# AnonymizedModelName[1]: tiny and powerful multilingual encoder for scientific texts

Paper #2662

**Abstract.**

LLM-based representation learning is widely used to build effective information retrieval systems, including scientific domains. For making science more open and affordable, it is important that these systems support multilingual (and cross-lingual) search and do not require significant computational power. To address this we propose AnonymizedModelName-tiny (AMN-tiny), light multilingual encoder trained from scratch on 44M abstracts (15B tokens) of research papers and then tuned in a contrastive manner using citation data. AMN-tiny outperforms SciNCL, English-only SOTA-model for scientific texts, on 13/24 tasks, achieving SOTA on 7, from SciRepEval benchmark. Furthermore, AMN-tiny is much more effective than SciNCL: it is almost 5x smaller (23M parameters vs 110M), having approximately 2x smaller embeddings (312 vs 768) and 2x bigger context length (1024 vs 512). In addition to the tiny model, we also propose the AMN-small (61M parameters and 768 embeddings size), which is more powerful and can be used for complicated downstream tasks. We further study different ways of contrastive pre-training and demonstrate that almost SOTA results can be achieved without citation information, operating with only title-abstract pairs.

## 1 Introduction

Processing large amounts of scientific papers to obtain knowledge in a specific domain can be time-consuming and resource-intensive. Acceleration and enrichment of the discovery process is essential for making scientific research more approachable. Recently, valuable advancements in information retrieval and document processing were made with the help of Large Language Models (LLM). However, most language models require a significant amount of computational power, even for fine-tuning and inference. This can be a challenge for scientists from small labs and individual researchers who have limited computational resources. To mitigate this issue and make LLM-based tools for research available to a wide community, we propose a new, comparatively small model designed specifically for the scientific domain that can be used even without a GPU. On the contrary, even if significant computation resources are available, our small model can filter most of irrelevant documents, leaving a more complicated reranking task with a much shorter list of options.

While many scientific articles are written in English, a substantial part of research is still done in other languages. Moreover, some scientists who do not speak English fluently cannot access scientific texts to the full extent. To overcome this problem, we propose a cross-lingual language model. While maintaining SOTA performance for English, our model can perform cross-lingual search. As for the second language, we chose Russian because it does not belong to the group of Romano-Germanic languages, which makes the cross-lingual task more difficult. We assume that training a cross-lingual language model that supports, for example, French or German in addition to English is less complicated and can be successfully done using the methods examined in this paper.

One of the most revealing signals in scientific articles is citation information. Many recent works, like SPECTER [3] and SciNCL [16], adapt it in different ways, including sophisticated approaches based on Graph Neural Networks. In the meantime, for low-resource languages, citation information may be hard to find, especially in cross-lingual scenarios. However, titles and abstracts are usually available and also have translations to English. Thus, we propose a model that is trained in a contrastive manner with title-abstract pairs similarly to the first stage of E5 [24] model. Our approach allows us to achieve 95% on the SOTA level of benchmarks for evaluating scientific document representations with no need to gather citation information.

All in all, our main contributions can be summarized as follows:

1. We propose a model called AnonymizedModelName-tiny[1] (AMN-tiny), that outperforms SciNCL [16], English-only SOTA-model for scientific texts, on 13 out of 24 tasks from SciRepEval benchmark, achieving SOTA on 7 of them. AMN-tiny is almost 5 times smaller (23M vs 110M), has more than twice smaller embedding size (312 vs 768) and twice bigger context length (1024 vs 512).
2. We also propose a bigger model AMN-small (61M), which is still almost half the size of SciNCL. With its increased capacity, AMN-small can be successfully tuned for various downstream tasks.
3. We present a two-stage method of training the AMN family of models, which includes MLM and contrastive pre-training. This technique can be used in cross-lingual scenarios for any language pair.
4. We show that contrastive pre-training with title-abstract pairs can be nearly as effective as pre-training with citation pairs.
5. We demonstrate that AMN models are superior to a wide list of competitors, including multilingual E5, LaBSE, SciNCL, SciRus, and ruSciBERT on ruSciBench [5] — a benchmark of scientific articles from the Russian electronic library. We further notice that if there are two abstracts for a paper (e.g., in Russian and in English), random choice of language allows for high scores in cross-lingual search even without direct training of the translation model.

[1] The model name will be revealed after a blind review. Weights were added to the supplementary material on the anonymous Google Drive (the maximum size of supplementary materials on the easychair.org website is only 100M, which is not enough). We also attach the Jupyter Notebook with code to reproduce results on the ruSciBench and SciRepEval benchmarks. This notebook was prepared to run easily in Google Colab.

## 2 Related work

Semantic search is commonly deployed in information retrieval systems. For solving the task, different Transformer-based [23] models that are fine-tuned on domain-specific texts are used. For example, Sentence-BERT [18] can be adapted to specific domains and used for different retrieval tasks. Further, such techniques as self-retrieval [21] can also be used. In this method, the index of the given corpus is built using self-supervised learning, then the self-retrieval model generates the index and passage in natural language, and finally, self-assessment is used to score and rank the generated passages.

Semantic search in multilingual (and cross-lingual) scenarios is even less explored and requires strong multilingual alignment. For example, Multilingual sentence-level semantic search algorithm MAML-Align [13] distils knowledge from a Teacher model, specialized in transferring from monolingual to bilingual semantic search, to a Student model, which meta-transfers from bilingual to multilingual semantic search.

