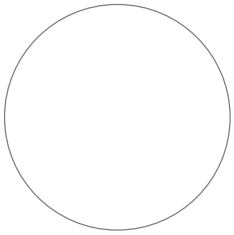
AIRI

2025

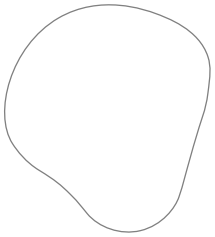
2024

Annual report



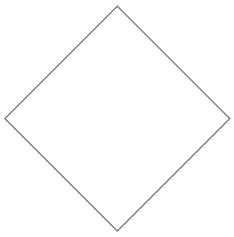
Nature

Constant movement and evolution, regularity and systematicity, laws and rhythms, chaos and life



Human

A living organism, part of nature, capable of thought, feeling and speech, imperfect



Technology

Human creation and its extension, logical and predictable, utilitarian, designed to help

AIRI

The unity of nature, human beings and technology

Table of Contents

The Institute’s fourth year

Ivan Oseledets
CEO

In 2024, the Institute experienced significant expansion, doubling its membership. The Institute boasts a cadre of over 200 dedicated professionals, including nine individuals who successfully defended their Ph.D. and Doctor of Science theses during the year. Notably, many of these professionals have come to Russia from abroad to join AIRI and contribute to the advancement of AI science both in Russia and in the global community.

At the same time, the number of successfully conducted researches increased, with 90 of these being presented at A and A* level conferences. 17 papers were accepted to the leading international conference on artificial intelligence, NeurIPS.

At AIRI, we continue to build universal AI systems that solve real-world problems. This year, we focused on the concept of applied AGI, which is based on scientific validity and business benefits. Together with industrial partners, we launched a number of ambitious projects on generative design and the application of AI technologies in medicine.

New collaborations with leading universities and technology companies are emerging. We were joined by partners from GIAP, SberMedAI, ISP RAS, KAMAZ Digital, Sibur Digital and many others. A notable highlight of 2024 was the opportunity to establish collaborative relationships with regional research centers.The sheer number of these centers, including NEFU, ASU, and TSU, is a testament to the rich diversity and strong competencies present in Russia. I am confident that thanks to AIRI’s commitment to connecting colleagues from diverse geographical locations and academic domains, the future will yield a multitude of unpredictable but pleasant results.

I am grateful to the AIRI staff and our partners for their hard work. Together, we are not only bringing AGI closer, but also helping to train the next generation of researchers and inspiring each other to make discoveries.

I invite you to learn more about the results of AIRI’s work in 2024, as they are thoroughly delineated in this report.

AIRI Mission

To create universal AI systems that solve real-world problems

The main goal of AIRI is to find opportunities to use artificial intelligence for solving complex scientific, social and economic problems. The Institute's researchers conduct both core and applied research with the aim of making significant advancements in the field of AI and its applications. Their work contributes to shaping the global research agenda.

AIRI Values

Human-centricity

Scientific freedom

Responsibility
& contribution

Openness
& transparency

Collaboration
& partnership

Focus areas

Research

Conduction of breakthrough research in the field of artificial intelligence and work towards the formation of a global center of expertise

Contribution to the development of artificial intelligence

Participation in the global development of artificial intelligence through the creation, development and support of open source projects

Scientific and industrial partnerships

Development of partnerships with scientific organizations, industry and government, development and commercialization of technologies in the field of artificial intelligence

Laboratories

Cooperation with institutes, universities and industrial partners to launch joint research laboratories in the field of artificial intelligence

Popularization of AI

Holding specialized conferences and events, creating and supporting competitions, promoting AI technologies

Key results

236
publications

66
conference papers (A*)

25
conference papers (A)

59
journal papers (Q1)

Research group leaders

Andrey Kuznetsov

FusionBrain Lab

Vladislav Shakhuro

Robotics

Aibek Alanov

Controllable Generative AI

Konstantin Sobolev

Video Generative AI

Aleksandr Panov

Cognitive AI Systems Lab

Alexey Kovalev

Embodied agents

Yury Kuratov

Memory-augmented models

Alexey Skrynnik

RL-based agents

Dmitry Dylov

AGI Med Lab

Dmitry Umerenkov

Foundation Models

Iaroslav Bespalov

Multimodal AI Architectures

Nazar Buzun

Representation Learning

Dmitrii Kriukov

Biomarker Research

Research group leaders

Ivan Oseledets

Computational Intelligence

Evgeny Burnaev

Learnable Intelligence

Anton Konushin

Spatial Intelligence

Egor Ershov

Computational color photography

Olga Kardymon

Bioinformatics

Marina Munkhoeva

Self-supervised
and representation learning

Ilya Makarov

Industrial AI

Alexander Panchenko

Computational Semantics

Alexander Tyurin

Optimization in Machine Learning

Evgeny Frolov

Personalization Technologies

Aleksandr Gorban

AI, Data Analysis and Modeling
Laboratory

Oleg Rogov

Reliable and Secure Intelligent Systems

Semen Budennyy

New Materials Design

Elena Tutubalina

Domain-specific NLP

Alexey Ossadtchi

Neurointerfaces

Artur Kadurin

DL in Life Sciences

Vladislav Kurenkov

Adaptive Agents

Artem Shelmanov

Weakly Supervised NLP

Management

Ivan Oseledets

CEO

Maksim Kuznetsov

Director of Strategic Development and Partnerships

Olga Surovegina

Science & Technology Partnerships Director

Stepan Mamontov

Head of scientific development department

Olga Popova

Head of project management office

Anton Rizaev

Chief Financial Officer

Ekaterina Sarafanova

Accountant officer

Nikol-Maria Cook

Head of purchasing department

Alexandra Broytman

Marketing and communications director

Maria Zvonareva

HR director

Yuliya Nikitina

Director of legal support

Konstantin Katanov

Chief Information Officer

Scientific results

Towards applied AGI

AGI =

Engineering approach

- → Scientific validity: Publications at A/A* and Q1
- → Transfer of ideas into the product pipeline / testing in the business environment

+

Business benefit

- → Functional customer for conducting a scientific study
- → Preliminary business impact analysis and prerequisite research conducted on the customer's side

Areas of applied AGI

Compute and Data	Multimodality	Agency	Embodiment	World Model
Effective training & inference	Modality	Self-learning	Hardware	Knowledge
Effective data storage	Omnimodality	Self-reflection/Reasoning	Software	Forecasting
Synthetic data	Perception Augmentation	Self-alignment	Interface	Causality
		Goal-setting & planning		
		Multiagency		

Main result

Dmytry Dylov
Head of AGI Med Lab

AGI Med

“Digital Assistant 2.0” and “Moonshot”, designed for primary patient reception and to assist radiology diagnosticians

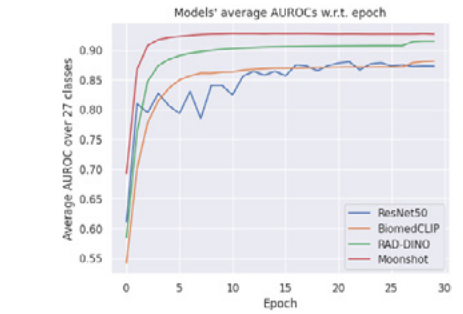
The mobile application of the medical company SberHealth (part of Sberbank’s healthcare industry) now has an AI assistant to help Russians take care of their health. The technology was developed by scientists from AIRI and SberMedAI on the basis of Sberbank’s GigaChat neural network model.

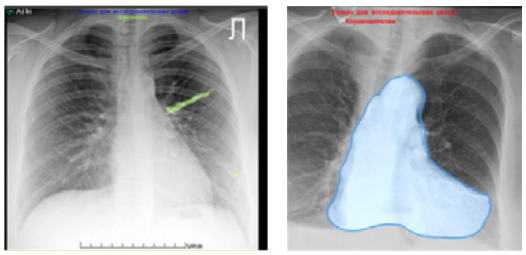
- → Pre-diagnosis and referral to target specialist
- → Decoding of tests and doctor’s reports
- → A set of intelligent security filters is implemented

At the same time, the “Moonshot” fundamental model proved to be a powerful model for analyzing radiological images. Trained on a large Moscow dataset, the model learned to generate the Findings and Conclusion sections of chest radiologic reports, predict pathology grade, and automatically contour them. The method combines a visual encoder (contrast learning) trained on medical images and a specialized biomedical language model. The approach effectively overcomes the complexities associated with medical semantics and accommodates the diversity of findings in images, providing an automated reporting process.

The model is trained on large datasets such as PadChest, BIMCVCOVID19, CheXpert, OpenI, and MIMIC-CXR, ensuring its robustness and adaptability to a variety of clinical cases. Evaluation metrics highlight the accuracy and ability of the model to account for fine medical details, demonstrating it’s great potential to optimize the work of radiologists. This achievement not only automates an important diagnostic step, but also facilitates the introduction of advanced AI solutions into medical practice.

At the major conference on medAI (MICCAI 2024), one of the key achievements was the creation of a new approach to interactive segmentation of complex elongated objects such





as wires, catheters, or veins without the use of traditional masks. The data structure developed is centered on a set of central lines of contoured objects and is particularly effective. Significant efforts have also been made to develop new approaches to medical image analysis. In particular, we proposed anatomical position embeddings for 3D images that can accurately predict the anatomical position of regions, opening up new possibilities for localizing organs and pathologies. These results were presented at MICCAI 2024, and 2 out of 3 Russian papers at the conference were presented by AGI Med.

AIRI AGI Med Lab participated in two international competitions: RRG24 at BioNLP and the MIDRC XAI Challenge, where they placed second and fifth, respectively, with the first generation of their Moonshot model. This was made possible by developing methods for occlusion classification and lung segmentation on a dataset with limited labeling. The team won the $5000 prize with the solution based on the CLIP model and also showed excellent results in the classification and segmentation tasks. The team was an order of magnitude smaller than the leading IT giants.

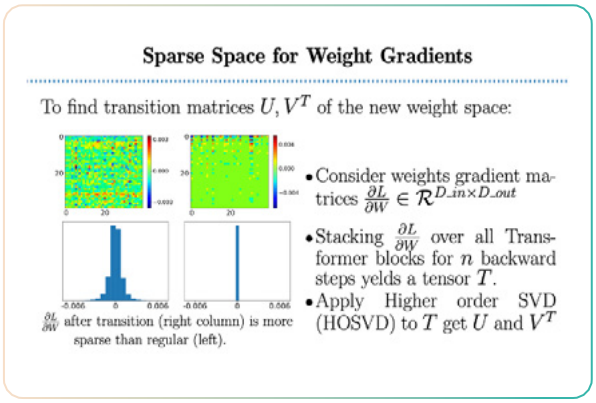
Majoir results

The SparseGrad method for parameter-efficient training of large language models based on the HOSVD tensor decomposition

Alexander Panchenko
Principal research scientist

Ivan Oseledets
CEO

In the SparseGrad method, which utilizes HOSVD decomposition, a space is identified where the matrix of linear layer weights is highly sparse (~1% of all parameters remain).By creating a new linear layer class and rewriting the autograd to account for the transition to this space, the authors reduced the number of parameters used in training Transformers and, consequently, the memory used. With a constant memory consumption on encoder architectures (BERT, RoBERTa) and decoder architectures (LLaMa2-7B), the model exhibits superior performance in comparison to LoRA and MeProp, a baseline that selects the most relevant parameters in the linear layer and trains exclusively those parameters. An empirical investigation was conducted to assess the efficacy of the proposed approach by implementing it on LLaMa 2 7B. The experimental results demonstrated that the method consistently yielded the optimal validation loss and the most optimal metric on the I-Bench question-answer benchmark. A qualitative analysis of the generated text corroborated these findings, validating the conclusions drawn from the experimental results.



SparseGrad: A Selective Method for Efficient Fine-tuning of MLP Layers
Viktoriia Chekalina, Anna Rudenko, Gleb Mezentsev, Alexander Mikhalev, Alexander Panchenko, Ivan Oseledets
EMNLP'24, A*

AMORE method for assessing the reliability and interpretability of large language models in chemistry based on augmentation of molecular structures

Elena Tutubalina
Principal research scientist

Artur Kadurin
Research engineer

Ivan Oseledets
CEO

Andrey Kuznetsov
Head of Laboratory

Denis Dimitrov
Scientific advisor

The recent integration of chemistry with natural language processing (NLP) has advanced drug discovery. The representation of molecules in language models (LMs) is crucial to improve the understanding of chemical processes.We have developed Augmented Molecular Retrieval (AMORE), a flexible zero-shot system for assessing the validity of chemical language models of different natures. These models can be trained solely on molecules for chemical tasks and on a combined corpus of natural language texts and structures represented as strings. The system's efficacy hinges on modifications to molecules that preserve their chemical properties, such as cecularization and cycle substitution.The metric employed to gauge the similarity between distributed representations of molecules and their modifications is based on the concept of molecular similarity.

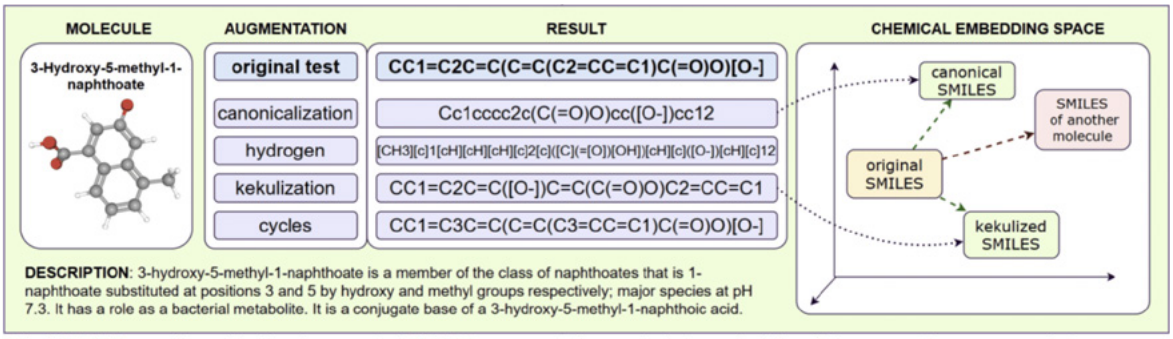


Fig.1 AMORE encodes original and augmented SMILES representations, calculates embedding distances, and assesses model performance based on top-1 accuracy, where the correct augmented SMILES is retrieved first

A*: Ganeeva, V., Sakhovskiy, A., Khrabrov, K., Savchenko, A., Kadurin, A. & Tutubalina, E. Lost in Translation: Chemical Language Models and the Misunderstanding of Molecule Structures. In Findings of the Association for Computational Linguistics: EMNLP 2024

A: The Shape of Learning: Anisotropy and Intrinsic Dimensions in Transformer-Based Models", Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, Andrey Kuznetsov, EACL 2024

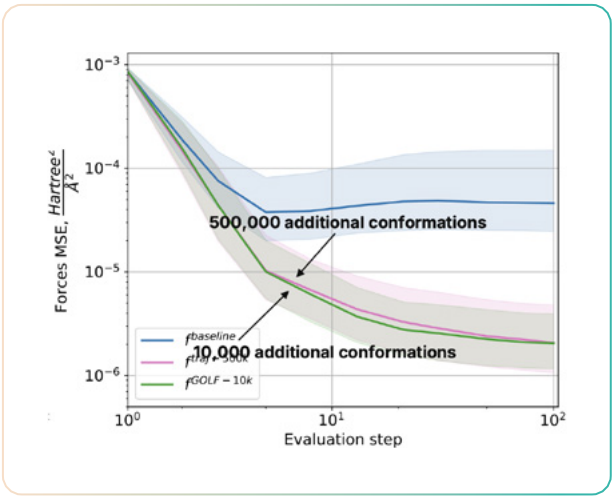
GOLF framework for active learning of neural potentials focused on local optimization of drug molecules

Aleksandr Panov
Laboratory director

Elena Tutubalina
Principal research scientist

Artur Kadurin
Research engineer

The GOLF approach that was developed enabled a 50-fold reduction in the amount of data necessary to pre-train neural networks on the task of optimizing conformations, which are spatial representations of molecules. A surrogate oracle is used to select conformations for subsequent partitioning using DFT. The energy is then evaluated to determine whether it has decreased. If the energy has decreased, the optimization is continued, or the NNP prediction is considered incorrect and the previous conformation is added to the training sample.



A*: Tsypin, A., Ugadiarov, L. A., Khrabrov, K., Telepov, A., Rumiantsev, E., Skrynnik, A., Panov, A., Vetrov, D., Tutubalina, E., Kadurin, A. Gradual Optimization Learning for Conformational Energy Minimization. In The Twelfth International Conference on Learning Representations (ICLR).

A*: Tsypin, A., Ugadiarov, L. A., Khrabrov, K., Telepov, A., Rumiantsev, E., Skrynnik, A., Panov, A., Vetrov, D., Tutubalina, E., Kadurin, A. Gradual Optimization Learning for Conformational Energy Minimization. In The Twelfth International Conference on Learning Representations (ICLR).

Study of internal properties of transformer models (internal dimensionality and anisotropy) characterizing the presence of linear dependence of a number of adjacent layers of embeddings

Ivan Oseledets
CEO

Denis Dimitrov
Scientific Advisor

Andrey Kuznetsov
Head of Laboratory

In this study, the investigation into decoder models was continued, and a linear dependence between successive layers of models based on the Transformer architecture was revealed. A model pruning process was proposed to replace part of the decoder layers by a linear transformation of the input vector. Benchmark measurements demonstrated that replacing 5-10% of the layers with a linear transformation does not have a significant impact on the model results.

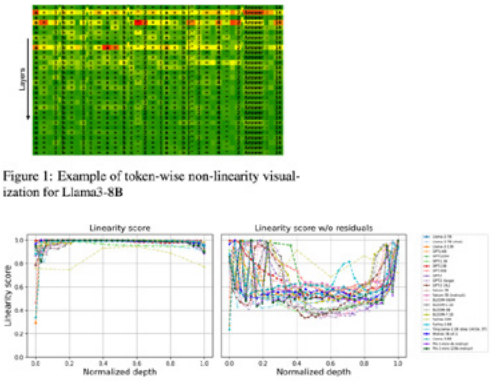


Figure 1: Example of token-wise non-linearity visualization for Llama3-8B

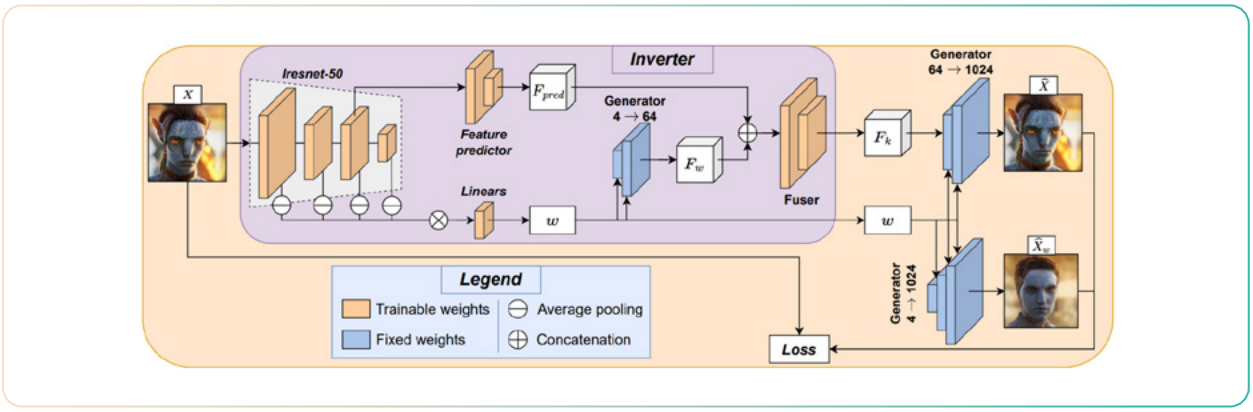
Figure 4-1: Linearity profiles for different open source models. Normalized depth is the layer index divided by the total depth.

