

# Dmytro Kovalchuk

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

---

## Summary

8 years building AI platforms — from infrastructure and API architecture to multi-agent orchestration at enterprise scale. At Parcha, built the entire platform from scratch as founding engineer: multi-tenant Kubernetes infrastructure, agent orchestration APIs, evaluation pipelines, and security controls for SOC 2 certification. Led technical decisions and worked directly with enterprise clients on integration and rollout. The system I architected — Grep.ai — is ranked **#1 on DeepResearch Bench** with **99.7% accuracy**, running 300+ agent skills across 16 industries on Claude.

---

## Skills

- **Infrastructure:** Kubernetes, Docker, Terraform, GCP, AWS, Prometheus, Grafana, GitHub Actions, secret management
  - **Platform & APIs:** Multi-tenant architecture, API design, workflow orchestration, rate limiting, state management, CI/CD
  - **Backend:** Python (async), REST APIs, gRPC, FastAPI, PostgreSQL, BigQuery, Redis, Celery
  - **AI/Agents:** Multi-agent orchestration, MCP tools, Claude/Anthropic Agent SDK, RAG, FAISS, pgvector
  - **Security & Compliance:** SOC 2 Type II, access management, audit logging, isolation controls
- 

## Experience

### Parcha — Founding AI Engineer

Sep 2023 – Present | San Francisco, CA

First engineering hire. Technical #2 behind CTO. Built the AI platform from zero to production serving enterprise clients.

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

**Agent Orchestration & API Architecture** - Core architect of **Grep.ai** — ranked **#1 on DeepResearch Bench** (54.37), achieving **99.7% accuracy** on PhD-level research tasks - Designed multi-agent orchestration with configurable depth modes: FAST (single agent, sub-second), DEEP (expanded tooling), ULTRA-DEEP (spawns sub-agents for parallel research with state checkpointing) - Built 50+ MCP tool integrations wrapping internal APIs and external data sources (OpenCorporates, sanctions lists, SEC, courts across 200+ jurisdictions) - Designed 300+ agent skills across 16 industries with fail-safe execution and graceful degradation

**Evaluation & Developer Experience** - Built “Capture First, Measure Later” telemetry pipeline for production model monitoring; integrated Braintrust and Scorecard for trace-based evals and LLM-as-judge monitoring -

Identified false negatives during enterprise pilot; validated fix on 100 prod cases, then productized into self-service evaluation UI for compliance teams - Early Claude Code adopter (Feb 2025 research preview); developed plan-then-execute workflow, converted entire team from Cursor through organic adoption

**Client-Facing & Leadership** - Joined CEO on enterprise sales calls to explain AI architecture, model governance, and reliability - Supported client compliance teams through integration rollouts, post-go-live tuning, and ongoing case reviews - Early access participant across Anthropic programs (Flannel EAP, Sessions API EAP, Agent SDK tracing); provided direct product feedback

---

## **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 fine-tuned models
  - Built gRPC inference services for chatbot; designed message processing pipeline in Scala
  - Custom fine-tuning of LayoutLM family of models for document processing 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.

- Established Airflow as primary MLOps platform for model training and deployment pipelines
- 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

*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