TANISHQ TEMBHURNE

Hyderabad, India
(+91) 9119530237 | tanishq.aidev1310@gmail.com | LinkedIn | GitHub

Education

Woxsen University, Hyderabad **2021 – 2025**
Computer Science Engineering (B.Tech)

Skills

Programming: Python, JavaScript/Node.js
AI/ML: Machine Learning, Deep Learning, PyTorch, NLP, Computer Vision, Fine-Tuning, Model Evaluation, Model Optimization, Generative AI, LLMs, SLMs
LLM & Agentic AI: RAG, LangChain, LangGraph, LlamaIndex, AI Agents, Agentic Workflows, Prompt Engineering, Tool Calling, Multi-Agent Systems, Function Calling
AI Infrastructure & Deployment: Azure, AWS, GCP, Azure Databricks, Docker, REST APIs, Model Serving, CI/CD, MLOps
Retrieval & Data: Vector Databases, FAISS, ChromaDB, Neo4j, Embeddings, Semantic Search, Hybrid Search, Reranking, Pandas, NumPy
Observability & Automation: Azure Monitor, Application Insights, Log Analytics, KQL, Azure Logic Apps, Jira Automation, Production Monitoring, Incident Response
Data Visualization: Power BI, Matplotlib, Seaborn

Experience

Centific **April 2025 – Present**
SRE – Engineer
- Engineered an AI-driven incident investigation system across 3 Azure production environments, integrating Azure Monitor, Application Insights, Log Analytics/KQL, runbooks, and LLM-based reasoning to correlate telemetry and generate evidence-backed root-cause analyses.
- Designed an agentic investigation workflow that dynamically analyzes alerts, queries production telemetry using KQL, retrieves relevant runbooks and historical context, and synthesizes multi-source evidence into structured RCAs, reducing MTTR by 30%.
- Automated 30+ production alerts per month into an AI-assisted investigation pipeline using Azure Monitor, Logic Apps, and Jira, automatically initiating investigation workflows and publishing RCA findings and incident updates while reducing manual incident-management effort by 50%.
- Built production observability and AI investigation workflows across Azure Monitor, Application Insights, and Log Analytics, combining logs, metrics, traces, dependency telemetry, and deployment activity to improve incident detection and accelerate AI-assisted diagnosis by 30%.
- Designed an AI SRE agent with human-in-the-loop validation, generating detailed RCAs and three progressive incident updates within the 30-minute SLA, reducing L1/L2 investigation dependency by 50%.

Conquest Technology Solutions **Feb 2024 – June 2024**
Data Science Intern
- Developed and analyzed KPI dashboards in Power BI to provide actionable insights.
- Boosted sales up to 15% in low-performing areas through effective data visualization.
- Conducted advanced EDA to identify trends and anomalies, enhancing business intelligence.
- Simplified complex datasets with clear visualizations and ensured accurate reporting through stakeholder collaboration.

Projects

Autonomous Meeting Agent Python, Node.js, SLM, RAG, Chromium, OpenAI Realtime, ElevenLabs, Redis
- Built a real-time AI virtual employee capable of autonomously joining Chromium-based meetings, actively listening to participants, answering and asking questions, following meeting instructions, sharing/presenting content, and generating and distributing meeting minutes (MoM) to attendees.
- Fine-tuned domain-specific Small Language Models (SLMs) to power the agent's reasoning and conversational behavior, reducing dependency on large general-purpose models while maintaining contextual response quality across meeting interactions.
- Improved context retrieval Recall@5 from 0.52 → 0.76 by tuning chunking, embedding strategy, metadata filtering, and top-K retrieval across a 500-query evaluation set.
- Reduced average prompt context size by 45% through conversation summarization and selective memory retrieval while maintaining 90%+ answer consistency on multi-turn evaluation scenarios.
- Designed hierarchical short- and long-term memory for conversational context, meeting decisions, action items, and participant information; reduced average prompt context size by 45% through summarization and selective memory retrieval.
- Implemented Redis-based caching for frequently accessed knowledge and retrieval results, reducing redundant retrieval/LLM calls and improving median response latency by 40% through cache reuse and controlled invalidation.

