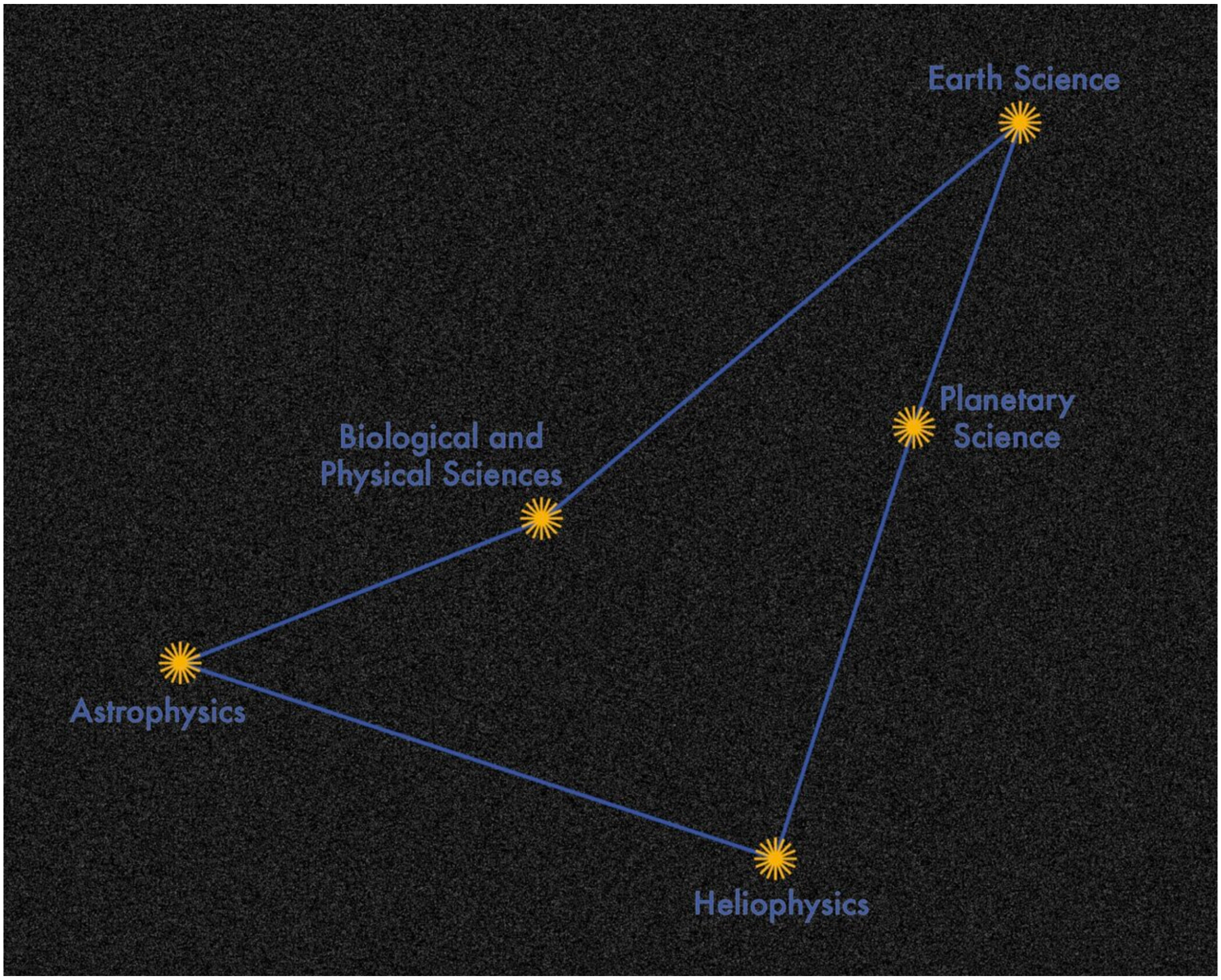
TechXplore

NASA-IBM collaboration develops INDUS large language models for advanced science research

June 25 2024, by Derek Koehl



Named for the southern sky constellation, INDUS (stylized in all caps) is a comprehensive suite of large language models supporting five science domains. Credit: NASA

Collaborations with private, non-federal partners through Space Act Agreements are a key component in the work done by NASA's Interagency Implementation and Advanced Concepts Team (IMPACT). A collaboration with International Business Machines (IBM) has produced INDUS, a comprehensive suite of large language models (LLMs) tailored for the domains of Earth science, biological and physical sciences, heliophysics, planetary sciences, and astrophysics and trained using curated scientific corpora drawn from diverse data sources.

