Muhammad Taha Mukhtar

tahamukhtar20@gmail.com | LinkedIn | GitHub | Google Scholar

Summary

Machine learning engineer and researcher working on representation learning, interpretability and RL environments for coding agents. Visiting Researcher at DFKI since 2024, with 14,000+ GPU-hours of experiments on up to 115 GPUs. Also ships production systems, from generative AI backends at Vyro to AI agents for clients, and built mapcv, an open-source remote-sensing dataset library with 10,000+ downloads.

Experience

AI and Software Consultant June 2026 to Present
Multiple Contracts (Remote)

- Moved production document parsing from cloud APIs to a self-hosted MLX cluster, porting the Qianfan-VL OCR model and serving local LLMs, cutting the AI bill to near zero.
- Designed and shipped AI agents across production systems, from browser automation to document workflows, using Python, Temporal, Docker and Kubernetes on GCP.
- Performed a web application security audit of a production marketplace, remediating IDOR, broken access control and privilege escalation across its APIs and Stripe payment flows.
- Compressed a foundation model into a small edge model through pruning, retraining and quantization, deployed on Raspberry Pi for dashcams.

Visiting Researcher June 2024 to Present
German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany (Hybrid)

- Studied what representations encode and how to measure it.
- Post-trained a medical LLM for local deployment using PPO, DPO, RLHF, RLVR, PEFT, LoRA and quantization.
- Developed a semi-supervised framework for mapping informal settlements from satellite imagery, with a freshly annotated dataset and a class-adaptive pseudo-labeling method (arXiv:2602.04525).
- Researched self-supervised, semi-supervised and generative learning across SimCLR, MoCo, DINO, GANs, FixMatch-style methods and JEPA, including noised-target JEPA variants and latent world models.
- Ran 14,000+ GPU-hours, up to 115 concurrent GPUs, on NVIDIA B200, H200, H100, A100 and RTX A6000.

Research Associate November 2025 to May 2026
National Center for Artificial Intelligence (NCAI), Islamabad, Pakistan (Hybrid)

- Worked on E-TCAV, a framework for efficiently approximating TCAV scores with penultimate-layer proxies, evaluated across different architectures and datasets (arXiv:2605.10261).
- Studied adversarial neural cellular automata, where independently trained networks compete on a shared grid, explaining their bistable dynamics through mean-field analysis and targeted ablations.
- As a BS graduate, co-supervised two MS theses, on periocular biometrics (student graduated) and NSCLC immunotherapy response (preparing for publication and thesis defense).

Backend Engineer January 2025 to May 2025
Vyro.ai, Islamabad, Pakistan (Onsite)

- Developed the ImagineArt developer platform (platform.imagine.art) and its public generative AI APIs.
- Built low-latency LLM and generative media streaming with goroutines for chatlyai.app and imagine.art.
- Refactored the LLM backend around dependency injection, and optimized ChatlyAI's slow Cassandra queries.

Freelancer April 2022 to January 2025
Fiverr and Direct Clients (Remote)

- Delivered 50+ Next.js, MERN and Python projects, including e-commerce stores, admin dashboards and client websites, reaching Level 2 seller status.
- Built and deployed full-stack client products with payments and cloud storage across Vercel, AWS EC2 and ECS, Cloud Run, Firebase and Hetzner.
- Sped up PostgreSQL by 25x through indexing on a system handling 10,000+ daily transactions, and helped mitigate a production DDoS attack with rate limiting and IP blocking.

Open-Source Contributor March 2024 to September 2024
CVAT, Google Summer of Code 2024 (Remote)

- Designed a configurable keymap system for CVAT's React and TypeScript annotation platform, with scope- and prefix-aware conflict detection and persistent user preferences (project page).

- Re-architected shortcut registration into modules across the annotation interface, with Cypress end-to-end tests and documentation.
- Contributed before the program by fixing warped frames from rotated videos and adding a shortcut for pinned properties, with 8 PRs merged in total (final report).

Software Engineer January 2024 to June 2024
Thinksoft (Remote)

- Co-developed WebbsAI (webbsai.com), a generative AI platform that turns prompts into complete websites.
- Built product features across the stack with TypeScript, Next.js, Zustand, Firebase, MongoDB and AWS S3, supporting concurrent generation workflows.