A*: Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. Your Transformer is Secretly Linear. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)

Controllable Generative AI

Aibek Alanov
Research scientist

- The utilization of the StyleGAN model for the purpose of image manipulation has been demonstrated. The development of a StyleFeatureEditor model has been accomplished, thereby enabling the editing of images while concurrently reconstructing them with a high degree of accuracy. The state-of-the-art approach has been enhanced by a factor of four in terms of reconstruction accuracy, as measured by the LPIPS metric[1].
- The development of a model for transferring hairstyles to human faces is presented. A novel model was proposed, demonstrating superior quality in terms of hairstyle transfer and realistic results in comparison to existing baselines. Additionally, the model exhibits a significant increase in processing speed, operating at an order of magnitude faster[2].
- The development of efficient parameterization for the pre-training of neural network models has been proposed. This method allows for more efficient, in terms of memory utilization, pre-training of neural network models, including large diffusion models[3].



[1] A*: "The Devil is in the Details: StyleFeatureEditor for Detail-Rich StyleGAN Inversion and High Quality Image Editing" на конференции CVPR 2024.
Авторы: Денис Бобков, Вадим Титов, Айбек Аланов, Дмитрий Ветров

[2] A*: "HairFastGAN: Realistic and Robust Hair Transfer with a Fast Encoder-Based Approach" на конференции NeurIPS 2024. Авторы: Максим Николаев, Михаил Кузнецов, Дмитрий Ветров, Айбек Аланов

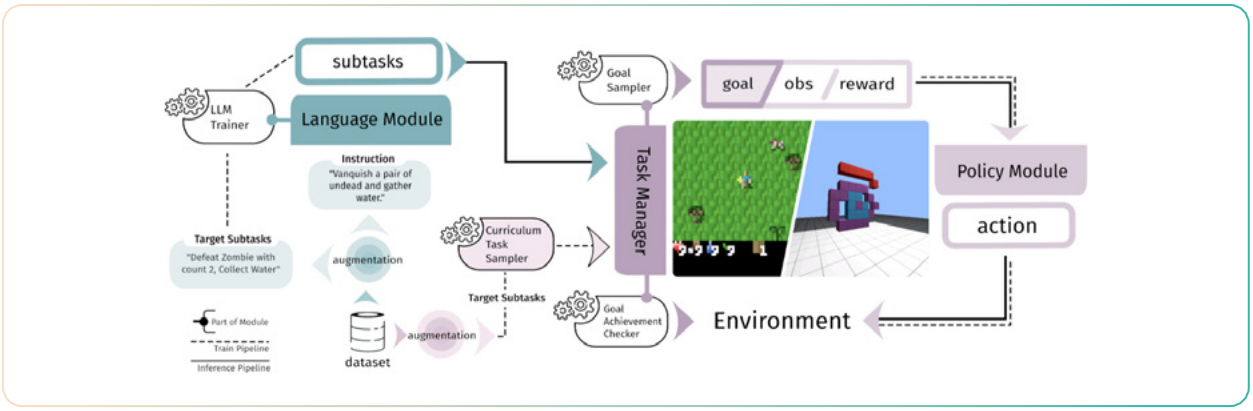
[3] A*: "Group and Shuffle: Efficient Structured Orthogonal Parametrization" на конференции NeurIPS 2024.
Авторы: Михаил Горбунов, Николай Юдин, Вера Соболева, Айбек Аланов, Алексей Наумов, Максим Рахуба

An original neurosymbolic architecture for controlling embodied agents in complex, dynamic environments. This architecture supports language-based planning and includes two types of world models to facilitate effective learning environments

Aleksandr Panov
Laboratory director

The findings of the study are presented in a doctoral dissertation defended at the MIPT dissertation council on the specialty 1.2.1 “Artificial Intelligence and Machine Learning”.

The IGOR approach for strategy building in multimodal environments utilizes a large language model to transform text and instructions into a high-level plan, which is then implemented by an RL-agent, building a strategy for behavior in the environment. The efficacy of the proposed method has been substantiated in the IGLU environment, where it achieved a superior performance compared to the top-ranked results in the NeurIPS competition. Additionally, in the Crafter environment, it surpassed the state-of-the-art solution provided by Dynalang.

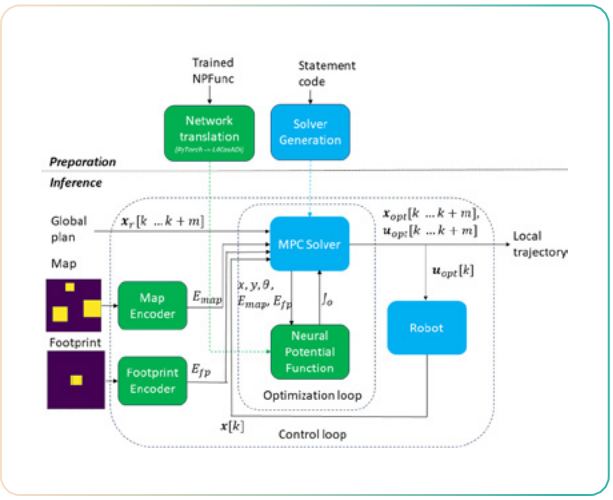


A: Volovikova, Z., Skrynnik, A., Kuderov, P. and Panov, A.I., 2024. Instruction Following with Goal-Conditioned Reinforcement Learning in Virtual Environments. In ECAI 2024 (pp. 650-657). IOS Press. [Core A]

A model for the problem of local planning considering dynamic obstacles

Aleksandr Panov
Laboratory director

A model was developed to compute the neural potential employed in MPC for the task of local path planning with dynamic obstacles. The model is based on a transformer architecture inspired by GPT. The constructed control system was evaluated through experimentation on a real mobile robot, demonstrating a competitive advantage over all existing world analogues. This research was presented at the prestigious robotics conference, ICRA 2024.



A*: Alhaddad, M., Mironov, K., Staroverov, A., Panov, A. Neural Potential Field for Obstacle-Aware Local Motion Planning, IEEE International Conference on Robotics and Automation (ICRA) 2024, pp. 9313–9320.

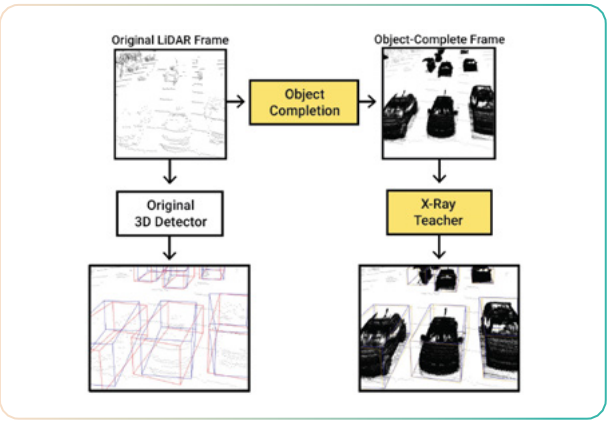
An approach combining LIDAR frame sequences through point cloud registration algorithms, achieving SOTA results in a semisupervised setting and enhancing the performance in a supervised setting

CVPR 2024 (A*), IEEE Access (Q1), demo IJCAI 2024 (A*)

Ilya Makarov
Principal research scientist

The approach has been demonstrated to enhance the accuracy and efficiency of 3D detection models to a considerable extent. This enhancement is of critical importance for a variety of applications, including autonomous vehicles and robotics, as well as other fields that rely on high-precision object recognition.

The technology has the potential to be utilized in autonomous transportation systems, warehouse automation, and robotics.



Gambashidze, A., Dadukin, A., Golyadkin, M., Razzhivina, M., & Makarov, I. Weak-to-Strong 3D Object Detection with X-Ray Distillation. CVPR 2024

LLM's strategic and ethical decision making research

Ilya Makarov
Principal research scientist

A framework has been developed for testing hypotheses concerning the correspondence between people's emotional reactions and LLM decisions. The findings indicate that emotional alignment with humans is predominantly weak in most cases. However, emotional prompts, particularly negative ones, have been observed to significantly impact the behavior of models, leading to a reduction in cooperation in games and a decline in quality in ethical tasks. NeurIPS (A*)

It is imperative to comprehend the emotional ramifications on models to ensure the development of safe and reliable AI systems in ethical and strategic applications.

The utilization of the framework facilitates the optimization of LLMs in domains where emotional context must be taken into account, such as in assistants or interactive systems.

Development of an empathic LLM assistant

An InsideOut approach was developed in which basic Ekman emotions were involved in formulating responses. Based on GigaChat and GPT-3.5, an improvement of 12-20% in emotion recognition and up to 3.9% in the quality of empathic responses was achieved.

ECAI demo (A)

The enhancement of empathy in large language models (LLMs) has been demonstrated to facilitate the development of assistants that exhibit greater human-likeness and efficacy. These assistants find applications in domains such as psychology, training, and client support services.

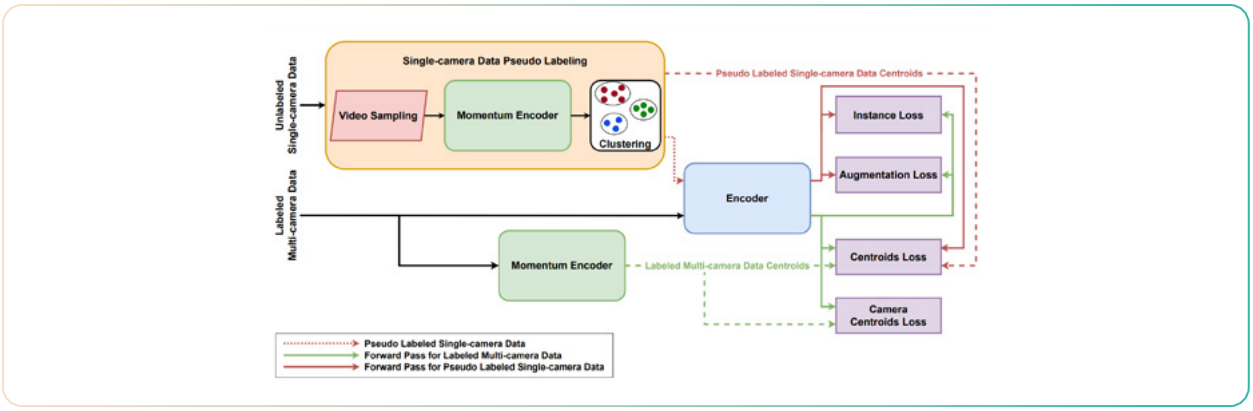
Mozikov, M., Severin, N., Bodishtianu, V., Glushanina, M., Nasonov, I., Orekhov, D., Pekhotin, V., Makovetskiy, I., Baklashkin, M., Lavrentyev, V., Tsvigun, A., Turdakov, D., Shavrina, T., Savchenko, A., & Makarov, I. EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas. NeurIPS 2024

Deeb, B. M., Savchenko, A., & Makarov, I. (2024). CA-SER: Cross-Attention Feature Fusion for Speech Emotion Recognition. In ECAI 2024 (pp. 4479-4482). IOS Press. ECAI 2024 Demo.

Smart learning on blended data for human recognition

Anton Konushin
Senior research scientist

A set of methods was developed for training human reidentification models on mixtures of unlabeled univariate and labeled multivariate data. This development enabled a substantial enhancement in the generalization ability of the models, as validated through testing in cross-dataset scenarios.



Mamedov T., Konushin V., Konushin A. ReMix: Training Generalized Person Re-identification on a Mixture of Data // arXiv preprint arXiv:2410.21938. — 2024 (принята на WACV 2025)

The first open environment and the world's largest dataset for In-Context Reinforcement Learning

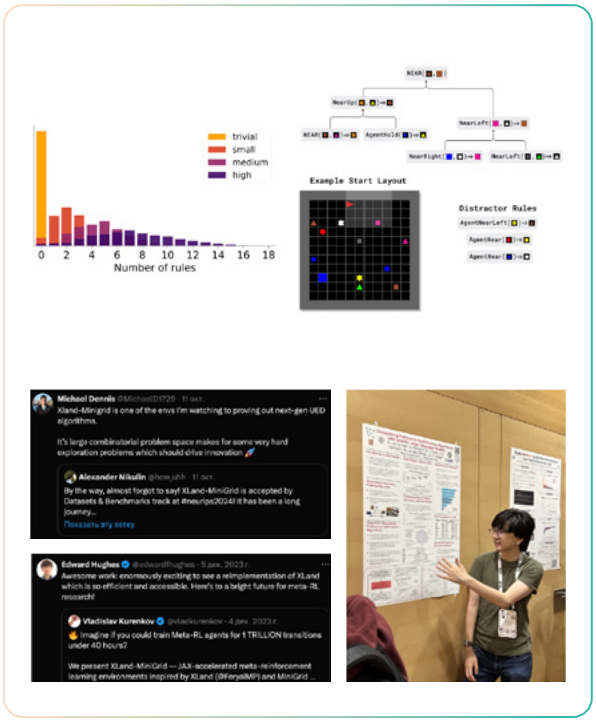
Vladislav Kurenkov

Research engineer

The XLand-MiniGrid environment, developed for the purpose of facilitating contextual learning, has been shown to significantly accelerate the rate at which experiments are conducted, thereby rendering the domain more accessible. By migrating the environment's computation to the GPU, the training of agents for experiments and the collection of data has been expedited tenfold, reducing the time required from weeks to minutes. In comparison to existing analogues, the environment offers a greater degree of diversity, encompassing millions of unique tasks of varying complexity, albeit abstract. The environment has garnered recognition within the research community and has been endorsed by multiple researchers, including those at Google DeepMind.

The XLand-100B dataset, the most extensive existing dataset in the field, was constructed on the basis of the environment. It will enable researchers to study in detail the scaling laws of current and future approaches.

Furthermore, current approaches in general are ineffective in solving the problems presented in the dataset, so it will serve as a good benchmark and a guiding star in the field.

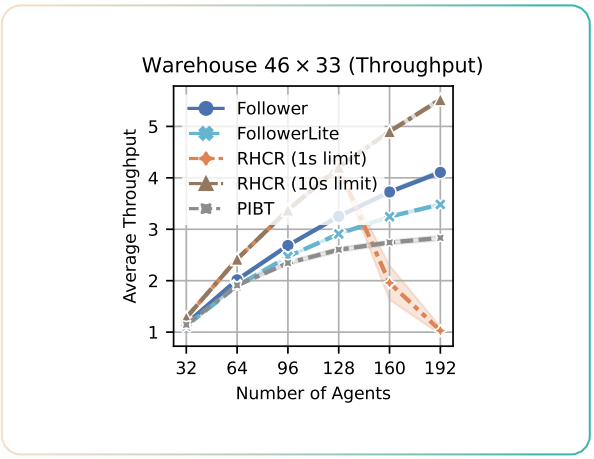


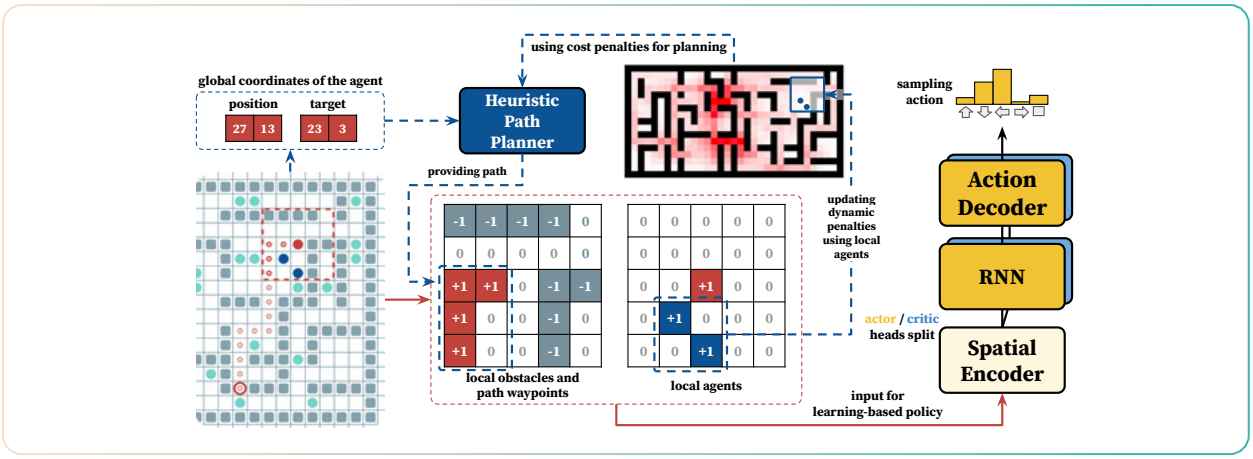
A*: XLand-MiniGrid: Scalable Meta-Reinforcement Learning Environments in JAX, NeurIPS 2024

Decentralized methods for solving the multi-agent planning problem

Aleksandr Panov
Laboratory director

A novel approach to addressing the challenge of multi-agent trajectory planning in a decentralized framework is presented. This approach integrates established techniques for addressing this challenge (e.g., heuristic search) with contemporary learning methods (e.g., model-based reinforcement learning, large-scale imitation learning with transformers, etc.). This research is a joint effort with A.A. Andreychuk, A.A. Skrynnik, M. Nesterova, and K.S. Yakovlev.





Skrynnik A., Andreychuk A., Yakovlev K., Panov A. Decentralized Monte Carlo Tree Search for Partially Observable Multi-agent Pathfinding // In Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI 2024). pp. 17531-17540.

