NIRANJAN VSKS

Senior Agentic AI Engineer | Forward Deployed Engineering

niranjan.vsks@gmail.com +91 7893312735 Hyderabad, India niranjanvsks.xyz linkedin.com/in/niranjanvsks

github.com/niranjan-vsks

Summary

Senior Agentic AI Engineer with **7+ years** of expertise. I build **production agentic systems** inside the customer's own cloud, under their **identity**, **compliance constraints** and **data boundaries**. I own the **MCP** and integration surface, **deployment** and identity **architecture**, and the **evaluation**, **latency** and **cost** controls that govern them at scale. **Core backend and systems engineering** throughout. Blank page to production.

Experience

Senior Engineer - GenAI & Agentic AI, *Coforge* 03/2024 – Present | Hyderabad

- **Architected** and **shipped** a **SOC 2-compliant enterprise agentic platform** used by **17 QA teams**, building the **containerized** backend and **agent services** orchestrated on **Kubernetes** inside **customer-managed AWS, Azure and GCP** tenants; adapted compute sizing, identity through **Azure AD, IAM and SSO,** and **data boundaries** per cloud while retaining **one reference architecture** instead of customer-specific forks.
- **Fine-tuned LLaMA** and **Qwen** models with **LoRA** and aligned them with **RLHF**. For a **finance-sector** client, adapted their in-house **LLM** with **LoRA** so their own **model** generated test cases directly inside the platform, under the **data-residency constraints** that ruled out **hosted providers**.
- **Built the multi-model gateway and secure AI execution control plane**: provider routing, model selection and fallback, tiered routing, **context caching** and **prompt-payload** reduction; enforced tenant-aware RBAC and approval controls, with **per-tenant token and cost telemetry plus per-request attribution** for auditable operation.
- **Owned evaluation, observability and FinOps as one production control plane**: defined Acceptance Criteria Coverage, Test Design Coverage and **RAGAS** as standing quality thresholds, continuously monitored model and **retrieval quality** against them, and implemented **per-tenant cost telemetry with per-request attribution**; reduced token and serving cost through **context caching, prompt-payload reduction** and tiered **model routing.**
- Cut production **hallucination from ~15% to under 5%** with **GraphRAG** on **Neo4j** over live **Jira**, **Confluence , ServiceNow** and **Azure DevOps** data; built ingestion, **entity normalization**, **query transformation**, **reranking** and **context compression** so outputs remained grounded in linked requirements and acceptance criteria, with **retrieval** and response quality **tracked** continuously.
- **Lead engineer and architect** on the platform, **directing the engineering team** from architecture through delivery and owning the technical decisions end to end across **three cloud rollouts**; drove the **cross-team migration** to a **single reference architecture** across three clouds without formal authority over the teams involved, and set the **code review, RFC and release standards** the practice now works to.
- The practice's **AI technical authority**: **L1 technical interviewer** for **AI/ML, GenAI & Agentic AI** hiring across **50+ interviews**, setting the technical bar, and **mentoring engineers** on agent design, retrieval architecture and **evaluation** practice.

Data Scientist - GenAI, ML & Conversational AI, *Hewlett Packard Enterprise* 11/2022 – 12/2023 | Bangalore

- Architected an enterprise **document intelligence and retrieval system**, migrating from rule-based FAQ to **LLM-assisted RAG** over **~1,500-1,700 internal documents**: ingestion, structured understanding and grounded generation over unstructured enterprise content, cutting repeat user escalations **~40%**
- Built the **ETL and ELT pipelines** behind the retrieval system, ingesting unstructured enterprise content into a **governed lakehouse layer** on **Databricks**: **schema design**, incremental refresh, **data validation and reconciliation**, and **versioned, reproducible datasets** rather than a one-time load. Owned **throughput, latency and cost** on the pipeline end to end.
- Designed the **retrieval and ranking pipeline** end to end: chunking with **overlap calibration**, **hybrid search** combining **semantic and keyword** signals, **BM25** and metadata filtering tuned for enterprise-jargon queries. Added evaluation loops that measured **retrieval precision** against a curated held-out **evaluation** set, separating **retrieval** error from **generation** error, and improved response relevance **30-40% in pilot**.

