Training Biomedical Retrievers From Large-Scale Citation Contexts

Xing David Wang and Duy Le Thanh and Ulf Leser
Humboldt-Universität zu Berlin
Department of Computer Science
{xing.david.wang,leser}@hu-berlin.de

Abstract

The MedCPT model has demonstrated that strong biomedical retrievers can be trained using proprietary PubMed search logs. In this work, we study whether freely available citation sentences are sufficient to train similarly effective models. We construct a large-scale training dataset of ~62 million citation sentence-abstract pairs extracted from PubMed Central. We train a lightweight BERT-based retriever-reranker model called CiteRec on this dataset and evaluate it across three benchmark settings: (a) the biomedical subset of BEIR for information retrieval, (b) SciRepEval for generalizable scientific document embeddings, and (c) CitancePlus, a new set of ~90 thousand citation sentence-abstract pairs for PubMed-scale citation recommendation.

We show that CiteRec performs competitively with MedCPT on the biomedical BEIR subset and outperforms it on SciRepEval. On CitancePlus, CiteRec achieves strong performance for citation recommendation over the full PubMed corpus, outperforming both MedCPT and a substantially larger Qwen3-Embedding-8B retriever.

1 Introduction

Dense retrievers have become increasingly effective in first-stage information retrieval (IR) as shown on several IR benchmarks (Thakur et al., 2021; Muennighoff et al., 2023; Enevoldsen et al., 2025). They also have been successfully adopted to the biomedical domain (Jin et al., 2023; Sinha et al., 2025). Dense retriever are trained using query-document relevance pairs, which can be obtained either from human-annotated datasets such as MS MARCO (Bajaj et al., 2016) or BioASQ (Tsatsaronis et al., 2015), or automatically mined pairs like search engine logs (Jin et al., 2023; Rekabsaz et al., 2021). Automatically mined relevance pairs are much easier to obtain, however, they may lack in quality compared to manually annotated pairs. Citation graphs and citation edges provide another potential source for automatically mining relevance pairs (Cohan et al., 2020; Medić and Snajder, 2020; Li et al., 2023), either using abstract-abstract, or citation sentences-abstract pairs. In the following, we call the sentence surrounding an in-text citation a "citance" (see examples in Table 1). These citances are attractive supervision signals because they often summarize or contextualize the cited work, representing potential queries answerable through the referenced document.

Among these sources, search logs have proven particularly effective in training dense retrievers. The MedCPT model (Jin et al., 2023) has demonstrated that strong biomedical retrievers can be trained using only PubMed search logs, highlighting the effectiveness of large-scale weak supervision but relying on data that is not publicly available. In this work, we study whether freely available citances alone can provide sufficient supervision to train similarly effective models for both citation recommendation and general biomedical information retrieval. Prior work leveraging citances as training data (Medić and Snajder, 2020; Gu et al., 2022) mainly focused their evaluation on the citation recommendation task. In contrast, we also evaluate their effectiveness for general biomedical retrieval. For training our model, we use the OPCitance dataset (Hsiao and Torvik, 2023) which is derived from all citances and their referenced PubMed IDs in PubMed Central full-text articles published until May 2019. After pre-processing the OPCitance dataset, our resulting dataset contains about 62 million citance-abstract pairs.

We train a lightweight BERT-based retriever-reranker model, CiteRec, on our dataset of citance-abstract pairs. Despite relying solely on freely available and partially noisy citation data, CiteRec achieves competitive performance with MedCPT on biomedical information retrieval in BEIR, outperforms it on general scientific embedding tasks

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 75–83

Query	Relevant Document
The macroscopic appearance of the desmoid tumor is a pale surface with scant vascularization.	(Janssen et al., 2017)
Carboxylation is a common strategy in the assimilation of compounds that lack a functionalizable terminal carbon group.	(Erb, 2011)
High-grade tumors and pre-operative urinary cytologypositive cases are more likely to have recurrence.	(Yamada et al., 2009)

Table 1: Example citation sentences (citances) and their referenced documents from our training dataset.

