



Ramprasad Mondal

AI Engineer | Immediate Joiner

Software Engineer with 6+ years of experience, now specializing in **Generative AI applications, multi-agent systems, and knowledge graph pipelines**. Strong foundation in **scalable distributed backends in Python** across fintech and AI domains, with hands-on AI engineering experience delivering **bilingual AI platforms**. **Founding-engineer background** with a bias for **shipping fast** and iterating in production, combining deep backend fundamentals with **cross-functional leadership** and **production-grade, security-conscious system design**, delivering to **enterprise and government clients** across **UAE and Indian markets**.

+971 562388292 Abu Dhabi/Dubai, UAE m.ramoffice@gmail.com linkedin.com/in/ramprasad-mondal github.com/ramprasad-mondal

Skills

Languages	Python, JavaScript
AI Stack	Generative AI, Agentic AI, GraphRAG, LangChain, LangGraph, LangSmith, LLM
Databases	Neo4j, Qdrant, PostgreSQL, MySQL, Redis, MongoDB, MinIO, ClickHouse
Frameworks	FastAPI, Django, Flask, Hono, Bun, Celery, OpenTelemetry
APIs & Protocols	REST, GraphQL, gRPC, OpenAPI
Cloud & Infra	AWS, Azure, GCP, Docker, Kubernetes, vLLM, Air-Gapped Deployment
Auth & DevOps	Keycloak, Zod, OpenAPIHono, Git, Temporal, Airflow, Prometheus
Architecture	Multi-Agent Systems, Knowledge Graphs, Microservices, Event-Driven, Distributed Systems
Core Skills	Problem Solving, Feature Development, Cross-functional Collaboration, Ownership

Experience

OpenInnovation AI – AI Platform

Dec 2025 – Jul 2026 (8 months)

Domain: Multi-Agent Systems, GraphRAG, Air-Gapped Enterprise LLM Deployment

AI Engineer

Abu Dhabi, UAE

- Architected a **multi-tenant, multi-agent AI platform** where each agent maintains an isolated **Neo4j-backed knowledge graph** and **GraphRAG retrieval pipeline**, implemented dynamic query routing across agents using a knowledge map and **active-threats dashboard** for real-time situational awareness.
- Built **knowledge ingestion pipeline** with **semantic chunking**, entity extraction, and **hybrid RAG** (dense vector + graph traversal) – improved **response quality by 50%+** and **overall processing efficiency by 50–60%**. Used async **FastAPI** + **MinIO** for scalable, memory-efficient document processing.
- Engineered **multi-agent orchestration** with structured meeting rooms for collaborative agent reasoning. Build a parallel **AI report generation pipeline** synthesising graph context and LLM output into **stakeholder-ready deliverables** in English and Arabic.
- Delivered **full English–Arabic bilingual platform** – backoffice service for static content translation and **LLM-powered dynamic translation** across all agent responses, reports, and documents with **RTL support**.
- Developed **multi-modal AI document extraction** microservice supporting scanned PDFs, handwritten notes (Arabic + English), and images – enabling high-accuracy knowledge ingestion across **heterogeneous enterprise document types**.
- Led **air-gapped deployment** onto **NVIDIA DGX Spark, ASUS GX10, and H200** servers – **zero cloud dependency**, fully offline Docker-based installation. Engaged directly with **UAE government ministries and private sector clients** for requirements gathering, on-site delivery, and post-deployment support.
- Deployed **production LLMs up to 35B parameters** (including MoE) using **vLLM on Kubernetes** with GPU – worked with **tensor parallelism** and continuous batching to enable enterprise-grade inference within air-gapped environments.
- Contributed to a **TypeScript API** built on **Hono/Bun** with Zod validation and auto-generated OpenAPI docs, developing feature tickets within the platform's **Keycloak**-based JWT auth and **RBAC** structure.

Buildwealth Technologies – Fintech

Jul 2020 – Nov 2025 (5 years 5 months)

Domain: Fintech Backend, Multilingual RAG, Partner & Revenue Platforms

Senior Software Engineer II (1y 4m) | Senior Software Engineer I (1y 9m) | Software Engineer (2y 4m)

- Built a **multilingual conversational AI assistant** with retrieval-augmented responses, intent classification, and entity extraction across 8+ Indian languages, helping partners discover products, create orders, and access personalized recommendations.

- Developed a **RAG-based Q&A service** using embeddings and vector retrieval, enabling users to query uploaded documents and receive contextual, grounded responses.
- Led a **cross-functional team of 6** to design, deliver, and operate a production partner-payout system, coordinating backend engineering with product, finance, and operations stakeholders.
- Designed and scaled a **Python/Django fintech backend** on AWS ECS, using Redis caching and Celery/SQS background processing to support peak traffic of approximately **10K requests per minute** for **250K+ registered and 20K+ daily active partners**.
- Reengineered a production **daily revenue pipeline** to asynchronously retrieve NAVs, commission rates, and available units before performing computation with Dask and Pandas on a dedicated worker, reducing end-to-end processing time for **2M transaction records by over 99%**, from 10–12 hours across 6 workers to 3–4 minutes.
- Built a secure **authentication gateway and login workflows** using OAuth, OTP, Redis-backed token and session management, authorization-code exchange, and request-signature validation, reliably processing **50K+ authenticated requests daily**.
- Automated and optimized a production **invoice-generation workflow** using Celery workers, Airflow orchestration, bulk database operations, and caching, reducing processing time by **94%**, from 4 hours to 15 minutes, for **10K+ monthly partner invoices** with GST, TDS, ARN, and PAN requirements.
- Implemented centralized **observability** using request-ID tracing, structured logging, CloudWatch, GCP Logging, ClickHouse, Grafana dashboards, and automated Slack alerts, accelerating root-cause analysis across production APIs, services, and databases.
- Developed **payment and payout infrastructure** integrating Cashfree and Razorpay, with batch transfers, partial payments, retry handling, status reconciliation, and failure recovery to improve the reliability of partner disbursements and order settlement.
- Designed **ETL pipelines** for CAMS/KARVY data reconciliation–improved revenue accuracy by **15%** through **analytical** processing.
- Integrated **Digio KYC** and secure webhook workflows, automating verification-status tracking, callback processing, and exception handling to streamline partner onboarding.
- Designed modular Python services and REST/GraphQL APIs for a **multi-product fintech platform** spanning mutual funds, unlisted stocks, MLDs, PMS, loans, and credit cards, supporting shared onboarding, payments, and reporting workflows.

MountBlue Technologies – IT Services

Feb 2020 – Jun 2020 (5 months)

Domain: Full-Stack Web Apps (Internship)

Software Engineer Intern

Bengaluru, India

- Built **COVID Q&A platform** with user auth, comments, and real-time dashboards.
- Developed e-commerce platform for a local shop demonstrating **problem solving** and **feature development** skills.

Education & Achievements

- **B.Tech, Techno India Saltlake** – Electrical Engg. – 7.7 CGPA
- **Diploma, Durgapur Polytechnic** – Electrical Engg. – 8.0 CGPA
- **Best Employee Award (2025)** - Buildwealth
- **Best Team Award (2025)** - Buildwealth

Jun 2016 – Jul 2019

Jun 2013 – May 2016