Debarghya Saha

MACHINE LEARNING ENGINEER

debarghyasren@gmail.com · +91 9147031684 · GitHub · LinkedIn · Portfolio

TECHNICAL SKILLS

Languages:	Python, TypeScript, SQL
MLOps / Cloud:	Docker, Kubernetes, Terraform, GitHub Actions, CI/CD, MLflow, Prometheus, Grafana, AWS, GCP, Linux
ML / AI:	PyTorch, HuggingFace Transformers, LangChain, scikit-learn, ONNX Runtime, vLLM, llama.cpp, Stable Baselines3, LLM Fine-Tuning, Quantization, RAG, NLP, Computer Vision, Prompt Engineering
Data / Serving:	FastAPI, Django, Apache Beam, Apache Kafka, Airbyte, Pandas, NumPy, REST APIs, Microservices, System Design
Databases:	PostgreSQL, MongoDB, Milvus, Pinecone, Vector Databases (HNSW)
Speech / Realtime:	FastSpeech2, TTS/STT, SSML, FreeSWITCH, Asterisk

PROFESSIONAL EXPERIENCE

Ethical Den (Hivyr.ai), Junior Data Scientist *Oct 2025 – Present*

- Led a cross-functional team of 10 engineers (ML, backend, frontend, and AI interns) across model architecture, training optimization, and deployment strategy, keeping research work on a shippable path to production.
- Removed GPU hardware from the serving path entirely, measured by production inference running on Raspberry Pi class devices instead of CUDA servers, by converting PyTorch models to ONNX Runtime and re-architecting CUDA-bound workflows into a quantized CPU-only pipeline.
- Built and deployed production ML pipelines for automated valuation, forecasting, document extraction, and record deduplication using OCR, vector embeddings, and supervised models, replacing manual review with automated scoring.
- Designed a RAG assistant for natural-language querying across structured and unstructured enterprise data, plus agentic workflows on self-hosted LLMs, keeping inference on-premise for privacy-sensitive, offline decision automation.
- Shipped a production AI voice agent at sub-200 ms end-to-end response latency with human-level prosody and pronunciation, by fine-tuning and quantizing TTS models down to a 3 GB deployable artifact and parallelizing the streaming inference pipeline.

Ethical Den (Hivyr.ai), AI Intern *Jul 2025 – Sep 2025*

- Shipped a real-time multilingual conversational voice agent chaining speech-to-text, an on-device LLM (Gemma 3n E4B), sentiment detection, and text-to-speech into one low-latency streaming pipeline with no cloud inference dependency.
- Cut TTS model training time by 32% by building a FastSpeech2 pipeline in PyTorch over Epitran-generated IPA phoneme sequences, improving prosody and emotion fidelity with pitch, energy, and duration embeddings.
- Engineered an Apache Beam pipeline that curated 24 kHz multilingual voice datasets with SSML markup, scaling dataset preparation from manual collection to an automated, repeatable job.

TECHNICAL PROJECTS

RagProd: Self-Hosted Retrieval-Augmented Generation QA System · GitHub

- Built a self-hosted RAG question-answering service that retrieves across 500,000 documents in under 6 seconds, using HNSW indexing in Milvus behind an async FastAPI backend, PostgreSQL for relational storage, and vLLM (Qwen2.5-3B-Instruct) for streaming inference.
- Containerized the full stack with Docker Compose using network-segregated zones and health-checked orchestration, eliminating startup race conditions across services.
- Provisioned on-demand GPU infrastructure as code with Terraform on RunPod, enabling one-command switching between local and cloud-hosted inference.

Deep Reinforcement Learning Agent for Pokémon Red · GitHub

- Built a custom OpenAI Gym environment from ROM disassembly and memory mapping to simulate battle mechanics as a trainable RL task.
- Reached an 82% win rate and 35% higher training throughput by training PPO and A2C agents with reward shaping, Generalized Advantage Estimation, and action-space optimization.

EDUCATION

B.Tech, Computer Science and Business Systems, Meghnad Saha Institute of Technology, West Bengal, India *2022 – 2026*

ACHIEVEMENTS

- **Smart India Hackathon 2024 Winner:** led a team to first place on Ministry of Education Problem 1742 with an AI-powered educational solution, ranking in the top 0.3% of 88,221+ competing teams.
- **AI4Andhra Police Hackathon Finalist:** built connectors for police-database retrieval and automated summarization.
- **Smart India Hackathon 2023 Internal Winner:** placed first institution-wide with a disaster-management simulation application.