INDUS: Effective and Efficient Language Models for Scientific Applications

Bishwaranjan Bhattacharjee[1*], Aashka Trivedi[1], Masayasu Muraoka[1], Muthukumaran Ramasubramanian[3], Takuma Udagawa[1], Iksha Gurung[3], Nishan Pantha[3], Rong Zhang[1], Bharath Dandala[1], Rahul Ramachandran[2], Manil Maskey[2], Kaylin Bugbee[2], Mike Little[4], Elizabeth Fancher[2], Irina Gerasimov[5], Armin Mehrabian[5], Lauren Sanders[6], Sylvain Costes[6], Sergi Blanco-Cuaresma[7], Kelly Lockhart[7], Thomas Allen[7], Felix Grezes[7], Megan Ansdell[8], Alberto Accomazzi[7], Yousef El-Kurdi[1], Davis Wertheimer[1], Birgit Pfitzmann[10†], Cesar Berrospi Ramis[1], Michele Dolfi[1], Rafael Teixeira de Lima[1], Panagiotis Vagenas[1], S. Karthik Mukkavilli[1], Peter Staar[1], Sanaz Vahidinia[8], Ryan McGranaghan[9], Tsendgar Lee[8]

[1]IBM Research AI, [2]NASA MFSC, [3]UAH, [4] Navteca, [5] NASA GSFC, [6]NASA Ames, [7]Harvard-Smithsonian CfA, [8]NASA HQ, [9] JPL, [10]Smart City & ERZ Zurich

Abstract

Large language models (LLMs) trained on general domain corpora showed remarkable results on natural language processing (NLP) tasks. However, previous research demonstrated LLMs trained using domain-focused corpora perform better on specialized tasks. Inspired by this insight, we developed INDUS, a comprehensive suite of LLMs tailored for the closely-related domains of Earth science, biology, physics, heliophysics, planetary sciences and astrophysics, and trained using curated scientific corpora drawn from diverse data sources. The suite of models include: (1) an encoder model trained using domain-specific vocabulary and corpora to address NLP tasks, (2) a contrastive-learning based text embedding model trained using a diverse set of datasets to address information retrieval tasks and (3) smaller versions of these models created using knowledge distillation for applications which have latency or resource constraints. We also created three new scientific benchmark datasets, CLIMATE-CHANGE NER (entity-recognition), NASA-QA (extractive QA) and NASA-IR (IR) to accelerate research in these multi-disciplinary fields. We show that our models outperform both general-purpose (RoBERTa) and domain-specific (SCIBERT) encoders on these new tasks as well as existing tasks in the domains of interest. Furthermore, we demonstrate the use of these models in two industrial settings- as a retrieval model for large-scale vector search applications and in automatic content tagging systems.

1 Introduction

Large language models (LLMs) trained on huge amounts of data have demonstrated impressive capabilities on natural language understanding and generation tasks. Most popular LLMs rely on the transformer architecture (Vaswani et al., 2017) and are trained using general-purpose corpora like Wikipedia or CommonCrawl (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020; Raffel et al., 2020; Brown et al., 2020; Touvron et al., 2023). Although these general-purpose models exhibited strong performance, the distributional shift of vocabulary led to sub-optimal performance on domain-specific natural language understanding and generation tasks (Beltagy et al., 2019). Following this observation, several domain-specific LLMs like SCIBERT (Beltagy et al., 2019), BIOBERT (Lee et al., 2019), MATBERT (Walker et al., 2021), BATTERYBERT (Huang and Cole, 2022) and SCHOLARBERT (Hong et al., 2023) were developed to improve accuracy on in-domain NLP tasks.

In this research, we specifically focused on interdisciplinary scientific topics related to astrophysics, physics, Earth science, heliophysics, planetary sciences and biology. While the training corpora of existing domain-specific models such as SCIBERT, BIOBERT and SCHOLARBERT partially cover some of these fields, there is no model available that encompasses all of the fields of interest collectively.

Thus, we developed INDUS, a collection of encoder-based LLMs focused on these domains of interest (Figure 1) trained using curated corpora from diverse sources. Specifically, we make the following contributions:

1. Utilizing the byte-pair encoding (BPE) algorithm, we constructed INDUSBPE, a customized tokenizer from the curated scientific corpus.
2. We pretrained **encoder-only LLMs** using curated scientific corpora and the INDUSBPE tokenizer (§2, §3). We further created **sentence-**

*Correspondence: bhatta@ibm.com, mr0051@uah.edu, aashka.trivedi@ibm.com, rahul.ramachandran@nasa.gov

†Work done while at IBM Research AI

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 98–112

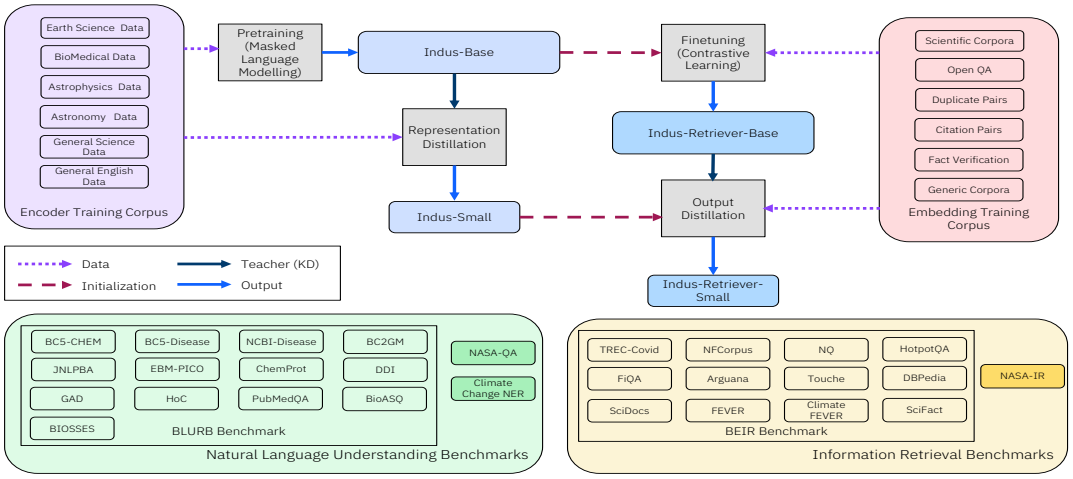

Figure 1: Overview of INDUS models: the general-purpose encoder model and the retriever built from it, and their distilled counterparts. Also shown are the benchmarks used for evaluation, highlighting our new benchmarks, NASA-QA, CLIMATE-CHANGE NER and NASA-IR.

embedding models by fine-tuning the encoder-only models with a contrastive learning objective (§4). We also trained **smaller, efficient versions** of these models using distillation.

