TOMASZ KORBAK

Department of Informatics, University of Sussex, UK
tomekkorbak.com ⋄ t.korbak@sussex.ac.uk

INTERESTS

language models, reinforcement learning, probabilistic programming, Bayesian inference

FEATURED WORK

1. **Korbak, T.**, Elsahar, H., Kruszewski, G. Dymetman, M. (2022). On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. NeurIPS 2022.
2. **Korbak, T.**, Elsahar, H., Kruszewski, G. Dymetman, M. (2022). Controlling conditional language models without catastrophic forgetting. ICML 2022.
3. **Korbak, T.**, Perez, E. Buckley, C. (2022). RL with KL penalties is better viewed as Bayesian inference. RLDM 2022 workshop on RL as a Model of Agency.
4. Kuciński, Ł., **Korbak, T.**, Kołodziej, P., Miłoś, P., (2021). Catalytic role of noise and necessity of inductive biases in emergence of compositional communication. NeurIPS 2021.
5. **Korbak, T.** Elsahar, H., Dymetman, M. Kruszewski, G (2021). Energy-Based Models for Code Generation under Compilability Constraints. ACL 2021 workshop NLP4Prog.
6. **Korbak, T.**, Zubek, J., Rączaszek-Leonardi, J. (2020). Measuring non-trivial compositionality in emergent communication. NeurIPS 2020 workshop "Talking to Strangers: Zero-Shot Emergent Communication".
7. Główka, K., Niklewski, M., **Korbak, T.**, Zubek, J., Rączaszek-Leonardi, J. (2020). Emergence of Action-grounded Compositional Communication. CogSys 2020.
8. **Korbak, T.**, Zubek, J., Kuciński, Ł., Miłoś, P., & Rączaszek-Leonardi, J. (2019). Developmentally motivated emergence of compositional communication via template transfer. NeurIPS 2019 workshop Emergent communication.

RESEARCH EXPERIENCE

CILVR Lab, **Visiting researcher**
New York University *February 2022 - ongoing*

I'm working with Sam Bowman and Ethan Perez on offline RL and conditional training for aligning language models with human preferences.

Department of Informatics, **PhD researcher**
University of Sussex *September 2020 – ongoing*

I'm working on aligning language models with Christopher Buckley and Anil Seth.

Redwood Research, **Teaching assistant**
MLAB *August 2022 - September 2022*

I was a TA for Machine Learning for Alignment (MLAB) bootcamp organised by Redwood Research in Berkeley.

Naver Labs Europe **Research intern**
Naver Corporation *February 2021 – August 2021*

I worked on energy-based models for controllable language generation with Marc Dymetman

Human Interactivity and Language Lab, **Research assistant**
Faculty of Psychology, University of Warsaw *February 2019 – October 2020*

I investigated compositional generalisation in neural networks with Joanna Rączaszek-Leonardi and Piotr Miłoś.

Institute of Computer Science, **Research intern**
Polish Academy of Sciences *April 2017 — November 2017*

I worked on NLP tools for Polish (e.g. aspect-based sentiment analysis) as part of Clarin-PL project.

INDISTRY EXPERIENCE

Daftmobile **Senior Machine Learning Engineer**
February 2019 – October 2020

I led development of ML matchmaking pipelines for Elympics. This included designing a Kubernetes microservice architecture and automating model training and deployment and conducted RL experiments.

Sigmoidal **Machine Learning Engineer**
2018 – 2020

I worked on several NLP projects for clients in finance and education, e.g. information retrieval and multilingual classification. I managed engineering teams as a technical leader and owning data labeling.

Samsung R&D Centre **Junior NLP Engineer**
2017 – 2017

I was developing representation learning techniques for error analysis for Bixby, the voice assistant.

Inteliclinic **Junior Python Developer**
2015 – 2017

I worked on data processing pipelines for EEG and EMG signals from wearable devices.

Webinterpret **Junior Python Developer Intern**
2015 – 2015

I developed infrastructure for machine translation for e-commerce platforms.

EDUCATION

PhD in Informatics, University of Sussex, UK *2020 – ongoing*
MSc in Cognitive Science, University of Warsaw, Poland *2016 – 2019*
BSc in Cognitive Science, University of Warsaw, Poland *2013 – 2016*
BAs in Philosophy, University of Warsaw, Poland *2012 – 2015*

ADDITIONAL TRAINING

Machine learning for Alignment (MLAB) bootcamp, Redwood Research *2022*
Diverse Intelligences Summer Institute, University of California, Los Angeles *2020*
Bayesian Methods in Deep Learning, National Research University (Moscow) *2018*
School of Pioneers (tech entrepreneurship workshops), University of Cambridge *2018*
Computational Psychiatry Course, ETH (Zurich) *2017*

SKILLS

Python (web frameworks and data science ecosystem), C++, PyTorch, tensorflow, git, Docker, Kubernetes, slurm, cloud computing, GNU/Linux, LaTeX

AWARDS AND FELLOWSHIPS

Leverhulme Doctoral Scholarship (Leverhulme Trust)	*2020-2023*
Diverse Intelligences Summer Institute Fellowship (Templeton Foundation)	*2020*
Collegium Invisibile Fellowship	*2017 – 2021*
Minister of Science and Higher Education (Poland) scholarship for exceptional students	*2016*
Diamond grant award (Ministry of Science and Higher Education, Poland)	*2016 – 2020*

Last updated on September 21, 2022.