Harsh Vyas

(647) 219-6530 | hnv.hnvyas@gmail.com | linkedin.com/in/harshnvyas | github.com/HarshNVyas | Canada

Professional Experience

Agentic AI Developer | *Capria Ventures, Toronto, ON* Sep 2024 – Present

- Designed and deployed **production-grade GenAI applications** using **Python**, **Azure AI Foundry**, LLMs, RAG pipelines, and agentic workflows across finance, healthcare, education, logistics, agriculture, and technology domains.
- Developed a production **multi-agent automation framework** using **Google ADK**, **LangGraph**, **A2A protocols**, and **MCP tool integrations** to automate Value Add Metrics, Dropped Balls analysis, OKR tracking, lead generation, and operational workflows.
- Built an internal **Recruitment Agent** using **Azure Durable Functions**, **Azure OpenAI Service Azure Cosmos DB** to process up to **5,000 candidates per run** and sync validated results into **Google Sheets**.
- Architected a multilingual **RAG platform** with **RBAC**, semantic search, guardrails, content filtering, prompt-injection defense, and source-grounded responses using **LangChain**, **Azure AI Search**, **Azure Blob Storage**, **Azure Cache for Redis**, and **Azure Cosmos DB**.
- Deployed scalable AI services on **Azure App Service**, **Azure Container Apps**, **Azure Virtual Machines**, **Azure Load Balancer**, **Azure Application Gateway**, **Azure Front Door**, and **Azure API Management** with routing, caching, load balancing, and rate limiting.
- Integrated AI capabilities into enterprise applications using **REST APIs**, **GraphQL**, **gRPC**, **FastAPI**, asynchronous workflows, and secure service-to-service communication for business-critical production systems.
- Containerized GenAI services using **Docker** and deployed workloads across **AKS**, **ACI**, and **Azure Container Registry** for model APIs, agent workers, retrieval services, and background processing jobs.
- Built event-driven AI/data pipelines using **Kafka**, **Azure Service Bus**, **Azure Event Grid**, batch processing, vector indexing, retrieval workflows and validation layers for real-time ingestion, document intelligence and enterprise Q&A.
- Reduced **LLM hallucinations and latency** by improving retrieval quality, context filtering, caching, chunking strategies, guardrail design, and evaluation workflows across production RAG and agentic AI systems.
- Implemented **CI/CD and release automation** using **GitHub Actions**, **GitLab CI/CD**, **Docker pipelines**, **Azure Container Registry**, and automated validation gates, improving release cadence to **2–4 releases per week**.
- Established production **monitoring, logging, observability, and governance** using **Azure Monitor**, **Application Insights**, **Log Analytics**, **Grafana**, **Prometheus**, audit logs, access controls, and human-in-the-loop review.

Projects

Language Proficiency Board Platform | Next.js, Python, WebSockets, Deepgram, ElevenLabs

- Architected a production **AI-powered IELTS and CELPIP exam platform** with **Next.js**, **JWT authentication**, **PostgreSQL**, role-based access, subscriptions, admin workflows, and student-facing mock/live exam delivery.
- Designed a real-time **AI speaking exam system** using a persistent **WebSocket server**, streaming speech-to-text, LLM-driven examiner responses, text-to-speech playback, transcript persistence, and asynchronous AI scoring for 10–15 minute speaking sessions.
- Built an asynchronous **OCR and AI evaluation pipeline** to process uploaded exam PDFs into structured content, store exam data, evaluate responses with LLM rubrics, and generate band scores, feedback, and performance insights.

Publications

Accident Prone System using YOLO | Paper Link

- Published a real-time computer vision system using **YOLO**, achieving **99.4% accuracy** and optimizing for **low-latency** detection on traffic footage.

Education

Post Graduate Diploma Information Technology Business Analysis Apr 2024
Conestoga College, Kitchener, Ontario

- Coursework: Software Engineering, Advanced Data Mining, DBMS, Project Management, System Testing

Bachelor of Engineering in Computer Science and Engineering May 2022
Gujarat Technological University, Gujarat, India

- Coursework: Machine Learning, Deep Learning, Object-Oriented Programming, Distributed Systems, Cloud Computing

Technical Skills

GenAI, Agents & RAG: Azure AI Foundry, Azure OpenAI Service, Azure AI Search, OpenAI, Hugging Face, Azure Machine Learning, LLMs, RAG, Advanced RAG, Agentic AI, Multi-Agent Systems, LangChain, LangGraph, Scikit-learn, TensorFlow, PyTorch, Prompt Engineering, LLM Evaluation, Guardrails, Prompt-Injection Defense

Azure, Cloud & MLOps: Azure Compute (App Service, Container Apps, Functions, Durable Functions, Kubernetes Service, Container Instances), Container Registry, Cosmos DB, Blob Storage, Azure Service Bus, Azure Monitor, Application Insights, Key Vault, Managed Identities, Docker, Kubernetes, MLflow, GitHub Actions, GitLab CI/CD

Programming, APIs, Data & Vector Search: Azure API Management, Python, TypeScript, FastAPI, REST APIs, Kafka, Redis, PostgreSQL, SQL, NoSQL, Vector Databases (Weaviate, ChromaDB, Pinecone, FAISS, PG Vector)