

VAMSI KRISHNA UNNAM

AI Engineer · Backend Agentic Systems · Production Reliability

vk86294140@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

SUMMARY

Backend AI Engineer with 5 years building production LLM systems, document-intelligence workflows, Python services, and AWS reliability automation. Built parallel LLM agent workflows for incident intelligence that cut complex-case resolution time by 40%, and automated multi-region AWS failover across 32 enterprise applications, cutting per-application cutover time by 70%. Strong in Python/FastAPI backends, schema-validated LLM outputs, RAG, evaluation harnesses, CI/CD, and observability.

TECHNICAL SKILLS

AI Agents & LLM Systems: LLM integration, agentic workflows, parallel orchestration, RAG, prompt engineering, structured response parsing, source-grounded outputs

Backend Engineering: Python, FastAPI, Flask, REST APIs, Pydantic, PostgreSQL, SQL, Bash, Go, TypeScript, JavaScript, microservices

Evaluation & Quality: Eval harnesses, golden test sets, regression testing, pytest, CI quality gates, prompt/model versioning, A/B testing

Cloud & Reliability: AWS (EC2, RDS, S3, Lambda, Route 53, CloudWatch, ECS Fargate, ECR, ALB, IAM), Terraform, Docker, Kubernetes, Gremlin, Datadog, Splunk, Grafana

ML & Data: PyTorch, TensorFlow, Scikit-learn, LightGBM, OpenCV, Pandas, NumPy, Spark/PySpark, Kafka, SageMaker

EXPERIENCE

Software Engineer — Backend AI Systems & Reliability, Fortune 100 Financial Services

Jul 2023 – Present

- Reduced complex-case resolution time by 40% by building parallel LLM agent workflows where specialized agents analyze operational data sources concurrently and merge into structured, source-grounded incident summaries with confidence scoring.
- Cut manual document-processing time by 60% by designing and deploying a production generative-AI summarization pipeline across 5 concurrent workstreams with 3 business stakeholders.
- Own backend implementation for production AI workflows: Python/FastAPI inference adapters, Pydantic schema validation, context assembly, typed response parsers, and observability hooks.
- Reduced per-application AWS failover cutover time by 70% through Python-driven runbook automation with pre/post health checks across EC2, RDS, Route 53, and CloudWatch for 32 enterprise applications, validating RTO/RPO with Gremlin chaos experiments.
- Built reusable AI evaluation and validation harnesses (pytest, schema checks, regression tests, CI/CD quality gates) that cut integration-testing cycles by 35% and became the standard framework for 6+ downstream AI projects.

Machine Learning Engineer, Global IT Services Firm — Client: ZF (Automotive)

Jul 2020 – Aug 2022

- Achieved 91% object-detection accuracy by translating ADAS requirements into YOLO-based perception models for vehicles, pedestrians, and traffic signs in a live autonomous-driving inference pipeline, alongside U-Net and DeepLab segmentation models in TensorFlow for lane and road-boundary detection.
- Reduced segmentation dataset generation time by 50% by automating TFRecord serialization and annotation pipelines for large-scale camera and LiDAR road-scene data, and surfaced geospatial alignment defects with a Python LiDAR point-cloud visualization tool.

PROJECTS

Document RAG Assistant

Python · FastAPI · Claude API · Docker

- Document-intelligence service answering questions over uploaded documents with source-grounded citations — chunking, embeddings, hybrid keyword + vector retrieval, and a FastAPI backend.
- Offline-testable with an echo LLM stub and a retrieval eval harness scoring recall@k and citation grounding, plus pytest coverage over chunking, retrieval, and API contracts.

Credit-Risk MLOps Pipeline

Scikit-learn · FastAPI · Pydantic · GitHub Actions

- Reproducible pipeline that trains, versions, and serves a credit-default model behind a schema-validated FastAPI service, with train/serve parity and CI gates on ROC-AUC thresholds, monotonic risk behavior, and API contracts.

Chaos-Forge

Python · HTTP proxy · Docker · Pytest

- Chaos-engineering toolkit that injects latency, errors, blackholes, and CPU/memory pressure, then validates steady-state hypotheses with p50/p95 probes and pass/fail resilience reports.

EDUCATION

M.S. Computer Science, Stevens Institute of Technology

May 2024 · GPA 3.5/4.0

B.E. Electronics & Communication, RMK Engineering College / Anna University

Mar 2021 · CGPA 8.7/10