and achieves strong results on citation recommendation. Our contributions are summarized as follows:

- We train the CiteRec model, comprised of a BERT-based dense retriever and cross-encoder reranker, on a large corpus of over 60 million citance-abstract pairs collected from PubMed Central.
- We systematically evaluate performance on a diverse set of biomedical and scientific benchmarks: (a) the biomedical subset of BEIR for information retrieval, (b) SciRepEval for generalizable scientific document embeddings, and (c) CitancePlus, a new citation recommendation dataset mined from PubMed Central after the OPCitance cutoff date, complementing it with over 90 thousand unseen citance-abstract pairs .
- We show that a lightweight BERT-based retriever-reranker trained solely on citance–abstract pairs generalizes well across tasks. The model performs competitively with MedCPT on biomedical BEIR and outperforms it on SciRepEval. On CitancePlus citation recommendation, it outperforms both MedCPT as well as substantially larger decoder-based embedding models such as the Qwen3-Embedding-8B retriever.

Our results suggest that citation sentences alone provide a surprisingly strong supervision signal for learning biomedical document representations.

The rest of the paper is structured as follows: We give an overview of Related Work in Section 2 and describe our training set and our model design in Section 3. In Section 4, we show our evaluation setup and how to reproduce our experiments. We report our results in Section 5 and discuss ablation studies in Section 6 before concluding our work in Section 7.

2 Related Work

Scientific citations have been widely used as training data for generalizable document embedding models. Prior work such as SPECTER (Cohan et al., 2020) and SciNCL (Ostendorff et al., 2022) learn document representations from abstract–abstract citation pairs derived from the global citation graph. To keep only high-quality abstract pairs, (Liang et al., 2025) extend this approach by using the corresponding local citation sentence to filter out potentially noisy pairs before large language model training. In contrast to these approaches, our model directly trains on citance-abstract pairs instead of indirect abstract-abstract supervision.

Leveraging citances for citation recommendation has been introduced by He et al. (2010). Current models typically follow an IR workflow, consisting of a first-stage retriever before reranking the retriever results (Gu et al., 2022). Newer approaches (Medić and Snajder, 2020; Gu et al., 2022) have incorporated the article title and abstract to the query in addition to the citance for a richer flow of information. In our work, we want to examine how well such models perform not only on citation recommendation but also on general IR queries such as found in BEIR (Thakur et al., 2021). Thus, we stick to shorter, asymmetrical queries consisting of citances only.

Dense IR models are usually trained on large-scale annotated datasets such as MS Marco (Bajaj et al., 2016) and evaluated on annotated data as well, such as the BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) benchmarks. In the biomedical domain, however, annotated data is scarce motivating alternative sources. A key contri-

bution in this direction is MedCPT (Jin et al., 2023) which demonstrated that training biomedical retrievers solely on large-scale weak supervision data from PubMed search and click logs achieves strong performance on biomedical IR benchmarks. MedCPT establishes a strong baseline for biomedical retrieval using weak supervision, but relies on proprietary data. In contrast, we use citance-abstract pairs mined from PubMed Central full-text articles as a fully open alternative source of supervision, enabling large-scale training without privacy constraints.

Besides domain-specific fine-tuning, recent approaches (Li et al., 2023; Xiao et al., 2023; Zhang et al., 2025) have also explored scaling dense retrievers through larger models and larger training datasets. The gte model family (Li et al., 2023) employs a pre-training step by training first on automatically mined query-document pairs before fine-tuning on high-quality, manually annotated retrieval pairs. Among the pre-training data, they also employ a set of citance-abstract pairs from the S2ORC corpus (Lo et al., 2020) as auxiliary supervision signal. (Sinha et al., 2025) continue fine-tuning the gte models on pairs of document abstracts and synthetic queries targeting those. To assess how weak supervision from one source compares to these larger models, we also compare our approach against these embedding models.

In contrast to prior work, we investigate whether large-scale citance–abstract pairs alone are sufficient to train biomedical retrievers that generalize across both citation recommendation and general information retrieval tasks.

