BlackboxNLP 2025

The 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP

Proceedings of the Workshop

November 9, 2025

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 979-8-89176-346-3

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 eighth 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:

- Mechanistic interpretability, reverse engineering approaches to understanding particular properties of neural models.
- Understanding how language models use context by measuring their context-mixing processes (e.g., their token-to-token interactions).
- Scaling up analysis methods for large language models (LLMs).
- Probing methods for testing whether models have acquired or represent certain linguistic properties.
- 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.
- Explanation methods such as saliency, attribution, free-text explanations, or explanations with structured properties.
- Evaluation of explanation methods: how do we know the explanation is faithful to the model?
- Evaluation of techniques for steering LLM output behavior
- Uncovering the reasoning processes of LLMs
- Understanding under the hood of memorization in LLMs.
- Insights into LLM Failures
- Opinion pieces about the state of explainable NLP.

The eighth BlackboxNLP workshop will be held in Suzhou, China on November 9, 2025, hosted by the Conference on Empirical Methods in Natural Language Processing (EMNLP). 28 full papers, 26 non-archival extended abstracts, and 3 shared tasks were accepted for in-person and online presentations, from a total of 99 submissions. This year's workshop will also feature two invited talks and a shared task roundtable. BlackboxNLP 2025 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. Our invited speakers, 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 Suzhou and online.

Organizing Committee

Organizing Committee

Dana Arad, Technion
Yonatan Belinkov, Technion
Hanjie Chen, Rice University
Najoung Kim, Boston University
Aaron Mueller, Boston University
Hosein Mohebbi, Tilburg University
Gabriele Sarti, University of Groningen

Program Committee

Reviewers

David Demitri Africa, Aryaman Arora, Leila Arras

Lorenzo Basile, Vamshi Krishna Bonagiri

Ruidi Chang

Verna Dankers, Chunyuan Deng

Zhouxiang Fang, Nils Feldhus

Aryo Pradipta Gema, Hila Gonen, Michael Eric Goodale, Aviral Gupta, Sarang Gupta, Balint Gyevnar

Tal Haklay, Michael Hanna, Maria Heuss

Richard Johansson

Patrick Kahardipraja, Jonathan Kamp, To Eun Kim, Marianne De Heer Kloots, Laura Kopf, Kunal Kukreja, Vinayshekhar Bannihatti Kumar, Jenny Kunz

Sewoong Lee, Alessandro Lenci, Guanlin Li, Sheng Liang, Yang Janet Liu

Vera Neplenbroek, Yaniv Nikankin

Francesco Ortu

Gonçalo Paulo, Lis Pereira, Antonin Poché, Christopher Potts, Charlotte Pouw, Nikhil Prakash

Elena Sofia Ruzzetti

Dong Shu, Sanchit Sinha

Martin Tutek

Dennis Ulmer

Oskar Van Der Wal, Jithendra Vepa

Zhengxuan Wu

Haotian Xia

Qiyuan Yang

Yong Zhang

Keynote Talk

Quanshi Zhang
Shanghai Jiao Tong University

Keynote Talk

Verna Dankers
McGill University

Table of Contents

Program

Sunday, November 9, 2025

09:00 - 09:15 *Opening Remarks*

09:15 - 10:00 *Invited Talk 1*

10:00 - 10:30 *Session 1 (Orals)*

Language Dominance in Multilingual Large Language Models
Nadav Shani and Ali Basirat

Circuit-Tracer: A New Library for Finding Feature Circuits
Michael Hanna, Mateusz Piotrowski, Jack Lindsey and Emmanuel Ameisen

10:30 - 11:00 *Break*

11:00 - 12:00 *Session 2 (Posters)*

CE-Bench: Towards a Reliable Contrastive Evaluation Benchmark of Interpretability of Sparse Autoencoders
Alex Gulko, Yusen Peng and Sachin Kumar

Char-mander Use mBackdoor! A Study of Cross-lingual Backdoor Attacks in Multilingual LLMs
Himanshu Beniwal, Sailesh Panda, Birudugadda Srivibhav and Mayank Singh

Evil twins are not that evil: Qualitative insights into machine-generated prompts
Nathanaël Carraz Rakotonirina, Corentin Kervadec, Francesca Franzon and Marco Baroni

