

Muhammad Hassan Haider

AI/LLM Engineer — Python Backend Engineer

☎ +923107380784 — ✉ itshassanbutt@gmail.com — 🗨 WeChat: hassanhaider4356 — 🌐 m-hassan-haider — 🏠 github.com/Hassan-Butt4356

Summary — Results-driven AI/LLM and Backend Engineer with 4+ years of experience building scalable backend systems and intelligent AI-powered applications. Hands-on expertise in Retrieval-Augmented Generation (RAG), agentic AI workflows, LLM fine-tuning (LoRA/DoRA), and local inference optimization (Ollama, vLLM, GGUF quantization). Skilled across the full stack — from multi-tenant Django/DRF microservices and FastAPI backends to production-grade RAG pipelines with ChromaDB and pgvector, LangChain/LlamaIndex orchestration, and real-time voice AI systems. Delivered systems reducing manual effort by 70–90% across ETL automation, payment integrations, and document intelligence. Open to AI/LLM Engineer, Generative AI Developer, and Backend+AI roles where scalability, privacy, and intelligent automation matter.

Skills

– Languages	– AI / LLM	– ML/NLP	– Cloud/DevOps
– Python	– RAG & Agentic AI	– TensorFlow / Keras	– AWS (S3, Lambda, Glue, Athena)
– SQL	– LLM Fine-Tuning (LoRA/DoRA)	– spaCy & FastText	– Docker & Docker Compose
– Frameworks	– Prompt Engineering	– Scikit-Learn	– Celery & Redis
– Django & DRF	– ChromaDB	– PEFT / Transformers	– Snowflake
– Flask	– vLLM / Ollama Inference	– Data Science	– Web & Scraping
– FastAPI	– Backend Tools	– Numpy & Pandas	– BeautifulSoup
– LangChain	– Django ORM	– Matplotlib & Seaborn	– Requests
– LlamaIndex & Ollama	– SQLAlchemy	– Streamlit	
– Tkinter & PyQT5	– Pydantic	– Testing	
		– Locust (Load Testing)	

Experience

QubitDance Chengdu China

June 2025 – January 2026

Backend Developer

- Developed and maintained microservices architecture with Django and DRF for scalable backend solutions
- Collaborated with frontend and DevOps teams to design RESTful APIs for seamless microservice communication
- Containerized Django apps using Docker for consistent and portable environments
- Implemented JWT authentication for secure API access and role-based access control
- Optimized database queries to enhance system performance and reduce latency
- Provided technical support for critical issues, collaborating with cross-functional teams to improve processes

Techmate Technologies

October 2022 – June 2025

Python Developer

- Automated ETL processes for large datasets using Python and Pandas
- Developed data processing pipelines integrated with AWS and Snowflake
- Created dynamic reporting tools (Excel/CSV) with metadata tracking
- Implemented payment gateways (Stripe, PayPal) for e-commerce solutions
- Developed Streamlit dashboards with real-time data synchronization
- Automated routine tasks using Python scripts to improve efficiency
- Implemented geospatial data validation using HereAPI and configured AWS S3 for secure cloud storage solutions

Code Embedded

November 2021 – October 2022

Python Django Developer

- Developed scalable Django web applications with RESTful APIs using DRF
- Implemented JWT authentication and role-based access control systems

Education

University of Gujrat (Hafiz Hayat Campus)

September 2016 – October 2020

BS in Software Engineering

Superior College (Qila Didar Singh)

September 2014 – May 2016

Projects

AgentForge – Multi-Tenant Agentic RAG Platform

Technologies: python, django, drf, llamaindex, ollama, chromadb, celery, redis, postgresql, jwt, docker

- Built an open-source multi-tenant platform where each user creates AI agents backed by their own private PDF knowledge bases, with full tenant isolation and JWT-secured REST API
- Implemented a LlamaIndex ReActAgent with QueryEngineTool, enabling the LLM to autonomously decide whether to retrieve from ChromaDB or answer from its own parametric knowledge
- Engineered incremental vector indexing and background PDF ingestion using Celery + Redis, ensuring non-blocking document processing and real-time agent readiness
- Designed a fully local LLM stack (Ollama + LLaMA3/Mistral) with zero cloud dependency for privacy-first deployments
- Containerized the full stack with Docker Compose (Django, Celery, Redis, ChromaDB, PostgreSQL) for one-command reproducibility

