Transferring Scientific English Pre-Trained Language Models to Multiple Languages Using Cross-Lingual Transfer

Nikolas Rauscher[1,2], Fabio Barth[2], Georg Rehm[2,3]
[1]Technische Universität Berlin, Germany
[2]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany
[3]Humboldt-Universität zu Berlin, Germany
nikolas.rauscher@dfki.de, fabio.barth@dfki.de, georg.rehm@dfki.de

Abstract

In this paper, we present a pipeline for domain-adaptive pre-training and cross-lingual transfer of scientific language models from English to non-English languages. Starting from the multilingual scientific corpus SciLaD, we construct a cleaned English pre-training split and continually pre-train a T5-base encoder–decoder model, resulting in EN-T5-Sci. Our model achieves consistent zero-shot improvements on the Global-MMLU English benchmark, outperforming its base model, with particularly strong gains in STEM and Social Sciences. Despite its moderate size, it performs comparably to the much larger BLOOM model on scientific categories. Building on EN-T5-Sci, we transfer scientific knowledge to German, Japanese, Russian, Polish, Spanish, and Portuguese using the WECHSEL method. Our approach reinitializes language-specific embedding layers via aligned static embeddings while retaining the pre-trained Transformer weights, yielding six monolingual scientific T5 models. In zero-shot evaluation in each respective language, the transferred models generally outperform monolingual baselines. These results demonstrate that scientific domain knowledge acquired through English pre-training can be effectively transferred across languages, enabling competitive non-English scientific language models without training large multilingual systems from scratch.

Keywords: T5, scientific language models, pre-training, cross-lingual transfer, multilingual scientific NLP

1. Introduction

In recent years, large language models have shown remarkable improvements in scientific English language processing (Hu et al., 2025). Many of these language models are pre-trained on scientific text, which provides a complementary data source to generic web pages, as it is factually dense and stylistically consistent. They follow a standardized structure and undergo peer review, which ensures greater validity. Prior work has shown that adapting models to scientific literature improves performance on domain-specific tasks, even when using significantly less data than broad web scrapes (Phan et al., 2021; Taylor et al., 2022). However, the availability of high-quality scientific text is unevenly distributed, with English being the predominant language (Tardy, 2004). Even high-resource languages, such as German or Japanese, can be considered low-resource languages in the scientific domain compared to English. As a result, scientific language models have been English-centric, and multilingual language models are trained on noisier non-scientific data or machine-translated content (Phan et al., 2021; Taylor et al., 2022; Ali et al., 2025).

An approach for generating high-quality language models for low-resource languages is domain-adaptive pre-training (DAPT) with cross-lingual transfer (Minixhofer et al., 2022). Here, a language model is trained on a high-resource language using a large corpus and then efficiently and effectively transferred into a low-resource language using cross-lingual parameter transfer. Although this process has been a well-established method for generating low-resource language models in the general domain, this method has not been applied to the scientific domain.

We propose a pipeline for DAPT with cross-lingual transfer from a pre-trained English scientific language model to generate non-English scientific language models. We therefore continually pre-train an English T5-base model on a newly composed scientific corpus (Foppiano et al., 2025) and then transfer the acquired knowledge to six non-English languages by applying the well-established WECHSEL method to the scientific domain (Minixhofer et al., 2022). We present monolingual scientific language models in German, Japanese, Russian, Polish, Spanish, and Portuguese. Our transferred scientific language models outperform their baselines and achieve results comparable to those of a much larger multilingual LLM trained on general-domain and scientific data. Our contributions can be summarized as follows:

1. We create a pipeline for scientific knowledge transfer from English to non-English languages.
2. We train a scientific English T5-base model (EN-T5-Sci), which achieves consistent

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 261–269
12 May 2026.

zero-shot gains on Global-MMLU benchmarks, especially in science-related categories.

3. We present six monolingual scientific T5 models in German, Japanese, Russian, Polish, Spanish, and Portuguese by transferring EN-T5-Sci using WECHSEL. The transferred models outperform almost all general-domain monolingual models in their language and achieve strong performance on scientific question answering.

