HARSHVARDHAN CHOUDHARY

Machine Learning Engineer

+91 9329516048 harshvardhan.yashvardhan@gmail.com github.com/Harshvardhan-To1

linkedin.com/in/harshvardhan-choudhary

SUMMARY

ML Engineer at fal working across inference optimization, on-device AI, and customer-facing solutions for generative-media products. I take systems from idea to production and support large enterprise customers end to end.

EXPERIENCE

Machine Learning Engineer

fal

03/2025 - Present · Remote

- Build and ship production inference systems for enterprise customers including Adobe, Shopify, and Amazon, owning the work from research through deployment and meeting customer targets for latency, throughput, and reliability.
- Designed AI automations for support and operations workflows using MCP, cutting manual effort on repetitive tasks.
- Handle technical support and solutions for major customers such as Canva, Netflix, Epic Games, Artlist, Freepik, and World Labs, debugging production issues with Grafana and Datadog.
- Joined as employee no. 17 and contributed through fal's growth from an early-stage startup to a multi-billion-dollar company, as the team scaled from 17 to 140 people.

Machine Learning Intern

Styldod Inc.

08/2024 - 10/2024 · Remote

- Sped up text-to-image and image-to-image inference by 2.5x using DeepCache and OneDiff.
- Built a data collection pipeline with Scrapy, BeautifulSoup, and Selenium, producing a training dataset of 16,000+ images with automated scraping and quality filtering.

PROJECTS

Chitra (event photo finder)

2025 · chitra.cloud

- Designed and built an event-photo finder end to end from idea to production: users take a selfie and instantly get their photos from large event galleries through face matching.

Curio (in-browser website chat)

2025 · curio-umber-five.vercel.app

- Crawls any website and turns it into a chatbot you can talk to, with the language model running entirely in the browser through WebGPU (WebLLM) so nothing you ask leaves your device. Can also be embedded into a site as a widget with a single line of code.

LLM Distillation and Pruning (Cynaptics Club)

09/2024 - 10/2024 · GitHub

- Two-stage knowledge distillation for large models such as Llama 3.1 70B: cache the teacher's logits, then train the student against them with KL divergence to keep VRAM usage low while preserving performance.

POSITIONS OF RESPONSIBILITY

President, Cynaptics Club, IIT Indore 03/2025 - 04/2026

UG SCC Coordinator, Counseling Cell, IIT Indore 04/2024 - 04/2025

EDUCATION

B.Tech, Electrical Engineering
IIT Indore
CGPA **8.27 / 10**

2023 - 2027

Senior Secondary
CBSE Board
SCORE **94.2%**

2023

Secondary
CBSE Board
SCORE **96.2%**

2021

SKILLS

Languages

Python · C/C++ · JavaScript · HTML · CSS

ML / AI

PyTorch · TensorFlow · Hugging Face Transformers · Diffusers · RAG · model distillation · inference optimization · on-device / in-browser ML (WebLLM)

Backend & Tools

FastAPI · Django · Docker · React · MongoDB · MCP · Scrapy · Selenium · BeautifulSoup · Git

Observability & Platforms

Grafana · Datadog · Linux · cloud computing

AWARDS

- Silver Medal, Inter IIT Tech Meet 14.0, 2025
- 3rd Prize, Databricks Hackathon (IIT-level), 2025
- 1st Prize, Enosium Hackathon, Fluxus, IIT Indore, 2024
- Silver Medal, IITISoC, IIT Indore, 2024
- Solved 200+ problems across Codeforces and LeetCode