Sampreeth Avvari

LinkedIn: sampreeth-avvari | **Portfolio:** sampreethavvari.github.io | **Email:** spa9659@nyu.edu | **Phone:** +1 9083077763

Education

Master of Science, Computer Engineering — Aug 2023 - May 2025

- New York University, New York — CGPA: **3.9 / 4**

Bachelor of Technology, Computer Science and Engineering — July 2017 - June 2021

- Jawaharlal Nehru Technological University, India — **Dean's List** | CGPA: **3.92 / 4**

Skills

Languages:	Python, Go, Java, C++, C#, React, Next.js, TypeScript, Node.js, Express.js, JavaScript
Frameworks:	MCP, Angular, Node.js, Microservices
Databases:	MySQL, PostgreSQL, Milvus, DynamoDB, MongoDB, Redis, InfluxDB
Tools:	FlaskAPI, Gin, Spring Boot, Figma, AWS, Kafka, GraphQL, Docker, Kubernetes, N8N, Jenkins, Git, SVN
Libraries:	PyTorch, Transformers, TRL, Huggingface, Unsloth, Accelerate, TensorFlow, OpenCV, Pandas, NumPy, Scikit, Matplotlib, NLTK, WandB
Specialization:	Agentic AI, Unsloth, Model Training

Work Experience

Lead AI Engineer | Hybridge Implants, New York — August 2025 – Present

- Project 1: Multi-modal RAG & Agentic Orchestration for Clinical Consultations
  Architected a **Multi-modal RAG** system via **LangChain** and **Gemini 3 Pro** to analyze clinical consultations, driving a **130% increase** in treatment acceptance and **43% revenue growth**.
  - Engineered a **Hybrid Search** retrieval layer (Keyword + Vector) using **Supabase (pgvector)** and **GraphQL** to query clinical metadata and dental rubrics with high precision via **Python** middleware.
  - Built **Agentic Stateful Chains** to identify ”Clinical Friction Points” through audio-transcript sentiment analysis, optimizing **Context Windows** and chunking to reduce hallucinations by **35%** while maintaining **HIPAA compliance**.
- Project 2: Agentic AI Automation Framework for NPCs
  Engineered an **Agentic AI framework** in **n8n** using **Gemini 3 Pro** and **JavaScript** to automate QA for New Patient Coordinators; boosted intake conversion from **3% to 12%** via real-time coaching.
  - Orchestrated an end-to-end **CoT (Chain of Thought)** grading engine using **n8n** and **pgvector** to analyze 6-phase performance metrics and trigger automated coaching feedback email.
- Project 3: Enterprise AI Infrastructure & Data Pipelines
  Deployed self-hosted **N8n** on **AWS** via **Docker** to unify AI initiatives, cutting costs by **20%** and automating C-Suite analytics pipelines to recover **500+ annual hours** of manual labor.

Software Automation Test Engineer | Shure Incorporated, India — Aug 2021 - Aug 2023

- Built and deployed RESTful APIs with **Flask (Python)** to streamline audio analytics data pipeline for Shure Cloud, leveraging **AWS (DynamoDB, S3, MSK)**. Also optimized data flow and integration, slashed processing time by **20%**
- Developed back-end services and **CI/CD pipelines** using **Python** and **Jenkins**, streamlining build, testing using scalable frameworks using **Python** and **Selenium**, and release workflows, reducing deployment errors by **40%**.

Research Experience

Lead Machine Learning Researcher & Author | New York University, New York — May 2024 – Sep 2025

- **Fine-tuned** LLaMA 3.1 8B with **Reinforcement Learning using Human Feedback(GRPO)** via QLoRA, improving argument persuasiveness scores by **15%** in human evaluation and build causal inferences in argument mining.
- Built **ETL pipelines** to transform Reddit (ChangeMyView) posts into chat-style templates, generating **20k+ samples**.
- Benchmarked models using **BLEU, ROUGE**, and human evaluation on **Qualtrics**, achieving a **50% gain** in model argumentativeness compared to state of the art debate models; tracked 500+ experiments with **WandB**.

Projects

Loan Radar (GitHub) — Jan – May 2025

- Solved real-time latency bottlenecks in financial risk scoring by engineering a distributed MLOps pipeline using **Ray Tune** and **FastAPI**; reduced inference latency to **<200ms**, enabling the system to serve **5k+** predictions with high availability.

RAG-IPL — Jan – May 2025

- Engineered a **Natural-Language-to-SQL** architecture using **LangChain** and **RapidFuzz** to translate complex cricket queries into **SQLite** operations; eliminated LLM hallucinations via **OpenAI embeddings** and a **Streamlit** interface to deliver **100% verifiable** player insights from indexed IPL statistics.

Customer Segmentation & Recommendation System (GitHub) — Feb – May 2024

- Engineered a scalable recommendation pipeline using **Apache Spark** on **Hadoop**, utilizing **MinHash LSH** and **ALS matrix factorization** to process **22M+ records**, achieving a **20% lift in Precision@K** over popularity baselines.