BlackboxNLP 2024

# The 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP

## Proceedings of the Workshop

November 15, 2024

The BlackboxNLP organizers gratefully acknowledge the support from the following sponsors.

**Google**

**Apple**

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-170-4

# 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 seventh 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:

- Explanation methods such as saliency, attribution, free-text explanations, or explanations with structured properties.
- Mechanistic interpretability, reverse engineering approaches to understanding particular properties of neural models.
- Scaling up analysis methods for large language models (LLMs).
- Probing methods for testing whether models have acquired or represent certain linguistic properties.
- Analysing context mixing (e.g., token-to-token interactions) in deep learning architectures.
- Adapting and applying analysis techniques from other disciplines (e.g., neuroscience or computer vision).
- Examining model performance on simplified or formal languages.
- Proposing modifications to neural architectures that increase their interpretability.
- Open-source tools for analysis, visualization, or explanation to democratize access to interpretability techniques in NLP.
- Evaluation of explanation methods: how do we know the explanation is faithful to the model?
- Understanding under the hood of memorization in LLMs.
- Opinion pieces about the state of explainable NLP.

The seventh BlackboxNLP workshop will be held in Miami, Florida on November 15, 2024, hosted by the Conference on Empirical Methods in Natural Language Processing (EMNLP). 35 full papers and 18 non-archival extended abstracts were accepted for in-person and online presentations, from a total of 91 submissions. This year's workshop will also feature papers on interpretability from the Findings of the ACL: EMNLP 2024, as well as two invited talks and a panel discussion with experts in the field. BlackboxNLP 2024 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 Miami and online.

# Organizing Committee

## Organizing Committee

Yonatan Belinkov, Technion-Israel Institute of Technology
Najoung Kim, Boston University
Jaap Jumelet, University of Amsterdam
Hosein Mohebbi, Tilburg University
Aaron Mueller, Northeastern University, Technion
Hanjie Chen, Rice University

# Program Committee

**Program Committee**

Aswathy Ajith, Carolyn Jane Anderson, Leila Arras, Pepa Atanasova

Ioana Baldini, Laurent Besacier, Jannik Brinkmann, Jonathan Brophy, Lisa Bylinina

Ruidi Chang

Anubrata Das, Subham De, Chunyuan Deng

Yanai Elazar

Nils Feldhus, Javier Ferrando, Richard Futrell

Michael Eric Goodale, Sarang Gupta

Tal Haklay, Michael Hanna, Maria Heuss

Robin Jia, Richard Johansson

Jonathan Kamp, Ken Kawamura, Kunal Kukreja, Jenny Kunz, Tatsuki Kuribayashi

Anna Langedijk, Jiaxuan Li, Tomasz Limisiewicz, Yining Lu

Pranav Mani, Jack Merullo, Lama Moukheiber, Mira Moukheiber

Anmol Nayak

Byung-Doh Oh, Hadas Orgad

Lis Pereira, Tiago Pimentel, Yuval Pinter, Charlotte Pouw, Adithya Pratapa

Sara Rajaee, Yasaman Razeghi, Kiamehr Rezaee, Juan Diego Rodriguez, Rudolf Rosa

Gabriele Sarti, Arnab Sen Sharma, Rico Sennrich, Mattia Setzu, Natalie Shapira, Dimitar Shterionov, Sanchit Sinha, Pia Sommerauer, Shane Steinert-Threlkeld

Aarne Talman, Jörg Tiedemann

Saujas Vaduguru, Jithendra Vepa

Lucas Weber

Yichu Zhou

# Keynote Talk

**Jack Merullo**
Brown University

# Keynote Talk

**Himabindu Lakkaraju**
Harvard University

# Table of Contents

# Program

**Friday, November 15, 2024**

| | |
|---|---|
| 09:00 - 09:10 | *Opening Remarks* |
| 09:10 - 10:00 | *Invited Talk 1* |
| 10:00 - 10:30 | *Session 1 (Orals)* |
| | *Routing in Sparsely-gated Language Models responds to Context*<br>Stefan Arnold, Marian Fietta and Dilara Yesilbas |
| | *Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models*<br>Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko and Anna A Ivanova |
| 10:30 - 11:00 | *Break* |
| 11:00 - 12:30 | *Session 2 (Posters)* |
| | *Optimal and efficient text counterfactuals using Graph Neural Networks*<br>Dimitris Lymperopoulos, Maria Lymperaiou, Giorgos Filandrianos and Giorgos Stamou |
| | *Routing in Sparsely-gated Language Models responds to Context*<br>Stefan Arnold, Marian Fietta and Dilara Yesilbas |
| | *Are there identifiable structural parts in the sentence embedding whole?*<br>Vivi Nastase and Paola Merlo |
| | *Learning, Forgetting, Remembering: Insights From Tracking LLM Memorization During Training*<br>Danny D. Leybzon and Corentin Kervadec |
| | *Language Models Linearly Represent Sentiment*<br>Oskar John Hollinsworth, Curt Tigges, Atticus Geiger and Neel Nanda |
| | *LLM Internal States Reveal Hallucination Risk Faced With a Query*<br>Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie and Pascale Fung |