Many embedding models are based on E5 [24] — text embeddings that transfer well to a wide range of tasks and domains. Multilingual E5 embeddings [25] can be more easily transferred to different languages using supervised fine-tuning techniques with paired bilingual data. The original mE5 model is pre-trained in a contrastive manner on 1 billion multilingual text pairs, including 50 million pairs retrieved from Semantic Scholar (the Semantic Scholar Open Research Corpus, S2ORC [11]) and then fine-tuned on a combination of specific datasets. For mE5 pre-training, InfoNCE contrastive loss [15] with only in-batch negatives is used. We utilize contrastive training similar to E5 in the second stage of AMN training (section 4.2).

We focus on information retrieval for scientific applications, inducing knowledge from scientific domain via training on eLibrary and Semantic Scholar (S2AG) datasets. Recent works on encoding scientific texts include models such as SPECTER [3], SciNCL [16] and SPECTER2 [19]. SPECTER unravels its document-level representation capabilities by involving citation graphs. Starting from the trained SciBERT [1] used as an initialization model, SPECTER is pretrained on the citation objective using cite and co-cite pairs of papers. However, only title and abstract pairs are required for inference with no need to include any citation information, which makes it possible to use SPECTER for the papers not cited yet.

While SPECTER relies on discrete direct citations, SciNCL [16] uses controlled nearest neighbor sampling over citation graph embeddings for contrastive learning and outperforms SPECTER [3], SciBERT [1] and CiteBERT [26] on the SciDocs benchmark. Close to SciNCL is Citeomatic [2] that relies on bag-of-words representations and learns a triplet-based document embedding model, where positive samples are papers cited in the query.

The SPECTER2 family of models includes not only base models from which general-purposed embeddings can be retrieved but also a retrieval specific adapter that can be used for tasks where, given a paper query, other relevant papers have to be retrieved from a corpus. SPECTER2 is built upon SciRepEval [19] and evaluated on this benchmark too.

There have also been other attempts to train a model related to a scientific domain, for example, based on BERT and GPT architectures. Thus, in [26] SciBERT was fine-tuned using CiteWorth — a dataset with labelled cite-worthiness, balanced across 10 diverse scientific fields, and curated from S2ORC data. Three ways of tuning SciBERT are explored: MLM fine tuning on CiteWorth, fine-tuning for the task of citeworthiness detection, and joint option combining the first two approaches. On the other hand, SciGPT2 is introduced in [12] and is used to build SciGEN — a domain-adapted left-to-right generative language model. However, these models are left behind by SPECTER and SciNCL on scientific-oriented benchmarks, so we do not include them in our comparisons.

We compare our approach not only to model oriented to scientific domain, but also to general ones, including MPNet [20], E5 [24, 25] and Language-agnostic BERT Sentence Embedding (LaBSE [4]). MPNet combines masked language modeling and permuted language modeling, inheriting the advantages of the two approaches. It was trained on a large-scale dataset (over 160GB text corpora) and fine-tuned on a variety of down-streaming tasks (GLUE, SQuAD, etc.). Multilingual E5 is another powerful encoder model. At the moment of writing this manuscript, a model trained on top of E5-mistral-7b-instruct held first place on the MTEB benchmark [14]. Finally, LaBSE supports 109 languages and performs strongly even on those languages that LaBSE has not seen during training. It achieves 83.7% bi-text retrieval accuracy over 112 languages on Tatoeba [22], while still performing competitively on monolingual transfer learning benchmarks. This makes LaBSE a good baseline for multi- and cross-lingual encoders.

All the aforementioned models are usually evaluated on English benchmarks. Our paper is devoted to cross-lingual models, which we explore in the context of English-Russian language pair. Thus, we further study approaches designed specifically for encoding scientific papers written in Russian, such as RuSciBERT [7, 8] and SciRus [6]. RuSciBERT is an encoder with 125M parameters that was trained on 6.5GB of Russian texts. The SciRus-tiny model is much smaller and contains only 23M parameters. It was trained on eLibrary data with contrastive techniques similarly to the way we train AMN models during the second stage.

## 3 Datasets and Benchmarks

### *3.1 Datasets.*

We use two multilingual datasets for the experimentation: a set of papers and their relatedness from Semantic Scholar (S2AG [11]) and a set of scientific papers from the web-site eLibrary.ru.

#### *3.1.1 Semantic Scholar Academic Graph Dataset (S2AG)*

S2AG dataset [11] is a heterogeneous knowledge graph of Semantic Scholar articles that also comprises information about authors and citations. Originally, it consists of 205M publications, 121M authors, and nearly 2.5B citation edges, integrating knowledge from Crossref, PubMed, Unpaywall, and other sources. For our experiments, we took 30M title-abstract pairs out of this dataset. This amount of data appeared to be sufficient for small-sized models even for cross-lingual tasks. The chosen subset contains texts in different languages with the majority of them ($83.26\%$) being in English (Table 1). The tail on languages distribution include Chinese, French, Spanish, German and others. Russian language is present too, making up just $0.42\%$. The distribution over different academic fields in our chosen subset is presented in Table 3 and statistics about average title and abstract lengths is shown in Table 2.

