Jayant Khandebharad

+91 7517678382 • jntkhandebharad@gmail.com • linkedin.com/in/khandebharad • github.com/Jayantkhandebharad • Pune, India

SUMMARY

Agentic AI & Platform Engineer with 3+ years building **production-grade agentic AI systems, RAG pipelines, and multi-agent orchestration platforms** on Python, LangGraph, LangChain, FastAPI, and Azure. Designed end-to-end intelligent workflows with planning, memory, tool use, and multi-provider LLM abstraction — deployed at scale on containerized cloud infrastructure with full observability across multi-tenant production environments.

WORK EXPERIENCE

CloudLex — SDE II, Generative AI & Platform *Jan 2023 – Present*

Legal-tech SaaS — multi-tenant AI platform for personal-injury law firms (Azure, Python).

- **Agentic AI Platform (multi-service ecosystem):** Designed and shipped a monorepo of coordinated Python services powering a production agentic-AI platform — FastAPI APIs, LangGraph-orchestrated agents, background workers, and Azure Functions — with planning, memory (Cosmos DB checkpoints), tool use (vector search, CloudLex REST, document intelligence), and multi-provider LLM abstraction (Azure OpenAI, Anthropic Claude, Groq Llama, Google Gemini). Owned architecture, implementation, CI/CD, and operations end-to-end.
- **Agentic workflows & multi-agent orchestration:** Built category-aware LangGraph *StateGraph* agents with dynamic tool binding, context-switching state machines, and Cosmos DB-backed conversation memory; agents autonomously plan retrieval strategies, select tools per query category, and chain reasoning across medical-document analysis, case summarization, and draft generation.
- **RAG pipelines & vector search:** Engineered a GraphRAG pipeline indexing **10K+** documents (retrieval precision **~35%**, latency **~40%** lower) — an 8-section medical-overview extraction over Azure AI Search (vector + semantic + hybrid, 1024-dim embeddings), and two-phase deduplication (deterministic identity-tuple + Gemini semantic cleanup); retrieval tools with `InjectedState` for secure, hallucination-resistant grounding and page-level citations.
- **Production voice agent (PSTN IVR):** Architected a multi-tenant PSTN-connected voice AI agent on Azure Communication Services — real-time bidirectional streaming with Azure Neural TTS/STT, LLM tool-calling for human escalation (Bandwidth SMS context-briefing), and a TTL-windowed fallback sweeper that recovers orphaned conversations when disconnect webhooks never arrive — handling **400+** calls/month autonomously with **~60%** faster response.
- **Observability & distributed tracing:** Authored `lexee_logging`, an in-house structured-logging package with OpenTelemetry, correlation-ID propagation across HTTP/queue/worker boundaries, ContextVar scopes, and a 9-category error-code registry (`LEX-XNNN`) — enabling end-to-end tracing via a single KQL query in Azure Log Analytics.
- **Containerized deployment & CI/CD:** Operated a parameterized multi-stage Azure DevOps pipeline deploying Docker containers to Azure Container Apps, Container App Jobs, and Functions — multi-Dockerfile monorepo, dual-tag strategy, revision-based rollback, Key Vault secret injection, and parallel stages so only changed components rebuild — cutting cloud cost **~22%** at **99.9%** uptime.
- **Data security & governance:** Implemented HS512-JWT middleware with cross-tenant isolation (`fid/uid` claim validation against query params and body), Azure Key Vault, managed identity, and per-firm data-access policies — blocking cross-tenant leakage at the edge.

Eaton — Software Development Engineer Intern *Jun 2022 – Jul 2022*

- Delivered 4+ customer-management and access-control screens in Angular 8 to enterprise coding standards; profiled and optimized grid rendering with Chrome DevTools.

TECHNICAL SKILLS

Agentic AI & LLM Platforms: LangChain, LangGraph, LlamaIndex, CrewAI, DSPy, Semantic Kernel, LiteLLM, agentic workflow design (planning, memory, tool use), multi-agent systems, AI agent orchestration, state machines, tool calling, MCP, Agent2Agent (A2A)

RAG & Vector Databases: RAG, GraphRAG, vector search & embeddings, Azure AI Search (vector + semantic + hybrid), FAISS, Pinecone, Weaviate, Qdrant, prompt engineering (CoT, few-shot), LLM evaluation & guardrails

LLM Providers & ML: Azure OpenAI (GPT-4o), OpenAI, Anthropic Claude, Groq (Llama), Google Gemini, fine-tuning (PEFT, LoRA, SFT), deep learning, NLP, Hugging Face Transformers, PyTorch, scikit-learn

Languages & Frameworks: Python, FastAPI, Flask, Pandas, NumPy, Pydantic, SQLAlchemy, C++, SQL, Bash

Databases: PostgreSQL, MySQL, Cosmos DB (NoSQL), Redis, Azure Blob Storage, DuckDB

Cloud & DevOps: Azure (Container Apps, Functions, Cognitive Services, Key Vault, DevOps Pipelines), Docker, Docker Compose, CI/CD, Azure Container Registry, managed identity

Observability & Governance: OpenTelemetry, Azure App Insights, Log Analytics, KQL, Prometheus, Grafana, structured logging, distributed tracing, correlation-ID propagation, PII handling, data governance, multi-tenant security, JWT/RBAC

CS Fundamentals: Data Structures & Algorithms, System Design, OOP, Design Patterns, Distributed Systems, Concurrency & Locking, Event-Driven Architecture

KEY PROJECTS

Lexee Ecosystem — Production Agentic-AI Platform: LangGraph agents, FastAPI APIs, background workers (PostgreSQL `SELECT FOR UPDATE SKIP LOCKED`), Azure Functions (queue sync, token tracking), React SPA — multi-tenant case-aware chat, document intelligence, and draft generation; provider-agnostic LLM layer (Azure OpenAI, Claude, Groq, Gemini) with per-request token telemetry (Redis→Queue→Cosmos) for per-tenant cost attribution.

NL-to-SQL Analytics Chatbot: Streamlit + PandasAI + DuckDB natural-language analytics chatbot with LangGraph state-machine orchestration for interactive data exploration.

EDUCATION

B.E., Computer Science — Pune Institute of Computer Technology (PICT), Pune | CGPA: **9.08 / 10** Jul 2019 – Jul 2023

ACHIEVEMENTS

Smart India Hackathon 2022 — National Finalist • **620+** LeetCode problems, CodeChef **1801** • **MHT-CET 99.71 percentile** (top 1% of ~300,000).