*Enhancing adversarial robustness in Natural Language Inference using explanations*
Alexandros Koulakos, Maria Lymperaiou, Giorgos Filandrianos and Giorgos Stamou

*MultiContrievers: Analysis of Dense Retrieval Representations*
Seraphina Goldfarb-Tarrant, Pedro Rodriguez, Jane Dwivedi-Yu and Patrick Lewis

*Can We Statically Locate Knowledge in Large Language Models? Financial Domain and Toxicity Reduction Case Studies*
Jordi Armengol-Estapé, Lingyu Li, Sebastian Gehrmann, Achintya Gopal, David S Rosenberg, Gideon S. Mann and Mark Dredze

*Attend First, Consolidate Later: On the Importance of Attention in Different LLM Layers*
Amit Ben Artzy and Roy Schwartz

*Enhancing Question Answering on Charts Through Effective Pre-training Tasks*
Ashim Gupta, Vivek Gupta, Shuo Zhang, Yujie He, Ning Zhang and Shalin Shah

*Faithfulness and the Notion of Adversarial Sensitivity in NLP Explanations*
Supriya Manna and Niladri Sett

*Transformers Learn Transition Dynamics when Trained to Predict Markov Decision Processes*
Yuxi Chen, Suwei Ma, Tony Dear and Xu Chen

*On the alignment of LM language generation and human language comprehension*
Lena Sophia Bolliger, Patrick Haller and Lena Ann Jäger

*An Adversarial Example for Direct Logit Attribution: Memory Management in GELU-4L*
Jett Janiak, Can Rager, James Dao and Yeu-Tong Lau

*Uncovering Syllable Constituents in the Self-Attention-Based Speech Representations of Whisper*
Erfan A Shams, Iona Gessinger and Julie Carson-Berndsen

*Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations*
Róbert Csordás, Christopher Potts, Christopher D Manning and Atticus Geiger

*Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models*
Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko and Anna A Ivanova

*Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2*
Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah and Neel Nanda

*Self-Assessment Tests are Unreliable Measures of LLM Personality*
Akshat Gupta, Xiaoyang Song and Gopala Anumanchipalli

*How Language Models Prioritize Contextual Grammatical Cues?*
Hamidreza Amirzadeh, Afra Alishahi and Hosein Mohebbi

*Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads*
Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath and Neel Nanda

*WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions*
Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro and Manas Gaur

*Do Metadata and Appearance of the Retrieved Webpages Affect LLM's Reasoning in Retrieval-Augmented Generation?*
Cheng-Han Chiang and Hung-yi Lee

*Attribution Patching Outperforms Automated Circuit Discovery*
Aaquib Syed, Can Rager and Arthur Conmy

*Pruning for Protection: Increasing Jailbreak Resistance in Aligned LLMs Without Fine-Tuning*
Adib Hasan, Ileana Rugina and Alex Wang

*IvRA: A Framework to Enhance Attention-Based Explanations for Language Models with Interpretability-Driven Training*
Sean Xie, Soroush Vosoughi and Saeed Hassanpour

*Counterfactuals As a Means for Evaluating Faithfulness of Attribution Methods in Autoregressive Language Models*
Sepehr Kamahi and Yadollah Yaghoobzadeh

**Friday, November 15, 2024 (continued)**

*Investigating Layer Importance in Large Language Models*
Yang Zhang, Yanfei Dong and Kenji Kawaguchi

*Mechanistic?*
Naomi Saphra and Sarah Wiegreffe

*Toward the Evaluation of Large Language Models Considering Score Variance across Instruction Templates*
Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito and Taro Watanabe

*Accelerating Sparse Autoencoder Training via Layer-Wise Transfer Learning in Large Language Models*
Davide Ghilardi, Federico Belotti, Marco Molinari and Jaehyuk Lim

*Wrapper Boxes for Faithful Attribution of Model Predictions to Training Data*
Yiheng Su, Junyi Jessy Li and Matthew Lease

*Multi-property Steering of Large Language Models with Dynamic Activation Composition*
Daniel Scalena, Gabriele Sarti and Malvina Nissim

*Probing Language Models on Their Knowledge Source*
Zineddine Tighidet, Jiali Mei, Benjamin Piwowarski and Patrick Gallinari

