

Khushi Malik

+91-9536110472 — khushimalik13569@gmail.com — LinkedIn — GitHub — Portfolio

Education

Kirori Mal College, University of Delhi

Bachelor of Science in Mathematics

CGPA: 8.27/10.0

Delhi, India

Aug 2023 – May 2026

Work Experience

Planto AI

Lead AI Researcher

Remote, New Delhi

Aug 2025 – Nov 2025

- Spearheaded optimization of **Retrieval-Augmented Generation (RAG)** pipelines, achieving an **18% increase in LLM accuracy** and a **25% reduction in latency** through advanced re-ranking and semantic chunking strategies.
- Led a **cross-functional team of 5 engineers** to ship an AI-powered VS Code extension utilizing **multi-agent workflows**, driving a **30% boost in developer productivity**.
- Engineered **cost-efficient model fine-tuning** workflows using Hugging Face Transformers and **PEFT (LoRA)**, reducing cloud infrastructure **costs by 40%**.
- Collaborated with executive stakeholders to define AI strategy and presented data-driven insights that **accelerated adoption of AI initiatives by 15%**.

Planto AI

AI Researcher

Remote, New Delhi

May 2025 – July 2025

- Optimized **LLM inference** by implementing tokenization improvements and **prompt compression**, increasing **speed by 20%** and response **relevance by 15%**.
- Architected scalable end-to-end ML pipelines using Python and Hugging Face, enabling production rollout within 4 weeks handling **10k+ daily queries**.
- Designed and deployed **enterprise-grade RAG systems** for efficient retrieval across organizational datasets.

Freelance Projects

UNIDO AfricaRice Edge AI Mobile App

GitHub

Flutter, Dart, TensorFlow Lite, ONNX, PyTorch

- Built a **100% offline** Flutter-based mobile AI system to classify rice quality on low-end Android devices by optimizing a **ConvNeXt Vision Transformer** via ONNX graph surgery and **FP16 quantization**, reducing model **memory footprint by 50%** and tile input size by 75%.
- Prevented OOM crashes on MediaTek Helio CPUs using Dart background isolates and ARM big.LITTLE core tuning, enabling **real-time predictions from 12MP images** without UI lag on **6GB RAM devices**.

Amini GeoFM – Decoding the Field Challenge

GitHub

Python, PyTorch, Google Earth Engine

- Built a **hybrid satellite time-series** classification system to predict crop types from Sentinel-2 imagery by optimizing seasonal pattern learning using **PatchTST transformer embeddings** fused with vegetation indices (NDVI, NDRE, SAVI).
- Achieved robust classification of irregular multivariate field data using a **soft-voting ensemble** of Random Forest, SVM, and Logistic Regression, achieving **0.7473 Macro F1**.

Projects

AI Finance News System

GitHub

LangGraph, RAG Pipelines, Python, SQL

- Built a **multi-agent financial news system** to reduce information overload through redundancy removal and **actionable insight extraction**.
- Delivered context-aware Q&A for traders and managed end-to-end development from problem definition to ML optimization.

AI Medical Assistance System

GitHub

LLMs, LoRA/PEFT, Hugging Face, Python

- Fine-tuned a **medical LLM** on clinical reasoning datasets to generate accurate and explainable responses using **Chain-of-Thought prompting**.
- Integrated **real-time LLM inference** into a Python backend for clinical question answering.

Achievements & Certifications

- **Zindi Competition: Global Rank 120** as a Data Scientist among 100k Data Scientists, awarded prize money of **USD 2000** across Data Science Competitions.
- Received **Letter of Recommendation** from Coding Jr.
- **Certifications:** Excel Business Intelligence, SQL Beginner to Advanced, Python Beginner to Advanced.

Technical Skills

Languages & Frameworks: Python, SQL, TypeScript, FastAPI, PyTorch, Hugging Face

AI/ML & GenAI: RAG, PEFT/LoRA, NLP, LLMs, Transformers

Cloud & DevOps: Docker, Git/GitLab, Weights & Biases

Databases: Pinecone (Vector DB), PostgreSQL