Bhaswata Choudhury

bhaswata08@gmail.com | +91 700210456

GitHub | LinkedIn

Professional Summary

Generative AI Engineer with strong AI fundamentals and experience in developing and implementing cutting-edge AI applications. Proven expertise in Python, Large Language Models (LLMs), and the LangChain framework, specializing in designing, building, and optimizing LLMs for generative AI solutions. Expert in designing and building AI agents and AI Agent frameworks, with deep knowledge of Model Context Protocol (MCP) servers, prompt engineering, structured language modeling, and LLMOps. Passionate about collaborating with cross-functional teams to deliver scalable, efficient, and innovative AI-driven systems.

Technical Skills

Programming: Python, Rust

Backend: Development of Microservices using FastAPI

LLMOps: Langchain, Langgraph

Vector Databases: Pinecone, Weaviate, Chroma, Qdrant, Milvus

Model Serving & Deployment: Hugging Face TGI, VLLM, Xinference, Infinity servers, Docker, Kubernetes

Observability & Monitoring: Weights & Biases (W&B), Langfuse, Langsmith

Cloud AI Platforms: AWS SageMaker, Azure ML, GCP Vertex AI

Experience

Project Engineer Aug 2024 – Jul 2025

Centre for Development of Advanced Computing (C-DAC) *Bangalore*

- Led the development and deployment of AgriCHAT, a high-profile generative AI solution for a key client, managing the entire project lifecycle independently from design to production deployment.
- Engineered a proprietary Agentic Splitter technology that improved RAG recall by 40% over conventional methods through advanced header identification and optimized document chunking strategies.
- Achieved State-of-the-Art (SOTA) performance in PDF extraction to Markdown conversion, significantly enhancing data ingestion pipelines for downstream AI processes.
- Designed and implemented a sophisticated multi-agent AI framework orchestrating complex components including PDF extraction, embedding management (Infinity Server), vector storage (Milvus), and LLM services (VLLM).
- Spearheaded the creation of a novel synthetic data generation pipeline for Indic NER datasets, producing 1.5 million high-quality tagged sentences across multiple Indic languages to address critical data scarcity challenges.
- Developed innovative workflows leveraging LLMs to generate English sentences from RDF Triplets, implementing accurate entity alignment mechanisms for cross-lingual NER model training.
- Deployed full-stack AI solutions using FastAPI backends, Streamlit frontends, and Docker Compose for containerization, ensuring robust production-ready systems with comprehensive observability through Langfuse.

Intern Jan 2024 – Jun 2024

Centre for Development of Advanced Computing (C-DAC) *Bangalore*

- Developed an innovative agentic chatbot for Project MANAS (mental health platform), serving as an intelligent router with capabilities to trigger external and internal APIs for seamless user navigation.
- Pioneered a novel graph-based memory system for personalized user interactions, enabling advanced agentic reasoning and memory operations months before similar public solutions like Mem0 were available.
- Architected and implemented complex AI agent frameworks with sophisticated context-aware responses, demonstrating early expertise in agentic AI systems and memory management.
- Enhanced user experience through intelligent routing mechanisms and personalized interactions, contributing to improved mental health support delivery through AI-driven solutions.
- Collaborated with cross-functional teams to integrate the chatbot seamlessly into the existing MANAS ecosystem, ensuring smooth user experience and system reliability.

Projects

AgriCHAT Nov 2024 – Jun 2025
Centre for Development of Advanced Computing (C-DAC) *Bangalore*

- Engineered and delivered a high profile generative AI solution for a key client, independently managing the entire project lifecycle from design to deployment.
- Developed a proprietary Agentic Splitter technology that improved RAG recall by 40% over conventional methods by leveraging advanced header identification for optimized document chunking.
- Achieved State-of-the-Art (SOTA) performance in PDF extraction to Markdown, enhancing data ingestion for downstream AI processes.
- Designed and implemented a sophisticated AI agent framework orchestrating multiple complex components including PDF extraction, embedding management (using Infinity Server for dense and sparse embeddings), and an LLM service (VLLM).
- Managed robust data storage solutions utilizing Milvus vector database and ensured system reliability with Langfuse for observability.
- Spearheaded the deployment of the full-stack solution using FastAPI for backend services, Streamlit for front-end, and Docker Compose for containerization, ensuring seamless operationalization.

Synthetic Indic NER Dataset Generation Aug 2024 – Nov 2024
Centre for Development of Advanced Computing (C-DAC) *Bangalore*

- Led the end-to-end development of a novel synthetic data generation pipeline, creating a 1.5 million sentence high-quality Named Entity Recognition (NER) Indic dataset to address critical data scarcity.
- Architected a sophisticated workflow leveraging LLMs to generate English sentences from RDF Triplets and applied an existing English NER tagging model.
- Developed and implemented a unique mechanism to ensure accurate entity alignment and high-quality translation of tagged entities across multiple Indic languages.
- Collaborated with linguists to validate and ensure the high quality of the generated parallel NER sentences across multiple Indic languages.
- The generated dataset is actively being used by colleagues to train robust NER models for Indic languages.

Agentic AI for Mental Health (Project MANAS) Jan 2024 – Jun 2024
Centre for Development of Advanced Computing (C-DAC) *Bangalore*

- Developed an agentic chatbot as a mental health router for Project MANAS, integrating capabilities to trigger external and internal APIs to guide users to relevant application sections.
- Engineered a novel, graph-based memory system for personalized user interactions, enabling advanced agentic reasoning and memory operations (predating public solutions like Mem0).
- Enhanced user experience through intelligent routing and context-aware responses, demonstrating expertise in complex AI agent design and implementation.

Education

Tezpur University Aug 2020 – Jun 2024
Bachelor of Technology — CGPA: 8.04 *Tezpur*

Certifications & Achievements

- **Hasgeek Open source Hackathon Winner 2024:** Winner for Open source hackathon 2024 *2023-2024*
- **AI Programming with Python Nanodegree with AWS and Udacity:** Top 500 Global Rank *2023-2024*
- **Machine Learning Fundamentals with AWS and Udacity:** - Top 500 Global Rank *2022 - 2023*
- **Deep Learning:** - NPTEL *2022*
- **Fundamentals of Artificial Intelligence:** - NPTEL *2022*