BOHDAN CHUPRYNKA

bohdan.chuprynka@gmail.com

AI Engineer

Cleveland, OH

SUMMARY

Applied AI Engineer building retrieval, agent, and backend systems across web and macOS products. Founding AI Engineer at Remly | Cybersecurity Intern at Cleveland Clinic. Combining RAG, semantic search, MCP, and AI infrastructure with data protection, classification, and secure system design.

EDUCATION

High School Diploma

Cleveland, OH

- High School Diploma (Expected June 2027)
- Weighted GPA: 3.97
- National Honor Society Student
- AP Scholar with Distinction

TECHNICAL SKILLS

Core ML & AI Engineering

• **Deep Learning:** PyTorch, TensorFlow, HuggingFace Transformers
• **Generative AI & RAG:** LangChain, LangGraph, LlamaIndex, Embeddings, Vector Search (Qdrant, Pinecone), Peft/LoRA, Prompt Engineering
• **Computer Vision:** Image Segmentation, Object Detection, Feature Extraction
• **Classical ML:** Time-Series Forecasting, Feature Engineering, Cross-Validation, Hyperparameter Tuning, A/B Testing

Deployment & MLOps

• **Containerization & Serving:** Docker, FastAPI, APIs
• **Cloud Platforms:** AWS (S3, EC2, Lambda, SageMaker)
• **Experiment Tracking & Versioning:** MLflow, Weights & Biases, DVC
• CI/CD: GitHub Actions
• Reproducible ML Pipelines

Backend & Data Engineering

• Advanced Python (OOP, Packaging), SQL (CTEs, Window Functions, Query Optimization)
• ETL/ELT Pipelines
• Vector Database Management
• Bash

Analytics & Evaluation

• Pandas, NumPy, Scikit-learn
• Matplotlib, Seaborn
• Experimental Design & Model Evaluation

Languages

• English (Fluent), Ukrainian (Native), Russian (Fluent)

CREDENTIALS

- **NCWIT Aspirations in Computing – Honor Mention (x2)**
- **Google Data Analytics** (June 26, 2025)
- **IBM AI Engineering Specialization** (October 2, 2024)
- **Tensorflow for Deep Learning** (December 14, 2025)
- **SQL For Data Science** (December 5, 2025)

EXPERIENCE

Founding AI Engineer - Remly

June 2026 - Present

- Built retrieval, agent, and product features across Remly's Next.js web app and Swift macOS client, making local Mac files and connected apps searchable by meaning and available to AI agents with cited answers.
- Implemented exact-filename search, relevance ranking, Ask context and citations, connector integrations, Mac ingestion batching, and OAuth 2.1 MCP retrieval with access controls and cross-platform tests.

Cybersecurity Intern - Cleveland Clinic

June 2026 - August 2026

- Reviewed 17 sensitive-data classifiers, engineered 13 custom regex classifiers, and built a privacy-safe validation harness with synthetic tuning, holdout datasets, and six-format tests, achieving approximately 100% precision and recall on synthetic evaluations.

Co-Founder & CTO - Aximon (Antler S26)

March 2026 - May 2026

- Built AI tutoring and platform features across a TypeScript extension and backend, including learner and code-context injection, Python validation, automated checks, interactive IDE previews, model-provider routing, shared LLM-output caching, cost controls, cloud-synced chat, and editor analytics.

AI Engineer - Cleveland Clinic Internship

Vertebrae X-ray Segmentation for Neurosurgical Planning

June 2025 - August 2025

- Built an nnU-Net pipeline for vertebral segmentation in spinal X-rays, covering preprocessing, training, post-processing, and evaluation with labeled masks and clinical collaborators.

PROJECTS

Policy RAG System | Github | Live Demo

Python, NumPy, FastAPI, OpenAI API, Docker, Streamlit, GitHub Actions

• **Built** a RAG-based Q&A system that ingests policy documents (TXT/PDF) and returns citation-grounded answers via hybrid dense + lexical retrieval.
• **Implemented** the retrieval pipeline from scratch (embeddings, cosine similarity, BM25, hybrid score fusion) using Python + NumPy. No LangChain or LlamaIndex.
• **FastAPI** backend with session-scoped indexing, TTL policies, and input validation. Containerized with Docker + GitHub Actions CI, deployed via Streamlit.

Persona-RAG | Github | Paper

Python, PyTorch, Unsloth/LoRA, LangGraph, Qdrant, MLflow

• Fine-tuned a Qwen2.5-3B LoRA, served locally via llama.cpp (GGUF), that matches a GPT-4o-mini hybrid-retrieval product on voice register at $0 vs. $0.37/1K replies and ~11× fewer tokens.
• Built both arms behind a single LangGraph pipeline with hybrid BM25 + dense retrieval, MMR, recency reranking, and base-model/length-cap controls to isolate the fine-tune's impact.