

EMNLP 2026

**The Seventh Workshop on Insights from Negative Results in
NLP**

Proceedings of the Workshop

October 29, 2026

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

Silver



©2026 Association for Computational Linguistics

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 979-8-89176-480-4

Introduction

Publication of negative results is difficult in most fields, and the current focus on benchmark-driven performance improvement exacerbates this situation and implicitly discourages hypothesis-driven research. As a result, the development of NLP models often devolves into a product of tinkering and tweaking, rather than science. Furthermore, it increases the time, effort, and carbon emissions spent on developing and tuning models, as the researchers have little opportunity to learn from what has already been tried and failed.

The mission of the workshop on Insights from Negative Results in NLP is to provide a venue for many kinds of negative results, with the hope that they could yield useful insights and provide a much-needed reality check on the successes of deep learning models in NLP. In particular, we solicit the following types of contributions:

- broadly applicable recommendations for training/fine-tuning, especially if X that didn't work is something that many practitioners would think reasonable to try, and if the demonstration of X's failure is accompanied by some explanation/hypothesis;
- ablation studies of components in previously proposed models, showing that their contributions are different from what was initially reported;
- datasets or probing tasks showing that previous approaches do not generalize to other domains or language phenomena;
- trivial baselines that work suspiciously well for a given task/dataset;
- cross-lingual studies showing that a technique X is only successful for a certain language or language family;
- experiments on (in)stability of the previously published results due to hardware, random initializations, preprocessing pipeline components, etc;
- theoretical arguments and/or proofs for why X should not be expected to work;
- demonstration of issues with under-reporting of training details of pre-trained models, including test data contamination and invalid comparisons.

The seventh iteration of the Workshop on Insights from Negative Results attracted 37 submissions. We accepted 19 papers of these, resulting in a 51% acceptance rate. We hope the workshop will continue to contribute to the many reality-check discussions on progress in NLP. If we do not talk about things that do not work, it is harder to see what the biggest problems are and where the community effort is the most needed.

Organizing Committee

Chairs

João Sedoc, New York University, USA
Aleksandr Drozd, AI-KIP, Japan
Arjun Reddy Akula, Google DeepMind, USA
Raphael Shu, Acenta AI, USA
Anna Rogers, ITU Copenhagen, Denmark
Xiaoli Shen, Microsoft, Japan
Salvatore Giorgi, Vanderbilt University, USA
Adrian Benton, Google, USA

Program Committee

Chairs

João Sedoc, New York University, USA
Aleksandr Drozd, AI-KIP, Japan
Arjun Akula, Google DeepMind, USA
Raphael Shu, Acenta AI, USA
Anna Rogers, ITU Copenhagen, Denmark
Xiaoli Shen, Microsoft, Japan
Salvatore Giorgi, Vanderbilt University, USA
Adrian Benton, Google Research, USA

Program Committee

Sharada Acharya, Khushan Adatiya, Wazir Ali, Abhinav Arora

Stefanie Anna Baby, Susanna Maria Baby, Sanket Badhe, Nihal Balani, João Batista, Uri Berger

Suyash Damle, Aleksandr Drozd

Tamás Ficsor

Adithya V Ganesan, Sheela Ramprakash Goenka, Santhosh Guntupalli, Bhaskar Gurram

Vanya Jauhal

Lakshmojee Koduru, Gaurav Kumar, Laida Kushnareva

Seolhwa Lee, Yifei Li, Zilinghan Li

Afshaan Mazagonwalla, Vilobh Meshram, Kunal Mulwani

Rahul Nagurtha, Juan Navarro-Horniacek

John E. Ortega

Rohit Patwa, Sergey Pletenev, Lahari Poddar, Giovanni Puccetti

Sunny Rai, Tejasvi Ravi

Mario Sanz-Guerrero, Vivek Sardeshmukh, Stephanie Schoch, Erica K. Shimomoto, Mantek Singh, Shashanka Subrahmanya

