

# Shreyas Pradeepkumar Khandale

US Work Authorized | Open to relocate | 607-774-7417 | [skhandale@binghamton.edu](mailto:skhandale@binghamton.edu) | [linkedin](#) | [github](#)

## EDUCATION

### State University of New York Binghamton

Dec 2026

*Master of Science in Computer Science, Machine Learning and Artificial Intelligence* **Cumulative GPA: 3.60/4.00**

## EXPERIENCE

### AI Systems Integrator Intern

Feb 2026 – Present

*Zoo Media Inc.*

*(remote) Severance, CO*

- Architect a multi-tenant Voice AI Sales Agent platform, as measured by recognition as a **3-time “Outstanding Team Lead”** winner, by building an autonomous pipeline from CRM webhooks to AI-powered outbound calling
- Restore a non-functional codebase with zero working routes across roughly 45 files to a fully operational platform, as measured by a **90.9% test pass rate**, by debugging 79 runtime errors and resolving 40 missing dependencies
- Lead a cross-functional team of 9 developers, as measured by a 6-workstream production roadmap across AI/LLM, voice (ElevenLabs), telephony (RingCentral), and cloud, with branching strategy and code-review workflows

### Computing Services Admin

Jan 2025 – Present

*Information Technology Services, Binghamton University*

*Binghamton, NY*

- Administer and improve a production AI-powered conversational system, as measured by reliable service for **roughly 20,000 students and 4,000 staff**, by applying SRE practices and automated health checks
- Engineer a continuous NLP retraining and evaluation pipeline, as measured by a **sustained ticket-resolution rate of nearly 90%**, by analyzing resolved tickets, automated testing, and user-feedback loops
- Maintain **24/7 availability** of campus network and AI services, as measured by real-time PRTG monitoring of routing and repeater infrastructure, by triaging and escalating incidents to Network Engineering

## PROJECT EXPERIENCE

### arXiv Paper Curator - Agentic RAG | *Python, LangGraph, LLMs, BM25+KNN, Redis, FastAPI, Docker, AWS*

- Cut repeated-query latency **150-400x** (16s to sub-100ms), as measured by Langfuse traces, by adding SHA-256 Redis caching over a hybrid BM25 + 1024-dim vector pipeline fused via RRF
- Shrank the Docker image **8GB (70%)** for a 174s cloud build, as measured by AWS EC2 deployment benchmarks, by diagnosing CUDA wheel bloat and pinning CPU-only PyTorch wheels via uv
- Shipped a public agentic RAG system at **\$0**, as measured by 10 live REST endpoints, by orchestrating a 6-node LangGraph agent across a 7-service Docker stack on AWS EC2

### Cashflow Management | *Python, FastAPI, MongoDB, XGBoost, Pinecone, Gemini, Docker*

- Built a 13-week cashflow forecasting backend, as measured by **\$1,450/week MAE** on a held-out backtest, by engineering a FastAPI/MongoDB pipeline with category-level XGBoost and payables deferral
- Added a grounded Q&A layer over forecasts, as measured by **86.7% (26/30)** on an eval set, by building a RAG pipeline indexing forecast packages in Pinecone and answering via Gemini
- Delivered an auditable financial QC system, as measured by enforced invariants (non-negative balance; paid + deferred = outflow), by building a Streamlit/Plotly dashboard over REST endpoints

## TECHNICAL SKILLS

**Languages:** Python, SQL, C++, TypeScript, JavaScript, Bash

**ML & DL:** PyTorch, scikit-learn, XGBoost, LightGBM, CatBoost, SHAP, Transformers, CNNs, LSTMs

**GenAI & LLMs:** RAG, Agentic AI (LangGraph, LangChain), Prompt Engineering, Hybrid Search (BM25+KNN, RRF), Vector Embeddings, Langfuse, Ollama

**Data & Backend:** Pandas, NumPy, FastAPI, PostgreSQL, MongoDB, OpenSearch, Pinecone, Redis, Apache Airflow

**Cloud & Tools:** AWS (EC2/EBS), Docker, Git, GitHub Actions (CI/CD), Linux, Claude Code, Cursor

**Certifications:** NVIDIA DLI CUDA C/C++, Google Advanced Data Analytics, IBM Python for Data Science

## LEADERSHIP EXPERIENCE

**State University of New York Binghamton, Graduate Research Assistant**

Jan 2025 – May 2026

**AISSMS College of Engineering Pune, Teaching Assistant**

June 2022 – June 2024

**Codigo Madrid (Competitive Programming Club), Technical Lead**

May 2022 – June 2024