#### *3.1.2 eLibrary*

To adapt our model for the Russian language, we additionally used 18M title-abstract pairs from the Russian electronic library of scientific papers called eLibrary. It contains $8.6M$ papers in Russian

| Lang. | Count ↓ | Percent ↓ |
|---|---|---|
| EN | 25394300 | 83.26% |
| ZH | 844850 | 2.77% |
| FR | 802150 | 2.63% |
| ES | 728950 | 2.39% |
| ID | 603900 | 1.98% |
| PT | 512400 | 1.68% |
| DE | 347700 | 1.14% |
| KO | 274500 | 0.9% |
| RU | 128100 | 0.42% |

**Table 1:** Paper records in S2AG for different languages (only languages with more than $0.4\%$ presence are listed)

| Lang. | Text type | Count type | Percentile | | | |
|---|---|---|---|---|---|---|
| | | | 50th | 90th | 95th | 99th |
| EN | title | Character | 81 | 135 | 154 | 198,42 |
| | title | Word | 11 | 18 | 21 | 28 |
| | abstract | Character | 972 | 1924 | 2388 | 4012 |
| | abstract | Word | 144 | 285 | 354 | 607 |

**Table 2:** Average counts of characters and words for S2AG titles and abstracts

($\approx$2B tokens) and $8.8M$ papers in English ($\approx$1.2B tokens). A detailed analysis of average title and abstract lengths is presented in Table 5. English titles and abstracts are generally longer than Russian ones, while the number of articles written in English is just by $1\%$ greater than papers written in Russian. Most papers in eLibrary are written in Russian or English ($97.3\%$). Other papers are written in Ukrainian, Chinese, German, etc. (see Table 4 for detailed language distribution). Also, most papers written in Russian have translations into English. We also analyze the distribution over fields of study and show it in the Table 6

### *3.2 Benchmarks.*

There are several benchmarks for evaluation: some of them are designed specifically for the scientific domain, and others are just general ones for measuring encoders performance.

1. **SciDocs [3]** consists of four groups of tasks: document classification (medical subject headings classification and paper topic prediction), citation prediction (direct citations and co-citations), prediction of user activity (prediction of co-views and co-reads), and paper recommendation.
2. **SciRepEval [19]** is a benchmark of 24 tasks across four formats (classification, regression, proximity, and ad-hoc search) to train and evaluate multi-task embeddings of scientific papers. SciRepEval includes SciDocs as a subset, except for the recommendation task, which has limited power to distinguish different embeddings due to its noisiness. SciRepEval enhances SciDocs in terms of having more realistic tasks (search, author disambiguation, and paper-reviewer matching instead of nearest neighbor tasks) and increasing the diversity of the tasks (four of the tasks in SciDocs have over $0.99$ model-performance correlations among them). Moreover, SciRepEval has tasks that can be used for training, while all SciDocs tasks are designed to be used only for evaluation.
3. We also use a benchmark of Russian scientific papers called **RuSciBench [5]** which consists of two types of tasks: classification and translation search. Both tasks are available for English and Russian languages. While classification is also divided into two categories: OECD and GRNTI (different systems of arranging scientific articles), translation search checks whether the embedding of a paper is close to the embedding of the same paper written in another language.

| Field of Study | Count ↓ | Percent ↓ |
|---|---|---|
| n/a | 60.0M | 27.37% |
| Medicine | 31.8M | 14.50% |
| Biology | 20.4M | 9.30% |
| Physics | 11.6M | 5.29% |
| Engineering | 10.2M | 4.65% |
| Computer Science | 9.7M | 4.42% |
| Chemistry | 9.1M | 4.15% |
| Education | 7.4M | 3.38% |
| Material Science | 7.4M | 3.38% |
| Environmental Science | 7.0M | 3.19% |
| Economics | 6.2M | 2.83% |
| Psychology | 6.2M | 2.83% |
| Agricultural and Food Sciences | 5.9M | 2.69% |
| Business | 5.6M | 2.55% |
| Mathematics | 3.7M | 1.69% |
| History | 3.4M | 1.55% |
| Political Science | 2.9M | 1.32% |
| Art | 2.8M | 1.28% |
| Geology | 2.6M | 1.19% |
| Sociology | 1.4M | 0.64% |
| Philosophy | 1.4M | 0.64% |
| Law | 1.1M | 0.50% |
| Linguistics | 1.1M | 0.50% |
| Geography | 350k | 0.16% |

**Table 3:** Paper records in S2AG for different academic fields

| Lang. | Count ↓ | Percent ↓ |
|---|---|---|
| EN | 8827195 | 49.26% |
| RU | 8608748 | 48.04% |
| UK | 134965 | 0.75% |
| ZH | 99863 | 0.56% |
| UN | 77354 | 0.43% |
| DE | 33144 | 0.18% |

**Table 4:** Paper records in eLibrary for different languages (only languages with more than $0.1\%$ presence are listed)

## 4 Model

We propose several light multilingual encoders designed specifically for scientific text representation. Starting with random initialization of RoBERTa [10] model, we utilize a two-stage training process: first, pre-training on the masked language modeling task with dynamic masking, second, contrastive training on paired data.

