Andrew Koulogeorge

847-204-3005 | andrewkoulogeorge@gmail.com | Personal Website | GitHub | LinkedIn

Education

Carnegie Mellon University — Pittsburgh, PA
Master of Science in Computer Science, GPA: 4.11/4.00 — *Aug. 2024 – May 2026*

- **Coursework:** Convex Optimization, Machine Learning, Probabilistic Graphical Models, Parallel Programming & Computer Architecture, Distributed Systems, Machine Learning Systems

Dartmouth College — Hanover, NH
Bachelor of Arts in Mathematics, Summa Cum Laude, GPA: 3.98/4.00 — *Sept. 2020 – June 2024*

Experience

Pinterest — Sept. 2025 – Dec. 2025
Machine Learning Engineering Intern — *Remote*

- Pre-trained a **20-billion-parameter Foundation Model (FM)** to learn Pin representations for downstream ranking models; designed experiments demonstrating the ineffectiveness of impression data during pre-training
- Applied Parameter-Efficient Fine-Tuning (PEFT) methods to the FM to reduce serving costs via frozen pre-trained embedding tables, achieving neutral offline performance under a fixed training compute budget

AppLovin — May 2025 – Aug. 2025
Applied Scientist Intern — *Palo Alto, CA*

- Researched methods to improve AppLovin's bid prediction for impressions on the MAX auction house
- Designed, trained, and evaluated a hybrid Deep Learning Recommendation Model (DLRM) / Implicit Quantile Network (IQN) on historical auction data and ran A/B tests on live traffic. Deployed to production during internship, **contributing ∼$45M/year to margin with predictions serving 1.2B daily users**

Harpin AI — June 2024 – Aug. 2024
Applied Scientist Intern — *Bend, OR*

- Implemented, trained, and evaluated a Siamese Neural Network for profile similarity, enabling Harpin to bypass expert feature engineering and frictionlessly target customer use cases beyond identity data

Publications

A. Koulogeorge, S. Xie, S. Hassanpour, S. Vosoughi. *Bridging the Faithfulness Gap in Prototypical Models.* Insights Workshop, NAACL 2025 **(Oral Presentation)**.

W. Ma, H. Scheible, B. Wang, G. Veeramachaneni, P. Chowdhary, A. Sun, **A. Koulogeorge**, L. Wang, S. Vosoughi. *Deciphering Stereotypes in Pre-Trained Language Models.* EMNLP 2023.

Select Projects

FetchLess: Optimizing KV Cache Retrieval | *Python, PyTorch* | Code | Poster — Spring 2026

- Profiled FreeKV (ICLR 2026) KV cache retrieval policy to quantify fine-grained retrieval performance costs on LLM reasoning workloads
- Designed and implemented FetchLess retrieval policy, obtaining 2× speedup with no degradation on AIME benchmark

GPU-Accelerated N-Body Simulation | *CUDA, C++* | Code | Report — Spring 2026

- Implemented the Barnes-Hut N-body simulation algorithm from scratch in CUDA
- Profiled kernels with NVIDIA Nsight to quantify arithmetic intensity. Leveraged measurements to transition design from memory bound to compute bound, matching SOTA performance

Technical Skills

Languages: Python, CUDA, C/C++, Java
ML Frameworks: PyTorch, Ray, Megatron-LM, SGLang, NumPy, Pandas, Scikit-learn, XGBoost
Tools & Cloud: Git, Linux, Slurm, Weights & Biases, AWS, GCP