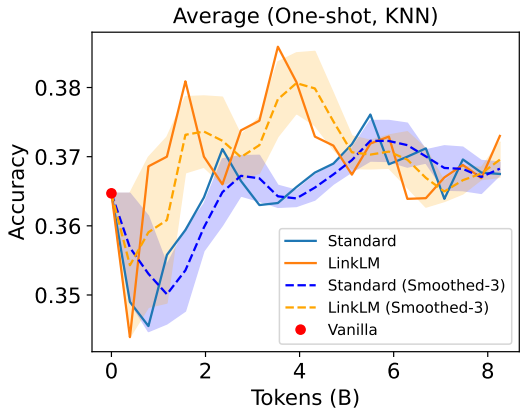
Auto-regressive biomedical LMs are usually employed under the in-context learning scenario, learning from the input-label mapping in previous demonstrations, which can be considered as the inter-sample referring ability. Therefore, we perform a few-shot evaluation on these four datasets, specifically conducting a three-shot evaluation.

Accuracy (%)		MedMCQA	MMLU-Medical	USMLE-QA	PubMedQA	Average (Micro)
Vanilla	(0 shot)	25.34	24.91	26.47	60.10	29.59
Standard	(0 shot)	29.55	25.98	28.83	62.80	32.63
LinkLM	(0 shot)	28.97	25.44	27.26	**66.00**	32.36
Vanilla	(3 shot, Random)	22.96±0.52	26.03±0.32	25.56±0.37	64.80±1.40	29.05
Standard	(3 shot, Random)	25.78±0.61	26.53±0.94	26.34±1.20	63.73±0.53	30.59
LinkLM	(3 shot, Random)	27.13±0.28	25.24±1.26	27.36±0.84	65.67±0.87	31.37
Vanilla	(1 shot, KNN)	30.10	**26.94**	26.55	62.30	32.69
Standard	(1 shot, KNN)	36.96	25.98	30.32	64.20	36.75
LinkLM	(1 shot, KNN)	**38.47**	25.01	**30.48**	64.10	**37.30**

Table 3: Quantitative performances of the vanilla TinyLlama and final checkpoints that are continually pre-trained in the standard way or with our LinkLM data on four biomedical MCQA datasets. The best and second-best performances are highlighted in bold and underlined, respectively. For standard few-shot evaluation, we run multiple times with three different random seeds to reduce the variant of the results.



(a) Smoothed average accuracy across four biomedical MCQA tasks under three-shot evaluation (Window size=3)



(b) Smoothed average accuracy across four biomedical MCQA tasks under one-shot evaluation using retrieved demonstration (Window size=3)

Figure 4: Comparison between different pre-training strategies under few-shot evaluation. The full and dotted lines represent the exact and smoothed values of performances, respectively. The colored area represents the standard deviation within a smoothing window.

Figure 4a illustrates that pre-training with LinkLM data significantly outperforms the standard pre-training under few-shot evaluation. Remarkably, 90.48% of the checkpoints have better average accuracy across the four datasets than standard pre-training, which confirms again the effectiveness of LinkLM data under continual pre-training. However, compared to zero-shot performance, TinyLlama-1.1B does not consistently benefit from the provided demonstrations in standard few-shot settings, as evidenced by its performance on MedMCQA and USMLE-QA. The average performances even drop slightly for TinyLlama pre-trained in the standard way (about 6.3% relative degradation) and TinyLlama pre-trained with LinkLM data (about 3.1% relative degradation). We hypothesize that it is due to the quality of randomly sampled demonstrations that fail to provide useful information and may even disrupt LM predictions.

Inspired by KATE (Liu et al., 2022), which retrieves similar demonstrations to boost few-shot performance, we use Contriever to retrieve the top-K similar demonstrations from each training set. Contrary to the findings reported in Min et al. (2022), our results suggest that it is possible to retrieve helpful demonstrations from the training set, whose input-label mapping can benefit the prediction of the query. We perform a one-shot evaluation here since adding more retrieved demonstrations does not improve the performance in our case. Figure 4b shows the comparison between our method and standard pre-training. Under this experimental setting, we observe obvious improvements over standard few-shot evaluation, highlighting the importance of high-quality demonstrations in the ICL scenario. Although the LM trained with LinkLM data only slightly outperforms standard pre-training at the end of continual pre-training, there

are 71.43% of checkpoints that have better average accuracy across four datasets than the standard pre-training. After pre-training for several steps, the LM pre-trained with LinkLM data can achieve good performance under this setting, indicating that LinkLM data can activate their potential on inter-sample referring ability when the demonstrations are closely related to the following query.

Table 3 demonstrates that using retrieved demonstrations instead of using randomly sampled ones as in standard ICL can significantly boost few-shot performance. With appropriate demonstrations, LMs perform significantly better than those under the zero-shot setting. Compared to zero-shot performance, LMs continually pre-trained in the standard way and with our designed LinkLM data achieve 12.4% and 15.3% of relative improvement, respectively. We believe the reason is that in the standard ICL setting, the sampled demonstrations may not be strongly related to the current question, so they can only provide shallow information like task format (Min et al., 2022). Sometimes, they even distract the LMs. However, when using retrieved demonstrations, current questions can not only understand the task format but also learn from the input-label mapping and knowledge shown in the demonstrations. LMs trained with LinkLM data can further improve inter-sample referring ability during the continual pre-training stage, thus achieving larger improvement in few-shot evaluation.

