Binaya Tripathi

Full-Stack AI Engineer | Agentic AI & LLM Systems | Claude • GPT-5 • LangGraph • MCP • Next.js • Python
San Francisco, CA | +1 (786) 471-8264 | binayatripathi.dev@gmail.com | linkedin.com/in/btripathi08

SUMMARY

Full-Stack AI Engineer with **12+ years** of experience architecting production-grade **LLM (Large Language Model)**-powered web applications, **agentic AI** systems, and **RAG (Retrieval-Augmented Generation)** pipelines using **Claude (Anthropic)**, **OpenAI GPT-5 / o3**, **LangGraph**, **MCP (Model Context Protocol)**, and **vector search** with **pgvector** and **Weaviate**. Expert in shipping end-to-end **GenAI** features that combine **React / Next.js / TypeScript** frontends with **Python (FastAPI)** and **Node.js** backends, deployed on **AWS Bedrock**, **Azure OpenAI**, and **GCP Vertex AI** via **Docker**, **Kubernetes**, and **CI/CD** pipelines. Delivered measurable outcomes including a **60%** reduction in manual workflow effort and **50%** faster project delivery through **AI agent** automation, **prompt engineering**, **fine-tuning**, and production **LLM observability**.

EXPERIENCE

Builders Academy 04/2023 - Present
Senior Full-Stack AI Engineer Remote

- Directed **AI agent** workshops and technical onboarding for **50+ developers** globally using **Claude** and **Cursor**, resolving complex technical queries and documenting reusable knowledge bases for multi-team project acceleration.
- Architected and shipped **5+ agentic AI** tools using **Claude (Anthropic)**, **LangChain**, **LangGraph**, and **OpenAI GPT**, including multi-step **AI agent** workflow automation solutions that reduced manual team effort by **60%**.
- Developed **AI-powered full-stack** web applications using **React**, **Next.js**, **Node.js**, and **Python**, leveraging **Cursor** and **Claude Code** to integrate **LLM** modules for real-time collaboration, rapid prototyping, and intelligent dashboards.
- Built end-to-end **LLM-powered pipelines** and **agentic AI** systems connecting **Claude** and **GPT** reasoning layers to user-facing interfaces and backend APIs, enabling production-grade autonomous **AI agent** behaviors for startup clients.
- Automated debugging workflows, deployment pipelines, and task tracking systems using **AI agents** and **Cursor**-driven development, cutting project delivery times by **50%** across 10+ active client teams.
- Mentored developers on **agentic AI** best practices, **prompt engineering**, **Claude** and **LangChain** agent design, **MCP** integrations, and AI-assisted full-stack architecture, creating documentation that accelerated onboarding for new contributors.

Co-FounderGPT 12/2022 - 11/2023
Full-Stack AI Engineer San Francisco, US

- Architected an **LLM**-powered co-founder matching platform using **OpenAI GPT-4**, **GPT-3.5**, **Claude 2 (Anthropic)**, and **LangChain** orchestration, serving **10K+** founders with **streaming responses** over **Server-Sent Events (SSE)**.
- Built end-to-end **RAG (Retrieval-Augmented Generation)** pipelines with **pgvector** and **Weaviate**, **OpenAI embeddings** (text-embedding-ada-002), and **hybrid search** to index **100K+** startup documents, reducing manual research time by **70%**.
- Shipped a full-stack **Next.js** + **FastAPI** + **PostgreSQL** web application in **TypeScript** and **Python**, with **TailwindCSS**, **shadcn/ui**, **Zustand**, and **React Query** for real-time founder dashboards and chat UIs.
- Deployed production **LLM** services on **AWS** (**Lambda**, **ECS**, **S3**) using **Docker** and **GitHub Actions CI/CD**, with **LangSmith LLM observability**, **prompt engineering** iterations, and offline **evals** that cut model cost by **45%**.
- Integrated **OpenAI** and **Anthropic API tool use** / **function calling** flows over **REST** / **OpenAPI** services for document parsing, web research, and CRM sync, automating **80%** of founder onboarding.

NFT Studios 08/2022 - 05/2023
AI Full Stack Engineer San Francisco, US

- Engineered a production **GenAI (Generative AI)** NFT discovery and valuation platform using **OpenAI GPT-3.5** / **GPT-4**, **LangChain**, and **Weaviate vector search** across **1M+** token metadata with **OpenAI embeddings** and price prediction.
- Built full-stack **React** + **Next.js** + **Node.js** (Express, **TypeScript**) marketplace with **GraphQL** and **REST APIs**, **PostgreSQL**, **Redis**, **WebSockets** for live bids, and **AWS** (**S3**, **Lambda**, **ECS**) serverless architecture.
- Orchestrated early **LLM** pipelines with **LangChain** and **LlamaIndex** to ingest on-chain events, call third-party APIs, and generate investor reports, blending **OpenAI GPT** with **Claude 1 (Anthropic)** for reasoning.

- Automated CI/CD with **GitHub Actions**, **Docker**, **Kubernetes** (EKS), and **Terraform**, launching **15+** features across 3 environments with **Playwright** E2E tests and **Sentry** monitoring — achieving zero production incidents over 9 months.
- Collaborated with product to rapid-prototype **LLM**-powered pricing models, applying **prompt engineering**, **few-shot** and **chain-of-thought** techniques with **Hugging Face Transformers** and **PyTorch**, raising valuation accuracy by **32%**.