Associate Software Engineer - Machine Learning & AI, *Mphasis* 07/2019 – 10/2022 | Bangalore

- Built **fraud-detection and anomaly-detection models** for financial services clients using tree based ensembles (**XGBoost, LightGBM, Random Forest**) with **SQL-based** feature extraction over **transactional data**, improving **precision ~15%** over legacy rule-based systems and cutting manual review volume.
- Owned the **production ML pipeline end to end**: feature engineering from raw transactional data, **training, validation against a held-out set, deployment and post-deployment monitoring**, with **scheduled retraining against feature drift** as transaction patterns shifted. Owned the classification metric decisions that make fraud models real: diagnosed the class imbalance where a model scores 99% accuracy and catches nothing, and chose precision as the operating metric because the cost asymmetry sat on false positives.
- Engineered the behavioural and transactional features from raw client data in **pandas**, across **scikit-learn, PyTorch and TensorFlow**, then **owned the models in production**: retraining against **drift** as transaction patterns shifted, and validating gains held on **held-out data** rather than the training split.

Education

Product Management in Generative & Agentic AI, *BITS School of Management (BITSoM)* 07/2025 – 03/2026 | India

PG Diploma in Machine Learning & Artificial Intelligence, *IIIT Bangalore* 06/2020 – 03/2022 | Bangalore

Bachelor of Technology in Computer Science, *Bharat Institute of Engineering and Technology* 08/2015 – 06/2019 | Hyderabad

Skills

Software Engineering & Cloud: Python | Go (Gin,Kratos) | React | TypeScript | FastAPI | Node.js | Full-Stack Solution Engineering | Advanced SQL | PostgreSQL | MongoDB | Redis | ETL Pipelines | Snowflake | Databricks | ServiceNow | Snowflake Cortex | Event-Driven and Async Architecture | Kafka | Microservice Architecture | Distributed Systems | Service Design and API Boundaries | Process Automation | Docker | Kubernetes | GitHub Actions | Jenkins | CI/CD | Terraform | Grafana | MLflow | AWS | AWS Lambda | Amazon S3 | GCP | Azure | Azure Key Vault | OpenTelemetry | Observability (Tracing, Metrics, Structured Logging) | FinOps and Cost Attribution | Code Review and Release Practice
Generative and Agentic AI: LLMs | VLLMs | MLOps | LLMOps | Agentic Workflows | Multi-Agent Orchestration | LangGraph | LangChain | CrewAI | OpenAI Agents SDK | Claude Code | Codex | LlamaIndex | MCP Servers | Tool and Function Calling | Structured Outputs | Prompt and Context Engineering | Langfuse | Model Gateway (Multi-Provider Routing, Fallback, Cost Attribution) | Agent Execution Observability | Agent Evaluation and Regression Harnesses | RAGAS | LLM Fine-Tuning (LoRA, QLoRA, RLHF) | TensorRT-LLM | TensorFlow | PyTorch | Guardrails, AI Safety and Governance
Retrieval & Data Architecture: Knowledge Graphs | GraphRAG | Neo4j | AWS Neptune | Entity Resolution | Hybrid Search | Semantic Querying | Reranking | Vector Databases (pgvector, FAISS, OpenSearch, Pinecone , Weaviate,Qdrant) | Agent Memory Architecture | Agentic RAG Pipelines | Citations & Grounding | Document Intelligence | Document Ingestion & Extraction | Classification | Palantir Foundry
Enterprise and Customer Facing: Forward Deployed Engineering | Customer Discovery | Stakeholder Alignment | Customer-Facing Technical Communication | Solutions Architecture | Enterprise Systems Integration (Dynamics 365, Jira, Confluence, Snowflake, Tableau, Power BI) | SOC 2 Compliance | Security Analytics (SIEM) | Customer Tenants Deployment (AWS, Azure, GCP) | Multi-Tenant Architecture | Identity & SSO Architecture | OAuth 2.0, OpenID Connect, JWT & Enterprise Authentication | RBAC Design | Financial Systems Integration | Approval Workflows & Policy Controls | Audit Trails | Regulated Environment Deployment (RBAC, audit, data residency)

Projects

Voice-First Financial AI Copilot, *Capstone, BITSoM | Architected, Built & Piloted*

- Designed a **multi-layer compliance architecture** with **SEBI advisory principles** and **DPDP requirements** as **first-class design constraints**: **RAG over a regulatory-safe knowledge base**, **system-prompt guardrails**, server-side governance, and **consent-first data handling**.
- Engineered a **central memory layer with advanced RAG** so user context, goals, and prior conversations persist **across sessions and across web and mobile** from a **single source of truth**.
- Integrated **Cashflow** and **asset-management , NOI , budgeting** frameworks in the application for end to end observability as well as robust delivery.

