

Alexander Choi

Software Engineer | ML Systems & Backend

choi.alexander.s@gmail.com | (201) 952-8222 | Arlington, Virginia

alexander-choi.com | [LinkedIn](#) | [GitHub](#)

TECHNICAL SKILLS

Languages: Python, C++

ML: PyTorch, Transformers, LLM Pretraining & SFT, Model Evaluation, Hugging Face, Pandas, scikit-learn

Backend & Cloud: REST APIs, AWS (ECS, Fargate, SageMaker, IAM, Elastic Beanstalk, EC2, S3), Docker, Terraform, FastAPI

Software Engineering & Developer Tools: Linux, Git, GitHub Actions, CI/CD, pytest, CMake

EDUCATION

Bachelor of Science in Computer Science, GPA: 3.67

George Mason University

May 2026

Fairfax, Virginia

SELECTED ENGINEERING PROJECTS

Acai Language Model | [GitHub](#) | alexander-choi.com/acai

Jul 2026 – Present

PyTorch, Python, Hugging Face

- Built a 153M-parameter decoder-only transformer from scratch in PyTorch, implementing attention, RoPE positional embeddings, custom training & inference pipelines, and autoregressive generation.
- Reduced median autoregressive inference latency by 73% (17.6s to 4.79s) on a representative workload and up to 95% at long contexts, by implementing KV caching, validated through controlled cached-vs-uncached benchmarks.
- Pretrained on 1B+ FineWeb-Edu tokens, and SFT on 500K+ SmolTalk examples.

Personal Website & Cloud ML Platform | alexander-choi.com

Aug 2026 – Present

AWS (Elastic Beanstalk, SageMaker, ECS, Fargate), Docker, Terraform, FastAPI, GitHub Actions

- Reduced AWS hosting cost by 85% (\$145/month to projected ~\$22/month) by architecting switchable ECS/Fargate & Elastic Beanstalk deployment paths in Terraform, selecting EB for production while retaining Docker based portability.
- Cut median post-submit cold-start inference latency 50.5% (38.3s to 19.0s) by implementing an interaction-triggered inference coordinator that warms AWS SageMaker Serverless endpoint while user composes prompt. Verified across controlled cold-start trials with CloudWatch telemetry.
- Built CI/CD with GitHub Actions to run pytest, build versioned Docker images, and publish to Amazon ECR for reproducible FastAPI deployments.

A Knight's Legacy & AChoiEngine | [GitHub](#)

Dec 2025 – Present

C++, SDL3

- Designed and built a custom C++ game engine and 2D adventure game as sole software developer.
- Developed modular engine architecture, AI path-finding algorithms such as vector fields, and collision physics.
- Profiled and optimized engine bottlenecks with Tracy Profiler, achieving a median 9.2ms/frame with 450+ AI-driven enemy agents.

WORK EXPERIENCE

George Mason Natural Language Processing Group | [EMNLP Publication](#)

Dec 2023 – Jan 2025

Research Assistant

Fairfax, Virginia

- Co-first author of a peer-reviewed paper published to EMNLP (2024) Main Conference.
- Built a Python pipeline integrating OpenAI APIs, Google Sheets API, and OAuth to analyze 10,000+ paragraphs on AI policy and governance.
- Led a Minerva funded (DoD) cross-department research project exploring AI integration into expert human workflows.
- Owned full scope of research project from end-to-end, including Topic Modeling research, study design, proposal to team, data cleaning, API integration, analysis via cosine similarity, and final publication writing.

Handshake AI Solutions

Dec 2025 – Present

Move Fellow

Arlington, Virginia (Remote)

- Maintained consistently high performance as a data annotator for AI labs, ranking top 5% among all fellows.