The rest of the paper is structured as follows: We first describe the pre-training dataset used for the continual pre-training of the English T5-base model and the language-specific data used for tokenizer training and cross-lingual transfer. We then describe the models and training setup, followed by the experimental evaluation of all models. Finally, we discuss related work and conclude the paper.

2. Dataset Construction & Pre-processing

The process of transferring a model from one language to another requires two data splits. A very large split in the initial language (English) for pre-training the language model, and a much smaller dataset in the target language to train the non-English tokenizer. As English has been the established lingua franca in the scientific domain, most scientific pre-training datasets already have a linguistic bias towards English. Many scientific datasets are published with English data only, or the text is not classified by language. For our approach, we use the recently published SciLaD dataset (Foppiano et al., 2025), which contains open-access articles from Unpaywall processed into plain text via GROBID. The corpus contains over 200 classified languages. The English split consists of approximately 11 million documents, accounting for 73.95% of the dataset. In contrast, the German split is only 1.99%, highlighting the language bias in scientific texts. The largest language split after the English split in the SciLaD dataset is Japanese, with 4.1%, followed by Portuguese, with 3.5%. Less than 1% of the data contains Polish-labeled text.

We build a cleaning pipeline using Hugging Face's DataTrove[1] library for the pre-training of the scientific English T5 model. Our cleaning pipeline removes non-semantic noise embedded during the PDF-to-text conversion. We remove approximately 184 million citation artifacts using regular expression (regex) filters. These include numeric brackets (e.g., "[1–3]"), parenthetical author-year patterns, and semicolon-separated author lists. We also delete artifact identifiers such as DOIs, ISBNs, and arXiv IDs. URLs and email addresses are replaced with placeholders.

Metric	Before	After	Change
Documents	10,030,761	10,027,111	-0.04%
Mean length (chars)	28,627	22,319	-22.0%
Mean perplexity	569.54	538.85	-5.4%
EN confidence	93.36%	93.75%	+0.4%
Paragraph duplicates	1.30%	0.75%	-42.3%
Digit ratio	1.56%	1.20%	-23.4%

Table 1: Impact of the DataTrove cleaning pipeline on the English SciLaD pre-training split, comparing corpus statistics before and after cleaning.

Through heuristic filtering, we also remove redundant text and tabular fragments. First, we delete structural noise, such as figure/table captions and repeated header metadata. We then remove tabular fragments if their numeric share exceeds 60%. Finally, we normalize the text by collapsing whitespace, standardizing ellipses to ASCII, and converting bullets to Markdown format. After processing, the average document length decreased by 22.0% while preserving 99.9% of the documents. The Wikipedia-LM perplexity improved by 5.4% and paragraph duplicates fell by 42.3% (see Table 1).

3. Model Training

In this section, we first describe the pre-training of the scientific English T5 model using the SciLaD dataset (Foppiano et al., 2025) and then outline our pipeline for scientific knowledge transfer from English to non-English languages.

3.1. Scientific English T5 Model

We perform continued pre-training on the clean English SciLaD data using the T5-base[2] Encoder–Decoder architecture (220M parameters), initialized from the original Google checkpoint. Following the approach of Raffel et al. (2023), we use span masked model training with a 15% masking rate and a mean noise span length of 3 tokens. We choose the T5 model over more recent generative decoder-only models to keep the computational cost for all monolingual language models reasonable. Note that, in this paper, we showcase the performance of knowledge transfer into non-English languages rather than building comparable language models for text generation, which is why

[1] https://github.com/huggingface/datatrove *accessed 2026-02-19*

[2] https://huggingface.co/google-t5/t5-base *accessed 2026-02-19*

we choose a smaller T5 model over an LLM with billions of parameters.

Sequence Preparation For the pre-training, we segment each document into fixed-length sequences of 512 tokens using a sliding-window strategy with 50% overlap to process long-form articles. This yields approximately 261M training datapoints, resulting in a span-corruption setup with an average decoder target length of about 114 tokens for 512-token inputs.

