

VINAYAK MATHUR

+91 9460041703 ◊ vinayak.k.mathur@gmail.com ◊ [linkedin.com/in/vinayakmathur2000/](https://www.linkedin.com/in/vinayakmathur2000/) · github.com/VnykzHub

OBJECTIVE

AI Engineer with 3+ years building production GenAI and agentic systems on GCP, Azure, and AWS. Specialized in LLM orchestration, multi-agent frameworks (LangGraph, ADK, Autogen), RAG pipelines, and scalable Python backends serving enterprise customers. Published research at ICCNT 2021. Looking to build customer-facing AI products at scale.

EDUCATION

Master of Data Science (Global), Deakin University	2023 – 2025
B.Tech in Computer Science (Information Technology), SRM Institute of Science and Technology	2018 – 2022

SKILLS

Languages	Python, SQL, JavaScript, TypeScript, Bash
GenAI & Agents	LLMs, RAG, Multi-Agent Systems, LangGraph, LangChain, Autogen, Google ADK, MCP, A2A, Vertex AI, Azure OpenAI, Nvidia NIM
Backend & Systems	FastAPI, Flask, REST APIs, Microservices, Pub/Sub, gRPC, PostgreSQL, MongoDB, Firestore, Vector Databases
Cloud & MLOps	GCP (Cloud Run, Vertex AI, Cloud Build, Cloud SQL, Pub/Sub, BigQuery), AWS, Azure, Docker, Kubernetes, GitHub Actions, CI/CD, SonarQube
ML & Data	PyTorch, TensorFlow, HuggingFace, Pandas, PySpark, NLP, Deep Learning

EXPERIENCE

Deloitte	Hyderabad, India
AI and Data Science Engineer	November 2025 – Present

- Building **Data Workbench**, a GCP-native platform consolidating 8+ AI-powered tools for data engineering and analytics workflows used across Deloitte client engagements.
- Architected and shipped Python/FastAPI backend services on Cloud Run with Cloud SQL and Firestore for persistence; designed user, project, and job orchestration layers integrated via GitHub Actions and Cloud Build CI/CD with SonarQube vulnerability scanning.
- Designed a Pub/Sub-based parallel execution layer for long-running LLM workloads on Vertex AI; benchmarked 100+ files processed concurrently per service with stable end-to-end latency, validating Vertex AI endpoint load balancing under burst traffic.
- Built **Dialect Shift Agent**, a Google ADK multi-agent system performing any-to-any SQL translation across 8 dialects (Teradata, BigQuery, PySpark, Snowflake, etc.), eliminating manual rewrites in client migrations.
- Built **IBG (Intelligent Business Glossary)**, a RAG service backed by Vertex AI vector search that auto-generates column descriptions and PII sensitivity classifications by connecting directly to BigQuery or ingesting metadata files.
- Refactored multiple Azure OpenAI-based legacy tools to GCP-native stacks (Vertex AI + ADK), redesigning APIs and deployment topology to fit a unified platform architecture.

Cognizant	Gurugram, India
Associate AI Engineer	July 2025 – October 2025
Junior Data Scientist	July 2022 – July 2025
AI/ML Intern	January 2022 – July 2022

- Architected a production GenAI data analyst for a global beverage leader using LangGraph, Azure OpenAI, and NL-to-SQL translation; orchestrated specialized agents for SQL generation, visualization, and Pandas-based computation to handle complex business rules (YTD, market share, channel splits).
- Engineered an autonomous multi-agent IT-support framework on Nvidia NIM, integrating MS Graph (Outlook/Teams) and a RAG debugging pipeline to triage and resolve tickets with near-zero human intervention.
- Built and deployed an automated interview platform on a multi-cloud Kubernetes architecture (GCP, Azure, AWS) integrating GPT-4 and Claude; achieved 99.9% uptime and a 40% reduction in operational cost over the prior system.
- Delivered a RAG-powered audit-automation tool (GPT-3/4 + Streamlit) for a Big 4 client, cutting audit preparation time by 65%.
- Trained ML models (Random Forest, GBM, NN) for customer churn prediction during the GENC Next program, improving baseline accuracy by 15%.

- Developed an anti-spoofing and speaker-verification model for on-device voice IoT using Residual Squeeze-and-Excitation networks in TensorFlow, achieving 4.44% EER on the ASVspoof benchmark; work published at ICCNT 2021 (IEEE).

SELECTED PROJECTS

- **Dialect Shift Agent (Deloitte)** – Multi-agent SQL translation system on Google ADK supporting any-to-any conversion across 8 SQL dialects, with reasoning agents for syntax mapping, function equivalence, and validation.
- **Pub/Sub Parallel Execution Layer (Deloitte)** – Asynchronous task pipeline on GCP Pub/Sub + Cloud Run enabling concurrent LLM-backed file processing; sustained 100+ parallel jobs per service against Vertex AI endpoints.
- **Intelligent Business Glossary (Deloitte)** – RAG service over Vertex AI vector search that auto-generates column descriptions and PII sensitivity tags; integrates directly with BigQuery and ingests heterogeneous metadata files.
- **Multi-Cloud Interview Platform (Cognizant)** – Kubernetes-orchestrated platform deployed across GCP, Azure, and AWS integrating GPT-4 and Claude for live candidate evaluation; sustained 99.9% uptime in production.

PUBLICATIONS

- Ray, R., Karthik, S., **Mathur, V.**, Kumar, P., Maragatham, G., Tiwari, S. “*Feature Genuinization based Residual Squeeze-and-Excitation for Audio Anti-Spoofing in Sound AI.*” 12th International Conference on Computing, Communication and Networking Technologies (ICCNT), IEEE, 2021. [[IEEE Xplore](#)]

CERTIFICATIONS

- **NVIDIA-Certified Associate:** Generative AI LLMs (NCA GEN-L)
- **Post Graduate Program in AI and Machine Learning,** University of Texas at Austin (McCombs)