Autonomous Wealth Operating System, *Multi-Asset Agentic Application | Architect. Engineer. Owner* Ongoing

- Architected and built an **autonomous multi-asset wealth platform** (stocks, mutual funds, ETFs, bonds, NPS) around a **21-agent analyst council**: isolated agent opinions, a **devil's advocate stage**, aggregation weighted by each agent's **historical accuracy**, **ReAct and self-reflection** and a **code-enforced veto gate** so **no money moves on LLM output alone**.
- Engineered a **five-ring risk system** (position, portfolio, drawdown, budget, and connection-level **broker locks**) plus a **frozen-threshold evolution gauntlet**: any strategy change must beat the incumbent across **fixed statistical gates** before promotion, which rules out **silent strategy drift**.
- Delivered it as a **multi-surface platform**: a shared **framework-free core** behind web and mobile clients, a **Telegram copilot** for conversational instructions and live trade, investment and report updates, **user-selectable LLM models per agent role**, and wallets integrated with **Stripe** and **banking APIs**.

Loop Copilot: AI CRM Copilot for Enterprise Sales, *Architect · Engineer · Sole Owner | Production, Fortune 500 pilot*

- **Embedded with an active beta user inside a Fortune 500 sales org**, from customer discovery through V2, and shipped against what the reps actually did rather than the documented process: **CRM logging from 4–6 min to ~45 sec (~85%)**.
- **Turned one customer's integration into a product surface**: delivered Dynamics 365, then designed the **multi-CRM expansion shell for Salesforce, HubSpot and Zoho** so the next customer is a configuration rather than a build.
- Architected and delivered the **full-stack enterprise solution** across React 19, async **FastAPI, MongoDB, Azure OpenAI, Azure AI Search, Azure Key Vault, Entra ID through OpenID Connect and MSAL, Microsoft Graph** and CI/CD. Chose **structured event sourcing for an append-only, auditable record**, eliminating hallucinated recall while preserving cross-session state across 10+ feature areas.
- Shipped a **voice agent pipeline** that listens to **sales-client calls (STT**), turns them into structured **PRD** and onboarding documents, and **auto-detects** newly mentioned accounts. It resolves them against the account repository using unique ID, then logs them straight into **D365** or the connected **CRM**, so reps never have to manually key in account details after a call.

Agentic Codebase Intelligence System, *Multi-Agent Platform for Enterprise & Architectural Level Codebase Auditing* 06/2026 – Present

- Architected and shipped a **multi-agent codebase-intelligence platform** end to end that ingests **polyglot repositories** and builds a **persistent structural model** of the system (entities, **call paths**, **cross-file dependencies using ReAct**) that agents reason over, rather than retrieving flat file text. Cut manual review and audit effort **~50%**.
- Built the **verification harness** the agents run inside and the **closed feedback loop** around it: deterministic checks, a **taste gate** that fails output on quality rather than correctness, and **one guard per defect class**, so every defect found in production becomes a permanent automated check and the system's reliability compounds instead of regressing.
- Built the **durable execution layer**: a leased job queue with **exponential backoff, checkpointed resumption, idempotency keys and a single drain loop**, so a killed worker resumes rather than repeats. Designed the **budget and timeout arithmetic** across nested stages so no stage can authorise work it cannot finish.
- Built an **audit orchestrator** coordinating **6 specialized agents** (retrieval, reasoning, verification) with **MCP tool integration**, compressing time-to-findings from **~2 days of manual review to under 20 minutes** and auto-filing prioritized issues into **Jira**.

Certifications

Claude Certified Architect
Anthropic : In-Progress

Agile Practices Certificate
SAFe