YAVAR KHAN

AI/ML ENGINEER
• San Jose, CA • (669)-339-8555 • yavarkhan1997@gmail.com
• linkedin.com/in/yavar-khan29 • github.com/yavar29 • yavarkhan.me

PROFESSIONAL SUMMARY

*AI/ML Engineer with 2.8 years of production engineering at **Accenture** and a **Master's in CS (AI/ML)** from SUNY Buffalo. Specialized in building production-grade **multi-agent systems**, **RAG pipelines**, and **LLM evaluation** frameworks deployed on **AWS** and **GCP** with **Terraform**, **Redis**, and **CI/CD**. Proven ability to ship and scale mission-critical systems across **globally distributed teams**.*

PROJECTS

WEATHERWISE.AI

- Architected a conversational weather intelligence system using **LangGraph** and **Vaisala Xweather MCP**, grounding LLM responses in NOAA, NASA, and NCA5 research data.
- Developed a **Corrective RAG pipeline** with Claude Vision chart grounding, embedding 429+ vectors from NOAA / NASA / NCA5 into **Pinecone** for scientifically grounded responses.
- Automated stateful interaction monitoring and evaluation via **LLM-as-Judge**, **LangSmith tracing**, and session checkpointing.
- Deployed via **Terraform** to **GCP Cloud Run** with **Redis** caching, **Anthropic prompt caching**, and real-time **WebSocket streaming**.

DEEP RESEARCH PRO - MULTI-AGENT ASYNC RESEARCH SYSTEM GITHUB

- Designed an asynchronous **multi-agent research engine** using the **OpenAI Agents SDK**, featuring an autonomous query planner and multi-step verification for cited synthesis.
- Optimized system throughput by **50x** using **bounded-parallelism** and concurrent search-summarization threads with system-level guardrails.
- Engineered a **dual-layer caching** strategy (LRU + SQLite, 24h TTL) with context-aware bypassing to reduce API overhead and latency.
- Ensured production reliability through **Pydantic schema validation**, automated **retry/fallback logic**, and real-time **observability logging**.

TRANSFORMER-BASED SENTIMENT ANALYSIS GITHUB

- Developed a **Transformer NLP architecture** from scratch in **PyTorch** with **multi-head self-attention** and custom positional encodings.
- Achieved **91.8% accuracy** (**0.92 F1-score**) on a **560K review** dataset using Dropout and Early Stopping for model regularization.
- Built a scalable **NLP preprocessing pipeline** using **spaCy** and **Gensim** for tokenization, embedding generation, and large-text handling.

PROFESSIONAL EXPERIENCE

ACCENTURE, PUNE, INDIA DEC 2020 - AUG 2023

Software Engineering Analyst (Promoted from Associate, 2022)

- Deployed **50+ production Java APIs** on Apigee and AWS, managing high-concurrency traffic for enterprise-scale consumer applications.
- **Reduced API errors from 25% to under 1%** within 2 months by debugging AWS Lambda/ElasticSearch logs and retiring faulty APIs.
- Built **ElasticSearch/CloudWatch dashboards** with anomaly detection alerting, reducing mean-time-to-detect from hours to minutes.
- **Collaborated** with cross-functional **global teams** on API performance tuning and secure governance (IAM, quotas, rate limits).
- Resolved **100+ client-reported bugs** using JIRA workflows, ElasticSearch log analysis, and Apigee trace sessions validated via Postman.
- **Mentored and led a 6-member team**, driving Agile/Scrum, CI/CD, and TDD practices to achieve on-time delivery for 12+ projects.

EDUCATION

UNIVERSITY AT BUFFALO, THE STATE UNIVERSITY OF NEW YORK (SUNY) AUG 2024 - DEC 2025

Master of Science - Computer Science (AI/ML Track) | GPA: 3.87

Relevant Coursework: Deep Learning, Machine Learning, Computer Vision & Image Processing, Pattern Recognition, Data Intensive Computing.

AMITY UNIVERSITY, INDIA JULY 2016 - MAY 2020

Bachelor of Technology - Information Technology

SKILLS

GenAI & Agents: Multi-Agent Systems, LangGraph, LangChain, RAG, LLMs (Claude, GPT, Gemini), OpenAI SDK, MCP, Prompt Engineering.

ML & Frameworks: PyTorch, TensorFlow, Scikit-learn, Transformers, NLP, Hugging Face, Pandas, NumPy.

Data & Retrieval: Vector DBs (Pinecone, FAISS, ChromaDB), ElasticSearch, Embedding/ETL Pipelines, Postgres, MongoDB, Redis.

Programming & Backend: Python, Java, FastAPI, AsyncIO, Pydantic, SQL, REST APIs, Git, Postman.

Cloud & LLMOps: AWS (Lambda, S3, IAM, CloudWatch), GCP (Vertex AI, Cloud Run, Cloud Build), Docker, Terraform, LangSmith.

CERTIFICATIONS

- AWS Certified Cloud Practitioner | AWS | 2023
- Machine Learning Specialization | Stanford/ DeepLearning.AI | 2023