3. We created **three new scientific benchmark datasets**, CLIMATE-CHANGE NER (an entity recognition task), NASA-QA (an extractive question answering task) and NASA-IR (a retrieval task) (§5) to further accelerate research in this multi-disciplinary field.
4. We demonstrate strong performance by our models on these benchmark tasks as well as on existing domain-specific benchmarks, outperforming general-purpose models like ROBERTa (Liu et al., 2019) as well as scientific-domain encoders like SCIBERT (Beltagy et al., 2019). We also show that the knowledge-distilled models achieved a significant reduction in latency while maintaining strong performance compared to the original models on most of these tasks.
5. We describe two industrial application areas of INDUS models in the scientific domain, where they outperform existing general-purpose models.

2 Data

Sufficient high-quality in-domain corpora is essential to develop models that perform better than their counterparts trained on open-domain corpora. We meticulously identified corpora for each of the aforementioned domains, and created English-only models for containment. Specifically, for each domain, we used open-source data which has a permissive license, and further augmented them with full text papers and material contributed by providers mentioned below. We now briefly describe each data source, and present statistics of the data in Table 1.

Dataset	Domain	#Tokens	Ratio
NASA CMR	Earth Science	0.3B	1%
AMS and AGU papers	Earth Science	2.8B	4%
English Wikipedia	General	5.0B	8%
PubMed Abstracts	Biomedical	6.9B	10%
PMC	Biomedical	18.5B	28%
SAO/NASA ADS	Astronomy, Astrophysics, Physics, General Science	32.7B	49%
Total		66.2B	100%

Table 1: Basic statistics of our pretraining dataset.

- **SAO/NASA Astrophysics Data System (ADS)**[1]: The biggest source of data, covering publications in astronomy and astrophysics, physics and general science including all arXiv e-prints.
- **PubMed Central (PMC)**[2] : An archive of biomedical and life science journal literature maintained by National Library of Medicine and National Institutes of Health. We used the portion of PMC that has a commercial-friendly license, along with the PubMed abstracts of all the articles in PMC.
- **American Meteorological Society (AMS)**[3]: We used full-text journal documents spanning topics in Earth systems, Earth interac-

[1]https://ui.adsabs.harvard.edu
[2]https://www.ncbi.nlm.nih.gov/pmc
[3]https://www.ametsoc.org/index.cfm/ams/publications/

Tokenizer	ADS	PMC	Wikipedia
RoBERTa	12,867,439	7,549,075	15,859
+lower_cased	12,862,227	7,557,868	16,901
INDUSBPE	12,309,023	6,920,659	16,056

Table 2: Number of tokens produced by RoBERTa and INDUSBPE tokenizers on 1k samples from each dataset. Fewer tokens lead to a smaller computation cost.

tions, applied meteorology, climatology, physical oceanography, atmospheric sciences, climate, hydrometeorology, weather, forecasting, and societal impacts.

- **American Geophysical Union (AGU)**[4]: Journal documents on the topics of atmospheres, biogeosciences, Earth surface, machine learning and computation, oceans, planets, solid Earth, and space physics.
- **NASA Common Metadata Repository (CMR)**[5]: A high-quality, continuously evolving metadata system that catalogs all data and service metadata records for NASA's Earth Science Data and Information System.

3 Methodology: Encoder Models

INDUSBPE Tokenizer We trained an uncased BPE tokenizer (Radford et al., 2019), INDUSBPE, using a subset of our training dataset (§2). We set the vocabulary size to 50265 (equal to that of the RoBERTa tokenizer (Liu et al., 2019)).

We performed a brief analysis of the differences between the vocabularies of INDUSBPE and the RoBERTa tokenizer. Out of 50265 tokens, 44.5% tokens are common in both tokenizers while the remaining 55.5% tokens are included only in either tokenizer, indicating a significant distributional shift in domain. To further understand this effect, we applied both tokenizers on 1000 randomly sampled text fragments from our datasets. As shown in Table 2, INDUSBPE tokenizer produced fewer tokens than the RoBERTa tokenizer, leading to an 8% drop in computation cost during training.

Encoder Model We trained $\text{INDUS}_{\text{BASE}}$[6] using a masked language modeling objective. The model architecture follows $\text{RoBERTa}_{\text{BASE}}$ (Liu et al., 2019), with 12 layers and 125M parameters.

Knowledge Distillation for Efficient Encoder Model We also trained a smaller model, $\text{INDUS}_{\text{SMALL}}$[7], with 38M parameters through knowledge distillation using $\text{INDUS}_{\text{BASE}}$ as the teacher. $\text{INDUS}_{\text{SMALL}}$ follows a 4-layer architecture recommended by the Neural Architecture Search engine (Trivedi et al., 2023) with an optimal trade-off between performance and latency. We adopted the distillation objective proposed in MiniLMv2 (Wang et al., 2021) to transfer fine-grained self-attention relations, which has been shown to be the current state-of-the-art (Udagawa et al., 2023).

4 Methodology: Sentence Embedding Models

Sentence embedding models represent text as low-dimensional vectors for efficient use in dense retrieval systems, such as Retrieval Augmented Generation, where relevant passages for a query are identified by the similarity between their embeddings (Karpukhin et al., 2020). Embedding models are trained using a contrastive learning objective (Khosla et al., 2020; Gao et al., 2021), which pushes the embeddings of a query closer to those of relevant passages and further away from those of non-relevant ones. We use the improved contrastive loss proposed in Li et al. (2023) which introduces an additional bidirectional signal to expand negatives.

Base Embedding Model We created our sentence embedding model, $\text{INDUS-RETRIEVER}_{\text{BASE}}$[8], by fine-tuning $\text{INDUS}_{\text{BASE}}$, following a bi-encoder framework (Reimers and Gurevych, 2019). Similar to prior work (Wang et al., 2022; Li et al., 2023; Xiao et al., 2023), we employed a stage-wise training approach. We first train on a large corpus of naturally occurring pairs collected from internet sources, and specifically include data from the science domain. Furthermore, we created a domain-specific dataset from the ADS data (§2) by including title-abstract pairs. Then, we finetune on high quality annotated datasets (e.g., question-answer pairs). Appendix B contains comprehensive details about the datasets used in training. For both stages, we used large batch sizes and in-batch negatives to better approximate the contrastive objective.

Knowledge Distillation for Embedding Model To optimize the latency for retrieval applications, we also created a small retriever model, INDUS-

[4]https://agupubs.onlinelibrary.wiley.com/

[5]https://www.earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/cmr

[6]https://huggingface.co/nasa-impact/nasa-smd-ibm-v0.1

[7]https://huggingface.co/nasa-impact/nasa-smd-ibm-distil-v0.1

[8]https://huggingface.co/nasa-impact/nasa-smd-ibm-st-v2

	Train	Validation	Test
Num. Abstracts	382	77	75
Num. Tokens	32,031	6,443	5,850
Entity Labels			
climate-nature, climate-greenhouse-gases, climate-assets, climate-problem-origins, climate-mitigations, climate-properties, climate-impacts, climate-datasets, climate-organizations, climate-observations, climate-models, climate-hazards, climate-organisms			

Table 3: CLIMATE-CHANGE NER statistics and entities.

