

CHIRANJEEVI GUNDU

Tempe, AZ | +1 (415) 816 7939 | chiranjeevigundu1@gmail.com | linkedin.com/in/chiranjeevigundu | chiranjeevigundu.github.io

SUMMARY

Forward Deployed AI Engineer with an M.S. in Computer Science (AI track) and four years of professional engineering. Makes AI quality measurable for enterprise clients — continuous evaluation loops (CI/CD for LLMs) on task-specific golden datasets, graded for relevance, faithfulness, correctness and coherence, so a quality regression fails a build instead of reaching users. Ships with backend rigor: typed FastAPI services, PostgreSQL data modeling, automated tests, and secure cloud and on-prem deployment.

PROFESSIONAL EXPERIENCE

Forward Deployed AI Engineer

Aug 2025 – Present

Axitem Software Solution Inc.

Remote, USA

- Made AI quality measurable for enterprise clients: designed and shipped continuous evaluation loops (CI/CD for LLMs) that score relevance, faithfulness, correctness and coherence of RAG and agent workflows on every change, so regressions fail the pipeline instead of reaching users.
- Turned vague "the answers are wrong" complaints into tracked, reproducible defects by building rubric-based grading pipelines in Python pairing deterministic PyTest checks with LLM-as-a-Judge, isolating and closing P1 hallucination bugs.
- Exposed retrieval failures that leaderboard benchmarks missed by curating task-specific Golden Datasets with client domain experts in backlog refinement workshops, replacing generic benchmarks with the client's own ground truth.
- Closed the loop from production back into the test suite: captured live user feedback and request traces, triaged edge cases in backlog refinement, and promoted each into the Golden Dataset so every reported failure became a permanent regression case.
- Shortened time-to-value on client AI features across 2-week sprints by catching failure modes — missed retrieval, prompt drift — before UAT rather than after release.
- Put grounded, source-cited answers in front of support teams daily: delivered RAG document-Q&A assistants (LangChain, OpenAI API) embedded inside the portals they already worked in, replacing hours of manual lookup.
- Unlocked privacy-sensitive enterprise deals by deploying evaluated open-source models (Llama 3, Mistral) entirely on client-owned infrastructure, with no client data leaving their environment.
- Cut deployment failures and reduced cross-cloud migration to a configuration change by standardizing delivery on a Dockerized FastAPI template with eval telemetry built in.

Student Worker — Administrative & Marketing Support

Aug 2024 – May 2025

Sodexo (on-campus role during M.S.)

Saint Louis, MO

- Drove AI tool adoption across the administrative team, demonstrating and training colleagues on Microsoft Copilot in Office applications and ChatGPT for planning, documentation, and drafting tailored communications, so the team completed routine work faster and took on additional activities.
- Gave management actionable themes instead of raw comment dumps by using AI assistants (ChatGPT, Gemini, spreadsheet AI) to cluster student dining feedback from portals, suggestion boxes and app reviews, and cut daily data-entry time with spreadsheet formulas and macros.
- Programmed daily digital menu boards across dining halls and retail locations, verifying dietary and allergen labeling was accurate before every service.

Backend Software Engineer

Aug 2021 – Jul 2023

Tata Consultancy Services (TCS)

Chennai, India

- Cut manual ticket-sorting time 40% for a Fortune 500 client over a 6-sprint timeline by integrating an ML text-classification model into their IT support system via Python (FastAPI) REST APIs.
- Designed a confidence-threshold fallback mechanism, ensuring low-confidence predictions defaulted to the manual queue and passed strict User Acceptance Testing (UAT).
- Improved mission-critical client workflow performance 60% by utilizing EXPLAIN/ANALYZE to profile PostgreSQL queries, adding targeted indexes, and resolving P2 timeout bugs.
- Enforced an Agile "Definition of Done" by building comprehensive PyTest unit-test suites, holding code coverage above 85% and cutting deployment failures by 30% through GitHub Actions CI/CD.
- Translated business-language Jira requirements from product owners into scoped technical tasks, enabling rapid development and reducing technical debt.