Optimization and Schedule We train on 4× NVIDIA H100 (80GB) for 487,500 steps (within the first epoch) with an effective batch size of 384 (per-GPU batch size 48 and gradient accumulation of 2). The learning-rate schedule uses a linear warmup for 20,000 steps to a peak of 10^{-3}, followed by an inverse-square-root decay. We apply gradient clipping of 0.5 and use the Adafactor (Shazeer and Stern, 2018) optimizer as the optimization method.

Monitoring and Checkpoint Selection We compute the validation loss and perplexity every 5,000 steps on a held-out validation split of 100,000 datapoints. We select the checkpoints with the minimal validation loss, referred to as EN-T5-Sci.

3.2. Cross-Lingual Transfer using WECHSEL

To transfer the EN-T5-Sci model to any non-English split, we build a pipeline that follows the WECHSEL approach (Minixhofer et al., 2022). The basic assumption in this approach is that Transformer layers primarily encode abstract, language-agnostic knowledge, while language-specific information is concentrated in the token embeddings. Accordingly, our pipeline transfers all pre-trained non-embedding weights from the EN-T5-Sci checkpoints to a non-English target model and reinitializes the embedding matrix for a non-English tokenizer. After the embedding matrix is re-initialized, the non-English tokenizer is trained on a language-specific scientific vocabulary. We run the pipeline for German, Japanese, Russian, Polish, Spanish, and Portuguese, resulting in six monolingual non-English scientific language models.

Tokenizer Training For each language, we first sample a 5 GB subset from the corresponding SciLaD language split and divide it into train, validation, and test splits. We then train a SentencePiece tokenizer on the train split to obtain a vocabulary that reflects the language- and domain-specific subword statistics. This is important because we replace the source embedding layer, and reusing an English tokenizer would likely yield suboptimal segmentation for non-English scientific text and a target embedding vocabulary that does not match the target distribution well.

During tokenizer training, we normalize line breaks to spaces and cap each document at 20k characters to limit outlier effects. We use SentencePiece BPE with a vocabulary size of 32k subwords, character coverage 1.0, and byte-fallback enabled. To ensure T5 compatibility, we fix special token IDs (`pad=0, eos=1, unk=2; bos=disabled`) and export the tokenizer with `extra_ids=100`, resulting in 32,100 tokens in total. The resulting tokenizer defines the target subword vocabulary U^t used in WECHSEL.

Embedding Initialization and Alignment To initialize the embedding matrix, we align English and non-English tokens in a shared static semantic space. Following the work of Minixhofer et al. (2022), we use multilingual fastText vectors and a bilingual dictionary to obtain aligned static subword embeddings for source and target vocabularies. Let U^s and U^t denote the source and target subword vocabularies, and let E^s and E^t denote the corresponding source and target model embedding matrices. For each target subword $x \in U^t$, we retrieve its k nearest source neighbors $y \in U^s$ by cosine similarity in the shared space,

$$s_{x,y} = \frac{u_x^t u_y^{s\top}}{\|u_x^t\| \, \|u_y^s\|},$$

and initialize the target embedding as a convex combination of the corresponding source embeddings, with mixture weights given by a softmax over similarities with temperature τ (k=10, τ=0.1). Concretely, letting $\mathcal{J}_x$ be the set of the k nearest source subwords for x,

$$e_x^t = \frac{\sum_{y \in \mathcal{J}_x} \exp(s_{x,y}/\tau) \cdot e_y^s}{\sum_{y' \in \mathcal{J}_x} \exp(s_{x,y'}/\tau)}.$$

4. Experimental Setup

As the number of publicly available benchmarks for the scientific domain is limited, we evaluate the models on a subset of Global-MMLU (Singh et al., 2025). Global-MMLU is a translation of MMLU (Hendrycks et al., 2021), a multiple-choice QA dataset spanning diverse subject areas, including elementary mathematics, computer science, and law. We evaluate each language model in its respective language. For comparability, we evaluate all models in a zero-shot setting using the EleutherAI `lm-evaluation-harness`[3], en-

