Denis V. Kazakov, **Data Scientist**

Web-site | dvkazakov@gmail.com | @denis_v_kazakov | denis.v.kazakov | Github| **LinkedIn**

Experienced language translator and data scientist specializing in natural language processing

Skills

- **Machine learning:**
  - Large language models (prompt engineering, fine-tuning)
  - Deep learning (PyTorch): RNN, transformers
  - Classical ML (scikit-learn, XGBoost): regression, classification, clustering, tree-based methods
  - Cross-validation, bootstrap
  - Hyperparameter optimization (Optuna)
  - Data transformation (PCA, SVD)
  - Ensembles, pipelines
- **Automatic speech recognition and text-to-speech**
- **Python** (NumPy, Pandas, Matplotlib, SciPy, StatsModels, Jupyter Notebooks)
- **Natural language processing (NLTK, SpaCy)**
  - Machine translation
  - Machine translation quality evaluation
  - Tokenization
  - Named entity recognition
- **Statistics:**
  - Hypothesis testing
  - ANOVA
  - Multiple testing
- **Git**

Education

Data Science reskilling course (2022). Tomsk State University. Achievements:

Recommender Systems boot camp. Higher School of Economics

Stepik courses (certificates): **(**statistics, Python, R, machine learning, SQL, Linux). Achievements:

- All courses were completed with distinction, ranking **among 1 to 6% top students**

Diploma of higher education (equivalent of Master of Physics). Faculty of Physics, Moscow State University

Experience (more info)

Language Researcher, Expert Center, AWATERA (language service provider)

- **Speech-to-speech translation**
  - Developing a prototype
  - Testing in various language pairs
- **Large language models (LLM)**, such as OpenAI GPT and its analogs:
  - Using LLMs for text translation and editing, translation quality assessment, and glossary compilation
  - Prompt engineering
  - Fine-tuning:
    - LLMs for text translation
    - Transformers (BERT, RoBERTa) for text domain classification
  - Comparative analysis of various LLMs (LangChain, Llama, SeamlessM4T, Mistral, etc.)
- **Machine translation quality assessment**. Comparison of various MT quality metrics (hLEPOR, COMET)
- **Natural language processing**
  - Tokenization. Tokenizer comparison and selection for different tasks (NLTK, SpaCy, tiktoken, BPE, WordPiece)
  - Named entity recognition (NER). Comparison of various algorithms
- **Statistics**
  - Experiment design
  - Data preparation for blind testing
  - Sample size estmation with numerical modeling (bootstrap)
  - Non-parametric statistical significance testing (bootstrap) of mean values, correlation coefficients, machine translation quality metrics, etc.
- **Automatic speech recognition and text-to-speech**
- **Processing data in the industry-specific formats** (tmx, xliff): format conversion, data modification for specific tasks.

Previous experience: translator and interpreter.

About me

Fluent in English (many years' experience as a professional translator and interpreter)