ALON JACOVI

Curriculum Vitae
Tel Aviv, Israel ⋄ alonjacovi@gmail.com

RESEARCH

Natural Language Processing, Machine Learning, Explainable Artificial Intelligence, Complex Reasoning in Artificial Intelligence

EDUCATION

Ph.D in Natural Language Processing and Machine Learning *2019–2023*
Bar Ilan University
Advisor: Prof. Yoav Goldberg
Dissertation Topic: Explaining Artificial Intelligence: Foundations and Practice

M.Sc in Natural Language Processing and Machine Learning *2017–2019*
Bar Ilan University
Advisor: Prof. Yoav Goldberg
Graduated *cum laude* with final grade 95.85.
Dissertation Topic: Understanding Convolutional Neural Networks for Text Classification

B.Sc in Computer Science *2014–2017*
Bar Ilan University
Graduated *cum laude* with final grade 95.2.
Final Project: Deep Reinforcement Learning Agent for the AI-BIRDS Competition

EXPERIENCE

Research Scientist, Google. 05/2024–current

Work on verification in cases of long-context, reasoning, and structured data.

Research Intern, Google. 05/2022–05/2024

Hosts: Dr. Roee Aharoni and Dr. Mor Geva
Topics: Evaluation of chain-of-thought verification methods, evaluation of tool-assisted complex reasoning strategies, and a user study of crowd-sourced evaluations of explanations.

Research Intern, Google. Fall 2021

Hosts: Dr. Jasmijn Bastings and Dr. Katja Filippova
Topic: Diagnosing AI explanation methods with folk concepts of behavior.

Research Intern, Allen Institue for Artificial Intelligence. 10/2020–1/2021

Host: Dr. Swabha Swayamdipta
Topic: Contrastive explanations for model interpretability.

Student Researcher, IBM. 07/2016–09/2020

- Work on improving graph-based virtual dialogue assistants, based on in-production conversation logs that were escalated to a human agent.
- Work on integration of non-differentiable systems in neural pipelines (e.g., knowledge bases and program interpreters in semantic parsing) for end-to-end learning.

Research Intern, RIKEN. Spring 2019

Hosts: Dr. Gang Niu and Prof. Masashi Sugiyama
Topic: Scalable evaluation and improvement of Document Set Expansion via neural Positive-Unlabeled Learning.

PUBLICATIONS

TACT: Advancing Complex Aggregative Reasoning with Information Extraction Tools.
Avi Caciularu, Alon Jacovi, Eyal Ben-David, Sasha Goldshtein, Tal Schuster, Jonathan Herzig, Gal Elidan, Amir Globerson.
In NeurIPS 2024.

Is It Really Long Context if All You Need Is Retrieval? Towards Genuinely Difficult Long Context NLP.
Omer Goldman*, Alon Jacovi*, Aviv Slobodkin*, Aviya Maimon*, Ido Dagan, Reut Tsarfaty. In EMNLP 2024.

Unpacking Human-AI Interaction in Safety-Critical Industries: A Systematic Literature Review.
Tita A. Bach, Jenny K. Kristiansen, Aleksandar Babic, Alon Jacovi.
In IEEE Access 2024.

A Chain-of-Thought Is as Strong as Its Weakest Link: A Benchmark for Verifiers of Reasoning Chains.
Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, Mor Geva.
In ACL 2024.

Diagnosing AI Explanation Methods with Folk Concepts of Behavior.
Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, Katja Filippova.
In JAIR 2023.

Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks.
Alon Jacovi, Avi Caciularu, Omer Goldman, Yoav Goldberg.
In EMNLP 2023. *(oral presentation)*

A Comprehensive Evaluation of Tool-Assisted Generation Strategies.
Alon Jacovi, Avi Caciularu, Jonathan Herzig, Roee Aharoni, Bernd Bohet, Mor Geva.
In Findings of EMNLP 2023.

Neighboring Words Affect Human Interpretation of Saliency Explanations.
Alon Jacovi*, Hendrik Schuff*, Heike Adel, Ngoc Thang Vu, Yoav Goldberg.
In Findings of ACL 2023.

Human Interpretation of Saliency-based Explanation Over Text.
Hendrik Schuff*, Alon Jacovi*, Heike Adel, Yoav Goldberg, Ngoc Thang Vu.
In ACM FAccT 2022.

Contrastive Explanations for Model Interpretability.
Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, Yoav Goldberg.
In EMNLP 2021.

Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI.
Alon Jacovi, Ana Marasović, Tim Miller, Yoav Goldberg.
In ACM FAccT 2021.

