

Pavel Vasilyev

LLM Engineer / AI Infrastructure Specialist

Jerusalem, Israel • marchdown@gmail.com • +972-54-343-1123

[marchdown](#) • [pavel-vasilyev-65105b149](#)

Summary

LLM engineer with 14 years ML/NLP experience and 3 years specializing in production language model systems.

Expert in LLM workflows, safety guardrails, prompt engineering, and scaling generative AI infrastructure.

Strong functional programming background (Clojure, Elixir) with production systems experience in Python, Rust, Go, Java, and R.

MSc Computational Linguistics, BSc Nuclear Physics.

Experience

Independent Consultant

Various Clients • Remote

Mar 2024 – Dec 2024

- **LevEhat NGO (CTO, 5 months):** Led cloud migration from GCP to AWS, coordinated team of 5 for volunteer management platform
- **Fintelligence (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

Stamina AI • Remote

Jan 2023 – Dec 2023

- 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

Various Clients • Remote

2022 – 2023

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

Senior Researcher

Spring Research • Remote

2020 – 2021

- 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

Nestlogic • Remote

2019

- 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

Data Scientist

Maverick Medical AI • Remote

2018

- Developed NLP system for medical named entity recognition in clinical text using spaCy and BiLSTM
- Built medical ontology framework for standardizing terminology across different hospital systems
- Created decision support tools for clinical workflows highlighting critical findings

Education

MSc Computational Linguistics

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

2016

Statistical NLP, Machine Translation, Information Extraction

BSc Nuclear Physics

Czech Technical University • Prague, Czech Republic

2011

Technical Skills

LLM Engineering

3 years production

APIs: OpenAI GPT-3.5/4, Anthropic Claude
Workflows: prompt engineering, RAG pipelines, context management, safety guardrails, content filtering
Infrastructure: production deployment, latency optimization, monitoring, cost optimization
Specialization: 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 • Functional background: Common Lisp, Haskell

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

Data Engineering

Apache Spark, Apache Airflow, Kafka • ETL pipelines, data quality, orchestration

Domain Expertise

LLM Production Systems (3 years): Therapeutic chatbots, safety systems, prompt engineering, RAG pipelines
Quantitative Finance (2+ years): Level 2 market data, PIN/VPIN algorithms, order flow analysis, backtesting
Healthcare AI (1 year): Medical NER, clinical terminology, HIPAA compliance, decision support tools
Financial NLP (1 year): SEC filings analysis, knowledge graphs, entity linking, sentiment analysis

Languages

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