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*

- B.S. in Computer Science & Statistics; M.S. in Computer Science
- **Relevant Coursework**: Deep Learning*,†, Machine Learning†, Operating Systems, Advanced Computer Architecture†, Data Structures and Algorithms, Distributed Systems*,†, Linear Algebra† ****TA***; †***Graduate/Ph.D***

Experience

Together AI | *AI Research Intern* *May 2025 - Present*

- Optimizing multi-node/GPU inference workloads w/ model parallelism for serving MoE models in **Python/Rust**
- Implementing scalable scheduler for disaggregated inference engine to serve thousands of concurrent requests
- Developed extreme key–value cache compression algorithms to reduce memory w/ stable perf. in **PyTorch,Triton**

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

- Reduced KPI prediction error by >10% by pretraining 10M parameter transformer/SSM for inference in **PyTorch**
- **Boosted Ad Review efficiency by 35%** via distributed fine-tuning of Multimodal LLM in Python w/ **LoRA**
- Reduced AWS ML pipeline time by **20%** through Spark/EMR distributed data prep. and Airflow orchestration

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
- Pretrained **1B parameter** LLM with **FSDP** custom attention module, beating GPT-3 in GSM8k, MMLU
- Optimizing GPU perf. for large-scale LLM workloads in **PyTorch, vLLM**, focusing on memory usage, power

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

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

Selected Publications

Language Modeling with Learned Meta-Tokens	ICML 2025 LCFM
Weak-to-Strong In-Context Optimization of Language Model Reasoning	NeurIPS 2024 ATTRIB

Projects & Activities

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

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

Distributed Cloud Search Engine | *Java, AWS, Node.js*

- Developed cloud search engine using Java/AWS, w/**fault-tolerant web crawler**, distributed storage system
- Serves queries on over **200K crawled web pages** with an average response time of **900 ms** using custom indexer
- Implemented custom PageRank algo., **search suggestions** query caching, achieving 6x indexing speed

Discus (Open Source) | *Next.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

UNIX Operating System | *C, C++, Unix I/O, ASM, FAT, Docker, Threading*

- Developed a UNIX-like OS in **C, C++**, featuring a **custom PCB** and **priority scheduler** with **0% starvation**
- Used **pthreads** for concurrency, and **custom FAT** file system for efficient process management and file handling

CUDAGrad | *C, 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

Technical Skills

Languages: Python, C, C++, CUDA, Triton, Rust, Javascript, SQL
Frameworks: PyTorch, JAX, vLLM, SGLang, Node.js
Developer Tools: Git, Docker, AWS, Nvidia Dynamo, Kubernetes, CUTLASS, NCCL
Concepts: Reinforcement Learning, Deep Learning, Distributed Systems, GPU Programming, Speculative decoding