

Dileep Reddy Battu

Software Engineer | AI/ML Engineer | LLM Applications, RAG & Agents

Sanford, FL | (928) 814-7228 | dileep151015@gmail.com | linkedin.com/in/dileepreddy27 | github.com/dileepreddy27 | dileepreddybattu.com

SUMMARY

Software engineer with 3+ years of experience building production LLM applications, data pipelines, and backend services in healthcare (teleradiology). Shipped an LLM-powered clinical assistant and GCP data services that process 20K+ imaging records daily. Builds end-to-end AI systems (RAG, agentic workflows, PyTorch models) with evaluation, audit trails, tests, and CI. M.S. in Computer Science; pursuing an M.S. in Data Science & AI.

TECHNICAL SKILLS

AI / ML: LLMs, RAG, AI Agents, LangChain, LangGraph, LlamaIndex, OpenAI API, Prompt Engineering, Embeddings, Vector Search (pgvector, Qdrant, LanceDB), PyTorch, MONAI, Keras, Reinforcement Learning (PPO), Model Evaluation

Languages: Python, TypeScript, JavaScript, SQL, Java, Go, C++

Backend & Frontend: FastAPI, Flask, Node.js, Express, Spring Boot, React, Next.js, REST APIs, WebSockets, JWT / OAuth 2.0

Data & Cloud: PostgreSQL, BigQuery, Supabase, Redis, Kafka, MongoDB, GCP (Cloud Run, Cloud Scheduler, Cloud Storage), Docker, GitHub Actions CI/CD, Git, pytest

PROFESSIONAL EXPERIENCE

Software Engineer, Vesta Teleradiology | Lake Mary, FL May 2025 – Present

- Built retrieval and prompt workflows for an **LLM-powered clinical assistant** (LangChain, OpenAI, React, TypeScript, OAuth 2.0) supporting radiology operations teams.
- Deployed Python data services on **GCP** (BigQuery, Cloud Run, Cloud Scheduler, Docker, CI/CD) processing **20K+ imaging records daily**, saving **300+ hours/month** of manual work and cutting data retrieval latency by **70%**.
- Designed Node.js and Python integration services connecting web apps to **PACS**, PostgreSQL, BigQuery, and external REST APIs.
- Evaluated LLM-enabled tools through prompt engineering, **model-output validation**, and monitoring before rollout to operations.
- Owned audit logging, SQL optimization, API testing, and production monitoring in **HIPAA-conscious** workflows; led root-cause analysis for database, API, and authentication incidents.

Skills: Python, SQL, LLMs, RAG, LangChain, OpenAI API, Prompt Engineering, React, Next.js, TypeScript, Node.js, Flask, REST APIs, OAuth 2.0, PostgreSQL, Supabase, BigQuery, GCP (Cloud Run, Cloud Scheduler, Cloud Storage), Docker, CI/CD, Git, PACS, HIPAA

Software Engineer Intern, Vesta Teleradiology | Lake Mary, FL Aug 2024 – May 2025

- Built API-driven reporting, provider-management, and validation features covering **300+ providers and 1,200+ facilities**.
- Extended RadMapping+ and CredMapping+ with UI components, backend services, and automation, reducing workflow overhead by **37%** for **90+ active users**.

Skills: Python, Flask, React, Next.js, TypeScript, Node.js, PostgreSQL, Supabase, BigQuery, SQL, REST APIs, Looker Studio, Git, Debugging

Graduate Teaching Assistant, Northern Arizona University | Flagstaff, AZ Aug 2023 – May 2024

- Led programming labs and debugging sessions for computer science courses on algorithms, data structures, and software engineering.

Skills: Python, Java, Data Structures, Algorithms, Debugging, Git

Software Engineer, Calibrage Info Systems | Hyderabad, India Jan 2022 – Dec 2022

- Built SmartLeads, a CRM platform (Next.js, TypeScript, Node.js, Express, PostgreSQL) for lead tracking and sales-pipeline analytics.
- Delivered SLKW, an e-commerce platform with auth, catalog search, cart, and payments, improving customer engagement by **30%**.

Skills: React, Next.js, TypeScript, JavaScript, TanStack Query, Node.js, Express, PostgreSQL, MongoDB, Tailwind CSS, REST APIs, Agile

AI / ML PROJECTS

EvidenceRAG: Auditable RAG Platform | *FastAPI, Next.js, PostgreSQL/pgvector, Docker* [GitHub](#)

- Hybrid retrieval that fuses vector cosine, keyword, and phrase scores with reranking and an evidence threshold; returns grounded citations, confidence, and trace IDs, and refuses when evidence is weak. Includes an offline evaluation runner and JWT/Argon2 auth.

RepoForge: Autonomous Code-Repair Agent | *LangGraph, Python AST, Docker, PostgreSQL* [GitHub](#)

- Issue-to-patch agent with a bounded LangGraph repair loop (up to 5 attempts), AST code context, sandboxed test execution in Docker, an event ledger for every step, and draft pull requests created only after the test suite passes.

MedImage Triage Lab: 3D Medical Image Segmentation | *PyTorch, MONAI, pydicom, SQLite* [GitHub](#)

- Trained a custom PyTorch 3D CNN with MONAI Dice + BCE loss on synthetic DICOM series; benchmarked against a threshold baseline on held-out data and fed predictions into an auditable triage queue with model SHA-256 provenance.

CodeContext MCP: Code Search for AI Agents | *Go, Tree-sitter, LanceDB, MCP* [GitHub](#)

- Model Context Protocol server that indexes Python, Go, and Rust code into AST-bounded chunks with a parallel Go worker pool and LanceDB vector search, returning cited snippets under a caller-set context budget.

EDUCATION

M.S., Data Science and Artificial Intelligence, Campbellsville University In progress

M.S., Computer Science, Northern Arizona University | GPA 3.5/4.0 May 2024

B.Tech., Computer Science and Engineering, JNTU Hyderabad May 2022