### *4.1 Stage 1: MLM Pre-training*

Our base model is RoBERTa [10] with 3 encoder blocks. We trained three variants of this base architecture: **tiny** with 23M parameters and embedding size equal to 312, **small** (61M parameters and embeddings of size 768) and **base** (85M parameters and embeddings of size 1024). Tiny and base models were trained for 2 epochs, while small models were trained for 4 epochs, but we noticed no significant improvements in their metrics after the second epoch. Thus, we report average SciDocs and RuSciBench scores for only 2 epochs of training for all the models in Fig. 1. We keep track of SciDocs average score instead of SciRepEval during training because SciDocs results are sufficient to get insights on how well the training goes, and

| Lang. | Text type | Count type | Percentile | | | |
|---|---|---|---|---|---|---|
| | | | 50th | 90th | 95th | 99th |
| ru | title | Character | 80 | 128 | 145 | 184 |
| ru | title | Word | 9 | 14 | 16 | 21 |
| ru | abstract | Character | 441 | 1284 | 1695 | 2330 |
| ru | abstract | Word | 50 | 151 | 201 | 278 |
| en | title | Character | 91 | 142 | 160 | 199 |
| en | title | Word | 12 | 19 | 22 | 27 |
| en | abstract | Character | 945.5 | 1808 | 2062 | 2649 |
| en | abstract | Word | 138 | 263 | 302 | 392 |

**Table 5:** Average counts of characters and words for eLibrary titles and abstracts

| Field of Study | Count ↓ | Percent ↓ |
|---|---|---|
| Physical sciences | 1.9M | 12.06% |
| Clinical medicine | 1.6M | 10.03% |
| Economics and business | 1.5M | 9.67% |
| Chemical sciences | 1.4M | 8.58% |
| Biological sciences | 1.4M | 8.52% |
| Educational sciences | 902k | 5.66% |
| Law | 726k | 4.55% |
| Engineering | 582k | 3.65% |
| Languages and literature | 520k | 3.26% |
| Earth and environmental sciences | 499k | 3.13% |
| Civil engineering | 478k | 3.00% |
| Agriculture, forestry, fisheries | 434k | 2.72% |
| Mathematics | 422k | 2.64% |
| History and archaeology | 359k | 2.25% |
| Other social sciences | 312k | 1.96% |
| Philosophy, ethics and religion | 297k | 1.86% |
| Environmental engineering | 279k | 1.75% |
| Nano technology | 231k | 1.45% |
| Psychology | 218k | 1.37% |
| Health sciences | 212k | 1.33% |
| Sociology | 202k | 1.26% |
| Mechanical engineering | 189k | 1.18% |
| Chemical engineering | 181k | 1.14% |
| Social and economic geography | 169k | 1.06% |

**Table 6:** Paper records in eLibrary for different academic fields (only fields with more than 1% presence are listed)

SciRepEval calculations take much more time. The final evaluation is done on the SciRepEval benchmark in Section 4.3.

As expected, the models with more parameters show better results. Moreover, when restricting RuSciBench data to texts written only in Russian, the difference in performance depending on the model size becomes more discernible. Also, dependence on the embedding size is more observable on classification tasks from SciDocs (MAG and MeSH) and is not so distinguishable on retrieval tasks. However, after contrastive training at Stage 2 nearly all the differences disappear (see next section for details).

### *4.2 Stage 2: Contrastive Training*

During the second stage of training, we utilize two types of pairs (title-abstract and cite-cocite) for contrastive learning.

#### *4.2.1 Title-Abstract pairs*

While citation graphs represent document-level relatedness very well, it is often difficult to find citation and co-citation information for non-English articles, which makes training SPECTER [3]

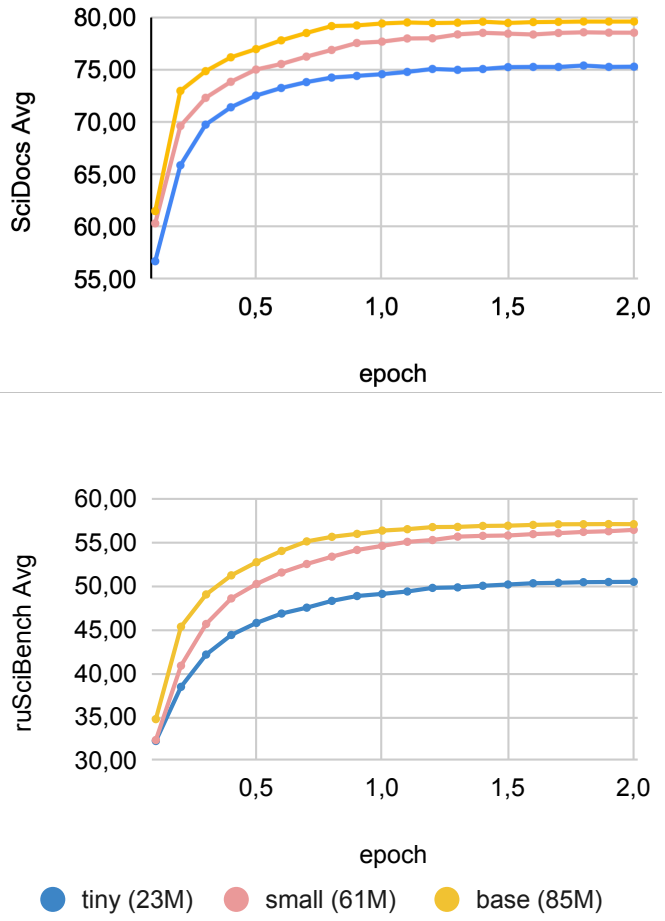


**Figure 1.** Average scores during MLM pre-training for three models of different sizes: tiny (23M), small (61M) and base (85M) on **SciDocs** (upper plot) and **RuSciBench** (lower plot) benchmarks.

or SciNCL [16] for other than English languages an intricate problem. Nevertheless, titles and abstracts of scientific papers are always available, in most cases in two languages: the initial one and the English translation. We collected $30,561,536$ title-abstract pairs from the S2AG dataset and $17,727,817$ pairs from the eLibrary dataset. Having such pairs, we adapt InfoNCE contrastive loss from E5 [24] for the second stage of pretraing:

$$L_{cont} = -\frac{1}{n}\sum_{i} \log \frac{e^{s_\theta(q_i,p_i)}}{e^{s_\theta(q_i,p_i)} + \sum_j e^{s_\theta(q_i,p_{\bar{ij}})}} \quad (1)$$

In eq.1 $s_\theta(q_i, p_i)$ is a scoring function between query $q$ and passage $p$ parametrized by $\theta$, $\{(q_i, p_i)\}_{i=1}^n$ is a collection of title-abstract pairs, and $\{p_{\bar{ij}}\}_{i=1}^n$ is a list of negative passages for the i-th example.

