Citation-Aware Continual Pre-Training for Biomedical Language Models

Masaki Asada, Tomoki Tsujimura, Tatsuya Ishigaki,
Shusaku Egami, Ken Fukuda, Hiroya Takamura,

National Institute of Advanced Industrial Science and Technology (AIST)
Correspondence: masaki.asada@aist.go.jp

Abstract

The biomedical literature contains rich structured knowledge, including citation links that encode relationships between scientific studies, but such information is typically ignored in standard language model pre-training. We propose a citation-aware continual pre-training method for decoder-only language models that incorporates citation graph information from PubMed into next-token prediction by placing citation-linked abstract pairs within a shared context. We evaluate our method on multiple biomedical QA benchmarks using two model families. Results show that citation-aware continual pre-training achieves higher average accuracy than both the original base models and citation-unaware pre-training across biomedical tasks.

1 Introduction

The biomedical literature constitutes a large and information-rich corpus, with PubMed indexing tens of millions of abstracts (Liang et al., 2021). The citation links among these documents encode a human-curated graph of conceptual relationships. When a researcher cites a prior study, it typically reflects some form of domain relevance, methodological connection, or thematic continuity between the two works. However, this structural information is largely ignored in standard language model pre-training pipelines, which treat each document as an independent sequence drawn uniformly at random.

The encoder-only era demonstrated that citation structure can be a powerful supervision signal. LinkBERT (Yasunaga et al., 2022) placed linked documents into shared BERT (Devlin et al., 2019) contexts and introduced a document relation prediction objective, yielding substantial gains on QA and multi-hop reasoning benchmarks. Its biomedical variant, BioLinkBERT, achieved state-of-the-art results on BLURB (Gu et al., 2021), confirming that awareness of inter-document relationships translates directly into improved clinical language understanding.

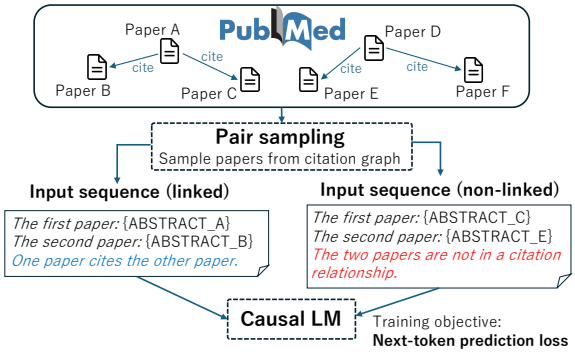

Figure 1: Overview of citation-aware biomedical pre-training. Each training example concatenates two abstracts with a relation text derived from the citation graph, and is trained with next-token prediction.

Decoder-only autoregressive large language models (LLMs), trained with a causal next-token prediction objective, have become the dominant modeling approach, and domain adaptation is typically pursued through continual pre-training on biomedical corpora such as PubMed Central. BioMistral (Labrak et al., 2024), for instance, adapts Mistral-7B with billions of biomedical tokens and reports improvements over general-purpose counterparts on medical QA benchmarks. Despite this progress, continual pre-training for causal LLMs remains document-independent, with each sample consisting of a single abstract or article and lacking structural signals about inter-document relationships. Citation-linked papers typically share concepts, methods, or disease entities, so placing them in a shared context window exposes the model to inter-document concept co-occurrences that no single abstract contains. Prior work has shown that citation structure encodes meaningful semantic relatedness at the document level (Cohan et al., 2020; Ostendorff et al., 2022); our hypothesis is that the same signal,

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 407–412

exposed at pre-training time through next-token prediction, supports integrative biomedical reasoning even when downstream questions are posed in closed-book form over a single passage. This raises a natural, to our knowledge unanswered question: does citation-aware pre-training confer the same benefits in causal LMs as it does for masked LMs?

As illustrated in Figure 1, we propose a method that incorporates information from the PubMed citation graph into next-token prediction pre-training. by constructing training samples that place citation-linked abstract pairs within the same context window together with a natural-language label indicating their relationship. We evaluate our approach on two small language models across multiple biomedical QA benchmarks, with a text-matched baseline that isolates the effect of citation structure from data quantity and lexical content. Our main contributions are as follows:

- We extend citation-aware pre-training from masked LMs to causal LMs by introducing a method that incorporates inter-document relationships while remaining compatible with standard next-token prediction training.
- We show that citation-aware continual pre-training improves accuracy over backbone models and citation-unaware pre-training across biomedical QA benchmarks and model families.

