

Tanishk Deo

404-908-9228 | tanishk.deo@gmail.com | [linkedin.com/in/tanishkdeo](https://www.linkedin.com/in/tanishkdeo) | github.com/TanishkDeo

EDUCATION

Georgia Institute of Technology

B.S./M.S. Computer Science, GPA: 4.0

Atlanta, GA

Graduation: December 2026

EXPERIENCE

Apple

May 2025 – August 2025

Machine Learning Engineer Intern

Cupertino, CA

- Research Engineering for Applied LLMs. Fine-tuning/model development, agent pipelines, and evaluation.
- Designed and fine-tuned large language models for retrieval-planning agents, tool-use pipelines, and new domain specific tasks.
- Developed synthetic data generation pipeline through hierarchal data generation and LLM evaluation benchmarks to measure factuality and tool performance.
- Selected to present to VP to demonstrate architecture for new Siri cross-domain features.

Apple

May 2024 – August 2024

Machine Learning Engineer Intern

Seattle, WA

- Engineered a dynamic LLM inference routing framework in Swift and Objective-C for Siri, enabling real-time selection between server/on-device language models based on confidence thresholds and query constraints.
- Leveraged multimodal auto-evaluation agents for data collection in Siri response framework and evaluation in Python pipeline to improve efficiency of multiple internal client teams with AI tooling.
- Released modularized Siri MVVM architecture across two UI versions to isolate ML evaluation components in iOS 18, reducing integration complexity and allowing model experimentation and testing.

Emory Healthcare

May 2022 – Aug 2023

Software Engineer Intern

Atlanta, GA

- Led development of medical applications, deployed in children's hospitals and patient clinics, to assist providers in communication of point-of-care tasks in pediatric hospitals with Swift, Flask, Lambda, S3, MySQL, AWS Medical Transcribe, and Translate.

RESEARCH

LIDAR Lab | ML Research for Robotics under Dr. Ye Zhao

- Developed an Inverse Reinforcement Learning framework integrating SAC, IRL, and Behavioral Cloning to enable reward policy learning from expert demonstrations with TorchRL, PyTorch, ONNX, and NVIDIA Isaac Sim.
- Improved prediction model for socially acceptable path planning for bipedal robots through modifying a Conditional Variable Autoencoder (CVAE) network and developing secondary network for predicting future centers to reduce calculation load.

Ubicomp Health Lab | HCI and ML Research for Health under Dr. Rosa Arriaga

- Designed, developed, and presented a mobile health application research project that aided children in measuring lung health through linear regression model. Awarded IEEE Atlanta Section Award & Augusta University Award at research conference.

PROJECTS

Global KV Cache for LLM Inference | Hugging Face Transformers, NVIDIA GPU Cloud, Docker

- KV Cache optimization for Hugging Face transformer models through attention tensors and a distributed semantic cache (gRPC/Redis) on NVIDIA GPU Cloud, for cache-augmented generation and 10% reduction in inference times.

OpenLend, HackGT | Python, Django REST, Go, Swift, Vertex AI

- Implemented P2P lending platform using Django REST and Golang real-time low-latency arbiter microservice on GCP Cloud Run. Trained and deployed a Random Forest credit risk model achieving 94% accuracy on Vertex AI.

TECHNICAL SKILLS

Languages: Python, Swift, Java, Go, C/C++, SQL, Node.js, Kotlin, PostgreSQL, JavaScript, React.js

Developer Tools: Git, AWS, GCP, Azure, MongoDB, Docker, CI/CD, AWS (S3, DynamoDB, Lambda, EC2), Firebase, Linux, iOS, Android, MongoDB

Libraries: numpy, matplotlib, pandas, Statsmodels, Scikit-Learn, SciPy, CVXPY, pytorch, torchrl, opencv