Upgraid

MLOps Engineer

Location: flexible (US time zones), some preference for Boston
Type: full-time, early technical hire

Why Upgraid exists

Buildings are the world's largest asset class, consume ~40% of energy globally (and generate the same share of greenhouse gas emissions), and shape the way we live, work, play, and interact. They are foundational to human societies. There are billions of them, from skyscrapers to data centers, malls, warehouses, and single-family homes.

Hundreds of millions - perhaps billions - of these buildings would benefit from upgrades. These upgrades would reduce energy costs, improve health, and create more attractive spaces for residents, consumers, students, patients, and more. But the way building upgrades are done today is archaic. Physical inspections, owners with no understanding of the systems in their buildings, and expensive manual energy audits of variable quality make the old way of doing things untenable.

Instead, imagine if every building could tell you exactly how it wants to be upgraded — what to fix, what it would save, and how fast it pays back. That's what we're building. Our AI model reads the built environment from space, runs advanced energy simulations, and delivers a ready-to-pitch upgrade proposal for every property. It's how we will accelerate building upgrades globally.

Who we are and where we're at

Our experienced founding team includes a former McKinsey partner and leader of built environment sustainability, an experienced product leader, and an MIT building scientist. We have rapidly closed our funding round, have advisors who have built companies from zero to IPO and senior leaders from the industry. We have recently been accepted to Greentown Labs, the world's leading climate tech incubator. We have paying customers who consider our product a quantum leap in how building upgrades are done. We are going places fast and would like incredibly bright and talented people to join us.

What you'll help build

- **Global Data Plane & Model Registry**
    - Design a centralized data lakehouse and API schema that is stable, versioned, and strictly typed.
    - Establish a multi-tenant data architecture with clear governance and isolation.
    - Implement a model registry to manage global model versions, artifacts, and lineage.
- **Federated Multi-Tenant Engine**
    - Architect a cost-efficient serving layer to support client-specific customization.
    - Implement dynamic serving to hot-swap client-specific adapters (PEFT/LoRA) on top of a global base model at runtime.
    - Ensure strict data isolation and privacy boundaries without maintaining separate infrastructure stacks for every tenant.
- **Async Job Orchestration**
    - Build the queuing architecture required for portfolio-scale analysis workloads.
    - Design the asynchronous POST /jobs API to manage long-running inference tasks and state management.
    - Implement robust failure handling (retries, dead letter queues) and event-driven notifications (webhooks/email).

Day-to-day (Your First 90 Days)

- **Month 1: Foundation & Infrastructure**
  - Establish the core cloud environment using Infrastructure as Code (Terraform/Pulumi).
  - Configure VPCs, IAM roles, secrets management, and secure CI/CD pipelines.
  - Set up basic observability (logs, metrics) and deploy a "Hello World" service securely.
- **Month 2: The Data & API Backbone**
  - Define the initial database schema (PostgreSQL) with strict tenant isolation logic.
  - Build the skeleton of the Async API (FastAPI).
  - Set up the message queue infrastructure (SQS/Kafka) to handle a basic job flow.
- **Month 3: The MVP Loop**
  - Deploy the global base model to a production inference endpoint.
  - Implement a v1 client feedback loop: ingestion of feedback data → storage → manual trigger of a fine-tuning job.
  - Deliver a working end-to-end flow where a user can submit a job and receive results.

In 6 months, success looks like

- We can onboard a new client without code changes, configure their local adapter, and push them into production in days, not months.
- Models are versioned, reproducible, and observable; you can compare global vs. local performance at a glance and roll back safely.
- Engineering velocity is high because infra is predictable, typed, and automated.

The kind of problems you'll enjoy

- Multi-tenant global → local model architectures (shared schemas, tenant overrides, RBAC).
- Geospatial pipelines and indexing for at-scale queries.
- Asynchronous job orchestration, fan-out/fan-in, idempotency, and cost-aware scaling.
- Turning fuzzy real-world building data into opinionated, API-ready insights that sellers can act on.

You might have done some of this

- Designed data platforms with Postgres/PostGIS, BigQuery/Snowflake, object storage, and a feature store; comfortable with schema evolution and backfills.
- Ran ML ops in production (model registry, evals, canary deploys, drift detection, feedback loops).
- Used tools like LoRA/Adapters or dynamic model loading
- Built job queues (Celery, BullMQ), worked with message brokers (Kafka/SQS), and handled state management for long-running processes
- Built secure APIs (OAuth2, JWT), async job APIs (status endpoints, webhooks) with strong documentation and versioning
- Built/Are familiar with architect network isolation (VPCs, PrivateLinks) to guarantee that Client A's data is physically or logically impossible for Client B to access.

Our tech stack is python based, AWS, pulumi, FastAPI, Amazon SQS, Coiled (Task),

What makes this role special

- Early team ownership: you'll set the architecture that powers our roadmap — from AI-powered audits to valuation-grade signals and marketplace rails.

- Greenfield + real traction: we’ve proven the MVP and have paying customers - now we’re scaling it to thousands of buildings and many customers.
- Surface area that matters: every pipeline you harden directly improves response rates, meeting quality, and conversion for our customers.
- A team that’s built to win: you’ll join an experienced, tightly aligned founding team — a former McKinsey partner who led the global built environment sustainability practice, a technical founder with deep mechanical and energy systems expertise, and an experienced operator and product leader driving growth and partnerships. It’s a mix of strategy, engineering, and execution that makes this the perfect moment to join.

What we value

- Enjoy cracking hard problems. The messier the data or the tougher the constraint, the better.
- Passion for decarbonization. Our aspiration is to visibly bend the curve down on global emissions from the built environment — and have fun doing it.
- Pragmatic rigor. Measure, ship, iterate.
- Security by default. Data trust is non-negotiable.
- Low-ego collaboration. Teach, learn, write things down.

Nice-to-haves

- Experience with geospatial data at scale (tiling, projections, spatial joins).
- Prior work on optimization or simulation pipelines.
- Building email, notification and other delivery pipelines (deliverability, templating, auditing).

Interview process

- 30-min intro (mutual fit + product deep dive).
- 60-min practical tech interview
- 60-min personal & team fit interview.
- CEO interview

Compensation

- Competitive salary + potential for equity (early-engineer level).
- Flexible location/hours, reasonable time off, and the tools you need.

Send your resume and a short note saying why you’re interested to:

Daniel Cramer – daniel@upgraid.us and Brodie Boland at brodie@upgraid.us