

Anubrat Bora

+91 60029 33403 | anubratbora25@gmail.com | anubratbora.com | [Google Scholar](#) | [GitHub](#) | [LinkedIn](#)

EDUCATION

Manipal Academy of Higher Education

Bachelor of Technology in Computer Science (Artificial Intelligence)

Bangalore, KA

Sep. 2022 – Jul. 2026

Central Board of Secondary Education

Intermediate (XII) : 93.6%, Matriculation (X) : 96.2%

Golaghat, AS

Jul. 2021

EXPERIENCE

AI Engineer Intern

Barsys

Jun. 2025 - May. 2026

New Delhi, DL

- Built scalable data ingestion and embedding pipelines using FastAPI, OpenAI Embeddings, and PostgreSQL (pgvector), creating a searchable vector database for semantic retrieval and recommendations; implemented HNSW-based vector search achieving sub-second latency and 95% recommendation accuracy.
- Engineered an AI support agent using Gemini 2.5 Flash for intent classification and routing across 10 customer-service intents; developed hybrid lexical-semantic RAG and evaluation pipelines for deployment.

Undergraduate Researcher (AI)

Manipal Institute of Technology Bangalore — Dept. of CSE

Sep. 2023 - Mar. 2024

Bangalore, KA

- Designed and evaluated AI systems combining LLMs, KGs, Ontologies, and Deep Learning to solve real-world problems in recommendation, document annotation, semantic classification, and automated knowledge discovery.
- Authored 6 peer-reviewed research publications in leading international venues, advancing state-of-the-art techniques in semantic intelligence, ontology synthesis, reinforcement learning, and knowledge-centric ML frameworks.

PROJECTS

Developing a Transformer from “Attention Is All You Need”

Feb. 2025 - Apr. 2025

- Implemented the Transformer architecture from scratch, referencing the seminal “Attention Is All You Need” paper, demonstrating expertise in advanced neural network design.
- Trained the model on a custom dataset using PyTorch, optimizing hyperparameters to achieve a BLEU score improvement of 15% over baseline performance.
- Analyzed model performance through detailed evaluation metrics, such as attention heatmaps and loss curves, gaining insights into Transformer interpretability and optimization strategies.

Fine-Tuning an Open-Source LLM for SQL Generation

Jan. 2024 - Mar. 2025

- Fine-tuned Llama on a dataset of English-to-SQL query pairs using techniques like LoRA and parameter-efficient fine-tuning, reducing inference errors by 25%.
- Developed a preprocessing pipeline to clean, tokenize, and balance the dataset, improving model training efficiency and SQL query accuracy.
- Evaluated model performance using execution accuracy and query-level metrics, identifying optimization opportunities and improving generated SQL quality.

RAG Workflow with LangChain, ChromaDB & Open-Source LLM

Dec. 2024 - Jan. 2025

- Built a Retrieval-Augmented Generation (RAG) pipeline by integrating LangChain, ChromaDB, and an open-source LLM, enabling context-aware, domain-specific text generation.
- Utilized ChromaDB indexing for efficient retrieval, achieving a 10% reduction in query latency and improving response relevance in real-time applications.
- Utilized comprehensive evaluation frameworks to measure retrieval accuracy, model performance, and end-to-end system reliability, ensuring production-readiness.

TECHNICAL SKILLS

Languages: Python, C, Java, SQL, JavaScript, HTML & CSS, LaTeX

Frameworks: LangChain, LangGraph, PydanticAI, React.js, Node.js, Express.js, PostgreSQL(pgvector), FastAPI

Developer Tools: Linux, Git, Docker, GCP, Firebase, Claude Code

Libraries: Transformers, Scikit-learn, Streamlit, Pandas, NumPy, SciPy, Matplotlib, TensorFlow, PyTorch, OpenCV