TECHNICAL SKILLS

AI Evaluation: LLM Evaluation, Eval Set & Golden Dataset Design, Rubric-Based Grading, LLM-as-a-Judge, Hallucination & Faithfulness Detection, Grounding & Citation, MRR / Recall@k / nDCG, Regression Gating in CI, Guardrails AI

Retrieval & LLM Systems: Retrieval-Augmented Generation (RAG), Hybrid Retrieval, Rank Fusion (RRF), Cross-Encoder Reranking, Structure-Aware Chunking, Embeddings, Semantic Search, Vector Search (pgvector, Pinecone), Postgres Full-Text Search, Prompt Engineering, Context Optimization, Tool-Calling Agents, Model Context Protocol (MCP), LangChain, LangGraph

Models & Inference: PyTorch, ONNX Runtime, Hugging Face, On-Device & Edge Inference, Ollama, LiteLLM, Llama 3, Mistral, OpenAI API, Anthropic Claude, OpenAI Vision, Fine-Tuning Fundamentals (LoRA, QLoRA), Scikit-learn, pandas, NumPy

Backend & Infrastructure: Python, SQL, Bash, FastAPI, Pydantic, REST APIs, PostgreSQL 16, pgvector, Redis, Row-Level Security, SQL Profiling (EXPLAIN/ANALYZE), ETL & Ingestion Pipelines, Docker, Docker Compose, AWS (ECS Fargate, RDS, S3, Secrets Manager, CloudWatch, VPC), AWS CDK (TypeScript), Azure, GitHub Actions, CI/CD, Git, PyTest

PROJECTS

One system across five open-source repositories: Aadyon Assist calls hybrid-rag over HTTP for retrieval and shares the llmkit model layer with Synapse, while floci-lab deploys both services to AWS. hybrid-rag holds no model itself — retrieval and measurement only.

hybrid-rag — Hybrid Retrieval Engine with a Measured Eval Harness (github.com/chiranjeevigundu/hybrid-rag)

- Built a hybrid retrieval engine fusing a dense arm (BAAI/bge-base-en-v1.5, 768-dim, served through ONNX Runtime) with a Postgres full-text arm using ts_rank_cd cover density, fused by Reciprocal Rank Fusion at k=60 over ranks, not raw scores.
- Demonstrated the two arms have opposite blind spots instead of assuming it: on a committed 22-case golden set the lexical arm returns 0.000 MRR on every paraphrase query, while fusion lifts exact-identifier retrieval from 0.917 to 1.000 — 0.920 MRR hybrid against 0.898 dense and 0.273 lexical.
- Measured cross-encoder reranking and shipped it disabled because it cost 0.032 MRR overall and 0.061 on paraphrase, retaining both the code path and the recorded measurement, so the default is a result rather than a guess.
- Wrote a reproducible evaluation harness computing MRR, recall@k and nDCG@10 positionally with no LLM judge, wired to a CI regression floor so a retrieval-quality drop fails the build instead of reaching a user.

Aadyon Assist — Self-Hosted AI Life-Ops Platform (github.com/chiranjeevigundu/aadyon-assist)

- Architected and open-sourced (MIT) a self-hosted AI platform end to end — data model, tool-calling agent layer, ingestion, REST API and installable PWA — running as six Docker Compose services on FastAPI and Postgres 16.
- Built a bounded agent loop capped at an explicit step limit, persisting every tool call, its arguments and its result to an append-only message log for full traceability and replay.
- Engineered a human-in-the-loop safety model where read-only analysis runs autonomously while any action with a real-world side effect becomes a proposal the user must approve, so no irreversible action executes on its own.
- Designed multi-tenant isolation with Postgres Row-Level Security forced at the database across 20 tables, running the API as a restricted role rather than the schema owner, since Postgres exempts owners from RLS.

Additional repositories: llmkit (Apache-2.0 shared LLM layer — tier-based routing over LiteLLM, one traced chat() chokepoint); floci-lab (AWS CDK deploying both services to ECS Fargate with zero NAT gateways); Synapse Storage System (vision-model document triage).

EDUCATION

M.S. Computer Science — Artificial Intelligence Track
Saint Louis University

Aug 2023 – May 2025
St. Louis, MO

B.Tech, Electronics & Communication Engineering
Jawaharlal Nehru Technological University

Jul 2017 – Aug 2021
India