SHREYAS JAGANNATH

Agentic AI Engineer | Multi-Agent Systems | MCP Infrastructure | LLMs | RAG | Computer Vision
+44 7438 193 743 | London, UK | shreyasjag@hotmail.com | shreyasjagannath.com |
linkedin.com/in/shreyasjagannath | github.com/jaggernaut007

PROFILE

Agentic AI Engineer with 8 years experience across the full digital transformation stack: from computer vision and AI consulting for global enterprises through production LLM systems to now designing and shipping multi-agent platforms, MCP infrastructure, and RAG pipelines at the frontier. Creator of Nexus-MCP, live on PyPI with 15 tools inbuilt; track record includes 10x LLM token reduction, 3x agent efficiency gain, and 100% safe AI outputs in production. I work at the intersection of deep technical build and stakeholder engagement: I diagnose the real problem, architect the solution, and ship it, having done this across founding teams, enterprise clients, and on-site client deployments. Seeking roles where agentic AI is the core challenge and where engineering judgment and delivery ownership matter.

WHAT SETS ME APART

Zero-to-Production Agentic AI Systems | MCP Server Development | Agent Guardrailing & Evaluation | LLM Cost & Token Optimisation | On-Site Client Deployment | Full-Stack AI & Infrastructure Ownership | Rapid Prototyping

CORE SKILLS

Agentic AI & LLMs: Multi-agent orchestration, MCP, RAG, LangGraph, LlamaIndex, CrewAI, Google ADK, LLM guardrailing, agent evaluation, RL-informed iteration, observability, hybrid search, knowledge graphs, prompt engineering

Computer Vision & 3D: Gaussian Splatting, Vision-Language Models, PyTorch3D, SLAM, OpenCV, ONNX Runtime, Mediapipe, Unity3D

Engineering & Cloud: Python, TypeScript, FastAPI, React, Next.js, Node.js, Docker, Terraform, GCP (Vertex AI, Cloud Run, Pub/Sub, BigQuery), AWS (Lambda, ECS, Bedrock, S3, Kinesis, DynamoDB)

Databases & Vector Stores: PostgreSQL, Neo4j, LanceDB, ChromaDB, Supabase, NeonDB, Redis, Neptune

EXPERIENCE

Lead Software Engineer (Acting CTO) Dec 2025 - Present

Move By Vision (Pro bono engagement taken during career transition)

- Inherited a Flutter + GCP platform in critical-failure alpha and brought it to a releasable state. Engineered \ a guardrailing and validation layer across all Vertex AI integrations (Pydantic schema validation, retry logic, safe fallback mechanisms, parallelised cloud functions), eliminating 100% of AI-induced crash failures and reducing generation latency by 55% (15s to 7s).
- Built an AI observability pipeline to continuously surface output quality issues across the full Vertex AI stack, and conducted a full GDPR / EU AI Act / App Store compliance and security audit producing a prioritised remediation roadmap of 19 critical vulnerabilities.
- Beyond the technical build, advised the founders on go-to-market strategy and phased release planning, aligning the technical roadmap with commercial milestones to ensure production readiness at each stage of growth

Lead AI Engineer Aug 2025 - Nov 2025

Klyft Technologies (Founding Team) | London

- Architected and delivered the entire AI platform from zero to production: multi-agent system design, LLM inferencing endpoints, session management, memory systems, RAG pipelines, evaluation and observability frameworks, and all MLOps practices. The full platform was built on Vertex AI and Cloud Run, delivered entirely as Infrastructure as Code using Terraform, with CI/CD via Google Cloud Build and GitHub Actions.
- Designed and shipped the workout calibration system, the most-used feature with universal adoption, delivering a 3x improvement in agent efficiency. Built RAG-based guardrailing with safety checks at every agent decision point, achieving 100% safe outputs across all workouts, eliminating injury risk and allergy exposure for users.
- Maintained 100% uptime and completely scalable across all GCP deployments throughout the product lifecycle, and achieved a 10x reduction in LLM token usage through agent-ops optimisation, prompt engineering, and intelligent caching.

Lead AI Engineer / CAIO Oct 2024 - Aug 2025

Scan and Sew Inc. (Founding Team)

- As CAIO and hands-on engineer, owned the full technical and AI direction of the company. Architected and built two production platforms in collaboration with the design team: a Designer AI Assistant using RAG over design

documents and Vision-Language Models (LLama AI and Nvidia Nemo), enabling designers to query and reason over their own design libraries in real time. Also delivered a real-time manufacturing lifecycle platform (ReactJS, AWS Lambda, GraphQL, DynamoDB, Amplify) with full AWS infrastructure set up and owned throughout.
- Both platforms shipped to production and saw active use: the Designer AI Assistant adopted by real designers in their daily workflow, and the manufacturing platform deployed end to end across the product lifecycle, generating significant stakeholder interests.
- Led Gaussian Splatting and PyTorch3D research to evaluate spatially aware CAD 2D-3D approaches, ultimately integrating a 3DGS API rather than building from scratch, a pragmatic engineering call that accelerated delivery without sacrificing capability.