Emil Vatai, Shubham Vatsal, Shreenivas Venkataramanan, Supriya Vijay

Yichen Wang

Letian Xu

Keynote Talk

Anya Belz

Dublin City University

Keynote Talk

Ryan Boyd

Vanderbilt University

Keynote Talk

Dirk Hovy

Bocconi University

Table of Contents

<i>Geometric Separation Without Functional Specialization: Auditing Subspace Disentanglement in Dual-Encoder LLM Paraphrase Detection</i>	
Govind Arun, Ashish Abraham and Gadha Lekshmi P	1
<i>Exploring Dowker Homology for Sentence Similarity</i>	
Marius Huber and Juri Opitz	7
<i>Does Pragmatic Transfer Help Euphemism Detection? Evidence from Target-Span Marking</i>	
Whitney Poh, Noriko Takahashi, Patrick Lee, Julia Sammartino, Libby Barak, Jing Peng and Anna Feldman	19
<i>Risk-Averse Failure Modes of Jensen-Shannon Distance in RLVR Distributional Alignment</i>	
Sebastian Loftus, Xinpeng Wang, Marie-Catherine De Marneffe and Barbara Plank	33
<i>Controlled Fine-Tuning for Two-Tower Recommenders: Why Gradual Unfreezing Can Fail and Cosine Regularization Suffices</i>	
Gabriel Dershowitz and Djallel Bouneffouf	43
<i>Do Soft Prompts Have Useful Wavelet Structure? Negative Evidence In A Controlled Setup</i>	
Soumyadeep Dey, Srikanth Tenneti, Bamdev Mishra, Shobhit Niranjana and Mayur Datar	48
<i>Decomposing LLM-Judge–Human Agreement: How Much Is Inherited Surface Preference?</i>	
Gaurav Kumar	60
<i>The Flip-Test: A Behavioral Diagnostic for Multi-Agent Debate Failures</i>	
Julia Hu, Alfred Shen and Kumar Lakshminarayanan	70
<i>Activation-Based Active Learning for In-Context Learning: Challenges and Insights</i>	
Yaseen Osman, Geoff Merrett and Stuart Middleton	76
<i>The Architecture Was Doing the Work: Why NLI Debiasing Results Do Not Generalise</i>	
Prabhjot Singh	84
<i>When Learned Context Planning Fails to Beat Strong Retrieval: A Controlled Study of Planning, Routing, and Reranking for Long-Context QA</i>	
Yingrui Li and Han Chen	90
<i>No Optimal Language Set Exists for Multilingual Instruction Tuning: Insights from a Linguistically-Informed Study</i>	
Gürkan Soykan and Gözde Şahin	95
<i>Short-Context Gains, Long-Context Drift: Negative Results for Token-Conditioned RoPE</i>	
Allan Kazakov	102
<i>When Gradient Importance Lies: Adaptive LoRA Rank Allocation Fails Under GRPO</i>	
Yash Sawant	110
<i>Do Not Score the Edited Answer: A Negative Result on Loose Evaluation for Verifiable Instruction Following</i>	
José Luis Delgado	115
<i>Perturbation, Not Provenance: Why Activation Steering Does Not Make a Robust Text Watermark</i>	
Sergey Pletenev	120

<i>PEFT Gains Need Hidden Baselines: Rank and Optimizer Sweeps Change LoRA Conclusions</i>	
Ding Wu	127
<i>When Relation-Free Retrieval Is Not Relation-Aware: A Directionality Stress Test for Linear Graph RAG</i>	
Ding Wu	133

Program

Thursday, October 29, 2026

09:05 - 09:50 *Invited Talk 1: Anya Belz*

09:50 - 10:30 *Oral Session 1: Semantic Representation*

Geometric Separation Without Functional Specialization: Auditing Subspace Disentanglement in Dual-Encoder LLM Paraphrase Detection

Govind Arun, Ashish Abraham and Gadha Lekshmi P

Exploring Dowker Homology for Sentence Similarity

Marius Huber and Juri Opitz

Does Pragmatic Transfer Help Euphemism Detection? Evidence from Target-Span Marking

Whitney Poh, Noriko Takahashi, Patrick Lee, Julia Sammartino, Libby Barak, Jing Peng and Anna Feldman

11:00 - 11:45 *Invited Talk 2: Ryan Boyd*

11:45 - 12:25 *Oral Session 2: Tuning Recipes that Fail*

Risk-Averse Failure Modes of Jensen-Shannon Distance in RLVR Distributional Alignment

Sebastian Loftus, Xinpeng Wang, Marie-Catherine De Marneffe and Barbara Plank

Controlled Fine-Tuning for Two-Tower Recommenders: Why Gradual Unfreezing Can Fail and Cosine Regularization Suffices

Gabriel Dershowitz and Djallel Bouneffouf

Do Soft Prompts Have Useful Wavelet Structure? Negative Evidence In A Controlled Setup

Soumyadeep Dey, Srikanth Tenneti, Bamdev Mishra, Shobhit Niranjana and Mayur Datar

14:05 - 14:50 *Invited Talk 3: Dirk Hovy*

14:50 - 15:30 *Oral Session 3: Test-Time Behavior and Evaluation*

Decomposing LLM-Judge–Human Agreement: How Much Is Inherited Surface Preference?

Gaurav Kumar

Thursday, October 29, 2026 (continued)

The Flip-Test: A Behavioral Diagnostic for Multi-Agent Debate Failures

Julia Hu, Alfred Shen and Kumar Lakshmipathi

Activation-Based Active Learning for In-Context Learning: Challenges and Insights

Yaseen Osman, Geoff Merrett and Stuart Middleton

16:05 - 17:30

Poster Session

The Architecture Was Doing the Work: Why NLI Debiasing Results Do Not Generalise

Prabhjot Singh

When Learned Context Planning Fails to Beat Strong Retrieval: A Controlled Study of Planning, Routing, and Reranking for Long-Context QA

Yingrui Li and Han Chen

No Optimal Language Set Exists for Multilingual Instruction Tuning: Insights from a Linguistically-Informed Study

Gürkan Soykan and Gözde Şahin

Short-Context Gains, Long-Context Drift: Negative Results for Token-Conditioned RoPE

Allan Kazakov

When Gradient Importance Lies: Adaptive LoRA Rank Allocation Fails Under GRPO

Yash Sawant

Do Not Score the Edited Answer: A Negative Result on Loose Evaluation for Verifiable Instruction Following

José Luis Delgado

Perturbation, Not Provenance: Why Activation Steering Does Not Make a Robust Text Watermark

Sergey Pletenev

PEFT Gains Need Hidden Baselines: Rank and Optimizer Sweeps Change LoRA Conclusions

Ding Wu

Thursday, October 29, 2026 (continued)

*When Relation-Free Retrieval Is Not Relation-Aware: A Directionality Stress Test
for Linear Graph RAG*

Ding Wu