PAVEL VASILYEV

LLM Engineer / AI Infrastructure Specialist

Jerusalem, Israel
marchdown@gmail.com
+972-54-343-1123
marchdown · pavel–vasilyev

Built production LLM systems for therapeutic chatbots, including prompt engineering, safety guardrails, and RAG pipelines using GPT-3.5/4 and Claude. Previously developed ML/NLP systems across fintech, healthcare, and quantitative research for over a decade. Work primarily in Python with strong functional programming background (Clojure, Elixir). MSc Computational Linguistics, BSc Nuclear Physics.

PROFESSIONAL EXPERIENCE

Independent Consultant — Mar 2024 – Dec 2024

Various Clients · *Remote*

LevEhat NGO (CTO, 5 months): Led cloud migration from GCP to AWS, coordinated team of 5 for volunteer management platform
Fintilligence (Technical Lead, 4 months): Built market microstructure analytics platform with PIN/VPIN algorithms and Level 2 data pipelines
General consulting: Time-series forecasting models, NLP classification systems, infrastructure architecture

AI Architect — Jan 2023 – Dec 2023

Stamina AI · *Remote*

Built complete LLM pipeline: prompt engineering, context management, response generation, safety guardrails
Integrated OpenAI GPT-3.5/4 APIs with custom safety layers and content filtering
Designed conversation state management and session handling for therapeutic context

Independent Consultant — 2022 – 2023

Various Clients · *Remote*

Projects: time-series forecasting, text classification, data pipeline optimization

Senior Researcher — 2020 – 2021

Spring Research · *Remote*

Developed ML models for trading signal generation using time-series analysis and statistical methods
Built data pipelines processing Level 2 market data (tick-by-tick, order book, market depth)
Researched topology-inspired approaches to market microstructure modeling

Data Scientist — 2019

Nestlogic · *Remote*

Built computer vision models for advertising creative optimization (image feature extraction)
Implemented A/B testing infrastructure using statistical hypothesis testing (t-tests, chi-square)
Deployed ML models to production on Google Cloud Platform with Kubernetes

TECHNICAL EXPERTISE

LLM Engineering · 3 years production

OpenAI GPT-3.5/4, Anthropic Claude

prompt engineering, RAG pipelines, context management, safety guardrails, content filtering

production deployment, latency optimization, monitoring, cost optimization

therapeutic chatbots, conversational AI, safety systems

Programming

Python (10+ years production) · Clojure (10+ years) · Rust (5+ years) · Go (3+ years) · Elixir (2+ years) · Java, R, SQL (advanced), bash

ML/AI

Frameworks: PyTorch, TensorFlow, Keras, scikit-learn, XGBoost, LightGBM, CatBoost · NLP: spaCy, NLTK, Hugging Face Transformers · Deep Learning: CNNs, RNNs, LSTMs, GRUs, Transformers, attention mechanisms, transfer learning, fine-tuning

Cloud & Infrastructure

AWS: EC2, S3, SageMaker, Lambda · GCP: GCE, GCS, GKE · Containers: Docker, Kubernetes

Databases

PostgreSQL (expert) · MongoDB, Redis, Neo4j, SQLite, MySQL

EDUCATION

MSc Computational Linguistics **2016**

Russian State University for the Humanities (RSUH) · *Moscow, Russia*

Statistical NLP, Machine Translation, Information Extraction

BSc Nuclear Physics **2011**

Czech Technical University · *Prague, Czech Republic*

Mathematical Modeling, Statistical Analysis, Computational Physics

LANGUAGES

English (Fluent) · Russian (Native) · French (Conversational) · German (Conversational) · Czech (Conversational) · Hebrew (Basic)