

ABHISHEK SHARMA

AI/ML Engineer - Agentic AI, Production RAG & Real-Time Voice

Email: abhi.aidev@gmail.com | LinkedIn: [linkedin.com/in/abhi-ai](https://www.linkedin.com/in/abhi-ai) | GitHub: github.com/Abhishek676062

Location: Indore, Madhya Pradesh, India | abhi.ai.tech

SUMMARY

AI/ML Engineer with 2+ years of experience shipping production multi-tenant RAG, real-time voice agents, and computer vision systems across 50+ enterprise tenants.

HEADLINE METRICS

- 287 ms p95 retrieval-to-answer latency across 50+ enterprise tenants on Synapse Arabic-English RAG
- 3 s to 200 ms voice turn latency after migrating cascaded pipeline to duplex speech-to-speech (LiveKit + Gemini Live)
- 90% multi-camera re-identification accuracy on retail/QSR occupancy analytics (YOLOv11 + BoT-SORT)
- 15+ ML models transitioned from experimental notebooks to containerized production AWS services
- 5 hours to 2 minutes case review time reduction on DECIMA pathology diagnostic system (90% accuracy on 100+ slides)

EXPERIENCE

Shanti Infosoft LLP (CMMI Level 5) - AI/ML Engineer (Dec 2025 - Present)

- Architected Synapse, a multi-tenant Arabic-English agentic RAG platform: LangGraph, Qdrant, retrieval RBAC.
- Built multilingual speech-to-speech sales agent handling 70% of inbound client conversations autonomously.
- Rebuilt multi-camera occupancy tracking with TensorRT FP16, NVDEC decoding, and tuned BoT-SORT.

Techlene Software Solutions Pvt. Ltd. - Python AI Developer (Aug 2024 - Nov 2025)

- Designed DECIMA GenAI + CV pathology system: WBC/RBC cell counting & LLM medical reports.
- Shipped schema-grounded NL2SQL ERP assistant with LangChain/LangGraph, clearing 2-day analyst backlog.
- Containerised 15+ models as FastAPI microservices with Airflow CI/CD on AWS.

EDUCATION

Rajiv Gandhi Proudhyogiki Vishwavidyalaya (RGPV) - B.Tech in Information Technology (2020 - 2024, CGPA 7.6/10)

SKILLS

LangGraph, LangChain, Qdrant, Pinecone, LiveKit, Gemini Live, YOLOv11, TensorRT, FastAPI, Docker, AWS, DeepEval, Python, TypeS