Saransh Agrawal

saranshagarwal2020@gmail.com | linkedin.com/in/saransh-agarwal8899 | Website | GitHub | Seattle, WA

Education

Texas A&M University *Master of Science in Data Science: (GPA: 3.7/4.0)* *College Station, TX,* ***May 2025***

Courses: Parallel Computing, Database Mgmt., Natural Language Processing, Reinforcement Learning, Data Stream & Algo.

Manipal Institute of Technology *Bachelors in Electronics and Instrumentation Engineering* ***Aug 2021***

Courses: Data Structures & Algorithms, Microprocessors, Digital Image Processing, Embedded Systems, Digital System Design

Technical Skills

Languages: Python, Go, Rust, C++, C, SQL, NoSQL(Mongo), Bash/Shell Scripting
Tools: Docker, Jenkins, Terraform, Kubernetes, PyTorch, Numpy, Scikit-Learn, MCP, Apache Spark, Hadoop, Git
Cloud: AWS, EC2, Lambda, ECS/EKS, S3, Redshift, EMR, RDS, DynamoDB, Sagemaker, Bedrock
Certificates: AWS Solutions Architect - Professional

Work Experience

Research Assistant **Aug 2024 – May 2025**

FLAIR Lab, Texas A&M University *College Station, TX*

- Engineered a distributed, multi-node/multi-GPU training pipeline for LLMs (**Llama3, Mistral-7B**) on a **SLURM HPC** cluster using **PyTorch**, **DeepSpeed**, and **Accelerate**, scaling training throughput by **4x**.
- Reduced inference latency by **21%** for a long-context multimodal model (**LLaVA**) by implementing **Streaming LLM** and **Cache Merging** techniques, increasing effective context length by **20%**.
- Architected and built a scalable framework to automate **LLM benchmarking** (MMLU) across diverse GPU configurations, managing job scheduling and resource allocation with **SLURM**.

Software Development Engineer **July 2021 – July 2023**

Viewzen Labs

- Architected a scalable MLaaS platform that streamlined deployment for **50+ users**, proven to reliably handle over **80 concurrent training jobs** during peak demand.
- Designed a robust microservices architecture using **REST APIs, Kafka, and Docker** to ingest and process over **5M** data points daily with **99.9%** uptime.
- Optimized a high-volume data transformation pipeline by **20%** by re-engineering critical components in **C++** and implementing concurrent processing.
- Developed and deployed an end-to-end prediction system to identify at-risk users from a highly imbalanced dataset, achieving an F1-score of **0.85** in production.
- Drove the adoption of engineering best practices by mentoring **2** junior developers on system design and automated **CI/CD pipelines** with **Jenkins**.

Machine Learning Engineer Intern **Feb 2021 – June 2021**

Centre for Digital Financial Inclusion

- Engineered and deployed an end-to-end **Speech-to-SQL** microservice using **Docker**, **Flask**, and **PostgreSQL** to power a natural language interface for a data dashboard.
- Enabled real-time, voice-activated data visualization, achieving a **90% success rate** in generating valid SQL queries from speech and delivering charts in under **4 seconds**.
- Achieved a **17% Word Error Rate (WER)** on a custom Indian Accent dataset by fine-tuning a **DeepSpeech** ASR model, ensuring high-fidelity transcription for the query generator.

Publications

[1] **Saransh Agrawal** and Kuan-Hao Huang. Selective Amnesia – Constrained Unlearning for Large Language Models via Knowledge Isolation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025) at ACL.*

[2] Ren-Wei Liang, Chin-Ting Hsu, Chan-Hung Yu, **Saransh Agrawal**, Shih-Cheng Huang, Shang-Tse Chen, Kuan-Hao Huang, and Shao-Hua Sun. Adaptive Helpfulness-Harmlessness Alignment with Preference Vectors. *arXiv:2504.20106* (under review)

Projects

- **Satellite-based Crop Monitoring System (SCMS)** — *Python, FastAPI, Docker, React.js, MongoDB*
  - Led backend system design for a real-time monitoring platform, architecting microservices with **FastAPI** and **Docker** to ingest and process terabytes of satellite imagery.
  - Deployed a forecasting microservice using an **LSTM** model, achieving **81% prediction accuracy** and enabling data-driven agricultural insights via a React dashboard.
- **Graph-Based Ranking & Semantic Search Engine** — *Go, Python, Apache Arrow, Sentence-Transformers*
  - Engineered an end-to-end semantic search engine in **Go** and **Python** to index and serve over **70k** research papers from the ACL Anthology. Boosted search relevancy by **30%** over a keyword baseline by implementing a hybrid ranking system combining **PageRank** and **Sentence-Transformers**.