

# Dmytro Kovalchuk

---

San Francisco Bay Area | kdmytro@gmail.com | 267-243-6871 | [linkedin.com/in/kdmytro](https://linkedin.com/in/kdmytro) | [github.com/KDmytro](https://github.com/KDmytro) | [kdmytro.com](https://kdmytro.com)

---

## Summary

8 years building AI systems — from model integration and evaluation to production infrastructure. At Parcha, built the AI platform from scratch: autonomous multi-agent systems (Grep.ai), evaluation pipelines, continuous model monitoring, plus the infra underneath (K8s, Terraform, Prometheus/Grafana). Platform serves enterprise fintech clients with high requirements for response time, reliability, and regulatory compliance.

---

## Skills

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

## Experience

### Parcha — Founding AI Engineer

Sep 2023 – Present | San Francisco, CA

First engineering hire. Technical #2 behind CTO. Built the AI platform from scratch.

#### Product & Architecture

- Core architect of **Grep.ai** — ranked **#1 on DeepResearch Bench** (54.37), achieving **99.7% accuracy** on PhD-level research tasks
- Early adopter of **Anthropic Agent SDK**; built MCP integrations for 50+ data sources (OpenCorporates, sanctions lists, adverse media, SEC, patents, courts across 200+ jurisdictions)
- Designed multi-agent system with 300+ skills across 16 industries; depth modes from 10s to 15min replacing days of manual research

#### Evals & Model Governance

- Created regulatory-grade documentation for enterprise clients — dataset sourcing, evaluation methodology, assumptions, limitations — used for compliance and regulatory approval
- Built “Capture First, Measure Later” observability pipeline; integrated Braintrust and Scorecard for trace-based evals and LLM-as-judge monitoring on production samples
- Identified false negatives during enterprise pilot; validated fix on 100 prod cases (Low Risk: 17%→11%, High Risk: 12%→18%), then productized into self-service “Evaluate Criteria Changes” UI for compliance teams

#### Client-Facing

- Joined CEO on sales calls to explain model governance and AI reliability to enterprise prospects
- Supported client compliance teams through rollouts, post-go-live tuning, and ongoing case reviews

## Infrastructure

- Replaced external DevOps consultant; owned Terraform, GCP, Kubernetes, Prometheus/Grafana, BigQuery, GitHub Actions, secret management
- Orchestrated million-record migrations with rate-aware control plane, manually balancing K8s scaling against LLM token limits (RPM/TPM)
- Implemented technical controls for SOC 2 Type II certification — security controls, access management, audit logging

## Team & Culture

- Early Claude Code adopter (April 2025); developed feature.md planning workflow, converted team from Cursor through organic standup demos
- 

## 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 products.

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

*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