RETRIEVER$_{\text{SMALL}}$[9], with the aim to transfer the capability of the large teacher model (INDUS-RETRIEVER$_{\text{BASE}}$, with 12 layers and an embedding dimension of 768) to smaller student model (INDUS$_{\text{SMALL}}$, with 4 layers and an embedding dimension of 576), by distilling the teacher's distribution of similarity scores. Specifically, we use the distillation loss described in Xu et al. (2023)

Here, we find it beneficial to first conduct an embedding-oriented pretraining step, as presented in Retro-MAE (Xiao et al., 2022), on about 56M sentences from English Wikipedia, BooksCorpus, and StackExchange data. We observed that this step is not necessary in the larger model, but provides significant improvement in the smaller one. For distillation, we found that a stage-wise training approach does not benefit performance (ablation presented in Appendix E). We thus distilled in a single step with all the data described in Appendix B, also adding labelled pairs from FEVER (Thorne et al., 2018) and HOTPOTQA (Yang et al., 2018).

5 Creating Benchmarks

Benchmark datasets play a crucial role in assessing the language understanding capabilities of models. However, there is an absence of datasets tailored for the diverse and multidisciplinary fields under study. Thus, to effectively benchmark the proposed NLP models, we introduced three new datasets for NER, QA and IR. Appendix D compares the sizes of these datasets to popularly used benchmarks.

5.1 CLIMATE-CHANGE NER

CLIMATE-CHANGE NER[10] focuses on understanding and addressing climate-related topics across various domains. This comprises 534 abstracts sourced from Semantic Scholar Academic Graph (Kinney et al., 2023), collected using a seed set of climate-related keywords such as *wildfire* or *floods*. The abstracts were annotated with entities of interest that originate from complex taxonomies used in climate-related literature as shown in Table 3.

5.2 NASA-QA

We created NASA-QA[11], an extractive QA benchmark dataset focused on the Earth science domain (ES). Specifically, we sourced 39 paragraphs from ES papers appearing in AGU and AMS journals (§2), and subject matter experts formulated questions and annotated the spans of the paragraph that contain the answer. We used 29 paragraphs (145 questions) as the training set and remaining 10 paragraphs (50 questions) for evaluation. The training set was further augmented with paragraphs and QA pairs related to ES (oxygen, amazon rain forest and geology) from the SQUAD dataset (Rajpurkar et al., 2018). This resulted in a training set comprising 686 paragraphs with 5,081 questions (2,817 answerable and 2,264 unanswerable).

5.3 NASA-IR

Finally, we constructed a domain-specific information retrieval benchmark dataset, NASA-IR[12], spanning almost 500 QA pairs related to the Earth science, planetary science, heliophysics, astrophysics and biological physical sciences domains. We sampled a set of 166 paragraphs from AGU, AMS, ADS, PMC and PubMed (§2) and manually annotated them with 3 questions that are answerable from each of these paragraphs, resulting in 498 questions (398 questions in the test set and 100 in the validation set- this test is designed to be evaluated in a zero shot fashion). We also sampled random abstracts from ADS to enhance our corpus. Each question has only one relevant document, and we use the Recall@10 evaluation metric.

6 Experimental Results

Baselines We compared INDUS models against open source models of similar sizes (all models obtained from HuggingFace):

- INDUS$_{\text{BASE}}$ was compared to ROBERTa$_{\text{BASE}}$, SCIBERT, PUBMEDBERT, and BIOLINKBERT.
- INDUS$_{\text{SMALL}}$ was compared to MINILM (6-layer) and TINYBERT (4-layer).

[9] https://huggingface.co/nasa-impact/nasa-ibm-st.38m

[10] https://huggingface.co/datasets/ibm/Climate-Change-NER

[11] https://huggingface.co/datasets/nasa-impact/nasa-smd-qa-benchmark

[12] https://huggingface.co/datasets/nasa-impact/nasa-smd-IR-benchmark

Task	Metric	Dataset	Base model (125M params.)					Small model (~30M params.)		
			RoBERTa	SciBERT	PubMed	BioLink	$\text{INDUS}_{\text{BASE}}$	TinyBERT	MiniLM	$\text{INDUS}_{\text{SMALL}}$
NER	Entity F1	BC5-chem	90.3 (0.2)	91.4 (0.2)	93.2 (0.1)	**93.3** (0.2)	**93.3** (0.2)	84.6 (0.2)	86.1 (0.3)	**90.7** (0.1)
		BC5-disease	81.5 (0.3)	83.7 (0.3)	**85.4** (0.3)	85.3 (0.3)	85.2 (0.3)	74.0 (0.4)	77.4 (0.3)	**81.3** (0.3)
		NCBI-disease	87.6 (0.6)	87.6 (0.4)	88.2 (0.6)	88.2 (0.5)	**88.3** (0.4)	81.2 (0.4)	83.1 (0.5)	**85.6** (0.6)
		BC2GM	82.1 (0.3)	82.3 (0.2)	84.3 (0.3)	**84.7** (0.2)	84.0 (0.3)	74.7 (0.4)	77.1 (0.2)	**79.7** (0.3)
		JNLPBA	79.1 (0.2)	78.2 (0.2)	79.3 (0.2)	78.9 (0.2)	**80.3** (0.2)	70.3 (0.2)	73.4 (0.3)	**75.7** (0.2)
PICO	Macro F1	EBM PICO	72.3 (0.3)	72.4 (0.3)	72.9 (0.3)	**73.4** (0.2)	73.1 (0.2)	67.4 (0.2)	70.3 (0.1)	**73.1** (0.2)
Relation Extraction	Micro F1	ChemProt	50.4 (28.2)	73.9 (0.7)	77.2 (0.6)	**77.9** (0.4)	76.9 (0.5)	56.2 (3.2)	55.9 (2.1)	**71.7** (0.9)
		DDI	78.6 (1.5)	80.1 (1.0)	80.6 (1.1)	81.2 (0.6)	**81.7** (0.5)	39.3 (5.3)	51.5 (2.9)	**69.0** (1.2)
		GAD	80.0 (1.1)	81.6 (1.2)	**82.4** (1.2)	82.1 (1.5)	79.4 (5.6)	76.4 (1.3)	77.3 (1.0)	**81.3** (0.7)
Document Classification	Micro F1	HoC	82.2 (0.7)	83.1 (0.6)	**84.5** (0.4)	84.4 (0.5)	83.7 (0.5)	41.6 (6.8)	62.8 (4.7)	**80.2** (0.6)
Question Answering	Accuracy	PubMedQA	53.1 (3.3)	54.3 (3.8)	55.2 (5.5)	**59.1** (6.2)	58.2 (6.7)	50.3 (1.4)	51.6 (1.7)	**56.1** (1.4)
		BioASQ	69.1 (4.8)	74.6 (4.5)	84.3 (5.5)	**84.9** (10.5)	69.6 (5.8)	74.3 (3.6)	66.7 (2.3)	**75.4** (3.3)
Sentence Similarity	Pearson	BIOSSES	79.8 (6.3)	86.3 (3.5)	**92.2** (1.1)	91.1 (2.6)	72.2 (9.5)	**88.2** (1.1)	26.6 (8.7)	70.4 (3.3)
Micro Average	-	-	75.9 (3.7)	79.2 (1.3)	81.5 (1.3)	**81.9** (1.8)	78.9 (2.4)	67.6 (1.9)	66.1 (1.9)	**76.2** (1.0)
Macro Average	-	-	74.9 (3.7)	78.2 (1.6)	80.9 (1.4)	**81.2** (3.9)	76.4 (3.2)	65.6 (2.4)	60.6 (3.0)	**74.3** (1.3)

Table 4: Evaluation on BLURB. Standard deviation across 10 random seeds in parenthesis. Macro avg. reported across datasets and micro avg. computed by averaging scores on each task then averaging across task averages.

