Salvatore Petrolo

Munich, Bavaria, Germany toredev@outlook.it +39 3887808493 t0re199.github.io
linkedin.com/in/t0re199 github.com/t0re199

About

Machine Learning Engineer with a strong software engineering foundation, specializing in developing, fine-tuning, and deploying Large Language Models (**LLMs**) at scale for edge and cloud environments. Skilled in pre-training, supervised fine-tuning (**SFT**), and Reinforcement Learning from Human Feedback (**RLHF**) to optimize model performance and relevance. Expertise in transformer-based architectures, Retrieval-Augmented Generation (**RAG**), and **Computer Vision**.

Experience

Machine Learning Engineer *Tutzing, Bavaria, DE*
Knowlix GmbH *February 2023 – Present*

- **Improved documents key information extraction accuracy to 99%** by fine-tuning a Vision-Language Model (**VLM**) to eliminate OCR dependency and enhance document understanding.
- **Optimized cost, performance, and security of document intelligence workflows** by developing a serverless API with **AWS**, integrating a model zoo of **LLMs** on SageMaker, and implementing asynchronous job execution with callback and polling.
- **Achieved 4x speed-up and 98% accuracy in on-device key information extraction** by fine-tuning a Small Language Model (**SLM**) and optimizing it for Apple mobile deployment using **CoreML**.
- **Increased document summarization relevance and accuracy by 25%** by fine-tuning a Large Language Model (**LLM**) with Direct Preference Optimization (**DPO**) to align summaries with user preferences and domain-specific requirements.
- **Enhanced multilingual classification accuracy by 10%** by conducting transformer encoder pre-training for improved downstream performance.
- **Enabled precise document interaction via natural language queries** by building a Retrieval-Augmented Generation (**RAG**) system and integrating a vector search engine for efficient retrieval.

Education

University of Calabria *Sept 2020 – July 2022*
Master of Science in Artificial Intelligence and Machine Learning

- **Score:** 110/110 cum Laude.
- **Thesis:** Deep Anomaly Detection in ECG Signals to Detect Arrhythmia.

University of Calabria *Sept 2016 – Sept 2020*
Bachelor of Science in Computer Engineering

- **Score:** 110/110 cum Laude.
- **Thesis:** Implementation of a Language for developing dynamic data collection web applications.

Skills

Programming Languages: Python, Java, Swift, C, C++, CUDA, Javascript, Assembly.

Machine Learning: PyTorch, TensorFlow, Transformers, vLlm, Llama.cpp, CoreML, Mlx, JAX, ONNX.

LLMs: Pre-Training, SFT, RLHF, LoRA, Key-Value Caching, Grouped Query Attention, Quantisation.

Tools: Git, Jira, CI/CD, Jupyter, FastAPI, Flask, AWS, Azure, GCP, MLOps.

Languages: Italian (Native), English (Advanced), German (Intermediate)

Awards

- Selected as **the best student** of the Master's degree program in Computer Engineering at the **University of Calabria**.