

ANSH SHARMA

+91 8368065604 | New Delhi, India

Data Engineer | ML Engineer | Data Scientist

anshsharma2903@gmail.com | [Portfolio](#) | [LinkedIn](#) | [GitHub](#)

PROFESSIONAL SUMMARY

Data & ML Engineer specializing in privacy-focused, local-first AI infrastructure. Proven track record of designing memory-safe data ingestion pipelines, fine-tuning LLMs under strict hardware constraints, and building end-to-end NLP data platforms.

PROJECTS

Spore - Local-First AI Data Platform | Jan 2026 – Present | [WEB-PAGE](#) | [GITHUB](#)

- Architected an **open-source, local-first AI data platform** via Docker, targeting privacy-sensitive workflows with support for local and cloud LLMs. **Achieved 450+ pulls on docker hub.**
- Engineered **memory-safe** data ingestion pipelines utilizing **columnar formats** and chunking to **eliminate OOM errors**, successfully processing **15-20 GB+** datasets.
- Validated **end-to-end data workflows** by ingesting a **6.7 GB (~8M row) TSV** dataset directly into PostgreSQL in a memory-constrained environment.
- Designed a strict **human-in-the-loop execution engine** with CORS-locked WebSocket endpoints, securing session credentials via Fernet encryption in a containerized Redis instance.
- Integrated a **sandboxed Jupyter-compatible notebook** to **mitigate XSS/RCE risks**, supporting native 3D plots, EDA, and ML reporting.

XLN-V Tuktuk HIEN LID – Token-level LID model | 2026 | [GITHUB](#) | [HUGGING FACE](#)

- Addressed a scarce NLP sub-problem, **token-level language identification** on **code-mixed Hinglish social text** where sentence-level LID systematically fails on real-world input.
- Built end-to-end: **scraped ~30k Roman-script Hinglish comments**, engineered a pre-processing pipeline, and generated **pseudo-labels** across **~456k tokens** via a **3-model weak supervision ensemble (Fleiss' $\kappa \approx 0.65$)**
- Diagnosed and discarded an initial **Qwen 3.5 0.8B fine-tune** (hallucination failure due to architecture); switched to full-parameter fine-tuning of **facebook/xln-v-base**, the architecturally correct choice for token classification.
- Trained on a single Kaggle T4 GPU under memory constraints via **8-bit Adam** and gradient checkpointing; **Achieved 0.91 weighted F1 / 0.87 macro F1** on the **external LINCE benchmark**, within **0.01 macro F1** of XLN-RoBERTa-HIEN-LID and codeswitch-hineng-lid-lince, both professionally developed models.

Wingman - Open-Source GitHub Copilot Alternative | 2024 | [GITHUB](#)

- Built a VS Code extension leveraging **LLMs** to provide context-aware code completions, documentation generation, test-case creation, and real-time debugging suggestions.
- Integrates local model serving with **Ollama** for faster inference and fully privacy-focused code-assistance – **no data leaves** the user's machine.

ACADEMIC PROJECT

Data Annotation and Collection Project | 2025

- Collected **20k+ code-mixed Hinglish data records** from **Reddit, Twitter, Facebook, Instagram, YouTube**, etc. using **Selenium/Selenium-stealth, APIs, public JSON endpoints**. Preprocessed raw data for NLP workflows including **data annotation** using Excel and **Label Studio**, contributed to pipeline optimizations.
- Mentored 4 MCA/MTECH students across institutions** on data collection techniques and methodology.
- Designed an automated routing pipeline utilizing weak-model confidence scores to cluster records, significantly reducing manual cognitive load.

SKILLS

Languages: Python, SQL, NoSQL

ML/Data: Scikit-learn, Pandas, Numpy, Apache Arrow, Apache Superset, RAG, Vector Databases

LLM/NLP: Ollama, LM Studio, NLTK, langchain, huggingface, transformers

Infra: Docker, Flask, Git, REST APIs, WebSockets

Databases: PostgreSQL, MongoDB, Supabase, Redis, DuckDB

EDUCATION

Master of Computer Applications

Guru Gobind Singh Indraprastha University, Jagan Institute of Management Studies
New Delhi

8.1/10 CGPA

08/2024 – Ongoing

Bachelor of Computer Application

Guru Gobind Singh Indraprastha University, Maharaja Surajmal Institute
New Delhi

7.9/10 CGPA

11/2021 – 07/2024