Vijay Sheru

University of North Texas, Denton, Texas

+1 9405354790 | vijaysherr222@gmail.com | https://www.linkedin.com/in/vijaysheru/
https://github.com/vijaysheru

SUMMARY

AI Engineer with a Master's in Artificial Intelligence from the University of North Texas and 2+ years of software engineering experience at Infosys. Skilled in building and deploying real-world AI systems with expertise in LLMs, NLP, and MLOps. Experienced in developing GPT-powered tools, vector database applications, and autonomous AI agents. Successfully led multiple AI projects—including fraud detection, spiritual LLM agents, and an AI job platform—focused on scalability, performance, and user experience. Passionate about solving meaningful problems through intelligent automation, prompt engineering, and scalable APIs.

PROJECTS

Credit Card Fraud Detection using Neural Networks — Feb 2024 - Mar 2024

- Developed a deep learning pipeline using PyTorch to detect fraudulent credit card transactions with high precision.
- Engineered tensor-based preprocessing and modularized training pipelines for scalability and reproducibility.
- Boosted model accuracy by integrating XGBoost with ensemble learning and reinforcement learning-based threshold tuning.
- Reduced false positives by 22% and improved model interpretability through SHAP analysis and detailed evaluation metrics (AUC: 0.96, Precision: 0.88).

Art Vision Transfer (Neural Style Transfer) — Aug 2024 - Oct 2024

- Designed and implemented a real-time Neural Style Transfer system that transforms input images into artistic renditions using deep learning.
- Replaced the traditional VGG19 backbone with MobileNetV2, reducing inference time by 30% and optimizing for edge-device performance.
- Developed an interactive Flask web app integrated with OpenCV for live camera-style previews and seamless user interaction.
- Fine-tuned Gram matrix-based content and style loss functions to improve synthesis quality and preserve stylistic integrity across diverse images.

AI Response & Humanization System — Jan 2023 - Present

- Built an advanced multi-LLM response engine using Hugging Face APIs to aggregate content from Mistral-7B and Falcon-7B.
- Designed a dual-layer AI detection module integrating the Sapling API for real-time classification with an advanced heuristic fallback system.
- Engineered a flexible humanization pipeline using prompt-tuned rewriting strategies for conversational, formal, and narrative outputs.
- Implemented response scoring, natural language transformation, and re-evaluation logic to iteratively reduce AI detection probability.
- Developed a clean, responsive Streamlit UI with dynamic flow controls, AI percentage meters, and tone-aware rewriting customization.

Cold Email Generator with Prompt Chaining — Jan 2025 - Feb 2025

- Built an AI-powered cold email generation system using LangChain's prompt chaining and persona injection for highly personalized outreach.
- Created modular templates with context-aware memory, enabling multi-turn input adaptation and dynamic response generation.
- Deployed the tool as a fully functional web app with a responsive Streamlit frontend and FastAPI backend, supporting real-time generation and export of emails.
- Enabled use-case adaptability across sales, networking, and outreach campaigns by integrating Hugging Face LLMs for tone and style variation.

MythOS - Talk to Lord Krishna (Spiritual LLM Agent) — Feb 2025 - Mar 2025

- Built a Retrieval-Augmented Generation (RAG)-based Krishna chatbot capable of contextual Q&A with direct quotes from the Bhagavad Gita.
- Leveraged ChromaDB and LangChain to implement vector search and memory persistence for spiritual dialogue continuity.
- Enabled natural voice interaction using Coqui TTS, enhancing accessibility and engagement in immersive conversations.
- Added support for custom PDF/CSV uploads, allowing personalized question-answering with long-term knowledge integration.
- Designed and deployed the interface as a spiritual companion web app using Streamlit for seamless user interaction.

LaunchHire — Apr 2025 - Present

- Developed a full-stack AI-powered job application platform combining resume tailoring, job matching, mock interviews, and automated applications using GPT-4, FAISS, ElevenLabs, and FastAPI
- Implemented autonomous AI agents for resume generation, interview simulation, job ranking, and personalized feedback — with real-time voice + video playback.
- Built a secure resume vault system with local 48-hour auto-clean + 60-day server backup, and integrated PDF resume versioning and activity logging.
- Integrated job scraping and JSearch API, added FAISS-powered semantic filtering, and created a personalized endpoint with skill filters.

- Engineered an intelligent job apply agent using Playwright for browser autofill + Twilio SMS fallback when manual input is needed.

TECHNICAL SKILLS

- **Programming & Scripting**: Python, Java, SQL, Bash
- **AI & Machine Learning**: Supervised & Unsupervised Learning, Deep Learning, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Transformers, Generative AI, NLP, Computer Vision, Reinforcement Learning, Machine Learning, Natural Language Processing
- **Frameworks & Libraries**: TensorFlow, PyTorch, Scikit-learn, Hugging Face Transformers, LangChain, OpenCV, NLTK, SpaCy
- **Data Science & Analytics**: Pandas, NumPy, Matplotlib, Seaborn, Plotly, Data Cleaning, Feature Engineering, Exploratory Data Analysis (EDA), Data Visualization, Statistical Modeling
- **Model Deployment & MLOps**: MLflow, FastAPI, Flask, Docker, Kubernetes, CI/CD Pipelines, Model Monitoring, Model Versioning, Prompt Engineering, Containerization
- **Cloud & DevOps**: Amazon Web Services (S3, EC2, Lambda, Sage Maker), Heroku, Git, GitHub Actions, GCP
- **Databases & Vector Stores**: PostgreSQL, ChromaDB
- **Tools & Platforms**: Jupyter, VSCode, Google Colab, Streamlit, Tableau, Notion, LangFlow, Github, PowerBI

EDUCATION

University of North Texas, Denton, Texas **Aug 2023 - May 2025**
Masters in Artificial Intelligence *Denton Texas*

- **GPA:** 3.3

EMPLOYMENT EXPERIENCE

Infosys **Nov 2021 - Jul 2023**
Systems Engineer (AI Tools & Data Systems) *Hyderabad India*

- Designed and deployed scalable APIs with integrated performance optimization and automated testing, reducing server response time by 50% and improving system reliability.
- Improved messaging and alerting systems by implementing RESTful API patterns and clean code practices, leading to a 40% increase in user engagement and a 25% boost in operational efficiency.
- Collaborated on backend infrastructure with potential for AI model integration, setting up foundational services used in data-driven and predictive modules.
- Developed unit testing frameworks in JUnit to reduce code defects by 30% and accelerate software release cycles.
- Contributed to Agile AI-solution delivery processes, leading sprint planning, grooming, daily stand-ups, and demo sessions.
- Drove experimentation and iterative development, increasing engineering throughput by 30% and reducing time-to-market by 25% for AI-enabling backend tools.
- Led a team in designing a modular firmware architecture to support future intelligent features, improving memory usage by 30% and enhancing system stability.
- Contributed to the development of scalable backend systems and AI-ready infrastructure used in analytics, automation, and future ML model integration.

CERTIFICATIONS

- **Machine Learning Specialization**: Coursera - Machine Learning Specialization, Machine Learning Operations (MLOps)
- **AWS Cloud Practitioner**: AWS - AWS Cloud Practitioner
- **Internal Java Developer**: Infosys - Internal Java Developer, Infosys Internal Agile Certification