Steering Prepositional Phrases in Language Models: A Case of with-headed Adjectival and Adverbial Complements in Gemma-2
Stefan Arnold and Rene Gröbner

The Comparative Trap: Pairwise Comparisons Amplifies Biased Preferences of LLM Evaluators
Hawon Jeong, ChaeHun Park, Jimin Hong, Hojoon Lee and Jaegul Choo

Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models
Iuri Macocco, Nora Graichen, Gemma Boleda and Marco Baroni

Language Dominance in Multilingual Large Language Models
Nadav Shani and Ali Basirat

Interpreting Language Models Through Concept Descriptions: A Survey
Nils Feldhus and Laura Kopf

Investigating ReLoRA: Effects on the Learning Dynamics of Small Language Models
Yuval Weiss, David Demitri Africa, Paula Buttery and Richard Diehl Martinez

When LRP Diverges from Leave-One-Out in Transformers
Weiqiu You, Siqi Zeng, Yao-Hung Hubert Tsai, Makoto Yamada and Han Zhao

Understanding the Side Effects of Rank-One Knowledge Editing
Ryosuke Takahashi, Go Kamoda, Benjamin Heinzerling, Keisuke Sakaguchi and Kentaro Inui

Emergent Convergence in Multi-Agent LLM Annotation
Angelina Parfenova, Alexander Denzler and Jürgen Pfeffer

PrivacyScalpel: Enhancing LLM Privacy via Interpretable Feature Intervention with Sparse Autoencoders
Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang and Xuebing Zhou

Circuit-Tracer: A New Library for Finding Feature Circuits
Michael Hanna, Mateusz Piotrowski, Jack Lindsey and Emmanuel Ameisen

The Lookahead Limitation: Why Multi-Operand Addition is Hard for LLMs
Tanja Baeumel, Josef Van Genabith and Simon Ostermann

Can LLMs Detect Ambiguous Plural Reference? An Analysis of Split-Antecedent and Mereological Reference
Dang Thi Thao Anh, Rick Nouwen and Massimo Poesio

Normative Reasoning in Large Language Models: A Comparative Benchmark from Logical and Modal Perspectives
Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima and Mitsuhiro Okada

A Theorem-Proving-Based Evaluation of Neural Semantic Parsing
Hayate Funakura, Hyunsoo Kim and Koji Mineshima

A Pipeline to Assess Merging Methods via Behavior and Internals
Yutaro Sigrist and Andreas Waldis

From BERT to LLMs: Comparing and Understanding Chinese Classifier Prediction in Language Models
Ziqi Zhang, Jianfei Ma, Emmanuele Chersoni, You Jieshun and Zhaoxin Feng

Mechanistic Fine-tuning for In-context Learning
Hakaze Cho, Peng Luo, Mariko Kato, Rin Kaenbyou and Naoya Inoue

Understanding How CodeLLMs (Mis)Predict Types with Activation Steering
Francesca Lucchetti and Arjun Guha

The Unheard Alternative: Contrastive Explanations for Speech-to-Text Models
Lina Conti, Dennis Fucci, Marco Gaido, Matteo Negri, Guillaume Wisniewski and Luisa Bentivogli

Exploring Large Language Models' World Perception: A Multi-Dimensional Evaluation through Data Distribution
Zhi Li, Jing Yang and Ying Liu

On the Representations of Entities in Auto-regressive Large Language Models
Victor Morand, Josiane Mothe and Benjamin Piwowarski

Can Language Neuron Intervention Reduce Non-Target Language Output?
Suchun Xie, Hwichan Kim, Shota Sasaki, Kosuke Yamada and Jun Suzuki

Fine-Grained Manipulation of Arithmetic Neurons
Wenyu Du, Rui Zheng, Tongxu Luo, Stephen Chung and Jie Fu

What Features in Prompts Jailbreak LLMs? Investigating the Mechanisms Behind Attacks
Nathalie Maria Kirch, Constantin Niko Weisser, Severin Field, Helen Yannakoudakis and Stephen Casper

Sunday, November 9, 2025 (continued)

BlackboxNLP-2025 MIB Shared Task: Improving Circuit Faithfulness via Better Edge Selection
Yaniv Nikankin, Dana Arad, Itay Itzhak, Anja Reusch, Adi Simhi, Gal Kesten and Yonatan Belinkov