Especially, on MedMCQA, LM trained with LinkLM data significantly outperforms LM trained in a standard way, no matter whether the demonstrations are randomly sampled or retrieved. By conducting a case study on MedMCQA, shown in Table 5, we find that retrieved demonstrations from the training set are highly related to the following question and usually provide pertinent knowledge. Since TinyLlama pre-trained with LinkLM data can memorize knowledge and learn to refer to necessary information across different documents meanwhile during continual pre-training, it is also encouraged to refer to some information from previous contexts in downstream tasks after pre-training.

Note that in domain adaptation, we usually use documents in a single focused domain, and therefore even the *standard* approach concatenates documents with similar topics within the context, helping LMs to refer to necessary information across document boundaries (i.e., $[EOS]$ token). In our method, we explicitly arrange the related references in the context, improving the inter-sample referring ability further. From another aspect, our pre-training method narrows the gap between pre-training phases and ICL with retrieved demonstrations. Therefore, we can expect that the inter-sample referring ability will be improved further and more robust if we construct more LinkLM data for further training.

One-shot example (with retrieved demonstration) for MedMCQA

```
Question: A 60 year old male presents with a creamy curd
like white patch on the tongue. The probable diagnosis is -
A. Candidiasis
B. Histoplasmosis
C. Lichen planus
D. Aspergillosis
Answer: Candidiasis

Question: In a 6-month-old child, thick curd like white
patch appears on the buccal mucosa. On rubbing it leaves an
erythematous patch. Most likely diagnosis is:
A. Tuberculosis
B. Lichen planus
C. Lupus erythematous
D. Candidiasis
Answer:
```

Prediction: *Candidiasis*

Figure 5: Example of one-shot ICL with the retrieved demonstration on the MedMCQA dataset

6 Conclusions

In this paper, we propose a pre-training sequence construction method for improving the referring ability of biomedical language models. Previous studies mostly focus on general domains and they train the LMs from scratch with designed pre-training sequences. In contrast, we explore this topic in a more challenging scenario, where the distinction of fine-grained topics is more difficult in the biomedical domain. Moreover, we explore it under the continual pre-training setting, since it is a prevalent method for developing domain-specific LMs now, filling the gap in this series of work. In this paper, we construct pre-training sequences by concatenating relevant references into the previous context using linking information from a citation network. Empirical studies show that compared to the standard pre-training (i.e., randomly packing documents), our method significantly improves the intra-sample referring ability and the inter-sample referring ability on biomedical MCQA tasks, which answers our research question: by carefully designing pre-training sequences, we can still improve the pre-training for biomedical language models by reordering the pre-training documents (using exactly the same corpus). Especially, pre-training using LinkLM data can further improve the performance when using retrieved demonstrations, revealing the future potential of our proposed methodology.

Limitations

Owing to limited computation resources, we only conducted experiments on a language model with 1.1B parameters (TinyLlama-1.1B) using up to 8B tokens, which may not be sufficient for biomedical LLM applications. Experiments on larger models with larger amounts of biomedical pre-training data are needed in the future. However, according to the current trend shown in our experiments, after training with more LinkLM data, the improvement compared to the standard pre-training would be larger.

Another limitation is that our methodology requires a citation network, restricting its applicability to other scientific domains where it is not easy to build the citation network. To address this, we believe that training a classifier for link prediction may be a possible solution. However, due to the constraints of this paper's length, we will not explore this direction in depth.

Besides, full papers from PubMed Central[5] are also commonly used for pre-training biomedical LMs. However, most of the full papers exceed the maximum input length of existing foundation LMs. Although these full papers are also linked to the citation network, how to construct LinkLM data for them remains a challenge. Future efforts will consider separating full papers into several paragraphs and constructing better pre-training sequences to improve the referring ability of biomedical LLMs.

Ethics Statement

Though using LinkLM data can improve the referring ability for biomedical language models, particularly in retrieval-augmented tasks (e.g., PubMedQA) and in-context learning scenarios, some potential issues for biomedical LMs may also apply to our case, such as generating inappropriate clinical suggestions accompanied by hallucinations. We strongly recommend conducting a thorough assessment and careful alignment (e.g., employing RLHF (Ouyang et al., 2022)) before deployment to the real world.

The involved pre-trained language model, TinyLlama, is licensed under Apache License 2.0[6]. We adhere strictly to this license during our experiments. Regarding the involved dataset, PubMed Abstract, we collected the raw data following instructions on the official website[7], ensuring not to violate their terms.

