

ARUN KUMAR CHUKKALA

Applied AI Engineer — LLM Systems · Agentic Workflows · Production AI

San Francisco, CA | +1 (409) 549-3003 | arunkiran721@gmail.com | github.com/arun3676 | arun7.com

SUMMARY

Applied AI engineer who ships production Large Language Model (LLM) systems end to end: agentic workflows with tool calling, Retrieval-Augmented Generation (RAG), evaluation harnesses, and guardrails. Builds the unglamorous layer that makes AI safe in production — evals that gate releases, cost and token budgets, grounded refusals instead of hallucinations, and observability that turns failures into regression tests. Fluent in agentic development tooling (Claude Code, Cursor, Codex, Model Context Protocol) as a daily engineering workflow. Live demos of every project at arun7.com.

EXPERIENCE

IBM — *Applied AI Engineer*

United States · Mar 2024 – Present

- Design and maintain evaluation infrastructure for LangGraph-based agent workflows in a regulated enterprise environment — golden datasets, regression gates, and human-in-the-loop (HITL) review — reducing manual review time by 30%.
- Build production observability that traces agent execution, tool calls, and failure modes end to end, converting production failures into permanent regression test cases.
- Reduced agent workflow latency by 62% (850 ms to 320 ms) through asynchronous execution and semantic caching of Large Language Model (LLM) calls, cutting redundant token spend on repeated queries.
- Develop and deploy Dockerized Python/FastAPI services that orchestrate LLMs, internal tools, and enterprise data through CI/CD pipelines, maintaining 98%+ uptime with retry, timeout, and fallback handling around non-deterministic model calls.
- Partner with non-technical stakeholders to translate operational workflows into AI solution requirements, documenting agent behavior, decision logic, and failure modes for the teams that adopt them.

Infosys — *Software Engineer, Applied AI / Risk Platform*

Hyderabad, India · Jan 2021 – Dec 2022

- Engineered ensemble anomaly-detection workflows for fraud decisioning, reducing false positives by 18% on a regulated risk platform by tuning precision against a labeled outcome set.
- Designed and deployed Dockerized FastAPI microservices and Azure Data Factory ETL pipelines processing 15M+ records for risk assessment and monitoring.
- Improved machine learning (ML) classification workflow efficiency by 25% and implemented output-consistency monitoring that alerted on model drift before it reached downstream decisions.

PROJECTS

Cork — *Multi-Tenant AI Retail Operations Platform* (app.poscork.com)

Jan 2026 – Present

- Architected and shipped a production multi-tenant platform — React, TypeScript, Node.js/Express, Prisma ORM, PostgreSQL on Vercel/Render/Supabase — with Sentry observability, CI/CD release gates, third-party API integrations (payments, messaging, external point-of-sale sync), and REST APIs secured by JSON Web Token (JWT) authentication and role-based access control (RBAC).
- Built an AI document-intelligence pipeline: Gemini Optical Character Recognition (OCR) structured JSON extraction feeding a five-stage matching cascade (UPC → item number → learned aliases → fuzzy matching → 768-dimension embedding vector search on pgvector), with human-in-the-loop confirmation so model output never mutates system-of-record data autonomously.
- Designed an offline evaluation harness over 25 real-world documents (42 ground-truth line items): 100% suggestion accuracy, zero false positives — enforced as 15 regression tests that block any release degrading quality; plus a conversational analytics agent with typed tool schemas, grounded refusal, per-tenant agent spend caps, and an audit trail of every agent action.

Selected AI Systems

2026 · live demos at arun7.com

- Claim Guardian** — AI medical-bill auditor (claimguardian.arun7.com): document-AI extraction priced deterministically against the CMS Medicare fee schedule (~17k codes); CI-gated evals — pricing 14/14, extraction recall 1.00, zero fabricated citations.
- Agentception** — AI job-search platform (agentception.vercel.app): LLM evaluation harness over 69 labelled job descriptions (match ranking AUROC 0.702, $p = 0.017$), replayed offline in CI; the evals surfaced five bugs already in production.
- Code Tribunal** — multi-agent verification of AI-generated code (code-council.vercel.app): seven agents cross-examine each diff against its ticket, emitting a deterministic 0–100 trust score that gates CI; web app, CLI, and MCP server.
- FinAgent** — compliance-grade financial RAG (finagent-landing.vercel.app): six-agent LangGraph workflow over SEC filings — hybrid BM25 + embedding retrieval, SHA-256 hash-chained audit trail, runtime data-egress governor; ~\$0.0005/query.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, SQL

AI / LLM Systems: LLM APIs (Anthropic Claude, Google Gemini, OpenAI), agent loops & tool calling, tool schema design, Retrieval-Augmented Generation (RAG), embeddings & vector search (pgvector), LangGraph, prompt engineering, structured JSON outputs, context-window management, token & cost optimization, evaluation harnesses & golden datasets, regression gates, guardrails & grounded refusal, human-in-the-loop (HITL), semantic caching, LLM observability & tracing, OCR / document intelligence, conversational AI & natural-language analytics

Agentic Development: Claude Code, Cursor, Codex, Model Context Protocol (MCP), multi-agent code review, AI skills / internal tooling

Engineering: FastAPI, Node.js/Express, React, PostgreSQL, Prisma ORM, Supabase, Docker, CI/CD, Git/GitHub, REST APIs, ETL pipelines, Azure, Vercel, Render, Sentry

EDUCATION

Lamar University — M.S. in Computer Science

Beaumont, TX · Dec 2024

Sri Indu Institute of Engineering & Technology — B.Tech. in Computer Science

Hyderabad, India · May 2020