

Shreyas Kulkarni

☎ (413) 800-8124 ✉ shreyas.gp269@gmail.com ✎ [LinkedIn](#) ✎ [GitHub](#) 📍 New York City, NY

Applied Researcher focused on Foundation Model inference optimization and training efficiency

WORK EXPERIENCE

Capital One

Principal Associate, Applied Research

July 2026 - Present

- Proposed a trajectory-aware consistency distillation method for diffusion speculative decoding, distilling multi-step token-unmasking trajectories into a single-step drafter and improving acceptance length by 20% over DFlash and DSpark.
- Training masked diffusion language models to serve as speculative drafters across agentic workflows, RL rollouts, and synthetic data generation, achieving a $2.4\times$ inference speedup in agentic workflows.
- Developing Flow-Map based speculative decoding approaches for one-step language modeling in continuous denoising space.

Senior Associate, Applied Research

April 2025 - June 2026

- Optimized inference performance for Agentic AI systems with FP8 quantization and custom EAGLE adapter training, delivering a $1.55\times$ inference speedup over off-the-shelf adapters.
- Productionized EAGLE speculative decoding pipeline using TensorRT-LLM and vLLM backends on Triton Inference Server with tree attention and quantized draft models.

Apple

Machine Learning Intern

Summer 2024

- Trained a ControlNet adapter for diffusion-based synthetic face generation, producing targeted training data with fine-grained control over facial attributes and expressions.

Adobe Research

Research Intern

Summer 2023

- Trained compact diffusion models via model distillation for efficient few-shot personalization.

National University of Singapore

Research Fellow

Sept 2019 - Sept 2020

- Verification systems for Neural Networks using Differential Privacy.

Symantec (acquired by Broadcom)

Software Engineer

July 2018 - Sept 2019

- Stream processing for real-time security vulnerability and malicious app detection using Kafka, Storm, and Spark.

EDUCATION

University of Massachusetts Amherst

2021 - 2024

Master of Science in *Computer Science* | Former Ph.D. candidate; transitioned to industry

Victor Lesser AI Graduate Scholarship

Indian Institute of Technology (IIT), Guwahati

2014 - 2018

Bachelor of Technology in *Engineering Physics*, Minor in *Mathematics*

OPEN SOURCE CONTRIBUTIONS

vLLM (Contributor)

- Added lookahead-slot block allocation to prevent DFlash prefix cache corruption in speculative decoding [PR #42971](#).
- Added EAGLE draft model quantization support for speculative decoding [PR #28435](#).

PATENTS

- Progressive Marginal-Consistency Distillation of Block-Diffusion Drafters for Multi-Agent LLM Acceleration
- Multi-Agent Debate-Driven Data Curation, Verification and Training Pipeline

PUBLICATIONS AND PREPRINTS

- Towards Scalable Customization and Deployment of Multi-Agent Systems for Enterprise Applications
Shreyas Kulkarni, Paresh Dashore, Uttam Gurram, Nadia Bathaee, Kartik Balasubramaniam, Genta Indra Winata, Sambit Sahu, Shi-Xiong Zhang
Preprint. Under Review
- Can Large Language Models Understand, Reason About, and Generate Code-Switched Text?
Genta Indra Winata, David Anugraha, ..., **Shreyas Kulkarni**, ..., Adril Putra Merin, Emmanuele Chersoni
Preprint. Under Review
- Reconstruction Attacks on Interaction-based Federated Learning with Adversarial Data Manipulation
Dzung Pham, **Shreyas Kulkarni**, Amir Houmansadr
Networks and Distributed Systems Security (NDSS), 2025
- Adaptive Client Selection in Federated Learning using Reinforcement Learning
Shreyas Kulkarni, Dhawal Gupta, Virat Shejwalkar, Amir Houmansadr
Master's Thesis, UMass Amherst
- Improving Out-of-distribution Detection in Neural Networks via t -distribution based Generative Classifier
Shreyas Kulkarni, Virat Shejwalkar, Amir Houmansadr
Preprint.
- Determining the Optimal Fuzzifier Range for Alpha-Planes of General Type-2 Fuzzy Sets
Shreyas Kulkarni, Rishabh Agrawal, Frank Rhee
IEEE International Conference on Fuzzy Systems 2018
- Phase Transitions in Erdos-Renyi Random Graph
Shreyas Kulkarni, Sameer Kamal
Bachelor's Thesis, IIT Guwahati
- Fractals, Newton's method and Julia set
Shreyas Kulkarni, Shubham Ugare, Aneesh Barthakur, Akash Rakheja
SIAM Chapter 2017

TECHNICAL SKILLS

Programming Languages & Libraries	Python, NumPy, JAX, Triton, pandas, scikit-learn
Deep Learning Frameworks	PyTorch, Accelerate, peft, transformers
Training & Distributed Systems	Megatron-LM, FSDP, DeepSpeed
Inference & Serving	vLLM, TensorRT-LLM, SGLang, Triton server, FastAPI, Ray Serve, gRPC

ACADEMIC SERVICE

- Primary/secondary reviewer for ACL, EMNLP, IEEE S&P, NDSS, CCS.
- Served on the MS/PhD Graduate Admissions Review Committee at the College of Information and Computer Sciences, UMass Amherst.
- Mentored in the PhD Applicant Support Program at the College of Information and Computer Sciences, UMass Amherst.

TEACHING

- Teaching Assistant, CS689 Graduate Machine Learning, UMass Amherst (Fall 2021, Fall 2022).
- Teaching Assistant, CS660 Advanced Information Assurance, UMass Amherst (Fall 2023).
- Teaching Assistant, CS240 Reasoning Under Uncertainty, UMass Amherst (Spring 2022).
- Teaching Assistant, CS250 Introduction To Computation, UMass Amherst (Fall 2024).