Erina Karati

763-900-0570 | erina.karati@gmail.com | linkedin.com/in/ekarati | github.com/ekarati | Bellevue, WA

Education

University of Minnesota, Twin Cities | M.S. Computer Science, **4.0/4.0 GPA** *Minneapolis, MN | Aug 2024 – May 2026*

Techno India University | B.Tech Computer Science, **9.11/10 GPA** *West Bengal, India | Aug 2017 – May 2021*

Industry Experience

LyftBridge Innovation | GenAI Engineer *Minneapolis, MN | Jul 2025 – Aug 2025*

- Built a cloud-hosted agentic assistant using **FastAPI**, **Google Cloud Run**, **ElevenLabs** real-time voice agents, and **Zapier MCP** to automate Google Workspace workflows; reduced response latency by **70%** versus prior local deployment.
- Designed modular tool-invocation architecture with scoped Google Workspace permissions, structured logging, retry/fallback guardrails, and API failure tracing across voice handoffs and MCP/Zapier workflows; **reduced failed tool executions by 35%** and improved real-time agent debuggability.

Supercell | AI Engineer *San Francisco, CA | Apr 2025 – Jun 2025*

- Developed a modular **multi-agent LLM framework** using **LLaMA 3**, **Agentic RAG**, **FastAPI**, **MCP**, and **FAISS**, enabling planning, collaboration, memory, personality, and emotion systems across local/cloud endpoints for game AI.
- Built FAISS-based per-agent vector memory supporting **1M+ semantic embeddings**; evaluated retrieval quality, memory consistency, and latency-recall-cost tradeoffs for cross-agent grounding across multi-agent simulations and endpoint behavior.
- Evaluated social-agent memory/trust modeling; research accepted to **WiML Symposium @ ICML 2026** for AI systems.

Microsoft | Software Engineer *Bangalore, India | Sep 2021 – Aug 2024*

- Served in a **customer-facing, forward-deployed engineering role** for **200+ enterprise and Fortune 500 clients** across finance, healthcare, government, automotive, and consumer goods, resolving high-severity Windows, Azure, networking, identity, and security failures at scale.
- **Scoped ambiguous customer failures** and led RCA across TCP/IP, Wi-Fi, VPN, SDN, firewalls, Active Directory, and Azure using packet captures, protocol traces, telemetry, and vendor evidence to drive Sev-A/Sev-B remediation.
- Converted recurring field failures into reusable **PowerShell automation** and cross-vendor remediation playbooks across Cisco, Palo Alto, VMware, Juniper, and Aruba; recognized as Wi-Fi SME, partnered with global engineering teams on cross-stack escalations, and earned **Global CSS Impact Award** ($1K) and **Star ACE Team** (Top 5%) awards across FY22/FY23.

Selected Projects

MinneDigest: AI-Powered News & Podcast Platform | *FastAPI, ChatGPT, LLaMA 3, Ollama, ElevenLabs, Whisper*

- Awarded **$10K AI x Journalism grant**; built and deployed scalable FastAPI backend for automated newsletter/podcast generation, multilingual content workflows, and white-labeled hosting for local outlets with publishing pipeline support.
- Implemented content-generation, podcast narration, and outlet-specific deployment flows with LLaMA/Ollama experimentation, Whisper transcription, ElevenLabs audio, and FastAPI orchestration for multi-outlet publishing support.

On-Device Health AI App | *Swift, Gemma, Core ML, MLX, VisionKit, PDFKit, SQLite, CryptoKit*

- Won **$1K Vanta Best Security Award**; built encrypted iOS Health Memory RAG with biometric access, on-device storage, hybrid SQL + semantic retrieval, and token-budgeted context over private health documents without server upload.

Agentic Code Repair: Multi-Agent Feedback Routing | *Qwen2.5, Python, Hugging Face*

- Built a planner/executor/critic code repair system with feedback routing and failure-mode analysis; achieved **78% accuracy** on 400+ LeetCode problems, matching 32B performance with a 7B architecture while reducing inference cost by **75%**.

Academic Experience

University of Minnesota, Twin Cities | Graduate RA & TA *Minneapolis, MN | Jan 2025 – May 2026*

- Built and evaluated NLP/RAG pipelines over **500K+ research abstracts** using **LangChain**, Word2Vec, and semantic clustering; performed error analysis to improve trend-detection accuracy to **85%** for scientific trend analysis at scale.
- Developed **GenAI automations** and web-based tools for educators and researchers, translating applied AI workflows into reusable demos, tutorials, and instructional assets for ML, Responsible AI, and GenAI course labs and assignments.
- Created GenAI tutorials and demos for **Azure AI Foundry**, **Copilot Studio**, AutoGen, agentic workflows, and multimodal RAG; taught/evaluated **120+ students** in ML and Responsible AI coursework across labs, lectures, and assessments.

Technical Skills

AI / LLM: LLMs, Generative AI, Agentic AI, Multi-Agent Systems, RAG, Tool Calling, MCP, Prompting, Context Engineering, Embeddings, Vector Search, LLM Evaluation, Multimodal AI

Frameworks / Cloud: Python, Java, C++, SQL, Swift, FastAPI, LangChain, LangGraph, AutoGen, Hugging Face, PyTorch, scikit-learn, FAISS, ChromaDB, MLflow, Azure AI Foundry, Google Cloud Run, AWS, Docker, Linux

Systems / Security: Distributed Systems, REST APIs, PostgreSQL, SQLite, TCP/IP, VPN, Firewalls, Active Directory, Wireshark, PowerShell, RCA, Production Debugging, Core ML, MLX, CryptoKit