3 Methods

3.1 CiteRec Overview

We first explain our methodology to obtain biomedical citances before describing the two main components of our CiteRec model: (i) the lightweight dense retriever to quickly obtain an initial set of candidate documents and (ii) the cross-encoder reranker to refine the ranking of the candidates.

3.2 Training Data

For training, we use the OPCitance dataset created by Hsiao and Torvik (2023) who extracted all pairs of citances and referenced documents from PubMed Central until the cutoff date of May 2019. Compared to other full-text corpora such as S2ORC (Lo et al., 2020), OPCitance is solely focused on the biomedical domain. In our setup, citances function as queries while the title and abstract of references serve as the documents to be retrieved. We employ two steps for data pre-processing: First, we filter out all documents that were cited but were not assigned any valid PubMed ID as the reference. Secondly, we filter out citation contexts longer than 64 tokens discarding long queries (i) with potentially malformed or irrelevant context and (ii) to ensure that the reranker obtains sufficient document context (we allow up to 512 tokens for the combination of citation and reference document). After preprocessing, we retain roughly 62M citance-abstract pairs and show some example citance-abstract pairs in Table 1.

3.3 Retriever

For our CiteRec retriever, we train separate query and document encoders both initialized from a PubMedBERT-base model (Gu et al., 2021). We take the final hidden state of the [CLS] token as the respective query and document embeddings. For retriever training, we optimize a query-to-document loss function

$$\ell_i^{q2d} = -\log \frac{\sum_{k \in Pos(q_i)} sim(q_i, d_k)}{\sum_{m=1}^{|B|} sim(q_i, d_m)},$$

whereas $|B|$ denotes the batch size. q_i is a citance query, d_i its associated positive document and the similarity $sim(q, d)$ is implemented by the dot product of both embeddings. Extending the standard InfoNCE loss formulation (Oord et al., 2018), we allow multiple positive documents $d_k \in Pos(q_i)$ for a given query q_i. For negatives, we use other citation pairs in the same batch as in-batch negatives. However, instead of relying on batching random pairs together, we group citation pairs from the same full text into the same batch. This allows us to mine potentially more informative negatives as documents referenced by the same paper are more likely to be thematically related, hence acting as harder negatives than other, random documents.

Similarly to MedCPT, we also define a complementary document-to-query loss ℓ_i^{d2q} which models relevance of individual queries to a given document d_i. We combine both losses to a batch loss ℓ_B:

$$\ell_B = \frac{\alpha}{|B|} \sum_{1}^{|B|} w_i \ell_i^{q2d} + \frac{1-\alpha}{|B|} \sum_{1}^{|B|} w_i \ell_i^{d2q}$$

where the hyperparameter $\alpha = 0.8$ controls the influence of each individual loss term and the weight w_i models the importance of citing paper and cited references according to their citation counts. To normalize citation counts across fields, we use the Relative Citation Ratio rcr[1] (Hutchins et al., 2016) and set the weights $w_i = \log(1+rcr_i)$. The rcr value takes into account the year of a publication and its citation numbers relative to its field.

3.4 Reranker

For CiteRec reranker training, we initialize a cross-encoder from the PubMedBERT-base model (Gu et al., 2021). The cross-encoder optimizes the same InfoNCE loss ℓ_i^{q2d} as the retriever but uses random in-batch negatives as the reranker can leverage random negatives more effectively due to its cross-attention mechanism. It takes as input the sequence [CLS] q_i [SEP] d_i [SEP] with the query q_i, the document d_i and the special BERT tokens [CLS] and [SEP], and calculates the similarity score

$$sim(q, d) = W^T e([CLS]) + b$$

where e([CLS]) is the final hidden state embedding of the [CLS] token and $W^T \in \mathbb{R}^d, b \in \mathbb{R}$ are learnable weights. The reranker takes into account the top-100 documents returned by our retriever.

4 Experiments

4.1 Evaluation Datasets

We conduct our experiments on three diverse biomedical IR and document embedding benchmarks.

