YAVAR KHAN

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

PROJECTS

DEEP RESEARCH PRO - MULTI-AGENT ASYNC RESEARCH SYSTEM GITHUB

- **Architected a Multi-Agent Research System** using the OpenAI SDK, implementing an **autonomous query planner** and **multi-step verification** for cited synthesis.
- Optimized system throughput by **50x** via **bounded-parallelism** and concurrent search-summarization threads with system-level guardrails.
- Designed a **two-layer caching stack** (LRU + SQLite, 24h TTL) with context-aware bypassing and hit-rate logging to cut redundant API calls.
- Implemented reliable structured outputs and system observability through **Pydantic validation**, **retry/fallback logic**, detailed **performance logging**, **real-time live logs**, and **trace-based debugging** via the OpenAI Agents SDK..

MULTI-AGENT GENAI WEATHER ASSISTANT GITHUB

- Created a **multi-agent weather system** using **Gemini LLMs** to transform user queries into precise, data-grounded insights.
- Integrated **Google Agent Development Kit (ADK)** with Open-Meteo APIs for **geocoding**, **temporal inference**, **variable mapping**, and **dynamic endpoint routing**.
- Engineered robust workflows with **timezone-aware parsing** and structured JSON planning for deterministic query handling.
- Deployed on **Google Vertex AI** with **Retrieval Augmented Generation (RAG)** support and evaluation pipeline.

CONTEXT-AWARE AI RAG ASSISTANT GITHUB

- Engineered a **RAG pipeline** using **FAISS vector indexing** for persistent memory, achieving a **10x reduction in retrieval latency**.
- Developed a **multi-persona assistant** with **runtime persona switching** for adaptive tone, context reasoning.
- Added **real-time Pushover alerts** for connection requests and unhandled queries to enhance automation and user engagement.
- Built a **modular (Python + FastAPI + Gradio) stack** with private **dataset-backed retrieval** and deployed on **Hugging Face Spaces**.

TRANSFORMER-BASED SENTIMENT ANALYSIS GITHUB

- Developed a **Transformer-based NLP architecture** from scratch in **PyTorch**, implemented **multi-head self-attention** and custom positional encodings.
- Regularized training on **560K reviews** with **Dropout, Early Stopping, and L2**, achieving **91.8% accuracy** and **0.92 F1-score**.
- 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

- **Deployed 50+ Java APIs on Apigee and AWS**, boosting scalability and reducing release downtime.
- **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** to improve system observability and monitoring data-driven services.
- **Collaborated with global teams** to optimize performance and ensure secure API governance (IAM, quotas, rate limits).
- **Resolved 200+ client issues** with 40% faster turnaround through JIRA tracking and ElasticSearch analysis.
- **Mentored and led a 6-member team**, driving Agile/Scrum, CI/CD, and TDD practices to achieve 100% on-time delivery for 12+ projects.
- **Promoted from Associate to Analyst** in recognition of leadership, mentorship, and consistent project delivery.

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

AMITY UNIVERSITY, INDIA JULY 2016 - MAY 2020

Bachelor of Technology - Information Technology

SKILLS

Programming & Backend: Python, Java, C++, SpringBoot, FastAPI, REST APIs, AsyncIO, JSON Schema (Pydantic), SQL.
AI/ML & GenAI: PyTorch, TensorFlow, Scikit-learn, LLMs, RAG, Multi-Agent Systems, Prompt Engineering, Transformers, FAISS.
Data & Retrieval Systems: Pandas, NumPy, ETL Pipelines, ElasticSearch, Embedding Pipelines, PostgreSQL, MySQL, MongoDB.
Cloud & Deployment: AWS (Lambda, S3, IAM, CloudWatch), Vertex AI, Docker, Hugging Face Spaces, CI/CD Workflows.
Tools & Platforms: Git, JIRA, Kibana, Confluence, Postman, Gradio, Streamlit, Hugging Face, LangChain, LangGraph, Pushover Integrations.

CERTIFICATIONS

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