

# **7th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP 2024)**

Miami, Florida, USA  
15 November 2024

ISBN: 979-8-3313-0841-4

**Printed from e-media with permission by:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571

**Some format issues inherent in the e-media version may also appear in this print version.**

Copyright© (2024) by the Association for Computational Linguistics  
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics  
at the address below.

Association for Computational Linguistics  
209 N. Eighth Street  
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006

Fax: 1-570-476-0860

[acl@aclweb.org](mailto:acl@aclweb.org)

**Additional copies of this publication are available from:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571 USA  
Phone: 845-758-0400  
Fax: 845-758-2633  
Email: [curran@proceedings.com](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

## Table of Contents

<i>Optimal and efficient text counterfactuals using Graph Neural Networks</i> Dimitris Lymperopoulos, Maria Lymperaïou, Giorgos Filandrianos and Giorgos Stamou . . . . .	1
<i>Routing in Sparsely-gated Language Models responds to Context</i> Stefan Arnold, Marian Fietta and Dilara Yesilbas . . . . .	15
<i>Are there identifiable structural parts in the sentence embedding whole?</i> Vivi Nastase and Paola Merlo . . . . .	23
<i>Learning, Forgetting, Remembering: Insights From Tracking LLM Memorization During Training</i> Danny D. Leybzon and Corentin Kervadec . . . . .	43
<i>Language Models Linearly Represent Sentiment</i> Oskar John Hollinsworth, Curt Tigges, Atticus Geiger and Neel Nanda . . . . .	58
<i>LLM Internal States Reveal Hallucination Risk Faced With a Query</i> Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie and Pascale Fung . . . . .	88
<i>Enhancing adversarial robustness in Natural Language Inference using explanations</i> Alexandros Koulakos, Maria Lymperaïou, Giorgos Filandrianos and Giorgos Stamou . . . . .	105
<i>MultiContrivers: Analysis of Dense Retrieval Representations</i> Seraphina Goldfarb-Tarrant, Pedro Rodriguez, Jane Dwivedi-Yu and Patrick Lewis . . . . .	118
<i>Can We Statically Locate Knowledge in Large Language Models? Financial Domain and Toxicity Reduction Case Studies</i> Jordi Armengol-Estapé, Lingyu Li, Sebastian Gehrmann, Achintya Gopal, David S Rosenberg, Gideon S. Mann and Mark Dredze . . . . .	140
<i>Attend First, Consolidate Later: On the Importance of Attention in Different LLM Layers</i> Amit Ben Artzy and Roy Schwartz . . . . .	177
<i>Enhancing Question Answering on Charts Through Effective Pre-training Tasks</i> Ashim Gupta, Vivek Gupta, Shuo Zhang, Yujie He, Ning Zhang and Shalin Shah . . . . .	185
<i>Faithfulness and the Notion of Adversarial Sensitivity in NLP Explanations</i> Supriya Manna and Niladri Sett . . . . .	193
<i>Transformers Learn Transition Dynamics when Trained to Predict Markov Decision Processes</i> Yuxi Chen, Suwei Ma, Tony Dear and Xu Chen . . . . .	207
<i>On the alignment of LM language generation and human language comprehension</i> Lena Sophia Bolliger, Patrick Haller and Lena Ann Jäger . . . . .	217
<i>An Adversarial Example for Direct Logit Attribution: Memory Management in GELU-4L</i> Jett Janiak, Can Rager, James Dao and Yeu-Tong Lau . . . . .	232
<i>Uncovering Syllable Constituents in the Self-Attention-Based Speech Representations of Whisper</i> Erfan A Shams, Iona Gessinger and Julie Carson-Berndsen . . . . .	238
<i>Recurrent Neural Networks Learn to Store and Generate Sequences using Non-Linear Representations</i> Róbert Csordás, Christopher Potts, Christopher D Manning and Atticus Geiger . . . . .	248

<i>Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models</i>	
Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko and Anna A Ivanova	
	263
<i>Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2</i>	
Tom Lieberum, Senthoran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah and Neel Nanda	
	278
<i>Self-Assessment Tests are Unreliable Measures of LLM Personality</i>	
Akshat Gupta, Xiaoyang Song and Gopala Anumanchipalli	
	301
<i>How Language Models Prioritize Contextual Grammatical Cues?</i>	
Hamidreza Amirzadeh, Afra Alishahi and Hosein Mohebbi	
	315
<i>Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads</i>	
Callum Stuart McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath and Neel Nanda	
	337
<i>WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions</i>	
Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro and Manas Gaur	
	364
<i>Do Metadata and Appearance of the Retrieved Webpages Affect LLM's Reasoning in Retrieval-Augmented Generation?</i>	
Cheng-Han Chiang and Hung-yi Lee	
	389
<i>Attribution Patching Outperforms Automated Circuit Discovery</i>	
Aaquib Syed, Can Rager and Arthur Conmy	
	407
<i>Pruning for Protection: Increasing Jailbreak Resistance in Aligned LLMs Without Fine-Tuning</i>	
Adib Hasan, Ileana Rugina and Alex Wang	
	417
<i>IvRA: A Framework to Enhance Attention-Based Explanations for Language Models with Interpretability-Driven Training</i>	
Sean Xie, Soroush Vosoughi and Saeed Hassanpour	
	431
<i>Counterfactuals As a Means for Evaluating Faithfulness of Attribution Methods in Autoregressive Language Models</i>	
Sepehr Kamahi and Yadollah Yaghoobzadeh	
	452
<i>Investigating Layer Importance in Large Language Models</i>	
Yang Zhang, Yanfei Dong and Kenji Kawaguchi	
	469
<i>Mechanistic?</i>	
Naomi Saphra and Sarah Wiegrefe	
	480
<i>Toward the Evaluation of Large Language Models Considering Score Variance across Instruction Templates</i>	
Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito and Taro Watanabe	
	499
<i>Accelerating Sparse Autoencoder Training via Layer-Wise Transfer Learning in Large Language Models</i>	
Davide Ghilardi, Federico Belotti, Marco Molinari and Jaehyuk Lim	
	530
<i>Wrapper Boxes for Faithful Attribution of Model Predictions to Training Data</i>	
Yiheng Su, Junyi Jessy Li and Matthew Lease	
	551

<i>Multi-property Steering of Large Language Models with Dynamic Activation Composition</i>	
Daniel Scalena, Gabriele Sarti and Malvina Nissim . . . . .	577
<i>Probing Language Models on Their Knowledge Source</i>	
Zineddine Tighidet, Jiali Mei, Benjamin Piwowski and Patrick Gallinari . . . . .	604