Bitfari 03/2022 - 10/2022
Senior Full Stack Engineer San Francisco, US

- Developed a full-stack smart-city campaign platform in **React**, **TypeScript**, **Node.js**, and **Python (FastAPI)** with **PostgreSQL** and **Redis**, delivering real-time content to **500+** displays across 12 US markets.
- Built an early AI creative-assistant module integrating the **OpenAI GPT-3 API** (davinci) for ad-copy generation, **Hugging Face Transformers** for content classification, and **scikit-learn** predictive models, cutting creative turnaround from 3 days to 4 hours.
- Containerized microservices with **Docker** and **Kubernetes** on **AWS** (**ECS**, **Lambda**, **S3**, **CloudWatch**), managed infrastructure with **Terraform**, and shipped via **GitHub Actions CI/CD**, moving deployment cadence from weekly to daily.
- Designed **REST** and **OpenAPI**-documented backend services across **PostgreSQL** and **MongoDB**, adding **Datadog** and **Sentry** observability that reduced mean-time-to-detect incidents by **55%**.

WANAMAKER 03/2013 - 08/2021
Full Stack Engineer San Francisco, US

- Built and maintained a customer-facing e-commerce platform in **React**, **Redux**, **TypeScript**, **Node.js** (**Express**), and **PostgreSQL**, scaling to **2M+** monthly sessions and **200K+** SKUs across B2B and B2C channels.
- Designed **RESTful APIs**, **GraphQL** endpoints, and microservices on **AWS** (**EC2**, **S3**, **RDS**, **Lambda**) with **Docker**, **Jenkins**, and **GitHub Actions CI/CD**, plus **Redis** caching and NGINX load balancing — cutting checkout latency from **3s to 1s**.
- Built classical **ML** and **NLP** data pipelines in **Python** using **scikit-learn**, **PyTorch**, **spaCy**, and **Pandas** for product recommendations, search ranking, and demand forecasting across **200K+** SKUs.
- Led code reviews, pair programming, and **TDD** practices with **Jest**, **Playwright**, and **Pytest**, mentoring **6+** junior engineers and establishing coding standards that reduced production defects by **40%**.

TECHNICAL SKILLS

AI / LLM & Agents: Claude (Anthropic — Opus 4, Sonnet 4.6, Haiku 4.5), OpenAI GPT-5 / o3 / o4-mini / GPT-4o, Gemini 2.5, Llama 4, Hugging Face Transformers, Claude Agent SDK, OpenAI Agents SDK, LangChain, LangGraph, LlamaIndex, Vercel AI SDK, AI Agents, Agentic AI, Multi-Agent Systems, MCP (Model Context Protocol), Tool Use / Function Calling, Structured Outputs, RAG (Retrieval-Augmented Generation), Prompt Engineering, Prompt Caching, Fine-Tuning (LoRA, QLoRA, DPO), LLM Observability (LangSmith, LangFuse), Evals, Guardrails, Streaming (SSE), GenAI (Generative AI), Cursor, Claude Code

Vector Databases & Search: pgvector, Weaviate, Chroma, Qdrant, FAISS, Semantic Search, Hybrid Search, Embeddings, Re-ranking, Contextual Retrieval

ML / Data Science: PyTorch, Hugging Face Transformers, TensorFlow, scikit-learn, Pandas, NumPy, MLflow, Weights & Biases, MLOps, NLP, spaCy, Predictive Analytics

Languages: Python, TypeScript, JavaScript (ES6+), SQL, Go, Rust, Java, PHP

Frontend: React, Next.js (App Router, Server Components, Server Actions), TypeScript, TailwindCSS, shadcn/ui, Radix UI, Zustand, Redux, React Query / TanStack Query, SWR, React Native, Streaming UI

Backend & APIs: Python (FastAPI, Flask, Django), Node.js (Express, NestJS), REST APIs, OpenAPI, GraphQL, tRPC, gRPC, WebSockets, Server-Sent Events (SSE), Microservices, Serverless, Event-Driven Architecture

Databases: PostgreSQL, MySQL, MongoDB, Redis, DynamoDB, Supabase, Firebase, Prisma, Drizzle, SQLAlchemy

Cloud & AI Platforms: AWS (Bedrock, SageMaker, Lambda, EC2, S3, ECS, EKS, RDS, API Gateway, Step Functions, CloudWatch), Azure (OpenAI Service, Functions, AKS), GCP (Vertex AI, Cloud Run, GKE, BigQuery), Vercel AI SDK, Cloudflare Workers

DevOps & CI/CD: Docker, Kubernetes, Terraform, Helm, GitHub Actions, GitLab CI, Jenkins, CircleCI, CI/CD Pipelines, Nginx, Linux

Testing & Monitoring: Jest, Vitest, Playwright, Cypress, Pytest, React Testing Library, Datadog, Sentry, Prometheus, Grafana, OpenTelemetry, LangSmith, LangFuse

Tools: Cursor, Claude Code, GitHub Copilot, Git, GitHub, GitLab, Jira, Linear, Notion, Figma, VS Code, Postman

EDUCATION

Master's Degree, Data Science 2021 - 2022
Eastern University

Bachelor of Science, Physics 2009 - 2013
Winona State University

CERTIFICATIONS

- JavaScript for Web Designers
- Frontend Web Development Certificate