Bella Nicholson

San Francisco Bay Area, California, USA · bellanich.github.io · in bella-nicholson · bellanich · American

A resourceful Machine Learning Engineer (3+ years) adept at engineering robust software systems for ML research and deployment. A life-long learner who's well-versed in LLM inference and edge deployment, now keenly focusing on hardware-aware software design. Excited about the intersection of cutting-edge ML research, software engineering, and hardware-aware design.

EXPERIENCE

Machine Learning Software Engineer // **Netflix** // May 2025 – Present — Los Gatos, California, USA

For an iconic streaming entertainment service, I supported ML researchers in foundation models development for improved recommendations:

- Engineered and maintained the underlying robust software systems and infrastructure required for foundation model development
- Collaborated closely with researchers to implement, optimize, and scale novel foundation model architectures

Machine Learning Engineer // **Brenntag** // Nov 2023 – Apr 2025 — Amsterdam, Netherlands

At the world's leading chemical distributor, I deployed and maintained ML products across 70+ countries:

- Migrated a €30M+ annual revenue AI assistant to a **more cost-effective and secure AWS** platform, reducing operational costs by €50k+/month
- Developed a **real-time notification system** to monitor critical ML jobs and model metrics, **improving system visibility** and reliability
- Validated and refined the new company ML Platform design in a close collaboration with cross-team data and cloud engineers

Machine Learning Consultant // **Deloitte** // Sept 2021 – Oct 2023 — Amsterdam, Netherlands

At a global consulting firm, I delivered and optimized production-ready ML solutions for diverse clients. Achievements include:

- **Stabilized** a Dutch e-classified ads platform's "For You" **recommendation engine** (2000+ lines of code) with a 65% increase in test coverage
- Centralized tracking of **1000+ models** and associated experiments for a German steel conglomerate, improving **model reproducibility**
- Launched a self-paced, ML-focused coding training website to standardize and improve code quality across Deloitte NL

Computer Vision Intern // **Cubelizer** // June 2017 – July 2017 — Madrid, Spain

As part of a Google-backed **edge computer vision** startup, I improved customer detection by 12% for retail space price optimization:

- Developed a video stream-based **object detection** method in compliance with EU privacy regulations
- Applied image processing and classical machine learning techniques to low-resolution images

PROJECTS

Pocket Multi-Modal Large Language Model

- Deployed a **custom embedded, vision-text foundation model** and Google's Gemma 2B model on various **edge devices (laptop, phone, tablet)**
- Extended an open source LLM hardware-optimization framework to quantize and optimize a new multi-modal LLaVA foundation model
- Documented the project implementation, including application solution prototyping, in a detailed 4-part blog post series

Transformers Decoded: A Guide to Optimizing Large Language Models

- Developed a comprehensive study guide on Large Language Models (LLMs), explaining underlying concepts and modern optimization strategies
- Covered LLM inference **optimization techniques (speculative decoding, flash attention, continuous batching; etc.)** for efficient deployment

EDUCATION

Master of Science, Artificial Intelligence

University of Amsterdam · Cum laude (8.0/10.0) · Sept 2018 – Dec 2020 · Amsterdam, Netherlands

- Courses on AI, including Deep Learning, Computer Vision, Natural Language Processing, Information Retrieval, and Reinforcement Learning
- Thesis on "Interpretable Representation Learning for Relational Data" in collaboration with Crunchr

Bachelor of Science, Biomedical Engineering

The College of New Jersey · Magna cum laude (3.8/4.0) · Sept 2014 – May 2018 · Ewing, New Jersey, USA

SKILLS

Programming Languages & Tooling

Python, Git (2016-present), Bash (2018-present), Terraform (2023-present), SQL (2019-present), Docker (2022-present)

MLOps Platforms

Amazon Web Services (2021-present), Google Cloud Platform (2022-present), Databricks (2022-present)

ML Frameworks

PyTorch (2018-present), PySpark (2022-present), FastAPI (2021-present), Tensorflow (2021-2022), Keras (2022)

English 100% · Spanish 75% · German 60% · Russian 50%