EMNLP 2023

BlackboxNLP Analyzing and Interpreting Neural Networks for NLP

Proceedings of the Sixth Workshop

December 7, 2023

©2023 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL) 317 SIDNEY BAKER ST S STE 400-134 KERRVILLE TX 78028 United States of America

TEL: +1-570-476-8006 FAX: +1-570-476-0860 acl@aclweb.org

ISBN 979-8-89176-052-3

BlackboxNLP gratefully acknowledges financial support from the following sponsors.





Message from the Organizing Committee

As researchers achieve unprecedented technological breakthroughs in natural language processing, the need to understand the systems underlying these advances is more pertinent than ever. BlackboxNLP, now in its sixth iteration, has played an important role in bringing together scholars from a diverse range of backgrounds in order to rigorously study the behavior, representations, and computations of "black-box" neural network models. Our workshop showcases original, cutting-edge research on topics including but not limited to:

- analysis of representations via probing and related techniques;
- explanation methods such as feature attribution, free-text explanations, or structured explanations;
- interdisciplinary methods (e.g., from neuroscience, cognitive science, computer vision, etc.);
- interpretable architectures and neural network modules;
- mechanistic interpretability and reverse engineering of neural computations;
- open-source tools for analysis, visualization, and/or explanation;
- opinions about the state of interpretability and explainable NLP; and
- targeted evaluations using simplified or formal languages.

The sixth BlackboxNLP workshop will be held in Singapore on December 7, 2023, hosted by the Conference on Empirical Methods in Natural Language Processing (EMNLP). 29 full papers and 17 non-archival extended abstracts were accepted for in-person and online presentations, from a total of 66 submissions. This year's workshop will also feature 19 papers on interpretability from the *Findings of the ACL: EMNLP 2023*, as well as two invited talks and a panel discussion with experts in the field.

BlackboxNLP 2023 would not have been possible without the high-quality peer reviews submitted by our program committee, as well as the logistical assistance provided by the EMNLP organizing committee. We gratefully acknowledge financial support from our sponsors, Google and Apple. Our invited speakers, panelists, authors, and presenters have allowed us to put together an outstanding program for all participants to enjoy.

Welcome to BlackboxNLP! We look forward to seeing you in Singapore and online.

Yonatan Belinkov Sophie Hao Jaap Jumelet Najoung Kim Arya McCarthy Hosein Mohebbi

Workshop Organizers

Organizing Committee

Yonatan Belinkov	Technion–Israel Institute of Technology
Sophie Hao	New York University
Jaap Jumelet	University of Amsterdam
Najoung Kim	Boston University
	Google
Arya McCarthy	Johns Hopkins University
Hosein Mohebbi	Tilburg University

Program Committee

Badr Abdullah	Dieuwke Hupkes	Rudolf Rosa
Carolyn Anderson	Yangfeng Ji	Naomi Saphra
Leila Arras	Robin Jia	Gabriele Sarti
Pepa Atanasova	Jaap Jumelet	Sebastian Schuster
Arianna Bisazza	Lis Kanashiro Pereira	Mattia Setzu
Jonathan Brophy	Eugene Kharitonov	Tatiana Shavrina
Lisa Bylinina	Saurabh Kulshreshtha	Gaofei Shen
Hanjie Chen	Jenny Kunz	Pia Sommerauer
Benoit Crabbé	Anna Langedijk	Shane Steinert-Threlkeld
Verna Dankers	Alessandro Lenci	Vinitra Swamy
Subham De	Sheng Liang	Aarne Talman
Yanai Elazar	Tomasz Limisiewicz	Jörg Tiedemann
Jean-Philippe Fauconnier	Nelson F. Liu	Dennis Ulmer
Nils Feldhus	Kanishka Misra	Saujas Vaduguru
Ghazi Felhi	Hosein Mohebbi	Oskar van der Wal
Javier Ferrando	Anmol Nayak	Eva Vanmassenhove
Bob Frank	Joakim Nivre	Jithendra Vepa
Richard Futrell	Siddharth Patwardhan	Johannes Welbl
Michael Goodale	Mohammad Taher Pilehvar	Peter West
Sarang Gupta	Tiago Pimentel	Sarah Wiegreffe
Michael Hanna	Yuval Pinter	Zhouhang Xie
David Harwath	Adithya Pratapa	Fabio Massimo Zanzotto
Christian Herold	Sara Rajaee	Yian Zhang
John Hewitt	Shauli Ravfogel	Yichu Zhou

