Sagnnik Biswas

+91-8158867604 | sagnnikbiswas2002@gmail.com | Sagnnik Biswas | Sagnnik

EDUCATION

Manipal Institute of Technology **Oct 2020 – Nov 2024**
BTech - Electronics and Communication (Data Science minor) - ***CGPA - 8.34*** *Manipal, Karnataka, India*

EXPERIENCE

Research Internship **Jan 2024 - Feb 2025**
Rakuten India *Bangalore*

- Designed a cross-modal GNN–LLM architecture for structured banner layout generation.
- Built spatial layout GNNs and aligned graph embeddings with LLM representations to improve multimodal reasoning.
- Scaled end-to-end training using multi-GPU DDP and DeepSpeed ZeRO-3, enabling efficient large-scale training and inference.

Research Internship **Jan 2023 – Dec 2023**
IIT Kharagpur - Under the guidance of Prof. Somak Aditya

- Constructed entity–event causal scene graphs from collision video datasets to represent spatio-temporal interactions.
- Designed counterfactual QA tasks on these graphs to evaluate LLM reasoning under interventions.
- Analyzed LLM failure cases in temporal and counterfactual reasoning using synthetic video data.

Robotics Internship **Jul 2023**
Aeonix Research and Innovations LLP *Kolkata*

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

RoboManipal **Dec 2020 – Dec 2023**
Electronics Subsystem Member *MIT, Manipal*

- Designed and debugged circuits and control systems for competition robots.
- Integrated computer vision–based automation pipelines.
- Selected for ABU Robocon Nationals (2022, 2023).

PROJECTS

Docker MCP Bridge — LLM Tool Orchestration Infrastructure **Dec 2025**

- 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**

- Designed a LangGraph-based stateful multi-agent research engine with Redis checkpointing and resumable execution.
- Implemented real-time SSE streaming, citation synthesis, and multi-source search pipelines.
- Built a Streamlit frontend for interactive research workflows.

Ollama ScaleWright — GPU LLM Serving Platform **Nov 2025**

- Built a GPU-aware LLM serving platform with Celery-based async inference and Redis-backed model tracking
- Implemented a knapsack-based VRAM eviction algorithm for dynamic model loading/unloading.
- Enabled real-time streaming inference for low-latency, multi-user LLM workloads.

TECHNICAL SKILLS

Languages: Python, SQL
Machine Learning & Research: PyTorch, Transformers, Graph Neural Networks, Multimodal Learning
Systems for ML: Distributed Training (DDP, FSDP, DeepSpeed ZeRO-3), Multi-GPU Inference, Model Serving
Backend & Infrastructure: FastAPI, Docker, Redis, Celery, LangGraph