11:00 - 12:30 *Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly*

11:00 - 12:30 *Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals*

11:00 - 12:30 *Inducing Induction in Llama via Linear Probe Interventions*

11:00 - 12:30 *Implicit Meta-Learning in Small Transformer Models: Insights from a Toy Task*

11:00 - 12:30 *Latent Concept-based Explanation of NLP Models*

11:00 - 12:30 *Exploring Alignment in Shared Cross-Lingual Spaces*

**Friday, November 15, 2024 (continued)**

| | |
|---|---|
| 11:00 - 12:30 | *Compositional Cores: Persistent Attention Patterns in Compositionally Generalizing Subnetworks* |
| 11:00 - 12:30 | *How LLMs Reinforce Political Misinformation: Insights from the Analysis of False Presuppositions* |
| 11:00 - 12:30 | *Does Alignment Tuning Really Break LLMs' Internal Confidence?* |
| 11:00 - 12:30 | *How Does Code Pretraining Affect Language Model Task Performance?* |
| 11:00 - 12:30 | *ToxiSight: Insights Towards Detected Chat Toxicity* |
| 11:00 - 12:30 | *Clusters Emerge in Transformer-based Causal Language Models* |
| 11:00 - 12:30 | *Quantifying reliance on external information over parametric knowledge during Retrieval Augmented Generation (RAG) using mechanistic analysis* |
| 11:00 - 12:30 | *Mind Your Manners: Detoxifying Language Models via Attention Head Intervention* |
| 11:00 - 12:30 | *Can One Token Make All the Difference? Forking Paths in Autoregressive Text Generation* |
| 11:00 - 12:30 | *Exploring the Recall of Language Models: Case Study on Molecules* |
| 11:00 - 12:30 | *How do LLMs deal with Syntactic Conflicts in In-context-learning ?* |
| 11:00 - 12:30 | *Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models* |
| 12:30 - 14:00 | *Lunch* |
| 14:00 - 15:00 | *Invited Talk 2* |
| 15:00 - 15:30 | *Session 3 (Orals)* |
| | *Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2*<br>Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah and Neel Nanda |

*Mechanistic?*
Naomi Saphra and Sarah Wiegreffe

15:30 - 16:00 *Break*

15:30 - 16:30 *Session 2 (Posters)*

*Optimal and efficient text counterfactuals using Graph Neural Networks*
Dimitris Lymperopoulos, Maria Lymperaiou, Giorgos Filandrianos and Giorgos Stamou

*Routing in Sparsely-gated Language Models responds to Context*
Stefan Arnold, Marian Fietta and Dilara Yesilbas

*Are there identifiable structural parts in the sentence embedding whole?*
Vivi Nastase and Paola Merlo

*Learning, Forgetting, Remembering: Insights From Tracking LLM Memorization During Training*
Danny D. Leybzon and Corentin Kervadec

*Language Models Linearly Represent Sentiment*
Oskar John Hollinsworth, Curt Tigges, Atticus Geiger and Neel Nanda

*LLM Internal States Reveal Hallucination Risk Faced With a Query*
Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie and Pascale Fung

*Enhancing adversarial robustness in Natural Language Inference using explanations*
Alexandros Koulakos, Maria Lymperaiou, Giorgos Filandrianos and Giorgos Stamou

*MultiContrievers: Analysis of Dense Retrieval Representations*
Seraphina Goldfarb-Tarrant, Pedro Rodriguez, Jane Dwivedi-Yu and Patrick Lewis

*Can We Statically Locate Knowledge in Large Language Models? Financial Domain and Toxicity Reduction Case Studies*
Jordi Armengol-Estapé, Lingyu Li, Sebastian Gehrmann, Achintya Gopal, David S Rosenberg, Gideon S. Mann and Mark Dredze

*Attend First, Consolidate Later: On the Importance of Attention in Different LLM Layers*
Amit Ben Artzy and Roy Schwartz

*Enhancing Question Answering on Charts Through Effective Pre-training Tasks*
Ashim Gupta, Vivek Gupta, Shuo Zhang, Yujie He, Ning Zhang and Shalin Shah

*Faithfulness and the Notion of Adversarial Sensitivity in NLP Explanations*
Supriya Manna and Niladri Sett

*Transformers Learn Transition Dynamics when Trained to Predict Markov Decision Processes*
Yuxi Chen, Suwei Ma, Tony Dear and Xu Chen

*On the alignment of LM language generation and human language comprehension*
Lena Sophia Bolliger, Patrick Haller and Lena Ann Jäger

*An Adversarial Example for Direct Logit Attribution: Memory Management in GELU-4L*
Jett Janiak, Can Rager, James Dao and Yeu-Tong Lau