INDUS contains two types of models; encoders and sentence transformers. Encoders convert natural language text into numeric coding that can be processed by the LLM. The INDUS encoders were trained on a corpus of 60 billion tokens encompassing astrophysics, planetary science, Earth science, heliophysics, biological, and physical sciences data. Its custom tokenizer developed by the IMPACT-IBM collaborative team improves on generic tokenizers by recognizing scientific terms like biomarkers and phosphorylated.

Over half of the 50,000-word vocabulary contained in INDUS is unique to the specific scientific domains used for its training. The INDUS encoder models were used to fine tune the sentence transformer models on approximately 268 million text pairs, including titles/abstracts and questions/answers.

By providing INDUS with domain-specific vocabulary, the IMPACT-IBM team achieved superior performance over open, non-domain specific LLMs on a benchmark for biomedical tasks, a scientific question-answering benchmark, and Earth science entity recognition tests. By designing for diverse linguistic tasks and retrieval augmented generation, INDUS is able to process researcher questions, retrieve relevant documents, and generate answers to the questions. For latency sensitive applications, the team developed smaller, faster versions of both the encoder and sentence transformer models.

Validation tests demonstrate that INDUS excels in retrieving relevant passages from the science corpora in response to a NASA-curated test set of about 400 questions. IBM researcher Bishwaranjan Bhattacharjee commented on the overall approach, "We achieved superior performance by not only having a custom vocabulary but also a large specialized corpus for training the encoder model and a good training strategy. For the smaller, faster versions, we used neural architecture search to obtain a model architecture and knowledge distillation to train it with supervision of the larger model."

INDUS was also evaluated using data from NASA's Biological and Physical Sciences (BPS) Division. Dr. Sylvain Costes, the NASA BPS project manager for Open Science, discussed the benefits of incorporating INDUS, "Integrating INDUS with the Open Science Data Repository (OSDR) Application Programming Interface (API) enabled us to develop and trial a chatbot that offers more intuitive search capabilities for navigating individual datasets. We are currently exploring ways to improve OSDR's internal curation data system by leveraging INDUS to enhance our curation team's productivity and reduce the manual effort required daily."

At the NASA Goddard Earth Sciences Data and Information Services Center (GES-DISC), the INDUS model was fine-tuned using labeled data from domain experts to categorize publications specifically citing GES-DISC data into applied research areas.

According to NASA principal data scientist Dr. Armin Mehrabian, this fine-tuning "significantly improves the identification and retrieval of publications that reference GES-DISC datasets, which aims to improve the user journey in finding their required datasets." Furthermore, the

INDUS encoder models are integrated into the GES-DISC knowledge graph, supporting a variety of other projects, including the dataset recommendation system and GES-DISC GraphRAG.

Kaylin Bugbee, team lead of NASA's Science Discovery Engine (SDE), spoke to the benefit INDUS offers to existing applications, "Large language models are rapidly changing the search experience. The Science Discovery Engine, a unified, insightful search interface for all of NASA's open science data and information, has prototyped integrating INDUS into its search engine. Initial results have shown that INDUS improved the accuracy and relevancy of the returned results."

INDUS enhances scientific research by providing researchers with improved access to vast amounts of specialized knowledge. INDUS can understand complex scientific concepts and reveal new research directions based on existing data. It also enables researchers to extract relevant information from a wide array of sources, improving efficiency. Aligned with NASA and IBM's commitment to open and transparent artificial intelligence, the INDUS models are openly available on Hugging Face.

For the benefit of the scientific community, the team has released the developed models and will release the benchmark datasets that span named entity recognition for climate change, extractive QA for Earth science, and information retrieval for multiple domains. The INDUS encoder models are adaptable for science domain applications, and the INDUS retriever models support information retrieval in RAG applications.

The work is published on the *arXiv* preprint server.

More information: Bishwaranjan Bhattacharjee et al, INDUS: Effective and Efficient Language Models for Scientific Applications,

arXiv (2024). DOI: 10.48550/arxiv.2405.10725

Provided by NASA

Citation: NASA-IBM collaboration develops INDUS large language models for advanced science research (2024, June 25) retrieved 29 June 2024 from https://techxplore.com/news/2024-06-nasa-ibm-collaboration-indus-large.html