BEIR (Thakur et al., 2021) is a zero-shot information retrieval benchmark to assess the generalization capabilities of newly developed retrieval models. Similarly to (Jin et al., 2023), we evaluate our model on the same subset of five BEIR tasks as (Jin et al., 2023) which include TREC-COVID (Voorhees et al., 2021), containing questions related to the COVID-19 pandemic, NFCorpus (Boteva et al., 2016), questions regarding nutrition facts, BioASQ (Tsatsaronis et al., 2015), a general-purpose biomedical question answering dataset over the entirety of PubMed, SciFact (Wadden et al., 2020), retrieving supporting documents that verify scientific claims and SciDocs (Cohan et al., 2020), finding similar documents given the title and abstract of a paper.

SciRepEval (Singh et al., 2023) is aimed at assessing the quality of document embedding models and includes tasks besides retrieval/search, such as classification and regression. Exemplary classification tasks include assigning documents to a MeSH descriptor and to its field of study. Exemplary regression tasks include predicting the citation count and the year of a publication. SciRepEval distinguishes between tasks whose datasets were used during the training of their embedding models (in-train) and tasks that were not (out-of-train, none of our evaluated models were trained on any of the in-train datasets.).

Our third benchmark deals with the task of recommending references for given citation contexts from PubMed. For this, we create a new, independent test set called CitancePlus mined from PubMed Central full-text articles published after May 2019 to prevent overlaps with the OPCitance dataset used for our model training. To obtain query-document pairs, we randomly sampled 10,000 full texts published between June 2019 and June 2024 and for each of them, sampled up to 10 in-line citations as queries. We then extract all referenced PubMed articles of each citance as valid reference documents. The resulting dataset contains 32,000 query-document pairs in the development split and 93,000 pairs in the test split. We used the official PubMed baseline of 2025 as the document corpus (containing all PubMed abstracts until January 2025).

4.2 Baselines

We compare the results of our CiteRec model to several baselines: BM25 (Robertson and Zaragoza, 2009) is a lexical retriever serving as a strong baseline in IR tasks (Thakur et al., 2021). MedCPT (Jin et al., 2023) is the baseline closest to ours, training a dual bi-encoder setup with a separate query and document embedder initialized from PubMedBERT-base, while also training a cross-encoder reranker on top. While we exclusively train on freely available citance-abstract pairs, MedCPT exclusively trains on a large proprietary set of 255 M PubMed search logs containing pairs of user query and abstract[2].

[1]We download the most recent dump from March 2025 under https://nih.figshare.com/articles/dataset/iCite_Database_Snapshot_2025-03/28789178.

[2]We use the MedCPT models provided via Huggingface (https://huggingface.co/ncbi/MedCPT-Article-Encoder) which differ a bit from the results reported in the original paper.

SciNCL (Ostendorff et al., 2022) and SPECTER (Cohan et al., 2020) are document embedding models that both leverage the citation graph for training similar to our approach. Note that they are using abstract-abstract pairs instead of citances for training and thus have not been explicitly designed for short search queries.

The last group of baselines are retriever models trained on a much larger mixture of data including search logs, web links and manually annotated pairs of questions and answer documents: (i) BERT-based encoder models such as the gte family (Li et al., 2023) and bge (Xiao et al., 2023) have around 100 to 300 million model parameters resulting in a similar size as MedCPT and CiteRec but training on more diverse retrieval pairs; the gte models use around 800 million retrieval pairs during unsupervised pre-training and around three million high-quality pairs for additional fine-tuning, the bge models use a similar pipeline of pre-training and subsequent fine-tuning with around 200 million pairs. (ii) Decoder-only embedding models like Qwen3 (Zhang et al., 2025) are much larger in size with around 8 billion model parameters and trained on a carefully selected set of around 150 million retrieval pairs during pre-training and 12 million pairs during fine-tuning.

