YASSINE KRAIEM

New York, NY | +1 (808) 570-8151 | yassinekraiem08@gmail.com | linkedin.com/in/yassinekraiem | github.com/Yassinekraiem08

EDUCATION

Columbia University, New York *Aug 2026 — May 2028*
M.S. in Artificial Intelligence

Alma College, Alma *Aug 2022 — Dec 2025*
B.S. in Computer Science *GPA: 3.88, Summa Cum Laude*

EXPERIENCE

RecipeOne *Jul 2025 — Jan 2026*
Founder & AI Software Engineer *Alma, MI*

- Built a multi-tenant B2B SaaS platform for institutional foodservice ops (recipe planning, procurement, inventory across multi-site facilities), with PostgreSQL backend modeling 15+ entities and 30+ relational tables.
- Implemented automated unit conversion and yield computation pipelines, processing 7,000+ ingredient records to generate purchasing quantities and cost estimates across inventory systems.

ListFlowAI - https://listflowai.com *Oct 2023 — Apr 2025*
Co-Founder & AI Software Engineer *Remote - Washington, D.C.*

- Co-founded an AI-powered B2B SaaS automation platform used by 200+ users, processing 5,000+ inbound leads through LLM-driven enrichment, classification, and routing workflows.
- Architected backend services and data pipelines using Python, FastAPI, and PostgreSQL, extracting structured data from emails, forms, and CRM exports and reducing manual lead processing by 80%.

Eat Pro Japan *Jun 2024 — Aug 2024*
Software Engineering Intern *Tokyo, Japan*

- Shipped features for the booking and reservation system (Node.js, Next.js, MySQL on AWS) on a 20-engineer team, hardening the reservation flow against availability conflicts as the platform scaled from 5K to 20K users.
- Built an internal admin tool for the operations team to manage restaurant availability and resolve booking conflicts in real time, replacing a manual spreadsheet-based workflow and eliminating a recurring class of user-facing reservation failures.

PROJECTS

AI Decision Support System (Retrieval-Augmented Generation) *Oct 2025 — Feb 2026*

- Built a production RAG system using Python, FastAPI, and pgvector, enabling hybrid retrieval across 1,600+ document chunks and reducing hallucinated responses from 80% to 0% through citation-grounded generation.
- Designed parallel reranking and semantic caching pipeline, cutting p95 latency by 76%; conducted confidence calibration study (ECE = 0.256) and documented 9 failure modes via CI/CD evals.

Natural Language Analytics System (Text-to-SQL) | querymind-demo.vercel.app *Nov 2025 — Mar 2026*

- Built a full-stack NL-to-SQL analytics system (FastAPI, React, SSE streaming), adapting ideas from DIN-SQL and CHESS for execution-guided refinement and schema-aware generation over complex relational schemas.
- Engineered a self-correcting SQL pipeline with up to 3 retry attempts, semantic query cache (cosine similarity ≥ 0.92, 0 LLM calls on hits), Spider-style execution-accuracy benchmark with hardness classification, and 212 automated tests across 15 modules.

AI Workflow Orchestrator (Agents + Tool Calling) | https://ai-workflow-orchestrator.vercel.app *Nov 2025 — Mar 2026*

- Built a multi-agent AI orchestration system for incident triage across logs, tickets, and emails using a classify → plan → execute → replan loop with human-in-the-loop escalation; deployed on AWS ECS Fargate with observability via Jaeger, Prometheus, Grafana, and CI/CD through GitHub Actions OIDC.
- Designed fault-tolerant execution with Celery workers, retries with exponential backoff, and dead-letter queues; implemented multi-model routing and LLM-as-judge evaluation (5 metrics), achieving 95% task success vs 68% baseline and reducing per-task cost by 21%.

SKILLS

Languages: Python, C++, Java, Go, Swift, TypeScript, SQL
Systems: Distributed Systems, Concurrency, Data Structures & Algorithms, System Design
Backend: gRPC, Protocol Buffers, REST APIs, Microservices, FastAPI
Infrastructure: Linux, Docker, Kubernetes, AWS (ECS, Lambda, S3), CI/CD (GitHub Actions)
Databases: PostgreSQL, Redis, Vector Search
AI/ML: LLM Systems, RAG, Agent Orchestration, PyTorch, Core ML, Model Evaluation