

NAACL 2024

**The 2024 Conference of the North American Chapter of the  
Association for Computational Linguistics**

**Proceedings of the Student Research Workshop**

June 18, 2024

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

**SRW Sponsor**



©2024 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](mailto:acl@aclweb.org)

ISBN 979-8-89176-117-9

## Introduction

Welcome to the NAACL 2024 Student Research Workshop.

The NAACL 2024 Student Research Workshop (SRW) is a forum for student researchers in computational linguistics and natural language processing. The workshop provides a unique opportunity for student participants to present their work and receive valuable feedback from the international research community as well as from faculty mentors.

Continuing the tradition of previous student research workshops, we offer archival and non-archival tracks tailored for research papers and thesis proposals. The research paper track welcomes submissions from Ph.D. students, Masters students, and advanced undergraduates, providing a venue to showcase their completed or ongoing work, alongside preliminary results. Additionally, the thesis proposal track caters to advanced Masters and Ph.D. students who have identified their thesis topic, offering them a platform to receive feedback on their proposal and guidance on potential future avenues for their research.

This year, we received 67 submissions in total. We accepted 40 of them, resulting in an acceptance rate of 60%. Out of the 40 accepted papers, 10 were non-archival and 30 are presented in these proceedings.

Mentoring is at the heart of the SRW. In line with previous years, we had a pre-submission mentoring program before the submission deadline. A total of 13 papers participated in the pre-submission mentoring program. This program offered students the opportunity to receive comments from an experienced researcher to improve the writing style and presentation of their submissions.

We are immensely grateful to the National Science Foundation for their sponsorship. Their support has played a significant role in ensuring the success of the conference and facilitating the participation of students from a wide range of backgrounds to attend. We also express our sincere gratitude to the program committee members for their thorough reviews of each paper and all of the mentors for donating their time to provide valuable feedback to the student authors. We are also deeply appreciative of the NAACL 2024 organizing committee, the NAACL 2024 SRW organizing committee, and our faculty advisors Marcos Zampieri, Frank Ferraro, and Swabha Swayamdipta, for their essential guidance which was key to organizing this year's workshop. Lastly, we thank all the student authors for submitting their work and participating in the NAACL 2024 edition of the SRW.

# Organizing Committee

## **Student Research Workshop Student Chairs**

Yang (Trista) Cao, University of Maryland, College Park  
Isabel Papadimitriou, Stanford University  
Anaelia Ovalle, University of California, Los Angeles

## **Faculty Advisors**

Marcos Zampieri, George Mason University  
Frank Ferraro, University of Maryland Baltimore County  
Swabha Swayamdipta, University of Southern California

## Program Committee

### Pre-submission Mentors

Ivan Habernal, Paderborn University, Germany  
Myra Cheng, Stanford University  
Lisa Li, Stanford University  
Sarah Masud, IIT-Delhi  
Aditya Yadavalli, Karya  
Qingcheng Zeng, Northwestern University  
Farhan Samir, University of British Columbia  
Karel D'Oosterlinck, Ghent University / Stanford University  
Pranav Goel, Northeastern University  
Forrest Davis, Colgate University  
Sweta Agrawal, Instituto de Telecomunicações  
Hua Shen, University of Michigan  
Dezhi Ye, Tencent  
Robert Vacareanu, University of Arizona  
Will Held, Georgia Institute of Technology

