Vladislav Lialin, Ph.D.

 guitaricet | guitaricet | Scholar | vlialin | vladlialin.com | vlad.lialin@gmail.com | 978 746 1818

SUMMARY

Ph.D. in Computer Science and Natural Language Processing, **3 years of commercial ML Engineering experience** (pre-PhD). M.S. in Math and CS and B.A. in Physics from Moscow Institute of Physics and Technology (top 41 in QS Ranking By Subject). Vladislav has 10+ years of programming experience, 7+ years of Data Science & Deep Learning experience, 5+ years of Research experience. Vladislav contributed to PyTorch, Fairseq, HuggingFace Transformers, and T0 on GitHub and loves making ML pet projects, excited about developing the next generation of multi-modal AI systems, AGI and AI safety. Hobbies: piano, music theory, cooking, hiking, learning languages.
Experience: Google, Apple, Amazon (internships), Parla (ed-tech startup, ML engineer), iPavlov (ML engineer). Has 10+ publications, publications at **ACL, EMNLP**, workshops at **NAACL, WACV**, and **NeurIPS**.
Professional Interests: AI Research, Generative AI, Scaling, Multimodal, Efficient Training/Inference, AI Safety
Technical Skills: DL, NLP, RL, Transformers, LLM, Large-Scale Distributed Training, RLHF, PEFT

WORK EXPERIENCE

Amazon, Research Intern at Alexa AI   New York City, NY, USA, May 2022 - Aug 2022

- Developed a video-conditioned LM using adapters to condition a pre-trained OPT on TimeSformer video representations using video captioning datasets. Accepted to a workshop at WACV with an oral presentation.
- Improved training throughput by 16x on 128GPUs using novel attention and modality mixing techniques and trained a 2 billion parameters multimodal LLM. Developed a method to utilize 50 million unlabeled videos.

Skills: NLP, Deep Learning, PyTorch, Transformers, LLMs, Multimodal, AWS, Kubernetes, Distributed Training

Apple, Research Intern at Siri Natural Language   Remote, USA, May 2021 - Aug 2021

- Developed TextMDETR — an model for in-scene text visual question answering that uses OCR information and generates answer using a fused transformer decoder and a pointer network.
- Evaluated pre-trained vision-language models (VideoBERT, ViLBERT, . . .) on Apple Vision Pro-related tasks.

Skills: NLP, Deep Learning, PyTorch, Multimodal, OCR, VQA, Object Detection, Distributed Training

Google, SWE Intern at Google Assistant   Remote, USA, May 2020 - Aug 2020

- Improved semantic parsing retraining speed by a factor of 20 by developing a computationally efficient continual learning recipe: paper, github.
- Implemented a state-of-the-art sequence-to-sequence network pointer network for semantic parsing.

Skills: NLP, Deep Learning, Google Cloud, PyTorch, Lightning, Continual Learning, Semantic Parsing

UMass Lowell, Research Assistant, PI: prof. Anna Rumshisky   Lowell, MA, USA, Aug 2019 - present

- Developed the first parameter-efficient **pre-training** method – ReLoRA, developed class-attention architecture, published the most well-known parameter-efficient fine-tuning survey.
- Developed a new NLP course to improve exposure to Transformers (students got hired by FAANG).
- Demonstrated that model architecture, dataset size and model size of pre-trained LMs are not predictive of their linguistic capabilities (ACL 2022 paper).
- Published Clinical NLP papers in collaboration with MIT MEDG. Currently working in collaboration with MIT, Harvard Medical School, and UCLA on evaluating GPT4 on clinical case reports and developing EHR LLMs.

Skills: NLP, Deep Learning, Distributed Training, Prompt engineering, Model Analysis, Clinical NLP, PEFT, RLHF

MIT, PRIMES Mentor   Cambridge, MA, USA, Feb 2020 - present

- Mentored high-school students and tutored them in NLP. Our projects were published at MIT and ACL 2022.
- My students got accepted to MIT.

Skills: Mentoring, NLP, Model Analysis, RLHF

iPavlov, Machine Learning Engineer   Moscow, Russia, Jul 2018 - Aug 2019

- Developed and deployed to production a customer support automation system (outsource at Sberbank, the largest bank in Russia). Tech Lead role with a team of 5 engineers.
- Led a non-profit course Deep Learning in Natural Language Processing in Moscow. The course engaged more than 400 people from Russia, Ukraine and other countries. Course videos are available on YouTube.

