

Aditya Nair

github.com/workforadityanair-code | nairadi17@gmail.com | linkedin.com/in/aditya-nair-088a982ab

PROFESSIONAL SUMMARY

Computer Science (Data Science) student and AI-focused builder with hands-on experience in machine learning, software engineering, and data workflows, including dataset development for a production text-to-speech (TTS) model. Experienced in taking technical projects from data collection and preprocessing through automation, validation, and working implementations, with independent projects in Python and Rust. Strong interest in AI products, rapid experimentation, automation, and practical problem solving.

EDUCATION

Manipal University Jaipur <i>B.Tech in Computer Science & Engineering (Data Science) – CGPA: 78.60%</i> Army Public School (APS) <i>Class 12 (CBSE) – 80%</i> Army Public School (APS) <i>Class 10 (CBSE) – 94%</i>	Ongoing – 2027 Jaipur, India 2023 2021
---	---

TECHNICAL SKILLS

Programming: Python, C, Java, Rust, SQL, Bash
Machine Learning: Generative AI, Large Language Models, Dataset Curation, Text-to-Speech (TTS) Pipelines, PyTorch, Hugging Face
Software Engineering: Data Pipelines, Automation, Tokenization, API Development, Git
Data & Systems: Data Preprocessing, Data Quality, Model Evaluation, Operating Systems, Linux/Unix Environments

EXPERIENCE

Data Science Intern <i>Core Dataset & Preprocessing Pipelines</i> <i>EvoTalk</i> - Built and curated the core training dataset for EvoTalk, the company's text-to-speech (TTS) model, later released on Hugging Face and GitHub. - Sourced raw audio-text pairs and defined data quality criteria to filter noisy, misaligned, or clipping samples before ingestion into the training pipeline. - Designed and ran automated preprocessing pipelines for text normalization, forced alignment, phoneme mapping, and deduplication. - Performed quality assurance audits on transcription accuracy and audio integrity, improving the reliability of ground-truth data. - Collaborated with ML researchers on end-to-end dataset preparation, validation, and release workflows supporting EvoTalk's training and public release.	Mistyoz AI
--	------------

PROJECTS

Sparse Tokenizer <i>Python, Rust</i> - Built a Python and Rust tokenizer that generates sparse token representations, prioritizing infrequent but significant tokens for efficient handling of large vocabularies. - Designed a configurable sparsity threshold to control the trade-off between memory footprint and vocabulary coverage. - Benchmarked the Rust implementation against Python on large-scale text corpora, evaluating runtime performance against development speed.	GitHub
ScriptGuard <i>Python, ML, Computer Vision</i> - Developed an ML pipeline to analyze scanned historical documents, estimate authenticity, and detect potential forgery anomalies. - Extracted handwriting style features and linguistic patterns, comparing them against verified authentic documents to flag stylistic and linguistic anomalies. - Built an interpretability dashboard with authenticity confidence scores and regional heatmaps for historians and archivists.	GitHub

CERTIFICATIONS

- Introduction to Large Language Models – Google (2024)
- Generative AI for Healthcare – University of Glasgow (2025)
- Introduction to Operating Systems – NPTEL (2025)
- Introduction to Image Processing – MathWorks (2025)
- Programming in Java – NPTEL (2024)