Yash Nilesh Gori

Mumbai, MH | yashnileshgori@gmail.com | (+91)7718081766 | Portfolio | linkedin.com/in/yashgori20 | github.com/yashgori20 | Huggingface

Summary

AI Engineer with expertise in architecting and deploying large-scale GenAI and backend systems. Skilled in building RAG pipelines, API-driven architectures, and production-grade ML integrations with strong command of embedding systems and vector databases. Experienced in implementing CI/CD workflows, and complex AI prototypes into scalable enterprise solutions.

Experience

AI Engineer (IC) | Webotix IT Consultancy (Early-stage AI SaaS startup;GenAI QA) *Dec 2024 – Sep 2025*

- **Architected and deployed** low-latency RESTful APIs (FastAPI, Flask) powering GenAI inference and compliance validation across multi-tenant enterprise environments, achieving **sub-500ms** average response time.
- **Built scalable microservices** with Azure Cosmos DB, Redis caching, and containerized Docker workflows, improving throughput by **2.3×** and cutting server costs by **30%** through optimized data handling.
- **Developed RAG-based retrieval system** combining Azure Document Intelligence OCR and vector search, reducing document parsing time by **90%** while sustaining **80%+ validation accuracy**.
- **Implemented CI/CD automation** via GitHub Actions and Docker, ensuring zero-downtime and improving velocity by **40%**.
- **Delivered production-ready prototype** within 4 weeks using open-source infrastructure, demonstrating rapid execution under constrained resources **securing $5,000 Microsoft AI Founders Hub grant** for the same
- **Developed Power BI dashboards** to monitor system KPIs (latency, uptime, accuracy), enabling real-time visibility.
- Demoed **AI workflows** to Dubai-based pilot clients and implemented improvements that boosted user adoption by **30%**

Product Management Intern | MetaRizz (IT consultancy & product solutions firm) *Dec 2023 – May 2024*

- Operated as **product owner for two client apps** GuestInMe (UX revamp) and MediNobel (hospitality platform) defining feature requirements, release scope, and sprint deliverables in collaboration with engineering teams.
- Worked closely with designers to revamp GuestInMe's UX and ship new features including table booking and club pass purchases, driving a **40%** increase in engagement and growing the active user base to **1K+ users**
- Led the **0→1** development of MediNobel, coordinating design-engineering workflows from prototype to production.

Business & Growth Manager | Watermelon Gang (Social media & marketing agency) *Aug 2022 – Nov 2023*

- Acquired and managed **5 B2B clients** in fintech and AI, driving growth and efficiency through structured problem-solving.
- Scaled Ali Solanki's YouTube channel from **50K → 70K** subscribers by implementing KPI-based content iteration cycles.

Projects

Swift Check AI: QC Platform | *Azure OpenAI, Flask, Document Intelligence, Cosmos DB, Redis, Docker*

- Architected **multi-tenant RAG** microservice (Azure OpenAI + FAISS-style retrieval) for compliance validation.
- Exposed platform as RESTful APIs with FastAPI for template creation and compliance checks.
- Cut document generation time **90%**; supported **25+** configurable parameters and maintained **80%** compliance accuracy with sub-second cached responses.

Financial Compliance Automation Tool | *Python, Mixtral LLM, GROQ Cloud, FAISS, Model Training* Link

- Built **custom RAG pipeline** trained on RBI docs; iterative refinement achieved **80% compliance prediction accuracy**.
- Demonstrated ₹10% potential annual cost-savings opportunity for banks via automated checks.

Inhance: LinkedIn and Resume Profile Optimization Platform | *Streamlit, GROQ Cloud, Mixtral LLM, Multi-Agent System* Link

- Implemented **multi-agent evaluation** and ATS-scoring pipeline; produced role-optimized suggestions and LaTeX resume.
- Delivered automated, actionable recommendations and an interactive enhancement agent for profile tuning.

DocuTalk: AI-Powered Document Intelligence Platform | *Python, FAISS, LangChain, Gemini Embeddings, Flask* Link

- Engineered semantic search + conversational layer using **Gemini embeddings + FAISS**; handled context-aware Q&A.
- Built REST endpoints with FastAPI and integrated with cross-platform **Flutter** frontend, reducing manual review from hours to minutes & Optimized retrieval latency with caching strategies.

Skills

Programming & Frameworks: Python, REST, FastAPI,SQL/NoSQL (Cosmos DB, Redis, MySQL basics)
AI/ML Core: Retrieval-Augmented Generation (RAG) systems, prompt engineering, vector databases (FAISS, Pinecone), LLM integration (GPT, Mixtral, LLaMA, Gemini), OCR processing, NLP, multimodal AI (text + voice)
DevOps & Cloud: CI/CD (GitHub Actions), Google Cloud (Vertex AI, BigQuery ML), Docker, Azure (OpenAI, Cosmos DB)
Data & Analytics: Pandas, NumPy, model evaluation (AUC, F1), Power BI, Hadoop, Spark
Professional Skills: Translating business needs into AI solutions, pre-sales demos, solution architecture.

Education

K. J. Somaiya College of Engineering Mumbai, Maharashtra
B.Tech. in Information Technology | CGPA: 8.12/10 *2021–2025*

Impact and Achievements

Volunteer Instructor | Vacha NGO

- Designed and delivered structured educational programs for **30+ underprivileged students** creating lesson plans, interactive activities, and learning materials to enhance engagement and understanding.
- Conducted a **career guidance session** to help students identify strengths, set academic goals, and future opportunities.

Semi-finalist | Devopia 7.0 Hackathon - Selected among the **top 3 teams** for delivering an AI-powered solution.