[5] https://www.ncbi.nlm.nih.gov/pmc

[6] http://www.apache.org/licenses/LICENSE-2.0

[7] https://pubmed.ncbi.nlm.nih.gov/download/

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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *EMNLP-IJCNLP*, pages 3615–3620.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

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 Trans. Comput. Healthcare*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Pre-training to learn in context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4849–4870.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Xanh Ho, Anh Khoa Duong Nguyen, An Tuan Dao, Junfeng Jiang, Yuki Chida, Kaito Sugimoto, Huy Quoc To, Florian Boudin, and Akiko Aizawa. 2024. A survey of pre-trained language models for processing scientific text. *arXiv preprint arXiv:2401.17824*.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. 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.

Yoav Levine, Noam Wies, Daniel Jannai, Dan Navon, Yedid Hoshen, and Amnon Shashua. 2021. The inductive bias of in-context learning: Rethinking pretraining example design. In *International Conference on Learning Representations*.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2023. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pages 22188–22214. PMLR.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style,

high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A Smith, Luke Zettlemoyer, Wen-tau Yih, and Mike Lewis. 2023. In-context pretraining: Language modeling beyond document boundaries. In *The Twelfth International Conference on Learning Representations*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *Preprint*, arXiv:2211.09085.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045.

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.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yu Zhao, Yuanbin Qu, Konrad Staniszewski, Szymon Tworkowski, Wei Liu, Piotr Miłoś, Yuxiang Wu, and Pasquale Minervini. 2024. Analysing the impact of sequence composition on language model pre-training. *arXiv preprint arXiv:2402.13991*.

A Perplexity Evaluation

In addition to evaluating on downstream tasks, we also tracked the loss on the evaluation set. We sampled 10,000 abstracts from the excluded isolated data points to serve as the evaluation set for perplexity evaluation. As shown in Table 4, no significant difference was observed between the standard pre-training and pre-training with our LinkLM data, which is consistent with the findings of Liu et al. (2023) stating that LMs with similar pre-training losses may perform differently on downstream tasks.

Strategy	Eval Loss	Eval PPL
Standard	1.871	6.49
LinkLM	1.874	6.51

Table 4: Loss and perplexity on evaluation set.

B Experimental Details

B.1 Implementation Details

We chose TinyLlama-1.1B[8] as our experimental subject, which had been pre-trained sufficiently using 3T tokens (Zhang et al., 2024). After tokenization, we obtain approximately 8B tokens for continual pre-training. We follow most of the original hyperparameters for pre-training TinyLlama, using a context length of 2048 tokens. The global batch size we use is 0.5M tokens. According to the conclusions from Goyal et al. (2017), we use a smaller learning rate of 1e-4.

[8]We used `TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T` checkpoint from Hugging Face.

We used PyTorch (Paszke et al., 2019) and transformers library (Wolf et al., 2020) for implementation. Pre-trained checkpoints were downloaded from Hugging Face[9]. We also adopted Deepspeed Zero3 (Rajbhandari et al., 2020), flash-attention (Dao et al., 2022; Dao, 2024), and checkpointing techniques to speed up training. All experiments were conducted on 8 NVIDIA A100 (40GB) GPUs. Continual pre-training TinyLlama-1.1B with approximately 8B tokens cost approximately 24 hours on these 8 NVIDIA A100 GPUs.

B.2 Prompt Engineering

In our zero-shot and few-shot evaluation, we used the prompts following Gao et al. (2023) to complete the multi-choice question-answering tasks as shown in Table 5. And Table 6 shows an example for MedMCQA under the few-shot evaluation (#Shot=3). With the help of a retriever, we can retrieve relevant demonstrations from the training set to assist the prediction of the following queries, as shown in Figure 5, where we also find that the retrieved demonstrations actually provide not only the task format but also relevant knowledge, and therefore benefits the in-context learning.

Prompt template for MedMCQA, USMLE-QA, and MMLU-Medical

```
Question: {question}
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}
Answer:
```

Prompt template for PubMedQA

```
Abstract: {context}
Question: {question}
Answer:
```

Table 5: Prompt templates for MCQA tasks.

[9] https://huggingface.co/models

Three-shot example for MedMCQA

```
Question: Claw sign on x-ray is seen in?
A. Ischemic colitis
B. Intussusception
C. Sigmoid volvulus
D. Crohn's disease
Answer: Intussusception

Question: All of the following are microsomal enzyme inhibitors except
A. Glucocoicoids
B. Cimetidine
C. Ciprofloxacin
D. INH
Answer: Glucocoicoids

Question: A young female presents with a history of dyspnoea on exertion. On
examination, she has wide, fixed split S2 with ejection systolic murmur (III/VI)
in left second intercostal space. Her ECG shows left axis deviation. The most
probable diagnosis is –
A. Total anomalous pulmonary venous drainge.
B. Tricuspid atresia.
C. Ostium primum atrial septal defect.
D. Ventricular septal defect with pulmonary arterial hypertension.
Answer: Ostium primum atrial septal defect.

Question: Which of the following is not true for myelinated nerve fibers:
A. Impulse through myelinated fibers is slower than non-myelinated fibers
B. Membrane currents are generated at nodes of Ranvier
C. Saltatory conduction of impulses is seen
D. Local anesthesia is effective only when the nerve is not covered by myelin
sheath
Answer:
```

Table 6: An example of three-shot in-context learning for MedMCQA.