Software Engineer June 2023 to January 2024
OneScreen, Islamabad, Pakistan (Hybrid)

- Built the React frontend of an LLM presentation maker that generates complete slide decks from prompts.
- Prototyped LLM agents and RAG pipelines with Python and LangChain, using BERT embeddings, MongoDB vector search and reranking.
- Implemented Redis-backed leaky bucket rate limiting for LLM API calls.

Skills

Languages: Python, TypeScript, JavaScript, Go, SQL, Rust, Bash
Backend & Databases: FastAPI, NestJS, Express, Chi, Fiber, Temporal, PostgreSQL, MongoDB, Redis, Cassandra
Frontend & Web Performance: React, Next.js, Redux, Zustand, Tailwind CSS, Vite, Webpack, SSR, SEO, Cypress
AI Agents & RAG: Orchestration, Browser Agents, Tool Calling, MCP, Vector Search, Reranking, LangChain
Agent Evaluation: RL Environments, Verifiers, Reward Design, Reward-Hacking Defense, Gymnasium, PettingZoo
LLMs: Hugging Face, Vision LLMs, PPO, DPO, RLHF, RLVR, PEFT, LoRA, Quantization, MLX, llama.cpp
ML Libraries: PyTorch, scikit-learn, torchvision, Segmentation Models PyTorch, Weights & Biases, TensorBoard
ML Concepts: Attention, Contrastive Learning, Pruning, Mixed Precision, Interpretability
Infrastructure: Docker, Enroot, Kubernetes, Helm, GCP, AWS, GitHub Actions, Linux, SLURM

Projects

mapcv April 2026 to Present
Open-source library for fast and reliable remote-sensing dataset creation (GitHub)

- Turns imagery and labels into ready-to-train datasets for six computer vision tasks, from XYZ tiles, GeoTIFF/COG, Sentinel-2 and Earth Engine, with leakage-safe spatial splits and byte-identical resume.
- Wrote the rasterizer, decoders and writer in Rust with no GDAL dependency, running 1.5 to 34x faster than a rasterio script, GDAL and leafmap on the same output, using 84% less RAM than the script at 10,000 patches.
- Released on PyPI with cross-platform wheels and a Zenodo DOI, reaching 10,000+ downloads and 14 GitHub stars, and published an MCP server that lets coding agents plan and build datasets.

Faultline August 2026 to November 2026
Open-source RL environment that tests whether coding agents really fix distributed-systems bugs (GitHub)

- Injects seeded faults into a FastAPI order pipeline and grades repairs in Docker sandboxes by killing workers mid-payment, redelivering events and dropping provider responses, scored from a ledger agents cannot touch.
- Ships as Gymnasium, PettingZoo, Harbor, Inspect AI, Verifiers and OpenEnv environments over a five-level curriculum, with shaped and sparse rewards, checkpoint and restore, and replayable trajectories.
- Built reward-hacking tests that reject fake providers, faked fulfillment, skipped charges and hard-coded orders, and benchmarked Claude Opus and Haiku against reference, random and rule-based baselines.

Research

- **M. T. Mukhtar**, S. M. A. Kazmi, K. Naseem, M. A. Chattha, A. Dengel, S. Ahmed, M. N. Bajwa, M. I. Malik. "SLUM-i: Semi-supervised Learning for Urban Mapping of Informal Settlements and Data Quality Benchmarking." arXiv:2602.04525, 2026.
- H. Aslam, M. A. Chattha, **M. T. Mukhtar**, M. I. Malik, A. Dengel, S. Ahmed. "E-TCAV: Formalizing Penultimate Proxies for Efficient Concept Based Interpretability." arXiv:2605.10261, 2026.

Selections and Awards

- Google Summer of Code 2024, CVAT. Selected as 1 of 1,220 contributors from 43,984 registered applicants.
- Awarded DAAD funding for research internships at DFKI in 2024 and 2025, selected in 2024 as 1 of 2 from about 250 students in the computing department.

- SOFTEC, 2nd place nationwide among 64 teams, 24-hour software hackathon, 2024.
- Three High Achiever Medals, NUST, 2025, for national and international co-curricular achievements.

Education

National University of Sciences and Technology (NUST) November 2021 to June 2025
Bachelor of Science in Computer Science, Islamabad
Thesis at DFKI: Semi-Supervised Learning for Slum Segmentation using Remote Sensing and Deep Learning
Advisor: Dr. Muhammad Imran Malik
Co-advisor: Dr. Muhammad Naseer Bajwa
Relevant coursework: Reinforcement Learning, Artificial Intelligence, Digital Image Processing, Linear Algebra, Probability and Statistics, Differential Equations, Calculus