4.3 Implementation Details

We conducted training for the retriever on two NVIDIA A100 GPUs with a batch size of 128, gradient accumulation step size of 4, using bf16 precision, a cosine scheduler with a linear warm-up phase, and a target learning rate of 2e-6. We trained the model for one epoch resulting in a runtime of around four days. As cross-encoder reranker training converges much faster, we applied early stopping on the development set to find our best model there. We set the batch size for the reranker training to 1, the gradient accumulation step size to 32 and the target learning rate of 5e-6. The CitancePlus corpus, the CiteRec retriever, the CiteRec reranker and the accompanying code is available at under `https://github.com/WangXII/CiteRec`.

5 Results

In Table 2, we report the results of our CiteRec model and of the baselines on the subset of biomedical tasks in BEIR. As expected, even though the SciNCL model was trained on abstract-abstract citation links, it does not perform well on the biomedical BEIR subset except for SciDocs. Our CiteRec retriever is better suited for IR queries than SciNCL by learning from citance-abstract pairs instead.

Looking at individual subtasks and retriever-only models, our CiteRec retriever achieves particularly good results on the SciFact dataset with an NDCG@10 score of 0.765 outperforming all other first-stage retrievers. On average across all subtasks, our CiteRec retriever performs competitively against the MedCPT retriever (0.425 NDCG@10 compared to 0.431) showing that training on public citation data alone can match up to training on proprietary search logs. Adding a reranker on top of the CiteRec retriever improves retrieval performance on all examined datasets resulting in an average increase of around 5.3 percentage points (pp). However, here, bigger gaps are observed on the BioASQ and TREC-COVID datasets when compared to the corresponding MedCPT reranker indicating that real-world user queries might help more in reranker training for biomedical question answering tasks. Compared to the gte and bge models which trained on a richer mixture of (high-quality) data, the MedCPT and CiteRec retrievers trained only on one (noisy) source of data show lower performances. However, by adding the rerankers, both models close the gap with CiteRec achieving comparable performances and MedCPT even surpassing them.

When evaluating quality of document embeddings on the SciRepEval benchmark (see Table 3), we can see that our CiteRec retriever is able to outperform the MedCPT retriever by over 3.3 pp (also reflecting the more citation-focused subtasks of SciRepEval). As expected, the SPECTER and SciNCL models perform well on this task as they are designed to jointly optimize document abstract embeddings. SciNCL achieves an average score of 0.688 and SPECTER a score of 0.675 compared to 0.673 for CiteRec. Still, the bge-large model achieves the best results overall with a score of 0.694. CiteRec performs competitively overall with the models trained on a larger set of data, achieving the joint second best results on the In-Train part of SciRepEval together with gte-large (both 0.559).

Our last main experiment focuses on the CitancePlus dataset ranking relevant abstracts in PubMed given a local citation context. We present the results on CitancePlus in Table 4 measured by NDCG and MAP scores @10 and @100. We observe that BM25 is a strong baseline for this task outperforming neural models like gte-base and Med-

	Param Size	TREC-COVID	NFCorpus	BioASQ	SciFact	SciDocs	Avg
BM25	-	0.656	0.325	0.465	0.665	0.158	0.454
gte-base	100 M	0.676	0.374	0.355	0.753	0.229	0.477
gte-large	300 M	0.687	**0.381**	0.358	0.750	**0.237**	0.482
bge-large	300 M	**0.744**	0.343	0.400	0.736	0.218	0.488
SciNCL	100 M	0.342	0.228	0.108	0.568	0.194	0.288
MedCPT	100 M	0.616	0.365	0.325	0.736	0.114	0.431
+ reranker	200 M	0.727	0.365	**0.498**	**0.793**	0.176	**0.512**
CiteRec (Ours)	100 M	0.601	0.299	0.314	0.765	0.148	0.425
+ reranker (Ours)	200 M	0.675	0.347	0.392	0.792	0.184	0.478

Table 2: Biomedical BEIR results measured by NDCG@10. Best results in each column in bold, second best underlined.

	In-Train	Out-of-Train	Avg
gte-large	0.559	0.720	0.683
bge-large	**0.575**	**0.736**	**0.694**
SPECTER	0.547	0.720	0.675
SciNCL	0.556	0.734	0.688
MedCPT	0.531	0.678	0.640
CiteRec (Ours)	0.559	0.713	0.673

Table 3: SciRepEval Results. Best results in each column in bold, second best underlined.

