VENKATKUMAR RAJAN

+91 88703 11826 | Madurai, Tamil Nadu, India | venkatkumarr.vk99@gmail.com
linkedin.com/in/venkatkumarvk | kaggle.com/venkatkumar001 | vk-ant.github.io/Venkatkumar

PROFESSIONAL SUMMARY

Generative AI Engineer with **3.5+ years** of experience designing and deploying production-grade AI systems across multimodal applications. Proven ability to translate cutting-edge research into scalable, real-world solutions that deliver measurable business impact, with a strong focus on secure, responsible AI systems and governance. **Kaggle Master** and author of open-source LLM evaluation library **llmevalkit**.

EXPERIENCE

Generative AI Engineer | *Business Analyst* — **Dec 2023 – Present**
EXL Service (India) Pvt. Ltd. — *Noida, India (Remote)*

- **AI-Powered Credential Extraction System** *(Unified Text RAG + Multimodal RAG, Azure):* Architected and owned end-to-end design, development, and production deployment of a **unified** dual-mode RAG pipeline that automates extraction of **46+ structured fields** from complex, unstructured documents (native PDFs, scanned images, mixed-format files). Built on Azure OpenAI (GPT-5), Azure Document Intelligence (OCR, layout parsing), Azure AI Search (per-field vector indexing, semantic retrieval), and Azure OpenAI Embeddings (chunking + embedding pipeline). Implemented confidence-based routing, automated aggregation, and human-in-the-loop review — achieving **70% extraction accuracy** and **reducing manual effort by 30%**. Deployed via Azure DevOps CI/CD with Azure Blob Storage, Azure Key Vault, and structured logging.
- **Unified GenAI Document Intelligence & Automation System** *(Azure):* Owned end-to-end design, development, and production deployment of a real-time, **unified** Generative AI document parsing and intelligence system built on Azure OpenAI, NLP, and computer vision. A single extensible codebase extracts and structures data from multiple complex document types — onboarding a new document type requires only a **config entry and prompt update, with zero redevelopment**. Achieved $>$**95% extraction confidence**, **reduced manual processing effort by 40–50%**, deployed with Azure Blob Storage, Azure Key Vault, Azure DevOps CI/CD, and full audit logging.
- **AI-Based Document Classification System** *(Unified, Azure):* Built a scalable, **unified** AI document classifier supporting **15+ document categories** with $>$**97% accuracy**. New categories added instantly via a reference document — no engineering or prompt changes required.
- **Custom AI Agent Builder** *(Databricks, Internal):* Designed and built an internal framework enabling users to **create and configure custom AI agents** — connecting multiple Unity Catalog tables, attaching PDF knowledge bases, and defining agent behaviour without writing code. Built specifically to support **denial management analytics**, enabling analysts to perform natural-language queries, root-cause analysis, and insight generation over structured and unstructured data. Deployed as a governed, reusable Databricks App with RAG-powered responses.
- **Internal AI Productivity Tools:** Built a suite of internal tools to accelerate team workflows — **Excel AI assistant** (natural-language queries over spreadsheets), **SQL chatbot** (plain-English to SQL), **Power App** integrations, and **Document Intelligence field extraction** for automated form processing — reducing manual effort across data, operations, and analyst teams.
- Delivered **10+ GenAI PoC presentations** to prospective clients; broad experience across AI system design, enterprise data pipelines, and production LLM deployment.

Machine Learning Engineer — **Sept 2022 – Oct 2023**
AugRay — *Chennai, Tamil Nadu (Onsite)*

- **Real-Time Ball Fault Detection** *(Edge AI, Computer Vision):* Designed and deployed a lightweight real-time defect detection system for a leading sports goods manufacturer using MobileNetV2, optimised for NVIDIA Jetson AGX. Achieved **90%+ detection accuracy**, **reduced QA error rates by 40%+**, and **increased manufacturing throughput by 3**$\times$ through efficient edge inference and model optimisation.
- **Automated 3D Reconstruction** *(Generative AI + Photogrammetry + NeRF):* Implemented an end-to-end automated 3D reconstruction pipeline combining traditional photogrammetry with Generative AI and Neural Radiance Fields (NeRF). Achieved **95% reconstruction accuracy** for non-reflective objects and $\sim$**50%** for reflective objects, significantly reducing manual processing time and improving 3D asset scalability.
- **Wall & Floor Segmentation with Intelligent Colour Schemes:** Fine-tuned **SegFormer-B0** (segformer-b0-finetuned-ade-512-512) for wall and floor semantic segmentation, achieving **80% accuracy** and enabling a paint manufacturer to automate interior colour preview generation. **Reduced designer effort by 20+ hours/month** and improved design turnaround time and visualisation consistency.
- **Foot Detection & Virtual Shoe Try-On** *(Computer Vision, AR):* Led development of a CV-based virtual shoe try-on system trained on **10,000+ annotated foot images**. Achieved **79% placement accuracy**, enabling realistic footwear visualisation and contributing to a **5% reduction in customer return rates** during pilot trials.
- **Generative AI Proofs of Concept:** Built multiple GenAI PoCs including automated flyer generation, face texture synthesis, and a site-based conversational chatbot — driving innovation in interactive marketing, customer engagement, and AI-powered digital experiences.

