

SYED MAHAD EHSAN

Senior Software Engineer | Backend & AI/RAG Systems

Karachi, Pakistan · Open to remote (UTC+5) · +92 342 2170411 · mahad.ehsan@gmail.com
linkedin.com/in/syed-mahad-ehsan · mahadehsan.com

SUMMARY

Senior Backend Engineer with 5+ years building multi-tenant SaaS backends and, most recently, enterprise RAG systems. Designed and delivered a hybrid-retrieval knowledge platform using Microsoft Foundry and PostgreSQL/pgvector, and scaled async and realtime workloads with Django, Celery, Redis and SSE on AWS. Primary expertise is in Python (Django, FastAPI), with hands-on experience delivering ASP.NET Core services, and a strong focus on tenant isolation, reliability and test coverage.

TECHNICAL SKILLS

Languages: Python, JavaScript

Backend: Django, Django REST Framework, FastAPI, ASP.NET Core, EF Core, REST APIs, JWT / OAuth

AI / LLM: RAG, hybrid search (vector + full-text, RRF), agentic retrieval, query rewriting, contextual chunking, citation grounding, embeddings, HNSW, Docling / OCR ingestion, retrieval evaluation, Azure OpenAI, Microsoft Foundry (Azure AI Foundry)

Data: PostgreSQL, pgvector, MySQL (range partitioning), Redis, MongoDB, DynamoDB

Async & Realtime: Celery, Celery Beat, Server-Sent Events (SSE), Firebase Cloud Messaging

Cloud & DevOps: AWS (Elastic Beanstalk, S3, RDS, CloudWatch), Azure (VMs, Container Registry), Docker, Docker Compose, GitHub Actions, nginx, Gunicorn, DigitalOcean

Testing & Tools: pytest, Django TestCase, xUnit, Git, GitHub, GitLab, Bitbucket, Azure DevOps

Frontend: React, Vite

PROFESSIONAL EXPERIENCE

Diya Interactive

Senior Software Engineer

Software Engineer

Karachi, Pakistan

Jul 2024 - Present

Jun 2023 - Jul 2024

Enterprise RAG Knowledge Platform (industrial / energy sector) | ASP.NET Core, FastAPI, PostgreSQL/pgvector, Microsoft Foundry, Azure

- Led architecture and backend development of a multi-tenant RAG platform enabling engineers to query OEM manuals and IOMs through conversational search with page-level citations.
- Engineered hybrid retrieval: multi-turn query rewriting, an **agentic tool-using retrieval loop**, pgvector HNSW and Postgres full-text search fused with RRF, parent/child context expansion and structural citation verification, streamed to clients over SSE.
- Integrated **Microsoft Foundry (Azure OpenAI)** chat and text-embedding-3-large models behind provider-agnostic adapters for LLMs, embeddings, rerankers and document parsers.
- Built a fault-tolerant ingestion pipeline on Docling with OCR, contextual chunk enrichment, SHA-256 deduplication, embedding reuse and leased workers with retries; **ingested a 1,300+ page manual into 3,800+ searchable chunks**.
- Enforced tenant isolation with **23 permissions across Global, Organization and Unit scopes**, a unit filter on every retrieval query, and dedicated isolation tests in both the .NET and Python services.
- Shipped **91 API endpoints and ~680 automated tests**, a retrieval evaluation harness (hit@k, page accuracy, latency) with a CI recall-regression gate, and CI/CD via GitHub Actions, Azure Container Registry, Docker Compose and Azure VMs.

Security Workforce Management SaaS | Django/DRF, Celery, Redis, MySQL, AWS

- Core backend engineer on a multi-company platform for scheduling, attendance, leave, compliance and notifications, spanning 11 Django apps, 31 DRF ViewSets and **1,200+ automated tests**.
- Replaced in-process APScheduler with Celery Beat, **resolving a memory leak that was growing Gunicorn workers**; migrated 7 jobs and scaled the system to 12 periodic jobs and 24 async task types.
- Cut repeat bulk-schedule generation **from ~9s to 0.06s (thousands of queries down to 16)** through batched bulk_create and query reduction, and moved the schedule table to yearly MySQL range partitions.
- Built realtime delivery with Redis-backed SSE (Last-Event-ID replay) and Firebase push across 14 event types, with deduplicated notification fan-out and retries.
- Owned production AWS Elastic Beanstalk infrastructure across Gunicorn, nginx, Celery worker/Beat, Redis, deployment hooks and CloudWatch monitoring/logging.

Online Counselling Platform | Django/DRF, MySQL, Vonage OpenTok

- Rebuilt in-app chat as a versioned REST API (v2) with OpenTok session tokens, replacing the legacy WebSocket implementation.
- Migrated Django 2.2 → 4.2, DRF 3.15 and SimpleJWT upgrade, created the GitHub Actions test pipeline (MySQL 8, per-module test runs), and delivered production fixes across appointments, auth and video sessions.

KNYSYS, LLC

Associate Software Developer

Junior Software Developer

Karachi, Pakistan

Sep 2022 - May 2023

Mar 2021 - Sep 2022

- Built Python/Django backends and REST APIs supporting client business workflows.
- Implemented asynchronous processing, scheduled jobs and background tasks with Celery and Redis.
- Worked on database integrations and deployments across AWS and Azure environments.

Bitrupt

Full-Stack Developer Intern

Feb 2020 - May 2020

- Built React and React Native applications with Firebase and SQLite, from implementation to deployment.

EDUCATION

BS Computer Science, DHA Suffa University

CGPA 3.5

2016 - 2021