Legal-Aid Conversational Chatbot LangChain, LangGraph, FAISS, Neo4j, Groq
- Built a retrieval-augmented legal chatbot using FAISS and Neo4j, combining dense vector retrieval with knowledge-graph relationships to improve contextual retrieval across legal documents.
- Improved Recall@5 from 0.41 → 0.70 by replacing basic keyword retrieval with hybrid BM25 + dense retrieval and metadata filtering.
- Reduced unsupported LLM responses by 32% by implementing retrieved-context validation, source grounding, and response constraints across a 500-query evaluation dataset.
- Improved retrieval quality to 0.70 Recall@5 and 0.62 MRR through iterative evaluation of chunk size, embedding strategy, metadata filtering, and top-K retrieval parameters.
- Implemented LangGraph-based conversation state and memory management, maintaining context across 10+ conversational turns while restricting responses to retrieved legal knowledge.

Resume Ranker Langchain, Python, NLP, Embeddings, LLMs
- Built an NLP-based resume ranking pipeline extracting candidate skills, experience, education, and job requirements to generate relevance scores against job descriptions.
- Created a labeled evaluation dataset of 100 resumes across 30 job descriptions and evaluated ranking quality using Precision@K, Recall@K and NDCG@K.
- Improved NDCG@10 from 0.61 → 0.82 by combining semantic embedding similarity with structured skill and experience matching instead of relying solely on keyword overlap.
- Improved Precision@10 from 0.68 → 0.84 through score normalization, skill-weight optimization, and semantic similarity-based candidate ranking.

- Reduced candidate screening effort by 60%, ranking approximately 100 resumes in under 2 minutes compared with manual resume-by-resume screening.

Multi-Agentic AI Marketing System CrewAI, LLMs, Streamlit

- Developed a multi-agent AI marketing system using CrewAI to automate and optimize marketing strategies.
- Integrated Ollama 3.1:8b (local) as the LLM to enhance decision-making and content generation.
- Implemented intelligent agents for campaign management, audience targeting, and performance analysis.
- Improved engagement and conversion rates through data-driven marketing strategies.
- Deployed the system on Streamlit for seamless user interaction and accessibility.

Vectorization and Validation Vector Databases, Azure Databricks, RAG

- Developed an AI-powered Q&A system processing different types of documents for efficient query resolution.
- Implemented ChromaDB & Mosaic AI for vector search, achieving a 70% improvement in retrieval accuracy.
- Validated results using cosine similarity and Euclidean distance, achieving up to 92% query accuracy.
- Deployed on Azure Databricks for scalability and real-time performance.

Medical Imaging AI System BLIP + RAG + Vector Database

- Developed a multimodal medical imaging system using BLIP (Bootstrapped Language-Image Pretraining) to detect abnormalities in chest X-ray images.
- Designed and implemented a Retrieval-Augmented Generation (RAG) pipeline integrating FAISS/ChromaDB, enhancing contextual explanation quality and increasing diagnostic insight relevance by 25%.
- Engineered an end-to-end workflow for image embedding generation, semantic retrieval, and structured report synthesis, reducing manual interpretation effort by 30–40%.
- Deployed a scalable backend architecture for real-time inference and medical report generation, achieving sub-2-second response latency and improving processing efficiency by 40%.

Deep Learning Model Optimization & Finetuning (Image Generation) PyTorch + Weights & Biases

- Built and fine-tuned a deep learning model using PyTorch, optimizing hyperparameters through structured experimentation for domain-specific image generation and improving model accuracy by 12%.
- Leveraged Weights & Biases (W&B) for experiment tracking, hyperparameter sweeps, and model comparison, reducing experimentation time by 35%.
- Conducted systematic tuning of learning rate, batch size, optimizer configuration, and regularization techniques, reducing validation loss by 20%.
- Implemented performance monitoring using accuracy, precision, recall, F1-score, and loss tracking, enabling reproducible training pipelines and improving model stability by 20–30%.

Soft Skills

- Strong analytical and problem-solving skills with a structured approach to troubleshooting and root cause analysis.
- Effective communication and stakeholder collaboration in cross-functional technical environments.
- Ownership-driven mindset with a focus on reliability, accountability, and continuous improvement.
- Ability to work under pressure in production environments while maintaining SLA commitments.
- Adaptability and quick learning capability in fast-paced AI and cloud-native ecosystems.