Invited Speakers

Zhijing Jin	Max Planck Institute for Intelligent Systems
	ETH Zurich (Swiss Federal Institute of Technology Zurich)
Antoine Bosselut	EPFL (Swiss Federal Institute of Technology Lausanne)

Table of Contents

<i>Knowledge-Grounded Natural Language Recommendation Explanation</i> Anthony Colas, Jun Araki, Zhengyu Zhou, Bingqing Wang and Zhe Feng
<i>Emergent Linear Representations in World Models of Self-Supervised Sequence Models</i> Neel Nanda, Andrew Lee and Martin Wattenberg
<i>Explaining Data Patterns in Natural Language with Language Models</i> Chandan Singh, John X. Morris, Jyoti Aneja, Alexander Rush and Jianfeng Gao
Probing Quantifier Comprehension in Large Language Models: Another Example of Inverse Scaling Akshat Gupta
Disentangling the Linguistic Competence of Privacy-Preserving BERT Stefan Arnold, Nils Kemmerzell and Annika Schreiner
"Honey, Tell Me What's Wrong", Global Explanation of Textual Discriminative Models through Coop- erative Generation Antoine Chaffin and Julien Delaunay
Self-Consistency of Large Language Models under Ambiguity Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier and Jacob Pfau 89
Character-Level Chinese Backpack Language Models Hao Sun and John Hewitt
Unveiling Multilinguality in Transformer Models: Exploring Language Specificity in Feed-Forward Net- works Sunit Bhattacharya and Ondřej Bojar
Why Bother with Geometry? On the Relevance of Linear Decompositions of Transformer Embeddings Timothee Mickus and Raúl Vázquez
Investigating Semantic Subspaces of Transformer Sentence Embeddings through Linear Structural Prob-
<i>ing</i> Dmitry Nikolaev and Sebastian Padó
Causal Abstraction for Chain-of-Thought Reasoning in Arithmetic Word Problems Juanhe (TJ) Tan
<i>Enhancing Interpretability Using Human Similarity Judgements to Prune Word Embeddings</i> Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot and Uri Hasson
When Your Language Model Cannot Even Do Determiners Right: Probing for Anti-Presuppositions and the Maximize Presupposition! Principle Judith Sieker and Sina Zarrieß180
Introducing VULCAN: A Visualization Tool for Understanding Our Models and Data by Example Jonas Groschwitz
<i>The Self-Contained Negation Test Set</i> David Kletz, Pascal Amsili and Marie Candito

Investigating the Effect of Discourse Connectives on Transformer Surprisal: Language Models Under- stand Connectives, Even So They Are Surprised
Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu and Philippe Blache
METAPROBE: A Representation- and Task-Agnostic Probe Yichu Zhou and Vivek Srikumar 233
How Much Consistency Is Your Accuracy Worth? Jacob K. Johnson and Ana Marasović
<i>Investigating the Encoding of Words in BERT's Neurons Using Feature Textualization</i> Tanja Baeumel, Soniya Vijayakumar, Josef van Genabith, Guenter Neumann and Simon Ostermann
<i>Evaluating Transformer's Ability to Learn Mildly Context-Sensitive Languages</i> Shunjie Wang and Shane Steinert-Threlkeld
Layered Bias: Interpreting Bias in Pretrained Large Language Models Nirmalendu Prakash and Roy Ka-Wei Lee
Not Wacky vs. Definitely Wacky: A Study of Scalar Adverbs in Pretrained Language Models Isabelle Lorge and Janet B. Pierrehumbert
<i>Rigorously Assessing Natural Language Explanations of Neurons</i> Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu and Christopher Potts 317
NPIs Aren't Exactly Easy: Variation in Licensing across Large Language ModelsDeanna DeCarlo, William Palmer, Michael Wilson and Bob Frank332
Memory Injections: Correcting Multi-Hop Reasoning Failures During Inference in Transformer-Based Language Models
Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard and Ian Foster
Systematic Generalization by Finetuning? Analyzing Pretrained Language Models Using Constituency Tests
Aishik Chakraborty, Jackie CK Cheung and Timothy J. O'Donnell
On Quick Kisses and How to Make Them Count: A Study on Event Construal in Light Verb Constructions with BERT
Chenxin Liu and Emmanuele Chersoni
Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language