[3] https://github.com/EleutherAI/lm-evaluation-harness *accessed 2026-02-19*

Language	Abbrev.	Base model checkpoint (Hugging Face)
German	DE-Base	GermanT5/t5-efficient-gc4-german-base-nl36
Japanese	JA-Base	megagonlabs/t5-base-japanese-web
Spanish	ES-Base	vgaraujov/t5-base-spanish
Polish	PL-Base	allegro/plt5-base
Portuguese	PT-Base	unicamp-dl/ptt5-base-portuguese-vocab
Russian	RU-Base	ai-forever/ruT5-base

Table 2: Non-English monolingual base checkpoints and abbreviations used in the experiments.

suring the prompt template and scoring configuration are fixed across all models. As a metric, we use accuracy using multiple-choice log-likelihood scoring. In multiple-choice log-likelihood scoring, the option with the highest log-probability is selected as the prediction. We use fixed seeds for all evaluations and report overall scores, as well as subject-level changes in Global-MMLU, to characterize domain gains and potential trade-offs.

As a baseline for the scientifically continued-pretrained English T5 model (EN-T5-Sci), we evaluate the original Google T5-base checkpoint (EN-T5-Base) and the SciFive model, a T5 model trained on biomedical literature, on the same setup. We additionally include BLOOM-176B, a multilingual decoder-only LLM trained on general and scientific data. Because BLOOM is architecturally different (decoder-only vs. encoder-decoder) and substantially larger, it is not a controlled baseline comparison. Therefore, it serves as a multilingual reference point. For the non-English experiments, we use one monolingual T5 checkpoint per language as the corresponding base model (Table 2). These checkpoints define the DE-Base, JA-Base, ES-Base, PL-Base, PT-Base, and RU-Base in the experiments.

For each target language, we also report BLOOM results on Global-MMLU as an additional multilingual reference. To establish domain-adapted monolingual baselines, we continuously pre-train the non-English base models for 15,000 steps on the respective 5GB SciLaD subsplit (DE-Base-CP, JA-Base-CP, RU-Base-CP, PL-Base-CP, ES-Base-CP, and PT-Base-CP). For this adaptation phase, we use the same sequence preparation as in English continued pre-training, i.e., 512-token sequences with a sliding-window strategy and 50% overlap. We also use the same span-corruption objective and peak learning rate of 10^{-3}. We reduce warmup to 1,500 steps and apply gradient clipping with a max norm of 1.0. We compare these controls against our WECHSEL-initialized models (DE-Trans-Init, JA-Trans-Init, RU-Trans-Init, PL-Trans-Init, ES-Trans-Init, and PT-Trans-Init). This setup enables a controlled comparison between standard monolingual adaptation and our initialization-based transfer approach.

5. Evaluation Results

5.1. English Language Model

The EN-T5-Sci model achieves a higher score on the English MMLU split compared to its base model, with an average accuracy score increase of 4 pp (see Table 3). The largest category-level gain is in Social Sciences (+9.4 pp), followed by STEM (+7.2 pp). Compared to the SciFive model, our model also performs on average 4 pp better. However, compared to the BLOOM model, the EN-T5-Sci only outperforms it in the Social Sciences category with a 0.9 pp higher score, while having an average 1.7 pp lower score. Note that the BLOOM model has 176 billion parameters, 800 times as many as the T5 model. When comparing only the scientific categories (STEM, Humanities, and Social Sciences), the EN-T5-Sci is only 0.6 pp worse than the BLOOM model.

Model	Avg	STEM	Hum	SocSci	Other
English Monolingual Models					
SciFive-Base	22.9	21.3	24.2	21.7	24.0
EN-T5-Base	22.9	21.3	24.1	21.7	23.9
EN-T5-Sci	26.9	28.5	24.2	**31.1**	25.1
Multilingual Model					
BLOOM-176B	**28.6**	**29.8**	**25.6**	30.2	**30.1**
Δ (EN-T5-Sci - EN-T5-Base)	+4.0	+7.2	+0.1	+9.4	+1.2

Table 3: Zero-shot accuracy (%) on the English **Global-MMLU-EN**, for SciFive-Base, EN-T5-Base, EN-T5-Sci, and BLOOM-176B, reported as overall and category-level results.