LLM Fine-Tuning Pipeline (LoRA / DoRA)

Technologies: python, peft, transformers, qwen3, gguf, ollama, pytorch

- Built an end-to-end fine-tuning pipeline for Qwen3-4B-Instruct on custom PDF datasets, covering data preparation, LoRA/DoRA training with PEFT, model merging, and GGUF export for local deployment via Ollama
- Implemented DoRA (Weight-Decomposed Low-Rank Adaptation) alongside standard LoRA, toggled via a config flag, to improve fine-tuned model accuracy on domain-specific tasks
- Engineered a PDF-to-training-data conversion script that chunks, cleans, and formats documents into instruction-following pairs for supervised fine-tuning
- Exported merged models to GGUF format and validated inference quality locally via Ollama, completing a full fine-tune-to-deployment loop without cloud GPU dependency

Real-Time Voice AI System (ASR → LLM → TTS)

Technologies: python, sensevoice, ollama, vllm, indextts, websockets, fastapi, docker

- Architected a production multi-service voice AI pipeline on a Compshare Tesla V100S (32GB) GPU server, integrating SenseVoice ASR, Ollama/Qwen3-4B for LLM inference, and IndexTTS 1.5 via vLLM for high-speed speech synthesis
- Built a FastAPI orchestration layer (port 8002) managing the full ASR → LLM → TTS flow with SQLite-backed conversation memory, coordinating three independently running tmux-managed microservices
- Implemented sentence-level response chunking with a TTSSequenceManager to stream ordered, low-latency audio back to clients, minimizing perceived response time in real-time voice conversations
- Converted IndexTTS 1.5 from ModelScope format to vLLM/HuggingFace format and tuned GPU memory utilization (25%) to run TTS alongside Ollama within the 32GB VRAM budget

X-NFT Microservice Platform

Technologies: python, django, django rest framework, postgresql, pydantic

- Developed a Django + Django Rest Framework microservice for NFT and Collection management
- Integrated with a Solidity-based blockchain microservice to handle NFT mint, buy, transfer, and burn operations
- Implemented multi-tenant architecture to support multiple organizations with separate NFTs and collections
- Built role-based access control with user, role, and permission management
- Optimized database queries and indexing to improve NFT listing and search performance
- Implemented advanced search, filter, and sorting for NFTs and Collections (similar to OpenSea)
- Used Pydantic for data validation and API request/response handling

PayUni – E-commerce Payment Microservice

Technologies: python, django, django rest framework, postgresql, jwt

- Developed a multi-tenant e-commerce microservice where merchants can list and manage products
- Implemented dynamic role management for merchants to create custom user roles
- Integrated crypto and fiat payment support using a separate payment microservice
- Designed and prepared secure payment payloads using Django and Django Rest Framework
- Implemented complete exception handling and error validation for payment workflows

APIFY – API Marketplace Platform

Technologies: python, django, django rest framework, alipay, wechat pay, postgresql, jwt

- Built an API marketplace platform similar to RapidAPI using Django and Django Rest Framework
- Developed API, versions, endpoints, and pricing management system
- Implemented demand creation feature with functional requirements and budget limits
- Enabled users to make offers on demands and purchase APIs directly from the marketplace
- Implemented group purchase feature for shared API subscriptions

- Integrated Alipay and WeChat Pay through an existing payment microservice

Automation for Transport Surveys

September 2023 – June 2025

Technologies: python, pandas, here api, aws s3

- Developed a comprehensive suite of Python scripts to automate ETL processes, including data extraction, munging, cleaning, and preparation for downstream analysis and reporting
- Designed and implemented advanced Pandas-based ETL pipelines for handling missing values, outliers, data inconsistencies, and format standardization across diverse datasets
- Integrated HERE Routing API for real-time route validation and geospatial enrichment, enhancing data accuracy and enabling reliable logistics/operational insights
- Reduced manual data cleaning and preparation effort by 70% through automated ETL validation scripts and reusable transformation modules
- Built dynamic file generation functionality to export customized, clean datasets as Excel and CSV files, supporting seamless data delivery and stakeholder consumption

