

KRITI GOYAL

in [linkedin.com/in/kritigoyal16](https://www.linkedin.com/in/kritigoyal16)

kg16.github.io

📍 Seattle, WA

Experience

Apple

Seattle, WA

Machine Learning Engineer, Systems Intelligence & Machine Learning

Jan 2023 – Present

- Build and maintain core components of Apple Foundation Models, the key framework behind Apple Intelligence experiences—enabling 60+ product teams (Siri, Messages, Photos, Accessibility, etc.) to ship on-device and server-side AI with performant models and infrastructure.
- Developed and publicly released Apple’s custom ML library (tamm) on PyPI, empowering millions of App Store developers to build AI features for their own apps using Apple’s on-device foundation model, strengthening privacy-preserving AI across an ecosystem generating over \$1.3 trillion in billings.
- Led the development of a model publication and evaluation system in the critical pathway for publishing on-device and server models to millions of devices globally, streamlining pre-deployment validation.
- Developed a high-performance, vLLM-based inference engine for serving multi-modal foundation models (over 150 billion parameters); integrated Ray for auto-scaling, load balancing, and fault tolerance to maximize GPU utilization and minimize downtime.
- Accelerated AI development by reducing AI model deployment time from several days of manual work to minutes via single-click execution, saving thousands of engineering hours annually.

Machine Learning Engineer – Intern, ML Platforms, AI/ML

May 2022 – Sept 2022

- Integrated T5 large language model in FlowRunner, enabling developers to use this state-of-the-art model optimized for internal systems to solve various text-to-text problems using a simple YAML config.

Software Engineer, Global Business Intelligence

Jul 2019 – Aug 2021

- Redesigned Apple Retail’s big data analytics pipelines to be schema-less, migrated from Teradata to separate the compute (Spark) & storage (AWS S3) to scale them independently, achieving a 90% reduction in operational costs, saving Apple millions annually.
- Architected systems to handle 100x traffic spikes during Apple’s New Product Introductions and engineered recovery automations that reduced system downtime from over 4 hours to under 1 minute, ensuring uninterrupted performance for revenue-critical dashboards.
- Implemented retrieval-augmented generation (RAG) with contextual embeddings to improve search for AppleCare advisors.

Amazon

Bengaluru, India

Machine Learning Engineer – Intern, Consumer Behavior Analytics

Jan 2019 – Jul 2019

- Partnered with research scientists to productionize Machine Learning Attribution model which guides the apportionment of Amazon’s budget of over \$10 billion on various advertising media.
- Implemented the snapshot creation & featurization modules to process 1 billion+ records per day & scale to handle spikes in traffic during sales and holiday seasons.
- Optimized the machine learning pipeline to reduce the attribution model’s execution time by 75%.

Industry Engagement

- Reviewer for 60+ manuscripts for top-tier AI/ML conferences (ICML, CVPR, ACL, KDD) & IEEE journals.
- IEEE Senior Member: Presenter at IEEE conferences in Seattle.
- Women in Data Science (WiDS) Ambassador: Dedicated to encouraging women in data science.
- Invited Talks: “AI Data Techniques in Robotics”; “MedGemma: Small Foundation Model Transforming Medicine”; “AI Agents: Past, Present, and Future”.

Education

University of Wisconsin-Madison

Madison, WI

Master of Science, Computer Sciences

Aug 2021 – Dec 2022

Coursework: Adv. Deep Learning, Adv. NLP, Machine Learning, Deep Learning for Visual Recognition

Birla Institute of Technology and Science, Pilani

Hyderabad, India

Bachelor of Engineering (Hons.), Computer Science

Jul 2015 – Jan 2019

Coursework: Information Retrieval, Optimization, AI, Data Mining, Data Structures, Fuzzy Logic, Probability & Statistics, Adv. Calculus & Algorithms

Technical Skills

ML & Foundation Models: LLMs, Multimodal Models, RAG, LoRA, PEFT, Quantization, Pruning, Efficient Transformers, Distributed Training & Inference

Frameworks & Tools: PyTorch, JAX, Hugging Face, vLLM, Ray, Spark, Pandas, Git

Cloud & Infra: AWS (S3, EC2, EMR, DynamoDB, SWF)

Languages: Python, Scala, C++, SQL, Bash

Selected Projects

Benchmarking Efficient Transformers for Medical Imaging

Aug 2022 – Dec 2022

- Benchmarked 6 attention mechanisms keeping the same architecture and training settings on the RadImageNet dataset.
- Demonstrated efficient transformer models can even outperform top-1 accuracy of full self-attention ResNet-50 baseline.

Implementation and Analysis of MPI Primitives

Aug 2021 – Dec 2021

- Implemented tree and SAG broadcast from scratch in C++ and performed α - β cost analysis for mesh and ring topologies.
- Outperformed standard broadcast in Gloo by 2.2x and OpenMPI by 2x for 400 KB message, ring topology by SAG broadcast.

Achievements

Distinguished Alumni, BITS Pilani — Featured on university website; invited for alumni talks.

Best Innovation Award, Govt. of India — 1 of 64 nationwide finalists in Smart India Hackathon.

ACM-ICPC, Baylor University — Ranked 2/54 teams; Amritapuri regionals.

Kishore Vaigyanik Protsahan Yojana (KVPY) — Top 0.2%, Rank 246 of 135,000+ candidates.

National Talent Search Examination (NTSE) — Top 0.05%, Rank in top 500 of 800,000+.