

Suyash Kumar Dubey

Applied / Forward-Deployed AI Engineer · Agents, RAG, Production LLM Systems

dsuyash57@gmail.com | +91 91180 23581 | linkedin.com/in/suyash-kumar-dubey-410533211 | github.com/suyashdube
Remote, India (IST) · Overlaps US Eastern mornings and full AU/APAC day · Available full-time

40%	20%	60%	3 yrs
less clinician documentation time after agent rollout	accuracy lift on real-time EEG classification	manual steps removed from a content ops pipeline	shipping LLM systems in regulated, data-sensitive domains

PROFILE

Engineer who ships LLM products end to end and sits with the customer while doing it. Three years building agentic systems in healthcare and public digital infrastructure, where a wrong output is a clinical risk rather than a bad demo. Comfortable in the forward-deployed pattern: scope with the stakeholder in plain language, build the agent, instrument it, then iterate against real usage. Production stack is LangGraph, LangChain, FastAPI, AWS Bedrock and Langfuse, with Kubernetes and Kafka underneath when the load demands it.

EXPERIENCE

AI/ML Consultant · SpinX AI (independent practice) Mar 2026 – Present

- Deliver LLM products for clients across the US, UK, Europe and Australia, owning the arc from discovery call to deployed system.
- Translate governance requirements (AI management-system standards, national AI ethics and cybersecurity frameworks) into concrete technical controls, then ship against them.
- Build RAG and agentic systems on LangGraph and LangChain with evaluation harnesses and monitoring wired in from the first commit, not bolted on before launch.

LangGraph · LangChain · FastAPI · EKS · Kafka · Redis · Langfuse

AI Research Engineer · Brainwave Science (contract) Sep 2025 – Present

- Raised real-time EEG signal detection accuracy 20% with a hybrid CNN/RNN architecture, benchmarked against published research and validated for reliability, not just headline accuracy.
- Designed a modular inference pipeline with hot-swappable model components so experiments run without disrupting production streams.
- Read current papers and turned them into live model iterations, then explained the results to non-technical stakeholders without the jargon.

PyTorch · TensorFlow · Keras · Python

SDE, Python & Generative AI · Asha Health (AI medical scribe) Mar 2023 – May 2025

- Built and ran a production medical scribe that extracts clinical entities from physician-patient conversations, cutting manual transcription time roughly 40%.
- Designed LangGraph stateful agent workflows with conditional branching across documentation, coding and review stages, so the agent handles the messy middle instead of failing over to a human every time.
- Shipped a RAG pipeline on AWS Bedrock that grounds every generated note in patient history and clinical guidelines, which is what made the output usable on protected health data.
- Owned prompt governance and observability through Langfuse: versioned prompts, traced every model call, watched token spend, latency and quality KPIs to catch regressions before clinicians did.
- Gathered requirements directly from clinical stakeholders and converted them into agent behaviour specs, no engineering jargon in the room.

Python · LangChain · LangGraph · AWS Bedrock/Lambda/S3/DynamoDB · Langfuse · TypeScript

SDE II, Backend & Generative AI · EMB Global (Rayo AI platform) May 2025 – Jul 2025

- Built the agentic content pipeline behind an AI SEO platform: multi-step agents for research, drafting and optimisation, with brand-voice and compliance checks enforced at generation time.

- Cut manual intervention roughly 60% by orchestrating content jobs, webhook triggers and third-party integrations through n8n.
- Owned latency and scalability of the LLM serving path on FastAPI and Azure under high content throughput.

Python · FastAPI · LangChain · n8n · GPT-4 / Claude · Azure · PostgreSQL

Backend Developer · National Health Authority, India (C4GT open source) Jul 2023 – Sep 2023

- Built the subscription management microservice for the ABHA national health portal from scratch, automating consent-driven subscription flows across health services.
- Selected for Code for GovTech, a competitive national mentorship programme for Digital Public Goods.

Spring Boot · Java · MongoDB · Docker

SELECTED BUILDS

Compliance Audit Agent

Agentic API that ingests a URL, scrapes it autonomously, judges it against a written policy with an LLM and returns structured non-compliant findings with justification.

LangChain, Flask, Pydantic, OpenAI

WhatsApp Virtual Try-On

Zero-human-in-the-loop automation: Twilio WhatsApp inbound, diffusion try-on model, result delivered back to the user. Flask, Twilio, Gradio

RAG Document Intelligence API

End-to-end retrieval-augmented QA over unstructured documents with embedding retrieval, grounded generation and an evaluation harness, exposed as REST. LangChain, FAISS, FastAPI

In-Memory-Compute NN Accelerator

Evaluated LeNet, AlexNet, VGG and GoogLeNet for in-memory computation substitution, cutting average execution time by 3 minutes. TensorFlow, Keras, Simulink

TECHNICAL

Agents & LLM	LangGraph, LangChain, CrewAI, n8n, Make, Flowise, Dify, OpenAI API, Anthropic API, Claude Code, prompt engineering and evaluation
RAG & Infra	RAG pipelines, FAISS, Chroma, AWS Bedrock, Langfuse, LLM observability, eval harnesses, PII/PHI redaction
Backend	Python, FastAPI, Django, Node.js, Go, Java, REST, GraphQL, PostgreSQL, MongoDB, Redis, microservices, event-driven design
Cloud & DevOps	AWS (Lambda, EKS, ECS, RDS, DynamoDB, S3, SQS), Azure, GCP, Docker, Kubernetes, Terraform, ArgoCD, GitHub Actions
Data & Monitoring	Kafka, Spark, Prometheus, Grafana, Loki, Langfuse, K6 load testing, Metabase

EDUCATION & RECOGNITION

- **B.Tech, Electronics and Communications Engineering**, IIIT Delhi (2020 – 2024). Undergraduate research assistant, VLSI Lab.
- Deep Learning Specialization and TensorFlow Developer Specialization (Coursera). Generative AI for Data Privacy and Protection (Coursera).
- Google Hash Code global rank under 900 of 2,447 teams. CodeChef 4-star. Open-source contributor to India's national digital health infrastructure.