G. NITESH KUMAR REDDY

niteshreddy257@gmail.com — LinkedIn — **8249829096**

SKILLS

LLM Evaluation Frameworks, **LLM Observability**, **Agentic Workflows**, Distributed Tracing, Span Processing, Telemetry Pipelines, **OpenTelemetry**, LangChain, AWS Bedrock, Google Vertex AI OpenAI APIs, Anthropic Claude, **Python**, **TypeScript**, **Node.js**, Golang, **ClickHouse**, **PostgreSQL**, MongoDB, **Docker**, **AWS (EC2, S3)**, Microservices Architecture, CI/CD, Data Structures & Algorithms

EXPERIENCE

Software Engineer Sep 2025 – Present
BrowserStack *Mumbai, Maharashtra*

- Worked on **AI Evals & LLM Observability**, architecting and developing a dual-language **AI Observability SDK (Node.js/TypeScript & Python)** enabling zero-code auto-instrumentation of LLM applications using OpenTelemetry across OpenAI, Anthropic, AWS Bedrock, Google Vertex AI, and LangChain.
- Engineered a **custom telemetry processing pipeline** with intelligent span filtering, caller-context extraction via stack inspection, distributed trace propagation, and declarative decorators (`@workflow`, `@agent`, `@task`, `@tool`) for structured tracing of agentic workflows.
- Built a scalable **LLM evaluation platform** supporting 20+ evaluation metrics including faithfulness, answer relevancy, contextual precision/recall, bias detection, safety scoring, RAGAS, and G-Eval, with asynchronous job processing and concurrent workers.
- Designed a **multi-variant experimentation framework** with dataset versioning, prompt lifecycle management, batch evaluation, human-review annotation queues, and cost-aware model comparison across quality, latency, and inference/evaluation costs.
- Subsequently worked on **Test Companion**, an AI coding agent for testers similar to Claude Code, focusing on improving production agent **performance, reliability, and cost** through telemetry-driven analysis and optimization.
- Reduced agent execution cost and latency by removing an optimization that increased model calls, taking a representative workflow from **15 → 4 model calls, $0.372 → $0.229 per task, and 149s → 36s runtime**.
- Optimized prompt-cache usage using production telemetry, **halving cache keepalive and saving approximately $240–300/month** while preserving warm-resume performance; evaluated the trade-off and rejected further reduction due to its impact on resume reliability.
- Diagnosed and improved agent reliability by identifying that **58% of production tasks repeated identical tool calls**, uncovering eight distinct looping root causes, and building safeguards including **parallel-safety gates, tool-call batching, and completion validation** to prevent unsafe execution and false task completion.
- Root-caused a production agent failure where a task **read the same file 254 times without completing**, identifying a deduplication flaw caused by caching only on file path instead of file path and line range; reproduced the issue at the wire level and fixed the underlying read-tracking logic.

Software Development Engineer 1 Aug 2024 – Aug 2025
Credflow *Noida, Uttar Pradesh*

- Implemented intelligent **LLM agents using LangChain and Google Vertex AI** to automate credit underwriting workflows by parsing and analyzing complex GSTR-3B CAM sheets containing 10–15 subsheets. Integrated the system with **Kafka for event-driven processing** and automatically updated summaries in the database, improving operational efficiency and decision turnaround time.
- Designed and developed backend services for **Loan Origination System (LOS)** and **Loan Management System (LMS)** in a fintech environment using **Node.js, Express.js, TypeScript, MongoDB, and PostgreSQL**.

EDUCATION

Veer Surendra Sai University of Technology, Sambalpur, Odisha Dec 2020 – May 2024
Bachelor of Technology - Computer Science and Engineering, CGPA (8.52)