Skills: Deep Learning, NLP, RL, Teaching, Management of ML projects: Formalization, Data Collection, Metrics

Parla, NLP Tech-Ed Startup, Machine Learning Engineer   Moscow, Russia, Feb 2017 - May 2018
– Developed backend and machine learning algorithms for language learning application Parla.ai.
– Parla was in top10 most downloaded apps in Spanish and Russian AppStore and Google Play.
Skills: ML, Deep Learning, NLP, Sklearn, XGBoost, Time Series, PostgreSQL, AsyncIO, ML Deployment

Moscow Institute of Physics and Technology   Moscow, Russia, Jun 2016 – Jan 2017
Research Assistant at Laboratory of Fundamental Interactions
– Analysis of Large Hadron Collider data (ALICE experiment)
– π^0-hadron correlations in proton-lead collisions at $\sqrt{s(NN)} = 7$TeV

SELECTED PUBLICATIONS

Lialin, Vladislav and Anna Rumshisky (n.d.). "ReLoRA: High-Rank Training Through Low-Rank Updates". In: Workshop on Advancing Neural Network Training, **NeurIPS 2023**. arXiv: 2307.05695.

Lialin, Vladislav, Vijeta Deshpande, and Anna Rumshisky (n.d.). "Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning". In: arXiv: 2303.15647.

Lialin, Vladislav, Kevin Zhao, Namrata Shivagunde, and Anna Rumshisky (n.d.). "Life after BERT: What do Other Muppets Understand about Language?" In: **ACL 2022**. URL: https://aclanthology.org/2022.acl-long.227.

Lialin, Vladislav, Stephen Rawls, David Chan, Shalini Ghosh, Anna Rumshisky, and Wael Hamza (n.d.). "Scalable and Accurate Self-supervised Multimodal Representation Learning without Aligned Video and Text Data". In: **IEEE WACVW** 2023. DOI: 10.1109/WACVW58289.2023.00043. arXiv: 2304.02080.

Shivagunde, Namrata, **Lialin, Vladislav**, and Anna Rumshisky (n.d.). "Larger Probes Tell a Different Story: Extending Psycholinguistic Datasets Via In-Context Learning". In: To be presented at **EMNLP 2023**. arXiv: 2303.16445.

Deshpande, Vijeta, Dan Pechi, Shree Thatte, **Vladislav Lialin**, and Anna Rumshisky (n.d.). "Honey, I Shrunk the Language: Language Model Behavior at Reduced Scale". In: Findings of **ACL 2023**. Toronto, Canada: Association for Computational Linguistics. URL: https://aclanthology.org/2023.findings-acl.326.

Lehman, Eric and **Lialin, Vladislav** et. al (2022). "Learning to Ask Like a Physician". In: Clinical NLP Workshop at **NAACL** 2022. Association for Computational Linguistics, pp. 74–86. URL: https://aclanthology.org/2022.clinicalnlp-1.8.

Lialin, Vladislav, Rahul Goel, Andrey Simanovsky, Anna Rumshisky, and Rushin Shah (2020). *Update Frequently, Update Fast: Retraining Semantic Parsing Systems in a Fraction of Time*. arXiv: 2010.07865 [cs.CL].

AWARDS

TechCrunch Disrupt Startup Alley 2018, Parla was featured as top3 AI & Machine Learning startup
Producthunt Golden Kitty 2017, Parla was featured as Bot of the Year.

EDUCATION

2019 - present	PhD in Computer Science, NLP at **University of Massachusetts Lowell**	(GPA: 4.0/4.0)
2017 - 2019	M.S. in Applied Mathematics and Computer Science at **Moscow Institute of Physics and Technology (top50 in QS ranking by Subject)**	(GPA: 4.8/5.0)
2013 - 2017	B.S. in Applied Mathematics and Physics at **Moscow Institute of Physics and Technology (top50 in QS ranking by Subject)**	(GPA: 4.5/5.0)

SKILLS

Prog. Languages	Python (strongest), C/C++, Go, TypeScript, Swift, Java, Wolfram, Bash, Rust, Triton, CUDA
Machine Learning	NLP, Deep Learning (Transformers, RNN, CNN, GAN), Reinforcement Learning (Q-learning, PPO, SAC, Decision Transformer, RLHF), Time Series, Topic Modeling, Classical ML (linear models, gradient boosting, clustering), Distributed Training, Megatron-DeepSpeed
Frameworks	PyTorch, Jax/Flax, HF Transformers, HF Datasets, Deepspeed, FairSeq, PEFT, Wandb, Scikit-Learn, NumPy, SpaCy, PIL, OpenCV, PyAV
Languages	English (fluent), French (B1), Russian (native), German (A1), Spanish (A1), Mandarin (A1)