HAYK STEPANYAN

haykstepanyan02@gmail.com
stepanyanhayk.github.io

EDUCATION

Columbia University New York, NY
Master of Science in Computer Science *Aug 2025 - May 2026*

- Concentration: Machine Learning

Georgia Institute of Technology Atlanta, GA
Bachelor of Science in Computer Science *Aug 2020 - Dec 2023*

- Concentration: Embedded Devices and Artificial Intelligence

EXPERIENCE

Columbia University New York, NY
Graduate Research Assistant (Prof. Matthew McDermott) *Aug 2025 – Present*

- Training a retrieval-augmented pretraining (RAP) model on 10M+ biomedical documents (PubMed, clinical notes) to improve domain-specific LLM generation and factual accuracy.
- Designed and implemented an efficient dense retriever (FAISS), reducing retrieval latency 6x and preserving 90% recall
- Set up and managed multi-node SLURM training clusters (8xL4 GPU nodes).

Google Sunnyvale, CA
Software Engineer *Apr 2024 - Aug 2025*

- ***Research:*** Led engineering of a multilingual dataset of 2M cultural artifacts to benchmark LLMs, streamlining Gemini's ability to evaluate cultural understanding and mitigate cultural biases across 7 dimensions.
- ***Cloud:*** Scaled a high-performance C++ service for multi node GPU testing, enabling 150+ internal users to process 10K+ qualification tests and accelerate scalability of new hardware.
- ***Cloud:*** Created statistical anomaly-detection models with an integrated PostgreSQL-backed data pipeline for TPU/GPU qualification testing, reducing issue triage time from 1 week to under 4 hours.

College of Computing, Georgia Tech Atlanta, GA
Undergraduate Researcher *Aug 2021 - Dec 2023*

- Developed scalable infrastructure for the GTSfM (Global Structure from Motion) framework under *Prof. Frank Dellaert*, establishing large-scale 3D reconstruction experiments.
- Prototyped and optimized a distributed AI compute cluster using Dask, lowering reconstruction runtimes by 5x through parallelized GTSfM computations across multiple machines.

Meta Menlo Park, CA
Software Engineer Intern *May 2022 - July 2022*

- Increased Facebook Group Reels watch time by 0.5% (12 million additional user exposures) by building personalized video generators that tailored content to user interests.
- Built a user interest-based Group Video generator for the Facebook Watch Tab, leveraging user behavior signals to enhance content relevance and engagement.
- Designed and deployed an originality-controlled Group Video generator for In-Feed Recommendations to ensure content diversity and cut down repetition by 15%.

PUBLICATIONS

1. **Stepanyan, H.**, Verma, A., Zaldivar, A., et al. (2025). *Scaling Cultural Resources for Improving Generative Models*. arXiv preprint: arXiv:2510.25167
2. Baid, A., Lambert, J., Driver, T., Krishnan, A., **Stepanyan, H.**, Dellaert, F. (2023). *Distributed Global Structure-from-Motion with a Deep Front-End.* arXiv preprint: arXiv:2311.18801

SKILLS

Python, C++, PyTorch, TensorFlow, Hugging Face Transformers, RAG/RAP, FAISS, Dask, SLURM, CUDA (basics), Distributed Training, LLM Fine-Tuning, Evaluation Pipelines (factuality, grounding, bias), PostgreSQL, Docker, GCP, Azure.