Prototyping Large Language Models for CMS Event Analysis in the HL-LHC Era

1. Introduction

The High-Luminosity LHC (HL-LHC) will increase the integrated luminosity delivered to the CMS detector by roughly an order of magnitude compared to Run-2 and Run-3, producing petabytes of data per year and introducing pile-up levels approaching 200 interactions per bunch crossing [1]. The CMS Phase-2 upgrade will feature a high-granularity tracker, extended coverage, and precision timing layers, resulting in significantly larger and more complex event topologies [2].

Track reconstruction is among the most computationally demanding components of CMS event processing. Current ML approaches, such as graph neural networks (GNNs), are effective but require explicit graph construction and can be difficult to scale efficiently to full Phase-2 conditions [3]. Recent work (LM4Tracking, ACAT 2024) has demonstrated that transformer architectures—originally developed for language modeling—can successfully reconstruct tracks in simplified pixel-only geometries with low latency [4]. However, transformer-based tracking has not yet been demonstrated at full-detector scale in realistic Phase-2 conditions.

At the same time, new developments in efficient transformer design for scientific point-cloud data, such as Locality-Sensitive Hashing (LSH)-based attention and sparse neighborhood kernels [6, 7], have shown that it is possible to reduce attention complexity to near-linear while maintaining accuracy.

We propose to investigate transformer-based tracking for full-detector Phase-2 CMS events using doublets—geometrically consistent pairs of hits from adjacent detector layers—already defined in the CMS reconstruction workflow. Doublets provide a physics-motivated sequence compression, reducing the number of tokens from thousands of hits to hundreds of doublets. While this compression mitigates input size, it does not by itself solve the $O(N^2)$ cost of transformer attention. Therefore, we will explore the combination of doublet tokenization with efficient attention mechanisms (e.g. LSH) as a promising path to scalable transformer tracking at HL-LHC conditions.

2. Objectives

- Leverage CMS Phase-2 doublets as the model's tokens, embedding spatial and curvature features into each token.

- Pretrain a compact (~30–50 M parameter) transformer using a masked-token prediction objective (MLM-style), teaching the model the geometric and

kinematic structure of true tracks versus combinatorial noise.

- Fine-tune for track reconstruction in full-detector, high pile-up (200) environments.

- Benchmark against CMS GNN-based tracking in terms of efficiency, fake rate, inference speed, and scalability.

3. Methodology

Phase 1 – CMS Phase-2 Prototyping

- Data preparation: Use doublets directly from CMS Phase-2 simulated data. Since doublets are already produced in the reconstruction chain (via PC modules in the outer tracker), no custom doublet-building step is required. This ensures realistic coverage across barrel and endcap regions while staying aligned with CMS workflows.

- Pretraining: Apply BERT-style masked-token prediction. A fraction of doublet tokens will be masked, and the model will learn to reconstruct them, capturing geometric and kinematic correlations of tracks.

- Fine-tuning: Train the model for track linking and full track reconstruction in high pile-up conditions.

Phase 2 – Feature Engineering and Scaling

- Feature set: Doublets will embed spatial and curvature features. Tracker hits themselves do not provide per-hit timing; timing information is supplied by the MIP Timing Detector (MTD) at the track level [2]. Timing can be considered later as an auxiliary feature, but it is not part of the raw doublet tokens.

- Efficient attention: While doublets reduce sequence length, they do not remove the quadratic scaling of attention. To address this, we will incorporate algorithmic sparsification strategies such as LSH or sparse kernels [6, 7]. The combined approach—physics-driven compression plus efficient attention—offers the best path to HL-LHC feasibility.

- Evaluation: Compare tracking efficiency, fake rate, and inference latency with CMS GNN tracker [3].

Model Details

- Transformer: ~30–50 M parameters with geometry-aware positional embeddings.
- Sequence compression: Doublets reduce the token count significantly (hits → doublets), but efficient attention remains necessary for Phase-2 scale.
- Extensions: Integration of LSH-based or sparse neighborhood attention for scalability.
- Compute requirements: 1–4 × A100/H100 GPUs (40–80 GB), 5–10 TB SSD/NVMe storage, 32–64 CPU cores for preprocessing.

4. Expected Outcomes

- A CMS Phase-2–ready doublet-based transformer pipeline, leveraging existing doublets in the reconstruction chain.
- A compact transformer model showing competitive performance with GNN-based tracking under high pile-up conditions.
- Quantitative benchmarks of efficiency, fake rate, latency, and scalability relevant for both offline and potential online/trigger contexts.
- Documentation and recommendations for extending transformer-based methods to other Phase-2 reconstruction tasks and Large Physics Models [5].

References

1. CMS Collaboration, HL-LHC Computing and Data Challenges, CERN, 2025.
2. CMS Collaboration, The Phase-2 Upgrade of the CMS Tracker, CERN-LHCC-2017-009, CMS-TDR-014, 2017.
3. CMS Collaboration, Graph Neural Networks for Particle Tracking, JINST 16 (2021) P03017.
4. X. Ju, Y. Melkani et al., LM4Tracking: Large Models for Particle Tracking, ACAT 2024.
5. Lopez et al., Large Physics Models: Roadmap for Foundation Models in HEP, arXiv:2501.05382.
6. L. Huang et al., Locality-Sensitive Hashing-Based Efficient Point Transformer with Applications in High-Energy Physics, arXiv:2402.12535.

7. [arXiv:2402.10239] Efficient Transformer Variants for Large-Scale Scientific Data Processing.