BlackboxNLP-2025 MIB Shared Task: IPE: Isolating Path Effects for Improving Latent Circuit Identification
Nicolò Brunello, Andrea Cerutti, Andrea Sassella and Mark James Carman

BlackboxNLP-2025 MIB Shared Task: Exploring Ensemble Strategies for Circuit Localization Methods
Philipp Mondorf, Mingyang Wang, Sebastian Gerstner, Ahmad Dawar Hakimi, Yihong Liu, Leonor Veloso, Shijia Zhou, Hinrich Schuetze and Barbara Plank

Findings of the BlackboxNLP 2025 Shared Task: Localizing Circuits and Causal Variables in Language Models
Dana Arad, Yonatan Belinkov, Hanjie Chen, Najoung Kim, Hosein Mohebbi, Aaron Mueller, Gabriele Sarti and Martin Tutek

11:00 - 12:00 *Probabilistic Soundness Guarantees in LLM Reasoning Chains*

11:00 - 12:00 *Formal Semantic Control over Language Models*

11:00 - 12:00 *GPT-2 prefers ambiguous utterances following informative contexts*

11:00 - 12:00 *FIRM: Fairness Interventions at Runtime and Model-training*

11:00 - 12:00 *Understanding the Read-Write Functionalities of Gated Neurons in Transformers*

11:00 - 12:00 *ToxiSight: Behavioral Insights for Human-AI Toxicity Annotation*

11:00 - 12:00 *WeightLens: Input-Independent Interpretability for LLM Transcoders*

11:00 - 12:00 *Layer Importance for Mathematical Reasoning is Forged in Pre-Training and Invariant after Post-Training*

11:00 - 12:00 *Neighborhood Selection is Critical for Explaining Large Language Models*

11:00 - 12:00 *Mechanisms of In-Context Syntactic Generalization in Language Models*

Sunday, November 9, 2025 (continued)

11:00 - 12:00 *Accelerating Path Patching with Head Pruning for Efficient Circuit Discovery*

11:00 - 12:00 *T-FIX: Text-Based Explanations with Features Interpretable to eXperts*

11:00 - 12:00 *Can Explainability and Privacy Coexist in Natural Language Processing? An Empirical Study*

11:00 - 12:00 *Characterizing Mamba's Selective Memory using Auto-Encoders*

11:00 - 12:00 *Towards discovering linguistic indicators for misalignment in language models*

11:00 - 12:00 *MIND: Multi-Granular INterpretable Detection of Mental Manipulation*

11:00 - 12:00 *SCOPE: Semantic Entropy Probes for LLM-as-a-Judge*

11:00 - 12:00 *Scratchpad Thinking: Alternation Between Storage and Computation in Latent Reasoning Models*

11:00 - 12:00 *Probing the Diffusion: An Interpretability Toolkit for Text-to-Image Models*

11:00 - 12:00 *Mitigating Sycophancy in Language Models via Sparse Activation Fusion*

11:00 - 12:00 *Mitigating Sycophancy in Language Models via Multi-Layer Activation Steering*

11:00 - 12:00 *Emergent World Beliefs: Exploring Transformers in Stochastic Games*

11:00 - 12:00 *Pivotal Tokens Encode Reasoning Shifts in Large Language Models*

11:00 - 12:00 *Sycophancy as compositions of Atomic Psychometric Traits*

11:00 - 12:00 *Prompt Genotyping: A Large-Scale Meta-Analysis of Linguistic and Structural Features Predictive of LLM Performance*

11:00 - 12:00 *Under the Hood of Graph Transformers: Explanations and Linguistic Probing on ISGs*

Sunday, November 9, 2025 (continued)

12:00 - 13:45 *Lunch*

13:45 - 14:30 *Invited Talk 2*

14:30 - 14:45 *Session 3 (Oral)*

Normative Reasoning in Large Language Models: A Comparative Benchmark from Logical and Modal Perspectives
Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima and Mitsuhiro Okada

14:45 - 15:30 *Session 4 (Panel discussion)*

15:30 - 16:00 *Break*

15:30 - 17:00 *Session 5 (Posters)*

CE-Bench: Towards a Reliable Contrastive Evaluation Benchmark of Interpretability of Sparse Autoencoders
Alex Gulko, Yusen Peng and Sachin Kumar

