Jayant Khandebharad

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

SUMMARY

Software Engineer with 3+ years building production **Generative-AI, ML, and full-stack** systems end-to-end on Python, FastAPI, LangGraph/LangChain, and Azure. Ship agentic-AI platforms, RAG pipelines, real-time voice agents, and document-intelligence tools — owning design, development, deployment, and monitoring across the full life cycle. Strong in distributed, event-driven backends, **LLMOps** (token/cost telemetry, prompt evaluation), and grounding applied GenAI in ML fundamentals (PyTorch, transformers from scratch). Focused on reliability, guardrails, and cost.

TECHNICAL SKILLS

Languages: Python, TypeScript, SQL, C++, Bash

Generative & Agentic AI: LLMs, RAG, GraphRAG, vector search & embeddings, LangGraph, LangChain, LlamaIndex, LiteLLM, multi-agent orchestration, agentic workflows (planning/memory/tool-use), prompt engineering & evaluation (LLM-as-judge), guardrails, MCP

LLM Providers: Azure OpenAI (GPT-4o), Anthropic Claude, Google Gemini, Groq (Llama), Hugging Face Transformers

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

Voice & Document AI: Azure Communication Services (PSTN), Azure Speech (STT/TTS, SSML), OCR & document intelligence, Azure AI Search

Backend & Data: FastAPI, Azure Functions, async workers & queues (Service Bus / Storage Queues), REST APIs, PostgreSQL/pgvector, Cosmos DB, MySQL, Redis, Pandas, NumPy

Cloud, DevOps & LLMOps: Azure Container Apps & Jobs, Docker, Azure DevOps CI/CD, Bicep, Key Vault, Application Insights / Log Analytics / KQL, OpenTelemetry, token/cost & prompt-eval tracking

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

WORK EXPERIENCE

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

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

- **GraphRAG & retrieval:** Built a GraphRAG pipeline (LiteLLM + Cosmos DB on Azure) indexing **10K+** legal documents — improving retrieval precision **~35%** and cutting latency **~40%**; vector + semantic + hybrid search over pgvector with page-level citations for hallucination-resistant grounding.
- **Multi-agent LLM systems:** Designed LangGraph *StateGraph* agents for summarization, document review, and task generation — planning, dynamic tool-binding, and Cosmos DB-backed memory — automating client intake and cutting manual review workload **~50%**.
- **Real-time voice AI (PSTN):** Created a natural-language voice/IVR agent (Azure Communication Services + Azure Speech + GPT-4o) handling **400+** calls/month autonomously and reducing response time **~60%**, with tool-call human escalation.
- **Python backends & distributed workflows:** Engineered FastAPI services plus queue-driven **async workers** (Azure Service Bus / Storage Queues, PostgreSQL `SELECT FOR UPDATE SKIP LOCKED`) with idempotent retries and dead-lettering — sustaining throughput and correctness under load.
- **Full-stack delivery:** Built multiple LLM chatbots (helpdesk, multi-tenant intake, PIP/MVA) and an embeddable web widget (Azure Bot Framework / Direct Line) with FastAPI back-ends and React/TypeScript front-ends.
- **LLMOps, CI/CD & cost:** Built a provider-agnostic LLM layer with per-tenant token/cost telemetry and an LLM-as-judge prompt-evaluation service; shipped via multi-stage Azure DevOps pipelines to Container Apps/Jobs/Functions — cutting cloud costs **~22%** and maintaining **99.9%** uptime.

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.

KEY PROJECTS

Agentic AI Platform: Multi-tenant agentic-AI ecosystem — FastAPI APIs, a LangGraph supervisor with firm-scoped sub-agents, and a Service Bus queue-driven worker tier on Azure Container Apps; pgvector retrieval, multi-provider LLMs (Azure OpenAI / Claude / Gemini / Groq), and full observability.

GraphRAG Document Intelligence: 8-section medical-overview extraction — OCR (batch inference) → chunk/embed → Azure AI Search (vector + semantic + hybrid) with two-phase deduplication and page-level citations.

Chronological Medical Summary (full-stack): FastAPI + React/MUI app that turns a case's medical records into a dated clinical timeline via LangChain structured-output extraction (GPT-4o), with one-click branded PDF/DOCX export.

LLM from Scratch (PyTorch): Decoder-only Transformer — BPE tokenizer, RoPE, RMSNorm, SwiGLU, pre-norm blocks — targeting Stanford CS336 test contracts.

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).