# PhysBERT: A Text Embedding Model for Physics Scientific Literature

**Thorsten Hellert**
Lawrence Berkeley National Laboratory
Berkeley, California, USA
thellert@lbl.gov

**João Montenegro**
Lawrence Berkeley National Laboratory
Berkeley, California, USA

**Andrea Pollastro**
University of Naples Federico II
Naples, Italy

## Abstract

The specialized language and complex concepts in physics pose significant challenges for information extraction through Natural Language Processing (NLP). Central to effective NLP applications is the text embedding model, which converts text into dense vector representations for efficient information retrieval and semantic analysis. In this work, we introduce PhysBERT, the first physics-specific text embedding model. Pre-trained on a curated corpus of 1.2 million arXiv physics papers and fine-tuned with supervised data, PhysBERT outperforms leading general-purpose models on physics-specific NLP tasks.

## 1 Introduction

The field of physics encompasses a vast body of knowledge, spanning numerous sub-disciplines and theoretical frameworks. The specialized language used in physics publications [48] and the extensive corpus of information disseminated through academic journals, textbooks, technical reports, and online repositories present significant challenges for automated extraction of meaningful insights. To address these challenges, we introduce PhysBERT, a sentence embedding model specifically designed for the field of physics. Leveraging the BERT [9] architecture, PhysBERT is trained on a curated corpus of physics literature based on 1.2 million physics papers available on arXiv [4], encompassing a wide range of sub-disciplines within the field.

In this paper, we aim to validate the effectiveness of PhysBERT by creating specific datasets and downstream evaluation tasks such as information retrieval, classification, and semantic similarity, all tailored to the physics domain. The combination of comprehensive pre-training and targeted, supervised fine-tuning equips PhysBERT with a deep understanding of physics language, enabling it to significantly outperform general-purpose models on these physics-related NLP tasks. Additionally, we demonstrate that PhysBERT serves as an excellent starting point for fine-tuning in specific physics subdomains, highlighting its adaptability and potential for further specialization. A schematic overview of the workflow described in this paper is provided in Fig. 1. In addition to our model weights, we are releasing the training and evaluation datasets alongside this manuscript [3].

## 2 Related Work

Recent advancements in Natural Language Processing (NLP) are fundamentally transforming our ability to analyze and process textual data [26]. At the forefront of this transformation are text em-

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

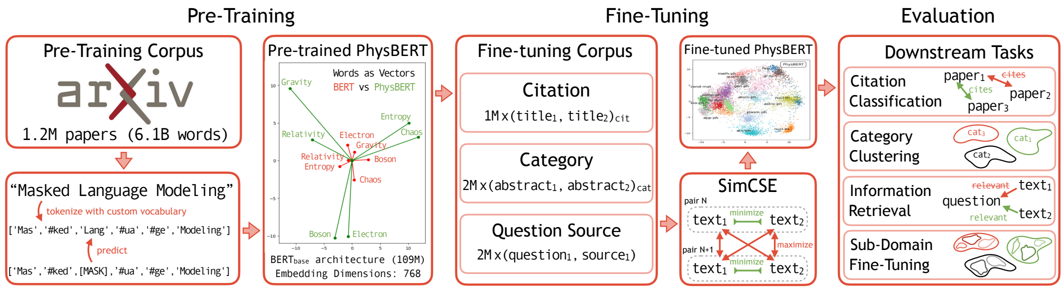


Figure 1: Schematic overview of the steps involved in developing the PhysBERT embedding model.

bedding models [25, 30], which convert textual data into dense vector representations, enabling computational analysis such as efficient information retrieval [23], text classification [16], and semantic similarity measurement [32]. In academic research, domain-specific embeddings can significantly enhance the accuracy of literature reviews by clustering related papers [38], identifying emerging trends [42], and improving the precision of reviewer matching tools for scientific journals [50]. In the last few years, Transformers [47] have become the foundation of these models [9, 32], which has significantly enhanced context awareness in NLP tasks. General-purpose text embedding models [11], typically trained on a diverse range of internet texts [27], lack the domain-specific modelling required to accurately represent the language of specific disciplines. Specialized embedding models have demonstrated significant improvements across various fields in natural science, including chemistry [37], material science [14], and the biomedical domain [18]. However, the domain of physics notably lacks embedding models specifically tailored to its unique semantic characteristics. Consequently, general-purpose embedding models are currently utilized in physics NLP applications due to the absence of specialized alternatives [15, 41, 31, 28, 40].