ETC Dashboard

December 2024 – June 2025

Technologies: python, streamlit, snowflake, aws s3, aggrid, jwt

- Developed a secure Streamlit dashboard with JWT authentication and role-based access control (admin/user), serving as the frontend for real-time ETL-processed data visualization
- Designed and implemented ETL pipelines integrating Snowflake as the cloud data warehouse, enabling efficient processing and storage of daily transactional data with real-time AgGrid synchronization
- Built automated ETL workflows using Python to extract data from operational sources, transform it, and load/upload to AWS S3 buckets, reducing manual data handling by 90%
- Engineered interactive AgGrid tables with server-side pagination, filtering, and sorting, optimized for handling large-scale ETL-loaded datasets in Snowflake

AsifMolBio Lead Platform

November 2024 – Present

Technologies: python, django, django rest framework, postgresql, paypal, stripe, aws s3, smtp

- Developed Django and Django REST Framework (DRF) backend with JWT authentication
- Integrated dual payment gateways (PayPal and Stripe) for payment processing
- Configured AWS S3 storage handling of media assets with secure permission management
- Optimized PostgreSQL queries using Django ORM, reducing API response times by 40%
- Automated email workflows (SMTP) for payment confirmations and admin alerts
- Implemented comprehensive admin dashboard with full CRUD control over testimonials, services, courses, and promotional content

TaskFlowAI: Task Management Platform

June 2025 – Present

Technologies: python, django, postgresql, jwt, langchain, RAG

- Developed a Django-based backend for a task management system with modular app structure (accounts, project, task) to manage users, projects, tasks, and comments
- Implemented a custom user model with role-based access (Admin, Member), leveraging Django's authentication system and signals for automatic profile creation
- Optimized database performance using PostgreSQL with indexes on frequently queried fields and `select_related` for efficient ORM queries
- Configured role-based permissions to restrict task and project operations, ensuring admins manage projects while members interact with assigned tasks

M-Bizz (AI Marketing Bots)

July 2024 – November 2024

Technologies: python, django, langchain, openai, rag, pandas, firebase, postgresql

- Built conversational AI chatbots using LangChain and OpenAI GPT models
- Designed scalable FastAPI backend for bot integration with business workflows
- Implemented NLP pipelines (Python/Pandas) to preprocess customer data
- Implemented Retrieval-Augmented Generation (RAG) for domain-specific knowledge

Text Generation Using LSTMs

Self-Learning Project

Technologies: python, pandas, tensorflow, keras, lstm, nlp

- Developed an LSTM-based language model to generate news headline-style text using New York Times dataset
- Preprocessed 9,000+ article headlines with pandas and created tokenized N-gram sequences
- Built and trained a deep learning model with Embedding + LSTM + Dense layers using TensorFlow/Keras
- Implemented text generation function that predicts next words based on seed text

Flask NLP Chatbot Enhancement and Maintenance

May 2023 – August 2023

Technologies: python, flask, stripe, openai

- Debugged and resolved critical bugs in an existing Flask-based NLP chatbot, improving stability and ensuring reliable context loading from JSON files for accurate response generation

- Integrated Stripe payment processing for secure subscriptions and transactions
- Implemented multi-role authentication and authorization (e.g., admin/user roles), adding granular access control
- Added caching mechanisms to frequently accessed contexts and responses, significantly reducing latency

MD-EXAM Pharmacy System

April 2023 – July 2023

Technologies: [python](#), [django](#), [mongodb](#), [django](#), [shipstation api](#), [pymongo](#)

- Developed Django pharmacy management system using Django/PyMongo for MongoDB integration
- Implemented ShipStation API integration for automated label generation, reducing processing time
- Designed role-based access control: Super Admin (pharmacy creation/user registration) and Staff (order management)
- Built order tracking system with status flags, filters, and bulk operations for received orders
- Created custom Django management command for daily ShipStation API sync to update order statuses automatically

Certifications

Python Certificate

- Udemy - Python Complete Course For Python Beginners (Click to verify)
- Udemy - Learn to Code in Python 3: Programming beginner to advanced (Click to verify)

Python Django Certificate

- Udemy - Python And Django Framework For Beginners Complete Course (Click to verify)

MySQL Certificate

- Udemy - SQL Programming and MySQL Certificate (Click to verify)

Web Scraping (BeautifulSoup) Certificate

- Udemy - Web Scraping in Python using Beautiful Soup (Click to verify)