SKILLS

- **Programming Languages:** Python, C++
- **LLM Systems & AI Engineering:** Prompt Engineering, LLM Evaluation, AI Agents, Retrieval-Augmented Generation (RAG), GraphRAG, Monitoring, End-to-End Orchestration
- **Frameworks & Libraries:** TensorFlow, PyTorch, OpenCV, scikit-learn, PySpark, NetworkX, Mediapipe, librosa, LangChain, LlamaIndex, LangGraph, CrewAI, AutoGen, Phidata, Groq, Ollama, Docling, MCP
- **Web & Backend Frameworks:** Django, FastAPI, Flask, REST APIs, Streamlit, Gradio

- **MLOps & LLMOps:** Model Deployment, CI/CD, Docker, Kubernetes, Kubeflow, MLflow, Kafka, Azure DevOps
- **Cloud & Platforms:** Azure (OpenAI, AI Search, Document Intelligence, Blob Storage, Key Vault), AWS, Databricks
- **Edge AI:** NVIDIA Jetson AGX, RaspberryPi
- **Database, Automation & Visualization:** MongoDB, MySQL, Power BI, Power Automate, Power Apps, Pytest, Unittest
- **Professional Skills:** Leadership, Project Management, Agile Methodologies, Business Analysis, Problem-Solving, Translating Research into Production Systems

PERSONAL PROJECTS & OPEN SOURCE

- **llmevalkit (PyPI v1.0.3):** Open-source Python library for evaluating LLMs across **15 metrics** (pure math-based + LLM-as-Judge) with no ground-truth requirement. Multi-provider: OpenAI, Anthropic, Azure OpenAI, Groq, Ollama. Grounded in SelfCheckGPT, G-Eval, FActScoring research. **v2.0 roadmap:** compliance metrics for HIPAA, GDPR, India DPDP Act, EU AI Act, NIST AI RMF, CoSAI — dual-mode (regex + LLM-judgment).
- **Responsible AI Series** (Medium/@VK_Venkatkumar, 6+ parts): Technical articles covering NIST AI RMF, GDPR, CoSAI, HIPAA, LLM Safety Guardrails, Bias Detection, Red Teaming — with original architecture diagrams and Python walkthroughs.
- **Text-to-3D Reconstruction Pipeline** *(Generative AI + NeRF, Research):* Proposed and built a fully automated pipeline that transforms natural language descriptions into detailed 3D models via a multi-stage workflow: text → Stable Diffusion image generation → RL-agent-based enhancement → reflection removal (Stable Delight) → image upscaling & background removal → volumetric 3D reconstruction. Addresses semantic coherence, geometric complexity, and visual fidelity for AR/VR and digital content creation applications. arXiv | Project Page
- **PBR Material Classification + Gaussian Splatting** *(Computer Vision, Independent Research):* Independent research on physically-based rendering (PBR) shader principles for interpretable material property extraction from single images — recovering albedo, smoothness, specular, and metallic channels without ground-truth labels. Extended to explore **3D Gaussian Splatting (3DGS)** for novel-view synthesis and single-image multi-view reconstruction. Medium Article
- **Comprehensive Computer Vision Portfolio** *(Object Detection, Tracking, Video Analysis):* Built a wide range of computer vision solutions spanning object detection, multi-object tracking, and real-time video analysis. Full project portfolio: github.com/VK-Ant/ComputerVision-Exploration-Project
- **GenAI RAG/CAG + Knowledge Graph System & Multimodal Chatbot:** Built a Generative AI-powered RAG/CAG and knowledge graph system integrating diverse data sources for efficient information retrieval and context-based generation. Also developed a multimodal chatbot supporting **PDF, CSV, and vision-based drag-and-drop input** with multilingual text + audio responses — applicable to real-time domains such as security and drone systems. GitHub

RESEARCH & PUBLICATIONS

- A Generative Approach to High Fidelity 3D Reconstruction from Text Data [arXiv] [GitHub]
- Advancing Audio Fingerprinting Accuracy with AI and ML: Addressing Background Noise and Distortion Challenges [IEEE]
- Implementation of PCB Layout using CNC Machine Controlling with Wireless Communication [IRJIET]

CERTIFICATIONS & ACCOMPLISHMENTS

- **Kaggle Master** – Competitions Expert (top ∼0.4% globally), Notebooks Master (top ∼1.3%)
- **Certifications:** TensorFlow Developer (Google), Nvidia Jetson AI Specialist, Microsoft Azure AI, Self-Driving Car Engineer (Udacity Nanodegree), Reinforcement Learning, Computer Vision, Generative AI, AI for Cybersecurity, AI Security

EDUCATION

- **M.Tech. in Artificial Intelligence & Machine Learning** — **Rajasthan, India**
  Birla Institute of Technology and Science (BITS), Pilani — *Mar 2023 – Apr 2025*
- **B.E. in Electronics and Communication Engineering** — **Madurai, Tamil Nadu, India**
  SACS MAVMM Engineering College, Anna University — *Jul 2016 – Oct 2020*