## 3 Downstream tasks

Due to the lack of publicly available benchmarks for scientific physics publications, we developed a custom set of assessments, closely following recognized text embedding benchmarks [27, 44].

**Category Clustering:** Sentences are paired with ground truth labels indicating their physics category. The sentences are first embedded into vector representations, and the KMeans [29] algorithm groups the embeddings into clusters, with the number of clusters matching the unique labels in the dataset. Clustering performance is evaluated using the V-measure score [35], with a stratified 10-fold cross-validation [36]. The final metric is the mean V-measure score across all test sets.

**Information Retrieval:** We follow common information retrieval benchmarking practices [44, 27]. Following standard RAG procedures, the embedding model transforms all queries and documents into embeddings, and cosine similarity scores are calculated between each query and all documents. Documents are then ranked based on these scores. Retrieval effectiveness is measured using the normalized Discounted Cumulative Gain at rank 10 (nDCG@10) [46].

**Citation Classification:** To evaluate the embedding models on the ability to correctly classify citing articles, we use a binary classification benchmark [32]. We use a balanced dataset with equal numbers of positive (citing) and negative (non-citing) pairs. Using cosine similarity, pairs are classified by identifying the optimal threshold separating positive and negative labels. The model's accuracy, referred to as cosine accuracy [39], is calculated based on the percentage of correct classifications.

**Fine-tuning on Physics Subdomains:** To demonstrate the effectiveness of PhysBERT as a foundation for domain-specific fine-tuning, we leverage the extensive nature of three large categories within arXiv—Condensed Matter, Astrophysics, and High Energy Physics—each of which comprises multiple subcategories. For instance, Astrophysics includes explicit subcategories such as 'Cosmology and Nongalactic Astrophysics' and 'Earth and Planetary Astrophysics' (see Ref. [5] for all categories). For the evaluation of this fine-tuning task we use a simplified setup akin to the supervised fine-tuning setup described above, with category clustering as the only evaluation metric.

# 4 Datasets

For unsupervised pre-training, we download all available papers from arXiv [4], including both PDFs and the available metadata using the provided bulk data access [6]. We source abstracts where full texts are not open access. We restrict the postprocessing to papers categorized by their authors under one of the 61 physics categories [5], which totalled to 1,25 million papers. All PDFs are processed using Nougat [8], and we utilize a postprocessed version containing only the full text of the sections, excluding captions, references, resulting in a corpus comprising 41 GB of text or about 6.1B words.

## 4.1 Supervised fine-tuning

**Abstract pairs from categories:** ArXiv publications are categorized based on the primary category assigned by the authors upon submission. To ensure robustness, we exclude categories with fewer than 5,000 papers and combine all subcategories under Astrophysics, Condensed Matter, and High Energy Physics—categories so extensive that they have subcategories—into their respective main categories. This approach leaves us with 21 categories, from which we draw 2 million abstract pairs, equally distributed across the categories to ensure a balanced dataset.

**Citation pairs:** We build a comprehensive citation tree using the Semantic Scholar [1] database API to query the references of papers in our arXiv database. By doing so, we can identify and pair the titles of papers that cite each other. We include 1M citation pairs in the training set.

**Synthetic Query-Source Data:** We use data augmentation, which artificially creates data to mimic real-world characteristics and patterns rather than directly collecting it [21]. Specifically, we generate 2M question-and-answer pairs from text chunks extracted from research papers, similar to standard RAG workflows [13]. We randomly select 1000-character text chunks from papers and use a locally running LLaMA3-70B [24] to generate three question-answer pairs exclusively answerable by the provided text.