5.2. Non-English Language Models

The best-performing non-English transferred models are DE-Trans-Init and ES-Trans-Init, both with 26.89% average accuracy. Both are numerically very close to the English source model EN-T5-Sci (26.87%). For German, STEM and Social Sciences show the largest gains, and the control model (DE-Base-CP) shows no improvement over DE-Base. This indicates that the German gains mainly come from the cross-lingual transfer, not from short continued pre-training alone. DE-Trans-Init is only 0.34 pp worse than BLOOM on Global-MMLU-DE, while scoring 0.79 pp higher on STEM.

Model	Avg	STEM	Hum	SocSci	Other
English Source Model on Global-MMLU-EN					
EN-T5-Sci	26.87	28.51	24.19	**31.07**	25.10
BLOOM-176B	**28.60**	**29.80**	**25.60**	30.20	**30.10**
German models on Global-MMLU-DE					
DE-Base	22.95	21.25	24.21	21.71	23.98
DE-Base-CP	22.95	21.25	24.21	21.71	23.98
DE-Trans-Init	26.89	**28.64**	24.14	**31.07**	25.14
BLOOM-176B	**27.23**	27.85	**25.27**	30.19	**26.65**
Japanese models on Global-MMLU-JA					
JA-Base	25.23	24.58	23.89	**28.57**	24.62
JA-Base-CP	22.95	21.25	24.23	21.71	23.98
JA-Trans-Init	25.51	**26.26**	**27.12**	23.79	24.01
BLOOM-176B	**26.04**	25.44	26.67	25.64	**26.07**
Russian models on Global-MMLU-RU					
RU-Base	26.01	28.35	24.55	28.18	23.72
RU-Base-CP	24.17	23.25	**25.36**	23.24	24.24
RU-Trans-Init	26.36	27.12	24.89	28.86	25.33
BLOOM-176B	**28.04**	**30.26**	25.31	**30.61**	**27.36**
Polish models on Global-MMLU-PL					
PL-Base	25.51	26.26	**27.12**	23.79	24.01
PL-Base-CP	24.65	23.88	24.51	23.43	**26.87**
PL-Trans-Init	24.66	23.91	24.51	23.43	**26.87**
BLOOM-176B	**27.17**	**27.40**	26.55	**29.54**	25.52
Spanish models on Global-MMLU-ES					
ES-Base	25.51	26.26	**27.12**	23.79	24.01
ES-Base-CP	25.51	26.26	**27.12**	23.79	24.01
ES-Trans-Init	26.89	28.61	24.17	31.07	25.14
BLOOM-176B	**29.23**	**30.29**	26.57	**31.59**	**29.84**
Portuguese models on Global-MMLU-PT					
PT-Base	23.55	22.58	23.91	23.14	24.40
PT-Base-CP	23.02	21.54	24.08	21.84	24.07
PT-Trans-Init	24.98	24.64	24.44	24.34	26.75
BLOOM-176B	**29.17**	**29.50**	**26.40**	**32.27**	**29.96**

Table 4: Zero-shot accuracy (%) on Global-MMLU benchmarks, grouped by evaluation language, for baseline, continued-pretrained, transfer-initialized, and BLOOM-176B models across all evaluated languages.

Except for the transferred Polish model, all other transferred non-English models improve over their corresponding baselines. For JA-Trans-Init, RU-Trans-Init, and PT-Trans-Init, the mean average gain is 0.69 pp. In contrast, PL-Trans-Init is 0.85 pp below PL-Base on average (24.66% vs. 25.51%), although it attains the strongest Polish *Other* score (26.87%, tied with PL-Base-CP and above BLOOM's 25.52%). JA-Trans-Init outperforms BLOOM in STEM and Humanities. PT-Trans-Init is the only transferred model that outperforms its base model across all categories (+1.43 pp on average), but it remains 4.19 pp below BLOOM in overall average.

Overall, the results show that the transferred language-specific models improve over the respective baselines across most languages. We even observe category-level gains over the much larger BLOOM model, for example, in German STEM and Social Sciences, as well as Japanese STEM and Humanities, although BLOOM remains stronger on most overall averages.