CPT across all metrics and gte-large in all but the NDCG@100 score (0.193 compared to 0.201). The MedCPT retriever struggles with the citation recommendation task only reaching an NDCG@10 of 0.097. Our CiteRec retriever model outperforms all other similarly sized retriever models on this task by more than 3 pp across all metrics, showing that the citation recommendation task benefits a lot from domain-specific finetuning.

We compare to the Qwen3 model only on the citation recommendation task as the gte and bge models have already proven to be challenging baselines on the biomedical BEIR and SciRepEval benchmarks and evaluating large decoder-based embedding models at PubMed scale is computationally expensive. On CitancePlus, the Qwen3-Embedding-8B model slightly outperforms our CiteRec first-stage retriever with an NDCG@10 score of 0.216 compared to 0.206. However, after adding the CiteRec reranker on top of the retriever, our overall CiteRec model even outperforms the Qwen3-Embedding-8B model by 2.4 pp in NDCG@10 score while being significantly more efficient in terms of parameter size and inference cost.

6 Discussion

In our first ablation study, we examine how retrieval performance of our CiteRec retriever changes when we modify our negative sampling from grouped batching of negatives from the same publication to random in-batch negatives and replacing the rcr values as individual weights for a document by the default value of 1. We report the corresponding results in Table 5. We observe that grouped batching improves NDCG@10 scores on both datasets, with the improvements being more pronounced on the CitancePlus dataset (0.206 vs 0.164) compared to the biomedical subset of BEIR (0.425 vs 0.412). This suggests that retriever training likely benefits from harder examples coming from the same documents whereas random negatives are easier and saturate model performance more quickly.

Our second ablation study examines the capabilities of our CiteRec models in a stricter setting: We constructed our original CitancePlus test dataset in such a way that no test query overlaps with any query encountered in the training set. However, the model might still have incorporated a slight bias toward documents already encountered in the OPCitance training data. To minimize such potential effects, we repeated the CitancePlus experiments in a stricter setting removing all citance-abstract pairs from the test dataset where the abstract document has already occurred at least once in the training dataset. We reduced the initial 93 thousand citance-abstract pairs to a subset of around 25 thousand after applying this filtering step.

Table 6 summarizes the results on the filtered test dataset. We see that the performance of BM25 increases on the filtered test dataset from 0.161 NDCG@10 to 0.185 NDCG@10. Our CiteRec

	Param Size	NDCG@10	NDCG@100	MAP@10	MAP@100
BM25	-	0.161	0.193	0.132	0.139
gte-base	100 M	0.143	0.178	0.121	0.127
gte-large	300 M	0.157	0.201	0.125	0.134
bge-large	300 M	0.170	0.212	0.137	0.145
Qwen3-Embedding-8B	8 B	<u>0.216</u>	<u>0.268</u>	<u>0.175</u>	<u>0.186</u>
SciNCL	100 M	0.076	0.106	0.057	0.063
MedCPT	100 M	0.097	0.132	0.075	0.082
+ reranker	200 M	0.174	0.191	0.146	0.149
CiteRec (Ours)	100 M	0.206	0.249	0.168	0.177
+ reranker (Ours)	200 M	**0.240**	**0.276**	**0.200**	**0.209**

Table 4: CitancePlus results. Best results in each column in bold, second best underlined.

	CitancePlus	BEIR
CiteRec	0.206	0.425
CiteRec (BatchNeg)	0.164	0.412

Table 5: Ablation on our CiteRec retriever model to train only using random in-batch negatives. Resulted scores are NDCG@10.