This approach uses no citation information and relies only on title-abstract pairs, but still, the decrease in overall performance compared to citation-based models is no more than 3% (see next section for more details).

After the second stage, we notice that the gap in performance between models of different sizes becomes much smaller. This effect is more salient for retrieval tasks, yet classification tasks show the same tendency too. Talking about evaluation on all the tasks, results on SciDocs (Fig. 2) show that all three discussed models perform on

par in the end of stage 2. Evaluation on RuSciBench (Fig. **??**) still demonstrates a significant gap between tiny and all the other models, while having nearly no difference between small and base models in terms of its performance.

The aforementioned approach has one more important feature — it learns the model to be cross-lingual for free. We achieve that by choosing a random combination of languages, if two of them are present in the training set. This leads to comparatively high results in terms of cross-lingual search.

### *4.2.2 Cite-Cocite pairs*

While being in short supply in many cases (mostly cross-lingual), citation information is still a valuable source of document relatedness that is important for representing scientific texts.

SPECTER [3] and SciNCL [16] rely on citation graphs. SPECTER uses direct citations for training but has a questionable method for sampling negatives (two papers are cocited but do not directly cite each other). In our experimentation, this way of mining pairs led to unstable training and also a significant decrease on RuSciBench benchmark.

SciNCL [16] uses sophisticated way of mining pairs for contrastive learning based on GNN (graph neural network).sophisticated based on GNN (graph neural network) way of mining pairs for contrastive learning. We designed a much simpler method and gathered a huge dataset of pairs of articles directly citing each other (citation) or often appearing together in other papers citations (cocitation). For each paper, we chose 5 random papers, which it cites. As for mining negatives, we follow the approach from E5 [24] embeddings training and treat random articles from the current batch as negative samples. Overall, we prepared $13,307,255$ citation and $61,950,491$ co-citation pairs from the S2AG dataset and $39,988,291$ citation and $33,682,590$ co-citation pairs from the eLibrary data. It turned out that such a simple technique of building pairs outperforms SPECTER and gives approximately the same result as the SciNCL on SciDocs benchmark with the number of parameters being 5 times smaller and the size of embedding being more than 2 times smaller.

We did two attempts at contrastive learning: one with just title-abstract pairs and another with both title-abstract and cite-cocite pairs. It appeared that title-abstract pairs are sufficient to get article representations in most cases. Fig. 2 and Fig. 3 show that addition of cite-cocite pairs improves metrics by 3% at most (small model, 61M parameters), while for the tiny model (23M) the difference does not exceed $1.5\%$. This effect can be clearly seen in lines "<tiny/small>-t-a" and "<tiny/small>-t-a-cite-cocite" at Fig. 2 and Fig. 3, where "t-a" stands for training with only title-abstract pairs and "t-a-cite-cocite" goes for contrastive training with both title-abstract and citation pairs.

Similarly to MLM pre-training, we noticed that scores fluctuate with no significant growth after the second epoch on both benchmarks and thus report results up until the end of the second epoch.

### *4.3 Benchmarking and comparison to competitors*

We compare AMN model trained on title-abstract and cite-cocite pairs to several competitors, including SPECTER [3], SciNCL [16], MPNet [20] and E5 embeddings [24, 25] on SciRepEval benchmark in Table 7. The results show that AMN beats all the presented competitors on 13 out of 24 tasks, achieving SOTA in 7 of them. Moreover, the AMN-tiny model is much more efficient than its counterparts: it is almost 5 times smaller than SciNCL (23M vs. 110M), has smaller embeddings (312 vs. 768) and supports twice longer context (1024 vs. 512).

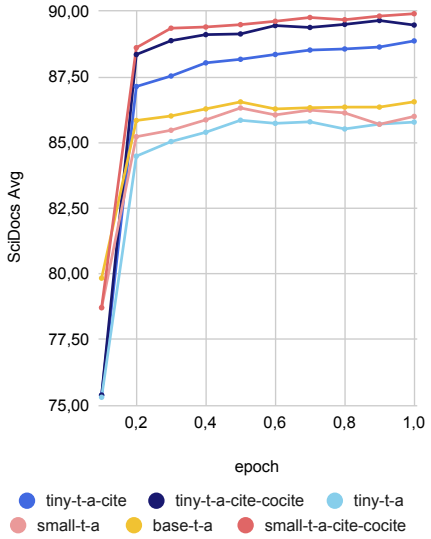


**Figure 2.** Contrastive training with just title-abstract pairs (t-a) and both title-abstract and citation pairs (t-a-cite-cotite) for tiny and small models on the SciDoc

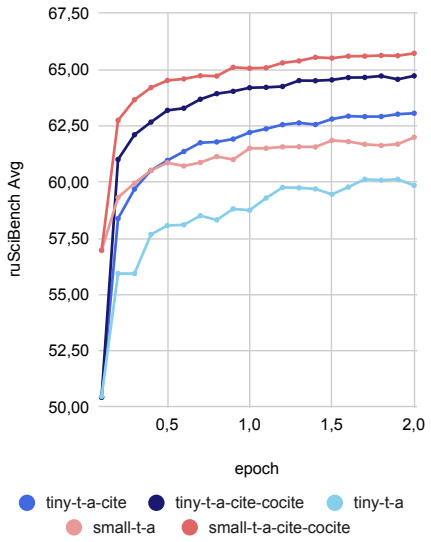


**Figure 3.** Contrastive training with just title-abstract pairs (t-a) and both title-abstract and citation pairs (t-a-cite-cotite) for tiny and small models on the RuSciBench benchmark.

