

Indra Prakash Gottipati

AI / Platform Engineer — Retrieval Systems · LLM Infrastructure · Multi-Cloud Deployment

Guntur, India · +91 63026 83750 · indraprakashgottipati61@gmail.com

linkedin.com/in/indra-prakash-gottipati-6a3b13226 · github.com/indra7777

SUMMARY

AI engineer who ships retrieval systems that are measured, not asserted. Built the evaluation harness and confidence-gated LLM cascade that took a production government search service from 40% to 88% answered-accuracy while cutting wrong answers from 46 to 2. Works end to end: .NET 10 and Python/FastAPI services, self-hosted LLM serving, Terraform-provisioned multi-cloud tenants, and the CI/CD to ship them.

EXPERIENCE

Unnanu Inc. — Austin, TX (remote) · Jan 2025 – Present

Lead AI Engineer (Oct 2025 – Present) · AI Engineer / Software Engineering Intern (Jan 2025 – Oct 2025)

- **Raised answered-accuracy from 40% to 88%** on a curated 85-query gold set for a **US state government agency's** public search service, cutting wrong answers 46 → 2 and lifting abstain-recall 18% → 97%. Designed a confidence-gated LLM catalog cascade in **.NET 10 / EF Core** where the model selects catalog rows rather than free-writing SQL and abstains honestly below threshold — shipped with a zero-false-answer release gate.
- **Built the retrieval evaluation discipline** behind that result: gold-set construction, p@1/p@3, facet-F1 and abstain-recall scoring, pooled candidates, LLM adjudication. Ran a **four-lane ablation** (rules / weighted catalog / LLM-authored SQL / vector) measuring vector retrieval at **97% corpus reachability vs the catalog's 35%** — but committing **36 wrong answers on 85 queries** — then shipped the composition, vector recall behind an evidence gate, over either lane alone.
- **Shipped multi-cloud tenant provisioning** so customers deploy a search project with no engineer on the call: ~24k LOC .NET orchestration service plus a Terraform-driven provisioning worker targeting **GCP Cloud Run, AWS and Azure** over a tri-dialect schema (SQL Server / PostgreSQL / MySQL). Added per-request search tracing and a per-call **LLM spend ledger with budget guardrails**.
- **Built a five-modality ingestion platform** (~56k LOC Python/FastAPI) for document, audio, image, video and web corpora — Docling with Tesseract/RapidOCR and Camelot extraction, Whisper plus pyannote diarization, Gemini enrichment — on **RabbitMQ** with dead-letter recovery and a worker supervisor. Cut a 100-file mixed batch from **~5 hours sequential to 45–90 minutes**.
- **Ran self-hosted LLM serving** on a 5-node Apple Silicon **exo/MLX** cluster (~52 GB pooled; Qwen3-30B and Qwen2.5-VL-72B 4-bit) behind an OpenAI-compatible endpoint consumed by production .NET jobs; added a LoRA fine-tuning and MLX conversion pipeline.
- **Owned platform and delivery**: migrated the core API **ASP.NET Web API 5 / EF 6 / .NET Framework 4.6.1** → **.NET 8** → **.NET 10** (64 projects) and rebuilt the **React 16 / Next 6** job board as **React 19 / Next 16** with strict TypeScript and a legacy-parity suite. Across ~60 Azure Pipelines and GitHub Actions: BuildKit layer caching to **~2.5-minute builds**, Key Vault with managed identity and **OIDC federated deploys** replacing long-lived secrets, daily Snyk/CodeQL/npm-audit, and ~310 xUnit plus 130 Vitest/Playwright tests behind an 80% coverage gate.

SELECTED PROJECTS

- **Unnanu Talent Slack App** (Slack Bolt 4) — apply to jobs, edit profile and upload résumés entirely inside Slack; re-architected 6 slash commands into a button-driven App Home surface, with OAuth 2.0 multi-workspace distribution and a dual-manifest UAT/production pipeline.
- **Agent Fleet Orchestrator** — Slack-driven scheduler routing engineering tasks to coding agents across a Mac-mini fleet: capacity- and priority-aware placement, SSH execution via a jump host, human approval gates.
- **PDF Accessibility Assessor** — automated **WCAG 2.1 AA / Section 508** gap analysis over PDFs (PDF/UA tagging, alt text, heading semantics, reading order), producing severity-scored compliance reports with LLM-generated remediation guidance.

SKILLS

Retrieval & LLM · RAG, hybrid search, BM25, cross-encoder reranking, BGE-M3, text-embedding-3-large, Qdrant, Chroma, FAISS, Azure AI Search, layout-aware chunking, eval harnesses (p@k, nDCG, abstention)

LLM Serving & Ops · exo, MLX, Ollama, LoRA fine-tuning, Groq, Azure OpenAI, Gemini, Whisper, spend/latency observability

Languages · Python, C#, TypeScript, JavaScript, SQL

Backend · FastAPI, ASP.NET Core (.NET 8/10), EF Core, LangGraph, RabbitMQ, Azure SQL, PostgreSQL, MySQL, Redis

Frontend · React 19, Next.js 16, Tailwind, Redux Toolkit, React Query

Cloud & DevOps · Azure (App Service, Key Vault, ACR, Container Apps), GCP (Cloud Run, Artifact Registry), AWS, Terraform, Docker, Azure DevOps, GitHub Actions, Playwright, xUnit

EDUCATION

B.Tech, Computer Science & Engineering (IoT) — R.V.R. & J.C. College of Engineering · 2021 – 2025 · CGPA **9.30 / 10**