## 4.2 Model evaluation data

For general physics clustering, we utilize 1,000 labeled paper abstracts from each of 21 major arXiv physics categories. Citation classification involves 50k pairs of citing and non-citing paper titles, while information retrieval uses 50k query-source pairs. All evaluation data is separate from the training sets. For subdomain fine-tuning, we focus on three large arXiv categories: Condensed Matter (10 subcategories), Astrophysics (7 subcategories), and High Energy Physics (4 subcategories). We create datasets with 10k abstract pairs per subcategory for training and 1k labeled abstracts per subcategory for evaluation, ensuring no overlap with training data.

# 5 Results

Given our extensive dataset of 40GB of text, which provides the capacity to train a new model from scratch, we build a custom tokenizer, following the BERT [9] approach with the standard vocabulary size of 30,523. We initialized the model with random weights corresponding to the $\text{BERT}_{\text{base}}$ architecture [9] and employed a pre-training strategy consistent with the RoBERTa methodology [22]. The training process was conducted with a batch size of 8, a Masked Language Modeling (MLM) probability of 15 %, and a learning rate of 1E-4, using the Adam optimizer [17]. This training was executed on 32 nodes, each equipped with 4 NVIDIA A100 GPUs at NERSC [2], utilizing PyTorch. The model was trained across four epochs with a sequence length of 128 tokens, followed by six epochs with a sequence length of 512 tokens, which takes about 10 hours. We refer to this model as $\text{PhysBERT}_{\text{MLM}}$. Following that we fine-tune [10] our model using Simple Contrastive Learning of Sentence Embeddings (SimCSE) [12] within the Sentence Transformer [32] framework, using semantically similar sentence pairs as described in Section 4, with all other sentences in the batch treated as negatives. We train for 2 epochs using eight A100 nodes which takes about 4 hours. Models are evaluated on all downstream tasks three times per epoch. After hyperparameter tuning, we set the learning rate to 2E-4, batch size to 256 per GPU, SimCSE temperature to 0.05, and weight decay to 0.01, using Adam as the optimizer. The best-performing model across three evaluation metrics, which we refer to as PhysBERT, is compared against models of particular interest to the

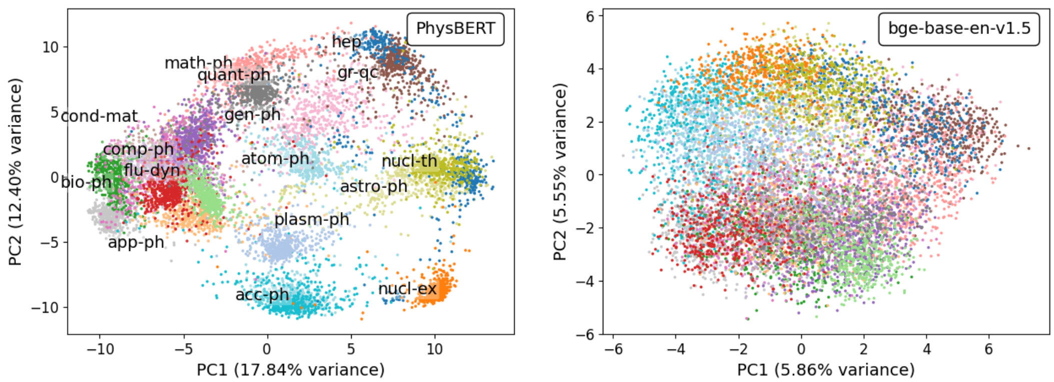


Figure 2: Comparison of embedding space visualizations for PhysBERT (left) and bge-base-v1.5 [49] (right, see also Table 1), using PCA on text embeddings from 500 random abstracts per physics category (as highlighted by different colors).

physics community [15, 41, 31, 28, 40] and top MTEB leaderboard models [11], including four derived from $\text{BERT}_{\text{large}}$.

The results in Table 1 demonstrate that PhysBERT surpasses existing models on all metrics. Notably, PhysBERT outperforms even larger models, highlighting its efficiency and superiority in handling complex physics-related NLP tasks despite its smaller size. Fig. 2 provides a visualization of the embedding space, where PCA is used to project 768-dimensional embeddings of 500 random abstracts from each physics category into two dimensions, providing a significantly better clustering.