2 Related Work

2.1 Biomedical Pre-trained Language Models

Dedicated pre-training on biomedical corpora has a long history. BioBERT (Lee et al., 2020) first demonstrated that continued pre-training of BERT on PubMed yields substantial gains across named entity recognition, relation extraction, and QA. With the rise of decoder-only architectures, the community has adapted these insights to the autoregressive setting. BioGPT (Luo et al., 2022) is a GPT-2-scale causal LM pre-trained on PubMed abstracts, while PMC-LLaMA (Wu et al., 2024) scales this to 75 billion tokens from PubMed Central and medical textbooks. BioMistral (Labrak et al., 2024) demonstrates that continual pre-training of a strong general-purpose model on PubMed Central is competitive with purpose-built biomedical models at a fraction of the compute cost. More recently, MedGemma (Sellergren et al., 2025) extends the Gemma 3 architecture with medical-domain training on diverse clinical data including radiology images and medical text, while ELAINE-medLLM (Yano et al., 2025) adapts Llama-3-8B for the biomedical domain across English, Japanese, and Chinese through continual pre-training and supervised fine-tuning. Despite the diversity of these approaches, all treat the corpus as a flat collection of documents, ignoring citation relationships. Our work is the first to exploit PubMed's citation graph in causal LM continual pre-training.

2.2 Graph-Aware and Structured Pre-training

Citation links encode semantically meaningful relationships between documents, motivating a line of graph-aware pre-training methods. SPECTER (Cohan et al., 2020) learns document embeddings using citation links as a proxy for semantic similarity in a triplet-loss framework, and SciNCL (Ostendorff et al., 2022) refines this with neighborhood-contrastive learning over citation graphs. LinkBERT (Yasunaga et al., 2022) is most directly related to our work. It augments BERT's masked language modeling objective with a document relation prediction task, classifying whether two text segments are contiguous, randomly paired, or citation-linked. Applied to PubMed, BioLinkBERT achieves state-of-the-art performance on PubMedQA (Jin et al., 2019), BioASQ (Nentidis et al., 2019) and the BLURB (Gu et al., 2021) benchmark. Our work translates the core insight of LinkBERT to the causal LM setting: rather than a classification auxiliary task, we encode document relation information as natural language tokens predicted autoregressively, making our method applicable to any decoder-only architecture without modification.

3 Method

3.1 Overview

We propose a citation-aware continual pre-training approach for decoder-only large language models. Our core idea is to exploit the PubMed citation graph, a large-scale human-curated signal of semantic relatedness between biomedical abstracts, by constructing training samples that place citation-linked document pairs within the same autoregressive context. Unlike LinkBERT (Yasunaga et al., 2022), which requires a separate classification head for document relation prediction, our formulation is architecturally transparent: the citation relation-

ship is expressed as a natural-language sentence appended to each paired-abstract sample and learned via standard next-token prediction.

3.2 Training Sample Construction

Given the PubMed citation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ corresponds to an abstract and each directed edge $(u, v) \in \mathcal{E}$ denotes that paper u cites paper v, we construct training samples of the following two forms.

Citation-linked sample:

```
The first paper: {ABSTRACT_A}
The second paper: {ABSTRACT_B}
One paper cites the other paper.
```

Non-linked sample:

```
The first paper: {ABSTRACT_C}
The second paper: {ABSTRACT_D}
The two papers are not in a
citation relationship.
```

For citation-linked pairs, we sample edges $(u, v) \in \mathcal{E}$ and place the abstracts of u and v as ABSTRACT_A and ABSTRACT_B, respectively. The order of the two abstracts within a pair is randomized to prevent the model from exploiting positional cues about citation directionality. For non-linked pairs, we sample two abstracts uniformly at random, verifying that no citation edge exists between them in either direction.

This formulation offers several practical advantages. First, it requires no modification to the model architecture or training objective: existing autoregressive pre-training pipelines can incorporate citation-linked samples without change. Second, the natural-language relation label is semantically informative; the model must associate the content of the two abstracts with the prediction of the concluding token sequence, implicitly encoding inter-document coherence.

3.3 Baseline: Citation-Unaware Pre-training

To isolate the effect of citation structure from the effect of data selection and lexical content, we construct a citation-unaware baseline as follows. Starting from the same set of paired-abstract samples used for our proposed model, we decompose each sample back into its constituent single-document units and shuffle these documents uniformly at random, forming standard single-document training samples. The resulting baseline corpus contains the same abstract texts as the proposed model but without citation-pairing information. The only difference between the two conditions is whether linked pairs are presented in context, enabling a clean ablation of citation structure.

