Spandan Das

spandand515@gmail.com • sd325.github.io • spandand • SD325

Education

Columbia University **2025 – 2026**
M.S. Computer Science

Carnegie Mellon University **2021 – 2025**
B.S. Computer Science

Relevant Coursework: (PhD) Deep Learning [Python], Deep Reinforcement Learning [Python], (PhD) Advanced NLP [Python], (PhD) Convex Optimization, Computer Systems [C], Algorithms, Probability, Statistics, Database Systems [C++]

Experience

DatologyAI **July 2025 –**
Member of Technical Staff
- Training large-scale models that power DatologyAI's data curation platform

NVIDIA **May – August 2024**
Artificial Intelligence Engineer Intern
- Developed an anomaly detection system for NVIDIA TEGRA chip production environment
- Implemented monitoring system to autonomously report issues in build, packaging, and testing processes via email and Slack
- Designed and built a real-time harmless error filter using **ElasticSearch** and **Flask** to perform LLM search over vector database of log embeddings

Apple **May – August 2023**
Machine Learning Engineer Intern
- Wrote **Golang** backend service to automate labeling queries with LLM-based natural language understanding results
- Created LangChain-inspired LLM integration library to filter and annotate semantic search results over 2.3B datapoints across Siri domains
- Technologies used: **AWS**, **Docker**, **Kubernetes**

NASA Goddard Space Flight Center **June – August 2020; June – August 2021**
Research Intern
- Trained ML models (**TensorFlow**, **Scikit-learn**, **XGBoost**) on GPM mission data to reduce satellite costs
- Used NASA NCCS supercomputing cluster to process 2016–2017 satellite data and optimize bagging models via multithreading
- Presented research to GSFC Climate and Radiation Lab and at AGU Fall Meeting
- Published in MDPI Remote Sensing Journal [`https://doi.org/10.3390/rs14153631`]

Research & Projects

CMU Language Technologies Institute (CX Group) **February – May 2024**
- Developed an active learning approach for data-efficient instruction tuning using data impact models
- Improved pretraining efficiency by adapting to models' evolving data preferences
- Published in NeurIPS 2024 [`https://arxiv.org/abs/2406.06046`]
- Technologies: **HuggingFace**, **PyTorch**

Visual Question Answering with LLMs **May – August 2023**
- Redesigned Winoground dataset as a visual question answering (VQA) problem
- Evaluated with MiniGPT4, PromptCap, ViperGPT, and LLaVA
- Submitted paper to EMNLP 2023 [Link to Paper]

CMU Robotics Institute (AirLab) **May – August 2022**
Research Assistant
- Developed an online camera calibration algorithm for multi-view stereo (6 cameras; Double Sphere model) on drones for real-time depth maps
- Technologies: **PyTorch**, **CUDA**, **OpenCV**, **Docker**

Achievements

- 2021 USA Math Olympiad – Top 2% (Top 550/30,000+; 232.5 USAMO Index)
- USA Computing Olympiad (USACO) – Top 600 in nation (Gold Division)

Skills/Extracurriculars

Technical Skills: Java, Python, Golang, C, C++, LaTeX, HTML, Linux

Extracurriculars: Tennis, Hindustani Classical Music, CMU Sahara (Bollywood Fusion Dance), Basketball, Card/Board Games