Anubrat Bora

+91 60029 33403 | anubratbora25@gmail.com | anubratbora.com | Google Scholar | GitHub | LinkedIn

Education

Manipal Academy of Higher Education — Bangalore, KA
Bachelor of Technology in Computer Science (Artificial Intelligence), CGPA: 8.3/10 — *Sep. 2022 – May. 2026*

Central Board of Secondary Education — Golaghat, AS
Intermediate (Class XII): 93.6%, Matriculation (Class X): 96.2% — *Jul. 2021*

Experience

AI Research Engineer — May. 2026 – Aug. 2026
Navdyut AI Labs — *Guwahati, AS*

- Built a native Assamese tokenizer (32k vocab, Unigram LM) on a 162M-token corpus with cross-lingual normalization repairing 11.4% of the data, achieving 3.6× fewer tokens/word than Llama 3 and 46.8% better compression than Gemma.
- Pretrained a 240M-parameter decoder-only LLM on 14.4B curated tokens using FSDP on a distributed A40 cluster, applying μP to zero-shot transfer hyperparameters from a 31M proxy and compressing vocabulary to 8.2k tokens to reallocate ∼108M embedding parameters into model width.

AI Engineer Intern — Jun. 2025 – May. 2026
Barsys — *New Delhi, DL*

- Built scalable LLM-powered RAG pipelines using FastAPI, OpenAI Embeddings, PostgreSQL (pgvector), and HNSW indexing, enabling sub-second semantic retrieval and achieving 95% recommendation accuracy.
- Engineered an AI support agent using Gemini 2.5 Flash for intent classification across 10 customer-service intents, combining hybrid lexical-semantic retrieval with recall@k and per-class accuracy evaluation.

Undergraduate Researcher (AI) — Sep. 2023 – Mar. 2024
Manipal Institute of Technology Bangalore — Dept. of CSE — *Bangalore, KA*

- Designed and evaluated LLM-powered recommendation and semantic-classification frameworks integrating Knowledge Graphs, ontologies, and generative AI for automated knowledge discovery.
- Authored 8 peer-reviewed publications across IEEE, ACM, Springer Nature, and CEUR-WS in LLMs, semantic AI, ontology engineering, recommendation systems, and knowledge-centric AI.

Projects

High-Throughput LLM Serving & Quantization Benchmark — Jul. 2026 – Aug. 2026

- Deployed Llama-3-8B behind a streaming FastAPI endpoint using vLLM with continuous batching and PagedAttention, serving 1,400 tokens/sec aggregate throughput at 180 ms p95 latency on a single A10 GPU.
- Benchmarked AWQ, GPTQ, and GGUF 4-bit quantization against FP16, reducing VRAM footprint by 68% with under 2% degradation on MMLU and HumanEval.
- Containerized the full stack with Docker and published reproducible load-test results (50 concurrent users via Locust) with per-config latency, throughput, and memory tables.

Self-Correcting Data-Analysis Agent with Task Evaluation — Aug. 2026 – Sep. 2026

- Built a LangGraph-based agent that plans, writes, and executes pandas/matplotlib code against arbitrary CSVs in a sandboxed runtime, with error-driven self-correction across up to 3 retry iterations.
- Designed an evaluation harness of 40 held-out analysis tasks with programmatic answer checking, achieving 87% task-completion rate and 92% recovery from injected tool failures.
- Deployed a live demo on Hugging Face Spaces with streaming intermediate reasoning, structured JSON outputs via Pydantic validation, and per-run cost/latency tracking.

Technical Skills

Languages: Python, C, Java, SQL, JavaScript, HTML & CSS, LaTeX
Frameworks: FastAPI, LangChain, LangGraph, PydanticAI, React.js, Node.js, Express.js
Developer Tools: Linux, Git, Docker, GCP, Firebase, PostgreSQL (pgvector), Weights & Biases, Claude Code
Libraries: PyTorch, Hugging Face (Transformers, Tokenizers, PEFT), TensorFlow, Scikit-learn, Pandas, NumPy, Matplotlib, Streamlit