### Program Committee

Prabhat Agarwal, Stanford University  
Abeer Aldayel, King Saud University  
Parsa Bagherzadeh, McGill University  
Gabriel Bernier-Colborne, National Research Council Canada  
Shaily Bhatt, Carnegie Mellon University  
Eduardo Blanco, University of Arizona  
Ting-Yun Chang, University of Southern California  
Xinyue Chen, Carnegie Mellon University  
Orchid Chetia Phukan, Indraprastha Institute of Information Technology, Delhi  
Xiang Dai, CSIRO  
Forrest Davis, Colgate University  
Chunyuan Deng, Georgia Tech  
Alphaeus Eric Dmonte, George Mason University  
Ritam Dutt, Carnegie Mellon University  
Koel Dutta Chowdhury, Saarland University  
Agnieszka Falenska, University of Stuttgart  
Felix Friedrich, TU Darmstadt  
Usman Gohar, Iowa State University  
Dhiman Goswami, George Mason University  
Venkata Subrahmanyam Govindarajan, University of Texas at Austin  
Ivan Habernal, Ruhr-Universität Bochum  
Pengfei He, University of Washington  
Yu Hou, University of Maryland, College Park  
Labiba Jahan, Southern Methodist University  
Harshit Joshi, Stanford University  
Abhinav Joshi, Indian Institute of Technology, Kanpur  
Lee Kezar, University of Southern California  
Dayeon Ki, University of Maryland, College Park

Fajri Koto, Mohamed bin Zayed University of Artificial Intelligence  
Mascha Kurpicz-Briki, BFH - Bern University of Applied Sciences  
Siyang Li, Columbia University  
Bryan Li, University of Pennsylvania  
Jasy Suet Yan Liew, Universiti Sains Malaysia  
Shicheng Liu, Stanford University  
Yanchen Liu, Harvard University  
Bruno Martins, Instituto Superior Técnico  
Sarah Masud, Indraprastha Institute of Information Technology Delhi  
Sandeep Mathias, Presidency University  
Tsvetomila Mihaylova, Aalto University  
Niloofer Mireshghallah, University of Washington  
Shubhanshu Mishra, shubhanshu.com  
Prakanya Mishra, AMD AI  
Masaaki Nagata, NTT Corporation  
Ranjita Naik, Microsoft  
Vincent Nguyen, CSIRO's Data61  
Isabel Papadimitriou, Stanford University  
Tanmay Parekh, University of California Los Angeles  
Rebecca Pattichis, University of California Los Angeles  
Adithya Pratapa, Carnegie Mellon University  
Dongqi Pu, Universität des Saarlandes  
Sunny Rai, School of Engineering and Applied Science, University of Pennsylvania  
Md Nishat Raihan, George Mason University  
Vyas Raina, University of Cambridge  
Lina Maria Rojas-Barahona, Orange Labs  
Hossein Rouhizadeh, University of Geneva  
Shubhashis Roy Dipta, University of Maryland, Baltimore County  
Sashank Santhanam, Apple  
Ryohei Sasano, Nagoya University  
Aditya Shah, Virginia Polytechnic Institute and State University  
Chenglei Si, Stanford University  
Vivek Srivastava, Tata Consultancy Services Limited, India  
Marija Stanojevic, WinterLightLabs  
Ashima Suvarna, University of California, Los Angeles  
Evgeniia Tokarchuk, University of Amsterdam  
Sowmya Vajjala, National Research Council Canada  
Sai P Vallurupalli, University of Maryland at Baltimore County  
Andrea Varga, Theta Lake Ltd  
Francielle Vargas, University of São Paulo  
Prashanth Vijayaraghavan, IBM Research  
Bonnie Webber, Edinburgh University, University of Edinburgh  
Jiannan Xu, University of Maryland, College Park  
Aditya Yadavalli, Karya Inc  
Dezhi Ye, Tencent PCG  
Qinyuan Ye, University of Southern California  
Tsung Yen Yeh, Shein  
Qingcheng Zeng, Northwestern University, Northwestern University  
Mike Zhang, Aalborg University

