Roman Koshkin

https://roman-koshkin.unit.oist.jp /RomanKoshkin roman.koshkin@gmail.com /in/romankoshkininterpreter

Skills

Programming languages: Python (highest proficiency), C++, Matlab, R, HTML, JavaScript
ML Frameworks: Pytorch (highest proficiency), HuggingFace, Lightning, scikit-learn, JAX
Frontend development: React, Next.js
Virtualization, Infra, HPC: Docker, Singularity, AWS, slurm
DB, LLMOps, MLOps, etc.: Neo4j, Redis, MongoDB, LangSmith, LangChain, LangGraph, wandb, Apache Spark

Education

OIST | *PhD (Machine Learning and Computational Neuroscience)* — Okinawa, Japan
Modeling memory in spiking networks and pattern detection — *Expected graduation: 11/2024*

HSE University | *Master (Cognitive Science)* — Moscow, Russia
with distinction, GPA: 8.9/10

VUMO University | *Bachelor (Applied Linguistics)* — Moscow, Russia
with honors, GPA: 4.9/5

Work Experience

Amazon — Tokyo, Japan
Research Scientist Intern — *09/2024 - 11/2024*
- Building multi-agent systems for open-ended automated market research and analysis.

NLP Group @ NAIST — Nara, Japan
Special Research Intern — *07/2023 - 10/2023*
- Developed speech-to-text SiMT models leveraging open-source causal LLMs
- Set up LLMOps/MLOps, parallel experiments to identify best design and HP choices

Reinforcement Learning Research Team @ Araya — Tokyo, Japan
Research Intern — *07/2023 - 10/2023*
- Conducted research towards using EEG for robot control with a brain-machine interface
- Compiled a sensor-aligned motor imagery EEG datasets, trained self-supervised EEG feature extractors
- Achieved competitive performance in downstream tasks (incl. MI imagery classification)
- Reimplemented and open-sourced an M/EEG speech-decoding model

Center for Bioelectric Interfaces @ HSE University — Moscow, Russia
Junior Research Fellow — *09/2017 - 07/2019*
- Conceptualized research agenda, coordinated a team of 3 researchers for 2 years
- Implemented EEG data collection, pre-processing pipelines, built ML models and other software
- Provided progress reports to the funding company, co-authored RF Patent 2747571

Selected Publications

Koshkin, R., Sudoh, K., Nakamura, S. (2024). LLMs Are Zero-Shot Context-Aware Simultaneous Translators. *EMNLP 2024*. [LINK]
Koshkin, R., Sudoh, K., Nakamura, S. (2024). TransLLaMa: LLM-based Simultaneous Translation System. *EMNLP 2024*. [LINK]
Koshkin, R., Fukai, T. (2024). convSeq: Fast and Scalable Method for Detecting Patterns in Spike Data. In *ICML 2024*. [LINK]
Koshkin, R., Fukai, T. (2023). Unsupervised Detection of Cell Assemblies with Graph Neural Networks. In *ICLR 2023*. [LINK]
Koshkin, R., Fukai, T. (2024). Astrocyte Regulation of Synaptic Plasticity Balances Robustness and Flexibility of Cell Assemblies. In *bioRxiv*. [LINK]

Awards, Grants and Fellowships

OIST Teaching Fellowship, 2023, ¥1.2M
KAKENHI Grant-in-Aid, 2023, ¥1.8M [LINK]
Japan Society for the Promotion of Science Fellowship, 2023 [LINK]
Google PhD Fellowship, 2021, $10K [LINK]

Patents

RF Patent 2747571. EEG method for estimating listeners' reaction to audio content. [LINK]

Projects

toLLMatch [LINK] *02/2024 - 05/2024*
- Multilingual LLM-based speech-to-text simultaneous machine translation with no costly pre-training or fine-tuning

TransLLaMa [LINK] *07/2023 - 10/2023*
- LLM-based speech-to-text simultaneous machine translation

convSeq [LINK] *01/2023 - 03/2023*
- Fast and scalable convolution-based method for unsupervised detection of patterns in neural recordings

SoNNet [LINK] *09/2020 – 06/2024*
- High-performance C++ library with a configurable user-friendly Python API for building spiking neural networks

graphSeq [LINK] *09/2022 - 12/2022*
- Graph neural network-based method for embedding and clustering of neural spiking patterns

