

Shishir Kallapur

Boston, MA, 02119 | (582)201-8592 | kallapur.shi@northeastern.edu | github.com/shishirkallapur | linkedin.com/in/shishirkallapur

Professional Summary

AI-focused Software Engineer and aspiring Machine Learning Engineer with 2+ years of experience delivering full-stack and intelligent solutions. Skilled in machine learning, reinforcement learning, NLP, transformers, large language models (LLMs), model fine tuning, prompt engineering, vector database integration and cloud-native development. Experienced in building GenAI and RAG pipelines for production-ready solutions. Passionate about translating cutting-edge AI research into robust, production-grade systems that deliver measurable business value.

Education

Northeastern University , Boston, USA	Sept. 2023 – May 2025
Master of Science in Artificial Intelligence	GPA: 3.91
Khoury College of Computer Sciences	
Courses: Foundations of AI, Programming Design Paradigm, Algorithms, Machine Learning, Reinforcement Learning, Natural Language Processing, Advanced ML, AI for HCI	
The National Institute of Engineering , Mysore, India	Aug. 2017 – Aug. 2021
Bachelor of Engineering in Computer Science and Engineering	GPA: 3.57

Technical Knowledge

AI/ML:	LLMs, GenAI, RAG, NLP, Transformers, Model Fine-Tuning, Prompt Engineering, MLOps ML System Design, Reinforcement Learning, Collaborative Filtering, Matrix Factorization
Frameworks:	PyTorch, TensorFlow, Scikit-Learn, NumPy, Pandas, Matplotlib, FAISS, MLflow, OpenCV
Backend:	FastAPI, Streamlit, Spring
Cloud/Tools:	AWS, Docker, Docker Compose, Git, JIRA, ServiceNow, Pinecone, Gspread
Databases:	MySQL, MongoDB, SQLite
Languages:	Python, SQL, Java, JavaScript, C++, C
Certifications:	AWS Cloud Practitioner

Work Experience

Amplifier Security	May 2024 – August 2024
AI Product Intern	
• Spearheaded a comprehensive benchmarking initiative for GPT models, significantly enhancing Ampy's response accuracy, speed and overall performance.	
• Implemented guardrails and prompts that boosted topical relevance by 35%, reducing hallucinations.	
• Automated response evaluation with custom Python scripts, improving testing speed by 3x.	
• Implemented a Retrieval-Augmented Generation (RAG) prototype with LangChain using Pinecone as Vector DB, enabling contextual replies from proprietary unstructured data.	
JP Morgan Chase & Co. , Bangalore, India	Sept. 2021 – August 2023
Software Engineer	
• Overhauled ServiceNow Knowledge module, enhancing request resolution speed by 20%.	
• Integrated JIRA with ServiceNow to automate SDLC tracking and reporting, incorporating CI/CD automation best practices and reducing manual effort by 40%.	
• Delivered 5 reusable UI macros to streamline HR documentation workflows and improved the team's document update efficiency by 40%.	
• Introduced and deployed catalog automation features, reducing request handling time by 30%.	

Projects

Movie Recommendation System	Dec. 2025 - Jan. 2026
• Engineered an implicit feedback recommendation system using ALS matrix factorization and cosine similarity-based collaborative filtering, with FAISS indexing for sub-100ms item similarity queries.	
• Architected a complete ML pipeline: data ingestion, time-based train/val/test splitting, feature engineering, MLflow-tracked training, model export, and FastAPI serving with cold-start fallback handling.	
• Designed a multi-service application with Streamlit frontend, SQLite request logging, and a monitoring dashboard tracking traffic, latency percentiles, and recommendation quality metrics.	
Local Document-Powered RAG Chatbot	May 2025 – June 2025
• Developed a local RAG-based chatbot that allows users to upload and conversational querying using Streamlit and Ollama.	
• Designed custom chunking and re-ranking pipelines to boost retrieval accuracy and relevance.	
• Integrated Chroma vector DB for fast similarity search and persistent conversational context.	
Relating Physical Activity to Problematic Internet Use in Youths	Sept. 2024 – Dec. 2024
• Developed a ML pipeline to identify at-risk youths, leveraging physical activity data to promote digital welfare.	
• Used transformer autoencoders and Random Forest based imputers to preprocess noisy, incomplete data.	
• Achieved 72% mean QWK score using a voting classifier that combined XGBoost, LightGBM, and CatBoost, effectively addressing dataset complexity and imbalance.	