

Caleb Key

calebkey121@gmail.com ♦ calebkey121.github.io ♦ (903) 910-8324 ♦ Austin, TX

Education

University of Texas at Austin
MS, Artificial Intelligence (Expected May 2027)

August 2025 – Present
Austin, TX

Texas A&M University
BS, Computer Engineering, Cum Laude

August 2018 – May 2022
College Station, TX

Work Experience

Air Force - 90th Cyber Operations Squadron
GS-12 Cyber Capabilities Developer, Active Top Secret Clearance

June 2022 – Sep. 2025
San Antonio, TX

- Contributed to NIPRGPT, DoD's flagship large-language-model platform (**360K+ users**), building and deploying retrieval-augmented generation capabilities in production.
- Designed and shipped "Shared Workspaces" (Sequelize/Flask/React) to share RAG projects and file uploads with link-based access, roles/permissions; reduced duplicate uploads by **80%**.
- Installed, benchmarked, and tuned DGX H100/H200 clusters; increased sustained tokens/sec by **30%** for training/inference, enabling faster iteration on generative model experiments and offline evaluations.
- Performed large-scale **data wrangling and cleaning** for hundreds of thousands of user interactions, reformatting and curating datasets for LLM/ML training pipelines.
- Deployed and maintained a locally hosted LLM platform and translation tool supporting private, on-prem inference in **security-sensitive environments**.
- **Top 1% selectee** for the AF-MIT AI Program, conducting applied ML research in security applications.

Portfolio Projects

Card-Based Strategy Game

- Developed a competitive card strategy game (Godot + Flask); evaluated heuristic baselines and trained reinforcement-learning agents that achieved **15–20% higher** win rates in direct head-to-head play.

Optimization of CPU Cache Eviction using Genetic Algorithms

- Created a C++ program leveraging genetic algorithms to optimize CPU cache eviction, achieving **up to 25% fewer** cache misses in simulation.

Handwritten Digit Recognition

- Built a neural network from scratch achieving **95% accuracy**; deployed alongside PyTorch and TensorFlow models in an interactive, web-based inference demo.

Skills & Interests

- **Skills:** Python, C, C++, SQL, PyTorch, TensorFlow, scikit-learn, Keras, Pandas, NumPy, Docker, HPC
- **Interests:** Evolutionary Algorithms, Computer Graphics, Reinforcement Learning, Computer Hardware