IPAC'25 - the 16th International Particle Accelerator Conferece

IPAC25
16th International Particle Accelerator Conference
TAIPEI, TAIWAN
1-6 June 2025
國家同步輻射研究中心
National Synchrotron Radiation Research Center

Contribution ID: **948** Contribution code: **THPM023** Type: **Poster Presentation**

Developing specialized text embedding models for accelerator physics

Thursday 5 June 2025 15:30 (2 hours)

The specialized terminology and complex concepts inherent in physics present significant challenges for Natural Language Processing (NLP), particularly when relying on general-purpose models. In this talk, I will discuss the development of physics-specific text embedding models designed to overcome these obstacles, beginning with PhysBERT—the first model pre-trained exclusively on a curated corpus of 1.2 million arXiv physics papers. Building upon this foundation, we turn our attention to accelerator physics, a subfield with even more intricate language and concepts. To effectively capture the nuances of this domain, we developed AccPhysBERT, a sentence embedding model fine-tuned specifically for accelerator physics literature. A key aspect of this development involved leveraging Large Language Models (LLMs) extensively to generate annotated training data, enabling AccPhysBERT to facilitate advanced NLP applications such as semantic paper-reviewer matching and integration into Retrieval-Augmented Generation systems.

Footnotes

Paper preparation format

LaTeX

Region represented

America

Funding Agency

Author: HELLERT, Thorsten (Lawrence Berkeley National Laboratory)

Co-authors: POLLASTRO, Andrea (Lawrence Berkeley National Laboratory); VENTURINI, Marco (Lawrence Berkeley National Laboratory)

Presenter: HELLERT, Thorsten (Lawrence Berkeley National Laboratory)

Session Classification: Thursday Poster Session

Track Classification: MC6: Beam Instrumentation and Controls,Feedback and Operational Aspects: MC6.D13 Machine Learning