Table 1: Downstream task results for various (uncased) text embedding models. Reported metrics include the average V-measure score for category clustering, cosine accuracy score for citation classification, and normalized Discounted Cumulative Gain at rank 10 (nDCG@10) for information retrieval. Additionally, the table presents the average V-measure scores for models fine-tuned in the physics subdomains of Condensed Matter, Astrophysics, and High Energy Physics, along with their overall average performance.

| | **Cit.Class.** | **Cat.Clust.** | **Inf.Retr.** | **Cond.Mat.** | **Astroph.** | **HEP** | **Avg.** |
|---|---|---|---|---|---|---|---|
| BERT[9] | 72.4 | 36.4 | 5.0 | 58.4 | 65.7 | 81.9 | 68.6 |
| bge-base-v1.5[49] | 89.5 | 58.1 | 46.3 | 60.0 | 67.5 | 84.9 | 70.8 |
| E5-base[45] | 83.4 | 54.8 | 52.5 | 58.7 | 67.3 | 82.8 | 69.6 |
| MiniLM-L6-v2[33] | 84.1 | 54.6 | 41.6 | 54.9 | 63.6 | 80.2 | 66.2 |
| mpnet-base[34] | 85.3 | 57.4 | 39.7 | 57.1 | 65.8 | 83.1 | 68.7 |
| PACuna[43] | 74.6 | 28.5 | 6.6 | 58.2 | 65.8 | 82.4 | 68.8 |
| RoBERTa[22] | 64.8 | 33.1 | 0.3 | 55.5 | 64.9 | 80.4 | 66.9 |
| SciBERT[7] | 75.5 | 44.8 | 4.1 | 59.7 | 66.4 | 85.0 | 70.4 |
| SPECTER2[38] | 83.4 | 52.0 | 6.6 | 60.0 | 67.2 | 85.0 | 70.7 |
| $\text{PhysBERT}_{\text{MLM}}$ (ours) | 60.1 | 49.1 | 6.9 | 60.9 | 68.5 | 86.8 | 72.1 |
| PhysBERT (ours) | **94.7** | **90.3** | **70.2** | **68.9** | **71.5** | **87.7** | **76.1** |
| **Large Models** | | | | | | | |
| E5-large[45] | 84.9 | 56.8 | 62.9 | 59.9 | 68.3 | 84.1 | 70.8 |
| UAE-Large-V1[20] | 89.7 | 58.3 | 50.0 | 60.3 | 68.0 | 85.0 | 71.1 |
| mxbai-large-v1[19] | 89.7 | 58.2 | 48.7 | 59.9 | 68.1 | 84.5 | 70.8 |
| bge-large-v1.5[49] | 89.6 | 58.3 | 52.3 | 60.1 | 67.9 | 84.1 | 70.7 |

We tested the fine-tuning capability of different models on three physics subdomains, training each model for 1 epoch using a linear learning rate decay. To ensure fair comparisons, we conducted a grid search to optimize learning rate and batch size within the ranges 1E-4, 2E-4 and 16, 32, respectively. Performance was evaluated three times during training on category clustering, and the checkpoint with the highest average V-measure score was reported in Table 1. Our fine-tuned PhysBERT outperformed other fine-tuned models, achieving the highest average V-measure across all categories, highlighting its potential as a robust foundation for domain-specific applications. Notably, $\text{PhysBERT}_{\text{MLM}}$, pre-trained only on MLM, outperformed larger reference models, demonstrating that unsupervised pre-training on a large physics corpus with domain-specific vocabulary provides a strong foundation for fine-tuning on specialized tasks.

## 6 Limitations

One limitation of our work is the reliance on self-created benchmark and training datasets due to the absence of publicly available physics datasets, which may impact generalizability. Additionally, while we focused primarily on text-based content, mathematical formulas, which can be important in certain physics literature, were not specifically addressed in this study.

## 7 Acknowledgments

The authors would like to express their gratitude to the NERSC team for their exceptional user support. Their dedication, patience and prompt responsiveness were instrumental in facilitating our computational endeavors, ensuring smooth resolution of any issues encountered.

Work supported by the Director of the Office of Science of the US Department of Energy under Contract no. DEAC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC award ERCAP0027412. The authors also acknowledge financial support from the PNRR Italian Ministry of University and Research (MUR) project PE0000013-FAIR.

