

Prateek Hitli

Bengaluru, Karnataka | +91-9606861693 | hitliprateek6@gmail.com
linkedin.com/in/prateek-hitli | github.com/Prateek1771 | prateekhitli.dev

Summary

AI Engineer with a B.E. in Artificial Intelligence and Machine Learning and production experience building LLM applications, RAG pipelines, multi-agent systems, AI-powered automation, and reliable backend services. Experienced with OpenAI, Claude, and Gemini APIs, LangChain, LangGraph, FastAPI, Python, Docker, and AWS. Focused on production AI engineering including evaluation, failure isolation, observability, security, guardrails, and scalable system design.

Technical Skills

AI Engineering: Agentic AI, Multi-Agent Systems, Agent Orchestration, RAG, Agentic RAG, Tool Calling, Function Calling, MCP, Prompt Engineering, LLM Evaluation, AI Guardrails, Opentelemetry, Logfire

LLMs & APIs: OpenAI API, Claude API, Gemini API, Groq, OpenRouter, Hugging Face, LoRA, QLoRA, Prompt Engineering

AI Frameworks: LangChain, LangGraph, NeMo Guardrails, Vercel AI SDK

Machine Learning: Transformers, Attention Mechanisms, PyTorch, TensorFlow, Computer Vision, OpenCV, YOLO

Retrieval: Pinecone, pgvector, Embeddings, Semantic Search, Retrieval Optimization

Backend & Databases: Python, TypeScript, FastAPI, Node.js, Express.js, REST APIs, PostgreSQL, MongoDB, DynamoDB, Redis

Cloud & Infrastructure: AWS, ECS/Fargate, SQS, Docker, Terraform, Git, GitHub Actions, CI/CD, Vercel

Testing & Evaluation: Pytest, Vitest, Ruff, Mypy, ESLint, Unit Testing, Integration Testing, Security Testing, Retrieval Evaluation, Response Evaluation

Experience

Junior AI Developer

DigiBull AI

Aug 2025 – Present

Bengaluru / Mysore, India

- Led the redesign of an Automated Optical Inspection system for industrial cable assembly, replacing a YOLO-based pipeline with an OpenCV detection system and increasing inspection accuracy from 65% to 99% while reducing deployment hardware requirements.
- Designed and implemented a multi-step agentic workflow that extracts required fields from inspection data and generates standards-conformant reports across multiple ISO guideline sets.
- Led development of an AI operations console proof of concept with role-based access control, audit logging, RAG-backed reporting, and AI guardrails; presented the system on-site to the client.
- Investigated a production API-key security incident involving 166M+ tokens and approximately \$758 in usage, identifying a CLI authentication path that allowed API-key revocation controls to be bypassed.
- Engineered a multi-platform AI chatbot integrating WhatsApp, website, and Instagram to provide conversational access to business data for customers, prospects, and employees; increased daily new-customer engagement by approximately 5 times (5–6 to 25–30 users per day) and reduced manual effort for internal teams.

Frontend Web Developer Intern

Unified Mentor

Sep 2024 – Feb 2025

Haryana, India

- Built a full-stack e-commerce platform using Next.js and optimized frontend performance through code splitting, lazy loading, and API optimization, reducing page load time by 40%.

Projects

AI-Powered Docker Repo Auditor

Python, FastAPI, OpenAI, Trivy, AWS, SQS, DynamoDB, Redis, Docker, Terraform, Next.js

GitHub | Live Demo

- Built a production-oriented AI security auditing platform that analyzes container images using Trivy, Docker history, and Docker inspect before feeding deterministic scanner results into a six-agent LLM pipeline.
- Implemented a fan-out/fan-in multi-agent architecture with four concurrent analysis agents and dependent Dockerfile optimization and risk-scoring agents, with per-agent timeouts and failure isolation so individual agent failures degrade the report instead of terminating the scan.
- Built an evaluation gate using seeded security defects and clean-control expectations, enforcing a minimum 90% recall and zero false positives while measuring score stability across repeated model runs.
- Implemented grounded vulnerability analysis by ranking and reducing scanner findings before model inference and rejecting vulnerability IDs not present in scanner-provided input.
- Deployed the distributed architecture on AWS Fargate using Terraform, SQS FIFO, DynamoDB, Redis, and GitHub OIDC; implemented CI/CD gates for testing, evaluation, infrastructure validation, container builds, deployment, and rollback.

GuardRAG

FastAPI, LangChain, OpenAI Embeddings, Pinecone, NeMo Guardrails, Logfire, Docker

GitHub | Live Demo

- Built a production RAG system using FastAPI, LangChain, OpenAI embeddings, and Pinecone with layered NeMo Guardrails for prompt injection detection, blocking, redaction, and human approval workflows.
- Implemented Logfire observability and inference optimizations using batching and caching, with retrieval and response quality evaluation.

AZ-CLAW

GitHub

Bun, TypeScript, Vercel AI SDK, Groq, OpenRouter, pgvector

- Built a terminal-first AI coding agent supporting Agent, Plan, Ask, and Second-Brain workflows with CLI and Telegram interfaces.
- Implemented staged file mutations with diff-based approval, pgvector-based semantic codebase recall, and automatic Groq-to-OpenRouter model failover.

Education

Bachelor of Engineering in Artificial Intelligence and Machine Learning <i>Coorg Institute of Technology</i> CGPA: 8.1/10	Nov 2021 – Jul 2025 <i>Ponnampet, Karnataka</i>
Pre-University – Physics, Chemistry, Mathematics, Computer Science <i>Shri Sharada PU College</i> Percentage: 92.33%	2019 – 2021 <i>Koppal, Karnataka</i>

Achievements & Certifications

- Oracle Generative AI Professional Certification.
- Co-created ISATVON, an open-source AI skill framework.
- Presented a research paper on AI-automated Cleanroom under Prof. Babita P. K.
- Represented college at JITHack-24 and inter-college coding events; attended Google DevFest 2025, Bengaluru.