ILLIA DYKANSKYI

ML/AI Engineer, Team Lead

Backend Engineer, Software Engineer, System Architect, Optimization Specialist

my.jobo@pm.me linkedin.com/in/illia-dykanskyi github.com/moon-strider

Citizenship: Ukrainian

PROFESSIONAL SUMMARY

Senior AI/ML engineer focused on turning LLMs and computer vision into pragmatic, production-ready systems. I work across the stack: from prototypes and data pipelines to evaluations and lightweight productization, comfortable solving ambiguous problems end-to-end.

Making complex AI approachable for engineers, PMs, analysts, and leadership. Documenting solutions so teams adopt them fast.

CORE STRENGTHS

- LLM/CV engineering
- Building agentic workflows and data analytics pipelines
- Rapid integrations (Telegram, web services)
- Technical writing, standardizing workflows
- Leading zero-to-one efforts
- Transforming ambiguous requirements into clear specifications and actionable plans

KEY SKILLS

Leadership	Project Management, Team Leadership, Technical Documentation
ML/AI	LLMs, Computer Vision, NLP, NER, OCR, Vector Search, RAG Systems, Model Distillation, LangGraph, LlamaIndex
Technologies	Python, FastAPI, Flask, BigQuery, LangChain, Ray, Aiogram, Kafka, RabbitMQ, Celery, dbt
Databases	MySQL, PostgreSQL, BigQuery, MongoDB, Milvus, Chroma, Weaviate, Pinecone
Tools	Git, Docker, Kubernetes, CI/CD, Jupyter, MLFlow, Airflow, Dagster, Kubeflow, Optuna, Automatic 11/11

TECHNICAL SKILLS

ML Frameworks	PyTorch, TensorFlow, Keras, scikit-learn, Hugging Face Transformers, TensorRT, XGBoost, CatBoost
LLM & APIs	LangChain, LangGraph, LlamaIndex, OpenAI, Anthropic Claude, xAI Grok, DeepSeek, OpenRouter, Groq, Chatsky (ex. DialogFlowFramework)
Computer Vision	OpenCV, ffmpeg, YOLO (v3-v12), RT-DETR, EfficientNet, EfficientDet, DeepSORT, OpenPose, Fine-tuning
Backend & APIs	Python, FastAPI, Flask, SQL, Go, HTML/CSS/JS, Svelte, Redis, PHP, JQuery, Bitrix
Data & Cloud	GCP, BigQuery, Vertex AI, AWS, Sagemaker, Hetzner, Cloudflare, MongoDB, pgvector, Spark, Snowflake, Databricks
MLOps & Data	MLflow, Airflow, Ray, DVC, Dagster, dbt
DevOps & Tools	Docker, Kubernetes, Helm, Terraform, Jenkins, Caddy, OvenMediaEngine, NVENC, ffmpeg, Git, CI/CD, Kubeflow
Observability	Grafana, Prometheus, ELK Stack, Loki, PostHog
Soft Skills	Adaptability, Stress Resistance, Collaboration, Technical Communication

LANGUAGES

- English – Professional Working Proficiency
- Ukrainian – Native or Bilingual
- Russian – Native or Bilingual
- German – Elementary

EDUCATION

Master's Degree in Digital Platforms and Big Data Analytics
Dubna State University | September 2022 – June 2024

Engineer's Degree in Big Data Analytics
Dubna State University | September 2020 – June 2022

Bachelor's Degree in Computer Science
Dubna State University | September 2018 – June 2022

CERTIFICATIONS

- Bitrix Platform Design and Configuration Integration (1C Company, September 2021)

PROFESSIONAL EXPERIENCE

Founder & Lead Engineer

Sayanara (sayanara.com) | November 2025 – Present | Remote

Stack: Python, FastAPI, PostgreSQL, pgvector, Redis, Svelte, Docker Compose, Caddy, Hetzner, S3, D2, Cloudflare,
OpenRouter, Groq, OvenMediaEngine, NVENC, TensorRT, Creem, PostHog, Alembic, ffmpeg, Argon2

