Shubham Singh
shubhammit21@gmail.com | (+91) 7774941127 | [Shubham Singh | LinkedIn]

SUMMARY

Results-driven Full Stack Developer with 6.2 years of experience in building scalable web applications. Proficient in Python, Django, and Azure cloud technologies, integrating LLM capabilities for innovative applications. Expertise in DevOps, RESTful APIs, and front-end development with ReactJS. Proficient in Large Language Model contextualisation using RAG and prompt engineering and LLM orchestration leveraging Lang chain framework. Recognised at Ernst & Young for contributions towards best-in-class Gen AI application development for Financial Services clients.

SKILLS

Languages / Scripting:	Python, Pyspark, Shell Scripts
Databases / NoSQL:	MySQL, PostgreSQL, SQL Server, MongoDB
Framework:	Django, Flask, Django Rest Framework (DRF), Lang Chain
Front-End Technologies:	HTML, CSS, JavaScript, Bootstrap, HTMX, React (basic)
Repository:	Git, code commit, Docker
Cloud Platforms:	AWS, Azure, Databricks GCP
Message Brokers:	RabbitMQ, Redis
AI/ML Technologies:	Generative AI (Gen AI), Prompt engineering, Retrieval-Augmented Generation (RAG)

EXPERIENCE

Software Engineer at PricewaterhouseCoopers Pune

Role: AI / Application Developer – Generative AI **July 2025 – Present**

Projects:

Physical Risk Verification (PRV) Automation -life Insurance Domain

- Automated interpretation of unstructured field investigation remarks into meaningful categories.
- Built LLM-based adverse risk classification (fraud, identity mismatch, medical flags) with explainable outputs.
- Designed **Risk Banding Framework** combining PRV classification and policy health
- Implemented **address validation and correction** using APIs and LLM logic for improved data quality
- **Impact:** Reduced manual review effort, improved address accuracy, and established a foundation for fraud detection

Business Problem:

Manual PRV review was slow and inconsistent, causing underwriting delays and missed fraud risks.

Built an AI-driven system to auto-classify risk from field verification remarks.

Projects: **Customer Segmentation for Sales** Optimization – Life Insurance

Business Problem:

Sales teams lacked clear customer targeting, leading to low conversion rates.

Applied ML-based segmentation to enable focused and personalized sales efforts.

Key Contributions:

• Built customer segmentation models using K-Means, K-Modes, and DBSCAN.

• Processed large customer datasets using PySpark on Databricks.

• Identified high-value customer segments to improve sales conversion and engagement.

Software Engineer at Ernst & Young, Mumbai **July 2023 – July 2025**

Projects:

- **Obligation Extraction**: Gen AI powered tool for automating Obligation register creation from regulatory circulars as published by Central Bank. Application enabled efficiency gain by 70-80% for compliance teams.
- **Compliance GPT:** A Gen AI-powered chatbot that provides real-time access to regulatory circulars, compliance guidelines, policies, and procedures. Built on a Retrieval-Augmented Generation (RAG) pipeline, it leverages internal prompts and proprietary data to deliver context-aware, accurate responses**.**

- **Internal Audit Report Automation**: Developed a Gen-AI tool for automating Internal Audit and Committee report creation using unstructured data within a RAG framework. Leveraged GPT-4 and Llama 3.2 to identify the superior LLM, utilising Lang chain for orchestration. This application improved the Internal Audit team's efficiency by 40-50%
- **Personal Discussion Tool:** Gen AI powered tool which automates the dynamic questioning of borrowers leveraging their banking records and demographic profile (using a RAG framework). Additionally utilizes external API for income benchmarking. Application helped in achieving significant improvement in credit lending decision making.
- **CAM (Credit Assessment Memo Automation**): Gen AI-powered tool to interpret and analyze borrowers' financial statements and banking records and automate the CAM writing which is manually done by a credit underwriter. Achieved 30% efficiency gain by implementing this for an NBFC.

Tech stack: Python, Django rest Api, HTMX, JavaScript, CSS, Azure File system Data Blob storage SDK, Azure app service, Azure Open Ai, Llama, Mistral

Key Responsibilities:

- Developed and maintained front-end and back-end components of web applications.
- Designed Retrieval-Augmented Generation (RAG) pipelines for GPT models using Lang Chain, contributing to Compliance GPT Projects for India and MENA Market.
- Worked with various advanced models, including **GPT-4.0 (40B)** and **Llama 3.3 (70B - Versatile)**, ensuring high-quality outputs tailored to project needs.
- Developed web scraping solutions using **Selenium** and **Beautiful Soup** to extract data from bank circulars and compliance documents efficiently.
- Leveraged **Generative APIs** to summarise extracted data, enhancing decision-making and compliance reporting processes.
- Implemented advanced data extraction using Azure Form Recogniser and Cognitive Services.
- Optimised application performance with asynchronous tasks and parallel processing for enhanced scalability.
- Containerised the application using **Docker**, optimising deployment processes and ensuring scalability.
- Designed and implemented **CI/CD pipelines** in **Azure Pipelines**, automating build, test, and deployment workflows
- Worked as a **Tech Lead** for the **Obligation Extraction** project, leading the development of advanced AI- based solutions to extract and analyse regulatory obligations efficiently.
- Successfully deployed the application, streamlining delivery and improving operational efficiency.

