Sagnnik Biswas

☎ +91-8158867604 ✉ sagnnikbiswas2002@gmail.com Sagnnik Biswas Sagnnik

Profile

Applied AI / ML Engineer with experience building multimodal LLM systems, GenAI applications, and scalable ML infrastructure. Proven ability to translate research ideas into production-ready solutions, with expertise in LLM alignment, multi-agent architectures, distributed training, and real-time model serving.

EXPERIENCE

Research Internship — Jan 2024 - Feb 2025

Rakuten India — *Bangalore*

- Designed and built a cross-modal GNN–LLM system for structured banner layout generation, improving multimodal reasoning capabilities in design workflows.
- Developed graph-based spatial encoders and aligned embeddings with LLM representations with CLIP based objectives to enable structured reasoning over visual layouts.
- Implemented large-scale distributed training using multi-GPU DDP and DeepSpeed ZeRO-3, reducing training time and supporting larger model configurations.

Research Internship — Jan 2023 – Dec 2023

IIT Kharagpur - Mentor Prof. Somak Aditya

- Modeled entity–event causal scene graphs to represent interventions in spatio-temporal systems.
- Created a counterfactual QA benchmark to evaluate robustness under distribution shifts and reasoning consistency of LLMs under hypothetical policy changes (manuscript prepared for EACL 2024).

Robotics Internship — Jul 2023

Aeonix Research and Innovations LLP — *Kolkata*

- Built an automated industrial metrology system using YOLOv8 + SAM + geometric CV pipelines and reduced inspection time from 15–20 min to less than 1 min.
- **Awarded Best Paper** – RCAAI 2023 (Springer LNEE).

PROJECTS

Docker MCP Bridge - LLM Tool Orchestration Infrastructure — Dec 2025 - Present

- Built a secure, container-native middleware enabling dynamic discovery, generation, and sandboxed execution of LLM tools for automated workflows via MCP.
- Implemented Redis-backed tool registry, user-scoped MCP sessions, and interrupt/resume workflows.
- Orchestrated isolated tool execution using Docker with FastAPI streaming APIs.

Multi-Agent Web Researcher — Nov 2025

- Built a LangGraph-based research agent that orchestrates planner, MCP-powered search/gather, RAG synthesis, quality-check loops, with Redis-backed checkpointing and streaming via FastAPI.
- Designed a ChromaDB-backed RAG pipeline that chunks and embeds multi-source search results, retrieves top-K relevant passages per query and generates citation-grounded reports.

Ollama Infer - GPU LLM Serving Platform — Nov 2025 - Present

- Engineered a LLM serving platform by extending Ollama with memory-aware model scheduling (knapsack-based VRAM eviction) to maximize GPU utilization and reduce inference latency.
- Designed asynchronous inference pipelines using Celery + Redis with retry logic, rate limiting, and failure isolation to ensure reliability under concurrent load.

TECHNICAL SKILLS

Languages & Frameworks: Python, PyTorch, JAX, SQL, LangGraph, Pydantic AI

Machine Learning: Transformers, GNNs, RAG, Multi-Modal Reasoning, Semantic Search & Ranking

Systems for ML: Fine-tuning & Alignment, Model Evaluation & Experimentation, Distributed Training (DDP, FSDP, DeepSpeed ZeRO-3), Multi-GPU Inference, Model Serving

Backend & Infrastructure: FastAPI, Docker, Redis, Celery, Github Actions, DigitalOcean (cloud deployment), MongoDB, PostgreSQL

EDUCATION

Manipal Institute of Technology — Oct 2020 – Nov 2024

BTech - Electronics and Communication (Data Science minor) - ***CGPA - 8.34*** — *Manipal, Karnataka, India*