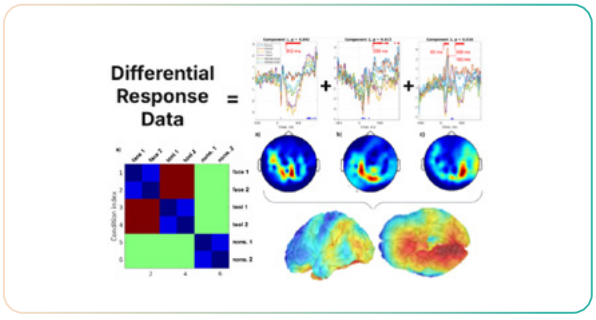
Skrynnik A, Andreychuk A, Nesterova M, Yakovlev K, Panov A. Learn to Follow: Decentralized Lifelong Multi-Agent Pathfinding via Planning and Learning. InProceedings of the AAAI Conference on Artificial Intelligence 2024 Mar 24 (Vol. 38, No. 16, pp. 17541-17549).

ReDisCA — a method for rapid representative similarity analysis of EEG and MEG brain activity data

Alexey Ossadtchi
Principal research scientist

The search for brain areas that exhibit a specific representational (geometric) structure in response to external stimuli. For example, responses to images of tools and images of faces exhibit a diametrical opposition, occupying “different sides” of the response spectrum when compared to responses to meaningless visual stimuli.

The development of the Representative Similarity Component Analysis (ReDisCA) method was completed, and it was applied to the analysis of MEG and EEG data to find neural sources with the desired representational structure. The ReDisCA method differs from traditional representational similarity analysis (RSA) in that it does not require an exhaustive search over a grid of neural sources. Rather, it uses an approximate closed-form optimization strategy to decompose EEG/MEG data into spatiotemporal components with specified representational properties. This innovation avoids the problem of multiple comparisons and significantly improves the detectability and localization of sources associated with user-defined representative similarity profiles. ReDisCA provides an accurate and interpretable framework for combining brain imaging data with computational models. It is 100 times faster than traditional RSA and 2 times more accurate. The work is published in Neuroimage (Q1), which is ranked in the top 1% of journals in its field.



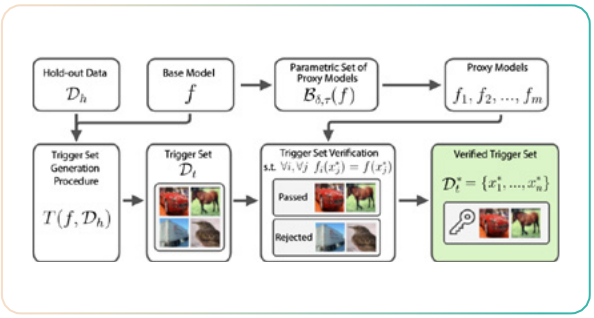
Q1: A. Ossadtchi, I. Semenkov, A. Zhuravleva, O. Serikov, E. Voloshina. Representational dissimilarity component analysis (ReDisCA). NeuroImage, 2024

Probabilistically Robust Watermarking of Neural Networks

Oleg Rogov
Senior research scientist

The result: a method of digital watermarking that is guaranteed to be resistant to the strongest black-box model stealing attacks. Article at IJCAI-2024 (A*)

The utilization of deep learning models on MLaaS (Machine Learning as a Service) product platforms has led to a surge of interest in digital watermarking methods for these models. These methods can be employed to substantiate the ownership of a specific model. However, existing methods often generate digital watermarks that are susceptible to attacks aimed at compromising model functionality. In this study, we have proposed a novel watermarking approach based on a trigger dataset, which exhibits remarkable resilience to functional theft attacks, particularly those involving knowledge distillation. This approach does not necessitate any additional model training and can be seamlessly integrated into any architecture.The fundamental principle of our method entails identifying a set of triggers that exhibit a high probability of transfer between the original model and a set of proxy models. Through experimental evaluations, it has been demonstrated that when the transferability probability of the set is sufficiently high, the method can be effectively utilized to verify the ownership of a stolen model. The efficacy of the proposed approach is evidenced by its superior performance in comparison to all existing state-of-the-art solutions for the digital labeling of machine learning models.

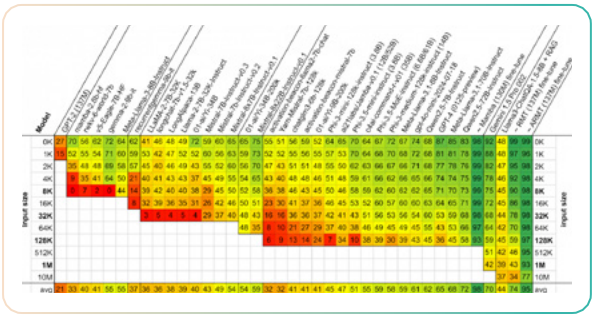


Pautov M, Bogdanov N, Pyatkin S, Rogov O, Oseledets I. Probabilistically Robust Watermarking of Neural Networks. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence 2024 August 3-9, (Vol. 33, pp. 4778-4787)

A novel benchmark for assessing the real-world performance of language models when confronted with large amounts of data

Yury Kuratov
Senior research scientist

Using the developed BABILong benchmark, we demonstrate the real efficiency of open-source and GigaChat family models at different training stages: short-context pre-train, long-contex tpre-train, supervised fine-tuning. High efficiency of the SFT stage on contexts less than 16k tokens and insufficient efficiency on large contexts is shown. This analysis is substantiated through its confirmation on both Russian and English tasks. It is concluded that BABILong/RuBABILong are the most suitable benchmarks for evaluating model training stages on task performance in large contexts.



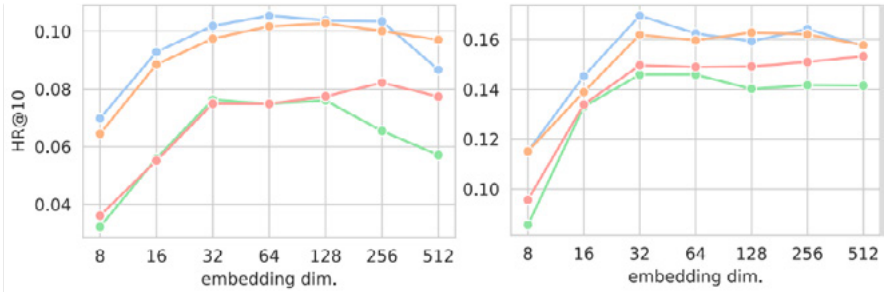
- As the context lengthens, there is a general decline in the quality of the models, though the extent of this degradation varies.
- The majority of models demonstrate effectiveness with only 10-20% of the declared context length.
- It has been demonstrated that even the most advanced models, such as the Gemini 1.5 and the Qwen 2.5, exhibit a decline in performance quality from 80% to 90% to 60% or less when executing extended tasks.
- BABILong demonstrates that even simple tasks with extensive contexts remain challenging for leading models to execute. The mere presence of a stated context comprising hundreds of thousands of tokens does not guarantee that the model possesses the capacity to process it effectively.

A*: BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack
Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Igorevich Sorokin, Artyom Sorokin, Mikhail Burtsev (NeurIPS).

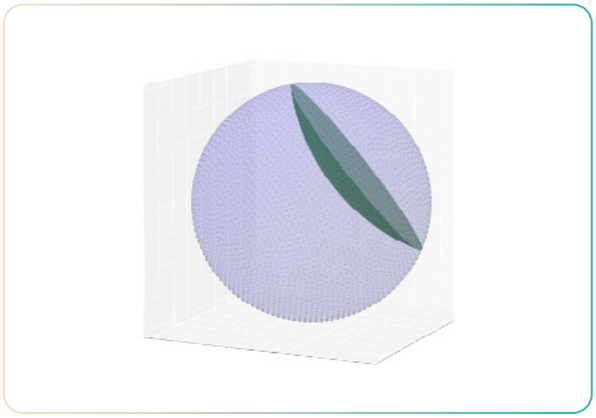
Several new algorithms, models and architectures for solving practical problems in recommender systems

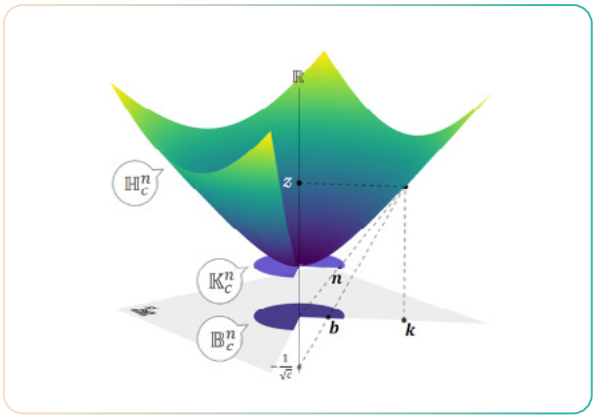
Evgeny Frolov
Research scientist

- A novel approach for scaling contemporary SOTA models to extensive product catalogs[2,3]. The solutions that have been developed represent a new milestone in the potential applications of previously proven transformer architectures in real-world problems. These solutions enable significant reductions in computational load during model training and concomitant improvements in model quality.
- A novel framework has been developed for integrating diverse sources of behavioral data within an end-to-end architecture for sequence learning [1]. The proposed solution's versatility offers significant opportunities for modeling heterogeneous sources of consumer behavior information within the framework of recommendation systems. Current efforts are underway to enhance the approach in highly dynamic environments.
- An effective cross-domain learning algorithm based on the ADMM[4], method has been developed. This algorithm facilitates the enhancement of recommendation systems by enabling the transfer of knowledge between different domains. This development is of particular practical interest in the context of marketplaces and online services, as it does not necessitate the transfer of personal consumer information in explicit form, while concurrently providing effective learning across multiple domains. Current efforts are underway to assess and expand the algorithm's capabilities for training on more than two domains.



→ A geometric approach founded on hyperbolic geometry was developed for training on sequences of user actions[5], thereby enabling the attainment of more compact models without loss of recommendation quality. The project was the first to develop expeditious methods for calculating the curvature of the input data space, scalable to real datasets. The results are already being utilized in related studies that require calculations on big data. It is planned to test the approach in the related area of large language models.





[1] A*: Baikalov, Vladimir; Frolov, Evgeny; “End-to-End Graph-Sequential Representation Learning for Accurate Recommendations”, Proceedings of the ACM on Web Conference 2024, 501-504, 2024.

[2] A: Gusak, Danil; Mezentsev, Gleb; Oseledets, Ivan; Frolov, Evgeny; “RECE: Reduced Cross-Entropy Loss for Large-Catalogue Sequential Recommenders”, Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 3772-3776, 2024.

[3] A: Mezentsev, Gleb; Gusak, Danil; Oseledets, Ivan; Frolov, Evgeny; “Scalable Cross-Entropy Loss for Sequential Recommendations with Large Item Catalogs”, Proceedings of the 18th ACM Conference on Recommender Systems, 475-485, 2024.

[4] A: Samra, Abdulaziz; Frolov, Evgeny; Vasilev, Alexey; Grigorevskiy, Alexander; Vakhrushev, Anton; “Cross-Domain Latent Factors Sharing via Implicit Matrix Factorization”, Proceedings of the 18th ACM Conference on Recommender Systems, 309-317, 2024.

[5] A: Frolov, Evgeny; Matveeva, Tatyana; Mirvakhabova, Leyla; Oseledets, Ivan; “Self-Attentive Sequential Recommendations with Hyperbolic Representations”, Proceedings of the 18th ACM Conference on Recommender Systems, 981-986, 2024.

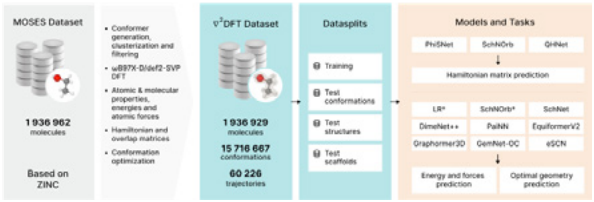
Applied results

A benchmark for the evaluation of LLM quality in the context of Russian language

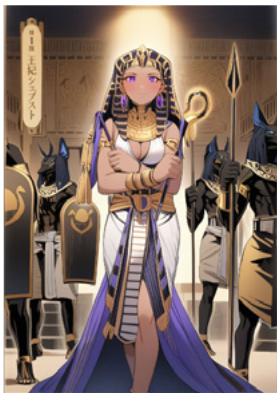
MERA: A Comprehensive LLM Evaluation in Russian. (ACL) In collaboration with colleagues from SberDevices and SberAI, the MERA benchmark was developed and published for the evaluation of large language models in tasks articulated in Russian. Furthermore, MathLogiQA and LogiQA tasks were meticulously crafted to assess the models' capacity to address mathematical and logical problems.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, et al.. 2024. MERA: A Comprehensive LLM Evaluation in Russian. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics

DFT: A Universal Quantum Chemistry Dataset of Drug-like Molecules and a Benchmark for Neural Network Potentials (NeurIPS D&B)



An approach for creating paired-end datasets is presented, offering a novel approach to creating a non-synthetic dataset for the manga colorization task. This methodology achieves state-of-the-art (SOTA) results. A benchmark was created for the AR Manga Colorization task using synthetic and real data. This benchmark was presented at ISMAR 2024 (A*) and WACV 2025 (A)



Golyadkin, M., Saraev, S., & Makarov, I. Benchmarking and Data Synthesis for Colorization of Manga Sequential Pages for Augmented Reality, ISMAR 2024.

Golyadkin, M., Plevokas, I., & Makarov, I. Closing the Domain Gap in Manga Colorization via Aligned Paired Dataset. WACV 2025

Optimization of neural networks for face descriptor extraction on mobile devices. The framework allows to adapt neural network architectures to specific devices taking into account their hardware limitations. The methodology includes the use of evolutionary algorithms and surrogate binary classifiers for fast sub-network selection

The development of device-optimized models is imperative to enhance the performance and privacy of mobile applications.

Savchenko, A., Maslov, D., & Makarov, I. (2024). Device-Specific Facial Descriptors: Winning a Lottery with a SuperNet. In ECAI 2024 (pp. 4439-4442)

ACL 2024 article “Fact-checking the output of large language models viatoken-level uncertainty quantification”, on the use of a new uncertainty estimation method for fact-checking the generation of LLMs.

Fadeeva, E., Rubashevskii, A., Shelmanov, A, Petrakov, S., Li, H., Mubarak, H., Tsymbalov, E., Kuzmin, G., Panchenko, A., Baldwin, T., Nakov, P., Panov, M. (2024): Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. In Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand. Association for Computational Linguistics.

An anatomical positional embedding for three-dimensional (3D) images that can accurately predict the anatomical position of regions opens new possibilities in image retrieval, organ localization, and generative modeling.These results were presented at the Medical Image Computing and Computer-Assisted Interventions (MICCAI) 2024 conference.

Three-dimensional medical images, such as computed tomography (CT) scans, are obtained by scanning a portion of a patient’s body. Typically, a substantial region is scanned (e.g., the entire chest or abdomen). The CT image is then sliced into three-dimensional patches, with each patch constituting a miniature CT image of a specific anatomical area.

Conventional positional patch embeddings encode the position of the patch relative to the image’s edge; however, given that the scanning area for different CT images may vary (one patient’s edge may correspond to the neck, while another’s may correspond to the diaphragm), conventional positional embeddings lack information regarding the anatomical position of the patches (the neck patch and the diaphragm patch will have the same conventional positional embedding).

In our work, “Anatomical Positional Embeddings” [1] (Core A), we proposed a model to produce data-driven embeddings that encode anatomical position information of patches (or pixels) of 3D medical images.The model is trained based on the premise that, in most patients, organs and anatomical regions are positioned relative to each other in a relatively consistent manner. The model is trained to predict 3D positional embeddings of patches based on their visual features, ensuring that the distances between embeddings correspond to the physical distances between patches in the original CT image. It has been demonstrated that embeddings trained in this manner are equivalent to 3D coordinates in a conventional coordinate system associated with the torso of an average patient.

Model applications:

- → Retrieval of images of a specific anatomical region.
- → Few-shot localization of organs.
- → Label-efficient classification of the anatomical region that contains the findings detected by other models (e.g., classification of pulmonary nodules into lobes and lung segments).
- → Tracking of findings on time-lapse images of a single patient.
- → Images registration.
- → Conditioning in generative and discriminative models.

Goncharov, M., Samokhin, V., Soboleva, E., et al. Anatomical Positional Embeddings. MICCAI 2024. Core A

A novel methodology for quantifying the aggregate uncertainty in the context of medical image classification and segmentation has been formulated, incorporating the insights of subject matter experts. (accepted at WACV'25)

A framework has been developed for testing hypotheses concerning the correspondence between people's emotional reactions and LLM decisions. The findings indicate that emotional alignment with humans is predominantly weak in most cases. However, emotional prompts, particularly negative ones, have been observed to significantly impact the behavior of models, leading to a reduction in cooperation in games and a decline in quality in ethical tasks. [EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas]

Mozikov, M., Severin, N., Bodishtianu, V., Glushanina, M., Nasonov, I., Orekhov, D., Pekhotin, V.,

Makovetskiy, I., Baklashkin, M., Lavrentyev, V., Tsvigun, A., Turdakov, D., Shavrina, T., Savchenko, A., &

Makarov, I. EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas. NeurIPS

New approaches to the application of optimal transport for machine learning tasks are developed.

- → This paper presents a reinforcement learning method for constructing autonomous agents that facilitates efficient learning on incomplete or noisy data.
- → Extending the neural optimal transport algorithm to knowledge transfer problems with class-preserving structure.
- → A method for solving the optimal transportation problem under data imbalance quickly and efficiently.

Rethinking Optimal Transport in Offline Reinforcement Learning
A Asadulaev, R Korst, A Korotin, V Egiazarian, A Filchenkov, E Burnaev
Neural Information Processing Systems 2024 (NeurIPS 2024)

Neural Optimal Transport with General Cost Functionals
A Asadulaev, A Korotin, V Egiazarian, P Mokrov, E Burnaev
The 12th International Conference on Learning Representations (ICLR 2024)

Light Unbalanced Optimal transport
M Gazdieva, A Asadulaev, E Burnaev, A Korotin
Neural Information Processing Systems 2024 (NeurIPS 2024)

Incomplete Reinforcement Learning
A Asadulaev, R Korst, A Korotin, V Egiazarian, A Filchenkov, E Burnaev
The 12th International Conference on Learning Representations (ICLR 2024 Workshop)

Publications

A* conference papers

5 AAAI	5 ACL	2 CVPR	6 ICLR
7 ICML	2 ISMAR	17 NeurIPS	1 ICRA
1 WWW	3 ACM KDD	2 ECCV	4 ICDM
4 IJCAI	7 EMNLP		

AAAI

Beyond Attention: Breaking the Limits of Transformer Context Length with Recurrent Memory

Aydar Bulatov, Yuri Kuratov, Yermek Kapushev, Mikhail Burtsev

A major limitation for the broader scope of problems solvable by transformers is the quadratic scaling of computational complexity with input size. In this study, we investigate the recurrent memory augmentation of pre-trained transformer models to extend input context length while linearly scaling compute. Our approach demonstrates the capability to store information in memory for sequences of up to an unprecedented two million tokens while maintaining high retrieval accuracy. Experiments with language modeling tasks show perplexity improvement as the number of processed input segments increases. These results underscore the effectiveness of our method, which has significant potential to enhance long-term dependency handling in natural language understanding and generation tasks, as well as enable large-scale context processing for memory-intensive applications.