## Table of Contents

<i>Systematic Analysis for Pretrained Language Model Priming for Parameter-Efficient Fine-tuning</i> Shih-Cheng Huang, Shih-Heng Wang, Min-Han Shih, Saurav Sahay and Hung-yi Lee	1
<i>Rephrasing Invokes Better Generations for Large Language Models</i> Haoran Yang, Hongyuan Lu and Wai Lam	8
<i>Exploring Compositional Generalization of Large Language Models</i> Haoran Yang, Hongyuan Lu, Wai Lam and Deng Cai	16
<i>Explainable CED: A Dataset for Explainable Critical Error Detection in Machine Translation</i> Dahyun Jung, Sugyeong Eo, Chanjun Park and Heuseok Lim	25
<i>SMARTR: A Framework for Early Detection using Survival Analysis of Longitudinal Texts</i> Jean-Thomas Baillargeon and Luc Lamontagne	36
<i>Fast Exact Retrieval for Nearest-neighbor Lookup (FERN)</i> Richard Zhu	42
<i>Start Simple: Progressive Difficulty Multitask Learning</i> Yunfei Luo, Yuyang Liu, Rukai Cai and Tauhidur Rahman	48
<i>LUCID: LLM-Generated Utterances for Complex and Interesting Dialogues</i> Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor and Anders Johannsen	56
<i>Fine-tuning Pre-trained Named Entity Recognition Models For Indian Languages</i> Sankalp Bahad, Pruthwik Mishra, Parameswari Krishnamurthy and Dipti Sharma	75
<i>Knowledge-centered conversational agents with a drive to learn</i> Selene Baez Santamaria	83
<i>Exploring Inherent Biases in LLMs within Korean Social Context: A Comparative Analysis of ChatGPT and GPT-4</i> Seungyeon Lee, Dong Jun Kim, Dahyun Jung, Chanjun Park and Heuseok Lim	93
<i>To Clarify or not to Clarify: A Comparative Analysis of Clarification Classification with Fine-Tuning, Prompt Tuning, and Prompt Engineering</i> Alina Leippert, Tatiana Anikina, Bernd Kiefer and Josef Van Genabith	105
<i>Detecting Response Generation Not Requiring Factual Judgment</i> Ryohei Kamei, Daiki Shiono, Reina Akama and Jun Suzuki	116
<i>Unknown Script: Impact of Script on Cross-Lingual Transfer</i> Wondimagegnhue Tufa, Ilia Markov and Piek Vossen	124
<i>Improving Repository-level Code Search with Text Conversion</i> Mizuki Kondo, Daisuke Kawahara and Toshiyuki Kurabayashi	130
<i>Improving Multi-lingual Alignment Through Soft Contrastive Learning</i> Minsu Park, Seyeon Choi, Chanyeol Choi, Jun-Seong Kim and Jy-yong Sohn	138
<i>Few-Shot Event Argument Extraction Based on a Meta-Learning Approach</i> Aboubacar Tuo, Romaric Besançon, Olivier Ferret and Julien Tourille	146



<i>Investigating Web Corpus Filtering Methods for Language Model Development in Japanese</i> Rintaro Enomoto, Arseny Tolmachev, Takuro Niitsuma, Shuhei Kurita and Daisuke Kawahara	154
<i>Referring Expressions in Human-Robot Common Ground: A Thesis Proposal</i> Jaap Kruijt	161
<i>Source Code is a Graph, Not a Sequence: A Cross-Lingual Perspective on Code Clone Detection</i> Mohammed Ataur Rahaman and Julia Ive	168
<i>Distilling Text Style Transfer With Self-Explanation From LLMs</i> Chiyu Zhang, Honglong Cai, Yuezhong Li, Yuexin Wu, Le Hou and Muhammad Abdul-Mageed	200
<i>Reinforcement Learning for Edit-Based Non-Autoregressive Neural Machine Translation</i> Hao Wang, Tetsuro Morimura, Ukyo Honda and Daisuke Kawahara	212
<i>Evaluation Dataset for Japanese Medical Text Simplification</i> Koki Horiguchi, Tomoyuki Kajiwara, Yuki Arase and Takashi Ninomiya	219
<i>Multi-Source Text Classification for Multilingual Sentence Encoder with Machine Translation</i> Reon Kajikawa, Keiichiro Yamada, Tomoyuki Kajiwara and Takashi Ninomiya	226
<i>A Reproducibility Study on Quantifying Language Similarity: The Impact of Missing Values in the URIEL Knowledge Base</i> Hasti Toossi, Guo Qing Huai, Jinyu Liu, Eric Khiu, A. Seza Doğruöz and En-Shiun Annie Lee	233
<i>Coding Open-Ended Responses using Pseudo Response Generation by Large Language Models</i> Yuki Zenimoto, Ryo Hasegawa, Takehito Utsuro, Masaharu Yoshioka and Noriko Kando	242
<i>Cross-Task Generalization Abilities of Large Language Models</i> Qinyuan Ye	255
<i>Commentary Generation from Data Records of Multiplayer Strategy Esports Game</i> Zihan Wang and Naoki Yoshinaga	263
<i>Facilitating Opinion Diversity through Hybrid NLP Approaches</i> Michiel Van Der Meer	272
<i>HybridBERT - Making BERT Pretraining More Efficient Through Hybrid Mixture of Attention Mechanisms</i> Gokul Srinivasagan and Simon Ostermann	285

