Jayant Khandebharad

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

SUMMARY

AI / Generative-AI Engineer (3+ yrs) building production **agentic-AI, RAG, voice, and document-intelligence** systems on Python, LangGraph/LangChain, FastAPI, and Azure. Design Python backends for **LLM inference orchestration**, async/distributed pipelines, and multi-agent orchestration (planning/memory/tool-use); build embedding/retrieval and **prompt-evaluation** systems with multi-provider LLMs. Comfortable owning **system design end-to-end** and the full ML lifecycle from data → deployment → monitoring.

TECHNICAL SKILLS

Languages: Python, SQL, C++, Bash

Generative & Agentic AI: LLMs, RAG, GraphRAG, vector search & embeddings, LangGraph, LangChain, LlamaIndex, multi-agent orchestration, agentic workflows (planning/memory/tool-use), prompt engineering & evaluation, LLM-as-a-judge evals, guardrails, MCP, LiteLLM; Azure OpenAI (GPT-4o), Anthropic Claude (incl. via AWS Bedrock), Google Gemini, Groq (Llama)

Document & Voice AI: OCR & document intelligence (PDF / vision extraction), Groq Batch API / batch inference, speech-to-text & TTS, real-time PSTN voice agents (Azure Communication Services), embeddable chatbots (Azure Bot Framework / Direct Line)

ML & Deep Learning: PyTorch, scikit-learn, NLP, Hugging Face Transformers, deep learning, transformers from scratch, regression, clustering, forecasting, ensemble methods, fine-tuning (LoRA/PEFT)

MLOps & LLMOps: CI/CD for models, experiment & token/cost tracking, model & prompt lineage, prompt-eval & regression, continuous retraining; MLflow & Databricks (working knowledge)

Cloud & Data: Azure (Container Apps, Container App Jobs, Functions, AI Search, Key Vault, DevOps Pipelines), Docker, PostgreSQL/pgvector, MySQL, Cosmos DB, Redis, DuckDB, Azure Blob/Queue; Pandas, NumPy

Observability & Responsible AI: OpenTelemetry, App Insights, Log Analytics, KQL, structured logging, distributed tracing, bias/guardrails, PII handling, data governance, multi-tenant security, JWT/RBAC

CS Fundamentals & System Design: end-to-end system design, distributed systems, concurrency & locking, event-driven architecture, design patterns, Data Structures & Algorithms

WORK EXPERIENCE

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

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

- **Agentic AI & multi-agent orchestration:** Built LangGraph *StateGraph* agents with planning, dynamic tool-binding, context-switching state machines, and Cosmos DB-backed **memory** that preserves context and tracks **embeddings** — agents autonomously plan retrieval, select tools per query, and chain reasoning across document analysis, summarization, and draft generation — cutting manual review workload **~50%**.
- **Python backends for inference orchestration & distributed workflows:** FastAPI services plus queue-driven **async workers** (Azure Service Bus / Storage Queues, queue-based autoscaling, PostgreSQL `SELECT FOR UPDATE SKIP LOCKED`) for high-throughput LLM/document processing — idempotent retries, dead-lettering, and a claim→submit→analyze state machine sustain throughput and correctness under load.
- **Document intelligence (RAG + OCR + batch):** Engineered an 8-section medical-overview pipeline that **OCRs** case files (PDF/vision extraction) via the **Groq Batch API** (batch inference), then chunks, embeds, and indexes into Azure AI Search (vector + semantic + hybrid, 1024-dim) over pgvector with two-phase dedup and page-level citations for hallucination-resistant grounding across **10K+** documents (retrieval precision **~35%**, latency **~40%** lower).
- **Real-time voice AI (PSTN callbots):** Built a multi-tenant PSTN-connected voice agent on Azure Communication Services — real-time bidirectional **speech↔LLM** (Azure Neural STT/TTS), tool-calling for live human escalation, and a TTL-windowed fallback sweeper that recovers orphaned calls when disconnect webhooks never fire — **400+** calls/month with **~60%** faster response.
- **Prompt evaluation & Responsible AI:** Built an **LLM-as-a-judge evaluation** service scoring live conversations against rubric metrics across bot surfaces, with audit logging and diff-reviewed config — surfacing prompt drift and enforcing guardrails, PII handling, and per-tenant isolation.
- **LLMOps, CI/CD & observability:** Built a provider-agnostic LLM layer (Azure OpenAI, Claude, Gemini, Groq via LiteLLM) with per-tenant **token/cost telemetry** (Redis→Queue→Cosmos) for model/usage lineage; ship Docker containers via a multi-stage Azure DevOps pipeline to Azure Container Apps/Jobs/Functions; authored `lexee_logging` (OpenTelemetry, correlation-IDs) for one-KQL-query production tracing — **~22%** lower cloud cost at **99.9%** uptime.

KEY PROJECTS

Lexee / Amicus — Production Agentic-AI Platform: multi-service Python platform (FastAPI, LangGraph supervisor + firm-scoped sub-agents, async Service Bus / queue-driven worker tier on Azure Container Apps) with embedding + memory subsystems, multi-provider LLMs, and full observability.

LLM Evaluation & Bot-Config Console: FastAPI + React tool running Azure-OpenAI LLM-judge evaluators (rubric-scored, prompt-aware) across four conversation stores, with diff-reviewed config CRUD and an audit trail — operationalizing prompt-evaluation/regression for production bots.

LLM-from-Scratch & ML Foundations: implemented transformer internals (BPE, RoPE, RMSNorm, SwiGLU, attention) and core ML/DL workflows in PyTorch — grounding GenAI work in fundamentals (regression, clustering, deep learning).

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).