- Architected, built, and shipped — solo, AI-agent-assisted — a media analytics/processing product that automatically detects and redacts sensitive content (PII, named entities, and user-defined policy/blacklist terms) across audio and video; live at sayanara.com as an audio alpha, with video redaction (face, object, OCR, and NSFW detection) and live-streaming redaction built, tested, and slated for launch.
- Engineered the core detection-and-redaction pipeline end-to-end: speech-to-text → topic and policy matching → entity and policy-term detection → a configurable final stage that either applies time-aligned redaction or emits detection-only logs.
- Engineered real-time streaming redaction with OvenMediaEngine, using GPU-accelerated encoding (NVENC) and inference-time optimizations to hold low latency under load.
- Designed a modular, fail-safe system built for high throughput, with automatic job recovery and resilient processing.
- Owned operations end-to-end — provisioning, deployment, security hardening, and ongoing maintenance of the production environment and release pipeline.
- Integrated subscription billing (Creem) with plan-based credit balances, replay-safe idempotent webhooks, a hosted customer portal, and refund/dispute handling wired to credit state.
- Designed and built the complete front-end application — the entire user-facing product, from onboarding and upload through review, results, and billing — with a clean, responsive UX.
- Built the product to be GDPR-compliant and privacy-preserving by design, including its product analytics, which capture only de-identified, aggregate signals and never user content.

Technical Lead [Contract]

StupidGoodAI | January 2026 – April 2026 | US, San Francisco

Stack: AWS, S3, Valkey, Next.js, fastmcp, Aurora (PostgreSQL), ChromaDB, Cerebras, FastAPI, ECS/EC2, Cognito,
Kubernetes, Terraform, Prometheus, Grafana, LangChain, LlamaIndex, OpenAI API

- Turned an unstructured concept into an architecturally sound restaurant recommendation platform; partnered with the founder on core functionality, roadmap, and technical specs, producing 170+ ADRs and module documentation.
- Acted as a full-cycle developer in a lean team (founder, data scraper, me): frontend, data engineering, analytics, system design, implementation, testing, and direct founder collaboration; delivered a working MVP for seed fundraising.
- Built an ETL pipeline that ingested, normalized, structured, and deduplicated millions of raw venue records from 100+ sources into 200k+ unique establishments with enrichment and integrity checks.
- Engineered a multi-stage RAG ranking system combining ETL data, vector search, reranking, explicit preferences, and behavioral signals for hyper-personalized recommendations.
- Designed and implemented an agentic AI concierge with persistent memory and custom MCP tools, enabling natural-language access to platform features and complex user workflows.
- Implemented observability with Prometheus and Grafana for system health, proactive issue detection, and high availability.

AI/ML Team Lead

Crypto Wallet | May 2025 – January 2026 | Dubai, UAE

Stack: Dagster, AWS, Sagemaker, Snowflake, Spark, Kubernetes, Helm, Ray Framework, Weaviate, Pytorch, OpenCV, PostgreSQL

- Built modular KYC platform with stateless ML Core: Document Recognition/ Classification/Extraction, Liveness Check, Face Matching, Deepfake Detection, Document Tampering Detection
- Achieved 100x cost reduction vs vendors ($7 → <$0.01 per 1000 verifications) by eliminating proprietary dependencies; from concept to production-ready MVP demo in 1 month
- Managing a team of 5 engineers; structuring research, development, and long-term planning
- Developed Python SDK for internal/external integrations, reducing client integration time from weeks to days
- Presenting technical roadmaps and progress to C-level stakeholders

Senior AI/ML Engineer

Crypto Wallet | December 2024 – May 2025 | Dubai, UAE

Stack: Ray Framework, GCP, Kubernetes, Helm, Chroma, Pytorch, OpenCV, Aiogram, Celery, BigQuery, Vertex AI, Grafana, Prometheus, LangGraph, Kafka, XGBoost, CatBoost

- Operated as a cross-functional IC across data, ML, and product
- Set up low-latency MySQL→BigQuery stream replication for 2.5+ TB; reduced heavy analytical queries from days/weeks to seconds, offloaded prod DB, eliminated stale aggregation tables
- Built anti-cheat analytics pipeline (800+ line SQL); ~90% precision, 0 false-positive bans, automated detection and banning of 100% of gross violators
- Built centralized DB API layer; eliminated data inconsistencies across 5+ dashboards and tools, new integrations no longer require writing custom queries
- Delivered Telegram admin bot (Aiogram + Go): real-time BigQuery analytics, role-based access (superadmin/admin/support), prod DB mutations; replaced legacy unmaintained web dashboard
- Built invoice parser for chat payments: extracts amount, currency, recipient from unstructured forwarded messages using NER+NLP, pre-fills transaction form
- Authored first centralized DB documentation (100+ tables); analyst onboarding cut from weeks to days
- Provided math support for game designers: in-game economy balancing, currency denomination

ML/Data Engineer [Contract]

FileMarketAI | September 2024 – October 2024

Stack: Terraform, Kubernetes, Helm, Kubeflow, Snowflake, GCP, Milvus, Vercel, Chroma, PostgreSQL,
LangChain, LangGraph, Dagster, LlamaIndex