5.3. Error Analysis

For the error analysis, we discuss in more detail the results of the English model and the German model as the best-performing non-English mod-

els. For both models, we analyze the best- and worst-performing categories in the MMLU evaluation in more detail. We highlight the largest accuracy gains and the categories in which the models fell short after pre-training or transfer, respectively.

English Model Table 5 lists the top- and bottom-performing subtasks for the English models. The largest performance gains of the English model are in subjects such as *high_school_statistics*, *professional_medicine*, and *college_chemistry*, underlining improved performance on science-related tasks. We see gains of over 30 pp in *high_school_statistics*, for instance, which shows the impact of specialization.

Subtask	EN-Base	EN-T5-Sci	SciFive	BLOOM
Top 4 (EN-T5-Sci)				
high_school_statistics	15.28	**47.22**	15.28	15.74
professional_medicine	18.38	**44.85**	18.38	18.75
college_chemistry	20.00	**41.00**	20.00	20.00
security_studies	18.78	**40.00**	18.78	18.78
Bottom 4 (EN-T5-Sci)				
human_aging	30.94	10.76	**31.39**	**31.39**
machine_learning	**31.25**	16.07	**31.25**	**31.25**
international_law	**23.97**	14.05	**23.97**	**23.97**
world_religions	**32.16**	17.54	**32.16**	31.58

Table 5: Zero-shot accuracy (%) on the top and bottom 4 **Global-MMLU-EN** subtasks, sorted by EN-T5-Sci performance. Scientific pre-training yields strong gains on STEM-oriented subtasks but regressions on unrelated subjects.

An interesting observation is the worst-performing subjects of the scientific T5 model. Here, the model performs worse on non-scientific tasks such as *human_aging*, *international_law*, and *world_religions*, with average scores 14.9 pp lower on those subjects. Moreover, the model performs worse on a science-related subject like *machine_learning*, dropping by 15.2 pp. One possible reason for this drop could be the structure of the *machine_learning* QA-pairs: most of them contain short answers with true/false options, which likely prompt the model to generate a broad probability distribution over the choices, since they seem more similar to the model.

Compared to BLOOM and SciFive, EN-T5-Sci also outperforms those models in the scientific categories (see Table 5). BLOOM and SciFive achieve similar scores to the base model in categories such as *human_aging*, *international_law*, and *world_religions*. These results highlight the transfer of scientific knowledge from the pre-training dataset.

German Model Similar to the scientific English model, the transferred German model has its strongest category-level results in STEM and Social Sciences. At the subtask level, the highest DE-Trans-Init accuracies are in *high_school_statistics*, *professional_medicine*, *college_chemistry*, and *security_studies*.

The lowest DE-Trans-Init accuracies are in *human_aging*, *international_law*, *machine_learning*, and *world_religions*. This pattern indicates that gains are uneven across subjects and include trade-offs in some tasks despite overall improvements in selected categories. One possible explanation is a specialization trade-off or catastrophic forgetting, in which the model emphasizes certain scientific patterns at the expense of broader knowledge (Haque, 2025). Since WECHSEL largely preserves the trained Transformer weights and mainly replaces and aligns the embedding layer, the transferred model keeps part of the source behavior while still exhibiting language-specific weaknesses.

Subtask	DE-Base	DE-Base-CP	DE-Trans-Init	BLOOM
Top 4 (DE-Trans-Init)				
high_school_statistics	15.28	15.28	**47.22**	37.50
professional_medicine	18.38	18.38	**44.85**	43.75
college_chemistry	20.00	20.00	**41.00**	31.00
security_studies	18.78	18.78	**40.00**	36.73
Bottom 4 (DE-Trans-Init)				
human_aging	**31.39**	**31.39**	10.76	15.25
international_law	**23.97**	**23.97**	14.05	19.01
machine_learning	**31.25**	**31.25**	16.07	25.00
world_religions	**32.16**	**32.16**	17.54	23.98

Table 6: Zero-shot accuracy (%) on the top and bottom 4 **Global-MMLU-DE** subtasks, sorted by DE-Trans-Init performance. Continued pre-training alone shows little effect, while transfer improves top subtasks similar to the source model.

