

# Sumith Sadanand Bhandari

AI Software Engineer | Production AI Agents | RAG | MCP | Python | TypeScript | AWS

London, UK | [sumithsbhandari@gmail.com](mailto:sumithsbhandari@gmail.com) | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

## Professional Summary

AI software engineer with 4+ years building and running production backend systems in Python and Node.js/TypeScript, now focused on AI agents that do real work for real users. At Goodspeed Studio (Jul 2026 to present, contract then freelance) I build and run a production Slack AI agent used daily by 16+ people, with tool calling, prompt caching, human-review steps and multi-tenant persistent memory, where I traced and fixed a cross-client confidentiality boundary. I also shipped a website AI chat agent through a GDPR, PECR and EU AI Act review, and I lead production engineering for client platforms. At TechMachinery I led a FastAPI backend for a multi-agent knowledge platform over PostgreSQL and pgvector, and built a document extraction pipeline that turns unstructured engineering specs into structured JSON at 94% accuracy. I measure what I build, from that extraction pipeline to a hybrid RAG system at MRR@6 0.91. In the open I build agent tooling: Claude Anchor, a cross-platform desktop app for Claude Code, and Brainvault (PyPI), which gives coding agents persistent memory through 12 MCP tools.

## Skills

|                          |                                                                                                                                                                                                                                                         |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>AI &amp; Agents</b>   | Tool-Calling Agents, Multi-Step Workflows, Multi-Agent Orchestration, Strands (AWS multi-agent), MCP Protocol (12-tool server), Agent Memory and Context, Prompt Caching, Structured Outputs, Document Extraction, Anthropic API, OpenAI API, LangChain |
| <b>RAG &amp; Evals</b>   | RAG Pipelines, Hybrid Dense + BM25 Retrieval, Embeddings, pgvector, FTS5, Retrieval Evaluation (MRR, Precision@k), Test-Set Evals                                                                                                                       |
| <b>Languages</b>         | Python, TypeScript, SQL, Node.js, Swift, Rust, Dart (Flutter)                                                                                                                                                                                           |
| <b>Backend</b>           | FastAPI, Flask, Express.js, REST APIs, Webhooks, Async APIs, Job Queues, Microservices, Event-Driven Architecture, WebSockets                                                                                                                           |
| <b>Databases</b>         | PostgreSQL, pgvector, Vector Databases, MongoDB, MySQL, Redis, SQLite, Supabase                                                                                                                                                                         |
| <b>Cloud &amp; Tools</b> | AWS (EC2, S3, Lambda, IAM, CloudWatch, VPC), GCP (GCE, Cloud Storage), Docker, NGINX, CI/CD, GitHub Actions, Railway, Vercel, Pytest, Playwright, Claude Code, Cursor, Codex                                                                            |

## Experience

**AI Engineer** Jul 2026 - Present

*Goodspeed Studio Ltd, London, UK (Contract, then Freelance, Hybrid)*

**Production AI Agents & Backend Integrations** *Python, TypeScript, Anthropic API, OpenAI API, Supabase, PostgreSQL, Railway, Vercel, Slack/Bolt*

- Built and hardened a production Slack AI agent used daily by **16+ team members**. Shipped backend integrations across web search, model upgrades, prompt caching, file reading, channel posting, human-review workflows, meeting-notes sync, a project-management tool, and end-of-day drafting; remediated a critical cross-client confidentiality boundary in the multi-tenant persistent-memory architecture.
- Built an internal Python lead-finding agent (Claude, job-board and contact-data APIs, Google Sheets, Slack webhooks) that identifies and prioritises commercial leads via REST and webhook integrations, processing about **20 qualified opportunities per day** after post-launch deduplication improvements.
- Led development and production launch of a website AI chat agent. Conducted a full compliance review covering **GDPR, PECR, and the EU AI Act**, including data-retention and data-handling decisions across Anthropic, OpenAI, Supabase, Railway, and the CRM.
- Led production engineering for external client platforms (identities confidential): Supabase Row Level Security audits, production-readiness reviews, PR review and release management, rollback drills, and client-channel communication for service reliability.

**Software Engineer** Jun 2025 - Sep 2025

*Visiminds Technologies, Remote, UK (Contract)*

**VAPT Automation Platform** *Python, Flask, MongoDB, Security Automation, OWASP Top 10*

- Built a VAPT platform that runs 14 security tools (Nmap, Nikto, Nuclei, TestSSL) and produces structured client reports, replacing a manual analyst workflow for enterprise clients.
- Designed a five-stage pipeline (Scan, Parse, Aggregate, Correlate, Report) that normalises output across tools into a single vulnerability schema with cross-tool deduplication.
- Set up a MongoDB-backed job queue with 20 concurrent workers; cut total scan and report time by **95%** (from roughly 4 hours to under 15 minutes). Platform in active production use for weekly and monthly VAPT cycles across **10+ enterprise clients**.

**Software Engineer** Oct 2023 - Dec 2024

*TechMachinery Labs, Bengaluru, India (Contract, Full-time)*

**Agentic AI Knowledge Platform for Heat Exchanger Design** *Python, FastAPI, Strands (AWS), RAG, PostgreSQL, pgvector*

- Led backend development of a FastAPI platform that lets engineers query and design heat exchanger systems using natural language; queries resolve in **under 5 seconds**.

