

NISARG R. KUDGUNTI

+91-7020631336 | [LinkedIn](#) | [GitHub](#) | [Kaggle](#) | ✉ nisarg6502@gmail.com

WORK EXPERIENCE

Techolution — Associate AI Engineer (*Hybrid*)

February 2025 - Present

- **AI Vulnerability Remediation (Fortune 500 Media Client):** Architected a multi-agent system integrating Checkmarx PR/Push webhooks to detect and remediate SAST, SCS, and IaC vulnerabilities across 50 GitHub repositories, with threshold-gated parallelism per vulnerability type, surfacing AI-generated fixes in a dashboard.
- Implemented a repo-scoped feedback loop on AlloyDB vector store, embedding accepted, regenerated, and PR comment feedback per vulnerability with language and type metadata as pre-filters, retrieving top fixes via cosine similarity at inference time to progressively align outputs to the client's codebase.
- **Project Cost Estimator (Enterprise Retail Client):** Built a Hybrid RAG conversational agent over 4,000+ cost estimation forms, using whole-form chunking to preserve structural integrity, LLM-based query entity extraction for metadata pre-filtering before retrieval, and a key-aware citation system to provide field-level citations for generated responses.
- **Construction Schedule AI Insights (Enterprise Retail Client):** Reduced manual report generation from 2-3 days to under 1 minute via an NL-to-SQL agent over 8,000+ construction forms, consolidating 80+ tables to 20 UI-mirrored tables and views to reduce LLM hallucination, retaining only form-visible columns.
- Engineered a multi-tool pipeline (table selector, SQL writer, validator, executor, LLM summarizer) with table-level schema metadata for table selection, DML blocklist validation, and cross-project comparisons, trend analysis, and XLSX export.
- **AI Knowledge Builder:** Building a multi-agent platform reducing client project delivery from 3–4 weeks to 3–4 days via automated data categorization and segmentation. Developed gap resolution with impact analysis that identifies low-confidence AI outputs, guides structured resolution, and applies fixes across dependent segments.

Techolution — AI Intern (*Onsite*)

July 2024 - January 2025

- Built Project Analyser, a 10+ language GitHub repository analysis pipeline benchmarked at 80k files against the Linux kernel, using Qwen on Cloud Run via vLLM for per-file summarization and Gemini on Vertex AI for executive summaries, service extraction, and dependency graphs.
- Extended the pipeline with HDBSCAN SOM-based clustering to derive hierarchical feature taxonomies and LLM-generated JSON schemas rendered as ReactFlow architecture diagrams.
- Reduced VertexAI inference cost by 90%+ by replacing proprietary LLMs with Qwen-2.5-Coder on Cloud Run via vLLM.

AlgoAnalytics — Data Science Intern (*Remote*)

February 2024 - June 2024

- Enhanced a RAG application using LangChain, FAISS, LLM reranking, and innovative chunking strategies to improve RAGAS metrics and Generative AI performance.

PROJECTS (LINK ON TITLE)

- **Custom Llama-3 150M PII Redactor** - Designed a 150M-parameter autoregressive transformer from scratch in PyTorch with Llama 3 internals (RoPE, RMSNorm, SwiGLU), pretrained on 1.8M documents, then supervised fine-tuned on a 200k PII dataset for on-premise PII redaction, enabling a local inference pipeline that scrubs sensitive data before it reaches any third-party LLM API.

SKILLS

- **Frameworks & Libraries:** LangChain, LangGraph, LlamaIndex, PyTorch, Flask, FastAPI
- **AI Systems:** Multi-Agent Systems, Retrieval Augmented Generation (RAG), LLM Fine-tuning, Prompt Engineering, Model Context Protocol (MCP), LLM Observability & Tracing (Langfuse)
- **Others:** Google Cloud Platform, Microsoft Azure, AWS, Docker, Ollama, vLLM
- **Languages and Databases:** Python, SQL, MongoDB, Redis, AlloyDB, FAISS, ChromaDB, Pinecone, Qdrant, Neo4j

EDUCATION

Savitribai Phule Pune University

January 2021 - May 2024

Pune Vidyarthi Griha's College of Engineering and Technology (PVGCOET)

CGPA **8.82**

Bachelor in Computer Engineering — Honors in Artificial Intelligence/Machine Learning

ACHIEVEMENTS

- **TEDxPVGCOET** – Led as the *Licensee and Organizer*, a 60-member team, hosted 8 distinguished speakers on theme Punarutthan (revival of ancient art forms and ideas).
- Passed all requirements for Microsoft Certified: Azure AI Fundamentals (AI-900). [Credential](#)