Notably, the short continued pre-training phase (*DE-Base-CP*) does not show measurable improvements over *DE-Base* on the reported subtasks. This suggests that 15k continuation steps are likely insufficient to change zero-shot behavior on Global-MMLU-DE. Similar to the English model, DE-Trans-Init shows large positive shifts on selected subtasks like *high_school_statistics* and on *professional_medicine*, but also large drops on others such as *human_aging* and *machine_learning*. Compared with DE-Trans-Init, BLOOM is lower on all selected top subtasks but consistently higher on all selected bottom subtasks.

6. Related Work

Scientific Datasets We chose the SciLaD dataset because it is the most recently published scientific dataset and contains multiple language splits (Foppiano et al., 2025). However, there are six corpora that have been used for scientific model pre-training in recent research that are similar to the SciLaD dataset. PubMed Abstract[4] and PubMed

[4] https://pubmed.ncbi.nlm.nih.gov *accessed 2026-02-19*

Central[5] (PMC) are two corpora that contain abstracts and full text of biomedical and scientific journal literature from the U.S. National Institutes of Health's National Library of Medicine. S2ORC (Lo et al., 2020) is, with over 81 million open-access papers, the largest corpus of scientific text. This dataset contains open-access papers from the Semantic Scholar literature and is, like SciLaD, processed using GROBID (Grobid). The UnarXive (Saier et al., 2023) corpus draws its texts from arXiv, covering multiple scientific fields. The corpus, with 1.9 million documents, is much smaller than the SciLaD dataset. The ACL Anthology Network is a much smaller scientific corpus containing 24.6 thousand papers on computational linguistics from the ACL Anthology. The only downstream pre-training corpora that cover multiple scientific tasks are the BigBio (Fries et al., 2022) datasets. This dataset contains normalized datasets covering various NLP tasks in the scientific domain. Note that none of these datasets focus on multilinguality.

Scientific LLMs Scientific language models cover domain-specific knowledge from scientific domains such as chemistry, physics, astronomy, material science, life science, and earth science. In recent years, there have been many efforts to publish language models in these fields with a major focus on the medical and biomedical domain (Workshop et al., 2023; Hu et al., 2025). The closest model to our pre-trained T5 models is the SciFive model, which is a T5 model trained on biomedical literature (Phan et al., 2021). This model is trained on two corpora: PubMed Abstract and PMC. However, the SciFive model is English-centric and performs best on classic NLP tasks such as named-entity recognition and relation extraction. Taylor et al. (2022) published the first generative LLM for the scientific domain named Galactica. As this model shows strong performance on mathematical MMLU and downstream tasks, such as PubMedQA (Jin et al., 2019), Galactica is also a monolingual English language model. The first multilingual language model, which is also trained on scientific data, is the BLOOM model (Workshop et al., 2023). This model has been trained on the vast BigBio dataset (Fries et al., 2022). Although it has strong capabilities in multilingual NLP tasks, BLOOM has not been applied to the scientific domain. In recent years, generative LLMs have shown strong capabilities for performing scientific tasks without being fine-tuned solely on scientific domain-specific data (Abdullah et al., 2025; Jiang et al., 2024; Hu et al., 2025).

Language Transfer Methods Language transfer from English-centric language models is an ongoing research topic that has been used mostly for two reasons in the past (Lee et al., 2025). First, to reduce computational costs for non-English language models (Bhukya et al., 2023) and second, to create language models for low-resource languages (Remy et al., 2024). Ostendorff and Rehm (2023) introduced CLP-Transfer, a cross-lingual, progressive transfer learning approach. In this method, cross-lingual transfer is combined with progressive transfer, meaning increasing the model size, which is beyond the scope of this project. Lee et al. (2025) present Cross-Lingual Optimization (CLO) for language transfer using translated data. This approach requires a translation model for performing cross-lingual transfer.