Char-mander Use mBackdoor! A Study of Cross-lingual Backdoor Attacks in Multilingual LLMs
Himanshu Beniwal, Sailesh Panda, Birudugadda Srivibhav and Mayank Singh

Evil twins are not that evil: Qualitative insights into machine-generated prompts
Nathanaël Carraz Rakotonirina, Corentin Kervadec, Francesca Franzon and Marco Baroni

Steering Prepositional Phrases in Language Models: A Case of with-headed Adjectival and Adverbial Complements in Gemma-2
Stefan Arnold and Rene Gröbner

The Comparative Trap: Pairwise Comparisons Amplifies Biased Preferences of LLM Evaluators
Hawon Jeong, ChaeHun Park, Jimin Hong, Hojoon Lee and Jaegul Choo

Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models
Iuri Macocco, Nora Graichen, Gemma Boleda and Marco Baroni

Language Dominance in Multilingual Large Language Models
Nadav Shani and Ali Basirat

Interpreting Language Models Through Concept Descriptions: A Survey
Nils Feldhus and Laura Kopf

Investigating ReLoRA: Effects on the Learning Dynamics of Small Language Models
Yuval Weiss, David Demitri Africa, Paula Buttery and Richard Diehl Martinez

When LRP Diverges from Leave-One-Out in Transformers
Weiqiu You, Siqi Zeng, Yao-Hung Hubert Tsai, Makoto Yamada and Han Zhao

Understanding the Side Effects of Rank-One Knowledge Editing
Ryosuke Takahashi, Go Kamoda, Benjamin Heinzerling, Keisuke Sakaguchi and Kentaro Inui

Emergent Convergence in Multi-Agent LLM Annotation
Angelina Parfenova, Alexander Denzler and Jürgen Pfeffer

PrivacyScalpel: Enhancing LLM Privacy via Interpretable Feature Intervention with Sparse Autoencoders
Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang and Xuebing Zhou

Circuit-Tracer: A New Library for Finding Feature Circuits
Michael Hanna, Mateusz Piotrowski, Jack Lindsey and Emmanuel Ameisen

The Lookahead Limitation: Why Multi-Operand Addition is Hard for LLMs
Tanja Baeumel, Josef Van Genabith and Simon Ostermann

Can LLMs Detect Ambiguous Plural Reference? An Analysis of Split-Antecedent and Mereological Reference
Dang Thi Thao Anh, Rick Nouwen and Massimo Poesio

Normative Reasoning in Large Language Models: A Comparative Benchmark from Logical and Modal Perspectives
Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima and Mitsuhiro Okada

A Theorem-Proving-Based Evaluation of Neural Semantic Parsing
Hayate Funakura, Hyunsoo Kim and Koji Mineshima

A Pipeline to Assess Merging Methods via Behavior and Internals
Yutaro Sigrist and Andreas Waldis

From BERT to LLMs: Comparing and Understanding Chinese Classifier Prediction in Language Models
Ziqi Zhang, Jianfei Ma, Emmanuele Chersoni, You Jieshun and Zhaoxin Feng

Mechanistic Fine-tuning for In-context Learning
Hakaze Cho, Peng Luo, Mariko Kato, Rin Kaenbyou and Naoya Inoue

Understanding How CodeLLMs (Mis)Predict Types with Activation Steering
Francesca Lucchetti and Arjun Guha

The Unheard Alternative: Contrastive Explanations for Speech-to-Text Models
Lina Conti, Dennis Fucci, Marco Gaido, Matteo Negri, Guillaume Wisniewski and Luisa Bentivogli

Exploring Large Language Models' World Perception: A Multi-Dimensional Evaluation through Data Distribution
Zhi Li, Jing Yang and Ying Liu

On the Representations of Entities in Auto-regressive Large Language Models
Victor Morand, Josiane Mothe and Benjamin Piwowarski

Can Language Neuron Intervention Reduce Non-Target Language Output?
Suchun Xie, Hwichan Kim, Shota Sasaki, Kosuke Yamada and Jun Suzuki

Fine-Grained Manipulation of Arithmetic Neurons
Wenyu Du, Rui Zheng, Tongxu Luo, Stephen Chung and Jie Fu