4 Experiments

4.1 Experimental Settings

Backbone Models We apply our continual pre-training procedure to two recent small language models representing different model families.

- **Phi-3.5-mini-instruct** (4B parameters) (Abdin et al., 2024): a decoder-only language model pre-trained on a mixture of web and synthetic data.
- **Llama-3.2-3B-Instruct** (Dubey et al., 2024): an instruction-tuned decoder-only language model from the Llama 3 family.

Training Data Composition We set the total number of tokens after tokenization to approximately 1.8 billion, using the same set of PubMed abstracts for both citation-aware and citation-unaware settings to ensure that any performance differences cannot be attributed to differences in data quantity.

In the citation-aware setting, the total number of training tokens is 1,877,396,650, while in the citation-unaware setting it is 1,807,869,680. Although both settings are constructed from the same set of PubMed abstracts, the citation-aware data contains slightly more tokens because it includes additional textual prefixes such as "One paper cites the other paper." or "The two papers are not in a citation relationship."

While our overall scale is smaller than that of prior biomedical language models such as BioMistral (Labrak et al., 2024), which was trained on 3B tokens, it remains on the same order of magnitude.

Benchmarks We evaluate on five established biomedical QA benchmarks implemented via `lm-evaluation-harness` [1], reporting accuracy on all tasks: **PubMedQA** (Jin et al., 2019), **MedMCQA** (Pal et al., 2022), **MedQA-4opt** (Jin et al., 2020), and medical subsets of **MMLU** (Hendrycks et al., 2021) (*anatomy*, *clinical knowledge*, *college biology*, *college medicine*, *medical genetics*, and

[1] https://github.com/EleutherAI/lm-evaluation-harness

Citation-Aware		BioMistral-7B† NA	BioGemma-4B† NA	Phi-3.5-4B-Inst NA	Phi-3.5-4B-Inst ✗	Phi-3.5-4B-Inst ✓	Llama-3.2-3B-Inst NA	Llama-3.2-3B-Inst ✗	Llama-3.2-3B-Inst ✓
MedQA-4opt		44.4	**64.4**	55.9	54.8	55.5	55.9	53.5	55.4
MedMCQA		43.9	**55.7**	53.3	52.8	53.6	54.2	53.9	54.7
PubMedQA		37.6	73.4	74.8	74.8	74.4	75.2	75.4	**76.4**
MMLU	anatomy	49.6	59.3	62.2	61.5	63.0	62.2	60.7	**63.7**
	clinical KG	60.9	71.3	64.2	65.3	64.2	75.1	76.2	**75.5**
	college biology	56.9	70.8	72.2	71.5	72.9	84.0	81.3	**84.7**
	college medicine	55.5	**65.3**	57.2	59.5	57.8	64.2	**65.3**	**65.3**
	medical genetics	61.7	**83.0**	76.0	75.0	76.0	75.0	74.0	76.0
	prof. medicine	55.1	**76.8**	69.5	70.2	72.8	74.6	72.4	73.2
Average		51.7	68.9	65.0	65.1	65.6	68.9	68.1	**69.4**

Table 1: Results on biomedical QA benchmarks without task-specific fine-tuning. "NA" indicates that the original released model is evaluated without additional continual pre-training. "✗" denotes Citation-Unaware continual pre-training, while "✓" denotes Citation-Aware continual pre-training (our proposed method). Blue values indicate improvements over the corresponding base model, while red values indicate performance degradation. **Boldface** highlights the best score in each row. † indicates scores reported from the original papers.

Citation-Aware		Phi-3.5-4B-Inst NA	Phi-3.5-4B-Inst ✗	Phi-3.5-4B-Inst ✓
PubMedQA		76.6	77.4	**78.6**
MedMCQA		58.2	58.9	**59.1**
MedQA-4opt		57.8	57.3	**58.1**
MMLU	anatomy	59.3	59.3	**63.0**
	clinical_KG	74.3	74.3	**76.2**
	college_biology	**85.4**	81.3	84.0
	college_medicine	64.7	65.9	**67.6**
	medical_genetics	75.0	73.0	**76.0**
	pro_medicine	71.0	71.0	**71.7**
Average		69.2	68.7	**70.5**

Table 2: Results on biomedical QA benchmarks with task-specific fine-tuning.

professional medicine). Average accuracy (AVG) across all nine evaluation settings is reported as our primary aggregate metric.