Software Engineer at Statvalu Pvt Ltd, Bangalore **Nov 2021 – July 2023**

Project Name: DocuExpert is an AI-powered tool that helps professionals quickly review and process large documents. It saves time by summarising and highlighting key information automatically. This makes document work faster, easier, and more accurate for industries like finance, law, healthcare, and more.

Tech stack: Python, Django rest Api Postgres Database, Open data sources, Azure, File system Data Lake storage SDK ORM, Azure app service, Azure Security group, virtual machine, Aws lambda

- Proficient in Django web framework, including Django ORM, Django Rest Framework, and Django Templates.
- Strong understanding of web development concepts such as MVC architecture, RESTful APIs, and database management.

As a backend developer worked on different document processing module.

- Worked on document metadata extraction before loading the files and storing the information into db.
- Load client document from GPC cloud, google drive, s3 and load it into azure data lake storage.
  Azure Cognito service was used to further generate the OCR using this document.
- Created Api for upload file folders from google drive link & progress bar
- Handled google auth social login.
- Work on Azure function for extracting data from pdf

- Worked on the optimisation of this whole process.
  - Using python parallel processing named celeryd and celery beats scheduler job, the task which use to take 3hrs to upload 1k documents and now it is optimised to **45-50min**. So, I was able to achieve 75% faster result than before.
  - Loading bulk records into UI which use to take 3-4min to load 4k records before and this was optimized to ~300 milliseconds. Where I was able to achieve 99% faster response.
- Worked on different role base structure for authentication
- In-depth architectural knowledge in building REST API and Securing them with JWT Authentication
- Devops work includes
  - handling security for the infrastructure.
  - Worked on attaching auto scaling.
  - Worked on the Infra cost optimisation by choosing alternate service for deployment.

Software Developer at Untramelled Business Solutions (OPC) Pvt Ltd, Hyderabad **Nov 2020 – OCT 2021**

Project Name: Mizynte

Tech stack: AWS s3, lambda, ASG, NLB, ALB, EC2, Django framework, Python, MY SQL Database, ORM

- Created AWS cloud infrastructure from scratch for the application. Single tier application was converted into 3 tiers in order in integrate isolation and security to the application.
- Used Django framework for developing application and hosted it in AWS
- Created and configured elastic load balancers and auto scaling groups to distribute the traffic and to have a cost efficient, fault tolerant and highly available environment.
- Used AWS services like ec2, lambda for server-less computation, ASG, load balancer for handing load from the client.
- Working on automation for loading data from **S3** into **RDS**
- Experience in Writing AWS IAM policies
- Using Aws Sage maker for deploying ai training data from s3 to ec2 instance.
- Created snapshots to take backups of the volumes and also images to store launch configurations of the EC2 instances.
- Implemented domain name service (DNS) through route 53 to have highly available and scalable applications.
- Worked with Nginx, Gunicorn, created role-based permission access for employee and manager role
- Worked with Api and tried integrating Django application from angular. Angular was used as front-end where files were uploaded and saved in DB using Django.
- Worked on **smtp** for sending bulk mail to customer with salary slip
- Worked on Django data models for storing, retrieving, and manipulating into database.
- Worked on **JWT token**-based authentication and different types of class-based views
- Build single page application with route in **Angular**.
- Used Jenkins for automated daily manual tasks, deploying application on Aws cloud and testing the application

INTERN

Data Engineer at Strategy Soft, Hyderabad **Aug 2019 - Jan 2020**

Project Name: Unibotai.com

Tech stack: Python, MY SQL Database, Open data sources, AWS S3

Strategy Implemented Voice / Speech recognition modules

- Implemented voice recognition and speech-to-text modules
- Designed Python scripts for real-time web page monitoring and data extraction using Beautiful Soup.

EDUCATION

Bachelor of Engineering

- MIT (MH), Electronic and Telecommunication. 2019

CERTIFICATION

- Extraordinary certificate at Ernst & Young 2024
- AWS (Solutions Architect) training & certification 2020
- Python training & certification 2021