Kudrin Vladislav

🤖 Machine Learning Engineer

💼 Work Experience

Petrel AI (B2B products) 📍 **Almaty**
February 2022 - April 2025 • 3 years 2 months

🐍 Python Intern (4 months)

▶ Code refactoring

▶ Small bug fixes

▶ Legacy rewriting

🧠 ML Engineer (2 years 10 months)

🚀 Examples:

▶ Chat bot (RAG + Chat agent) (~98% accuracy on answering right) (llama + chromadb)

▶ Text normalization (Few shot learning + RAG) (~88% on right mapping and ~97% for key-value) (qwen + chromadb)

▶ Commit reviewer (API Integration + Code agent) (~0.83 F1) (claude)

▶ AI tutor (TTS + STT + Image parsing + Internet search agent) (no metrics) (claude + llava + whisper + minimax audio)

▶ Spam detection (BERT fine tuning) (~0.99 F1 for NSFW, ~0.82 F1 for ad) (debert)

▶ HR Agent (API Integration + Document parsing + Scoring) (~0.71 F1 of right classification. mostly due to human factor) (claude + llava)

🎓 Education

Ural Federal University (2021 - 2025)

- Bachelor's degree in computer science
- Additional education in AI

🌍 Languages

- Russian (native)
- English (fluent)

💪 Hard Skills

- Creating python microservices
- Creating AI agents, RAG pipelines
- Working with text generation, text2text, classification models
- Working with databases
- Fine tuning
- Prompt engineering
- Deployment, working with cloud
- Integrating with other services

🤝 Soft Skills

- Responsibility
- Dedication
- Adaptivity
- Fast learning

⚡ Tech Stack

💻 Programming Languages

- ▶ Python (main)
- ▶ Javascript (a bit)
- ▶ C# (a bit)
- ▶ C++ (a bit)

🗄️ Databases

- ▶ PostgreSQL
- ▶ MongoDB (for document oriented)
- ▶ ChromaDB (production or testing)
- ▶ FAISS (production)

🤖 Agents

- ▶ Langchain (organizing pipelines)
- ▶ Langgraph (creating agents)
- ▶ Langfuse (agents monitoring)
- ▶ PydanticAI (creating agents)
- ▶ SmolAgents (creating light agents)
- ▶ LlamaIndex (integrating RAG)

🧠 Deep Learning

- PyTorch (fine tuning)
- Unsloth (fine tuning)
- PEFT (fine tuning)
- Transformers (testing)
- Sentence transformers (embedding)

⚙️ Backend and DevOps

- FastAPI (REST API)
- Docker (Images, compose, containers)
- Kubernetes (practiced in minikube)
- Git (github/gitlab)
- CI/CD (github actions)

🔧 MLOps

- TGI (production)
- vLLM (production)
- exllamav2 (production)
- llama.cpp (testing)
- ollama (testing)

🎯 Models

- Every opensource text generation (DeepSeek, qwen, mistral, llama, gemma...)
- Closed source text generation (GPT, Claude, Grok, Gemini)
- text2text and classification (DistilBERT, RoBERT, DeBERT, BART, T5, opus)

☁️ Cloud

▶ Google cloud (object storage, serverless containers, load balancers, sql clusters, firebase)

▶ Yandex cloud (For Russia)

▶ Openrouter (For fault tolerance and ease of use)

▶ DeepInfra, NovitaAI, Groq (Direct cheaper use)

🎭 Multimodality

▶ TTS (minimax audio API and kokoro)

▶ STT (whisper)

▶ Image text to text (llava, gemma, mistral small mostly for parsing)

▶ Video, audio (gemini flash)

💻 IDE

▶ Cursor (Now use claude 4)

▶ Windsurf (Sometime)

▶ Jupyter (For fine tuning or edge cases)

🌐 Frontend

▶ React (a bit)

▶ Streamlit

🖥️ OS

▶ Linux (for work and personal)

✨ More

- ▶ Built AI visual novel game in browser redcamptale.web.app
- ▶ Got 1st place in hackathon LLM coding challenge from sberbank (agriculture case)
- ▶ Contributing to open source projects (ollama, pydanticAI)
- ▶ Reading papers on arxiv (recent was codeact)
- ▶ Monitoring news on X* (Banned in Russia), Youtube channels (fireship, bycloud) and hugging face models (trending)
- ▶ Use Claude 4 as helper for coding.

🔗 Links

🐙 Github (some pet projects)

📊 Kaggle (participating in competitions, always top 10%)

🤗 Hugging Face (Have merges and GGUF conversions)

💻 LeetCode (python, C#, SQL)

📞 Contacts

📱 **Telegram**
@vladlen32230

📧 **Gmail**
vladlen32230@gmail.com