Source

Topological and Node Noise Filtering on 3D Meshes Using Graph Neural Networks

Vladimir Mashurov, Natalia Semenova

Topological and node noise filtration are typically considered separately. Graph Neural Networks (GNN) are commonly used for node noise filtration, as they offer high efficiency and low exploitation costs. This paper explores the solution of joint node and topological noise filtration through the use of graph neural networks. Since treating a 3D mesh as a graph is challenging, an indicator function grid representation is employed as input for GNNs to perform the joint filtering. The resulting machine learning model is inspired by point cloud to mesh reconstruction algorithms and demonstrates low computational requirements during inference, producing successful results for smooth, watertight 3D models.

Source

AAAI

Improved Anonymous Multi-Agent Path Finding Algorithm

Zain Alabedeen Ali, Konstantin Yakovlev

We consider an Anonymous Multi-Agent Path-Finding (AMAPF) problem where the set of agents is confined to a graph, a set of goal vertices is given and each of these vertices has to be reached by some agent. The problem is to find an assignment of the goals to the agents as well as the collision-free paths, and we are interested in finding the solution with the optimal makespan. A well-established approach to solve this problem is to reduce it to a special type of a graph search problem, i.e. to the problem of finding a maximum flow on an auxiliary graph induced by the input one. The size of the former graph may be very large and the search on it may become a bottleneck. To this end, we suggest a specific search algorithm that leverages the idea of exploring the search space not through considering separate search states but rather bulks of them simultaneously. That is, we implicitly compress, store and expand bulks of the search states as single states, which results in high reduction in runtime and memory. Empirically, the resultant AMAPF solver...

Source

Learn to follow: Decentralized lifelong multi-agent pathfinding via planning and learning

Alexey Skrynnik, Anton Andreychuk, Maria Nesterova, Konstantin Yakovlev, Aleksandr Panov

Multi-agent Pathfinding (MAPF) problem generally asks to find a set of conflict-free paths for a set of agents confined to a graph and is typically solved in a centralized fashion. Conversely, in this work, we investigate the decentralized MAPF setting, when the central controller that posses all the information on the agents' locations and goals is absent and the agents have to sequientially decide the actions on their own without having access to a full state of the environment. We focus on the practically important lifelong variant of MAPF, which involves continuously assigning new goals to the agents upon arrival to the previous ones. To address this complex problem, we propose a method that integrates two complementary approaches: planning with heuristic search and reinforcement learning through policy optimization. Planning is utilized to construct and re-plan individual paths. We enhance our planning algorithm with a dedicated technique tailored to avoid congestion and increase the throughput of the system. We employ reinforcement learning to discover the collision avoidance policies that effectively guide the agents along the paths. The policy is implemented as a neural network and is effectively trained...

Source

AAAI

Decentralized Monte Carlo Tree Search for Partially Observable Multi-agent Pathfinding

Alexey Skrynnik, Anton Andreychuk, Konstantin Yakovlev, Aleksandr Panov

The Multi-Agent Pathfinding (MAPF) problem involves finding a set of conflict-free paths for a group of agents confined to a graph. In typical MAPF scenarios, the graph and the agents' starting and ending vertices are known beforehand, allowing the use of centralized planning algorithms. However, in this study, we focus on the decentralized MAPF setting, where the agents may observe the other agents only locally and are restricted in communications with each other. Specifically, we investigate the lifelong variant of MAPF, where new goals are continually assigned to the agents upon completion of previous ones. Drawing inspiration from the successful AlphaZero approach, we propose a decentralized multi-agent Monte Carlo Tree Search (MCTS) method for MAPF tasks. Our approach utilizes the agent's observations to recreate the intrinsic Markov decision process, which is then used for planning with a tailored for multi-agent tasks version of neural MCTS. The experimental results show that our approach outperforms state-of-the-art learnable MAPF solvers. The source code is available at this https URL: https://github.com/AIRI-Institute/mats-lp

A*

Source

ACL

Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov

Large language models (LLMs) are notorious for hallucinating, i.e., producing erroneousclaims in their output. Such hallucinations canbe dangerous, as occasional factual inaccuracies in the generated text might be obscuredby the rest of the output being generally factual, making it extremely hard for the usersto spot them. Current services that leverageLLMs usually do not provide any means fordetecting unreliable generations. Here, we aimto bridge this gap. In particular, we proposea novel fact-checking and hallucination detection pipeline based on token-level uncertaintyquantification. Uncertainty scores leverage information encapsulated in the output of a neuralnetwork or its layers to detect unreliable predictions, and we show that they can be usedto fact-check the atomic claims in the LLMoutput.

A*

Source

Your Transformer is Secretly Linear

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, Andrey Kuznetsov

This paper reveals a novel linear characteristic exclusive to transformer decoders, including models such as GPT, LLaMA, OPT, BLOOM and others. We analyze embedding transformations between sequential layers, uncovering a near-perfect linear relationship (Procrustes similarity score of 0.99). However, linearity decreases when the residual component is removed due to a consistently low output norm of the transformer layer. Our experiments show that removing or linearly approximating some of the most linear blocks of transformers does not affect significantly the loss or model performance. Moreover, in our pretraining experiments on smaller models we introduce a cosine-similarity-based regularization, aimed at reducing layer linearity. This regularization improves performance metrics on benchmarks like Tiny Stories and SuperGLUE and as well successfully decreases the linearity of the models. This study challenges the existing understanding of transformer architectures, suggesting that their operation may be more linear than previously assumed.

A*

Source

ACL

MERA: A Comprehensive LLM Evaluation in Russian

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton A. Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Denis Dimitrov, Alexander Panchenko, Sergey Markov

Over the past few years, one of the most notable002 advancements in AI research has been in foun003 dation models (FMs), headlined by the rise of004 language models (LMs). However, despite re005 searchers' attention and the rapid growth in LM006 application, the capabilities, limitations, and as007 sociated risks still need to be better understood.008 To address these issues, we introduce a new009 instruction benchmark, MERA, for evaluating010 foundation models oriented towards the Rus011 sian language. The benchmark encompasses 21012 evaluation tasks for generative models covering013 10 skills and is designed as a black-box test to014 ensure the exclusion of data leakage. The pa015 per introduces a methodology to evaluate FMs016 and LMs...

Source

TaxoLLaMA: WordNet-based Model for Solving Multiple Lexical Sematic Tasks

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, Irina Nikishina

In this paper, we explore the capabilities of LLMs in capturing lexical-semantic knowledge from WordNet on the example of the LLaMA-2-7b model and test it on multiple lexical semantic tasks. As the outcome of our experiments, we present TaxoLLaMA, the everything-in-one model, lightweight due to 4-bit quantization and LoRA. It achieves 11 SotA results, 4 top-2 results out of 16 tasks for the Taxonomy Enrichment, Hypernym Discovery, Taxonomy Construction, and Lexical Entailment tasks. Moreover, it demonstrates very strong zero-shot performance on Lexical Entailment and Taxonomy Construction with no fine-tuning. We also explore its hidden multilingual and domain adaptation capabilities with a little tuning or few-shot learning. All datasets, code, and model are available online at this https URL

Source

CVPR

The Devil is in the Details: StyleFeatureEditor for Detail-Rich StyleGAN Inversion and High Quality Image Editing

Denis Bobkov, Vadim Titov, Aibek Alanov, Dmitry Vetrov

The task of manipulating real image attributes through StyleGAN inversion has been extensively researched. This process involves searching latent variables from a welltrained StyleGAN generator that can synthesize a real image, modifying these latent variables, and then synthesizing an image with the desired edits. A balance must be struck between the quality of the reconstruction and the ability to edit. Earlier studies utilized the low-dimensional W-space for latent search, which facilitated effective editing but struggled with reconstructing intricate details. More recent research has turned to the high-dimensional feature space F, which successfully inverses the input image but loses much of the detail during editing. In this paper, we introduce StyleFeatureEditor– a novel method that enables editing in both w-latents and F-latents.

A*

Source

Weak-to-Strong 3D Object Detection with X-Ray Distillation

Alexander Gambashidze, Aleksandr Dadukin, Maksim Golyadkin, Maria Razzhivina, Ilya Makarov

This paper addresses the critical challenges of sparsity andocclusion in LiDAR-based 3Dobject detection. Current methods often rely on supplementary modules or specific architectural designs, potentially limiting their applicability to new and evolving architectures. To our knowledge, we are the first to propose a versatile technique that seamlessly integrates into any existing framework for 3D Object Detection, marking the first instance of Weak-to-Strong generalization in 3D computer vision. We introduce a novel framework, X-Ray Distillation with Object-Complete Frames, suitable for both supervised and semi-supervised settings, that leverages the temporal aspect of point cloud sequences. This methodextracts crucial information from both previous and subsequent LiDAR frames, creating Object-Complete frames that represent objects from multiple viewpoints, thus addressing occlusion and sparsity. Given the limitation of not being able to generate Object-Complete frames during online inference, we utilize Knowledge Distillation within a Teacher-Student framework.

A*

Source

ICLR

Gradual Optimization Learning for Conformational Energy Minimization

Artem Tsypin, Leonid Ugadiarov, Kuzma Khrabrov, Manvel Avetisian, Alexander Telepov, Egor Rumiantsev, Alexey Skrynnik, Aleksandr I Panov, Dmitry Vetrov, Elena Tutubalina, Artur Kadurin

Molecular conformation optimization is crucial to computer-aided drug discovery and materials design. Traditional energy minimization techniques rely on iterative optimization methods that use molecular forces calculated by a physical simulator (oracle) as anti-gradients. However, this is a computationally expensive approach that requires many interactions with a physical simulator. One way to accelerate this procedure is to replace the physical simulator with a neural network. Despite recent progress in neural networks for molecular conformation energy prediction, such models are prone to distribution shift, leading to inaccurate energy minimization. We find that the quality of energy minimization with neural networks can be improved by providing optimization trajectories as additional training data. Still, it takes around 5×105 additional conformations to match the physical simulator's optimization quality. In this work, we present the Gradual Optimization...

A*

Source

Light Schrödinger Bridge

Alexander Korotin, Nikita Gushchin, Evgeny Burnaev

Despite the recent advances in the field of computational Schrödinger Bridges (SB), most existing SB solvers are still heavy-weighted and require complex optimization of several neural networks. It turns out that there is no principal solver which plays the role of simple-yet-effective baseline for SB just like, e.g., -means method in clustering, logistic regression in classification or Sinkhorn algorithm in discrete optimal transport. We address this issue and propose a novel fast and simple SB solver. Our development is a smart combination of two ideas which recently appeared in the field: (a) parameterization of the Schrödinger potentials with sum-exp quadratic functions and (b) viewing the log-Schrödinger potentials as the energy functions. We show that combined together these ideas yield a lightweight, simulation-free and theoretically justified SB solver with a simple straightforward optimization objective. As a result, it allows solving SB in moderate dimensions in a matter of minutes on CPU without a painful hyperparameter selection. Our light solver resembles the Gaussian mixture model which is widely used for density estimation. Inspired by this similarity, we also prove an important theoretical result showing that our light solver is a universal approximator of SBs.

A*

Source

ICLR

Chemical Language Models Have Problems with Chemistry: A Case Study on Molecule Captioning Task

Veronika Ganeeva, Kuzma Khrabrov, Artur Kadurin, Andrey V. Savchenko, Elena Tutubalina

Drug discovery has been greatly enhanced through the recent fusion of molecular sciences and natural language processing, leading these research fields to significant advancements. Considering the crucial role of molecule representation in chemical understanding within these models, we introduce novel probing tests designed to evaluate chemical knowledge of molecular structure in state-of-the-art language models (LMs), specifically MolT5 and Chem+Text T5. These probing tests are conducted on a molecule captioning task to gather evidence and insights into the language models' comprehension of chemical information. By applying rules to transform molecular SMILES into equivalent variants, we have observed significant differences in the natural language descriptions generated by the LM for a given molecule depending on the exact transformation used.

Source

Energy-guided Entropic Neural Optimal Transport

Petr Mokrov, Alexander Korotin, Alexander Kolesov, Nikita Gushchin, Evgeny Burnaev

Energy-based models (EBMs) are known in the Machine Learning community for decades. Since the seminal works devoted to EBMs dating back to the noughties, there have been a lot of efficient methods which solve the generative modelling problem by means of energy potentials (unnormalized likelihood functions). In contrast, the realm of Optimal Transport (OT) and, in particular, neural OT solvers is much less explored and limited by few recent works (excluding WGAN-based approaches which utilize OT as a loss function and do not model OT maps themselves). In our work, we bridge the gap between EBMs and Entropy-regularized OT. We present a novel methodology which allows utilizing the recent developments and technical improvements of the former in order to enrich the latter. From the theoretical perspective, we prove generalization bounds for our technique. In practice, we validate its applicability in toy 2D and image domains. To showcase the scalability, we empower our method with a pre-trained StyleGAN and apply it to high-res AFHQ unpaired I2I translation.

Source

ICLR

Neural optimal transport with general cost functionals

Arip Asadulaev, Alexander Korotin, Vage Egiazarian, Petr Mokrov, Evgeny Burnaev

We introduce a novel neural network-based algorithm to compute optimal transport (OT) plans for general cost functionals. In contrast to common Euclidean costs, i.e., or , such functionals provide more flexibility and allow using auxiliary information, such as class labels, to construct the required transport map. Existing methods for general cost functionals are discrete and do not provide an out-of-sample estimation. We address the challenge of designing a continuous OT approach for general cost functionals in high-dimensional spaces, such as images. We construct two example functionals: one to map distributions while preserving the class-wise structure and the other one to preserve the given data pairs. Additionally, we provide the theoretical error analysis for our recovered transport plans. Our implementation is available at \url{https://github.com/machinestein/gnot}

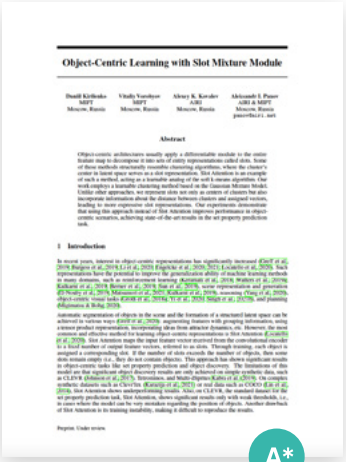
A*

Source

Object-Centric Learning with Slot Mixture Module

Daniil Kirilenko, Vitaliy Vorobyov, Alexey K. Kovalev, Aleksandr I. Panov

Object-centric architectures usually apply a differentiable module to the entire feature map to decompose it into sets of entity representations called slots. Some of these methods structurally resemble clustering algorithms, where the cluster's center in latent space serves as a slot representation. Slot Attention is an example of such a method, acting as a learnable analog of the soft k-means algorithm. Our work employs a learnable clustering method based on the Gaussian Mixture Model. Unlike other approaches, we represent slots not only as centers of clusters but also incorporate information about the distance between clusters and assigned vectors, leading to more expressive slot representations. Our experiments demonstrate that using this approach instead of Slot Attention improves performance in object-centric scenarios, achieving state-of-the-art results in the set property prediction task.



A*

Source

ICML

Estimating Barycenters of Distributions with Neural Optimal Transport

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, Alexander Korotin

Given a collection of probability measures, a practitioner sometimes needs to find an “average” distribution which adequately aggregates reference distributions. A theoretically appealing notion of such an average is the Wasserstein barycenter, which is the primal focus of our work. By building upon the dual formulation of Optimal Transport (OT), we propose a new scalable approach for solving the Wasserstein barycenter problem. Our methodology is based on the recent Neural OT solver: it has bi-level adversarial learning objective and works for general cost functions. These are key advantages of our method, since the typical adversarial algorithms leveraging barycenter tasks utilize tri-level optimization and focus mostly on quadratic cost. We also establish theoretical error bounds for our proposed approach and showcase its applicability and effectiveness on illustrative scenarios and image data setups.

Source

Neural operators meet conjugate gradients: The FCG-NO method for efficient PDE solving

Alexander Rudikov, Vladimir Fanaskov, Ekaterina Muravleva, Yuri M Laevsky, Ivan Oseledets

Deep learning solvers for partial differential equations typically have limited accuracy. We propose to overcome this problem by using themas preconditioners. More specifically, we apply discretization-invariant neural operators tolearn preconditioners for the flexible conjugategradient method (FCG). Architecture paired withnovel loss function and training scheme allowsfor learning efficient preconditioners that can beused across different resolutions. On the theoretical side, FCG theory allows us to safely usenonlinear preconditioners that can be applied inO(N) operations without constraining the formof the preconditioners matrix. To justify learningscheme components (the loss function and theway training data is collected) we perform severalablation studies. Numerical results indicate thatour approach favorably compares with classicalpreconditioners and allows to reuse of preconditioners learned for lower resolution to the higherresolution data.

Source

ICML

In-Context Reinforcement Learning for Variable Action Spaces

Viacheslav Sinii, Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, Sergey Kolesnikov

Recently, it has been shown that transformers pre-trained on diverse datasets with multi-episode contexts can generalize to new reinforcement learning tasks in-context. A key limitation of previously proposed models is their reliance on a predefined action space size and structure. The introduction of a new action space often requires data re-collection and model re-training, which can be costly for some applications. In our work, we show that it is possible to mitigate this issue by proposing the Headless-AD model that, despite being trained only once, is capable of generalizing to discrete action spaces of variable size, semantic content and order. By experimenting with Bernoulli and contextual bandits, as well as a gridworld environment, we show that Headless-AD exhibits significant capability to generalize to action spaces it has never encountered, even outperforming specialized models trained for a specific set of actions on several environment configurations.

Source

Emergence of In-Context Reinforcement Learning from Noise Distillation

Ilya Zisman, Vladislav Kurenkov, Alexander Nikulin, Viacheslav Sinii, Sergey Kolesnikov

Recently, extensive studies in Reinforcement Learning have been carried out on the ability of transformers to adapt in-context to various environments and tasks. Current in-context RL methods are limited by their strict requirements for data, which needs to be generated by RL agents or labeled with actions from an optimal policy. In order to address this prevalent problem, we propose ADε, a new data acquisition approach that enables in-context Reinforcement Learning from noise-induced curriculum. We show that it is viable to construct a synthetic noise injection curriculum which helps to obtain learning histories. Moreover, we experimentally demonstrate that it is possible to alleviate the need for generation using optimal policies, with in-context RL still able to outperform the best suboptimal policy in a learning dataset by a 2x margin.