Lead AI Engineer / CTO Jan 2020 - Aug 2023

AIMAGE Technologies (Seed-Stage, Team of 10)

- Built and led a cross-functional team of 10 across engineering, product, and design. Designed and built alongside the team: an NLP-driven visual search and recommendation platform combining TensorFlow/PyTorch with a vector-database-backed visual RAG knowledge graph, deployed on AWS (Lambda, ECS, Neptune, DynamoDB, S3, Kinesis) with GraphQL APIs and an AI/AR virtual try-on engine (SLAM, 3D geometry, OpenCV, PyTorch3D, Mediapipe, Unity3D)
- Deployed on-site with a key retail client across multiple engagements, embedding directly on the floor to understand their product discovery and try-on challenges firsthand. Iterated both platforms from real customer feedback, built and demonstrated working prototypes in client sessions, and achieved commercial validation.
- Both platforms shipped to production and saw active use: the try-on engine deployed with a retail client delivering measurable improvements in conversion rates and bounce rate reduction, and the search platform live with end users. Owned all AWS infrastructure and deployments throughout.

EARLIER CAREER

Software Consultant Oct 2019 - Mar 2021

Mage Ventures Pvt. Ltd.

- Delivered three production systems: a secure cross-platform payment gateway for Walmart-owned PhonePe (Node.js, React, AWS) that went live in production; an AI-powered grocery checkout engine using computer vision and TensorFlow deployed in store; and an automated data collection system for Varthana Finance that measurably improved operations across 1,000+ schools.

AI Consultant Oct 2018 - Oct 2021

Cellstrat Inc.

- Delivered AI/ML solutions and POCs for enterprise clients including TVS Motors, Target, Airbus, and Volvo in PyTorch, TensorFlow, Docker, and AWS. Partnered with NASSCOM (Government of India AI initiative) to evangelise enterprise AI adoption, delivering workshops, research talks, and POCs for 100+ professionals across India's leading organisations.

Full Stack Engineer (Founding Team) Jul 2017 - Jul 2018

EndGate Global Pvt. Ltd.

- Full-stack food delivery platform (Java, Spring, Hibernate, JavaScript, MySQL), owning order-processing and bulk-order automation for high-volume clients.

AI Research Intern Nov 2016 - Apr 2017

ISRO, Indian Space Research Organisation

- Mono Vision Depth Detection research for the Chandrayaan Moon Rover Mission, enhancing spatial navigation in lunar environments using computer vision.

NOTABLE PROJECTS & OPEN SOURCE

Nexus-MCP: Hybrid Code Intelligence Server *Live on PyPI*

An MCP server giving AI agents precise, token-efficient code understanding, with zero cloud dependencies, zero API keys. Hybrid search via Reciprocal Rank Fusion (vector + BM25 + code graph), FlashRank re-ranking, token-budgeted output. Structural analysis via rustworkx. Dual parsing: tree-sitter + ast-grep across 25+ languages. Persistent semantic memory with TTL expiration and 6 memory types.

Efficiency: 15,000-40,000 tokens saved per session (30-60% reduction vs standalone agentic file browsing). 15 tools, 461 tests, 14 ADRs, fully local, less than 350 MB RAM.

AI Mood Board: Multi-Agent Dynamic App Generator *Live | GCP Cloud Run*

A custom-built multi-agent platform on React and Vercel AI SDK where agents collaborate via a custom messaging service to generate fully functional mini-applications on demand. The pipeline covers intent understanding, UI/UX design, code writing, code review, and automated testing before rendering the component live on a canvas. Built on the latest OpenAI APIs, React, and Vercel, deployed on GCP Cloud Run. Users can iterate, improve, save, and modify components after generation.

JobScout: Agentic Job Discovery Platform & MCP Server *Live on GCP Cloud Run*
Agentic job discovery platform aggregating 1500+ companies across 10 ATS platforms and 7 web sources. LangGraph evaluation pipeline covering embedding generation, vector similarity, deduplication, LLM scoring. Dedicated MCP server for agent-compatible tool use.

Agentic Coding Playbook
Operational guide for IDE coding agents (Claude Code, Cursor, Copilot, Gemini CLI, Codex CLI). Covers AGENTS.md/CLAUDE.md architecture, MCP server integration, hook-based verification, spec-driven and eval-driven development, SAST/DAST, prompt regression CI, LLM judges, and golden dataset evaluation pipelines.
78 open-source repositories: github.com/jaggernaut007

EDUCATION

MSc Artificial Intelligence - Distinction 2023 - 2024
University of Surrey
Dissertation: human avatar reconstruction via Gaussian Splatting for edge computing.
Co-delivered a local LLM token-classification project that doubled render speed.
Winner, UKSEDS IOSM Competition 2024 (Peryton Space Society). Winner, Most Innovative Business Award, University of Surrey.

Master of Computer Applications - First Class 2013 - 2017
Christ University
Co-authored peer-reviewed research paper on AI and RL-based path-planning methods.
Built an IoT-based autonomous rover as a capstone engineering project.

Bachelor of Computer Applications - First Class 2010 - 2013
Bangalore University
Built a Healthcare Management System to streamline patient care workflows.
Led WEF Global Shapers team engaging C-level industry leaders.

Entrepreneurship Studies
IFA Paris
Winner, Most Innovative Company Award, presented by IBM Paris and IFA Paris.

CERTIFICATIONS

LangGraph & LangSmith (LangChain Academy) | Multi AI Agent Systems with CrewAI (Coursera) | Agentic RAG with LlamaIndex (Coursera) | Knowledge Graphs for RAG (Coursera) | Deep Learning & Computer Vision (Stanford Online) | Entrepreneurship: Idea to Launch (UC Berkeley) | Technology Entrepreneurship (Harvard Online)

LANGUAGES

English (fluent) | Hindi | Kannada | Telugu | French