| Universitetskaya 1, Innopolis  Tatarstan, Russia, 420500  <https://github.com/kezouke> | **Smirnov Elisey** | +7 (917) 888 5135  eliwhatthe@gmail.com  [linkedin.com/in/elisei-smirnov-ds](https://www.linkedin.com/in/elisei-smirnov-ds) |
| --- | --- | --- |
| **Briefly about me**  ML engineer in the field of NLP. I have experience in building intelligent LLM agents, optimizing inference and automating complex pipelines. I am distinguished by systems thinking, active participation in open-source projects and the ability to turn research ideas into scalable solutions.  **Education** | | |
| **Innopolis, Russia** | **Innopolis University** | **Autumn 2022 - May 2026** |
| * Bachelor's degree in Data Analytics and Artificial Intelligence. * Average score: 4.86. Current course: third year. * Undergraduate Courses: Operating Systems, Software Systems Design and Analysis, Databases, Algorithms, Computer Architecture, Mathematical Analysis, Discrete Mathematics, Differential Equations, Probability and Statistics, Optimization, Machine Learning, Statistical Methods for Data Science, MLOps, Deep Learning. | | |
| **Experience** | | |
| **Machine Learning (NLP) Engineer** | **Advanced Engineering School** | **April 2024 - Now** |
|  | | |
| **BetaGuidance** (Limited Generation LLM)   * Developed constrained text generation techniques that increased LLM output accuracy by 90%. * Integration of the project with the Hugging Face and VLLM models, increasing the compatibility and efficiency of asynchronous operations by 30%. * Input [contribution](https://github.com/vllm-project/vllm/pull/4985) to the open source VLLM library, allowing BetaGuidance to use token IDs in Logits Processors.     **graphite** (Automation of creation of educational courses)   * Created an ETL pipeline for course generation: PDF parsing (Nougat) → data deduplication → content enrichment using LLM. * Set up a cluster of 6 servers (4×A100 each, 24 GPUs in total) with monitoring via Prometheus/Grafana.   **Agnia** (Development of a multifunctional AI agent)   * Designed the architecture of an LLM agent that converts NLP requests into a sequence of API calls to automate tasks. * Implemented a hybrid mechanism for retrieving API call arguments, integrating the system with Google Calendar, Todoist and Gmail. * Input [contribution](https://github.com/vllm-project/vllm/pull/4985) to the open source VLLM library, allowing BetaGuidance to use token IDs in Logits Processors. | | |
|  | | |
| **Personal projects** | | |
| **Brain MRI Tumor Detection and Classification** - <https://github.com/kezouke/MRI-Diagnosis-API> | | |
| * Developed a system for classifying brain tumors using MRI images. * Created a CNN model API based on FastAPI and a web interface using Streamlit to interact with the model. * Deployed the system using Docker for seamless scalability and deployment. | | |
| **Skills** (Hard): Python, FastAPI, Streamlit, Docker, ООП, PyTorch, Hugging Face Transformers, VLLM, LLM Orchestration, Quantization, Speculative Decoding, CNN, NLP, PostgreSQL, Pandas, MLflow, Apache Airflow, Prometheus, Grafana, RabbitMQ, REST/JSON API, Git, Continuous Integration, Английский (C1)  **Skills** (Soft): Teamwork, Analysis and decomposition of tasks, Open to new ideas and ready to propose them. | | |