Source

ICML

Disentanglement Learning via Topology

Nikita Balabin, Daria Voronkova, Ilya Trofimov, Evgeny Burnaev, Serguei Barannikov

We propose TopDis (Topological Disentanglement), a method for learning disentangled representations via adding a multi-scale topological loss term. Disentanglement is a crucial property of data representations substantial for the explainability and robustness of deep learning models and a step towards high-level cognition. The state-of-the-art methods are based on VAE and encourage the joint distribution of latent variables to be factorized. We take a different perspective on disentanglement by analyzing topological properties of data manifolds. In particular, we optimize the topological similarity for data manifolds traversals. To the best of our knowledge, our paper is the first one to propose a differentiable topological loss for disentanglement learning. Our experiments have shown that the proposed TopDis loss improves disentanglement scores such as MIG, FactorVAE score, SAP score, and DCI disentanglement score with respect to state-of-the-art results...

A*

Source

Self-Supervised Coarsening of Unstructured Grid with Automatic Differentiation

Sergei Shumilin, Alexander Ryabov, Nikolay Yavich, Evgeny Burnaev, Vladimir Vanovskiy

Due to the high computational load of modern numerical simulation,there is a demand for approaches that would reduce the sizeof discrete problems while keeping the accuracy reasonable.In this work, we present an original algorithmto coarsen an unstructured grid basedon the concepts of differentiablephysics. We achieve thisby employing k-means clustering, autodifferentiation andstochastic minimization algorithms.We demonstrate performance of the designed algorithm on a linear parabolic equation which governsslightly compressible fluid flow in porous media.Our results show that in the considered scenarios, we reduced the number of grid points up to 10 times while preserving the modeled variable dynamics in the points of interest.The proposed approach can be applied to simulation of an arbitrary system described by evolutionary partial differential equations.

A*

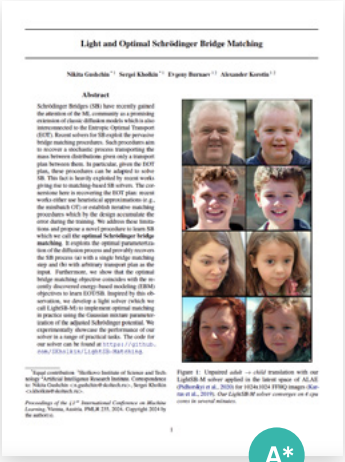
Source

ICML

Light and Optimal Schrödinger Bridge Matching

Nikita Gushchin, Sergei Kholkin, Evgeny Burnaev, Alexander Korotin

Schrödinger Bridges (SB) have recently gained the attention of the ML community as a promising extension of classic diffusion models whichis also interconnected to the Entropic OptimalTransport (EOT). Recent solvers for SB exploitthe pervasive bridge matching procedures. Suchprocedures aim to recover a stochastic processtransporting the mass between distributions givenonly a transport plan between them. In particular, given the EOT plan, these procedures can beadapted to solve SB. This fact is heavily exploitedby recent works giving rives to matching-basedSB solvers. The cornerstone here is recoveringthe EOT plan: recent works either use heuristical approximations (e.g., the minibatch OT) orestablish iterative matching procedures which bythe design accumulate the error during the training. We address these limitations and propose anovel procedure to learn SB which we call theoptimal Schrödinger bridge matching. It exploits the optimal parameterization of the diffusion process and provably recovers the SB process (a) with a single bridge matching step and(b) with arbitrary transport plan as the input.



Source

ISMAR

Benchmarking and Data Synthesis for Colorization of Manga SequentialPages for Augmented Reality

Maksim Golyadkin, Sergey Saraev, Ilya Makarov

This paper introduces an innovative approach to manga colorizationwithin augmented reality (AR) environments, focusing on theunique challenges posed by colorizing photos of manga books.We present a novel method using diffusion models to generatea synthetic dataset that accurately replicates photographed mangapages. Additionally, we have compiled a dataset of real manga photographs,capturing diverse environmental conditions. Integratingthese datasets, we established a comprehensive benchmark to evaluatecolorization models in scenarios that simulate AR applications.This benchmark was validated through a human study, confirmingthe accuracy of our metrics across both datasets. We also showedthat domain adaptation may improve model performance. Pavingthe way for practical applications, our framework enables...

Source

Pose Networks Unveiled: Bridging the Gap for Monocular Depth Perception

Yazan Dayoub, Anrey V. Savchenko, Ilya Makarov

Depth estimation is essential in Augmented Reality applications,enabling realistic object placement, scene understanding, spatialmapping, interaction, and environment awareness. This paper proposesa method to enhance depth model performance without increasinginference costs by improving the pose network in a selfsupervisedlearning setup. In particular, we enrich spatial informationin the pose network by incorporating features from differentscales and normalized coordinates. It is experimentally shown onthe KITTI dataset that our approach achieves a 2-7% improvementin the abs rel metric when compared to baseline techniques.

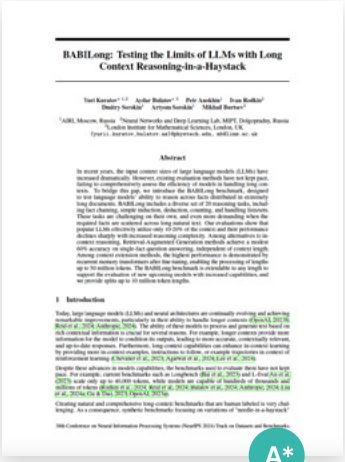
Source

NeurIPS

BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, Mikhail Burtsev

In recent years, the input context sizes of large language models (LLMs) have increased dramatically. However, existing evaluation methods have not kept pace, failing to comprehensively assess the efficiency of models in handling long contexts. To bridge this gap, we introduce the BABILong benchmark, designed to test language models' ability to reason across facts distributed in extremely long documents. BABILong includes a diverse set of 20 reasoning tasks, including fact chaining, simple induction, deduction, counting, and handling lists/sets. These tasks are challenging on their own, and even more demanding when the required facts are scattered across long natural text. Our evaluations show that popular LLMs effectively utilize only 10-20\% of the context and their performance declines sharply with increased reasoning complexity.



Source

ENOT: Expectile Regularization for Fast and Accurate Training of Neural Optimal Transport

Nazar Buzun, Maksim Bobrin, Dmitry V. Dylov

We present a new extension for Neural Optimal Transport (NOT) training procedure, capable of accurately and efficiently estimating optimal transportation plan via specific regularisation on conjugate potentials. The main bottleneck of existing NOT solvers is associated with the procedure of finding a near-exact approximation of the conjugate operator (i.e., the c-transform), which is done either by optimizing over maximin objectives or by the computationally-intensive fine-tuning of the initial approximated prediction. We resolve both issues by proposing a new, theoretically justified loss in the form of expectile regularization that enforces binding conditions on the learning dual potentials. Such a regularization provides the upper bound estimation over the distribution of possible conjugate potentials and makes the learning stable, eliminating the need for additional extensive finetuning. We formally justify the efficiency of our method, called Expectile-Regularised Neural Optimal Transport (ENOT). ENOT outperforms previous state-of-the-art approaches on the Wasserstein-2 benchmark tasks by a large margin (up to a 3-fold improvement in quality and up to a 10-fold improvement in runtime).

Source

NeurIPS

Light Unbalanced Optimal Transport

Milena Gazdieva, Arip Asadulaev, Alexander Korotin, Evgeny Burnaev

While the continuous Entropic Optimal Transport (EOT) field has been actively developing in recent years, it became evident that the classic EOT problem is prone to different issues like the sensitivity to outliers and imbalance of classes in the source and target measures. This fact inspired the development of solvers that deal with the unbalanced EOT (UEOT) problem – the generalization of EOT allowing for mitigating the mentioned issues by relaxing the marginal constraints. Surprisingly, it turns out that the existing solvers are either based on heuristic principles or heavy-weighted with complex optimization objectives involving several neural networks. We address this challenge and propose a novel theoretically-justified, lightweight, unbalanced EOT solver. Our advancement consists of developing a novel view on the optimization of the UEOT problem yielding tractable and a non-minimax optimization objective. We show that combined with a light parametrization recently proposed in the field our objective leads to a fast, simple, and effective solver which allows solving the continuous UEOT problem in minutes on CPU.

A*

Source

∇²DFT: A Universal Quantum Chemistry Dataset of Drug-Like Molecules and a Benchmark for Neural Network Potentials

Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin, Egor Rumiantsev, Alexander Telepov, Dmitry Protasov, Ilya Shenbin, Anton Alekseev, Mikhail Shirokikh, Sergey Nikolenko, Elena Tutubalina, Artur Kadurin

Methods of computational quantum chemistry provide accurate approximations of molecular properties crucial for computer-aided drug discovery and other areas of chemical science. However, high computational complexity limits the scalability of their applications. Neural network potentials (NNPs) are a promising alternative to quantum chemistry methods, but they require large and diverse datasets for training. This work presents a new dataset and benchmark called ∇²DFT that is based on the nablaDFT. It contains twice as much molecular structures, three times more conformations, new data types and tasks, and state-of-the-art models. The dataset includes energies, forces, 17 molecular properties, Hamiltonian and overlap matrices, and a wavefunction object.

A*

Source

NeurIPS

Rethinking Optimal Transport in Offline Reinforcement Learning

Arip Asadulaev, Alexander Korotin, Vage Egiazarian, Rostislav Korst, Andrey Filchenkov, Evgeny Burnaev

We present a novel approach for offline reinforcement learning that bridges the gap between recent advances in neural optimal transport and reinforcement learning algorithms. Our key idea is to compute the optimal transport between states and actions with an action-value cost function and implicitly recover an optimal map that can serve as a policy. Building on this concept, we develop a new algorithm called Extremal Monge Reinforcement Learning that treats offline reinforcement learning as an extremal optimal transport problem. Unlike previous transport-based offline reinforcement learning algorithms, our method focuses on improving the policy beyond the behavior policy, rather than addressing the distribution shift problem. We evaluated the performance of our method on various continuous control problems and demonstrated improvements over existing algorithms.

Source

RClicks: Realistic Click Simulation for Benchmarking Interactive Segmentation

Anton Antonov, Andrey Moskalenko, Denis Shepelev, Vlad Shakhuro, Alexander Krapukhin, Konstantin Soshin, Anton Konushin

The emergence of Segment Anything (SAM) sparked research interest in the field of interactive segmentation, especially in the context of image editing tasks and speeding up data annotation. Unlike common semantic segmentation, interactive segmentation methods allow users to directly influence their output through prompts (e.g. clicks). However, click patterns in real-world interactive segmentation scenarios remain largely unexplored. Most methods rely on the assumption that users would click in the center of the largest erroneous area. Nevertheless, recent studies show that this is not always the case. Thus, methods may have poor performance in real-world deployment despite high metrics in a baseline benchmark. To accurately simulate real-user clicks, we conducted a large crowdsourcing study of click patterns in an interactive segmentation scenario and collected 475K real-user clicks.

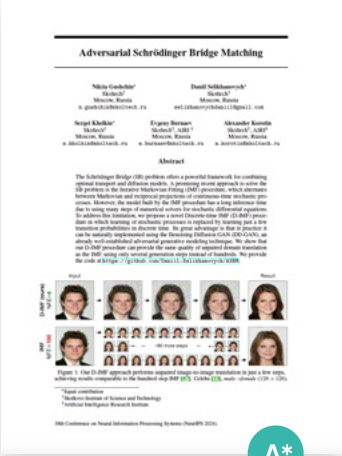
Source

NeurIPS

Adversarial Schrödinger Bridge Matching

Nikita Gushchin, Daniil Selikhanovych, Sergei Kholkin, Evgeny Burnaev, Alexander Korotin

The Schrödinger Bridge (SB) problem offers a powerful framework for combining optimal transport and diffusion models. A promising recent approach to solve the SB problem is the Iterative Markovian Fitting (IMF) procedure, which alternates between Markovian and reciprocal projections of continuous-time stochastic processes. However, the model built by the IMF procedure has a long inference time due to using many steps of numerical solvers for stochastic differential equations. To address this limitation, we propose a novel Discrete-time IMF (D-IMF) procedure in which learning of stochastic processes is replaced by learning just a few transition probabilities in discrete time. Its great advantage is that in practice it can be naturally implemented using the Denoising Diffusion GAN (DD-GAN), an already well-established adversarial generative modeling technique. We show that our D-IMF procedure can provide the same quality of unpaired domain translation as the IMF, using only several generation steps instead of hundreds.



A*

Source

EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Vladislav Pekhotin, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, Akim Tsvigun, Denis Turdakov, Tatiana Shavrina, Andrey Savchenko, Ilya Makarov

One of the urgent tasks of artificial intelligence is to assess the safety and alignment of large language models (LLMs) with human behavior. Conventional verification only in pure natural language processing benchmarks can be insufficient. Since emotions often influence human decisions, this paper examines LLM alignment in complex strategic and ethical environments, providing an in-depth analysis of the drawbacks of our psychology and the emotional impact on decision-making in humans and LLMs. We introduce the novel EAI framework for integratingemotion modeling into LLMs to examine the emotional impact on ethics and LLM-based decision-making in various strategic games, including bargaining and repeated games. Our experimental study with various LLMs demonstratedthat emotions can significantly alter the ethical decision-making...

A*

Source

NeurIPS

Energy-Guided Continuous Entropic Barycenter Estimation for General Costs

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, Alexander Korotin

Optimal transport (OT) barycenters are a mathematically grounded way of averaging probability distributions while capturing their geometric properties. In short, the barycenter task is to take the average of a collection of probability distributions w.r.t. given OT discrepancies. We propose a novel algorithm for approximating the continuous Entropic OT (EOT) barycenter for arbitrary OT cost functions. Our approach is built upon the dual reformulation of the EOT problem based on weak OT, which has recently gained the attention of the ML community. Beyond its novelty, our method enjoys several advantageous properties: (i) we establish quality bounds for the recovered solution; (ii) this approach seamlessly interconnects with the Energy-Based Models (EBMs) learning procedure enabling the use of well-tuned algorithms for the problem of interest; (iii) it provides an intuitive optimization scheme...

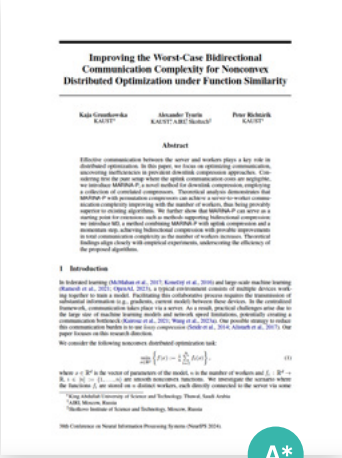
A*

Source

Improving the Worst-Case Bidirectional Communication Complexity for Nonconvex Distributed Optimization under Function Similarity

Kaja Gruntkowska, Alexander Tyurin, Peter Richtárik

Effective communication between the server and workers plays a key role in distributed optimization. In this paper, we focus on optimizing the server-to-worker communication, uncovering inefficiencies in prevalent downlink compression approaches. Considering first the pure setup where the uplink communication costs are negligible, we introduce MARINA-P, a novel method for downlink compression, employing a collection of correlated compressors. Theoretical analyses demonstrates that MARINA-P with permutation compressors can achieve a server-to-worker communication complexity improving with the number of workers, thus being provably superior to existing algorithms. We further show that MARINA-P can serve as a starting point for extensions such as methods supporting bidirectional compression. We introduce M3, a method combining MARINA-P with uplink compression and a momentum step, achieving bidirectional compression with provable improvements in total communication complexity as the number of workers increases. Theoretical findings align closely with empirical experiments, underscoring the efficiency of the proposed algorithms.



A*

Source

NeurIPS

XLand-MiniGrid: Scalable Meta-Reinforcement Learning Environments in JAX

Alexander Nikulin, Vladislav Kurenkov, Ilya Zisman, Artem Agarkov, Viacheslav Sinii, Sergey Kolesnikov

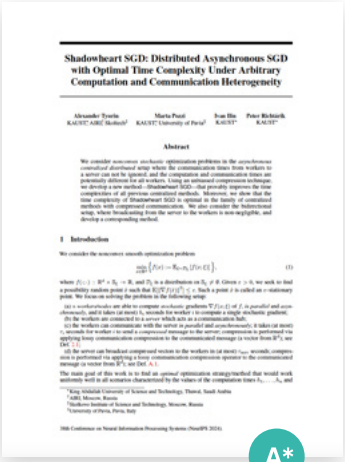
Inspired by the diversity and depth of XLand and the simplicity and minimalism of MiniGrid, we present XLand-MiniGrid, a suite of tools and grid-world environments for meta-reinforcement learning research. Written in JAX, XLand-MiniGrid is designed to be highly scalable and can potentially run on GPU or TPU accelerators, democratizing large-scale experimentation with limited resources. Along with the environments, XLand-MiniGrid provides pre-sampled benchmarks with millions of unique tasks of varying difficulty and easy-to-use baselines that allow users to quickly start training adaptive agents. In addition, we have conducted a preliminary analysis of scaling and generalization, showing that our baselines are capable of reaching millions of steps per second during training and validating that the proposed benchmarks are challenging.

Source

Shadowheart SGD: Distributed Asynchronous SGD with Optimal Time Complexity Under Arbitrary Computation and Communication Heterogeneity

Alexander Tyurin, Marta Pozzi, Ivan Ilin, Peter Richtárik

We consider nonconvex stochastic optimization problems in the asynchronous centralized distributed setup where the communication times from workers to a server can not be ignored, and the computation and communication times are potentially different for all workers. Using an unbiassed compression technique, we develop a new method-Shadowheart SGD-that provably improves the time complexities of all previous centralized methods. Moreover, we show that the time complexity of Shadowheart SGD is optimal in the family of centralized methods with compressed communication. We also consider the bidirectional setup, where broadcasting from the server to the workers is non-negligible, and develop a corresponding method.



Source

NeurIPS

Optimal Flow Matching: Learning Straight Trajectories in Just One Step

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, Alexander Korotin

Over the several recent years, there has been a boom in development of Flow Matching (FM) methods for generative modeling. One intriguing property pursued by the community is the ability to learn flows with straight trajectories which realize the Optimal Transport (OT) displacements. Straightness is crucial for the fast integration (inference) of the learned flow's paths. Unfortunately, most existing flow straightening methods are based on non-trivial iterative FM procedures which accumulate the error during training or exploit heuristics based on minibatch OT. To address these issues, we develop and theoretically justify the novel Optimal Flow Matching approach which allows recovering the straight OT displacement for the quadratic transport in just one FM step. The main idea of our approach is the employment of vector field for FM which are parameterized by convex functions.

