Gaurav Ghop

New York, United States | +1 929-628-4065 | gauravghopwork@gmail.com | **GitHub** | **Website** | **LinkedIn**

TECHNICAL SKILLS

Languages: Java, Python, TypeScript, JavaScript, C++, SQL

AI & LLM Systems: Agent orchestration, RAG, LLM guardrails, LangGraph, Langfuse, LoRA, local inference (MLX, Ollama), LangChain

Backend & APIs: Spring Boot, Node.js, Next.js, React Native, FastAPI, JUnit, REST, Webhooks, Microservices Architecture, OAuth

Databases & Storage: PostgreSQL (PgBouncer, PostGIS), LSM-trees, Cassandra, Bloom filters, Redis, InfluxDB, MongoDB

Cloud & Infrastructure: AWS (SQS, S3, EC2), Docker, Kubernetes, Kafka, Flink, GitHub Actions, CI/CD, Sentry, Grafana

EXPERIENCE

OurFreedom.ai Jun 2026 – Present

Software Engineer, Founding Team *New York, NY*

- Led development of our agent layer Carla, a LangGraph tool-calling agent across 6 tools, with Pydantic-validated tool outputs, retry and fallback handling, and a human review queue for flagged outputs, serving about 1,000 requests a day.
- Designed the agent's eval pipeline, built a 150-case gold set run nightly with Langfuse tracing, root-causing regressions to specific tool calls or prompts, cutting hallucinations from 16% to under 5% and raising tool-selection accuracy to 91%.
- Owned the AWS production backend, handling resource management, uptime, and rollbacks, and shipping GTM MVP features with the founders.
- Added Prometheus and Grafana monitoring for p95 latency, tool failure rate, and cost per request, with canary rollouts for every change.

Tiny Archives Jun 2025 – Aug 2025

AI Software Engineer, Summer Intern *New York, NY*

- Shipped a Python ingestion pipeline that normalized archival exports (EAD/XML, CSV, PDF) into a unified Pydantic schema, replacing 3 parsers.
- Built a RAG pipeline with LangChain and pgvector that grounded answers in 40,000 indexed documents and cut hallucination rate from 18 percent to 4 percent as measured by a weekly eval set of 500 queries.

Duende Ingenious Jul 2023 – Aug 2024

Founding Engineer *Pune, India*

- Introduced PgBouncer connection pooling with PostgreSQL, reducing peak connection count by 85% and eliminating exhaustion related incidents.
- Built an asynchronous order processing queue in Node.js on AWS SQS with exponential backoff retries and a dead letter queue, cutting silent order failures from 0.5% to 0.3% of daily volume
- Containerized backend services with Docker, automating CI/CD through GitHub Actions with integration tests, cutting release cycle from 2 weeks to 1.

eQ Technologic Aug 2021 – Jun 2023

Software Engineer I *Pune, India*

- Built idempotent Kafka consumers with automated reconciliation across 40M+ part records, detecting and resolving 1,200+ revision mismatches per month between PLM and ERP systems.
- Developed Java Spring Boot microservices for a low code API Gateway, adding retry logic, circuit breakers, and JUnit integration tests to REST, OData and SOAP adapters that reduced failed SAP and Teamcenter sync jobs by 40%.
- Built a Kubernetes operator to provision multi-tenant connector deployments from CRDs, eliminating 200+ hand maintained manifests.
- Integrated SAML SSO, Kerberos, LDAP, OAuth, and X.509 certificate authentication against customer identity providers for US defense deployments.

PROJECTS

Formulary Document Ingestion & Fine-Tuning Pipeline | *Python, MLX, LoRA, Docling, PostgreSQL* | **LINK**

- Built a PDF-to-model pipeline parsing formulary documents into overlapping page chunks, extracted into a prompt version-keyed store.
- LoRA fine-tuned Llama 3.1 8B with MLX and served the adapter on-device behind a chat UI, keeping inference fully local at zero per-token cost.

University Advisor Chatbot | *Python, FastAPI, Next.js, AWS Bedrock* | **LINK**

- Developed an agentic chatbot delivering source-cited university recommendations to international students, orchestrating Claude Sonnet through a tool-calling loop across 5 custom tools on a FastAPI backend with a Next.js chat UI.
- Sold and integrated the chatbot into an Indian study-abroad consultancy's web app serving 500+ students weekly.

Dispatch Engine | *Java, Spring Boot, PostGIS, Redis, Kafka* | **LINK**

- Engineered a delivery dispatch system for a food catering business that matches orders to nearby drivers using Redis GEO lookups with a PostGIS fallback, re-queuing failed dispatches through retry-safe state machines and publishing every state change to Kafka for live-order tracking.
- Shipped as freelance work and handed off to the client. Handles 300 orders/day and cut delivery-status inquiries by 90%.

EDUCATION

New York University May 2026

Master of Science in Computer Engineering *New York, NY*

Relevant coursework: Distributed Systems, High-Performance Machine Learning, Design & Analysis of Algorithms.

Vishwakarma Institute of Information Technology May 2021

Bachelor of Engineering in Electronics & Telecommunications Engineering *Pune, India*

ACHIEVEMENTS

Merit Scholarship for graduate tuition, NYU • **Runner-Up** AWS GenAI Hackathon 2025 • **Finalist** Google DeepMind Hackathon 2025 • **Runner-Up** Smart India Hackathon 2020 • Google KickStart Round B 2020 • ACM ICPC Regionals 2019 • CodeChef 4★