SHRIANSH MANHAS

• shriansh.manhas@gmail.com • +1 404-597-6891 • LinkedIn • GitHub • Portfolio

With a strong foundation from Georgia Tech, I specialize in large language models, retrieval-augmented generation (RAG), and data analytics. I have experience optimizing AI algorithms, building scalable ETL pipelines, and securing data with advanced encryption. My work includes designing RAG systems to improve LLM accuracy and co-authoring research on AI data security. Skilled at clear communication and cross-disciplinary collaboration, I am well-equipped to contribute to teams driving innovation in LLMs and data-centric AI.

EDUCATION

GEORGIA INSTITUTE OF TECHNOLOGY - MSCS	**GPA (current):** 3.75/4.0, Expected graduation - Dec 2025
NATIONAL INSTITUTE OF TECHNOLOGY, DELHI - BSCS	**GPA**: 8.12, Graduated: May 2024
Employment Authorization Document (EAD), I-766	**Expected Jan 2026**

INTERESTS AND SKILLS

Interests: Machine Learning, LLMs, Security, Software Engineering, Cloud Architecture, Data Science, Data Mining
Languages: C, C++, Python, Java, SQL, CUDA; ***Frameworks***: Pandas, Numpy, Tensorflow, Pytorch, Peft, LangChain, Huggingface
Tools: Bash, Spark, SQL, AWS, Azure, Linux, Kubernetes, Jenkins, Docker, Tableau, Power BI, Django, N8N, Blender

WORK/RESEARCH EXPERIENCE

SP TECH (Creating growth-driven digital environments powered by Salesforce) Atlanta, USA
Junior AI Developer May 2025 – Aug 2025

- ***Supervisor***: Mr. Neeraj Parikh
- **Led a team of 3 interns**, completing 100% of sprint deliverables on time, improving release velocity by 30%."
- Developed an **MCP server for a context-aware Slack bot** for summarization, availability checks, smart scheduling, and real-time Q&A—integrated enterprise-grade security (OAuth 2.0, RBAC, channel isolation) with a scalable PostgreSQL infrastructure hosted on Glama. Achieved **p95 latency of 320 ms**, **<$0.002 cost/request**, and **≈92% retrieval accuracy** via **OpenAI-embedding–based semantic search**. Deployed using **FastAPI, Docker, and GitHub Actions CI/CD** for reproducible builds and cloud scalability.
- Cut stale query responses by 40% by implementing **a self-refreshing RAG pipeline**, improving accuracy in financial queries. The system refreshes in real time on updates to the database, ensuring up-to-date responses and reducing stale or inconsistent outputs in financial queries.

SKIT.AI (Conversational voice AI solution provider in the accounts and receivables Industry) Bangalore, India
Software Developer Intern May 2023 – Aug 2023

- ***Supervisor***: Mr. Akshay Deshraj
- Built **ETL functionality** in the Docker pipeline to insert custom datasets for testing the LLM architecture instead of doing the train-test split, thus making debugging easier for MLOps in an Agile environment.
- Adding a fork from the component reduced error detection time by 16% for the NLP system

UNIVERSITY PROJECTS

CYCLIC PRECISION OPTIMIZATION(link) Feb 2025

- GPT-2 Fine-Tuning on SQuAD. How do I increase Fine Tuning efficiency of a model while ensuring it remains light weight?
- Trained the Model using LoRA, frozen weights. Evaluation compares Cyclic Precision to multiple forms of Dynamic Quantization.
- Resulted in a 14.4% increase in F1, 15% increase to EM with the tradeoff being Efficiency drop of 4.7%.

DND-DUNGEON MASTER (link) May 2025

- Designed and developed a fully autonomous AI Dungeon Master for Dungeons & Dragons campaigns.
- Focused on procedural world generation, dynamic story arcs, and real-time interaction.
- Leveraged a **quantized GPT-2 (124M)** fine-tuned with **LoRA**, deployed through a custom **RAG pipeline**, reducing latency by **40%** compared to API calls.
- Integrated a **PostgreSQL vector database** for semantic recall of in-game lore with **sub-300 ms retrieval** time

DETECTING AI-GENERATED SCIENTIFIC PAPERS(IBM) (link) Oct 2023

- Fine-tuned a **BERT-mini (11M parameters)** model to classify AI-generated research papers on the IBM dataset.
- Applied **magnitude pruning** to remove **35% of non-salient weights**, cutting inference latency by **~45%** preserving accuracy.
- Built a full preprocessing pipeline for tokenization, text cleaning, and class balancing to enhance data quality.
- Achieved an **F1-score of 0.9912** and placed **Top 5 globally** in IBM's *Detecting Generated Scientific Papers* - Kaggle.