What Features in Prompts Jailbreak LLMs? Investigating the Mechanisms Behind Attacks
Nathalie Maria Kirch, Constantin Niko Weisser, Severin Field, Helen Yannakoudakis and Stephen Casper

BlackboxNLP-2025 MIB Shared Task: Improving Circuit Faithfulness via Better Edge Selection
Yaniv Nikankin, Dana Arad, Itay Itzhak, Anja Reusch, Adi Simhi, Gal Kesten and Yonatan Belinkov

BlackboxNLP-2025 MIB Shared Task: IPE: Isolating Path Effects for Improving Latent Circuit Identification
Nicolò Brunello, Andrea Cerutti, Andrea Sassella and Mark James Carman

BlackboxNLP-2025 MIB Shared Task: Exploring Ensemble Strategies for Circuit Localization Methods
Philipp Mondorf, Mingyang Wang, Sebastian Gerstner, Ahmad Dawar Hakimi, Yihong Liu, Leonor Veloso, Shijia Zhou, Hinrich Schuetze and Barbara Plank

Findings of the BlackboxNLP 2025 Shared Task: Localizing Circuits and Causal Variables in Language Models
Dana Arad, Yonatan Belinkov, Hanjie Chen, Najoung Kim, Hosein Mohebbi, Aaron Mueller, Gabriele Sarti and Martin Tutek

15:30 - 17:00 *Probabilistic Soundness Guarantees in LLM Reasoning Chains*

15:30 - 17:00 *Formal Semantic Control over Language Models*

15:30 - 17:00 *GPT-2 prefers ambiguous utterances following informative contexts*

15:30 - 17:00 *FIRM: Fairness Interventions at Runtime and Model-training*

15:30 - 17:00 *Understanding the Read-Write Functionalities of Gated Neurons in Transformers*

15:30 - 17:00 *ToxiSight: Behavioral Insights for Human-AI Toxicity Annotation*

15:30 - 17:00 *WeightLens: Input-Independent Interpretability for LLM Transcoders*

15:30 - 17:00 *Layer Importance for Mathematical Reasoning is Forged in Pre-Training and Invariant after Post-Training*

15:30 - 17:00 *Neighborhood Selection is Critical for Explaining Large Language Models*

15:30 - 17:00 *Mechanisms of In-Context Syntactic Generalization in Language Models*

Sunday, November 9, 2025 (continued)

15:30 - 17:00 *Accelerating Path Patching with Head Pruning for Efficient Circuit Discovery*

15:30 - 17:00 *T-FIX: Text-Based Explanations with Features Interpretable to eXperts*

15:30 - 17:00 *Can Explainability and Privacy Coexist in Natural Language Processing? An Empirical Study*

15:30 - 17:00 *Characterizing Mamba's Selective Memory using Auto-Encoders*

15:30 - 17:00 *Towards discovering linguistic indicators for misalignment in language models*

15:30 - 17:00 *MIND: Multi-Granular INterpretable Detection of Mental Manipulation*

15:30 - 17:00 *SCOPE: Semantic Entropy Probes for LLM-as-a-Judge*

15:30 - 17:00 *Scratchpad Thinking: Alternation Between Storage and Computation in Latent Reasoning Models*

15:30 - 17:00 *Probing the Diffusion: An Interpretability Toolkit for Text-to-Image Models*

15:30 - 17:00 *Mitigating Sycophancy in Language Models via Sparse Activation Fusion*

15:30 - 17:00 *Mitigating Sycophancy in Language Models via Multi-Layer Activation Steering*

15:30 - 17:00 *Emergent World Beliefs: Exploring Transformers in Stochastic Games*

15:30 - 17:00 *Pivotal Tokens Encode Reasoning Shifts in Large Language Models*

15:30 - 17:00 *Sycophancy as compositions of Atomic Psychometric Traits*

15:30 - 17:00 *Prompt Genotyping: A Large-Scale Meta-Analysis of Linguistic and Structural Features Predictive of LLM Performance*

15:30 - 17:00 *Under the Hood of Graph Transformers: Explanations and Linguistic Probing on ISGs*

Sunday, November 9, 2025 (continued)

17:00 - 17:20 *Closing Remarks and Awards*