Model

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna and Oskar van der Wal....379

Workshop Program

Thursday, December 7, 2023

- 9:00–9:10 Opening Remarks
- 9:10–10:00 Invited Talk 1
- 9:10–10:00 Causal NLP: A Path towards Opening the Black Box of NLP Zhijing Jin

10:00–10:30 Session 1 (Orals)

- 10:00–10:15 *Knowledge-Grounded Natural Language Recommendation Explanation* Anthony Colas, Jun Araki, Zhengyu Zhou, Bingqing Wang and Zhe Feng
- 10:15–10:30 *Emergent Linear Representations in World Models of Self-Supervised Sequence Models* Neel Nanda, Andrew Lee and Martin Wattenberg

10:30-11:00 Break

- 11:00–12:30 Session 2 (Posters)
- 11:00–12:30 *Explaining Data Patterns in Natural Language with Language Models* Chandan Singh, John X. Morris, Jyoti Aneja, Alexander Rush and Jianfeng Gao
- 11:00–12:30 Probing Quantifier Comprehension in Large Language Models: Another Example of Inverse Scaling Akshat Gupta
- 11:00–12:30 *Disentangling the Linguistic Competence of Privacy-Preserving BERT* Stefan Arnold, Nils Kemmerzell and Annika Schreiner
- 11:00–12:30 "Honey, Tell Me What's Wrong", Global Explanation of Textual Discriminative Models through Cooperative Generation Antoine Chaffin and Julien Delaunay
- 11:00–12:30 *Self-Consistency of Large Language Models under Ambiguity* Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier and Jacob Pfau
- 11:00–12:30 *Character-Level Chinese Backpack Language Models* Hao Sun and John Hewitt

- 11:00–12:30 Unveiling Multilinguality in Transformer Models: Exploring Language Specificity in Feed-Forward Networks Sunit Bhattacharya and Ondřej Bojar
- 11:00–12:30 Why Bother with Geometry? On the Relevance of Linear Decompositions of Transformer Embeddings Timothee Mickus and Raúl Vázquez
- 11:00–12:30 Investigating Semantic Subspaces of Transformer Sentence Embeddings through Linear Structural Probing Dmitry Nikolaev and Sebastian Padó
- 11:00–12:30 *Causal Abstraction for Chain-of-Thought Reasoning in Arithmetic Word Problems* Juanhe (TJ) Tan
- 11:00–12:30 *Enhancing Interpretability Using Human Similarity Judgements to Prune Word Embeddings* Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot and Uri Hasson
- 11:00–12:30 When Your Language Model Cannot Even Do Determiners Right: Probing for Anti-Presuppositions and the Maximize Presupposition! Principle Judith Sieker and Sina Zarrieß
- 11:00–12:30 Introducing VULCAN: A Visualization Tool for Understanding Our Models and Data by Example Jonas Groschwitz
- 11:00–12:30 *The Self-Contained Negation Test Set* David Kletz, Pascal Amsili and Marie Candito
- 11:00–12:30 Investigating the Effect of Discourse Connectives on Transformer Surprisal: Language Models Understand Connectives, Even So They Are Surprised Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu and Philippe Blache
- 11:00–12:30 *METAPROBE: A Representation- and Task-Agnostic Probe* Yichu Zhou and Vivek Srikumar
- 11:00–12:30 *How Much Consistency Is Your Accuracy Worth?* Jacob K. Johnson and Ana Marasović
- 11:00–12:30 Investigating the Encoding of Words in BERT's Neurons Using Feature Textualization
 Tanja Baeumel, Soniya Vijayakumar, Josef van Genabith, Guenter Neumann and Simon Ostermann
- 11:00–12:30 *Evaluating Transformer's Ability to Learn Mildly Context-Sensitive Languages* Shunjie Wang and Shane Steinert-Threlkeld

