

Tanishk Singh

Phoenix, AZ • 520-270-0849 • tanishksingh@arizona.edu • linkedin.com/in/Tanishk-Singh • github.com/tanishksingh • tanishksingh.com

MS Data Science candidate with 2 years of experience, currently specializing in ML systems efficiency, production LLM/RAG pipelines. Proficient in PyTorch, TensorFlow, PySpark and container orchestration, with hands-on expertise in building end-to-end agentic pipelines and parallel GPU optimization. When I'm not debugging convergence curves, you'll probably find me on a climbing wall.

SKILLS

Languages: C++/C, Java, Python, R, SQL

ML/DL Frameworks: PyTorch, TensorFlow, CUDA, scikit-learn, Pandas, NumPy

ML/AI Techniques: Deep Learning, Federated Learning, NLP, LLM, RAG, LoRA/PEFT

Backend & Infra: SpringBoot, Docker, Kafka, PySpark, MongoDB, Postgres, pgvector, HPC, MLOps, A/B Testing

WORK EXPERIENCE

DASS Lab, University of Arizona • Tucson, USA

01/2026 – Present

Volunteer Graduate Researcher • Part-time

- Conducting research in AI efficiency, Federated Learning (edge ML) under Prof. Jyotikrishna Dass.
- Replicated and stress-tested published federated fine-tuning setups, ran fine-tuning and convergence experiments (FFA-LoRA, FLoRA, FedIT etc.) under homogeneous/heterogeneous federated settings in PyTorch on HPC clusters
- Helped in improving efficiency by developing aggregation methods, achieving 30 % improvement in performance and 70 % in communication overhead.

Newgen Software • Noida, India

07/2022 – 01/2024

Application Engineer

- Developed and integrated Spring Boot backend services and RESTful/SOAP APIs within Newgen's iBPS Retail Loan Origination System (RLOS) for a US credit-union client, enabling automated credential/KYC and credit-bureau checks.
- Worked on building and managing the CFCU database using PostgreSQL and developed server-rendered JSP screens wired to Spring/Servlet controllers and iBPS work-item attributes, streamlining the verification step for loan processors.

PROJECTS

McGraw Hill - Eller Immersion Learning Project

05/2026 – 08/2026

- Built a tool for instructors to help them plan classroom role-play activities by engineering an 8-step agentic LLM pipeline using a WorkflowStep protocol.
- Tackled a core AI-trust problem, models sounding confident while getting facts or math wrong, with a multi-layer review system combining automated checks, an independent AI reviewer, and a human-in-the-loop checkpoint for mid-generation review.
- Grounded every generated answer in the actual McGraw Hill textbook via a Postgres/pgvector retrieval system, and shipped the fully documented, production-ready version.

Real-Time Tourist Foot-Traffic Forecasting Pipeline (AWS)

- Built a real-time foot-traffic forecasting system for Tallinn's Old Town, streamed pedestrian-sensor and weather data via Kafka, forecast traffic 8 hours ahead with MSTL/AutoARIMA, and served it on a live dashboard panel/Bokeh.
- Re-architected the pipeline onto serverless AWS, lambda + EventBridge ingestion, RDS Postgres, and Fargate for retraining/dashboard hosting, building an running on demand infrastructure.

Hardware-Aware Selective Scan CUDA Kernel Implementation for Mamba State Space Models

02/2026 – 04/2026

- Identified and tackled the core bottleneck in Mamba state space models: a selective-scan recurrence that's inherently sequential and resists standard GPU parallelization.
- Built and progressively optimized five CUDA kernels, culminating in a Kogge-Stone parallel-prefix-scan implementation using warp-shuffle intrinsics.
- Pushed inference from 246ms down to 1.6ms per forward pass, a 156x speedup over the CPU baseline on an A100, and backed it with ablation studies showing shared-memory tiling alone cut global memory traffic by 40%

VoiceBridge - HackArizona (Hackathon)

03/2025

- Developed a multilingual speech-to-speech translation app using OpenAI Whisper ASR, Meta NLLB-200, and MMS-TTS for real-time, AI-driven communication into a low-latency, chunk-based pipeline (supporting 12+ languages)
- Shipped a production-style deliverable (live demo, hold-to-speak UX) using Flask + Socket.IO over WebSocket rooms.

EDUCATION

Master Of Science in Data Science

01/2025 – 12/2026

University of Arizona • Tucson, USA • GPA: 3.9

Bachelor of Technology in Electronics and Communications

07/2022

Dr. Akhilesh Das Gupta Institute of Technology & Management, GGSIPU • New Delhi, India • GPA: 3.22

LEADERSHIP

Basic Needs Centre (ASUA), Drive Chair • Tucson, USA

10/2025 – 05/2026

- Built partnerships with 10+ campus organizations within 6 months, boosting food-donation community support by 200%.