

Aniruddha Mandal

Email: aniruddha1.2.3.3.2.1@gmail.com | Phone: +91 6289835880 | Location: Bangalore, India | Portfolio: [github](#) | [Linkedin](#)

Experience

Founding Engineer, [WebZero.ai](#) (formerly Bridge AI), Bangalore, India (April 2026 - Present) [Part-time]

- Built and shipped **WARI (Webzero Agent Readiness Index)**, a FastAPI platform measuring whether AI agents can discover, understand, and safely complete real user journeys (sign-up, checkout, support) on any website; used for a **179-site online-broker benchmark**. Live at [wari.webzero.ai](#).
- Designed the **hybrid Webzero Score**, a 4-layer index: Discovery, Action Surface (OpenAPI, GraphQL, MCP, A2A) and Governance diagnostics (60%) blended with agent task outcomes (40%) from **AgentTrek**, a parallel sandboxed browser/computer-use agent harness on Docker/ECS, under non-compensatory gates (CAPTCHA walls, AI-crawler blocks, OWASP LLM Top 10 controls).
- Built and calibrated the **LLM judge** that grades agent traces: 5-rung enum verdicts placing quality on an invertible **153-value lattice**, a 7-case labelled ground-truth suite, vision-aware judging and cross-judge arbitration, fixing verdict instability that had swung one site from **45.7 to 0.02**.
- Run the production stack on AWS ECS Fargate and RDS Postgres; cut per-evaluation page rendering from **11.8s to ~1.8s (6.5x)**, halved sandbox cost and wall-clock per run by deleting a redundant evaluation arm, and track tokens and cost per run in **Vigil**, an internal telemetry dashboard.

Founding ML Engineer, [Sol Foundry](#), Bangalore, India (Jan 2026 - April 2026)

- Architected and built **Cerebro**, the multi-agent AI runtime service powering the Klair platform comprising **10+ specialized LLM agents** for message classification, callout extraction, enrichment, duplicate detection, and lifecycle management.
- Designed a sequential context-grounded **callout extraction pipeline (v3)** with entity relationship graphs, ephemeral context stores, and entity resolution — eliminating output fluctuation from batch processing.
- Built real-time enrichment and batch onboarding agent flows processing Slack/Gmail messages through multi-stage LLM pipelines with **Cerebras** inference, **LangSmith** tracing, and dynamic prompt management.
- Built a production **evaluation framework** with **27 graders** across 5 agent types, automated Slack alerting on failures, daily quality dashboards, CI quality gates, and auto-collection of failed traces as golden test cases.
- Designed and implemented an end-to-end **SLM fine-tuning pipeline** (QLoRA on LiquidAI LFM2.5-1.2B) to distill GPT-5-mini into a self-hosted model, reducing classification latency from **2.1s to <200ms** and inference costs by **~15x**.
- Fine-tuned **DistilBERT classifiers** achieving **87.2% accuracy** for multilingual message actionability scoring (English, Hindi, Kannada, Telugu, Tamil, Spanish).
- Engineered **topic identification agents** with content-based deduplication via fingerprinting, resolving cross-topic pollution and duplicate callout issues in production.

Founding AI Lead, [Vibe AI](#), Bangalore, India (Jan 2025 - Dec 2025)

- Created synthetic dataset of **65,356** records for emotional and empathy understanding using 3 different LLMs.
- Trained and developed the world's **15th** best AI conversational model on EQBench3.
- Led and managed the research and engineering team of **3** founding engineers and psychologists.
- Planned and researched 4 quantitative and qualitative benchmarks for making the model more human-like.
- Developed v1 and v2 algorithms of the conversational agent and engineered the external agents.
- Built the foundation behavioral model for user personalization.
- Conversational model quantization and optimization to increase TFFT and TPOT.