M/EEG-based zero-shot speech decoding [LINK] *09/2022 - 12/2022*
- Re-implementation of an algorithm that decodes speech from human brain recordings (M/EEG) 0-shot

Teaching Object Handling to a Robot [LINK] *04/2020 - 09/2020*
- Trained a robot to perform reach-and-grasp tasks by combining learned motor primitives

Backpropagation-free learning for classification tasks [LINK] *1/2020 - 4/2020*
- Built a spike-timing dependent plasticity-based spiking neural network for image classification

Neurobarometer [LINK] *10/2017 - 07/2019*
- Software & algorithm for EEG-based neuromarketing and consumer behavior research

Finding Weak Effects in Evoked Response Data [LINK] *09/2016 - 04/2017*
- Contributed to designing a novel projection-based method for identifying weak effects in noisy ERP data

Poster presentations

Koshkin, R., Fukai, T. (2022). Astrocytes facilitate self-organization and remodeling of cell assemblies under STP-coupled STDP. *SfN Conference*, Nov 14-16, San Diego. [LINK]

Koshkin, R., Fukai, T (2021). Leveraging Self-organized Structure for Memory Encoding in Binary Networks. *RIKEN-OIST Symposium*, Oct. 6-7, 2021, Japan [LINK]

Koshkin, R., Shtyrov, Y. & Ossadtchi, A. (2017). Testing One Aspect of the Efforts Model of Simultaneous Interpreting: An ERP Study. In *Proceedings of the Workshop "Neurobiology Of Speech And Language"*, Oct. 27-29, 2017, SPb, Russia [LINK]

Koshkin, R., Ossadtchi, A. & Shtyrov, Y.(2016). N1 ERP As an Index of Depth of Processing In Simultaneous Interpreting. In *Proceedings of Communication, Computation, and Cognitive Processes*, Sept. 28-29, 2016, Moscow, Russia [LINK]

Koshkin, R., Ossadtchi, A. & Shtyrov, Y.(2017). Working Memory Load In Simultaneous Language Interpretation: An ERP Study. *IEEE International Symposium «Video and Audio Signal Processing»*, Jun. 26-30, 2017, SPb, Russia [LINK]

Kuznetsova A., **Koshkin R**., Ossadtchi A. (2017). Localizing Hidden Regularities With Known Temporal Structure in the EEG Evoked Response Data. IEEE International Symposium «Video and Audio Signal Processing», Jun. 26-30, 2017, SPb, Russia [LINK]

Conference Proceedings, Book Chapters, and Other Publications

Koshkin, R., Ossadtchi, A. (2017). Working Memory Load in Simultaneous Language Interpretation: An ERP Study. In *Proc. of the 4th Conference "Cognitive Science in Moscow: New Research"*. July 15, 2017, Moscow, Russia. p. 434 [LINK]

Garcia, A., **Koshkin, R**., Paiva, T. (2023). EEG In Cognitive Translation and Interpreting Studies. (In review)

Koshkin, R., Shtyrov, Y. , Myachykov, A, & Ossadtchi, A. (2018). Testing the Efforts Model of Simultaneous Interpreting. *PLoS ONE* 13(10): e0206129. [LINK]

Koshkin, R., & Ossadtchi, A. (2017). Commentary: Functional Connectivity in the Left Dorsal Stream Facilitates Simultaneous Language Translation: An EEG Study. *Front. in Hum. Neurosci.*, 11(2), 273. [LINK]

Koshkin, R., Ossadtchi, A. & Shtyrov, Y. (2017). Attention, Working Memory And Listening In Simultaneous Interpreting. *Russian J. of Cognitive Sci.*, 4(4). [LINK]

Koshkin R. (2016). Comparative Analysis of English-Russian and Russian-English Simultaneous Interpreting. *Bulletin of Moscow University, Series 22: Theory of Translation.* Vol. 2, 28-43 [LINK]

Service

Reviewer for Language, Cognition and Neuroscience
Reviewer for Tiny Papers Track @ ICLR 2023
Science Mentor Introduction to Deep Learning with Python, Okinawa, Japan