Source

On the Optimal Time Complexities in Decentralized Stochastic Asynchronous Optimization

Alexander Tyurin, Peter Richtárik

We consider the decentralized stochastic asynchronous optimization setup, where many workers asynchronously calculate stochastic gradients and asynchronously communicate with each other using edges in a multigraph. For both homogeneous and heterogeneous setups, we prove new time complexity lower bounds under the assumption that computation and communication speeds are bounded. We develop a new nearly optimal method, Fragile SGD, and a new optimal method, Amelie SGD, that converge under arbitrary heterogeneous computation and communication speeds and match our lower bounds (up to a logarithmic factor in the homogeneous setting). Our time complexities are new, nearly optimal, and provably improve all previous asynchronous/synchronous stochastic methods in the decentralized setup.

Source

NeurIPS

Freya PAGE: First Optimal Time Complexity for Large-Scale Nonconvex Finite-Sum Optimization with Heterogeneous Asynchronous Computations

Alexander Tyurin, Kaja Gruntkowska, Peter Richtárik

In practical distributed systems, workers are typically not homogeneous, and due to differences in hardware configurations and network conditions, can have highly varying processing times. We consider smooth nonconvex finite-sum (empirical risk minimization) problems in this setup and introduce a new parallel method, Freya PAGE, designed to handle arbitrarily heterogeneous and asynchronous computations. By being robust to “stragglers” and adaptively ignoring slow computations, Freya PAGE offers significantly improved time complexity guarantees compared to all previous methods, including Asynchronous SGD, Rennala SGD, SPIDER, and PAGE, while requiring weaker assumptions. The algorithm relies on novel generic stochastic

A*

Source

HairFastGAN: Realistic and Robust Hair Transfer with a Fast Encoder-Based Approach

Maxim Nikolaev, Mikhail Kuznetsov, Dmitry P. Vetrov, Aibek Alanov

Our paper addresses the complex task of transferring a hairstylefrom a reference image to an input photo for virtual hair try-on. This taskis challenging due to the need to adapt to various photo poses, the sensitivity of hairstyles, and the lack of objective metrics. The current stateof the art hairstyle transfer methods use an optimization process for different parts of the approach, making them inexcusably slow. At the sametime, faster encoder-based models are of very low quality because theyeither operate in StyleGAN’s W+ space or use other low-dimensionalimage generators. Additionally, both approaches have a problem withhairstyle transfer when the source pose is very different from the targetpose, because they either don’t consider the pose at all or deal with it inefficiently. In our paper, we present the HairFast model, which uniquelysolves these problems and achieves high resolution, near real-time performance, and superior reconstruction compared to optimization problembased methods.

A*

Source

NeurIPS

Group and Shuffle: Efficient Structured Orthogonal Parametrization

Mikhail Gorbunov, Nikolay Yudin, Vera Soboleva, Aibek Alanov, Alexey Naumov, Maxim Rakhuba

The increasing size of neural networks has led to a growing demand for methods of efficient fine-tuning. Recently, an orthogonal fine-tuning paradigm was introduced that uses orthogonal matrices for adapting the weights of a pretrained model. In this paper, we introduce a new class of structured matrices, which unifies and generalizes structured classes from previous works. We examine properties of this class and build a structured orthogonal parametrization upon it. We then use this parametrization to modify the orthogonal fine-tuning framework, improving parameter and computational efficiency. We empirically validate our method on different domains, including adapting of text-to-image diffusion models and downstream task fine-tuning in language modeling. Additionally, we adapt our construction for orthogonal convolutions and conduct experiments with 1-Lipschitz neural networks.

Source

ICRA

Neural Potential Field for Obstacle-Aware Local Motion Planning

Muhammad Alhaddad, Konstantin Mironov, Aleksey Staroverov, Aleksandr Panov

Model predictive control (MPC) may provide local motion planning for mobile robotic platforms. The challenging aspect is the analytic representation of collision cost for the case when both the obstacle map and robot footprint are arbitrary. We propose a Neural Potential Field: a neural network model that returns a differentiable collision cost based on robot pose, obstacle map, and robot footprint. The differentiability of our model allows its usage within the MPC solver.
It is computationally hard to solve problems with a very high number of parameters. Therefore, our architecture includes neural image encoders, which transform obstacle maps and robot footprints into embeddings, which reduce problem dimensionality by two orders of magnitude. The reference data for network training are generated based on algorithmic calculation of a signed distance function. Comparative experiments showed that the proposed approach is comparable with existing local planners: it provides trajectories with outperforming smoothness, comparable path length, and safe distance from obstacles. Experiment on Husky UGV mobile robot showed that our approach allows real-time and safe local planning. The code for our approach is presented at https://github.com/cog-isa/NPField together with demo video.

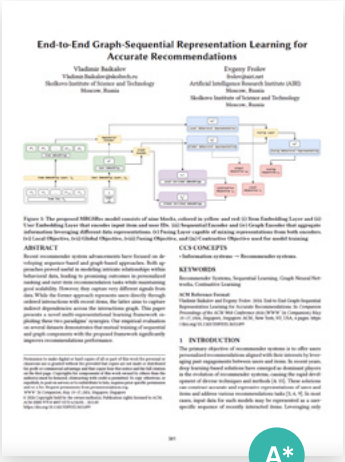
Source

WWW

End-to-End Graph-Sequential Representation Learning for Accurate Recommendations

Vladimir Baikalov, Evgeny Frolov

Recent recommender system advancements have focused on developing sequence-based and graph-based approaches. Both approaches proved useful in modeling intricate relationships within behavioral data, leading to promising outcomes in personalized ranking and next-item recommendation tasks while maintaining good scalability. However, they capture very different signals from data. While the former approach represents users directly through ordered interactions with recent items, the latter aims to capture indirect dependencies across the interactions graph. This paper presents a novel multi-representational learning framework exploiting these two paradigms' synergies. Our empirical evaluation on several datasets demonstrates that mutual training of sequential and graph components with the proposed framework significantly improves recommendations performance.



Source

ACM KDD

Uplift Modelling via Gradient Boosting

Bulat Ibragimov, Anton Vakhrushev

The Gradient Boosting machine learning ensemble algorithm, well-known for its proficiency and superior performance in intricate machine learning tasks, has encountered limited success in the realm of uplift modeling. Uplift modeling is a challenging task that necessitates a known target for the precise computation of the training gradient. The prevailing two-model strategies, which separately model treatment and control outcomes, are encumbered with limitations as they fail to directly tackle the uplift problem.This paper presents an innovative approach to uplift modeling that employs Gradient Boosting. Unlike previous works, our algorithm utilizes multioutput boosting model and calculates the uplift gradient based on intermediate surrogate predictions and directly models the concealed target. This method circumvents the requirement for a known target and addresses the uplift problem more effectively than existing solutions.Moreover, we broaden the scope of this solution to encompass multitreatment settings, thereby enhancing its applicability. This novel approach not only overcomes the limitations of the traditional two-model strategies...

Source

From Variability to Stability: Advancing RecSys Benchmarking Practices

Valeriy Shevchenko, Nikita Belousov, Alexey Vasilev, Vladimir Zholobov, Artyom Sosedka, Natalia Semenova, Anna Volodkevich, Andrey Savchenko, Alexey Zaytsev

In the rapidly evolving domain of Recommender Systems (RecSys),new algorithms frequently claim state-of-the-art performance basedon evaluations over a limited set of arbitrarily selected datasets.However, this approach may fail to holistically reflect their effectiveness due to the significant impact of dataset characteristicson algorithm performance. Addressing this deficiency, this paperintroduces a novel benchmarking methodology to facilitate a fairand robust comparison of RecSys algorithms, thereby advancingevaluation practices. By utilizing a diverse set of 30 open datasets,including two introduced in this work, and evaluating 11 collaborative filtering algorithms across 9 metrics, we critically examinethe influence of dataset characteristics on algorithm performance. We further investigate the feasibility of aggregating outcomes frommultiple datasets into a unified ranking. Through rigorous experimental analysis, we validate the reliability of our methodology under the variability of datasets, offering a benchmarking strategy thatbalances quality and computational demands. This methodologyenables a fair yet effective means of evaluating RecSys algorithms,providing valuable guidance for future research endeavors.

Source

ACM KDD

Learn Together Stop Apart: an Inclusive Approach to Ensemble Pruning

Bulat Ibragimov; Gleb Gusev

Gradient Boosting is a leading learning method that builds ensembles and adapts their sizes to particular tasks, consistently delivering top-tier results across various applications. However, determining the optimal number of models in the ensemble remains a critical yet underexplored aspect. Traditional approaches assume a universal ensemble size effective for all data points, which may not always hold true due to data heterogeneity.This paper introduces an adaptive approach to early stopping in Gradient Boosting, addressing data heterogeneity by assigning different stop moments to different data regions at inference time while still training a common ensemble on the entire dataset. We propose two methods: Direct Supervised Partition (DSP) and Indirect Supervised Partition (ISP). The DSP method uses a decision tree to partition the data based on learning curves, while ISP leverages the dataset's geometric and target distribution characteristics.An effective validation protocol is developed to determine the optimal number of early stopping regions or detect when the heterogeneity assumption does not hold. Experiments using state-of-the-art implementations of Gradient Boosting, LightGBM, and CatBoost, on standard benchmarks demonstrate that our methods enhance model precision by up to 2%, underscoring the significance of this research direction. This approach does not increase computational complexity and can be easily integrated into existing learning pipelines.

Source

ECCV

Scalar Function Topology Divergence: Comparing Topology of 3D Objects

Ilya Trofimov, Daria Voronkova, Eduard Tulchinskii, Evgeny Burnaev, Serguei Barannikov

We propose a new topological tool for computer visionScalar Function Topology Divergence (SFTD), which measures the dissimilarity of multi-scale topology between sublevel sets of two functions having a common domain. Functions can be defined on an undirected graph or Euclidean space of any dimensionality. Most of the existing methods for comparing topology are based on Wasserstein distance between persistence barcodes and they don't take into account the localization of topological features. On the other hand, the minimization of SFTD ensures that the corresponding topological features of scalar functions are located in the same places. The proposed tool provides useful visualizations depicting areas where functions have topological dissimilarities. We provide applications of the proposed method to 3D computer vision. In particular, experiments demonstrate that SFTD improves the reconstruction of cellular 3D shapes from 2D fluorescence microscopy images, and helps to identify topological errors in 3D segmentation.

Source

Guide-and-Rescale: Self-Guidance Mechanism for Effective Tuning-Free Real Image Editing

Vadim Titov, Madina Khalmatova, Alexandra Ivanova, Dmitry Vetrov, and Aibek Alanov

Despite recent advances in large-scale text-to-image generative models, manipulating real images with these models remains a challenging problem. The main limitations of existing editing methods are that they either fail to perform with consistent quality on a wide range of image edits or require time-consuming hyperparameter tuning or fine-tuning of the diffusion model to preserve the image-specific appearance of the input image. We propose a novel approach that is built upon a modified diffusion sampling process via the guidance mechanism. In this work, we explore the self-guidance technique to preserve the overall structure of the input image and its local regions appearance that should not be edited. In particular, we explicitly introduce layout-preserving energy functions that are aimed to save local and global structures of the source image. Additionally, we propose a noise rescaling mechanism...

Source

ICDM

LLM-KT: A Versatile Framework for Knowledge Transfer from Large Language Models to Collaborative Filtering

Nikita Severin, Aleksei Ziablitsev, Yulia Savelyeva, Valeriy Tashchilin, Ivan Bulychev, Mikhail Yushkov, Artem Kushneruk, Amaliya Zaryvnykh, Dmitrii Kiselev, Andrey Savchenko, Ilya Makarov

A*

We present LLM-KT, a flexible framework designed to enhance collaborative filtering (CF) models by seamlessly integrating LLM (Large Language Model)-generated features. Unlike existing methods that rely on passing LLM-generated features as direct inputs, our framework injects these features into an intermediate layer of any CF model, allowing the model to reconstruct and leverage the embeddings internally. This model-agnostic approach works with a wide range of CF models without requiring architectural changes, making it adaptable to various recommendation scenarios. Our framework is built for easy integration and modification, providing researchers and developers with a powerful tool for extending...

Source

Go-Kart Racing Simulator for Reinforcement Learning with Augmented Sim2Real Adaptation

Ildar Nurgaliev, Andre Kuzminykh, Andrey Savchenko, and Ilya Makarov

A*

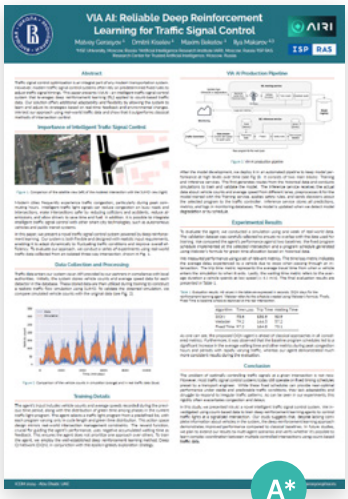
Training self-driving cars in real-world scenarios is inefficient due to the possibility of crashes with obstacles and borders. This paper introduces the virtual environment to enhance reinforcement learning training in a virtual Go-Kart racing simulator. The primary objective is to leverage augmented reality to enhance observations inside the simulation, improve policy networks, and make the Value function precise and robust. We develop the wrapper for the CARLA simulator, enabling a cost-effective sim2real transition. It is demonstrated that the augmented sim2real adaptation successfully integrates simulated training outcomes into real-world scenarios where the real Go-Kart can accomplish six laps in a single-race mode reaching the maximum speed of 11.5 m/s.

ICDM

VIA AI: Reliable Deep Reinforcement Learning for Traffic Signal Control

Matvey Gerasyov, Dmitrii Kiselev, Maxim Beketov, and Ilya Makarov

Traffic signal control optimization is an integral part of any modern transportation system. However, modern traffic signal control systems often rely on predetermined fixed rules to adjust traffic signal timings. This paper presents VIA AI — an intelligent traffic signal control system that leverages deep reinforcement learning (RL) applied to count-based traffic data. Our solution offers additional adaptability and flexibility by allowing the system to learn and adjust its strategies based on real-time feedback and environmental changes. We test our approach using real-world traffic data and show that it outperforms classical methods of intersection control.



Source

Video-based learning of sign languages: one pre-train to fit them all

Maxim Novopoltsev, Aleksandr Tulenkov, Ruslan Murtazin, Roman Akhidov, Iuliia Zemtsova, Emilia Bojarskaja, Daria Bondarenko, Andrey Savchenko, and Ilya Makarov

This paper presents a novel system for recognizing sign language from video, addressing a critical need for improved communication accessibility for the deaf and hard-of-hearing communities. We developed a foundation model for Isolated Sign Language Recognition that uses self-supervised pre-training to address data scarcity issues. By applying the VideoMAE algorithm and a specially prepared dataset of American Sign Language videos, we created a vision transformer for video classification that performs exceptionally well. Our model achieves state-of-the-art results for Greek (GSL) and Russian (Slovo) sign languages, and comparable results for American (WLASL) and Turkish (AUTSL) sign languages. The fine-tuning process was efficient, with optimal performance in under forty epochs for each language. We also built a sign language learning tool integrated with the sign language recognition algorithm, showcasing its practical use in educational settings.

IJCAI

Do You Remember the Future? Weak-to-Strong Generalization in 3D Object Detection

Alexander Gambashidze, Aleksandr Dadukin, Maxim Golyadkin, Maria Razzhivina, Ilya Makarov

This paper demonstrates a novel method for LiDAR-based 3D object detection, addressing major field challenges: sparsity and occlusion. Our approach leverages temporal point cloud sequences to generate frames that provide comprehensive views of objects from multiple angles. To address the challenge of generating these frames in real-time, we employ Knowledge Distillation within a Teacher-Student framework, allowing the Student model to emulate the Teacher's advanced perception. We pioneered the application of weak-to-strong generalization in computer vision by training our Teacher model on enriched, object-complete data. In this demo, we showcase the exceptional quality of labels produced by the X-Ray Teacher on object-complete frames, showing our method distilling its knowledge to enhance object 3D detection models.

A*

Source

Plug-and-Play Unsupervised Fault Detection and Diagnosis for Complex Industrial Monitoring

Maksim Golyadkin, Maria Shtark, Petr Ivanov, Alexander Kozhevnikov, Leonid Zhukov, Ilya Makarov

Today industrial facilities are equipped with lots of sensors throughout all the production line for monitoring means. Gathered data can be used to detect and predict failures; however, manual labeling of large amounts of data for supervised learning is complicated. This paper introduces an innovative approach to unsupervised fault detection and diagnosis tailored for monitoring industrial chemical processes. We showcase the efficacy of our model using two publicly accessible datasets from the Tennessee Eastman Process, each containing various faults. Furthermore, we illustrate that by fine-tuning the model on a limited amount of labeled data, it achieves performance close to that of a state-of-the-art model trained on the entire dataset.In our experiments, we show through human evaluation and quantitative analysis that the proposed method allows to produce desired editing which is more preferable by humans and also achieves a better trade-off between editing quality and preservation of the original image. Our code is available at this https URL.

A*

Source

IJCAI

AADMIP: Adversarial Attacks and Defenses Modeling in Industrial Processes

Vitaliy Pozdnyakov, Aleksandr Kovalenko, Ilya Makarov, Mikhail Drobyshevskiy, Kirill Lukyanov

The development of the smart manufacturing trend includes the integration of Artificial Intelligence technologies into industrial processes. One example of such implementation is deep learning models that diagnose the current state of a technological process. Recent studies have demonstrated that small data perturbations, named adversarial attacks, can significantly affect the correct predictions of such models. This fact is critical in industrial systems, where AI-based decisions can be made to manage physical equipment. In this work, we present a system which can help to evaluate the robustness of technological process diagnosis models to adversarial attacks, as well as consider protection options. We briefly review the system's modules and also consider some useful applications. Our demo video is available at: http://tinyurl.com/3by9zcj5

A*

Source

Probabilistically Robust Watermarking of Neural Networks

Mikhail Pautov, Nikita Bogdanov, Stanislav Pyatkin, Oleg Rogov, Ivan Oseledets

As deep learning (DL) models are widely and effectively used in Machine Learning as a Service(MLaaS) platforms, there is a rapidly growinginterest in DL watermarking techniques that canbe used to confirm the ownership of a particularmodel. Unfortunately, these methods usually produce watermarks susceptible to model stealing attacks. In our research, we introduce a novel trigger set-based watermarking approach that demonstrates resilience against functionality stealing attacks, particularly those involving extraction anddistillation. Our approach does not require additional model training and can be applied to anymodel architecture. The key idea of our method isto compute the trigger set, which is transferable between the source model and the set of proxy models with a high probability. In our experimentalstudy, we show that if the probability of the set being transferable is reasonably high, it can be effectively used for ownership verification of the stolenmodel. We evaluate our method on multiple benchmarks and show that our approach outperforms current state-of-the-art watermarking techniques in allconsidered experimental setups.

