Contribution ID: **177** Type: **Poster + Flashtalk**

Event Tokenization and Next-Token Prediction for Anomaly Detection at the LHC

Advances in Machine Learning, particularly Large Language Models (LLMs), enable more efficient interaction with complex datasets through tokenization and next-token prediction strategies. This talk presents and compares various approaches to structuring particle physics data as token sequences, allowing LLM-inspired models to learn event distributions and detect anomalies via next-token (or masked token) prediction. Trained only on background events, the model reconstructs expected physics processes. At inference, both background and signal events are processed, with reconstruction scores identifying deviations from learned patterns—flagging potential anomalies. This event tokenization strategy not only enables anomaly detection but also represents a potential new approach for training a foundation model at the LHC. The method is tested on simulated proton-proton collision data from the Dark Machines Collaboration and applied to a four-top-quark search, replicating ATLAS conditions during LHC Run 2 ($\sqrt{s} = 13$ TeV). Results are compared with other anomaly detection strategies.

AI keywords

anomaly detection; tokenization; Large-Language Model; transformers; next-token prediction

Primary author: VISIVE, Ambre (Nikhef - University of Amsterdam)

Co-authors: Dr NELLIST, Clara (Nikhef - University of Amsterdam); Mrs MOSKVITINA, Polina (Nikhef - Radboud University); Dr RUIZ DE AUSTRI, Roberto (Valencia University, IFIC); Dr CARON, Sascha (Nikhef - Radboud University)

Presenter: VISIVE, Ambre (Nikhef - University of Amsterdam)

Track Classification: Patterns & Anomalies