

Venkatesh Guttikonda

AI Engineer | Agentic Systems & LLM Applications

venkatesh.gtd1@gmail.com · +1 660-853-9110 · United States · venkatesh-guttikonda-ai-portfolio.netlify.app · github.com/Ven-Z8

SUMMARY

AI Engineer with **~2 years building production multi-agent systems**, on top of 3+ years of data engineering. Currently building an enterprise L2-triage platform at Infosys on LangGraph, MCP, and GraphRAG — observable from token to decision, deterministic by design, and onboardable to new business domains by adding an SOP document rather than code. Comfortable from agent runtime internals up to enterprise integration (Azure AD, MCP tool servers, audit trails).

TECHNICAL SKILLS

AI / LLM Stack — LangGraph, LangChain, Pydantic AI, Semantic Kernel, Instructor, MCP, Multi-Agent Systems, Context Engineering, GraphRAG, Agentic RAG, ReAct & Planning Agents, Tool Calling, Structured Outputs, AI Memory Systems, Long-Context Orchestration, Anthropic SDK (Claude), Azure OpenAI, Vertex AI, Hugging Face

Languages — Python, TypeScript, SQL

Frontend — React, Next.js, Streamlit

ML / NLP — PyTorch, Transformers (BERT, RoBERTa fine-tuning), Text Classification, NER, Semantic Search, Entity Linking, Summarization, Question Answering

AI Developer Tools — Claude Code, Cursor, GitHub Copilot, Goose

Data Engineering — Pandas, NumPy, PySpark, Kafka, Apache Airflow, Snowflake; Matplotlib, Seaborn, Plotly

Databases & Vector Stores — Postgres, MongoDB, Qdrant, Pinecone, FAISS

Cloud, DevOps & Observability — Azure (AD, AKS, Azure OpenAI), GCP (Vertex AI, BigQuery, Pub/Sub), AWS, Docker, Kubernetes, CI/CD (Jenkins), FastAPI, Phoenix, OpenTelemetry, LangSmith, RAGAS, DeepEval

AI Concepts — Prompt Engineering, RAG Architecture, Agent Memory, Context Window Management, Programmatic Tool Calling, Tool Use & Function Calling, Guardrails & Safety, Human-in-the-Loop, LLMOps, Eval Design, Token Economics, OWASP AI Security

PROFESSIONAL EXPERIENCE

AI Engineer — Infosys

May 2025 – Present

Houston, TX, United States

- Designed and deployed a **multi-agent AI platform** that automates L2 triage of batch-job failures, cutting engineer investigation time from **hours to minutes for 10+ engineers** across disconnected enterprise systems (Autosys, SQL, file servers, ServiceNow). Handles **8–10 tickets/day** at current load.
- Architected a unified **L2Ops layer with 4 specialized MCP tool servers** integrating directly with Autosys, SQL databases, file systems, and ServiceNow; secured the full API surface with Azure Active Directory and enterprise token management.
- Implemented **GraphRAG-based intent classification** to retrieve domain-specific SOPs and generate structured execution plans with explicit step dependencies — replacing free-form LLM responses with deterministic, auditable behavior.
- Engineered **context-engineering pipelines** with dynamic retrieval planning, memory-aware context injection, and structured execution routing to improve agent reasoning across long-running enterprise workflows.
- Built a **dynamic graph builder** in LangGraph that converts SOP execution plans into live agent graphs — auto-parallelizing independent steps and dispatching specialized agents against live enterprise systems.
- Engineered a **zero-code scaling model** where new workflows onboard by adding an SOP document with no code changes, enabling expansion to any team or business domain independently.
- Instrumented **end-to-end observability with Phoenix and ITrace**, tracing every LLM call, token usage, and agent decision for complete audit trails and enterprise compliance.

AI / NLP Engineer — X Node Inc.

Aug 2024 – May 2025

Remote

- Built and deployed advanced GenAI agents (**Chief of Staff, StandUp, Market Research, InfoGathering**) using LangGraph to automate enterprise workflows and boost team productivity.
- Leveraged **LangGraph for stateful multi-agent orchestration** with ReAct and Plan-and-Execute strategies, enabling robust task decomposition and agent autonomy across multi-step workflows.
- Built production-grade **RAG chatbots** with Pinecone, FAISS, and Azure OpenAI embeddings; improved retrieval accuracy through hybrid search and reranking.

- Engineered modular **NLP pipelines for PII masking, NER, and entity linking** in the legal domain, strengthening data privacy and regulatory compliance.
- Fine-tuned **BERT and RoBERTa** on Vertex AI for conversational and classification tasks, enabling context-aware responses on domain-specific workflows.
- Integrated enterprise AI observability reducing debugging time by **~40%** through token-level execution tracing and retrieval diagnostics.

Data Engineer — Walmart (Contract)

Jun 2023 – Jul 2024

Bentonville, AR

- Built and maintained **PySpark / Scala ETL pipelines** processing **100GB+ daily**; deployed near-real-time streaming with Spark Streaming and GCP Pub/Sub.
- Optimized data models and warehouse architecture, **cutting storage costs ~25%** while improving query performance for analyst and data-science teams.
- Partnered with cross-functional teams to deliver tailored datasets, **accelerating insight generation ~60%** and improving data-driven decision-making.

Data Analyst — Capital Numbers

Jun 2019 – May 2021

Noida, India

- Built a Python-based **recommendation model** with custom feature engineering; designed Spark/Kafka ingestion pipelines for large-scale structured and unstructured data.
- Conducted EDA and built dashboards generating a **~25% lift in actionable insights** and automating model evaluation via custom Python tooling.

PROJECTS

AgentOps Harness

github.com/Ven-Z8/agentops-harness

Python · LangGraph · Typer · FastAPI · MCP · SQLite

- Built a **LangGraph control plane** for AI coding-agent workflows: repo scanner → planner → structured worker handoff → diff attribution → test runner → risk guard → evidence guard → PR-ready report. Supports Claude Code, Codex, Cursor, and Goose as interchangeable workers — the harness record stays stable regardless of which worker ran.
- **Evidence Guard** cross-checks every model-claimed file and test change against the real git diff — ungrounded assertions are flagged before the report is finalized, preventing hallucinated claims from entering the record.
- **Risk Scorer** independently weights changed files, sensitive paths touched, and test pass rate — runs scoring ≥ 80 are automatically blocked. Full run history in SQLite with a FastAPI query layer and MCP stdio server exposing `scan / run / get_report / list_runs` tools.

NL2SQL Viz

github.com/Ven-Z8/nl2sql-viz

Python · FastAPI · Anthropic Claude · Next.js · Vega-Lite

- End-to-end analytics product: schema introspection → Claude SQL generation → deterministic string guard (SELECT/WITH only) → Postgres execution → chart planner → Vega-Lite rendering.
- Security boundaries aligned with **OWASP AI guidelines**: Argon2 key hashing, AES-256-GCM DSN encryption, blocked mutating keywords (INSERT/UPDATE/DELETE/DROP/ALTER), multi-statement rejection.

ContextIQ

github.com/Ven-Z8/contextiq

Python · FastAPI · Qdrant · FastEmbed · Anthropic Claude

- Context-engineered RAG for complex enterprise documents; token-aware context packing, citation-preserving retrieval, grounded answer synthesis.
- Measured: **Recall@10 0.556 → 0.736** vs lexical-only baseline; MRR 0.667 → 0.735.

EDUCATION

M.S., Applied Computer Science

Aug 2021 – May 2023

Northwest Missouri State University

B.E., Electronics & Communications

Jul 2015 – May 2019

Andhra University