- $\text{INDUS-RETRIEVER}_{\text{BASE}}$ was compared to BGE_{BASE} and a $\text{RoBERTa}_{\text{BASE}}$ model finetuned with the same method presented in §4.
- $\text{INDUS-RETRIEVER}_{\text{SMALL}}$ was compared to MINILM-V2 and $\text{BGE}_{\text{SMALL}}$.

6.1 Natural Language Understanding Benchmarks

6.1.1 BLURB

We evaluated our models on BLURB (Gu et al., 2021), a benchmark suite for natural language understanding and reasoning tasks in the biomedical domain. We followed the original work to compute the overall score (i.e., macro average).

Table 4 shows the evaluation results. Among base models, $\text{INDUS}_{\text{BASE}}$ significantly outperformed the general-purpose RoBERTa model while achieving competitive performance to the bio-domain-specific models, namely SciBERT, PubMedBERT, and BioLinkBert, in which the Macro Average of our model is still within two standard deviations ($76.4+3.2*2 = 82.8$), thus, the differences are not statistically significant. For smaller models, we noticed $\text{INDUS}_{\text{SMALL}}$ outperformed the baselines, TinyBERT and MiniLM, by a large margin in most cases, showing significant difference from second best models in NER, PICO, relation extraction, and document classification tasks. This demonstrates the effectiveness of knowledge distillation from our domain-specific teacher model, $\text{INDUS}_{\text{BASE}}$.

We noticed domain specific large baseline models tend to perform better than our model on paired input-text tasks, such as QA and semantic similarity tasks, although the results have relatively large standard deviations. We hypothesize that pretraining with paired texts in BERT-style models (e.g., SciBERT and PubMedBERT) in contrast to the RoBERTa-style models (e.g., RoBERTa and INDUS) may be beneficial for such paired input-text tasks. This is consistent with the observations of Tinn et al. (2023)[13].

Model	CLIMATE-CHANGE NER F1 (SD)	NASA-QA F1 (SD)
RoBERTa	60.8 (0.8)	66.8 (3.1)
SciBERT	61.8 (0.7)	63.5 (1.9)
$\text{INDUS}_{\text{BASE}}$	**64.0** (1.0)	**68.2** (2.9)
TinyBERT	34.3 (1.6)	43.2 (2.3)
MiniLM	44.7 (1.3)	**59.2** (3.9)
$\text{INDUS}_{\text{SMALL}}$	**54.8** (0.8)	47.4 (1.8)

Table 5: CLIMATE-CHANGE NER and NASA-QA benchmark results. Standard deviation for CLIMATE-CHANGE NER over 10 random seeds and NASA-QA over 3 random seeds in parenthesis.

6.1.2 CLIMATE-CHANGE NER

As shown in Table 5, our models clearly outperformed the corresponding baseline models on the CLIMATE-CHANGE NER task, suggesting the effectiveness of training on large domain-specific data.

6.1.3 NASA-QA

As mentioned in §5, we augmented the training set with relevant SQuAD pairs for fine-tuning. All models are fine tuned for 15 epochs, and the results are shown in Table 5. We observed that $\text{INDUS}_{\text{BASE}}$ outperformed all models of similar sizes, while $\text{INDUS}_{\text{SMALL}}$ had relatively strong performance compared to its counterparts.

[13] Specifically, as noted in their paper,"*pretraining with single sequences leads to a substantial performance drop in the sentence similarity task. ... therefore pretraining with 2 text segments helps.*"

Model	NASA-IR ↑ (Recall@10)	BEIR Avg. ↑ (NDCG@10)	Retrieval Time ↓ (s)
RoBERTa$_{BASE}$	0.66	0.37	1.20
BGE$_{BASE}$	0.67	**0.52**	1.18
INDUS-RETRIEVER$_{BASE}$	**0.71**	0.41	1.19
MINILM-V2	0.62	0.39	0.24
BGE$_{SMALL}$	0.66	**0.51**	0.42
INDUS-RETRIEVER$_{SMALL}$	**0.73**	0.42	0.26

Table 6: Evaluation results on NASA-IR and BEIR, and average retrieval time per query on the NQ test set on an A100 GPU. Retrieval time includes time to encode the query & corpus and time to retrieve relevant documents.

6.2 Information Retrieval Benchmarks

We evaluated our models on the NASA-IR dataset as well as BEIR Benchmark (Thakur et al., 2021), which consists of 12 retrieval tasks spanning a variety of domains. The BEIR benchmark used the Normalized Cumulative Discount Gain (nDCG@10) metric. As shown in Table 6, both of our sentence embedding models significantly outperform the baselines on the NASA-IR task while still maintaining good performance on several of the BEIR tasks (individual results on BEIR tasks shown in Appendix F). Notably, INDUS-RETRIEVER$_{SMALL}$ outperformed INDUS-RETRIEVER$_{BASE}$, on both NASA-IR and BEIR, while being about 4.6x faster.

7 Industrial Applications of INDUS

We show industrial applications of INDUS models for downstream tasks in the scientific domain.

7.1 Retrieval and Vector Search

NASA developed the Science Discovery Engine (SDE)[14], a search capability that enables the discovery of open data, software and documentation across astrophysics, biological and physical Sciences, Earth science, heliophysics and planetary science (Bugbee et al., 2022). To improve search performance, we developed a document retrieval and extractive QA pipeline using the finetuned INDUS models, with the following components:

- **Sentence Embedding Model:** We use INDUS-RETRIEVER$_{BASE}$ to encode a corpus into a vector database, enabling the retrieval of relevant documents based on a user query.
- **Document Re-Ranker Model:** To further improve the relevancy of search results, the retrieved documents are ranked using a document re-ranker model INDUS$_{RANKER}$[15]. This model is fine-tuned from INDUS$_{BASE}$ on the MS-MARCO dataset (Bajaj et al., 2016).
- **Extractive QA Model:** Answers are extracted using a QA model finetuned from INDUS$_{BASE}$.

This system is expected to be live by mid-December 2024.

	RoBERTa$_{BASE}$	INDUS$_{BASE}$
MS-MARCO (MRR@5)	35.9	**36.4**
NASA-QA (MRR@5)	31.1	**33.2**

Table 7: MRR@5 on re-ranking NASA-QA and MS-MARCO tasks using rerankers finetuned from different base models.