7. Conclusion

In this paper, we present six non-English monolingual scientific language models that have been transferred from a new, continued pre-trained scientific English T5-base model. The English scientific model has been transferred to German, Japanese, Russian, Polish, Spanish, and Portuguese using a pipeline that applies the WECHSEL (Minixhofer et al., 2022) method. The pipeline and the models have been published and can be used to create more non-English scientific language models[6]. We show that scientific knowledge can be preserved through transfer, demonstrating that even for high-quality scientific text, language transfer is beneficial. The non-English models were primarily evaluated in a zero-shot setting on the Global-MMLU benchmark, and the results show strong performance on scientific QA benchmarks, with most surpassing baselines and achieving results comparable to those of multilingual LLMs such as BLOOM. The models have also been published on Hugging Face and are linked in the GitHub repository.

Limitations

While our work provides valuable insights into scientific language transfer and the shortcomings of language bias in the scientific domain, it also has several limitations. First, there is a severe lack of multilingual or non-English scientific benchmarks. As English is the lingua franca in the scientific domain, there have been few efforts in the community to build benchmarks for classic NLP tasks. Most available benchmarks are automatically translated, limiting the interpretability of the results. A set of

[5] https://www.ncbi.nlm.nih.gov/pmc accessed 2026-02-19

[6] https://github.com/nikolas-rauscher/scientific-english-crosslingual-transfer

manually curated multilingual benchmarks would benefit the analysis of our models.

Second, we face computational limitations while performing our experiments. We are aware that larger models have been trained and released in the past, and that using them for evaluation would, based on the scaling law, increase performance. However, we want to emphasize the rising computational costs of training them and the economic and environmental impacts that come with them. The goal of this work is not to improve state-of-the-art scientific language models but to showcase the usability of language transfer in the scientific domain. However, based on the results of this work, we currently aim to release a multilingual generative language model for the scientific domain using language transfer.

Acknowledgements

This work was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[7] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.), funded by the Federal Republic of Germany and its states through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). Further support was provided from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101189745 (HIVEMIND).[8]

8. References

Abdulhady Abas Abdullah, Arkaitz Zubiaga, Seyedali Mirjalili, Amir H. Gandomi, Fatemeh Daneshfar, Mohammadsadra Amini, Alan Salam Mohammed, and Hadi Veisi. 2025. Evolution of meta's llama models and parameter-efficient fine-tuning of large language models: a survey.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, et al. 2025. Teuken-7b-base & teuken-7b-instruct: Towards european llms.

Ramesh K. Bhukya, Anjali Chaturvedi, Hardik Bajaj, Udgam Shah, Sumit Singh, and Uma Shanker Tiwary. 2023. Efficiently transferring pre-trained language model roberta base english to hindi using wechsel. In *2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Luca Foppiano, Sotaro Takeshita, Pedro Ortiz Suarez, Ekaterina Borisova, Raia Abu Ahmad, Malte Ostendorff, Fabio Barth, Julian Moreno-Schneider, and Georg Rehm. 2025. Scilad: A large-scale, transparent, reproducible dataset for natural scientific language processing.

Jason Alan Fries, Leon Weber, Natasha Seelam, et al. 2022. Bigbio: A framework for data-centric biomedical natural language processing.

Grobid. 2008–2026. Grobid. https://github.com/kermitt2/grobid.

Naimul Haque. 2025. Catastrophic forgetting in llms: A comparative analysis across language tasks.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Ming Hu, Chenglong Ma, Wei Li, et al. 2025. A survey of scientific large language models: From data foundations to agent frontiers.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. 2024. Mixtral of experts.

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

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. 2025. Cross-lingual optimization for language transfer in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15100–15119, Vienna, Austria. 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, Online. Association for Computational Linguistics.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

[7] https://www.nfdi4datascience.de
[8] https://hivemind-project.eu

Malte Ostendorff and Georg Rehm. 2023. Efficient language model training through cross-lingual and progressive transfer learning.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of llms for low-resource nlp.

Tarek Saier, Johan Krause, and Michael Färber. 2023. unarxive 2022: All arxiv publications preprocessed for nlp, including structured full-text and citation network. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, page 66–70. IEEE.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, et al. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

C Tardy. 2004. The role of english in scientific communication: lingua franca or tyrannosaurus rex? *Journal of English for Academic Purposes*, 3(3):247–269.

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

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.