# Shreya Menon

Frisco, TX | 469-486-0290 | shreyamenon8@gmail.com | LinkedIn | GitHub

## SUMMARY

**Aspiring machine learning engineer** with a strong foundation in **C++ and Python**, and hands-on experience in **optimizing large-scale AI systems** for high performance. Recently developed a Retrieval-Augmented Generation (RAG) pipeline using **LangChain, FAISS, and LLaMa-2 7B**, improving document retrieval accuracy by 35% and reducing inference latency by 30%. Skilled in **model optimization, vector search, and embedding-based retrieval**, with a deep understanding of Transformer architectures and practical exposure to deploying and scaling models. Passionate about efficient compute, inference acceleration, and contributing to cutting-edge systems in LLM and vision-language model optimization.

## EDUCATION

**UNIVERSITY OF TEXAS AT AUSTIN** — **Remote/ Austin, TX**

*Artificial Intelligence and Machine Learning: Post Graduate Program | GPA: 4.0* — *September 2024 – July 2025*

- **Concepts Studied:** Python, Data Science, Data Analysis, AI, SQL, Databases, Google Colab, Anaconda, Generative AI, Adv. Machine Learning, Neural Networks, Natural Language Processing, Computer Visions, Model Deployment

**UNIVERSITY OF TEXAS AT DALLAS** — **Richardson, TX**

*B.S. in Computer Science | GPA: 3.3* — *August 2019 – December 2023*

- **Relevant Courses:** Probability & Statistics, Adv. Data Structures & Algorithmic Analysis, Operating Systems Concepts, Digital Logic & Design, Introduction to Artificial Intelligence, Database Systems, Automata Theory, Comp Sci I & II

## SKILLS

- **Languages: Python, C/C++**, Java, JavaScript, SQL, MIPS
- **Frameworks/Libraries:** PyTorch, **Scikit-Learn**, OpenCV, **NumPy, Pandas**
- **High-Performance/Cloud: AWS (SageMaker, EC2, Lambda), Jupyter Notebook**, Google Colab
- **Tools: GitHub**, Anaconda, Flask, Django, **React Native**
- **Specializations: Model Optimization, NLP**, Large-Scale Inference, Data Pipelines, Deployment

## CERTIFICATIONS

AWS Certified Cloud Practitioner (CCP), **Coursera AI for Everyone, UT Decision Science & AI Immersion Program Certification**, LinkedIn Skill Assessment C++

## WORK EXPERIENCE

**Amazon Web Services** — **Dallas, TX**

*Cloud Support Intern* — *May 2022 – August 2023*

- Optimized cloud architectures by implementing efficient **resource allocation strategies**, reducing latency by **10%** and improving **scalability**.
- Worked to boost **cloud support** and increase **customer satisfaction by 15%** through online customer forums.
- Collaborated with **10+ engineers** to gain hands-on experience in **cloud engineering** and **cloud computing** through intern projects, focusing on **best practices** in **infrastructure design** and **deployment**.
- Learned and applied **15+ cloud technologies**, such as **EC2, S3, RDS, Lambda, CloudFormation, Sagemaker, etc.** to enhance project success.

## PROJECTS

**AI & Data Analytics Projects** — **Austin, TX**

*Personal Projects* — *January 2025 – Present*

- **Processed 25,480+ legal case records** and improved document retrieval efficiency using **advanced NLP techniques**.
- Built an **ML-powered financial data analysis pipeline**, processing **100K+ banking records**, achieving **98.3% accuracy and 100% recall** in predicting loan conversions.
- Optimized **food order analytics for 1,898 records**, enhancing data retrieval and customer behavior analysis.
- **Reduced document processing time by 30%** for visa applications through **automated text analysis** and data cleaning.
- **Improved AI model efficiency by 25%**, cutting computational costs while maintaining high accuracy.

**Retrieval-Augmented Generation (RAG) System** — **Austin, TX**

*Research Project* — *January 2025*

- Developed a RAG pipeline using **LangChain**, **FAISS**, and **LLaMa-2 7B**, improving document retrieval accuracy by **~35%** through semantic search.
- Split over **2,000 text chunks** from financial documents using **recursive character splitting** (chunk size: 2,000 tokens with 200 overlap).
- Integrated **sentence transformer embeddings**, enhancing semantic similarity matching performance by **28%**.
- Reduced inference latency by **30%** through model compression and **optimized vector search with FAISS**.
- Designed the pipeline to support real-time Q&A from dense financial texts, demonstrating scalable, cost-efficient generation.

**Nerveli** — **Richardson, TX**

*Mobile App Developer Assistant* — *January 2023 – May 2023*

- Collaborated in a **5-person team** to build a health-focused mobile app using **React Native**.
- Designed and implemented **5+ custom UI components** for symptom tracking, enhancing user interaction and input flow.
- Integrated a lightweight **AI model** to **predict user diagnoses**, increasing diagnostic accuracy by **15%**.
- Partnered with the UI/UX team to turn wireframes and prototypes into a **fully functional, user-ready application**.
- Implemented secure **user authentication** and state management, ensuring reliable data handling and **personalized user sessions**.