ARYA RAEESI

linkedin.com/in/arya-raeesi/ | +47 903 68 159 | Bergen, Norway
github.com/raeesiarya | raeesiarya@gmail.com | Citizenship: Norway

SUMMARY

M.Sc. student in Electronic Systems Design and Innovation at NTNU, including a one-year graduate exchange at UC Berkeley EECS. Industry experience from Sony Semiconductor Solutions and research experience from the NASA Small Explorers (SMEX) Program and NTNU SmallSat Lab. Current research in machine learning systems and language models, building on a foundation in hardware and embedded design.

EXPERIENCE

NASA Small Explorers (SMEX) Program / UC Berkeley Space Sciences Laboratory - Graduate Researcher - Berkeley, California, United States of America & Remote — 08/25 - Present

Conducting research on gamma-ray astrophysics using high-purity germanium detector data as part of the COSI (Compton Spectrometer and Imager) mission. Built a Naive Bayes classifier from scratch, including feature extraction, density-matrix likelihood, Bayes-rule decision, and evaluation pipeline. Exploring potential applications of neural networks for similar problems. Part of an international NASA project led by the UC Berkeley Space Sciences Laboratory in partnership with Northrop Grumman, SpaceX, the U.S. Naval Research Laboratory, etc.

Sony Europe - Analogue Design Engineering Intern - Oslo, Norway — 06/25 - 08/25

Designed and implemented an automated test software system for automotive image sensors, focusing on efficient data storage, repeatable test procedures, and system-level reliability. The system enables consistent sensor evaluation across revisions and supports scalable testing for future development cycles. Data output is structured in alignment with industry standards such as EMVA 1288, facilitating further analysis and integration.

NTNU SmallSat Lab - Student Guest Researcher (Project-Based) - Trondheim, Norway — 01/25 - 06/25

Conducted research on developing and training convolutional neural networks for sea, land, and cloud classification of hyperspectral satellite images from the HYPSO-2 mission. Utilized NVIDIA GPU cluster acceleration to optimize performance and training speed.

BlinkWeb - Founder - Bergen & Trondheim, Norway — 06/22 - 12/24

Founded and led a web and IT solutions company, delivering solutions for 10+ Norwegian businesses across construction, service, and research domains. Designed and implemented in-house automation and data pipelines that reduced manual workload and increased profitability.

EDUCATION

University of California, Berkeley (UC Berkeley) - Exchange Year (Graduate level), Electrical Engineering and Computer Science — 08/25 - 05/26

Nominated for a graduate exchange year at UC Berkeley, focusing on graduate/PhD courses in embedded systems, machine learning, and digital design to deepen technical expertise.

Norwegian University of Science and Technology (NTNU) - M.Sc. Electronic Systems Design and Innovation, Specialization in Design of Digital and Embedded Systems — 08/22 - 06/27

A multidisciplinary program combining electronics, software development, mathematics, and physics to design and innovate advanced electronic systems. My specialization in design of digital and embedded systems has given me deeper experience in designing integrated circuits, utilizing high-level languages for programming and constructing across hardware and software.

PROJECTS

Beyond Binary Priorities: Multi-Tier SLA Scheduling for Large Language Model Serving (UC Berkeley) — 09/25 - 12/25

Designed a multi-priority extension to Llumnix (OSDI 2024, Alibaba Group), enabling fine-grained SLA differentiation for multi-tenant LLM inference while preserving low tail latency and high GPU utilization. Reimplemented the architecture in Microsoft Research's Vidur simulator and evaluated latency, throughput, and cost efficiency through large-scale simulations. Achieved up to ~3× improvement in P99 latency, with diminishing returns beyond four priority tiers.

Auditing Forgetting in Limited Memory Language Models (UC Berkeley) — 02/26 - 05/26

Built a causal evaluation framework for memory separation in Limited Memory Language Models (Zhao et al., 2025), decomposing post-deletion correctness into parametric leakage, retrieval-mediated correctness, and retrieval artifacts. Ran 1404 datapoints per evaluation across six prompt formulations and thirteen database variants, orchestrated via Slurm and Weights & Biases. Found that retrieval artifacts dominate residual correctness after deletion, not parametric leakage. Developed with research feedback from NLP researchers at Stanford, Cornell, and UC Berkeley.

Four Stages, Two ISAs: A Pipelined RV32IF Core on PYNQ-Z1 FPGA (UC Berkeley) — 09/25 - 12/25

Developed a pipelined RISC-V SoC on the Digilent PYNQ-Z1 FPGA, implementing an RV32I core with CSR support and a pipelined RV32F floating-point unit. Designed a four-stage in-order pipeline with hazard detection, data forwarding, and precise control-flow handling, addressing end-to-end hardware–software co-design. Verified the system using RISC-V ISA tests and C benchmarks, and executed end-to-end workloads directly on FPGA hardware. Achieved 58 MHz, ~1.16 integer CPI, ~1.83 FP CPI, and an FOM of 12.3.Developed within a UC Berkeley hardware course supported by Apple's New Silicon Initiative and co-taught with industry researchers from NVIDIA.

End-to-End Client Acquisition System (BlinkWeb) — 12/22 - 02/23

Engineered a custom CRM and client acquisition engine using REST APIs, integrating the Brønnøysund Register and multiple phone registries (1881, 180, Gule Sider) for automated data enrichment and contact verification. Designed geospatial lead-matching algorithms to reduce false positives and built automated outreach and real-time contract generation workflows. The system saved ~450 hours/year, reduced third-party API costs by ~$10,000, and served as core infrastructure behind BlinkWeb's 1130% profit margin in 2024.

SKILLS ELECTRONICS & HARDWARE

Integrated circuits (Verilog HDL, Spice programming); Semiconductor physics (MOSFET and BJT physics, carrier transport physics, semiconductor quantum physics); Analogue circuit design; FPGA implementation of RTL design (Xilinx Vivado); Physical oscilloscope-, spectrum- and network analysis; Automotive image-sensor technology; Image-sensor testing; Sensor noise reduction; Multi-dimensional digital signal processing (Fourier analysis, feature extraction, pre-processing and filtering)

SKILLS SOFTWARE ENGINEERING

Algorithms and data structures; Deep learning (Python, PyTorch, CUDA, NVIDIA-GPU acceleration); Object-oriented programming (C++, C, Java); Databases (MySQL, PostgreSQL, Firebase); Concurrent and distributed systems; REST APIs; RISC-V Assembly; Natural Language Processing (transformers, word representations, sequence models, fine-tuning, RAG, evaluation); LLM inference and serving systems; ML experiment orchestration (Slurm, Weights & Biases); Git (GitHub, GitLab)