# Program

**Tuesday, June 18, 2024**

09:30 - 10:30 *Panel Discussion for Starting Researchers*

15:30 - 17:00 *In-Person Paper Poster Session*

*SMARTR: A Framework for Early Detection using Survival Analysis of Longitudinal Texts*

Jean-Thomas Baillargeon and Luc Lamontagne

*LUCID: LLM-Generated Utterances for Complex and Interesting Dialogues*

Joe Stacey, Jianpeng Cheng, John Torr, Tristan Guigue, Joris Driesen, Alexandru Coca, Mark Gaynor and Anders Johannsen

*Detecting Response Generation Not Requiring Factual Judgment*

Ryohei Kamei, Daiki Shiono, Reina Akama and Jun Suzuki

*Unknown Script: Impact of Script on Cross-Lingual Transfer*

Wondimagegnhue Tufa, Ilia Markov and Piek Vossen

*Improving Repository-level Code Search with Text Conversion*

Mizuki Kondo, Daisuke Kawahara and Toshiyuki Kurabayashi

*Few-Shot Event Argument Extraction Based on a Meta-Learning Approach*

Aboubacar Tuo, Romaric Besançon, Olivier Ferret and Julien Tourille

*Investigating Web Corpus Filtering Methods for Language Model Development in Japanese*

Rintaro Enomoto, Arseny Tolmachev, Takuro Niitsuma, Shuhei Kurita and Daisuke Kawahara

*Reinforcement Learning for Edit-Based Non-Autoregressive Neural Machine Translation*

Hao Wang, Tetsuro Morimura, Ukyo Honda and Daisuke Kawahara

*Evaluation Dataset for Japanese Medical Text Simplification*

Koki Horiguchi, Tomoyuki Kajiwara, Yuki Arase and Takashi Ninomiya

*Multi-Source Text Classification for Multilingual Sentence Encoder with Machine Translation*

Reon Kajikawa, Keiichiro Yamada, Tomoyuki Kajiwara and Takashi Ninomiya

**Tuesday, June 18, 2024 (continued)**

*Coding Open-Ended Responses using Pseudo Response Generation by Large Language Models*

Yuki Zenimoto, Ryo Hasegawa, Takehito Utsuro, Masaharu Yoshioka and Noriko Kando

*Commentary Generation from Data Records of Multiplayer Strategy Esports Game*

Zihan Wang and Naoki Yoshinaga

15:30 - 17:00

*In-Person Thesis Proposal*

*Knowledge-centered conversational agents with a drive to learn*

Selene Baez Santamaria

*Referring Expressions in Human-Robot Common Ground: A Thesis Proposal*

Jaap Kruijt

*Cross-Task Generalization Abilities of Large Language Models*

Qinyuan Ye

*Facilitating Opinion Diversity through Hybrid NLP Approaches*

Michiel Van Der Meer