

Jaswant Singh

GenAI Engineer | MLOps | Full-Stack (MERN)

Mandi, India | jaswantsinghjsn@gmail.com
linkedin.com/in/jaswant-singh009 | github.com/jaswantsinghh

SUMMARY

Full-Stack GenAI Engineer with **4+ years** shipping production AI/ML systems, including RAG pipelines, agentic workflows, and LLM automation on AWS, GCP, and Azure. Strong on both sides of the stack: **MLOps** (CI/CD, observability, model governance) and modern web engineering with React/Next.js, Node.js, and Python. Comfortable owning features from requirements through to deployment, with real cost and performance wins to show for it.

CORE SKILLS

Languages	Python, JavaScript, TypeScript, Java, C++, SQL
AI / GenAI	LLMs, LangChain, LangGraph, RAG, Agentic AI, prompt engineering, fine-tuning; Vector DBs (Pinecone, ChromaDB, FAISS); TensorFlow, PyTorch, Keras, Scikit-learn
MLOps & Cloud	AWS (Bedrock, SageMaker, Lambda, S3), GCP (Vertex AI, Cloud Functions), Azure OpenAI; Docker, Kubernetes, GitHub Actions, CI/CD; model versioning & governance; observability (metrics, alerts, drift detection)
Backend	Node.js, Express, Django, Flask, FastAPI; REST APIs, async & real-time streaming; Kafka, Redis, Firebase, serverless
Frontend	React, Next.js, React Native, Flutter; Redux, Zustand; Tailwind, ShadCN, MaterialUI, Framer Motion, SCSS
Databases	MongoDB, MySQL, Couchbase, Cassandra
Automation	Streamlit, Playwright, RPA orchestration, Power Platform (Apps / Automate / BI); Git, GitHub, GitLab

WORK EXPERIENCE

Intelligaia Technologies Pvt. Ltd., Senior Software Engineer (Full-Stack / AI)

Sept 2025 - Present

- Turn regulatory and business requirements into production AI tools for enterprise clients (**Cisco, Palo Alto Networks**); improved workflow efficiency by over **30%**.
- Own full-stack delivery of AI applications end to end: automated testing, CI/CD, and observability (metrics, alerts, health checks, drift detection).
- Build AI-driven automation and analytics with the OpenAI API, LangChain, Playwright, Flask, and Streamlit; also maintain secure Next.js front-ends (auth, middleware, dynamic API integrations) and present progress weekly to cross-functional client teams.

Key projects

- AuctionWire**: Graded trading-card auction platform (Next.js, Node.js) with a FastAPI OCR microservice that reads slab-certificate numbers from listing images across 7 auction houses and validates them against Gemrate. Fine-tuned a Qwen vision-language model with **QLoRA** and raised OCR accuracy from **85% to 98% (95%+ on rotated cards)**, removing the need for a third-party OCR provider.
- Cisco Outshift, Internet of Cognition (IoC)**: Built a proof-of-concept that turns a plain-language problem statement into a live multi-agent “cognitive fabric,” using GPT-4o to generate the agents, their staged reasoning, and the connection graph between them.
- Bhutan Smart Driving Test (Govt., via NGN)**: Real-time computer-vision system (YOLO11-seg with SAM segmentation and optical-flow tracking) that scores the parking/box test from live cameras, timing each run and counting how often a tyre touches the box boundary. Built with FastAPI and Angular.
- Saransh (Meeting Summariser)**: Self-hosted agent that joins a Google Meet through Playwright, transcribes the audio locally with faster-whisper, and writes a structured summary with an LLM. Runs fully in Docker with no SaaS dependencies.
- Nucleux**: Agentic-AI React component library (@nucleux) of composable, accessible, Tailwind-based building blocks for AI-chat and agent interfaces, shipped as tree-shakeable packages alongside a pattern-library site.
- Palo Alto Networks**: Built a React/TypeScript design system, converting Figma designs into production-ready code through an MCP-driven design-to-code workflow.
- Internal platforms**: An employee digital-profile and business-card system (Next.js, PostgreSQL/Prisma) with QR sharing, vCard export, and per-employee visitor analytics, plus a PM2 process-monitoring dashboard.

Superfine Cyber Security, GenAI / Full-Stack Developer

Oct 2023 - Aug 2025

Led GenAI and full-stack development of SaaS products on GCP and AWS.

- **PChat (PHI/PII privacy layer):** Built an Architecture Assistant that validates uploaded documents against compliance parameters, saving **3 FTEs (~US\$100K/yr)** for the client. Implemented a HIPAA-compliant assistant-thread architecture on the OpenAI API, cutting token usage for **5× efficiency** and reducing PHI/PII exposure by **90%+**. Integrated Kafka for real-time pipelines and Redis for in-memory state.
- **ComplyWise (regulatory compliance automation):** Built a security tool that detects and remediates vulnerabilities across GitHub repos and AWS workflows, cutting manual compliance verification time by **80%**. Designed a custom RAG pipeline for compliance-report generation and a vector-DB embedding store that sped retrieval **50%** while cutting cost **90%**. Added a real-time OpenAI voice agent over WebSockets.
- **DentalBuddy (AI dental marketplace & analytics):** Delivered a Dentist-Discovery directory with **95%** match accuracy and a fine-tuned AI chat for dental queries. Built multi-role authorization (Admin / Staff / Dentist) and real-time analytics that increased appointment bookings **60%** and reduced admin work **40%**.

Ryaz.io, Full-Stack Developer

Jun 2021 - Sep 2023

- Re-engineered a web-based teaching platform, cutting page-load time by **65%** (from 3.2s to 1.1s) and raising the Lighthouse accessibility score from **78 to 95**.
- Built a GenAI chatbot with model orchestration and a streaming response pipeline; migrated a ReactJS codebase to React Native.
- Shipped **10+** core features in React/Next.js, focused on responsive, accessible, SEO-friendly UI.

Freelance, Web Developer

Jan 2020 - May 2021

- Delivered responsive marketing sites and a custom Shopify storefront (**99.9%** uptime) for early clients including Wandering Compass, Gurpal Films, and Bistro 226, plus Bob The Builder, an audit-automation portal.

SELECTED PROJECTS & RESEARCH

- **Agentic AI Assistant Framework:** Multi-agent LangChain system with memory persistence and tool integration, deployed on AWS Lambda with async streaming (<500 ms) and CloudWatch observability.
- **Context-Window Optimization (research):** Prompt slicing, dynamic summarization, and embedding-fallback strategies that cut latency **40%** and removed context overflow in long conversations; published reusable components for the team.
- **AIOps Workflow Optimizer:** Observability system for LLM pipelines with automated retries, error handling, and drift detection; reduced model failures **70%**.

EDUCATION

B.Tech Computer Science, Guru Nanak Dev Engineering College (GNDEC), Ludhiana
 Schooling: CBSE, MPS Sundernagar (Class XII 85%, Class X 10.0 CGPA)

2019 - 2023, CGPA 8.8/10
 2019