

VAMSHIDHAR SINGH

Applied AI Engineer

(205) 861 4554 | vamshidharsingh00@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

Applied AI Engineer with 5+ years of experience shipping AI and data systems into production, most recently running **autonomous LLM agents** that operate unattended at **Bank of America**. I design **agent orchestration** in **LangGraph**, **long lived memory and retrieval** over complex operational and financial data, and **evaluation harnesses** that decide which prompt, model and graph changes reach production. Strong **Python** engineer focused on reliability through checkpointed recovery, fallback routing, guardrails and deep tracing. I work closely with business stakeholders and engineering teams to improve deployed systems and maintain **open source** AI repositories.

TECHNICAL SKILLS

Agents & Orchestration: LangGraph, LangChain, LlamaIndex, CrewAI, MCP servers, LiteLLM, tool calling, multi agent state machines

Memory & Retrieval: RAG, Graph RAG, Pinecone, ChromaDB, LanceDB, Neo4j, Elasticsearch, embedding pipelines, learning to rank

Evaluation & Observability: RAGAS, DeepEval, LangSmith, Langfuse, Galileo, MLflow, prompt A/B testing, OpenTelemetry, Grafana, Datadog

Reliability & Safety: NeMo Guardrails, prompt injection defense, PII masking, RBAC, checkpointing, retries with backoff, dead letter queues, on call

Languages & Backend: Python, SQL, TypeScript, Java, C++, FastAPI, gRPC, REST APIs, SQLAlchemy, pytest, TDD

LLM & ML: PyTorch, RLHF, RLVR, fine tuning, vLLM, TensorRT LLM, NVIDIA Triton, KServe, XGBoost, LSTM, SHAP

Cloud & Infrastructure: AWS (Lambda, ECS, EKS, SQS, SNS, Step Functions, DynamoDB), Docker, Kubernetes, Terraform, GitHub Actions, GitLab CI

Data Platforms: PySpark, Kafka, Flink, Databricks, Snowflake, Postgres, Airflow, dbt, Delta Lake

PROFESSIONAL EXPERIENCE

Bank of America | June 2024 to Present

AI/ML Engineer (Data Focus)

- **Agent Orchestration:** Built unattended **multi agent workflows** in **LangGraph** with state machine routing across intent classification, **RAG** retrieval and compliance review agents, lifting intent resolution to **92%**.
- **Long Lived Memory:** Designed **checkpointed agent state** plus vector memory in **Pinecone** and **ChromaDB** over HR policy, compliance and operational records so sessions resume mid task, making knowledge lookup **40% faster**.
- **Failure Recovery:** Added timeouts, bounded retries with backoff, **LiteLLM** fallback models, idempotent tool calls and **dead letter queues**, letting failed runs replay from the last checkpoint while agent APIs held **sub second** latency.
- **Evaluation Harness:** Built a release gate in **CI/CD** scoring golden sets, **RAGAS** and **DeepEval** metrics and guardrail tests against the production baseline in **MLflow**, blocking any prompt, model or graph change that regressed quality, latency or cost.
- **Guardrails & Governance:** Enforced **NeMo Guardrails**, prompt injection defense, RBAC and data masking on financial data, including real time PII redaction at **99.8%** accuracy.
- **Retrieval Data Pipelines:** Ran **PySpark** and SQL pipelines on **Databricks** and **AWS** over millions of daily records, with **200+ dbt models** and **500+ tests** keeping agent context fresh and validated.
- **LLM Microservices:** Shipped **FastAPI** LLM services on **Docker** and **Kubernetes** summarizing ledger commentary, audit logs and contract terms, with tracing in **LangSmith** and **Langfuse** plus drift monitoring.
- **Workflow Automation:** Automated multi way transaction reconciliations and exception handling with Python agents and FastAPI services, cutting manual break investigation **35%**.
- **Streaming AI:** Fine tuned streaming speech models for financial jargon and noisy audio over **gRPC** and **Kafka**, reducing word error rate **22%** at under **200ms** latency.

Tata Consultancy Services | Feb 2020 to July 2023

Machine Learning Engineer

- **Conversational AI:** Built a transformer intent classification chatbot and **XGBoost** retention models for a higher education platform, deployed through **Docker**, Flask and **GitHub Actions** with automated tests on every release.
- **NLP Classification:** Classified **1,000+** survey entries and **250+** policy filings across **8+** jurisdictions into an AI ethics and privacy taxonomy, feeding **2** published research briefs.
- **Entity Resolution:** Built **PySpark** fuzzy matching and NLP feature extraction normalizing **5M+** listings into canonical IDs, on **Kafka** pipelines ingesting **10M+** daily transactions at **99.8%** pricing accuracy.
- **Real Time Data for ML:** Engineered **GCP** and **Airflow** pipelines processing **150M+** daily events at **5,000** events per second peak, cutting pipeline failures **35%** across **2,000+** daily workflows.
- **Recommendations:** Developed Next Best Action and propensity models over basket and clickstream data spanning **300K+ SKUs** and **600K+** daily transactions.
- **Optimization Models:** Built routing models over **10M+** GPS points with live weather and traffic feeds, delivering **12%** faster turnaround and about **\$400K** yearly fuel savings.

EDUCATION

Master of Science, Business Analytics, Golden Gate University

Bachelor of Technology, Information Technology, Jawaharlal Nehru Technological University-H, India

CERTIFICATIONS

AWS Certified Developer Associate (DVA-C02)

Microsoft Certified Azure AI Engineer Associate (AI-102)