SAI ANIRUDH

+91 9004803007 • anirudhsai2005@gmail.com • linkedin.com/in/anirudh07 • github.com/aniyow07

Education

- **KJ Somaiya School of Engineering** *August 2023 – August 2027*
  *B.Tech in Artificial Intelligence and Data Science — **CGPA: 9.27/10*** *Mumbai, India*

Experience

- **Data Science Intern** *May 2026 – July 2026*
  Paradise Group Builders & Developers *Mumbai, India*
  - Cleaned and analyzed **25K+ construction project records** with SQL and Pandas.
  - Engineered **25+ predictive features** and benchmarked four models (Logistic Regression, Random Forest, Gradient Boosting, XGBoost) for delay-risk classification, achieving a **0.88 F1-score**.
  - Developed a completion-time forecasting model that reduced prediction MAE by **38% (from 9.4 to 5.8 days)** versus baseline and built a dashboard of delay probabilities, predicted timelines, and risk factors.

Projects

- **AgentReel: Multi-Agent AI Short-Video Generator** *2026*
  Python, Gemini API, PyTorch, Remotion, Manim, FFmpeg GitHub
  - Built a pipeline of **7 LLM agents** (planner, researcher, scriptwriter, storyboarder, asset generator, editor, QA) that turns a one-line brief into a narrated 30–60 s vertical video in 3 visual styles.
  - Grounded scripts in the user's files and Wikipedia, with hallucination checks that remove facts and numbers the sources don't support; covered by **170 automated tests**.
  - Stopped the LLM over-picking one template (**20 picks in 42 videos**) with no-repeat rules over ∼50 templates.
  - Cut encoding from **10.3 s to 3.1 s** per 10 s clip with NVENC, and replaced cloud TTS with a local 350M-parameter voice model running ∼**3**× **faster than realtime** in 2.9 GB VRAM.
- **Archon: Distilled Reasoning Model (Qwen3-8B)** *2026*
  PyTorch, Hugging Face PEFT/TRL, QLoRA, Unsloth, vLLM, lm-eval-harness, GCP GitHub
  - Built a **3,958-conversation** training set with a cost-tracked GCP batch pipeline, added step-by-step reasoning from a 32B model, and filtered it down to 3,366 clean conversations.
  - Fine-tuned **Qwen3-8B with QLoRA** on one L4 GPU for a concise reasoning style; eval loss **1.64 to 1.51**.
  - Benchmarked with lm-eval-harness and vLLM: kept **MMLU at 73.0%** while cutting reasoning tokens ∼**3**×; traced GSM8K and IFEval drops (88.7% to 81.7%, 85.2% to 76.9%) to uniform-length training data.
- **Halo: AI "Circle to Search" for the Desktop** *2026*
  Rust, Tauri 2, React, TypeScript, Google Gemini API GitHub
  - Built a desktop app: hit a hotkey, select part of the screen, and Gemini explains it in a floating panel.
  - Wrote the Rust backend for capture, compression and streaming, with request cancellation so old answers never mix with new ones; the API key never reaches the frontend.
  - Shipped for Windows and Linux (.exe, .deb, .rpm) with **114 automated tests**.

Technical Skills

- **LLMs & Agents:** Multi-agent pipelines, NLP, Fine-tuning & distillation, LLM evaluation, Gemini API
- **LLM Training & Serving:** PyTorch, Hugging Face (PEFT, TRL), QLoRA, Unsloth, vLLM, llama.cpp, Ollama
- **ML & Data:** scikit-learn, XGBoost, Pandas, NumPy, SQLite
- **Programming Languages:** Python, SQL, Rust, C++, TypeScript, Java
- **Cloud & Tools:** Google Cloud / GCP (Vertex AI, Compute Engine, Cloud Storage), Git, REST APIs

Achievements & Certifications

- **Amazon ML Challenge 2026** – Ranked **1,404 of 89K teams** (top 2%)
- **Codeforces Specialist** – Competitive programming in algorithms and data structures
- **Certifications** – Google Cloud Data Analytics Certificate; IUCEE Mastering Multi-Agent AI