

SHREYAS JAGANNATH

Senior AI Engineer | Forward Deployed / Applied AI | Agentic Systems, Full-Stack & Cloud

+44 7438 193 743 · London, UK · shreyasjag@hotmail.com · shreyasjagannath.com

linkedin.com/in/shreyasjagannath · github.com/jaggernaut007

PROFILE

Senior AI engineer with 8+ years across founding-team and product engineering roles, who builds and ships agentic systems end-to-end — and makes them hold up in production, not just demo. Across multiple founding-team roles I've done exactly what Forward Deployed Engineers are hired for: walked into ambiguous, half-built environments, scoped the real problem with incomplete information, and delivered production LLM platforms, multi-agent systems, and full-stack applications from zero. Cut LLM token costs by 10x, improved agent efficiency by 3x, and reduced latency by 55%, with reliability and auditability built in rather than bolted on. A deep learning and computer-vision background (TensorFlow, PyTorch, OpenCV) and an MSc in AI mean I reason at the model level as well as the systems level. Looking to bring that depth to a senior AI engineering, AI platform, or build-heavy applied AI role with real technical scope and a longer runway.

CORE SKILLS

Languages & Frameworks: Python, FastAPI, React.js, Node.js, TypeScript, GraphQL, Flutter, REST APIs

Cloud & Infrastructure: GCP (Vertex AI, Cloud Run, Pub/Sub, BigQuery), AWS (Lambda, ECS, Bedrock, S3, Kinesis, DynamoDB), Docker, Terraform, CI/CD

AI & LLM Systems: Machine Learning, Multi-agent orchestration, MCP, RAG, Claude Code, LangGraph, LangChain, LlamaIndex, Google ADK, Evaluations, DeepEval, Observability, Knowledge Graphs

Data & Storage: PostgreSQL, Neo4j, VectorDBs (LanceDB, ChromaDB, Pinecone)

Reliability & Governance: Evaluation-driven development (DeepEval), fail-closed CI gates, audit trails & compliance registers, human-in-the-loop safety checks, observability

EXPERIENCE

Lead Software Engineer | Move By Vision (Pro bono) | Dec 2025 – May 2026

- Stabilised a failing Flutter + GCP platform by building a guard-railing and validation layer across Vertex AI integrations
- Cut generation latency by 55% (15s to 7s) through optimised execution flows

Lead AI Engineer | Klyft Technologies (Founding Team, Pre-seed) | Aug 2025 – Nov 2025

- Improved agent efficiency by 3x (60s to 20s) on the core workout calibration feature, which drove early user adoption
- Cut LLM token usage by 10x through prompt optimisation and caching, balancing cost with performance
- Built a multi-agent health and wellbeing coaching platform (workouts, nutrition, mental wellbeing) from zero to production on GCP (Vertex AI, Google ADK, Gemini, BigQuery, Cloud Run, Terraform)

Lead AI Engineer | Scan and Sew Inc. (Founding Team) | Oct 2024 – Aug 2025

- Led AI architecture across two production platforms, hands-on from system design through to deployment
- Built a Designer AI Assistant using RAG and vision-language models, letting design teams query their internal libraries in real time
- Developed a real-time manufacturing lifecycle platform (React, AWS Lambda, GraphQL, DynamoDB), making pragmatic build-vs-buy calls (e.g. integrating external 3DGS APIs) to accelerate delivery without over-engineering

Lead Software Engineer / CTO | AIMAGE Technologies (Seed-Stage, Founder) | Jan 2020 – Aug 2023

- Founded and led a cross-functional team of 10 across engineering, product, and design
- Hired and mentored the engineering team, setting code-review and architecture standards and owning the technical roadmap
- Architected and built an NLP-driven visual search and recommendation platform on TensorFlow, backed by a vector-database RAG knowledge graph
- Built and deployed a virtual try-on engine using OpenCV, PyTorch, and Unity3D
- Delivered both systems to production on AWS (Lambda, ECS, Neptune, DynamoDB, S3, Kinesis) with GraphQL APIs and full infrastructure ownership, embedded on-site on the retail client's shop floor to improve conversion and

reduce bounce rates through visual search

EARLIER CAREER

Software Consultant (Part-time) · Mage Ventures Pvt. Ltd. · Oct 2019 – Mar 2021 — Delivered a payment gateway, CV-based retail checkout, and data automation platform.

AI Consultant (Founding Team) · Cellstrat Inc. · Oct 2018 – Jan 2020 — Built AI/ML POCs for enterprise clients (Airbus, Volvo, Target) and delivered workshops for 100+ professionals.

Full Stack Engineer (Founding Team) · EndGate Global Pvt. Ltd. · Jul 2017 – Jul 2018

AI Research Intern · ISRO – Indian Space Research Organisation · Nov 2016 – Apr 2017

Software Developer Intern · WebMobi Technologies Pvt. Ltd. · Oct 2014 – May 2015

SELECTED PROJECTS — AGENTIC SYSTEMS, RELIABILITY & GOVERNANCE

Quality Gate CI: AI-Powered Test Generation & Compliance Gate

A fail-closed compliance gate for regulated pharma: AI-generated tests plus audit-ready evidence (verdict, RCA, pass/fail counts) surfaced on every change before a human approves the merge. Prototyped end-to-end in under 5 hours.

- Two-job GitHub Actions pipeline: Claude Code Actions generates Gherkin and pytest-bdd/Playwright tests from acceptance criteria and code diffs, then a fail-closed gate enforces the deploy decision and surfaces audit evidence (verdict, pass/fail counts, RCA) directly on the PR before merge.
- Domain-Driven Design throughout to keep tests tied to business outcomes; append-only compliance register as the system of record, with a Streamlit dashboard for KPI tracking and the pipeline hardened against shell injection and race conditions.

Nexus-MCP: Hybrid Code Intelligence Server github.com/jaggernaut007/nexus-mcp

A published, local-first developer tool that cuts AI agent token costs by up to 90%, with no cloud, no API keys required.

- Published on PyPI (nexus-mcp-ci) and listed on the Glama MCP marketplace. Combines vector search, BM25, and structural code graph analysis to deliver 70-90% token reduction on code lookups; 15 tools in one local server, tree-sitter parsing across 25+ languages, under 350MB RAM, no API keys.

Health Intelligence Engine: Agentic Wellness Intelligence Platform github.com/jaggernaut007/healthAgent

Clinical AI platform that turns supplement queries into evidence-grounded recommendations, with a built-in Pharmacovigilance Critic acting as a safety guardrail on every output.

- 9-node LangGraph pipeline that grounds product data against NIH DSLD and PubMed in a Neo4j knowledge graph, running a Pharmacovigilance Critic that checks every recommendation against the user's allergies and medications, with DeepEval quality gates and a published healf CLI.

EDUCATION

MSc Artificial Intelligence (Distinction) | University of Surrey | 2023 – 2024

Dissertation: Human avatar reconstruction using Gaussian Splatting · Winner, UKSEDS IOSM 2024 (Most Innovative Business Award)

Master of Computer Applications (First Class) | Christ University | 2013 – 2017

Co-authored research on AI & RL-based path planning · IoT-based autonomous rover capstone

CERTIFICATIONS

Claude Code in Action, Model Context Protocol, Agent Skills, Building with the Claude API (Anthropic); LangGraph & LangSmith (LangChain Academy); Deep Learning & Computer Vision (Stanford)