*Uncovering Syllable Constituents in the Self-Attention-Based Speech Representations of Whisper*
Erfan A Shams, Iona Gessinger and Julie Carson-Berndsen

*Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations*
Róbert Csordás, Christopher Potts, Christopher D Manning and Atticus Geiger

*Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models*
Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko and Anna A Ivanova

*Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2*
Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah and Neel Nanda

*Self-Assessment Tests are Unreliable Measures of LLM Personality*
Akshat Gupta, Xiaoyang Song and Gopala Anumanchipalli

**Friday, November 15, 2024 (continued)**

*How Language Models Prioritize Contextual Grammatical Cues?*
Hamidreza Amirzadeh, Afra Alishahi and Hosein Mohebbi

*Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads*
Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath and Neel Nanda

*WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions*
Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro and Manas Gaur

*Do Metadata and Appearance of the Retrieved Webpages Affect LLM's Reasoning in Retrieval-Augmented Generation?*
Cheng-Han Chiang and Hung-yi Lee

*Attribution Patching Outperforms Automated Circuit Discovery*
Aaquib Syed, Can Rager and Arthur Conmy

*Pruning for Protection: Increasing Jailbreak Resistance in Aligned LLMs Without Fine-Tuning*
Adib Hasan, Ileana Rugina and Alex Wang

*IvRA: A Framework to Enhance Attention-Based Explanations for Language Models with Interpretability-Driven Training*
Sean Xie, Soroush Vosoughi and Saeed Hassanpour

*Counterfactuals As a Means for Evaluating Faithfulness of Attribution Methods in Autoregressive Language Models*
Sepehr Kamahi and Yadollah Yaghoobzadeh

*Investigating Layer Importance in Large Language Models*
Yang Zhang, Yanfei Dong and Kenji Kawaguchi

*Mechanistic?*
Naomi Saphra and Sarah Wiegreffe

*Toward the Evaluation of Large Language Models Considering Score Variance across Instruction Templates*
Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito and Taro Watanabe

**Friday, November 15, 2024 (continued)**

*Accelerating Sparse Autoencoder Training via Layer-Wise Transfer Learning in Large Language Models*
Davide Ghilardi, Federico Belotti, Marco Molinari and Jaehyuk Lim

*Wrapper Boxes for Faithful Attribution of Model Predictions to Training Data*
Yiheng Su, Junyi Jessy Li and Matthew Lease

*Multi-property Steering of Large Language Models with Dynamic Activation Composition*
Daniel Scalena, Gabriele Sarti and Malvina Nissim

*Probing Language Models on Their Knowledge Source*
Zineddine Tighidet, Jiali Mei, Benjamin Piwowarski and Patrick Gallinari

15:30 - 16:30 *Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly*

15:30 - 16:30 *Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals*

15:30 - 16:30 *Inducing Induction in Llama via Linear Probe Interventions*

15:30 - 16:30 *Implicit Meta-Learning in Small Transformer Models: Insights from a Toy Task*

15:30 - 16:30 *Latent Concept-based Explanation of NLP Models*

15:30 - 16:30 *Exploring Alignment in Shared Cross-Lingual Spaces*

15:30 - 16:30 *Compositional Cores: Persistent Attention Patterns in Compositionally Generalizing Subnetworks*

15:30 - 16:30 *How LLMs Reinforce Political Misinformation: Insights from the Analysis of False Presuppositions*

15:30 - 16:30 *Does Alignment Tuning Really Break LLMs' Internal Confidence?*

15:30 - 16:30 *How Does Code Pretraining Affect Language Model Task Performance?*

**Friday, November 15, 2024 (continued)**

| | |
|---|---|
| 15:30 - 16:30 | *ToxiSight: Insights Towards Detected Chat Toxicity* |
| 15:30 - 16:30 | *Clusters Emerge in Transformer-based Causal Language Models* |
| 15:30 - 16:30 | *Quantifying reliance on external information over parametric knowledge during Retrieval Augmented Generation (RAG) using mechanistic analysis* |
| 15:30 - 16:30 | *Mind Your Manners: Detoxifying Language Models via Attention Head Intervention* |
| 15:30 - 16:30 | *Can One Token Make All the Difference? Forking Paths in Autoregressive Text Generation* |
| 15:30 - 16:30 | *Exploring the Recall of Language Models: Case Study on Molecules* |
| 15:30 - 16:30 | *How do LLMs deal with Syntactic Conflicts in In-context-learning ?* |
| 15:30 - 16:30 | *Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models* |
| 16:30 - 16:40 | *Closing Remarks and Awards* |
| 16:40 - 17:30 | *Panel Discussion* |