Carlo A. Mussolini

ML Research Scientist
Carlomus@gmail.com

Education

University of Oxford, CERN - PhD, Particle and Accelerator Physics *2019 – 2023*
Thesis: Design of a Low-Energy Beamline in CERN's North Area for the NA61/SHINE Experiment

University of Oxford - MPhys, Theoretical and Particle Physics *2015 – 2019*
Graduated with First Class Degree

Professional Experience

ML Research Scientist
Fractile *Aug. 2023 – Present*

- Developed novel simulation and profiling tools in Python for large-scale model deployment on custom accelerators, enabling insights into performance
- Designed tooling for model performance evaluation and improvement across multiple use cases (reasoning, code), focusing on pre-training, fine-tuning, and distributed deployment
- Pioneered research into exotic number formats, quantization methods, and generative AI optimizations to enhance LLMs; applied these findings to improve performance on bleeding edge AI hardware
- Supervised a Master's student work focusing on implementing novel approaches to sparsity which aim at targeting novel hardware architectures, leading to significantly lower FLOPs required to complete inference
- Contributed to a cross-functional project building a custom compiler stack for distributed ML operations on proprietary hardware
- Contributed to securing a £5M+ research grant and co-authored patents for hardware acceleration in machine learning applications

PhD Researcher
University of Oxford & CERN *Sept. 2019 – Oct. 2023*

- Led the design of a low-energy secondary beamline for CERN's North Area, integrating both research and engineering aspects and requiring extensive cross functional collaborations and meetings
- Built and optimized frameworks for particle physics simulations, and enhanced data pipeline efficiencies
- Published 10+ papers on experimental physics and AI applications in accelerator physics, presented at IPAC and NBI conferences
- Collaborated on a project using GNNs to perform predictive maintenance on various components of the LHC
- Created a modular ML framework, implementing forward/backward passes for foundational layers (e.g., Linear, Convolution, ReLU, SoftMax)

Researcher
University of Oxford *Sept. 2018 – Jun. 2019*

- Used C++ to analyze simulation and real-world data discrepancies in the SNO+ experiment, enhancing model reliability for particle physics research

Research Interests

- **Generative AI and LLM Fine-Tuning:** Focused on refining LLMs for diverse applications and exploring optimization techniques for pre-training
- **Efficient Model Deployment:** Interested in designing systems for scalable model deployment, using advanced quantization, and low-level programming
- **Cross-Modal AI Models:** Enthusiastic about developing models that span text, image, and audio for comprehensive multimodal applications

Technical Skills

Programming Languages:
Python, C, C++

ML Frameworks:
PyTorch, TensorFlow, CUDA, Hugging Face

Distributed Systems:
Kubernetes, Docker,
Remote Execution Environments

Research & Experiments:
Profiling, Interpretability, Quantization, Optimization

Supervisions and Teaching

Supervisions
Fractile - Master's thesis on sparsity and LLM optimization

CERN - Master's thesis on radiation protection optimization

Teaching Assistant
University of Oxford - Held tutorials on Fluid Mechanics for 3^{rd}-year physics students

Internships

University of Sao Paulo
Research on biofuels, analyzing socio-environmental impacts

Gran Sasso National Labs
XENON-nT experiment data collection for dark matter research

Hobbies and Interests

Rock Climbing, Chess, Skiing, Basketball, Cooking