Model	Document Retrieval Score		Answer Quality
	MRR@1	MRR@3	Avg. Quality Score
RoBERTa$_{BASE}$	0.54	0.62	0.60
INDUS$_{BASE}$	**0.69**	**0.78**	**0.88**

Table 8: Avg. Document Retrieval and Answer Quality Scores for 26 questions formulated by experts across astrophysics, biology & physical science, Earth science, heliophysics & planetary science domains.

First, we compare the performance of INDUS$_{RANKER}$ to an identical re-ranker finetuned from RoBERTa$_{BASE}$ in Table 7. Here, we measure MRR@5 of correctly ranking the most relevant paragraph for the given question. While the INDUS$_{RANKER}$ has comparable performance to the RoBERTa-reranker on the MS-Marco dev set, it significantly outperforms the latter on the NASA-QA dataset, alluding to better domain contextualization of the INDUS$_{BASE}$ model.

We then evaluated the end-to-end performance of the domain-adapted model verses the generic RoBERTa model in the aforementioned pipeline. Both systems were queried with a set of questions spanning various thematic areas, and then manually scored by human annotators based on the document relevance and correctness of the extracted answers. For assessing document retrieval quality, we use the **MRR@1** and **MRR@3** metric, which computes the average reciprocal rank of the highest ranked document from the system's top-1 and top-3 retrieved documents respectively. For answer quality, experts mark an **Answer Quality Score**. A score of 1 indicates the correct answer is returned within the first three snippets (a contiguous chunk from the document), 0.5 indicates that the answer is returned in more than three snippets, and 0 indicates no relevant answer is returned. Table 8 shows the superior scores when using INDUS models, most likely due to the overlap in domain and verbiage of

[14] https://sciencediscoveryengine.nasa.gov/

[15] https://huggingface.co/nasa-impact/nasa-smd-ibm-ranker

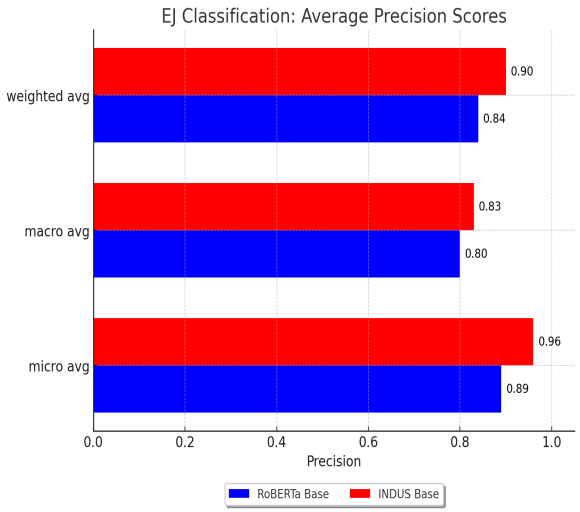

Figure 2: Average Precision Scores of the EJ Indicators Classification Test Set.

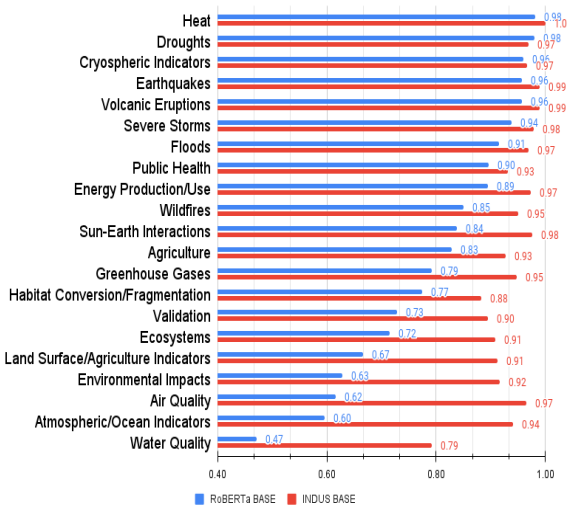

Figure 3: F1-Scores of the classes (GCMD Applied Research Areas) over 1036 test samples.

the content indexed by the SDE and training corpus of INDUS models. Example responses from both systems, and a screenshot of the system is shown in Appendix G.

7.2 Automated Content Curation

Environmental Justice Portal in SDE Content curation is a crucial step in providing a high-quality search experience the SDE, where Subject Matter Experts (SMEs) identify scientifically relevant information to make available for search and discovery. INDUS models are being used to automate this time consuming process, for example to identify datasets for specialized search applications like the Environmental Justice Data Search Interface[16], which focuses on data and metadata related to environmental justice (EJ). SMEs identified relevant EJ datasets and tagged them with eight indicators: *Human Dimensions, Health & Air Quality, Climate Change, Food Availability, Disasters, Urban Flooding, Extreme Heat, Water Availability*. This resulted in 139 classification samples which was used to finetune $\text{INDUS}_{\text{BASE}}$ to develop the multi-label classifier, $\text{EJ}_{\text{CLASSIFIER}}$[17]. We also added another "Not-EJ" class to identify documents that are not related to EJ. This classifier is being used to identify relevant EJ content from the SDE (live by mid-December 2024). To evaluate model performance, we used a held-out test set comprising 20% of the 139 samples, stratified equally across all indicators. As shown in Figure 2, the domain-specific model fine-tuned from $\text{INDUS}_{\text{BASE}}$ has higher precision than the general-purpose model $\text{RoBERTa}_{\text{BASE}}$.

GCMD Applied Research Area Tags Beyond SDE, we apply $\text{INDUS}_{\text{BASE}}$ to categorize scientific publications into 21 Applied Research Areas from the Global Change Master Directory (GCMD), as part of a collection that cites datasets from NASA's Goddard Earth Sciences Data and Information Services Center (GES-DISC) and have been annotated by experts. Each publication is annotated with multiple applied research areas allowing for multi-label classification, as detailed in Gerasimov et al. (2024). $\text{INDUS}_{\text{BASE}}$ was finetined to categorize scientific texts into the aforementioned categories, and is used to enhance publication and dataset discovery in GES-DISC Portal. We evaluate the model's performance on 1036 unseen publications, and show in Figure 3 that $\text{INDUS}_{\text{BASE}}$ outperforms finetuned $\text{RoBERTa}_{\text{BASE}}$ by 16% in terms of macro average F1 score.

8 Conclusion

In this work, we presented INDUS, a constellation of models for use in the science domain and show their applications in industrial settings. We demonstrated the effectiveness of a custom tokenizer and in-domain data for training high-quality encoder models and sentence embedding models. Further, we created smaller versions of the models suitable for applications with latency or resource constraints through state-of-the-art knowledge distillation techniques. For the benefit of the scientific community, we have released all models and benchmarks.