CitancePlus	Test	Filtered
BM25	0.161	0.185
gte-large	0.157	0.162
Qwen3-Embedding-8B	<u>0.216</u>	<u>0.221</u>
CiteRec (Ours)	0.206	0.205
+ reranker (Ours)	**0.240**	**0.248**

Table 6: Ablation on the CitancePlus dataset removing any paper from the test dataset which was also cited at least once in the training data. Result scores in NDCG@10. Best results in each column in bold, second best underlined.

retriever decreases only minimally in NDCG@10, from 0.206 to 0.205, indicating at most a slight bias toward documents already seen in the OPCitance training data. The larger retriever models gte-large and Qwen3-Embedding-8B improve by 0.5 pp in NDCG@10 score. However, when adding our reranker model to the top-100 retrieved documents, overall NDCG@10 score improves again by 0.8 pp compared to the original test dataset indicating that any such bias is not as pronounced in the reranker model.

7 Conclusion

In this work, we introduced a biomedical retriever-reranker model trained on a large set of PubMed Central citation-abstract pairs. We have shown that the resulting model generalizes well on general biomedical IR benchmark such as the BEIR subset when compared to models of similar size and training data such as MedCPT and outperforms both these baselines as well as substantially larger state-of-the-art models (e.g., Qwen3-Embedding-8B) when compared on the in-domain citation recommendation task.

Limitations

Training on citance-abstract pairs is inherently noisy as citations do not always match what is written in the referenced texts and not all information referred to in a given context is fully included in the title and abstract of the referenced document (Sarol et al., 2024). We did not further investigate any methods to filter out high-quality citations from lower-quality ones.

We also limited our experiments to smaller BERT-based model (100M parameters) and did not

train on larger decoder-only retrievers due to limited computational resources. It remains unclear whether scaling to larger models would improve performance or lead to overfitting on the noisy training data.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft [RTG2424].

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, pages 716–722. Springer.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S 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.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Tobias J Erb. 2011. Carboxylases in natural and synthetic microbial pathways. *Applied and environmental microbiology*, 77(24):8466–8477.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and SciBERT-based reranking. *Preprint*, arxiv:2112.01206 [cs].

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.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430.

Tzu-Kun Hsiao and Vetle I. Torvik. 2023. OpCitance: Citation contexts identified from the PubMed central open access articles. 10(1):243.

B. Ian Hutchins, Xin Yuan, James M. Anderson, and George M. Santangelo. 2016. Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level. 14(9):e1002541.

ML Janssen, DLM Van Broekhoven, JMM Cates, WM Bramer, JJ Nuyttens, Alessandro Gronchi, S Salas, S Bonvalot, DJ Grünhagen, and C Verhoef. 2017. Meta-analysis of the influence of surgical margin and adjuvant radiotherapy on local recurrence after resection of sporadic desmoid-type fibromatosis. *Journal of British Surgery*, 104(4):347–357.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arxiv:2308.03281 [cs].

Yuan Liang, Massimo Poesio, and Roonak Rezvani. 2025. Beyond citations: Integrating finding-based relations for improved biomedical article representations. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 297–306. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983. Association for Computational Linguistics.

Zoran Medić and Jan Snajder. 2020. Improved local citation recommendation based on context enhanced with global information. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 97–103. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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. Association for Computational Linguistics.

Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. Tripclick: the log files of a large health web search engine. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2507–2513.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

M Janina Sarol, Shufan Ming, Shruthan Radhakrishna, Jodi Schneider, and Halil Kilicoglu. 2024. Assessing citation integrity in biomedical publications: Corpus annotation and NLP models. page btae420.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566. Association for Computational Linguistics.

Aarush Sinha, Pavan Kumar S, Roshan Balaji, and Nirav Pravinbhai Bhatt. 2025. BiCA: Effective biomedical dense retrieval with citation-aware hard negatives. *Preprint*, arxiv:2511.08029 [cs].

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Yoshiaki Yamada, Yasusuke Inoue, Kogenta Nakamura, Katsuya Naruse, Shigeyuki Aoki, Tomohiro Taki, Motoi Tobiume, Kenji Zennami, Remi Katsuda, Kouji Hara, and 1 others. 2009. Long-term results and management of ureteral transitional cell carcinoma using the holmium: Yag laser via rigid-ureteroscopy. *Oncology reports*, 21(2):345–349.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arxiv:2506.05176 [cs].