

Sraghav Sharma

[412-475-4932](tel:412-475-4932) | sraghavs@andrew.cmu.edu | [linkedin.com/in/sraghav-sharma](https://www.linkedin.com/in/sraghav-sharma) | github.com/sraghav12 | Pittsburgh, PA

AI/ML engineer and data scientist (CMU MISB-BIDA, Dec 2026) with 2+ years shipping production ML - computer vision, LLM/RAG, and cloud analytics pipelines - with measured impact: +4% revenue, 300x query-cost reduction, 1,500+ store rollout.

EDUCATION

Carnegie Mellon University

August 2025 - Expected Dec 2026

Master of Information Systems Management - Business Intelligence and Data Analytics

Coursework: Introduction to Deep Learning, Artificial Intelligence Venture Studio, Unstructured Data Analytics

Thapar Institute of Engineering and Technology

August 2020 - June 2024

Bachelor of Engineering in Electronics and Computer

SKILLS

ML/LLM: PyTorch, scikit-learn, NLP, CLIP, LangChain, LangGraph, RAG (ChromaDB, AstraDB, Neo4j), evaluation/metrics

MLOps & Deployment: Docker, Airflow, FastAPI, Flask, model serving, monitoring/logging

Programming, Data & Cloud: Python, SQL, R, Java, BigQuery, GCP, AWS, Azure, Spark, Kafka, Power BI

PROFESSIONAL EXPERIENCES

Data Analytics Intern, Kinertia (Wing Jockey) - Pittsburgh, PA

Jun 2026 - Aug 2026

- Shipped a production Airflow DAG on GCP BigQuery aggregating 32M+ flight legs (2.7 TB of ADS-B data) into daily statistics for 183K aircraft and 13.6K airports; cut per-run scan cost 300x (\$6 to \$0.02) via partition and columnar pruning
- Built an LLM-powered pipeline classifying 847 aircraft types into a dimension table joined across all analytics; validated 96.8% coverage over 2.6M flights against an independent heuristic method
- Prototyped a public-facing analytics site (airport dashboards, per-aircraft stat cards, national heat map of 8,756 airports); demoed to company leadership and greenlit for publication
- Ran a user re-engagement experiment (email vs. in-app, 100 dormant users) and diagnosed the null result via BigQuery cohort analysis, surfacing a confirmed push-notification product gap

AI Engineer Research Assistant, Dynamic Decision Making Lab, Carnegie Mellon University - Pittsburgh, PA

Jun 2026 - Aug 2026

- Designed and built Brigade Kitchen, a parameterized 4-agent cooperative environment (Python, Gym-style) with 10 interdependent subtasks, hazards, and an order economy for human-AI teaming experiments; approved by lab PI
- Implemented rule-based agents (BFS pathfinding + priority state machines) and per-step JSONL telemetry for behavioral analysis; 36 unit tests, 3 experiment layouts

AI/ML Engineer, Sampark Softwares

Jun 2024 - Jun 2025

- Owled end-to-end rollout of a Dockerized, high-concurrency Flask inference API for YOLOv11, running 30s capture/inference cycles during business hours and scaling to 1,500+ stores
- Automated a daily web-scraping ETL pipeline for regional competitor data; shipped self-serve Power BI benchmarking dashboards contributing to a 4% national operational revenue increase

AI/ML Intern, Sampark Softwares

Jan 2024 - May 2024

- Reduced LLM query latency 30% (15s to 10s) during an OpenAI-to-Azure migration via caching and request-flow optimization

ACADEMIC PROJECTS

Beyond Vector Search - RAG Retrieval Benchmark

2026

Benchmark framework comparing four retrieval strategies (vector RAG, structure-aware PageIndex, a recursive LM agent that executes Python to navigate documents, long-context baseline) for QA over SEC 10-K filings (5 filings, up to 1.2M characters, corpora to 150 documents), scored via LLM-as-judge across OpenAI, Anthropic, and Gemini APIs.

EEG Visual Decoding

Spring 2026

Decoded viewed image categories from 122-channel EEG (13 subjects, 26K trials): 6.6M-parameter transformer with per-subject bias tokens reached 12.7% 20-class accuracy (2.5x chance, +3.2 pts over EEGNet); aligned EEG to CLIP ViT-H-14 embeddings for zero-shot image/caption retrieval (~97% class-aware recall), with attention implemented from scratch.

Adaptive RAG Router (LangGraph + Astra DB)

December 2025 - January 2026

LangGraph Q&A service routing queries between Astra DB vector retrieval and Wikipedia search; Dockerized end-to-end RAG pipeline (MiniLM embeddings, Cassandra vector store) with request logging for routing evaluation.