

Yash Malhotra

Backend + AI Systems Engineer | Full-Stack Product Engineering

Cleveland, OH | +1-216-224-4691 | yxm595@case.edu

LinkedIn | GitHub | Portfolio

Summary

Software/AI engineer with enterprise backend experience, AI product development, and graduate systems research. Builds Python/FastAPI APIs, data models, RAG and LLM workflows, and end-to-end product features while translating ambiguous business requirements into maintainable technical systems.

Technical Skills

Languages: Python, Go, Java, SQL, TypeScript, JavaScript, C/C++ **Backend and APIs:** FastAPI, Django, REST APIs, SQLAlchemy, Pydantic, authentication, authorization, RBAC, domain modeling

Frontend: React, Next.js, Tailwind CSS, HTML/CSS **AI/ML:** RAG, Graph RAG, LangGraph, LLM APIs, vector retrieval, multi-agent workflows, structured extraction, PyTorch

Data and DevOps: PostgreSQL, Cassandra, Redshift, ETL pipelines, AWS, Docker, GitHub Actions, CI/CD, Linux, Git

Experience

Senior AI Engineer/Researcher

Sep 2025 – Apr 2026

Decoupling.co

Remote

- Designed modular backend architecture for reusable AI products using Python, FastAPI, and Go, defining domain models, REST API contracts, provider abstractions, authentication, tenant boundaries, usage controls, and stateful workflows.
- Built RAG, Graph RAG, and LangGraph multi-agent pipelines combining document ingestion, metadata extraction, vector retrieval, structured outputs, evidence-aware generation, and human-review stages.
- Structured reusable services for document intelligence, customer-review analysis, synthetic-data generation, and AI artifact creation, allowing multiple workflows to share common backend capabilities.
- Worked directly with founders, product contributors, frontend engineers, and analytics teams to convert ambiguous requirements into technical roadmaps, data contracts, architecture decisions, and end-to-end product features.

Graduate Researcher

Jan 2024 – May 2026

Case Western Reserve University

Cleveland, OH

- Developed distributed AI and scientific-computing systems using Python, PyTorch, C++/CUDA, MPI4Py, and SLURM, including a multi-GPU Fourier backend that extended execution beyond single-GPU memory limits.
- Designed correctness, automatic-differentiation, backend-execution, capacity, and performance tests for complex GPU and distributed-computing workflows.

Adjunct Faculty

Oct 2025 – Dec 2025

Weatherhead School of Management, Case Western Reserve University

Cleveland, OH

- Taught graduate-level machine learning and AI topics, including PyTorch, deep learning, recommendation systems, transformers, and large language models, to technical and business-focused students.
- Designed lectures, programming exercises, and applied projects while explaining architecture decisions, model behavior, and engineering tradeoffs to audiences with varied technical backgrounds.

Senior Software Engineer

Aug 2021 – Jan 2024

Cognizant Technology Solutions – Johnson & Johnson MedTech

India

- Built enterprise backend APIs and data-processing workflows using Python, FastAPI, SQLAlchemy, Cassandra, Redshift, AWS, Spark, and Airflow; processed **8.5 million records in under 30 minutes** for global analytics.
- Designed reusable connectors and service layers across Redshift, PostgreSQL, MySQL, Oracle, AWS Glue, and Athena, standardizing database access, validation, and reconciliation across heterogeneous enterprise systems.
- Built an automated migration-validation framework that compared source and target systems, generated reconciliation reports, and reduced testing time from **more than 40 days to approximately 1.5 hours** while lowering production discrepancies to **below 2%**.
- Translated business requirements into REST APIs, data models, workflow automation, and maintainable backend services; reviewed designs, investigated production issues, and mentored more than five engineers.

Selected Projects

LLM-Based RAG and Graph RAG Systems

Python / FastAPI / LangGraph / LLM APIs

- Built source-aware retrieval workflows over documents and structured data with ingestion, metadata extraction, vector retrieval, schema-validated outputs, evidence traceability, and human-review stages.

Multi-Agent business Artifact Generation Platform

Python / FastAPI / Go / LangGraph

- Designed orchestrated agents that transform business context and analytical outputs into structured reports, presentations, dashboards, and follow-up actions through typed contracts and validation gates.

Education

Case Western Reserve University — M.S. Computer Science, Thesis Track, **GPA: 4.00/4.00**

May 2026

Birla Institute of Technology, Mesra — B.E. Electronics and Communication Engineering

May 2021