We further compare the AMN family of models on ruSciBench in Table 8. We report only Russian classification tasks for ruSciBERT and only English classification tasks due to their inability to encode in multilingual mode. It can be clearly seen that AMN-small (61M) trained on title-abstract and cite-cocite pairs outperforms in classification tasks all the listed competitors, including multilingual E5 embeddings [25], LaBSE [4], SciNCL [16] and two encoders designed specifically for Russian scientific articles: ruSciBERT [7, 8] and SciRus-tiny [6]. Importantly, AMN models have much fewer parameters compared to the other models (23M vs. 110M/280M/etc.) while achieving better quality on classification tasks.

| Task category | Task | Metric | SciNCL [16] | SPECTER [3] | all-mpnet-base-v2 [20] | multiling-E5-large [25] | AMN-tiny-t-a | AMN-tiny-t-a-cite-cocite | AMN-small-t-a-cite-cocite |
|---|---|---|---|---|---|---|---|---|---|
| Model size | | | 110M | 110M | 110M | 560M | 23M | 23M | 61M |
| Embedding size | | | 768 | 768 | 768 | 1024 | 312 | 312 | 768 |
| Max context length | | | 512 | 512 | 512 | 512 | 1024 | 1024 | 1024 |
| OoT | Biomimicry [CLF] | Wt. F1 | 50.22 | **51.22** | 50.93 | 33.83 | 33.82 | 36.76 | 36.62 |
| | DRSM [CLF] | Wt. F1 | 65.10 | **66.16** | 63.45 | 54.21 | 56.46 | 54.35 | 54.5 |
| | Relish | nDCG | 90.67 | 90.07 | **91.79** | 90.96 | 89.57 | 91.01 | 90.78 |
| | NFCorpus | nDCG | 70.85 | 64.9 | **76.65** | 75.89 | 67.26 | 71.4 | 72.12 |
| | TREC CoVID [QRY/SAL] | nDCG | 87.67 | 86.53 | **91.57** | 90.63 | 88.12 | 89.22 | 89.98 |
| | Peer Rev. Matching [SAL] | Avg P@k | 45.40 | 45.19 | 45.73 | 45.11 | 44.55 | **45.89** | 45.63 |
| | Review Score [RGN] | K Tau | 18.87 | 17.35 | 8.66 | 16.8 | 16.55 | **19.19** | 18.61 |
| | Max h-Index [RGN] | K Tau | 11.3 | 10.04 | 8.71 | 11.57 | 13.65 | 14.71 | **16.36** |
| | Tweet Mentions [RGN] | K Tau | 25.78 | 24.19 | 14.34 | 24.86 | 21.71 | **28.2** | 27.39 |
| Out-of-Train Avg | | | **51.8** | 50.6 | 50.2 | 49.32 | 47.97 | 50.08 | 50.22 |
| InT | MeSH [CLF] | F1 | 86.17 | 85.46 | 86.62 | **87.1** | 83.03 | 85.04 | 86.13 |
| | FoS Gold [CLF] | Wt. F1 | 35.19 | **35.32** | 34.58 | 29.17 | 31.9 | 34.48 | 34.8 |
| | Same Author Pred. [SAL] | MAP | 87.47 | 86.53 | 87.12 | 84.26 | 84.17 | 87.76 | **87.84** |
| | Search [QRY/SAL] | nDCG | 73.54 | 73.31 | 74.31 | **74.99** | 73.38 | 72.96 | 73.32 |
| | Citation Context [SAL] | MAP | 43.39 | 42.89 | 45.43 | **45.74** | 42.44 | 43.78 | 43.68 |
| | Citation Count [RGN] | K Tau | 34.61 | 33.21 | 22.74 | 35.3 | 33.63 | 35.87 | **38.06** |
| | Publishing Year [RGN] | K Tau | 29.0 | 25.96 | 21.02 | 33.0 | 25.21 | 40.51 | **40.74** |
| In-Train Avg | | | 55.6 | 54.7 | 53.12 | 55.65 | 53.39 | 57.2 | **57.8** |
| SD | MAG [CLF] | F1 | 81.11 | 79.4 | 82.61 | **83.4** | 81.43 | 82.46 | 83.33 |
| | MeSH [CLF] | F1 | 89 | 87.7 | 89.52 | 89.97 | 85.73 | 89.90 | **90.32** |
| | Co-View [SAL] | MAP | 85.28 | 83.4 | **85.67** | 81.76 | 81.53 | 84.43 | 84.64 |
| | Co-View [SAL] | nDCG | 92.23 | 91.4 | **92.44** | 90.5 | 90.54 | 91.88 | 91.99 |
| | Co-Read [SAL] | MAP | **87.69** | 85.1 | 87.18 | 82.48 | 81.91 | 86.02 | 86.27 |
| | Co-Read [SAL] | nDCG | **94.0** | 92.7 | 93.74 | 91.46 | 91.07 | 93.11 | 93.23 |
| | Cite [SAL] | MAP | **93.55** | 92 | 92.97 | 83.35 | 83.82 | 89.33 | 89.82 |
| | Cite [SAL] | nDCG | **97.35** | 96.6 | 97.04 | 92.74 | 92.87 | 95.52 | 95.72 |
| | Co-cite [SAL] | MAP | 91.66 | 88.0 | **92.37** | 85.74 | 84.02 | 89.83 | 89.94 |
| | Co-cite [SAL] | nDCG | 96.44 | 94.7 | **96.79** | 93.86 | 92.94 | 95.66 | 95.70 |
| SciDocs Avg | | | 90.84 | 89.10 | **91.03** | 87.53 | 86.59 | 89.81 | 90.10 |
| Avg | | | **66.08** | 64.8 | 64.78 | 64.17 | 62.65 | 65.7 | 66.04 |