[16] https://sciencediscoveryengine.nasa.gov/app/nasa-sba-ej/#/ej/home

[17] https://huggingface.co/nasa-impact/ej-classification

Acknowledgements

This work is supported by NASA Grant 80MSFC22M004. We thank all the SMEs who contributed towards the datasets introduced in the paper. We also thank American Geophysical Union (AGU) and the American Meteorological Society (AMS) for providing scientific papers and articles to help build the corpus for the pre-training the INDUS$_{\text{BASE}}$ model. We thank the ML team at Sinequa for providing assistance and expertise in finetuning the INDUS models for prototyping in the SDE. We also acknowledge support from the IBM Research AI Hardware Center and the Center for Computational Innovation (CCI) at Rensselaer Polytechnic Institute for computational resources on the AiMOS Supercomputer.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Kaylin Bugbee, Rahul Ramachandran, Ashish Acharya, Dai-Hai Ton That, John Hedman, Ahmed Eleish, Charles Driessnack, Wesley Adams, and Emily Foshee. 2022. Selecting approaches for enabling enterprise data search: Nasa's science mission directorate (smd) catalog. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 6836–6839. IEEE.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the use of arxiv as a dataset.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1156–1165, New York, NY, USA. Association for Computing Machinery.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

GCMD. Global change master directory (gcmd). `https://catalog.data.gov/dataset/global-change-master-directory-gcmd`. Accessed: 8 October 2024.

Irina Gerasimov, Andrey Savtchenko, Jerome Alfred, James Acker, Jennifer Wei, and KC Binita. 2024. Bridging the gap: Enhancing prominence and provenance of nasa datasets in research publications. *Data Science Journal*, 23(1).

GES-DISC Portal. Nasa publications. `https://disc.gsfc.nasa.gov/information/publications`. Accessed: 8 October 2024.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1265–1268, New York, NY, USA. Association for Computing Machinery.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the Knowledge in a Neural Network. In *NeurIPS Deep Learning Worksop*.

Zhi Hong, Aswathy Ajith, Gregory Pauloski, Eamon Duede, Kyle Chard, and Ian Foster. 2023. The diminishing returns of masked language models to science.

Shu Huang and Jacqueline M Cole. 2022. Batterybert: A pretrained language model for battery database enhancement. *J. Chem. Inf. Model.*, page DOI: 10.1021/acs.jcim.2c00035.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the ACL*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, and Georgios Paliouras. 2020. *Results of the Seventh Edition of the BioASQ Challenge*, page 553–568. Springer International Publishing.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Aashka Trivedi, Takuma Udagawa, Michele Merler, Rameswar Panda, Yousef El-Kurdi, and Bishwaranjan Bhattacharjee. 2023. Neural architecture search for effective teacher-student knowledge transfer in language models. *arXiv preprint arXiv:2303.09639*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, and Ion Androutsopoulos. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*.

Takuma Udagawa, Aashka Trivedi, Michele Merler, and Bishwaranjan Bhattacharjee. 2023. A comparative analysis of task-agnostic distillation methods for compressing transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 20–31, Singapore. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. 2021. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *Available at SSRN 3950755*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. DistillCSE: Distilled contrastive learning for

sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8153–8165, Singapore. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Training Details: Encoder Models

INDUS$_{\text{BASE}}$ was trained with the masked language modeling objective, using the default hyperparameters recommended in Table 9 of Liu et al. (2019).We change the effective batch size to 9216, training for 500K steps on 192 V100 GPUs.

INDUS$_{\text{SMALL}}$ was distilled using the MiniLMv2 approach (Wang et al., 2021), with an effective batch size of 480 for 500K steps on 30 V100 GPUs.

B Sentence Embedding Training Data

Table 9 shows the various data sources used for training embedding models. All data is presented in the form of text-pairs, where each item in the pair may be a sentence or a paragraph. We used about 360 million pairs for training and used in-batch negatives.

C Training Details: Sentence Embedding

For the base retriever model, we use the following loss: for a triple $\{q, p^+, P^-\}$ of a query, a relevant (positive) passage, and a set of non-relevant (negative) passages $P^- = \{p_j^-\}_{j=1}^m$, We define the InfoNCE loss (van den Oord et al., 2019) as:

$$\mathcal{L}_{IC} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s(q_i,p_i^+)}}{Z_i} \tag{1}$$

$$\begin{aligned} Z_i = &\sum_j e^{s(q_i,p_j)} + \sum_j e^{s(q_j,p_i^+)} \\ &+ \sum_{j\neq i} e^{s(q_i,q_j)} + \sum_{j\neq i} e^{s(p_i^+,p_j)^-} \end{aligned} \tag{2}$$

where $s(q,p)$ is a measure of temperature-scaled cosine similarity between the embeddings of query and a passage measured by (where $\mathbf{E}(\cdot)$ denotes the embedding function and τ is the temperature):

$$s(q,p) = \frac{1}{\tau}\frac{\mathbf{E}(q)\cdot\mathbf{E}(p)}{\|\mathbf{E}(q)\|\|\mathbf{E}(p)\|} \tag{3}$$

We trained each stage on 2 A100 GPUs with an effective batch size of 1,024. We first trained with unsupervised data for 300K steps followed by an additional 100K steps with the supervised data. We used a learning rate of $2e-5$ and $\tau = 0.02$ during both these steps.

We used knowledge distillation techniques introduced in (Xu et al., 2023) to create a smaller, more efficient retriever (INDUS-RETRIEVER$_{\text{SMALL}}$) through the supervision of INDUS-RETRIEVER$_{\text{BASE}}$. Specifically, for a sentence x_i and its corresponding in-batch element pairs $\{x_i, x_j\}_{j=1,j\neq i}^m$, we minimized the cross entropy between the teacher's distribution p_t of similarity scores between pairs and the student's distribution, p_s. Following Hinton et al. (2014), we also scaled the output distribution of both teacher and student by a temperature, τ_{KD}:

$$\mathcal{L}_{KD} = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_t(x_i,x_j)\log p_s(x_i,x_j) \tag{4}$$

$$p_s(x_i,x_j) = \frac{e^{s_s(x_i,x_j)/\tau_{KD}}}{\sum_{k=1}^m e^{s_s(x_i,x_k)/\tau_{KD}}} \tag{5}$$

$$p_t(x_i,x_j) = \frac{e^{s_t(x_i,x_j)/\tau_{KD}}}{\sum_{k=1}^m e^{s_t(x_i,x_k)/\tau_{KD}}} \tag{6}$$

Here, $s_s(x_i,x_j)$ and $s_t(x_i,x_j)$ represent the similarity scores between two pairs $\{x_i,x_j\}$, defined in Equation 3 for the student and teacher respectively. Note, τ_{KD} is the distillation temperature and is unrelated to the distance-temperature τ defined in Equation 3.

For the Retro-MAE style pretraining (Xiao et al., 2022), we trained on 8 A100 GPUs with an effective batch size of 128 for 2 epochs with a learning rate of $2e-5$. For the stage-wise distillation, we trained on 2 A100 GPUs for 300K steps with an effective batch size of 2,048, and learning rate of $7e-4$. Through experimentation, we found that $\tau_{KD} = 4$ performed the best, and we keep $\tau = 0.02$ as in the non-distilled case.