## References

[1] Semantic Scholar. `https://www.semanticscholar.org/`. Accessed: 2024-07-23.

[2] National energy research scientific computing center, October 2020.

[3] Physbert model card. `https://huggingface.co/`, 2024.

[4] arXiv. `https://arxiv.org`. Accessed: 2024-07-12.

[5] arXiv. arxiv category taxonomy. `https://arxiv.org/category_taxonomy`. Accessed: 2024-07-12.

[6] arXiv. arxiv category taxonomy. `https://info.arxiv.org/help/bulk_data_s3.html`. Accessed: 2024-07-12.

[7] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[8] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents, 2023.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[10] Ning Ding, Yujia Qin, Guang Yang, Furu Wei, Zhihao Yang, Yuxuan Su, Shuming Hu, Yichong Chen, Chi-Min Chan, Weizhu Chen, Jing Yi, Weifeng Zhao, Xuedong Wang, Zheng Liu, He Zheng, Jian Chen, Ying Liu, Jialiang Tang, Jiangtao Li, and Ming Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5:220–235, 3 2023.

[11] Hugging Face. Massively multilingual text embedding benchmark (mteb) leaderboard. `https://huggingface.co/spaces/mteb/leaderboard`, 2024. Accessed: 2024-07-28.

[12] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings, 2022.

[13] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[14] Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.

[15] Alex Hexemer. Exploration of a beamline chatbot. Unpublished, July 2024.

[16] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[18] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[19] Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embeddings model, 2024.

[20] Xianming Li and Jing Li. Angle-optimized text embeddings, 2024.

[21] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data for language models, 2024.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[23] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[24] Meta-Llama-3-70B-Instruct. `https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct`, 2024. Accessed: 2024-07-12.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[26] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey, 2021.

[27] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark, 2023.

[28] Daniel Murnane, Gabriel Facini, Runze Li, Daniele Del Santo, and Cary Randazzo. chATLAS - An AI Assistant for the ATLAS Collaboration. Talk at the 1st Large Language Models in Physics Symposium, 2024.

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python, 2018.

[30] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[31] Florian Rehm. AccGPT - The Current Vision for AI Assistance at CERN's Accelerator Control and Beyond. Talk at the 1st Large Language Models in Physics Symposium, 2024.

[32] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[33] Nils Reimers and Iryna Gurevych. Sentencetransformers/all-minilm-l6-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`, 2020. Accessed: 2024-07-21.

[34] Nils Reimers and Iryna Gurevych. Sentencetransformers/all-mpnet-base-v2. `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`, 2020. Accessed: 2024-07-21.

[35] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[36] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[37] Shokirbek Shermukhamedov, Dilorom Mamurjonova, and Michael Probst. Structure to property: Chemical element embeddings and a deep learning approach for accurate prediction of chemical properties, 2023.

[38] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

[39] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[40] Peter Steinbach, Timo Niehoff, and Tino Gottschall. Extracting Measurements from (legacy) Publications. Talk at the 1st Large Language Models in Physics Symposium, 2024.

[41] A. Sulc et al. Towards Unlocking Insights from Logbooks Using AI. In *15th International Particle Accelerator Conference*, 5 2024.

[42] Antonin Sulc, Annika Eichler, Gregor Kasieczka, and Tim Wilksen. Illuminating the Dark: Discovering in Dark Matter Research through Natural Language Processing. Talk at the 1st Large Language Models in Physics Symposium, 2024.

[43] Antonin Sulc, Raimund Kammering, Annika Eichler, and Tim Wilksen. Pacuna: Automated fine-tuning of language models for particle accelerators, 2023.

[44] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models, 2021.

[45] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024.

[46] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of ndcg type ranking measures, 2013.

[47] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[48] Peter Wulff. Physics language and language use in physics—what do we know and how ai might enhance language-related research and instruction. *European Journal of Physics*, 45(2):023001, jan 2024.

[49] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[50] Yu Zhang, Yanzhen Shen, SeongKu Kang, Xiusi Chen, Bowen Jin, and Jiawei Han. Chain-of-factors paper-reviewer matching, 2024.