**Table 7:** Results of evaluation on SciRepEval benchmark.

| Model/Metric | Size | Emb Size | Max Context Length | OECD-full | GRNTI-full | OECD-ru | GRNTI-ru | OECD-en | GRNTI-en | Translation search | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | F1 | | | | | | ru-en | en-ru |
| | | | | | | | | | | recall@1 | recall@1 |
| mE5-large [25] | 560M | 1024 | 512 | 58.19 | 70.44 | 58.23 | 69.67 | 56.77 | 68.93 | **99.19** | **99.37** |
| mE5-base [25] | 280M | 768 | 512 | 55.50 | 68.67 | 56.18 | 68.17 | 55.72 | 68.11 | 96.87 | 98.19 |
| LaBSE [4] | 471M | 768 | 512 | 54.26 | 66.77 | 54.60 | 66.16 | 53.67 | 65.81 | 98.31 | 97.20 |
| SciRus-tiny [6] | 23M | 2048 | 312 | 48.08 | 61.25 | 48.37 | 62.43 | 48.20 | 61.39 | 88.2 | 88.1 |
| ruSciBERT [7] | 125M | 768 | 512 | - | - | 53.80 | 66.69 | - | - | - | - |
| SciNCL [16] | 110M | 768 | 512 | - | - | - | - | 56.2 | 69.1 | - | - |
| AMN-tiny t-a | 23M | 312 | 1024 | 54.68 | 67.36 | 54.98 | 66.89 | 54.18 | 67.07 | 94.83 | 95.81 |
| AMN-tiny t-a-(co)cite | 23M | 312 | 1024 | 58.98 | 70.64 | 60.16 | **72.17** | 57.73 | 69.82 | 80.70 | 90.18 |
| AMN-small t-a-(co)cite | 61M | 768 | 1024 | **60.30** | **71.66** | **61.10** | 71.57 | **59.75** | **70.79** | 82.22 | 91.36 |

**Table 8:** Results of evaluation on ruSciBench benchmark.

| Model | MAG | MeSH | Cite | | CoView | | CoCite | | CoRead | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | map | nDCG | map | nDCG | map | nDCG | map | nDCG | |
| **S2AG + eLibrary** | | | | | | | | | | | |
| tiny-t-a | 82.01 | 85.98 | 84.17 | 93.03 | 81.37 | 90.46 | 83.31 | 92.59 | 81.45 | 90.88 | 86.53 |
| tiny-t-a-cite | 82.01 | 89.48 | 90.30 | 95.83 | 84.19 | 91.76 | 88.85 | 95.18 | 85.21 | 92.64 | 89.55 |
| tiny-cite | 80.82 | 89.83 | 88.88 | 95.17 | 84.05 | 91.77 | 89.14 | 95.24 | 85.02 | 92.59 | 89.25 |
| tiny-cocite | 81.54 | 89.87 | 87.46 | 94.54 | 84.08 | 91.84 | 89.48 | 95.39 | 85.64 | 92.92 | 89.28 |
| tiny-cite-cocite | 82.29 | 90.12 | 89.19 | 95.34 | 84.53 | 91.99 | 89.73 | 95.54 | 86.22 | 93.25 | 89.82 |
| tiny-all | 83.33 | 90.32 | 84.64 | 91.99 | 86.27 | 93.23 | 89.82 | 95.72 | 89.94 | 95.70 | 90.10 |
| **S2AG** | | | | | | | | | | | |
| tiny-all-s2-only | 82.18 | 89.27 | 88.22 | 94.89 | 84.37 | 91.90 | 89.37 | 95.39 | 85.92 | 93.07 | 89.46 |

**Table 9:** Ablations on SciDocs

| Model | OECD-full | GRNTI-full | OECD-ru | GRNTI-ru | OECD-en | GRNTI-en | Translation search | |
|---|---|---|---|---|---|---|---|---|
| | Macro F1 | | | | | | ru-en recall@1 | en-ru recall@1 |
| tiny-t-a | 54.06 | 66.52 | 54.70 | 66.36 | 53,72 | 66.57 | 94.13 | 94.73 |
| tiny-t-a-cite | 57.79 | 70.08 | 59.20 | 70.45 | 56.64 | 69.04 | 90.98 | 94.40 |
| tiny-cite | 58.04 | 69.53 | 59.96 | 70.66 | 56.65 | 68.51 | 61.41 | 66.87 |
| tiny-cocite | 57.99 | 69.17 | 58.97 | 69.80 | 56.70 | 68.52 | 50.61 | 54,68 |
| tiny-cite-cocite | 58.78 | 70.57 | 59.96 | 71.15 | 57.95 | 69.94 | 80.58 | 89.94 |
| tiny-all | 60.30 | 71.66 | 61.10 | 71.57 | 59.75 | 70.79 | 82.22 | 91.36 |

**Table 10:** Ablations on RuSciBench

### *4.4 Ablation study*

We conducted ablation experiments in two directions: over datasets to check how small-sized models work in cross-lingual scenarios while having little multilingual data and over possible ways to do contrastive training. We also show results only for the tiny model for simplicity purposes.