AI Engineer, [Maura.ai](#), San Francisco, US (Aug 2024 - Jan 2025) [Part-time]

- Created and implemented a novel architecture for real-time GraphRAG with contextual updation using OpenAI GPT-4o and Neo4j.
- Integrated 10+ core functionalities of the AI and added Slack API support for AI bot creation.

Senior AI Engineer, [Healtether Health](#), Chennai, India (Jan 2024 - Dec 2024)

- Developed AI backend and Hybrid RAG architecture for the clinical diagnostic system.
- Created one of India's novel disease-symptom mapping graphs in Neo4j using patient records from **20+** hospitals.
- Led a team of **5** AI engineers to fine-tune clinical models and develop the AI backend using FastAPI and Python.

ML Engineer, [LegalGraph.ai](#), Mountain View, US (Aug 2023 - Dec 2023)

- Built foundation model for LLM pipeline and integrated knowledge graphs in Neo4j to parse legal contracts.
- Implemented Retrieval-Augmented Generation (RAG) techniques for legal query answering.

Data Science Intern, [Echo \(acquired by Spreetail\)](#), Bengaluru, India (April 2023 - July 2023)

- Built 3 scrapers for Amazon, Nykaa, and Flipkart marketplaces and created an automated pipeline to scrape reviews and ratings in DataBricks, with regular monitoring system.
- Designed the backend system in Django, Python and SQL for the dashboard of product analysis and marketplaces.

Education & Honors

The Neotia University, School of Science & Technology, West Bengal, India

B.Tech in Computer Science and Engineering with Specialization in Data Analytics (Aug 2019 - July 2023)

Cumulative GPA: **9.0/10.0**

- Ranked **3rd** in Amazon ML Challenge 2021 hosted at HackerEarth.

Skills

- **Programming Languages:** Python, C/C++, R, Go.
- **Backend & Databases:** FastAPI, Django, PostgreSQL, MongoDB, Neo4j, Scylladb, Chromadb, Weaviate, Pinecone.

- **Machine Learning & AI:** TensorFlow, PyTorch, LangChain, HuggingFace Transformers, vLLM, Keras, JAX, Scikit-Learn, smolagents, LangGraph, OpenAI Agent SDK, trl, PEFT/LoRA.
- **Cloud & Data Engineering:** AWS, BigQuery, Terraform, Databricks, Docker, Kubernetes, Github Actions, Terraform.
- **LLM Ops & Evaluation:** LangSmith, Cerebras, QLoRA fine-tuning, model quantization (AWQ/GPTQ/GGUF), prompt engineering, custom evaluation frameworks, llama.cpp.

Projects & Publications

- **Impact of Non-Academic Parameters on Academic Performance** (DOI: [10.1007/978-3-031-24848-1_22](https://doi.org/10.1007/978-3-031-24848-1_22)): A Rough Set Theory-based study on how various factors affect student performance.
- **A Decadal Walk on BCI Technology** (DOI: [10.4018/978-1-7998-4706-9.ch006](https://doi.org/10.4018/978-1-7998-4706-9.ch006)): Discusses how human brain waves are read using BCI technology and how ML is applied.
- **CodeExplainer:** A multi-agent system that indexes, analyzes, and answers questions about codebases from GitHub using knowledge graph and LLM-powered reasoning.
- **LocalLoom:** Loom alternative for generating user guides from video recordings using VLM and ASR models.
- **VideoRAG:** Video based RAG pipeline with scene detection and qdrant vector database
- **Agentic Web Browser:** Agentic web browser that can perform browser based tasks in natural language.
- **DDPM App:** Implementation of Denoising Diffusion Probabilistic Model, and diffusion-based image to image model from scratch.
- **TicketBot:** An end-to-end pipeline for analyzing app reviews using LLM-powered topic discovery, clustering, and trend analysis.
- **DungeonForge:** End-to-end 8bit RPG dungeon video game builder from prompt using smolagents.
- Opensource contributor to **Cognee**: #1449 and #1333.