

Sarvesh Tiku

Atlanta, GA | sarveshtiku@gmail.com | (303)-827-9659 | linkedin.com/in/sarveshtiku | github.com/sarvesh-tiku

Education

Georgia Institute of Technology

Expected Graduation: May 2027

Bachelor of Science in Computer Science, Threads in Intelligence and People

GPA: 4.00/4.00

Relevant Coursework: Systems & Networks, Graduate Machine Learning, Computer Organization & Programming, Introduction to Artificial Intelligence, Database Systems, Linear Algebra, Data Structures & Algorithms, Objects & Design, Robotics & Perception

Leadership: AI/ML Workshop Lead, Big Data Big Impact; Head of Market Insights, Trading at Georgia Tech; Startup Exchange Board

Experience

Software Engineering Intern, Salesforce – San Francisco, CA

May 2026 – Aug 2026

- Reduced Sev-1 time-to-diagnosis by **100x** (7+ **engineer-hours** to **~4 minutes**) by productionizing a Root Cause Analysis agent over **20K+ production alerts** with evaluation and reliability infrastructure for production use across **20+ live Sev-1 incidents**.
- Built a ground-truth pipeline for **100K+ customer incidents** using SQL/Trino and Splunk APIs to capture and persist telemetry, incident context, and evaluation artifacts for new cases while reconstructing historical cases after source telemetry expired.
- Developed RCA replay and evaluation harness over persisted ground truth and execution traces, re-running **500+ historical incidents** per agent version to compare model behavior, surface regressions, and cut evaluation turnaround by **~90%**.

Machine Learning Research Assistant, Financial Services Innovation Lab – Atlanta, GA

Aug 2025 – Jan 2026

- Contributed to FinForge, a benchmark generation pipeline for evaluating LLMs across **11+ subdomains**. Developed data curation workflows over **100,000+** financial documents, using automated extraction and structured Q&A generation from verified sources.
- Generated **10,000** Q&A pairs, filtered them to **5,000** benchmark items, and validated a **500**-sample expert review that revealed a 30-point gap between automated and human approval; accepted to the **AAAI'26 Workshop on Agentic AI in Financial Services**

Software Engineering Intern, Salesforce – San Francisco, CA

May 2025 – Aug 2025

- Built and deployed a full-stack, multi-tenant Agentforce sandbox that reproduced production traffic, tenant isolation, and enterprise usage patterns, simulating peak-traffic environments to scale-test agents across **10K+ deployments** before customer rollout.
- Scaled the platform to **100K+ concurrent requests** using Redis caching, load balancing, rate limiting, and Kubernetes autoscaling, sustaining peak traffic while improving **p95 latency by ~28%** and reducing pre-deployment failures by **35%**.
- Built automated load-testing and release-validation infrastructure across **113K+ test runs**, flagging **13K+ degraded or failed runs** caused by latency, capacity, and reliability regressions and surfacing issues early enough for remediation before production rollout.

Computational Biology Research Assistant, Weigel Research Lab – Atlanta, GA

Jan 2025 – May 2025

- Built and benchmarked R- and OpenAI API-based research pipelines across **500+ academic papers**, evaluating zotero agent performance on statistical analysis, semantic retrieval, knowledge extraction, and automated literature synthesis.
- Received Georgia Tech's **President's Undergraduate Research Award (PURA)** to extend the work. Benchmarking framework accepted for publication at the 9th IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR).

Projects

Belief State - Temporal Belief in Financial Agents github.com/sarvesh-tiku/Belief-State

Jan 2026 – May 2026

- Built a **mechanistic-interpretability framework** to identify and causally test latent belief representations in language models using residual-stream hooks, activation patching, causal steering, and mediation analysis.
- Evaluated GPT-2 and **Qwen2.5-1.5B-Instruct**, localizing a causal belief direction across **6 deep layers** and validating steering against random and orthogonal controls; published at the **2nd Workshop on Advances in Financial AI, ICLR 2026**.

CachePilot: KV-Cache-Aware LLM Inference Gateway github.com/sarvesh-tiku/CachePilot

May 2026 – Aug 2026

- Built an **OpenAI-compatible LLM inference gateway** with KV-cache-aware routing, prefix fingerprinting, LRU cache tracking, load-aware scheduling, and Prometheus observability across simulated multi-worker deployments.
- Benchmarked round-robin, least-loaded, and KV-aware routing across **7 workload patterns**, improving simulated throughput by **16–25%**, reducing TTFT by up to **31%**; validated cache-residency estimates on Qwen w/ **204 ms vs. 400 ms** cache-hit/miss TTFT.

Quorum: Consensus-Gated Agent Reliability github.com/sarvesh-tiku/Quorum

May 2026 – June 2026

- Placed 3rd at **Inference-Time Compute Hackathon**; built **quorum-py**, a framework-agnostic Python runtime that gates irreversible agent actions behind parallel read-only juror consensus, with red-flag filtering, rollback, and recovery from failed actions.
- Integrated adapters for **Claude Agent SDK, LangGraph, and OpenAI Agents SDK**, stress-testing consensus gating and rollback under injected agent failures with **25+ automated tests** and publishing the runtime to TestPyPI.

Skills & Certifications

Languages: Python, Java, C++, C, JavaScript, TypeScript, SQL, Bash, HTML/CSS | **Frameworks and Libraries:** React, Node.js, Flask, FastAPI, PyTorch, TensorFlow, REST APIs | **Systems & Tools:** Linux, Git, Docker, PostgreSQL, MySQL, Google Cloud, CUDA

Certifications: AWS Certified Solutions Architect, NVIDIA-Certified AI Infrastructure & Operations Associate, Docker Associate

Relevant Publications

- Belief State: Interpreting Temporal Belief Dynamics in Agentic Financial Systems** 2nd Workshop on Advances in Financial AI Workshop: Towards Agentic and Responsible Systems, **ICLR - International Conference on Learning Representations (2026)**
- Mitigation of User-Prompt Bias in Large Language Models: A Natural Language Processing & Deep Learning Based Framework.** 3rd IEEE International Conference on Computing and Machine Intelligence (4 citations, 2025).