- 11:00–12:30 *Layered Bias: Interpreting Bias in Pretrained Large Language Models* Nirmalendu Prakash and Roy Ka-Wei Lee
- 11:00–12:30 Not Wacky vs. Definitely Wacky: A Study of Scalar Adverbs in Pretrained Language Models
 Isabelle Lorge and Janet B. Pierrehumbert
- 11:00–12:30 *The Hidden Space of Transformer Adapters* Jesujoba Alabi, Marius Mosbach, Matan Eyal, Dietrich Klakow and Mor Geva
- 11:00–12:30 Understanding and Mitigating Classification Errors through Interpretable Token Patterns
 Michael Hedderich, Jonas Fischer, Dietrich Klakow and Jilles Vreeken
- 11:00–12:30 Let the Models Respond: Interpreting Language Model Detoxification through the Lens of Prompt Dependence Daniel Scalena, Gabriele Sarti, Malvina Nissim and Elisabetta Fersini
- 11:00–12:30 *Where Exactly Does Contextualization in a PLM Happen?* Soniya Vijayakumar, Simon Ostermann, Tanja Baeumel and Josef van Genabith
- 11:00–12:30 Probabilistic Transformer: A Probabilistic Dependency Model for Contextual Word Representation Haoyi Wu and Kewei Tu

 11:00–12:30 Attention Lens: A Tool for Mechanistically Interpreting the Attention Head Information Retrieval Mechanism
 Mansi Sakarvadia, Arham Khan, Aswathy Ajith, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard and Ian Foster

- 11:00–12:30 *NeuroSurgeon: A Toolkit for Subnetwork Analysis* Michael Lepori, Ellie Pavlick and Thomas Serre
- 11:00–12:30 *Human-Understandable Knowledge Graph Completion* Zhao Xu, Wiem Ben Rim, Kiril Gashteovski, Timo Sztyler and Carolin Lawrence
- 11:00–12:30 Opening the Black-Box from a Regulatory Point of View—A Discussion on the EU AI Act
 Chia-Chien Hung, Wiem Ben Rim, Lindsay Frost, Lars Bruckner and Carolin Lawrence
- 11:00–12:30 *Quantifying the Plausibility of Context Reliance in Neural Machine Translation* Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim and Arianna Bisazza

11:00–12:30 Natural Language Explanation for Tabular Data Zhao Xu, Sascha Saralajew, Kiril Gashteovski and Carolin Lawrence

11:00-12:30	Learning to Add as a Way to Understand Transformers Harish Tayyar Madabushi and Nello Cristianini
11:00-12:30	Flexible Model Interpretability through Natural Language Model Editing Karel D'Oosterlinck, Thomas Demeester, Chris Develder and Christopher Potts
11:00-12:30	Investigating Information Provenance in Retrieval-Augmented Generation Alessandro Stolfo and Ari Kobren
11:00-12:30	Using Collostructional Analysis to Evaluate BERT's Representation of Linguistic Constructions Tim Veenboer and Jelke Bloem
11:00-12:30	Planning in Transformers: Evidence for Backtracking Circuits in Models Trained on Tree Traversal Abhay Sheshadri, Jannik Brinkmann and Victor Levoso
11:00-12:30	Measuring Moral Inconsistencies of Large Language Models Vamshi Krishna Bonagiri, Sreeram Vennam, Manas Gaur and Ponnurangam Ku- maraguru
11:00-12:30	The Architecture Gave the Language Model the Constituent Length Preferences Neil Rathi
11:00-12:30	[Findings] A Causal View of Entity Bias in (Large) Language Models Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou and Muhao Chen
11:00-12:30	[Findings] <i>IRFL: Image Recognition of Figurative Language</i> Ron Yosef, Yonatan Bitton and Dafna Shahaf
11:00-12:30	[Findings] Knowledge is a Region in Weight Space for Fine-tuned Language Models Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz and Leshem Choshen
11:00-12:30	[Findings] Contrastive Learning for Sentence Encoder Induces Word Weighting by Information-Theoretic Quantities Hiroto Kurita, Goro Kobayashi, Sho Yokoi and Kentaro Inui
11:00-12:30	[Findings] VISIT: Visualizing and Interpreting the Semantic Information Flow of Transformers Shahar Katz and Yonatan Belinkov
11:00-12:30	[Findings] Exploring Explainable Automated Student Answer Assessment with ChatGPT Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi and Yulan He

