Khush Gupta

(972) 832-5760 | khushgx@gmail.com | linkedin.com/in/khushg | github.com/khushgx

Education

University of Pennsylvania Philadelphia, PA

Jerome Fisher Management and Technology Program (M&T) *May 2026*

- Bachelor of Science in **Computer Science**, **Statistics**
- **Relevant Coursework**: Graduate Deep Learning, Graduate Machine Learning **(Current TA)**, Operating Systems, Computer Systems, Graduate Linear Algebra, Data Structures and Algorithms, Distributed Systems

Experience

Apple | *Machine Learning Engineering Intern* *May 2024 - August 2024*

- Reduced KPI prediction error by >10% by **pretraining** Mamba state space model for inference in **PyTorch**
- **Boosted Ad Review efficiency by 35%** via distributed fine-tuning of Multimodal LLM in Python w/ **LoRA**
- **Optimized AWS model deployments**, achieving a **20% reduction** in processing time with **Apache Spark** and **Hadoop** for distributed data pipelining, and **Apache Airflow** for fault-tolerant job automation

Machine Learning Research Lab - UPenn | *Undergraduate Researcher* *June 2024 - Present*

- Researched optimal placement of meta tokens to enhance **long-context reasoning** in Large Language Models
- Implemented **custom attention** module in PyTorch between meta tokens to maintain consistent attention scores
- Optimizing GPU perf. for custom machine learning workloads in **CUDA**, focusing on memory usage and **FLOPs**

Cypher Tech | *Software Engineering Intern* *August 2023 – May 2024*

- Developed **95% accurate** generative ML model w/ **a16z** using **Python, PyTorch** to predict bank runs
- **Reduced query times by 13%** in full-stack app with **JavaScript**, **PostgreSQL**, **Kafka** for concurrent reads

Advanced Health Academy (AHA) | *Software Engineering Intern* *June 2023 – August 2023*

- Increased **system scalability by 40%** and clientele by **creating REST API** w/ **Node.js, AWS Lambdas**
- Developed blood report interpretation LLM w/**98.4% accuracy** using distillation, Python MPT7B, ChromaDB

Publications

Weak-to-Strong In-Context Optimization of Language Model Reasoning NeurIPS 2024 ATTRIB

- K. Ramji, A. Shah, V. Gaur, K. Gupta. Weak-to-Strong In-Context Optimization of Language Model Reasoning." To appear in *NeurIPS 2024 Workshop on Attributing Model Behavior at Scale (ATTRIB)*, December 2024.

Investigating Language Model Dynamics using Meta-Tokens NeurIPS 2024 ATTRIB

- A. Shah, K. Gupta, K. Ramji, V. Gaur. "Investigating Language Model Dynamics using Meta-Tokens." To appear in *NeurIPS 2024 Workshop on Attributing Model Behavior at Scale (ATTRIB)*, December 2024.

Projects & Activities

CUDAGrad | *CUDA, PyTorch, Python*

- Developed automatic differentiation library in CUDA, achieving **100x speedup** from CPU implementations
- Implemented GPU-accelerated reverse-mode autodiff for tensors binded with PyTorch for deep neural networks

DiscusAI (Open Source) | *Node.js, Python, AWS Lambdas, PyTorch, DynamoDB, Docker, Kubernetes*

- Published open-source ML library for synthetic data generation, using **AWS/DynamoDB** for efficient storage
- Gained **60+ stars, 300+ users** in 2 months w/ features like data refinement, fine-tuning, and text-generation

KAN Transformer | *Python, PyTorch, CUDA, C, C++*

- Developed a decoder transformer from scratch in PyTorch, replacing MLP w/ a **Kolmogorov-Arnold Network**
- Achieved a log loss of **2.1** using a 300,000 parameters network with B-splines, adaptive grids, surpassing nanoGPT

Technical Skills

Languages: Python, Java, C, C++, PostgreSQL, CUDA, Triton, Rust
Frameworks: PyTorch, JAX, Kafka, Spark, Airflow, Tensorflow, DDP,
Developer Tools: Git, Docker, Google Cloud, AWS, VectorDBs, MongoDB, DynamoDB, Kubernetes
Concepts: Deep Learning, Machine Learning, Distributed Systems, Pretraining, Finetuning