Timur **Ganiev**

Yerevan, Armenia

+ 374 95 240 157 | ganiev.tmr@gmail.com | esceptico | esceptico | @postimortem

Skills

Programming	Python, SQL, Bash, Git
ML Frameworks	PyTorch, PyTorch Lightning, Transformers, NumPy, Pandas, PEFT, TRL, DeepSpeed, FSDP
Experiment Management	Weights & Biases, MLFlow, DVC
Infrastructure	FastAPI, Docker, Podman, Streamlit, OpenAPI, Ubuntu
LLM Engineering	SFT, DPO, RAG, Prompt Engineering, Model Distillation, Safety Alignment, VLMs, Quantization

Experience

Replika *Remote*

Lead ML Engineer *Jan. 2025 - Present*

- Led features end-to-end from design and implementation to A/B rollout and launch.
- Built real-time multimodal video calls and improved streaming latency for live conversations.
- Designed the long-term memory layer; applied context engineering to improve retrieval and sustain dialogue consistency.
- Built an agentic conversation workflow with parallel tool execution, reducing end-to-end latency by ~40%.

ML Engineer | NLP *Oct. 2022 - Jan. 2025*

- Maintained high-load LLM services (up to 100 RPS) with focus on throughput and latency.
- Boosted model safety by aligning on synthetic data, raising recall from 5% to 60% and reducing FN rate by 60%
- Fine-tuned open-source models with SFT and DPO on user feedback to improve response quality.
- Developed internal tools for synthetic data generation, filtering, offline evaluation, and fine-tuning.

Embedika *Remote*

ML Engineer | NLP *Feb. 2022 - Sep. 2022*

- Saw to completion an active learning service on multi-modal data.
- Introduced toxic classifier service. Achieved 94% F1-Score.
- Implemented and deployed spell checking service using BERT and PyTorch Lightning.

Sber *Moscow, Russia*

ML Engineer | NLP *May 2021 - Feb. 2022*

- Led a team of 4 engineers. Onboarded 3 new hires.
- Managed 3+ projects from scratch to production ready solutions.
- Optimized models by model distillation, ONNX and quantization, leading to 80% of inference time reduction.
- Designed and implemented model showcase system, which reduced time to production rate by 15-20%.
- Collaborated with DevOps and ML engineers to develop CI/CD pipelines for automatic unit and integration testing.

Data Scientist *Jan. 2020 - May 2021*

- Wrote a multi-target text classification model based on lightweight CNN architecture. Achieved 93% F1-Score on a heavily noisy and multi-task dataset.
- Gained model robustness by using Integrated Gradients method along with adversarial OCR-based dataset. Resulting in a 4% F1-Score increase.
- Enhanced a NER model by changing a token vectorizing method that led to a 5% F1-Score increase.

Advanced.Careers *Remote*

Data scientist *Aug. 2018 - Aug. 2019*

- Built resume and job posting parsing systems on top of custom PDF reader and text vectorizer, that improved CV upload rate by 10% and reduced time to apply by 8%.
- Enhanced job matching score by 10% by introducing a Smooth Inverse Frequency document vectorization.

Education

Kazan National Research Technical University named after A.N. Tupolev *Kazan, Russia*

B.S. in Computer Science *Sep. 2014 - Aug. 2018*

- Thesis theme: «Personality traits prediction based on Twitter account data»

Projects

Perceiver-IO *Maintainer*

Unofficial implementation of PerceiverIO *Aug. 2021*

- Implemented fully working PerceiverIO model from scratch and published to PyPi.

Squeezer *Maintainer*

Lightweight knowledge distillation pipeline *Oct. 2021*

- Built a fully customizable and model agnostic model distillation pipeline. This project was widely used in work and pet-projects.