11:00-12:30	[Findings] Verb Conjugation in Transformers Is Determined by Linear Encodings of Subject Number Sophie Hao and Tal Linzen
11:00-12:30	[Findings] Guiding LLM to Fool Itself: Automatically Manipulating Machine Read- ing Comprehension Shortcut Triggers Mosh Levy, Shauli Ravfogel and Yoav Goldberg
11:00-12:30	[Findings] Unnatural language processing: How do language models handle machine-generated prompts? Corentin Kervadec, Francesca Franzon and Marco Baroni
11:00-12:30	[Findings] Towards Concept-Aware Large Language Models Chen Shani, Jilles Vreeken and Dafna Shahaf
11:00-12:30	[Findings] Is Probing All You Need? Indicator Tasks as an Alternative to Probing Embedding Spaces Tal Levy, Omer Goldman and Reut Tsarfaty
11:00-12:30	[Findings] <i>Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-</i> <i>trained Language Models</i> Paul Youssef, Osman Alperen Koraş, Meijie Li, Jörg Schlötterer and Christin Seifert
11:00-12:30	[Findings] Causal Inference from Text: Unveiling Interactions between Variables Yuxiang Zhou and Yulan He
11:00-12:30	[Findings] Learning to Abstract with Nonparametric Variational Information Bot- tleneck Melika Behjati, Fabio James Fehr and James Henderson
11:00-12:30	[Findings] <i>Probing LLMs for Joint Encoding of Linguistic Categories</i> Giulio Starace, Konstantinos Papakostas, Rochelle Choenni, Apostolos Pana- giotopoulos, Matteo Rosati, Alina Leidinger and Ekaterina Shutova
11:00-12:30	[Findings] InterroLang: Exploring NLP Models and Datasets through Dialogue- based Explanations Nils Feldhus, Qianli Wang, Tatiana Anikina, Sahil Chopra, Cennet Oguz and Se- bastian Möller
11:00-12:30	[Findings] Evaluating Dependencies in Fact Editing for Language Models: Speci- ficity and Implication Awareness Zichao Li, Ines Arous, Siva Reddy and Jackie CK Cheung
11:00–12:30	[Findings] NERetrieve: Dataset for Next Generation Named Entity Recognition and Retrieval Uri Katz, Matan Vetzler, Amir David Nissan Cohen and Yoav Goldberg

11:00-12:30	[Findings] <i>Robustness of Named-Entity Replacements for In-Context Learning</i> Saeed Goodarzi, Nikhil Kagita, Dennis Minn, Shufan Wang, Roberto Dessi, Shub- ham Toshniwal, Adina Williams, Jack Lanchantin and Koustuv Sinha
11:00-12:30	[Findings] Can Retriever-Augmented Language Models Reason? The Blame Game Between the Retriever and the Language Model Parishad BehnamGhader, Santiago Miret and Siva Reddy
11:00-12:30	[Findings] Demystifying Prompts in Language Models via Perplexity Estimation Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith and Luke Zettlemoyer
11:00-12:30	[Findings] You Are An Expert Linguistic Annotator: Limits of LLMs as Analyzers of Abstract Meaning Representation Allyson Ettinger, Jena D. Hwang, Valentina Pyatkin, Chandra Bhagavatula and Yejin Choi
11:00-12:30	[Findings] Transparency at the Source: Evaluating and Interpreting Language Models With Access to the True Distribution Jaap Jumelet and Willem Zuidema
11:00-12:30	[Findings] Non-Compositionality in Sentiment: New Data and Analyses Verna Dankers and Christopher G. Lucas

12:30-14:00 Lunch

14:00–15:30 Session 3 (Orals)

- 14:00–14:15 Rigorously Assessing Natural Language Explanations of Neurons
 Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu and Christopher
 Potts
- 14:15–14:30 *NPIs Aren't Exactly Easy: Variation in Licensing across Large Language Models* Deanna DeCarlo, William Palmer, Michael Wilson and Bob Frank

 14:30–14:45 Memory Injections: Correcting Multi-Hop Reasoning Failures During Inference in Transformer-Based Language Models
 Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard and Ian Foster

14:45–15:00 Systematic Generalization by Finetuning? Analyzing Pretrained Language Models Using Constituency Tests Aishik Chakraborty, Jackie CK Cheung and Timothy J. O'Donnell

- 15:00–15:15 On Quick Kisses and How to Make Them Count: A Study on Event Construal in Light Verb Constructions with BERT Chenxin Liu and Emmanuele Chersoni
- 15:15–15:30 Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language Model Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna and Oskar van der Wal

15:30-16:00 Break

- 16:00–16:50 Invited Talk 2
- 16:00–16:50 From Mechanistic Interpretability to Mechanistic Reasoning Antoine Bosselut
- 16:50–17:00 Closing Remarks and Awards
- 17:00–18:00 Panel Discussion