First, we prepared a series of ablations to detect which pairs among title-abstract, cite, and cocite options contribute the most during contrastive training. It turned out that cite and cocite pairs are interchangeable: while using one of them, there is no need to add another. Moreover, training with only cite or cocite pairs leads to nearly the same results as training with cite (or cocite) pairs + title-abstract pairs. The only exception to this rule is cross-lingual scenarios, which need title-abstract pairs to get better results. For more details, see Table 9 for ablations on the SciDocs benchmark and Table 10 for RuSciBench ablations.

Second, we also checked how one language amplification for cross-lingual scenarios influences the performance of small-sized models on English benchmarks. To do so, we trained the contrastive stage on the S2AG dataset only. This led to nearly no decrease in performance. What is even more interesting is that the metrics on RuSciBench have increased, no matter that S2AG mainly consists of English texts. This leads to the conclusion that AnonymizedModelName has a basic understanding of all the languages from the S2AG dataset and, after specific tuning, can even yield SOTA results.

## 5 Conclusion

We proposed the AMN family of light multilingual encoders that show SOTA performance on 7 tasks from SciRepEval and also outperform its competitors, including SPECTER and SciNCL on RuSciBench — a benchmark constructed of translation search and document classification tasks on scientific papers from the Russian electronic library. While being relatively small (23M and 61M parameters for tiny and small models, respectively), these models can be successfully fine-tuned for different downstream tasks. We experiment with different ways of model training, combining MLM pretraining with contrastive training on title-abstract and citation pairs. Our research shows that contrastive learning with title-abstract pairs may provide nearly the same quality as citation and co-citation pairs, which are usually harder to collect. Our model is capable of working in cross-lingual scenarios without the direct usage of machine translators.

## 6 Future Work

There are several important directions not covered in this paper and left for future research:

- Training models with more parameters for more advanced tasks.
- Enforcing a high level of cross-lingual relations between many languages by inducing more specific datasets.
- Curriculum learning with different strategies: going from simple to more advanced papers, from less cited to more cited, etc.
- Exploration of how training influences Transformer inner states to get common patterns and choose the most optimal learning strategy based on them (like it is done in [17]).
- Usage of different adapters (like it is done in SPECTER2 [19]).
- Posttrain with Matryoshka [9] technique to make final models more effective and thus more available to the scientific community.

## References


[1] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[2] C. Bhagavatula, S. Feldman, R. Power, and W. Ammar. Content-based citation recommendation. *arXiv preprint arXiv:1802.08301*, 2018.

[3] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020.

[4] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

[5] N. Gerasimenko and A. Vatolin. Ruscibench benchmark. 2023. URL https://github.com/mlsa-iai-msu-lab/ru_sci_bench/tree/main.

[6] N. Gerasimenko and A. Vatolin. Scirus-tiny. 2023. URL https://huggingface.co/mlsa-iai-msu-lab/sci-rus-tiny.

[7] N. A. Gerasimenko, A. S. Chernyavsky, and M. Nikiforova. ruscibert: a transformer language model for obtaining semantic embeddings of scientific texts in russian. In *Doklady Mathematics*, volume 106, pages S95–S96. Springer, 2022.

[8] N. A. Gerasimenko, A. S. Chernyavsky, and M. Nikiforova. Ruscibert. 2023. URL https://huggingface.co/ai-forever/ruSciBERT.

[9] A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[11] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. S. Weld. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*, 2019.

[12] K. Luu, X. Wu, R. Koncel-Kedziorski, K. Lo, I. Cachola, and N. A. Smith. Explaining relationships between scientific documents. *arXiv preprint arXiv:2002.00317*, 2020.

[13] M. M'hamdi, J. May, F. Dernoncourt, T. Bui, and S. Yoon. Multilingual sentence-level semantic search using meta-distillation learning. *arXiv preprint arXiv:2309.08185*, 2023.

[14] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[15] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[16] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, and G. Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. *arXiv preprint arXiv:2202.06671*, 2022.

[17] A. Razzhigaev, M. Mikhalchuk, E. Goncharova, I. Oseledets, D. Dimitrov, and A. Kuznetsov. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. *arXiv preprint arXiv:2311.05928*, 2023.

[18] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[19] A. Singh, M. D'Arcy, A. Cohan, D. Downey, and S. Feldman. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*, 2022.

[20] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

[21] Q. Tang, J. Chen, B. Yu, Y. Lu, C. Fu, H. Yu, H. Lin, F. Huang, B. He, X. Han, et al. Self-retrieval: Building an information retrieval system with one large language model. *arXiv preprint arXiv:2403.00801*, 2024.

[22] J. Tiedemann. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[24] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pretraining. *arXiv preprint arXiv:2212.03533*, 2022.

[25] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

[26] D. Wright and I. Augenstein. Citeworth: Cite-worthiness detection for improved scientific document understanding. *arXiv preprint arXiv:2105.10912*, 2021.