Following prior work (Labrak et al., 2024), we adopt a 3-shot evaluation protocol without task-specific fine-tuning as our default setting. In addition, for the Phi-3.5-4B model, we also report results with supervised fine-tuning on the target tasks to assess the impact of downstream adaptation.

4.2 Results

Tables 1 and 2 present the results of continual pre-training under citation-unaware and citation-aware settings across different model families and evaluation conditions.

Overall, citation-aware pre-training consistently achieves the strongest performance across models and settings. In Table 1, citation-aware models outperform both the original released models (NA) and the citation-unaware baselines (✗) in terms of average accuracy. For instance, on Phi-3.5-4B, citation-aware pre-training improves the average score from 65.0 (NA) and 65.1 (✗) to 65.6, while on Llama-3.2-3B, it increases performance from 68.1 (✗) and 68.9 (NA) to 69.4. These gains are also reflected at the task level, with consistent improvements observed across multiple MMLU subsets and QA benchmarks.

Importantly, citation-unaware continual pre-training does not reliably improve performance over the base models and can even lead to slight degradation in several cases. This suggests that simply continuing pre-training on biomedical text without incorporating structural signals may not be sufficient, and can potentially harm previously acquired capabilities.

Table 2 further examines the effect of supervised fine-tuning (SFT) for Phi-3.5-4B-instruct model. The advantage of citation-aware pre-training not only persists but becomes more pronounced after fine-tuning, achieving the highest average accuracy (70.5) compared to both the base model (69.2) and the citation-unaware variant (68.7). Notably, citation-aware models achieve the best performance in the majority of individual tasks, indicating that the benefits of citation-aware representations transfer effectively to downstream adaptation.

While the magnitude in any single configuration is modest, the gain of citation-aware over citation-unaware pre-training has a consistent positive sign across two backbones, the majority of the nine tasks, and both the zero-shot and SFT regimes. Such cross-setting consistency is more readily explained by a systematic effect of the training signal than by configuration-specific noise.

5 Conclusion

We proposed a citation-aware continual pre-training method for decoder-only language models, incorporating citation graph information from PubMed into standard next-token prediction by placing citation-linked abstract pairs within a shared context. Experiments on biomedical benchmarks show that this approach improves performance over both the backbone models and citation-unaware pre-training across model families.

In future work, we plan to (i) compare citation-aware pre-training against a topically matched non-cited baseline to isolate the contribution of citation structure from topical co-occurrence, (ii) evaluate on multi-hop biomedical benchmarks to test the inter-document mechanism directly, and (iii) extend the approach to full-text articles and finer-grained citation contexts.

Limitations

This work has several limitations. First, our experiments use relatively small models (3B–4B parameters) and a limited pre-training scale (approximately 1.8B tokens), and the gains, while consistent, are modest in absolute terms; scaling to larger models and corpora may yield different outcomes. Second, our citation-unaware baseline draws abstract pairs uniformly at random, so citation-linked pairs differ from it in both citation status and topical similarity. The observed gains may therefore partly reflect co-presentation of topically related documents rather than citation structure per se, and a more controlled comparison would match non-cited pairs on topical similarity (e.g., via MeSH overlap or embedding similarity). Relatedly, all five benchmarks we use are closed-book single-passage QA and do not directly probe inter-document reasoning; multi-hop or cross-document biomedical benchmarks would test the proposed mechanism more directly. Finally, our method operates only on abstract-level data and cannot capture finer-grained relationships expressed at the sentence or passage level, which may limit its ability to fully exploit citation information.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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*, 3(1).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

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, Hong Kong, China. Association for Computational Linguistics.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

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.

Zhentao Liang, Jin Mao, Kun Lu, and Gang Li. 2021. Finding citations for pubmed: a large-scale comparison between five freely available bibliographic data sources. *Scientometrics*, 126(12):9519–9542.

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

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2019. Results of the seventh edition of the bioasq challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 553–568. Springer.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

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*, 31(9):1833–1843.

Ken Yano, Zheheng Luo, Jimin Huang, Qianqian Xie, Masaki Asada, Chenhan Yuan, Kailai Yang, Makoto Miwa, Sophia Ananiadou, and Jun'ichi Tsujii. 2025. ELAINE-medLLM: Lightweight English Japanese Chinese trilingual large language model for biomedical domain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4670–4688, Abu Dhabi, UAE. Association for Computational Linguistics.

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.