

# IBRAHIM SULTAN ABDULAZIZ

AI Software Engineer

ibrahimsultan4705@gmail.com | linkedin.com/in/-ibrahim-sultan | +91 8454080451 | github.com/xIBBAx

## EXPERIENCE

---

### Software Development Engineer (G2)

Oct 2025 – Present

63 Moons Technologies Ltd.

Mumbai, Maharashtra

- Hardened Qi Flash's **Python/FastAPI** retrieval backend with **Vespa** search routing, evidence-based web fallback and source-linked citations for legal research and uploaded documents.
- Implemented **Gemini fallback through LiteLLM** before the first model output, plus cancellation handling and sanitized **SSE** errors that retain partial responses.
- Added per-user **token-budget admission** with renewable **Redis locks**, and subscription checks for uploads, downloads and files carried through chat history.
- Restricted AI web tools to validated public HTTPS URLs and bounded searches; hardened request logging to keep private queries, responses and identifiers out of routine logs.
- Built and containerized **local CLM document and Granite embedding services**, with batched inference, bounded request queues and vector-dimension validation.

### AI/ML Integration Engineer

Aug 2024 – Sep 2025

Techpeek Magnus Solutions Pvt. Ltd.

Bangalore, Karnataka

- Built legal-AI chat features and integrated enterprise data sources with **Vespa-backed RAG** for document search and question answering.
- Automated tribunal-order ingestion with **Python, SQL and OCR**, structuring legal documents for search and case-outcome dashboards.

### AI/ML Software Engineer

Jun 2023 – Mar 2024

ILM UX Pvt. Ltd.

Ghansoli, Navi Mumbai

- Refined **IXCAN-A**, a colon-cancer prediction model, through data cleaning and clinician feedback.

## EDUCATION

---

### Pillai University

Jan 2021 – Jun 2024

B.Tech in Electronics & Computer Science | CGPA: 8.42

New Panvel, Navi Mumbai

## SKILLS

---

**Languages:** Python, SQL, TypeScript

**AI / Retrieval:** RAG, Vespa, hybrid search, BM25, embeddings, citation grounding

**LLM Systems:** Gemini, LiteLLM, tool calling, model fallback, SSE streaming

**Backend / Data:** FastAPI, async Python, Pydantic, PostgreSQL, Redis

**Infrastructure / Quality:** Docker, AWS S3, pytest, regression testing