

Nauman Mustafa

Sr. ML & AI Agentic Engineer | LLMs | Agentic AI | MLOps

Tokyo, Japan | +81-90-3194-6885 | nauman.mustafa.x@gmail.com

linkedin.com/in/naxalpha | github.com/naxalpha | medium.com/@NaxAlpha

PROFILE

Sr. ML & AI Agentic Engineer with 8+ years of experience turning ambiguous product problems into shipped AI systems. Expertise spans LLMs, agentic workflows, evaluation, computer vision, OCR, and production MLOps, with a track record of leading projects from experimentation through deployment.

Selected impact ~60% AI Revise win rate | 14M+ visual-regression requests | >95% OCR recognition

PROFESSIONAL EXPERIENCE

Sr. ML & AI Agentic Engineer

LegalOn Technologies

Tokyo, Japan

Jul 2025 - Present

Professional AI for Legal

- Built AI Revise V3, achieving an approximate **60% win rate** across most evaluation datasets.
- Designed and implemented the end-to-end AI Revise evaluation system, from metric computation through an **LLM-based judge**.
- Contributed to Project Re: Invent, rebuilding the company's core product around an **agentic AI foundation**.
- Shipped the Workflow Engine for Re: Invent, enabling users to author complex workflows in simple, natural language.

Senior Machine Learning Engineer

Autify

Tokyo, Japan

May 2020 - Jun 2025

Test automation for mobile and web applications

- Led the ML team by setting technical direction and driving delivery across multiple machine learning projects.
- Deployed an optimized Visual Regression API pipeline serving **14M+ requests** at approximately \$1,200/month using Knative on GKE.
- Developed and trained an in-house transformer-based OCR model from scratch, reaching **>95% character recognition** across 30K+ Japanese and English characters.
- Built LLM-powered tools for test creation and test generation, producing tests, Gherkin, and Playwright code from feature PRDs.

EARLIER EXPERIENCE

Machine Learning Engineer

VisionX

Islamabad, Pakistan

Jun 2018 - May 2020

Machine learning and computer-vision services

- Implemented lyrics extraction with EAST and Tesseract for a music-sheet parsing program.
- Improved a proprietary information-retrieval algorithm for parcel images captured by cameras.
- Bootstrapped infrastructure for a SaaS product that extracted information from receipts.
- Trained BERT-based classifiers on datasets of 100K+ examples and used GPT-2 and T5 for unstructured-text parsing.
- Deployed large language models in production with Docker, Cloud Run, and Kubernetes.

Research Assistant (Internship)

DAAD - RheinMain University of Applied Sciences

Wiesbaden, Germany

Jun 2017 - Sep 2017

- Implemented a forward-backward motion tracker using OpenCV for fish tracking in camera images.
- Developed a UI for debugging and annotating the fish-tracking algorithm and its ground truth.

SELECTED PROJECTS

Long Pythia | Apr 2023

Extended the context length of EleutherAI's Pythia language model to 8,000 tokens.

Token Compression | Aug 2023

Explored compressing multiple LLM tokens into a single embedding vector.

EDUCATION

Bachelor of Engineering in Machine Learning

National University of Sciences and Technology

3.7/4.0

Sep 2014 - Jun 2018

TECHNICAL SKILLS

Languages Python

ML / AI PyTorch, Deep Learning, Transformers, LLMs, Prompting, LoRA, Diffusion Models, OCR, OpenCV

Infrastructure Docker, Google Cloud, GKE, Kubernetes, Knative, MLOps

LANGUAGES & INTERESTS

Languages Urdu - Native or Bilingual Proficiency; English - Full Professional Proficiency

Interests AI, Philosophy, Technology