A*

Source

EMNLP

Lost in Translation: Chemical Language Models and the Misunderstanding of Molecule Structures

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, Elena Tutubalina

The recent integration of chemistry with natural language processing (NLP) has advanced drug discovery. Molecule representation in language models (LMs) is crucial in enhancing chemical understanding. We propose Augmented Molecular Retrieval (AMORE), a flexible zero-shot framework for assessment of Chemistry LMs of different natures: trained solely on molecules for chemical tasks and on a combined corpus of natural language texts and string-based structures. The framework relies on molecule augmentations that preserve an underlying chemical, such as kekulization and cycle replacements. We evaluate encoder-only and generative LMs by calculating a metric based on the similarity score between distributed representations of molecules...

Source

DeepPavlov 1.0: Your Gateway to Advanced NLP Models Backed by Transformers and Transfer Learning

Maksim Savkin, Anastasia Voznyuk, Fedor Ignatov, Anna Korzanova, Dmitry Karpov, Alexander Popov, Vasily Konovalov

Weintroduce DeepPavlov 1.0, an open-source framework designed for seamless use of Natural Language Processing (NLP) models, leveraging advanced transfer learning techniques. This frameworkoffers a modular, configurationbased approach, making it suitable for a wide range of NLP applications without requiring in-depth knowledge of machine learning or NLP. Built on PyTorch and supporting Hugging Face transformers, DeepPavlov 1.0 provides ready-to-use solutions for various NLP tasks. It is publicly available1 under the Apache 2.0 license and includes access to an interactive online demo2.

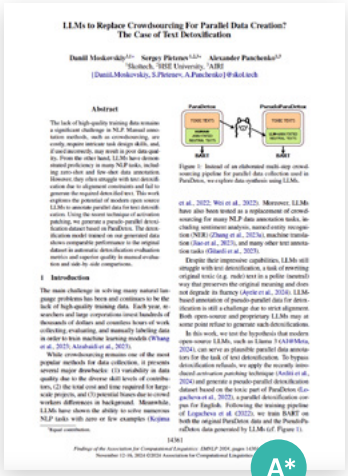
Source

EMNLP

LLMs to Replace Crowdsourcing For Parallel Data Creation: The Case of Text Detoxification

Daniil Moskovskiy, Sergey Pletenev, Alexander Panchenko

The lack of high-quality training data remains a significant challenge in NLP. Manual annotation methods, such as crowdsourcing, are costly, require intricate task design skills, and, if used incorrectly, may result in poor data quality. From the other hand, LLMs have demonstrated proficiency in many NLP tasks, including zero-shot and few-shot data annotation. However, they often struggle with text detoxification due to alignment constraints and fail to generate the required detoxified text. This work explores the potential of modern open source LLMs to annotate parallel data for text detoxification. Using the recent technique of activation patching, we generate a pseudo-parallel detoxification dataset based on ParaDetox. The detoxification model trained on our generated data shows comparable performance to the original dataset in automatic detoxification evaluation metrics and superior quality in manual evaluation and side-by-side comparisons.

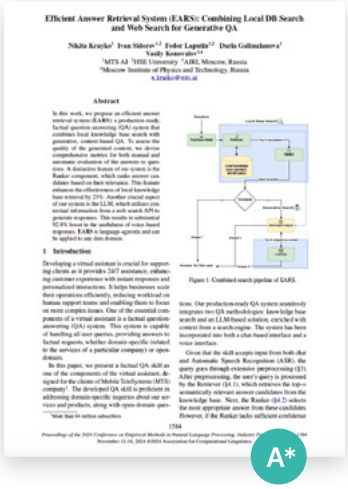


Source

Efficient Answer Retrieval System (EARS): Combining Local DB Search and Web Search for Generative QA

Nikita Krayko, Ivan Sidorov, Fedor Laputin, Daria Galimzianova, Vasily Konovalov

In this work, we propose an efficient answer retrieval system **EARS**: a production-ready, factual question answering (QA) system that combines local knowledge base search with generative, context-based QA. To assess the quality of the generated content, we devise comprehensive metrics for both manual and automatic evaluation of the answers to questions. A distinctive feature of our system is the Ranker component, which ranks answer candidates based on their relevance. This feature enhances the effectiveness of local knowledge base retrieval by 23%. Another crucial aspect of our system is the LLM, which utilizes contextual information from a web search API to generate responses. This results in substantial 92.8% boost in the usefulness of voice-based responses. **EARS** is language-agnostic and can be applied to any data domain.



Source

EMNLP

SparseGrad: A Selective Method for Efficient Fine-tuning of MLP Layers

Victoriia A. Chekalina, Anna Rudenko, Gleb Mezentsev, Aleksandr Mikhalev, Alexander Panchenko, Ivan Oseledets

The performance of Transformer models has been enhanced by increasing the number of parameters and the length of the processed text. Consequently, fine-tuning the entire model becomes a memory-intensive process. High-performance methods for parameter-efficient fine-tuning (PEFT) typically work with Attention blocks and often overlook MLP blocks, which contain about half of the model parameters. We propose a new selective PEFT method, namely SparseGrad, that performs well on MLP blocks. We transfer layer gradients to a space where only about 1\% of the layer's elements remain significant. By converting gradients into a sparse structure, we reduce the number of updated parameters. We apply SparseGrad to fine-tune BERT and RoBERTa for the NLU task and LLaMa-2 for the Question-Answering task. In these experiments, with identical memory requirements, our method outperforms LoRA and MeProp, robust popular state-of-the-art PEFT approaches.

A*

Source

Kandinsky 3: Text-to-Image Synthesis for Multifunctional Generative Framework

Arkhipkin Vladimir, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Bukashkin Anton, Konstantin Kulikov, Andrey Kuznetsov, Denis Dimitrov

Text-to-image (T2I) diffusion models are popular for introducing image manipulation methods, such as editing, image fusion, inpainting, etc. At the same time, image-to-video (I2V) and text-to-video (T2V) models are also built on top of T2I models. We present Kandinsky 3, a novel T2I model based on latent diffusion, achieving a high level of quality and photorealism. The key feature of the new architecture is the simplicity and efficiency of its adaptation for many types of generation tasks. We extend the base T2I model for various applications and create a multifunctional generation system that includes text-guided inpainting/outpainting, image fusion, text-image fusion, image variations generation, I2V and T2V generation. We also present a distilled version of the T2I model, evaluating inference in 4 steps of the reverse process without reducing image quality...

A*

Source

EMNLP

xCOMET-lite: Bridging the Gap Between Efficiency and Quality in Learned MT Evaluation Metrics

Daniil Larionov, Mikhail Seleznyov, Vasiliy Viskov, Alexander Panchenko, Steffen Eger

State-of-the-art trainable machine translation evaluation metrics like xCOMET achieve high correlation with human judgment but rely on large encoders (up to 10.7B parameters), making them computationally expensive and inaccessible to researchers with limited resources. To address this issue, we investigate whether the knowledge stored in these large encoders can be compressed while maintaining quality. We employ distillation, quantization, and pruning techniques to create efficient xCOMET alternatives and introduce a novel data collection pipeline for efficient black-box distillation. Our experiments show that, using quantization, xCOMET can be compressed up to three times with no quality degradation. Additionally, through distillation, we create an xCOMET-lite metric, which has only 2.6% of xCOMET-XXL parameters, but retains 92.1% of its quality. Besides, it surpasses strong small-scale metrics like COMET-22 and BLEURT-20 on the WMT22 metrics challenge dataset by 6.4%, despite using 50% fewer parameters. All code, dataset, and models are available online.

Source

AIRI publications

airi.net/articles/

AIRI open source repositories

Hugging Face

GitHub

Awards

MIDRC XAI Challenge

Researchers from AGI Med Lab and FusionBrain Lab placed in the top 5 of the MIDRC XAI Challenge. The competition was aimed at creating interpretable and reliable models of artificial intelligence. As a result, the team presented 3 variants of the problem and was ranked among the top 5 along with scientists from Johns Hopkins University, the University of Bern, a team from the Women's Hospital in Birmingham, as well as researchers from Stanford and the University of Tübingen.

Eedi — Mining Misconceptions in Mathematics

A team of researchers from AIRI's "Reliable and Secure Intelligent Systems" group, VeinCV, and Skoltech has won a silver medal in an international competition to create an algorithm that can help identify the causes of incorrect answers to math problems automatically.

Concordia Challenge

A team of scientists from AIRI, ISP RAS, ITMO, and the startup Coframe has been selected as a top 5 finalist in the Concordia Challenge competition, which was held as part of the prominent AI conference NeurIPS 2024. The team's work was recognized with an honorable mention from the competition's organizers, Google DeepMind and the Cooperative AI Foundation.

ASVspoof 2024 Challenge

AIRI's "Reliable and Secure Intelligent Systems" group and MTUSI's Intelligent Solutions Research Institute team, with the participation of a Skoltech graduate student, created a model for synthetic voice detection that was ranked among the top 10 best solutions in the international ASVspoof 2024 Challenge.

Awards

YandexML Prize

Ilya Zisman

Ilya Zisman, junior research scientist from the “Adaptive Agents” group, was nominated in the “First Publication” category.

Alexander Korotin

AIRI researcher and head of the research group at Skoltech, Alexander Korotin, won in the “Young Scientific Leaders” category.

Alexey Skrynnik

Alexander Tyurin

In the nomination entitled “Researchers,” two AIRI employees are highlighted: Alexey Skrynnik, who holds a Candidate of physical and mathematical sciences degree, serves as the head of the “RL Agents” group within the Laboratory of Cognitive AI Systems, and Alexander Tyurin, who holds a PhD in Computer Science and serves as the head of the “Optimization Methods in Machine Learning” group.

Aleksandr Panov

Elena Tutubalina

Two AIRI researchers were distinguished in the “Scientific Supervisors” category: Aleksander Panov, who holds a Candidate of physical and mathematical sciences degree and serves as the Director of Cognitive AI Systems Lab, and Elena Tutubalina, who possesses a Doctor of Science degree and functions as the Head of AIRI’s ‘Domain-specific NLP’ group and Senior Researcher at ISP RAS.

Anton Konushin

In the nomination “ML Teachers” — Anton Konushin, Candidate of physical and mathematical sciences, Head of the “Spatial AI” group.

Awards

International Olympiad in Artificial Intelligence (IOAI)

Andrey Gromyko

Andrey Gromyko, a trainee researcher at AIRI's FusionBrain Lab, was part of the Russian team that demonstrated the most outstanding result in the scientific round of the Olympiad. This team was awarded gold medals and also received silver in the practical stage. Additionally, they secured the top position in the sum of points for both stages of the competition.

ImageCLEFmed MEDVQA-GI

Mikhail Chaichuk

Research engineer Mikhail Chaychuk of the "Domain-Specific NLP" group created the best solution in the contest, which focused on generating medical images that simulate the results of endoscopic examinations of the stomach and intestines, such as gastroscopy and colonoscopy.

The Golden Names of Higher School

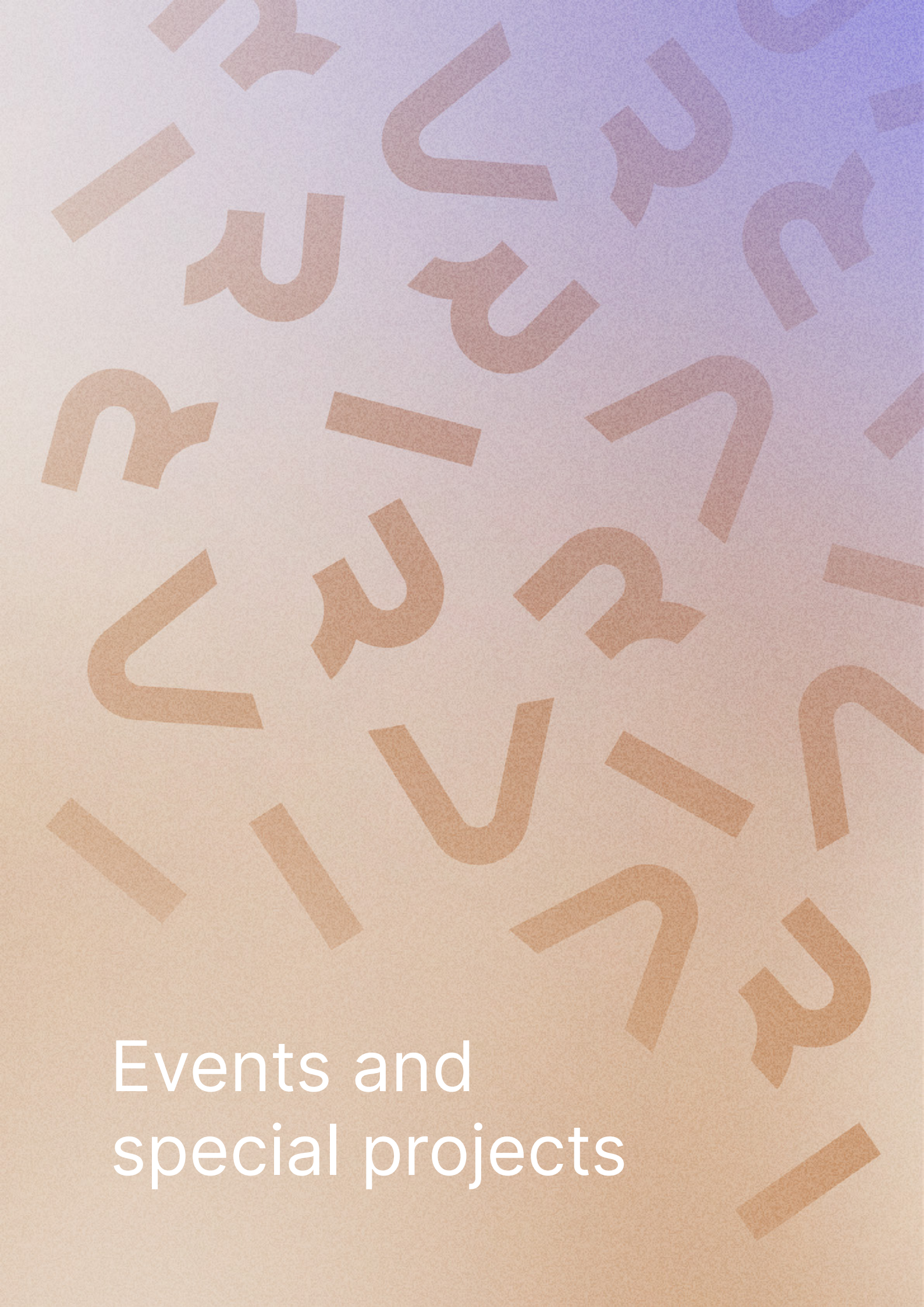
Anton Konushin

Anton Konushin won in the nomination "For Contribution to Science and Higher Education" of the All-Russian contest "Golden Names of Higher School".

National "AI Leaders" Award

Elena Tutubalina

One of the three winners of the National Award for Contribution to the Development of Artificial Intelligence Technologies in the category "Award to Scientists" was Elena Tutubalina, Doctor of Science, head of the "Domain-Specific NLP" group at AIRI and Senior Researcher at ISP RAS. Among the nominees in this category is Alexander Tyurin, PhD in Computer Science, Head of the "Optimization Methods in Machine Learning" group at AIRI and Senior Lecturer at Skoltech.



Events and special projects

AIRI Seminars playlist

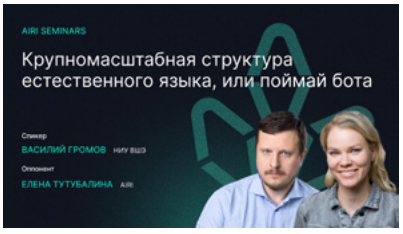
AIRI Seminars

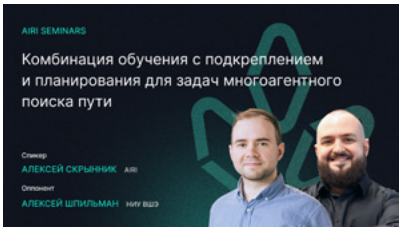
AIRI Seminars are a peer-to-peer scientific dialogue and an introduction of the professional community to the achievements in the field of Artificial Intelligence.

The seminar is designed to popularize and disseminate in the professional environment the principles and values adhered to by the Institute, as well as to promote the ideas that realize the mission of AIRI: the creation of universal artificial intelligence systems that solve real world problems.

Leading experts in the field of artificial intelligence from Russia and abroad are invited as speakers or opponents to present and constructively criticize research papers. This year the workshops were held not only online, but also offline.

In 2024, 19 seminars were held.





AIRI scientific seminars are available via VK Video

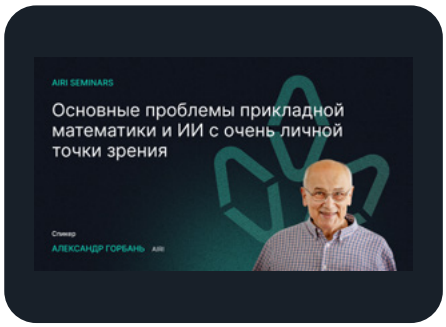
Statistics

19
seminars

22 000
views

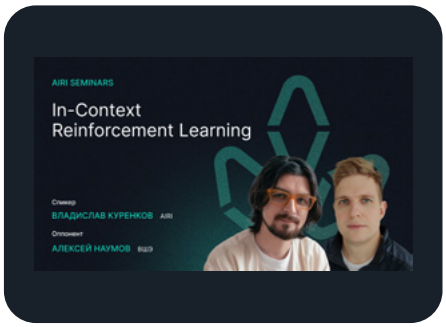
Most viewed seminars

Fundamental problems of applied mathematics and AI from a very personal point of view



Link

In-Context Reinforcement Learning



Link

Team

Aleksandr Panov

Alexey Skrynnik

Alexandra Broytman

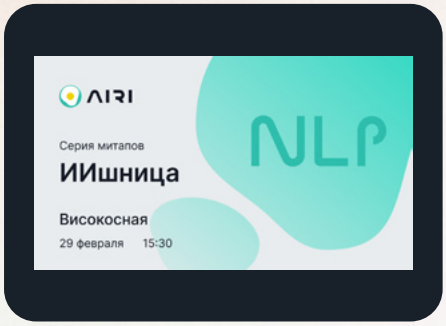
Ekaterina Mamontova

Yulia Trekhletova

Kristina Denisova

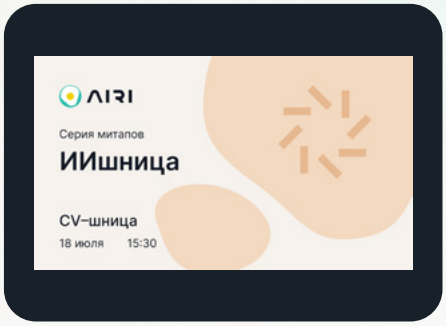
scrAImble meetup series

In 2024, we advanced the special project “scrAImble” — a series of online pitches where scientists discuss artificial intelligence in the framework of a 20-minute scientific report.