- Built crypto market intelligence platform: Twitter/YouTube ingestion pipeline with NER-based semantic analysis, RAG-powered chat interface for market research queries
- Implemented multi-model ensemble: intent classification routing to specialized pipelines, LLM agent with tools (SQL queries, chart generation, ML predictions fed back to context)
- Trained ML predictors (CatBoost, XGBoost, AutoML Forecasting) on Vertex AI for trend forecasting

Machine Learning Engineer

DeepPavlov.ai | October 2022 – December 2024 | Remote

Stack: Databricks, Jenkins, OpenCV, Ray Framework, ELK, Kubernetes, Helm, MLFlow, LangGraph,
Chatsky (ex. DialogFlowFramework), Grafana, Loki, MySQL, Pytorch, TensorFlow

- Open-source platform Dream (dialog systems) and tooling development
- Created dream_voice distribution (ASR/TTS, audio captioning); led Audio Captioning research: dataset collection/cleaning, model finetuning, benchmarking across architectures
- Completed CLIP model distillation (teacher-student): 50% size reduction with only 2-3% accuracy loss, significantly improving inference speed for production deployment
- Extended dp-agent to integrate Dream with Telegram: video/audio message handling, streaming responses
- Contributed to dream_multimodal, dream_mint, dream_ocean, and dream_robot – services, annotators, and pipelines used across the platform
- Built dream_robot ROS integration with Minecraft adapter using LLaVa/Fromage VLMs; implemented recursive task decomposition (goal→split→plan→execute→feedback→iterate) for agent control
- Built prototype dialog assistant for pick-up points using Chatsky (ex-DialogFlowFramework)
- 10+ merged OSS PRs across Dream ecosystem – code reviewed and adopted by internal/external users
- Master's thesis: "Distributed Real-Time Multimodal Dialog Systems" – resulted in productionized modules

ML/Data Engineer [Contract]

Virtual Showroom | June 2023 – July 2023

Stack: AWS, Sagemaker, Stable Diffusion, Automatic 11/11, Ray Framework, dbt, Pinecone, PostgreSQL, Docker, Kubernetes, OpenPose, C++

- Developed SaaS for automated product photography: input product photo → output marketplace-ready images, eliminating need for professional models, studios, and photographers
- Built proprietary data labeling pipeline using OpenPose (C++) for pose extraction from unlabeled photos; managed StableDiffusion finetuning experiments across hyperparameter configurations

Computer Vision Developer

Videointellect | December 2021 – September 2022 | Remote

Stack: OpenCV, Optuna, MLFlow, Docker, Kubernetes, Ray Framework, Airflow, Celery, Pytorch, TensorFlow

- Benchmarked object detectors/trackers (YOLO v3-v8, EfficientDet, DeepSORT); designed standardized internal evaluation pipeline (mAP/FPR/FNR), reducing model comparison time from days to hours
- Researched and finetuned specialized detectors: fire, smoke, PPE (safety equipment), prohibited symbols; managed full data pipeline – collection from open sources, cleaning, labeling, hyperparameter tuning
- Built open-source C# annotation tool (WinForms) with YOLO-format export; accelerated internal labeling workflow 3x compared to generic tools
- Delivered custom counting system for rubber briquettes on conveyor (on-premise, Bachelor's thesis): <1% counting error at ~30 FPS on 1080p, robust to lighting changes; algorithmic solution minimized hardware requirements, enabling deployment on client's existing edge devices
- Load-tested multi-camera streams on local servers; profiled throughput and latency bottlenecks
- Contributed to maintenance and incremental improvements for existing client deployments

Backend Developer

Hopper IT | September 2020 – November 2021 | Remote

Stack: PHP, JS, HTML, JQuery, Bitrix

- Implemented Bitrix managed/composite caching for several high-traffic pages; TTFB down ~25% and memory footprint down ~20% under peak
- Wrote small backend features and admin tools (PHP/JS), improved logs/alerts, and fixed integration issues with payment/forms

OPEN-SOURCE CONTRIBUTIONS & PUBLIC PROJECTS

- DeepPavlov Dream Platform: github.com/deeppavlov/dream/pull/505
- LLM-based NER: github.com/moon-strider/ner
- Swarm of Experts: github.com/moon-strider/swarm-of-experts
- Agent Injector: github.com/moon-strider/agent-injector
- MCP testing tool: github.com/moon-strider/mcp-probe
- RAG Experiments: github.com/moon-strider/miRAGe
- RL KYC Task Environment: github.com/moon-strider/rl-kyc-task-env
- CLIP model distillation: github.com/moon-strider/clip-comparison
- Perplexity MCP Server: github.com/moon-strider/perplexity-mcp