

# SHUBHAM

AI Engineer | 4 Production AI Systems on Azure | RAG | AI Agents | Text-to-SQL | Fine-Tuning

Bangalore, India | +917808968464 | shubh12pp@gmail.com

## PROFESSIONAL SUMMARY

AI Engineer with end-to-end production ownership across 4 production AI systems on Azure for a global cold logistics company, spanning design, deployment, fine-tuning, and LangSmith-based evaluation. Specializes in AI Agents, hybrid RAG, and Text-to-SQL, with enterprise integration experience supporting operations and legal teams.

## TECHNICAL SKILLS

**GenAI & LLMs:** RAG, AI Agents, Fine-Tuning, Prompt Engineering, LLM Evaluation, Tool Calling, ReAct, Tree-of-Thought, Workflow Orchestration

**Languages & Frameworks:** Python, SQL, PyTorch, Transformers, Hugging Face, LangChain, LangGraph, FastAPI

**Cloud & Deployment:** Azure, Azure AI Foundry, Azure OpenAI, Azure Web Apps, Azure Logic Apps, Azure Communication Services, Azure Speech Services, Docker

**Retrieval & Data Systems:** Hybrid (Dense + Sparse) Retrieval, Embeddings, Vector Databases, Metadata Filtering, FAISS, ChromaDB, Pinecone, ColBERT v2, SPLADE v2, Sentence Transformers, Text-to-SQL, Schema Linking

**Evaluation & Observability:** LangSmith, Precision/Recall/F1, Observability, Latency Monitoring, Token Usage Tracking, Tool-Call Tracing

## PROFESSIONAL EXPERIENCE

### NewCold

AI Associate

December 2024 - Present

- Deployed a LangGraph + ReAct voice/chat SOP AI Agent for workflow orchestration across **32 warehouse workflows** at **22 global sites** for 4 user groups, integrating with Azure Logic Apps and Azure Communication Services to automate workflow completion capture, manager email notifications, and follow-up call triggers, with **3s p50** and **4s p95 latency** while reducing manual SOP lookup and escalation overhead.
- Built a production legal-document RAG assistant for the legal team over 67 OSA agreements (~17,000 chunks) using hybrid dense+sparse retrieval (ColBERT v2 + SPLADE v2), topic-based chunking, query transformation, and metadata filtering on client and warehouse/site, supporting retrieval, document comparison, and summarization queries with **92% recall** and **73% precision**.
- Developed a production truck-booking AI Agent using LangGraph with Tree-of-Thought reasoning to evaluate 5 operational constraints across 4 data sources and present human-in-the-loop slot options to clients, achieving **85% recommendation acceptance**, **~2-minute** time to valid slot recommendation, and a **4.2/5** client rating.
- Developed a schema-aware Text-to-SQL assistant for operations teams across 4 SQL tables covering bookings, clients, rates, and bidding, enabling self-service natural-language querying for bookings, rates, revenue, and client analysis with **82% execution accuracy** and Chart.js-based visualizations for prompts like box plots over past 6 months, reducing analyst dependency and ad hoc SQL work.
- Fine-tuned a domain-specific model for an internal assistant and implemented LangSmith-based evaluation, monitoring, and observability across deployed applications, tracking precision, recall, token usage, cost, latency, and tool-calling behavior.

### Reliance Retail

Data Analyst Intern

April 2024 - September 2024

- Processed and analyzed 7,000+ rows of sales data using Python and SQL, building data preprocessing and feature-engineering workflows that improved data accuracy to 98%.
- Generated data-driven insights on customer behavior and business performance, supporting reporting workflows and contributing to a 10% profit margin increase.

### Razevm Private Limited

Business Analyst

May 2022 - June 2023

- Analyzed sales data and improved reporting workflows through data cleaning, preprocessing, and business insight generation, contributing to a 15% revenue increase and 30% faster data processing.

## EDUCATION

### Master of Business Administration (MBA) - Business Analytics

ITM Business School, Mumbai | Batch Topper | 2023 - 2025

### Bachelor of Arts

BBMKU University | Grade A | 2022