# 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, Terraform.

**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**.

## Academic Projects

- **Artificial Intelligence & Knowledge Representation and Reasoning:** Developed a Java-based automatic player for the Murus Gallicus game. Implemented a parallel version of the well-known **MiniMax** algorithm with **Alpha-Beta pruning** to optimize game-playing performance.
- **Machine and Deep Learning:** Applied advanced Machine and Deep Learning techniques to process image and text datasets. Focused on two main tasks: **Multi-Class Classification** and **Anomaly Detection**, using **TensorFlow** as the primary gradient computing framework.
- **Images and Videos Analysis:** Implemented **Multi-Class** and **Multi-Label Classification** for an unbalanced film trailer dataset in Python, using deep learning architectures like **ResNet** and **VGG**. Developed a custom modular architecture for image classification tasks with **PyTorch**.
- **GPGPU Programming:** Implemented a **CUDA** C parallel version of the Merge operation, achieving a 66x speedup over a serial implementation by optimizing memory access patterns and exploiting different types of GPU memory.
- **Data Mining:** Performed multi-class classification of Android applications using **scikit-learn** to analyze a Kaggle dataset. Applied data preprocessing techniques including **SMOTE** oversampling, feature extraction, and model optimization through hyperparameter tuning.
- **Big Data Management:** Designed and implemented a Python query tool interfacing with **Apache Spark** for big data processing and **MongoDB** for NoSQL storage. Leveraged **PySpark** and **PyMongo** libraries for efficient data management and querying.
- **Social Networks and Media Analysis:** Developed a sentiment analysis model using pre-trained **transformer** models from **Hugging Face** to process an Amazon English reviews dataset. Achieved high accuracy in sentiment prediction.
- **Distributed Systems and Cloud Computing:** Created a distributed contact-tracing system using an **Android Application** and a **JavaEE** REST backend. The system employed **Bluetooth Low Energy (BLE)** technology for proximity detection and a custom REST client for backend interaction.
- **Architectures and Programming of Processing Systems:** Optimized **Stochastic Gradient Descent** implementations using advanced programming concepts like **SIMD** parallelism and **Intel SSE/AVX** instruction sets, applied to **SoftSVM** and **Polynomial Kernel Method** algorithms.
- **Software Platforms for Web Applications:** Developed the **ShopCart** web application, including a frontend **Angular SPA** and backend REST services implemented using **JavaEE** stack with **JPA** and **Hibernate** for persistence. Utilized **PostgreSQL** as the database.