D Size of Proposed Benchmarks

The aim of our benchmark is to measure performance of models on three important yet orthogonal natural language understanding tasks, namely Named Entity Recognition, Extractive Question Answering and Information Retrieval. Each task further focuses on a different subset of domains of interest, specifically including those which are not covered by existing tests.

Dataset	Num. Pairs	Data Category	Data Format
StackOverflow†	18562443	Title-Body	s2p
StackExchange Math†	2201906	Title-Body	s2p
S2ORC [title - abstract] (Lo et al., 2020)	41769185	Title-Body	s2p
S2ORC Citation Pairs [Abstracts] (Lo et al., 2020)	52603982	Title-Body	p2p
StackExchange [title - body]†	5415570	Title-Body	s2p
Wikipedia (Fader et al., 2014)	6458670	Title-Body	s2p
Arxiv (Clement et al., 2019)	2358545	Title-Body	s2p
NASA ADS [title - abstract] (§2)	2633240	Title-Body	s2p
PubMed [title - abstract] (§2)	24001387	Title-Body	s2p
PMC [title - abstract] (§2)	2585537	Title-Body	s2p
StackExchange Duplicate Questions [title-body - title-body]†	250460	Duplicate Questions	p2p
StackExchange Duplicate Questions [body - body]†	250519	Duplicate Questions	p2p
StackExchange Duplicate Questions [title - title]†	304525	Duplicate Questions	s2s
WikiAnswer Pairs (Fader et al., 2014)	77427422	Duplicate Questions	s2s
Specter Pairs (Cohan et al., 2020)	684100	Citation Pairs	s2s
S2ORC Citation Pairs [Titles] (Lo et al., 2020)	52603982	Citation Pairs	s2s
SQuAD (Rajpurkar et al., 2016)	87599	Question Answers	s2p
NQ (Kwiatkowski et al., 2019)	100231	Question Answers	s2p
SearchQA (Dunn et al., 2017)	582261	Question Answers	s2p
StackExchange [title - answer]†	4067139	Question Answers	s2p
StackExchange [title-body - answer]†	187195	Question Answers	p2p
PAQ (Lewis et al., 2021)	64371441	Question Answers	s2p
FEVER (Thorne et al., 2018)*	109810	Fact Verification	s2p
HotpotQA (Yang et al., 2018)*	85000	Question Answering	s2p

Table 9: Training Data for Embedding Models. The training data totals to around 360M pairs. Data Format denotes s2p for sentence-to-paragraph mappings, s2s for sentence-to-sentence mappings, and p2p for paragraph-to-paragraph mappings. †Downloaded from https://huggingface.co/datasets/flax-sentence-embeddings/stackexchange_xml. *Only used for Distillation.

Moreover, we believe the size of each dataset to be comparable to other widely used domain-specific test sets in IR (eg. num. queries in BioASQ (Tsatsaronis et al., 2015), FiQA (Maia et al., 2018), DBPedia (Hasibi et al., 2017) and SciFact (Wadden et al., 2020) tasks from BEIR), QA (eg. num. questions in BioASQ (Nentidis et al., 2020) from BLURB), and NER (eg. num. entities in NCBI-disease (Doğan et al., 2014), BC5-Chem (Li et al., 2016), BC5-Disease (Li et al., 2016) from BLURB). We hope that the introduction of these datasets will serve as a much needed first step towards advancing benchmarking capabilities in this important field.

E Ablation Study: Stage-wise Distillation for Embedding Model

For the distilled embedding models, we find that stage-wise distillation does not benefit performance as much as a one-step process, combining all the supervised and unsupervised data. As shown in Table 10, the stage-wise approach underperformed the one-stage approach by 1 percentage point for both NASA-IR and on BEIR.

Model	Training	NASA-IR	BEIR Avg.
INDUS-RETRIEVER$_{\text{SMALL}}$	One-Stage	0.73	0.42
INDUS-RETRIEVER$_{\text{SMALL}}$	Stagewise	0.72	0.41

Table 10: Ablation Study: Evaluation results on NASA-IR and BEIR. NASA-IR showed Recall10 while BEIR reported nDCG10.

F Complete Results on BEIR Benchmark

Table 11 shows the per-dataset results on the BEIR tasks.

G Applications of INDUS for Retrieval: Performance and Interface

We show the interface for the Science Discovery Engine, the information retrieval system built with INDUS$_{\text{BASE}}$ in Figure 4, showing retrieved documents relevant to the seach query along with snippets with pertinant information.

Table 12 and Table 13 contain a few sample queries created for benchmarking by a human evaluator to compare the performance of the knowledge retrieval system leveraging INDUS$_{\text{BASE}}$ finetuned

Model	BEIR Eval												
	TREC-Covid	NFCorpus	NQ	HotPotQA	FiQA	ArguaAna	Touche	DBPedia	Scidocs	FEVER	Climate FEVER	SciFact	AVG. BEIR
RoBERTa$_{\text{BASE}}$	0.47	0.30	0.54	0.34	0.38	0.52	0.18	0.25	0.22	0.46	0.14	0.67	0.37
BGE$_{\text{BASE}}$	0.78	0.37	0.54	0.73	0.41	0.64	0.26	0.41	0.22	0.86	0.31	0.74	0.52
INDUS-RETRIEVER$_{\text{BASE}}$	0.56	0.32	0.54	0.49	0.36	0.54	0.17	0.31	0.21	0.56	0.14	0.74	0.41
MINILM-V2	0.47	0.32	0.44	0.47	0.35	0.50	0.17	0.32	0.22	0.52	0.25	0.65	0.39
BGE$_{\text{SMALL}}$	0.76	0.34	0.50	0.70	0.40	0.60	0.26	0.40	0.21	0.87	0.32	0.71	0.51
INDUS-RETRIEVER$_{\text{SMALL}}$	0.55	0.31	0.53	0.48	0.29	0.50	0.21	0.33	0.23	0.61	0.23	0.71	0.42

Table 11: Evaluation results BEIR.