- Designed a multi-agent system using Strands (AWS multi-agent orchestration) that combines pgvector retrieval, a structured materials database, and parsed equipment data across 3+ agent types for complex cross-domain queries as model-service integrations behind the API.
- Built a document parsing pipeline that converts unstructured engineering specs into structured JSON with 50+ fields, achieving **94% extraction accuracy** on a 50-document test set.

### Software Development Engineer

Jul 2022 - Sep 2023

*TechMachinery Labs, Bengaluru, India*

*Node.js, Python, Express.js, PostgreSQL, AWS S3, Microservices, RBAC*

- Built 10+ RESTful APIs across a microservices codebase; implemented data models, RBAC across 5 user roles, and auth middleware with unit test coverage.
- Integrated a payment gateway with full audit trails and moved media storage to AWS S3, replacing all on-premise infrastructure.
- Owned the full backend lifecycle on client projects: schema design, API development, testing, and AWS deployment with CI/CD discipline.

### Software Development Engineer - SDE 1

Nov 2021 - Jul 2022

*TechMachinery Labs, Bengaluru, India (Part-time, during final year BE)*

*Flutter, Python, Node.js, PostgreSQL, Express.js, REST APIs*

- Built Flutter mobile apps and REST backends in Python and Node.js; maintained PostgreSQL and MongoDB data layers across several client projects.

### Co-Founder & Engineering Lead

Nov 2020 - May 2022

*Drogher Technologies Pvt Ltd, Bengaluru, India | Quick-Commerce Platform*

*Node.js, PostgreSQL, Flutter, AWS, REST APIs, RBAC*

- Co-founded and engineered CornerCart, a quick-commerce platform (Node.js, PostgreSQL, Flutter, AWS, GitHub Actions CI/CD) connecting local stores to customers with real-time order management; led a team of 4, made the technical call to pivot from B2C to B2B consulting, and delivered full-stack products across healthcare, food delivery, and co-working verticals.

## Projects

### Claude Anchor - Multi-Account Desktop App for Claude Code

[github.com/SumithSB/claude-anchor](https://github.com/SumithSB/claude-anchor)

*Swift, Rust, Tauri, Shell Scripting, Anthropic OAuth API, Open Source (MIT)*

- Built and published an open-source desktop app for 3 platforms (Swift menu-bar app on macOS, Rust/Tauri tray app on Windows and Linux) that routes Claude Code to the right Claude account per project folder, with no telemetry.
- Implemented live tracking of 2 rate-limit windows (5-hour and weekly) per account via Anthropic's OAuth usage endpoint, with automatic failover to an account with headroom so sessions don't stall at the limit.
- Integrated with 5 shells and editors (zsh, bash, PowerShell, VS Code, Cursor) plus an anchor CLI; conversations move between accounts with full history, and installs ship via Homebrew and checksum-verified scripts.

### Brainvault - Local-First Memory Layer for AI Coding Agents

[github.com/SumithSB/brainvault](https://github.com/SumithSB/brainvault) | PyPI

*Python, SQLite, MCP Protocol, FTS5, Semantic Search, fastembed*

- Designed, built, and published an open-source MCP memory layer to PyPI; gives AI coding agents (Claude Code, Cursor) persistent local memory across sessions with no cloud dependency.
- Implemented 12 MCP tools (get-context, search-memory, save-memory, reflect, record-outcome) with FTS5 full-text search and optional semantic search via fastembed and sqlite-vec.
- Ships with a CLI of 20+ commands (bootstrap, bootstrap-git, doctor, embed, graph) to seed the vault from existing transcripts and git history.

### Fashion AI - Intelligent Personal Styling Assistant

MSc Dissertation, May 2025

*Flutter, FastAPI, PostgreSQL, pgvector, GCP (GCE), MediaPipe, OpenCV, LLM APIs, RAG, LAB Colour Science*

- Built an end-to-end mobile styling app: skin tone analysis pipeline (MediaPipe, LAB/ITA, CIEDE2000 Monk Scale mapping) validated against the MST-E dataset (1,542 images, 19 subjects); within-one-step accuracy of 34% on a 10-class classification problem.
- Recommendation engine ranks a 59,447-product H&M catalogue using a four-factor weighted scoring formula (colour compatibility 40%, style 30%, body type 20%, occasion 10%) with MMR diversity reranking.
- RAG conversational styling interface (hybrid dense + BM25 over 86-source knowledge base) achieved MRR@6 of 0.91 and P@1 of 0.89 across 53 evaluation queries.

## Education

### University of Leicester

Leicester, UK

*MSc Advanced Computer Science - Distinction (76.67%)*

Jan 2025 - Jul 2026

Dissertation: Fashion AI - Best Technical Project 2026 (Flutter, FastAPI, pgvector, Computer Vision, RAG)

Key Modules: Internet and Cloud Computing, Big Data and Predictive Analytics, Mobile and Web Applications

### Ramaiah Institute of Technology

Bengaluru, India

*BE Information Science and Engineering*

Aug 2018 - Jul 2022