Dmytro Kovalchuk

San Francisco Bay Area | kdmytro@gmail.com | 267-243-6871 linkedin.com/in/kdmytro | github.com/KDmytro | kdmytro.com

Summary

8 years building AI systems — from model integration and evaluation to production infrastructure. At Parcha, built the AI platform from scratch: LLM orchestration, evaluation pipelines, continuous model monitoring, plus the infra underneath (K8s, Terraform, Prometheus/Grafana). Platform serves enterprise fintech clients with sub-second inference across 50+ data sources.

Skills

- **Infrastructure**: Kubernetes, Docker, Terraform, GCP, AWS, Prometheus, Grafana, GitHub Actions
- **Backend**: Python (async), REST APIs, gRPC, PostgreSQL, BigQuery, Redis, Celery, FAISS, pgvector
- **AI/ML Systems**: LLM orchestration, model deployment, inference pipelines, RAG, embeddings
- **Practices**: SOC 2 compliance, observability, rate limiting, multi-tenant architecture

Experience

Parcha — Founding AI Engineer

Sep 2023 – Present | San Francisco, CA

First engineering hire. Built AI inference platform from zero to production serving enterprise clients.

Platform & Infrastructure

- Replaced external DevOps; own full infrastructure stack: Terraform, GCP, Kubernetes, secret management, CI/CD
- Built rate-aware control plane for million-record migrations — balances K8s pod scaling against LLM API limits (RPM/TPM) with automatic backpressure
- Implemented Prometheus/Grafana observability stack; created dashboards for latency percentiles, error rates, and resource utilization
- Designed multi-tenant architecture serving concurrent enterprise clients with isolated workloads
- Implemented technical controls for SOC 2 Type II certification — security controls, access management, audit logging

API & Service Design

- Architected core APIs for 50+ data source integrations (OpenCorporates, SEC, courts, sanctions lists across 200+ jurisdictions)
- Built deployment automation with configurable depth modes: FAST (single-agent, sub-second), DEEP (expanded tooling), ULTRA-DEEP (parallel sub-agent orchestration)
- Implemented queuing and caching layers for high-throughput batch processing

AI Platform

- Core architect of **Grep.ai** — ranked **#1 on DeepResearch Bench** (54.37), achieving **99.7% accuracy**
- Designed multi-agent orchestration system with 300+ skills across 16 industries
- Built "Capture First, Measure Later" telemetry pipeline for production model monitoring

---

Carvana — Team Lead, Machine Learning

Mar 2022 – Aug 2023 | Remote

Led 4-person ML team. Early LLM adopter pre-GPT-4.

- Led cross-functional initiative to deploy LLMs into production — document understanding, knowledge base Q&A, customer-facing chatbot with custom fine-tuned models
- Built gRPC services for chatbot inference; message processing pipeline in Scala
- Fine-tuned LayoutLM and DocumentAI models for document understanding; processed vehicle titles, registration, and financing paperwork at scale
- Promoted from Sr. ML Engineer to Team Lead within first year

---

Augment CXM — Senior Machine Learning Engineer

Aug 2017 – Jun 2022 | San Francisco, CA

Core ML team building conversational AI platform.

- Deployed FAISS vector database for semantic search over millions of conversational records
- Developed patented LSTM model for next utterance prediction — core product IP
- Shipped production NLP models: topic analysis, clustering, sentiment, NER
- Established Airflow as primary MLOps platform for model training and deployment pipelines

Previously: Core Algorithms and Data Science (Aug 2017 – Aug 2019)

Earlier Experience

- **Cornerstone OnDemand** — CSM/Business Analyst (2014–2017)
- **The Pep Boys** — System Analyst / Online Training Developer (2005–2013)

Education

- **Galvanize** — Data Science Immersive
- **Lakeland College, WI** — Computer Science, Economics
- **ICU University, Kyiv** — Management of Information Systems