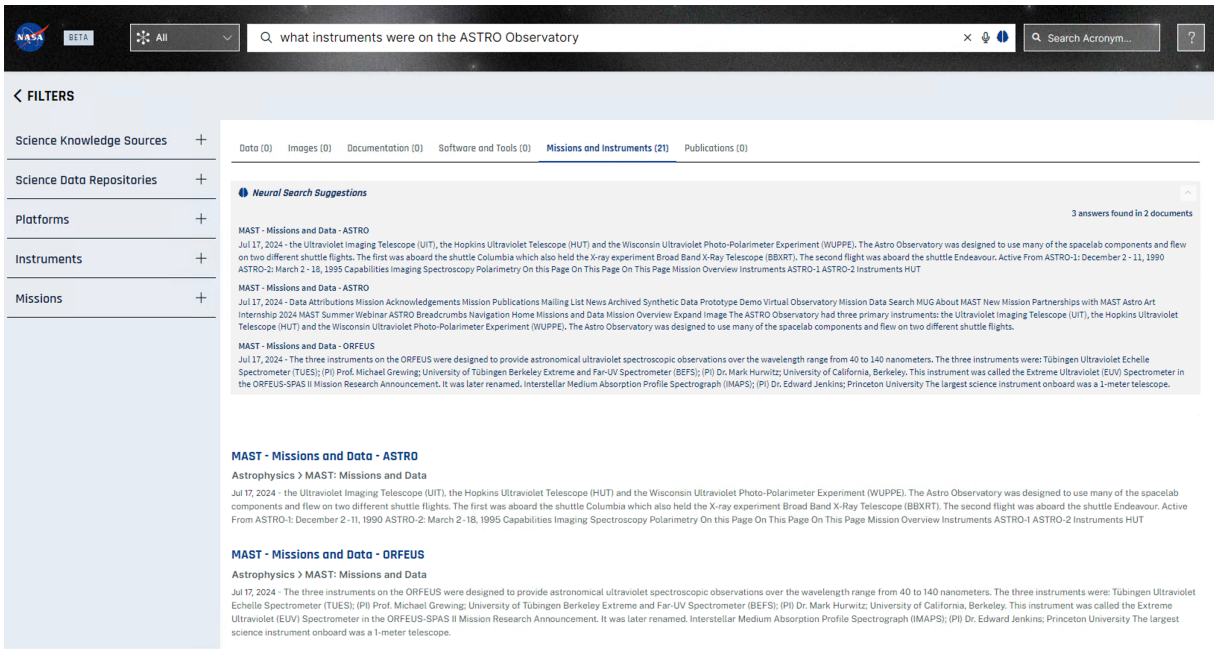

Figure 4: Interface to the Information Retrieval System built with INDUS. A user searches for a query and obtains snippets extracted from the document that contain relevant information, along with a list of relevant documents from which these snippets are extracted (screenshot edited to protect anonymity).

models with the one using generic RoBERTa$_{\text{BASE}}$ model. As shown, INDUS$_{\text{BASE}}$ usually provides a higher document and answer quality.

Question	Document Title	Retrieved Document Rank	Retrieved Document	Answer Quality Score
What does MODIS measure?	The MODIS Near-IR Water Vapor Algorithm	3	MODIS is a major facility instrument on the EOS polar orbiting satellite platforms (Asrar and Greenstone, 1995; King et al., 1992; Salomonson et al., 1989) designed to measure biological and physical processes on a global scale every 1 to 2 days. It is a 36-channel scanning radiometer covering the spectral region 0.4 - 15 μm. Five near-IR MODIS channels are useful for remote sensing of water vapor.	0.5
Which algorithm document describes the ZAVG product?	CERES ATBD Subsystem 8.0 - Monthly Regional, Zonal, and Global	1	Compute Regional, Zonal and Global Averages (Subsystem 8.0) This appendix describes the data products which are produced by the algorithms in this subsystem. The table below summarizes these products, listing the CERES and EOSDIS product codes or abbreviations, a short product name, the product type, the production frequency, and volume estimates for each individual product as well as a complete data month of production. The product types are defined as follows: Archival products:	0.5
Where did Perseverance land on Mars?	None	No relevant document retrieved	Perseverance's First Autonav Drive This image was taken during the first drive of NASA's Perseverance rover on Mars on March 4, 2021. Perseverance landed on Feb. 18, 2021, and the team has been spending the weeks since landing check... Perseverance Is Roving on Mars This map shows where NASA's Perseverance Mars rover will be dropping 10 samples that a future mission could pick up. A Map of Perseverance's Depot Samples This image is an edited version of the last 360-degree panorama taken by the Opportunity rover's Pancam from May 13 through June 10, 2018.	0.0
At what point in space is the JWST located?	#JwstArt Juried Art Show	1	Lines depict the direction of the waves reaching the telescope's instruments. Heat waves depicted highlight the temperature difference between the two sides of the solar shield. In order to analyze infrared light, the JWST needs to operate at 50 Kelvin (-223C/-370F) because the heat from the sun can interfere with the data entering the instruments. The bottom portion shows the relative location of the telescope after launch just outside earth umbra at the L2 Point about 1.5 million km from Earth.	1.0
What is the data policy for JWST?	Quick Start Guide - MAST Docs - STScI Outerspace	No Relevant Document Retrieved	No Answer Found	0.0

Table 12: Sample Questions from Human Evaluation of Vector Search pipeline leveraging $\text{RoBERTa}_{\text{BASE}}$ model.

Question	Document Title	Retrieved Document Rank	Retrieved Document	Answer Quality Score
What does MODIS measure?	DRAFT OF THE MODIS LEVEL 1B ATBD version 2.0 (ATBMOD - 01)	1	The MODIS raw output is a small, rapidly varying signal superimposed on a large background that varies more slowly, due to the thermal drifts and $1/f$ noise. Like its predecessor instruments, MODIS views space as its background subtraction reference and a full-aperture blackbody as its second reference for calibration. MODIS measures space and blackbody reference before and after each Earth view scan line. If $1/f$ noise is known at the time MODIS is viewing the space and blackbody reference then $1/f$ noise in the Earth view sector can be interpolated between four known	1.0
Which algorithm document describes the ZAVG product?	CERES ATBD Subsystem 8.0 Monthly Regional, Zonal, and Global	1	Monthly Zonal and Global Radiative Fluxes and Clouds (ZAVG). The Monthly Zonal and Global Radiative Fluxes and Clouds (ZAVG) product is a summary of the zonal and global averages of the radiative fluxes and cloud properties, probably most suitable for inclusion in the Earth. Observing System Data and Information System (EOSDIS) Information Management System (IMS) as a browse product. This product is the CERES equivalent to the zonal averages and global averages in the ERBE S-4 product. ZAVG is an archival product produced by the TISA subsystem for each instrument and for each combination of instruments.	0.5
Where did Perseverance land on Mars?	Sample Tube 266 - NASA Mars Exploration	1	Perseverance will land at the Red Planet's Jezero Crater a little after 3:40 p.m. EST (12:40 p.m. PST) on Feb... Perseverance on Mars NASA's Perseverance Mars rover is using its self-driving capabilities as it treks across Jezero Crater seeking signs of ancient life and gathering rock and soil samples for planned return to Earth. How Perseverance Drives on Mars This high-resolution image shows one of the six wheels aboard NASA's Perseverance Mars rover, which landed on Feb.18, 2021. The image was taken by one of Perseverance's color Hazard Cameras	1.0
At what point in space is the JWST located?	#JwstArt Juried Art Show Webb/NASA	3	None	1.0
What is the data policy for JWST?	Solar System Observation FAQ For Scientists Webb/NASA	1	The JWST Science & Operations Center will be located at the Space Telescope Science Institute (STScI) in Baltimore, MD. Competition will be fierce! What is my proprietary time? The baseline period for exclusive access to your JWST data is one year, as for HST and other missions. Some types of programs will have a shorter or zero exclusive access period. Proposers can also voluntarily reduce or waive their proprietary data rights. After the end of the exclusive access period the observations will be available for archival research.	1.0

Table 13: Sample Questions from Human Evaluation of Vector Search pipeline leveraging $\text{INDUS}_{\text{BASE}}$ model