Scalable Evaluation and Improvement of Document Set Expansion via Neural Positive-Unlabeled Learning.
Alon Jacovi, Gang Niu, Yoav Goldberg, Masashi Sugiyama.
In EACL 2021.

Aligning Faithful Interpretations with their Social Attribution.
Alon Jacovi, Yoav Goldberg.
In TACL 2021.

Amnesic Probing: Behavioral Explanations with Amnesic Counterfactuals.
Yanai Elazar, Shauli Ravfogel, Alon Jacovi, Yoav Goldberg.
In TACL 2021.

Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data.
Shachar Rosenman, Alon Jacovi, Yoav Goldberg.
In EMNLP 2020.

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?
Alon Jacovi, Yoav Goldberg.
In ACL 2020.

Improving Task-Oriented Dialogue Systems in Production with Conversation Logs.
Alon Jacovi*, Ori Bar El*, Ofer Lavi, David Boaz, David Amid, Inbal Ronen, Ateret Anaby-Tavor.
In KDD Converse @ KDD 2020.

Neural Network Gradient-based Learning of Black-box Function Interfaces.
Alon Jacovi*, Guy Hadash*, Einat Kermany*, Boaz Carmeli*, Ofer Lavi, George Kour, Jonathan Berant.
In ICLR 2019.

Learning and Understanding Different Categories of Sexism Using Convolutional Neural Network Filters.
Sima Sharifirad, Alon Jacovi, Stan Matwin.
In Widening NLP @ ACL 2019. *(extended abstract)*

Understanding Convolutional Neural Networks for Text Classification.
Alon Jacovi, Oren Sar Shalom, Yoav Goldberg.
In BlackboxNLP @ EMNLP 2018. *(oral presentation)*

IN SUBMISSION

CoverBench: A Challenging Benchmark for Complex Claim Verification.
Alon Jacovi, Moran Rachel Ambar, Eyal Ben-David, Uri Shaham, Amir Feder, Mor Geva, Dror Marcus, Avi Caciularu

Can Few-shot Work in Long-Context? Recycling the Context to Generate In-Domain Demonstrations.
Arie Cattan, Alon Jacovi, Alex Fabrikant, Jonathan Herzig, Roee Aharoni, Hannah Rashkin, Dror Marcus, Avinatan Hassidim, Yossi Matias, Idan Szpektor, Avi Caciularu.

($*$) - equal contribution.

PATENTS

Computerized dialog system improvements based on conversation data. *Published, 2023*
Ofer Lavi, Alon Jacovi, David Amid, David Boaz, Inbal Ronen, Ateret Anaby-Tavor, Ori Bar El.
US Patent: US20220059097A1 (11605386)

ACADEMIC SERVICE

Organizing Committee: Data Contamination Workshop (CONDA) 2024

Area Chair: ACL Rolling Review 2021 and 2022, ACL 2023

Reviewer: ACL 2020, HAMLETS 2020, EACL 2021, NAACL 2021, ACL 2021 *(outstanding reviewer)*, EMNLP 2021, BlackboxNLP 2021, TRAIT 2022, BlackboxNLP 2022, HMCaT 2022, Open Mind 2022, NAACL 2024, ACL Rolling Review 2024

Assistant Reviewer: IJCAI 2018, IJCAI 2019

INVITED TALKS

Tel Aviv University, Jul 2024: *Evaluating evaluations of things.*

DEEL & ANITI, Nov 2023: *What would a good explanation even look like?*

DNV, Mar 2023: *Formalizing Trust in Artificial Intelligence.*

Institute for Trusted Intelligence in Society Workshop, Sep 2022: *Formalizing Trust in Artificial Intelligence.*

Trustworthy ML Reading Group, Sep 2022: *Formalizing Trust in Artificial Intelligence.*

XAI Israel Meeting Group, May 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

Cambridge, May 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

Harvard, Mar 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

Oxford, Mar 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

NEC Labs Europe, Mar 2022: *Diagnosing AI Explanation Methods with Folk Concepts of Behavior.*

Allen Institute for Artificial Intelligence, Jun 2021: *Formalizing Trust in Artificial Intelligence.*

Microsoft, Oct 2020: *Formalizing Properties of Interpretability in NLP.*

HONORS AND SCHOLARSHIPS

2024	Runner-up, IAAI Best PhD Thesis Award
2021	Nadav Award for Excellence in Doctoral Studies
2019–2023	The President's Scholarship for Outstanding Doctoral Fellows
2016	The Miryam and Ezra Sofer Scholarship for Excellent Students
2016	Dean's Honors
2015	The Grace Shua and Jacob Ballas Scholarship
2015	Dean's Honors

LANGUAGES

Native	Hebrew
Fluent	English