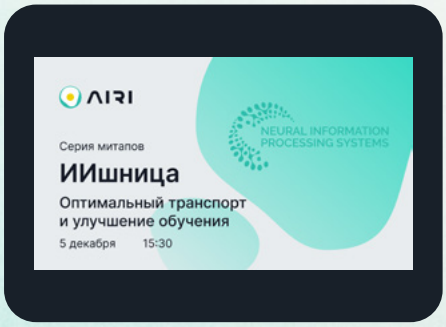
Feb 29

Leap scrAImble



Jul 18

CV-mble



Dec 5

NeurIPS 2024: Optimal Transport and Improved Learning

Link

Dec 6

NeurIPS 2024: Datasets, benchmarks and reviews

Link

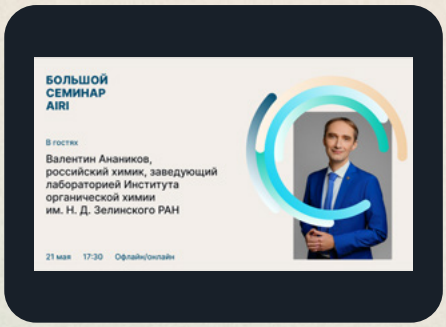
AIRI Grand Seminar

The AIRI Grand Seminar is an event designed to facilitate interactions among researchers from diverse fields and those with a general interest in science.

The seminar is chaired by Ivan Oseledets, Doctor of science, Russian Academy of Sciences professor, CEO of the AIRI Institute, and professor at Skoltech.

May 21

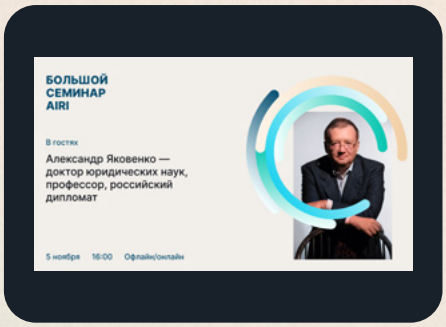
Artificial Intelligence in Chemistry



Link

Nov 5

Trends in global development



Link

AIRI scientists participated in podcasts

Aleksandr Panov
on the Big Data podcast.

Link

Andrey Kuznetsov
as a guest of the “OSNOVA”
podcast.

Link

Olga Kardymon and
Veniamin Fishman featured
on the Creative Science Lab
podcast.

Link

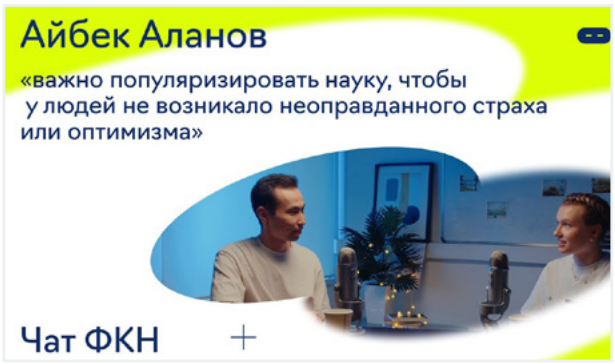
Ivan Oseledets on the
Money Loves Techno
podcast.

Link

AIRI scientists participated in podcasts

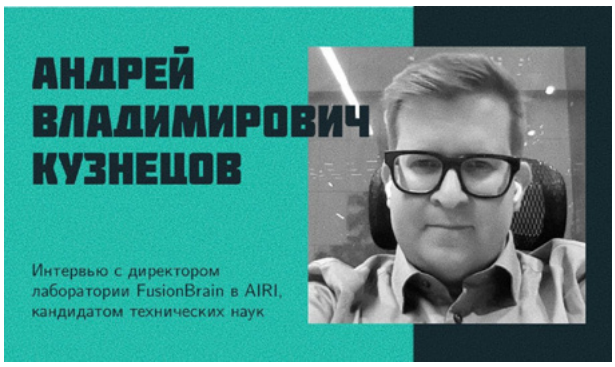
Ivan Oseledets
and Noosphere podcast

Link



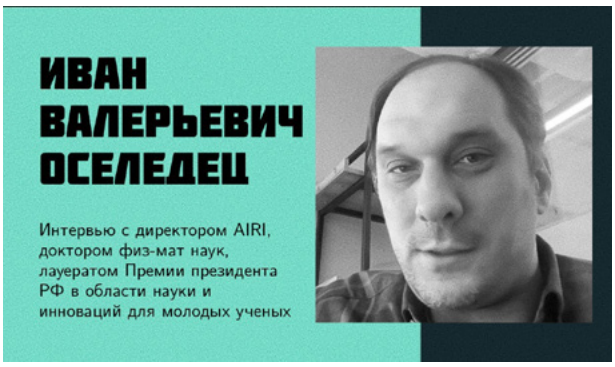
Aibek Alanov as a guest
of the ChatFCS podcast

Link



Andrey Kuznetsov visiting
Egor Bugaenko

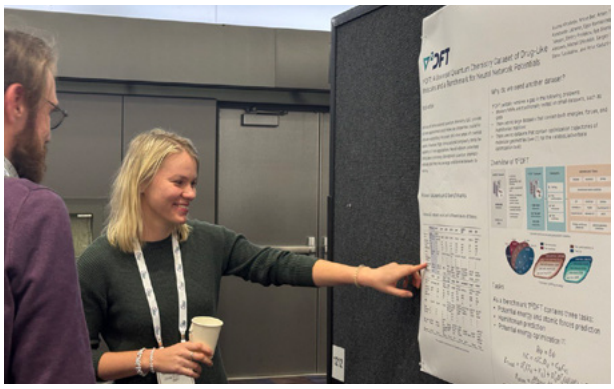
Link



Ivan Oseledets visiting
Egor Bugaenko

Link

AIRI researchers have presented their findings at a total of more than 70 scientific conferences



NeurIPS 2024 in Canada

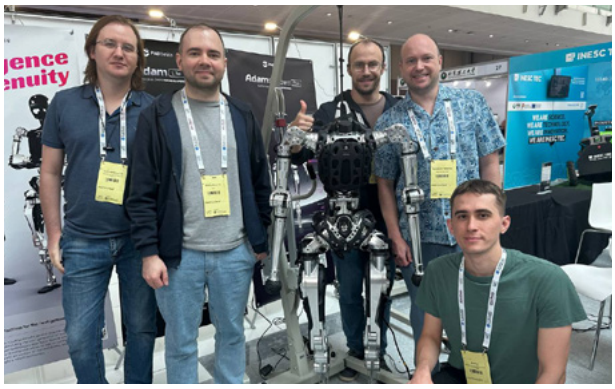
EMNLP 2024 in the USA

ECAI–2024 in Spain

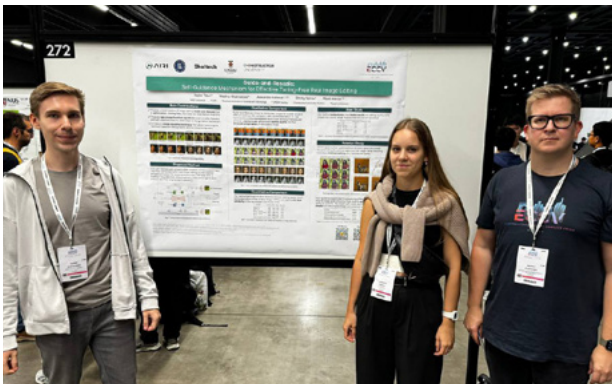
AIST 2024 in Kyrgyzstan

Fall into ML 2024 in Moscow

RecSys 2024 in Italy



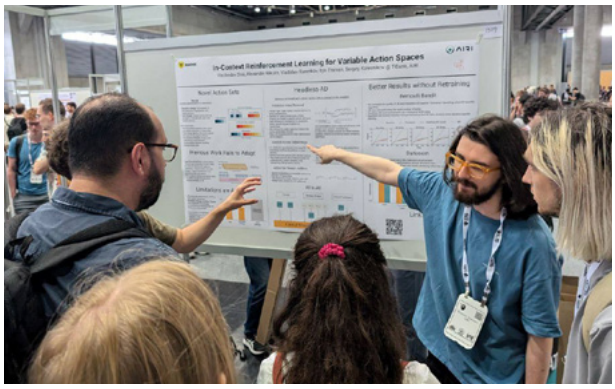
IROS 2024 in the UAE



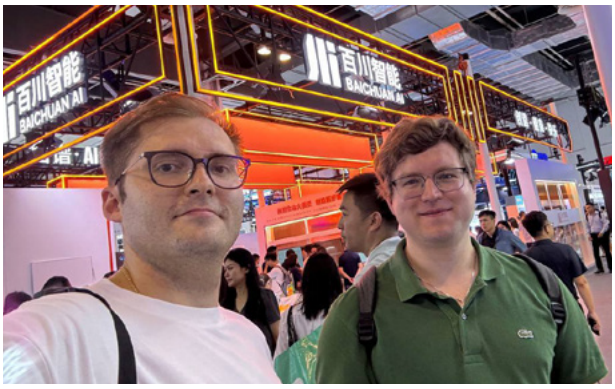
ECCV 2024 in Italy

ACL-2024 in Thailand

IJCAI 2024 in South Korea



ICML 2024 in Austria

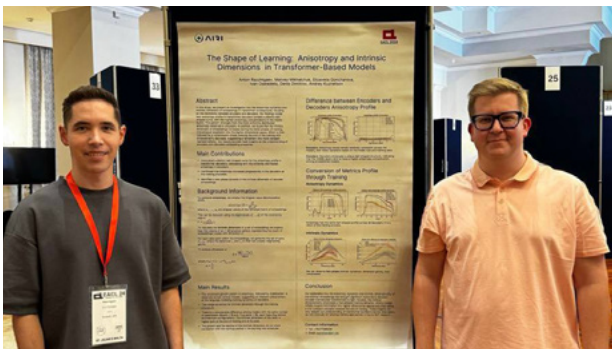


WAIC 2024 in China

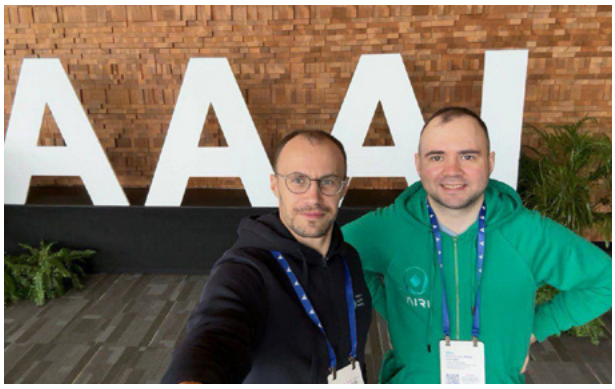
NAACL 2024 in Mexico

LREC-COLING 2024 in Italy

ICLR 2024 in Austria

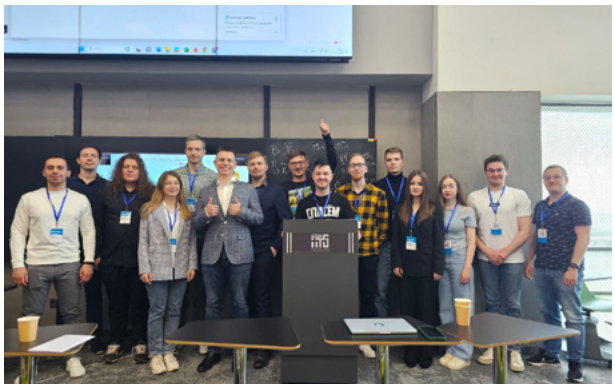


EACL 2024 in Malta



AAAI 2024 in Canada

IEEE International Conference on Robotics and Automation in Japan

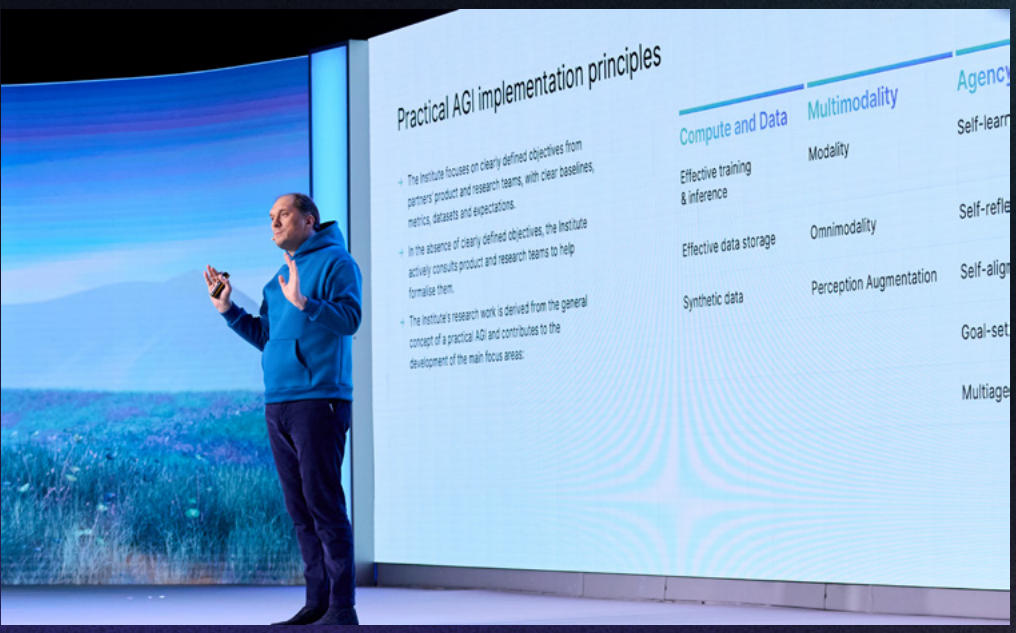


WAIT Trusted Artificial Intelligence Workshop in Kazakhstan

AI Journey 2024

18

researchers made
presentations
at the conference



AI Journey 2024

7

scientists
presented their
posters

AIRI summer school 2024

Summer with AIRI

AIRI Summer School for undergraduate and graduate students — is a deep dive into working with a wide range of state-of-the-art methods in artificial intelligence and machine learning.

The program participants spent 10 days with scientists from AIRI, ITMO, MIPT, HSE, Skoltech, Sber Artificial Intelligence Laboratory and other reputable research organizations and universities.

Researchers selected for the school not only met potential research mentors, but also received career counseling. They also applied their knowledge in hands-on workshops. At the program's finale, they defended 25 research projects on exploratory and practical topics. Among them: detecting anomalies in electrical measurements of industrial electric motors, studying adversarial attacks on secure models, pre-training multimodal models to work in a narrow domain, predicting absorption and emission frequencies of molecules using a large language model.

941
applications

37
professors

90
students

10
days

25
projects

45
lectures
and seminars

80
scientific
posters

наука меняет мир

лето открытий

Partners

SBER AI Lab

ITMO Advanced Engineering School

Scientific partners

Skoltech

HSE University

MIPT

MTUCI

Playlist
Reading Club | Embodied AI

Journal clubs

AIRI Reading Clubs are a place for lively discussion and the exchange of ideas. At AIRI Reading Clubs, researchers discuss new, important, or simply interesting research articles. At the meetings, participants delve deeper into the material, assess its relevance to their field, and broaden their research horizons.

Reading Club | ControlGenA

Led by
Aibek Alanov

32
meetings
for 2024

Reading Club | Embodied AI

Led by
Alexey Kovalev

10
meetings
for 2024

Reading Club | CV and Robotics

Led by
Vlad Shakhuro

31
meetings
for 2024

AIRI LEGO club

This year, the informal AIRI employee interest club continued to grow, with colleagues collecting Lego, getting to know each other, inviting friends and playing quizzes.

SafeSpeak-2024 hackathon

Deepfake detection for secure voice communications

SafeSpeak2024 is a hackathon dedicated to the development of audio spoofing detection technologies. The objective of this initiative is to solve current problems of secure voice authentication and to protect biometric systems from attacks.

>240

people have applied for participation

Young scientists from Russia, Ethiopia, Kazakhstan, Vietnam, Brazil, China, Uzbekistan and India applied for participation in the competition.

Hackathon “Typical AI vs Basic Models”

On December 14, the final of the hackathon “Typical AI vs Fundamental Models” took place. Twenty-three teams from leading Russian universities fought for the title of the best. Their main task was to create a machine learning model capable of automatically detecting pathology on chest radiographs.

AI Journey Contest 2024

AIRI researchers prepared 3 challenges for the AI Journey Contest 2024 hackathon:

- Emotional FusionBrain 4.0: Creating multimodal models for working with video, audio, and text.
- Multiagent AI: Creating a multi-agent RL system where agents can solve tasks by working together in different cooperation schemes.
- Embodied AI: Creating robots that can solve complex tasks that require interaction with the environment and the user, and communicate with them in natural language.

Night Photography Rendering Challenge 2024, NTIRE Workshop. CVPR 2024

Scientists from AIRI and the Institute for Information Transfer Problems (IITPI RAS) held a scientific competition on rendering night photos in conjunction with one of the most prestigious computer vision conferences, CVPR 2024. The goal was to use AI algorithms to process a night image from a smartphone camera and obtain a photo-quality image.

58 teams participated in the competition, including representatives from academia (Hong Kong, Singapore, Washington DC, Milan, etc.) and industry (Samsung, XiaoMi, Honor, Doshua, etc.).

AIRI in media

TASS, RIA, RBC, Forbes, Vedomosti, Kommersant, Rossiyskaya Gazeta, Gazeta.Ru, Lenta.ru, N+1, Izvestia, HiTech, Channel One, Russia 24, BFM, Durov’s Code, Radio «Mayak», Radio «Russia», Culture TV channel and many others.

Read expert contributions from AIRI staff

Forbes
Ivan Oseledets on how to evaluate the work of DeepTech researchers

Forbes
Anton Konushin on who to hire for an AI startup and what skills to look for

RBC Trends
Aleksander Panov on how scientists are using games to train artificial intelligence systems to solve real-world problems

RBC Trends
Yuri Kuratov on how the memory of neural networks is organized

RB
Maxim Kuznetsov on how to choose a neural network for business

RB
Alexandra Broytman on how to organize summer schools

AIRI in media

N+1

Partner material on what science produces and how it helps business

Kommersant

Evgeny Frolov on how researchers have improved the accuracy of recommender systems

Kommersant

Ivan Oseledets for Kommersant on how a team of Russian mathematicians refined the Nobel laureate's conclusions

TASS

Ivan Oseledets for Kommersant on how a team of Russian mathematicians refined the Nobel laureate's conclusions

Kommersant

Ilya Makarov on small neural networks being able to train large AI models better than humans

Kommersant

Aibek Alanov on how image editing will help science

New partnerships

Contact information

Webpage

airi.net

Social media

airi_research_institute

Airi_institute

habr

artificial-intelligence-research-institute

AIRI_inst

Address

Moscow, Presnenskaya Embankment, 6, buld. 2