

WikiNLP 2024

**The First Workshop on Advancing Natural Language  
Processing for Wikipedia**

**Proceedings of the Workshop**

November 16, 2024

©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 979-8-89176-188-9

# Program Committee

## Program Chairs

Angela Fan, Facebook  
Tajuddeen Gwadabe, Masakhane Research Foundation  
Isaac Johnson, Wikimedia Foundation  
Lucie-Aimée Kaffee, Hugging Face  
Fabio Petroni, Samaya AI  
Daniel van Strien, Hugging Face

## Reviewers

Saied Alshahrani, Pablo Aragón, Hiba Arnaout, Akhil Arora, Arnav Arora  
  
Bonaventure F. P. Dossou  
  
Srihari Jayakumar, Isaac Johnson  
  
Nithish Kannen  
  
Kartik Mathur, Jeanna Matthews  
  
Tiziano Piccardi  
  
Miriam Redi  
  
Marija Sakota, Sina Semnani, Indira Sen, Diego Sáez Trumper  
  
Harold Triedman, Mykola Trokhymovych, Houcemeddine Turki  
  
Thejas Venkatesh

## Table of Contents

<i>BordIRlines: A Dataset for Evaluating Cross-lingual Retrieval Augmented Generation</i> Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe and Chris Callison-Burch .....	1
<i>Multi-Label Field Classification for Scientific Documents using Expert and Crowd-sourced Knowledge</i> Rebecca Gelles and James Dunham .....	14
<i>Uncovering Differences in Persuasive Language in Russian versus English Wikipedia</i> Bryan Li, Aleksey Panasyuk and Chris Callison-Burch .....	21
<i>Retrieval Evaluation for Long-Form and Knowledge-Intensive Image-Text Article Composition</i> Jheng-Hong Yang, Carlos Lassance, Rafael S. Rezende, Krishna Srinivasan, Stéphane Clinchant and Jimmy Lin .....	36
<i>WikiBias as an Extrapolation Corpus for Bias Detection</i> K. Salas-Jimenez, Francisco Fernando Lopez-Ponce, Sergio-Luis Ojeda-Trueba and Gemma Bel- Enguix .....	46
<i>HOAXPEDIA: A Unified Wikipedia Hoax Articles Dataset</i> Hsuvas Borkakoty and Luis Espinosa-Anke .....	53
<i>The Rise of AI-Generated Content in Wikipedia</i> Creston Brooks, Samuel Eggert and Denis Peskoff .....	67
<i>Embedded Topic Models Enhanced by Wikification</i> Takashi Shibuya and Takehito Utsuro .....	80
<i>Wikimedia data for AI: a review of Wikimedia datasets for NLP tasks and AI-assisted editing</i> Isaac Johnson, Lucie-Aimée Kaffee and Miriam Redi .....	91
<i>Blocks Architecture (BloArk): Efficient, Cost-Effective, and Incremental Dataset Architecture for Wiki- pedia Revision History</i> Lingxi Li, Zonghai Yao, Sunjae Kwon and Hong Yu .....	102
<i>ARMADA: Attribute-Based Multimodal Data Augmentation</i> Xiaomeng Jin, Jeonghwan Kim, Yu Zhou, Kuan-Hao Huang, Te-Lin Wu, Nanyun Peng and Heng Ji .....	112
<i>Summarization-Based Document IDs for Generative Retrieval with Language Models</i> Alan Li, Daniel Cheng, Phillip Keung, Jungo Kasai and Noah A. Smith .....	126

# Program

## Saturday, November 16, 2024

- 09:00 - 09:05     *Opening Remarks*
- 09:05 - 09:50     *Keynote (Jess Wade)*
- 09:50 - 10:30     *Videos*
- 10:30 - 11:00     *Break*
- 11:00 - 12:00     *Poster Session*
- 12:00 - 12:45     *Lunch*
- 12:45 - 13:30     *Keynote (Scott A. Hale)*
- 13:30 - 14:15     *Panel (Misinformation)*
- 14:15 - 15:00     *Panel (Impact of LLMs)*
- 15:00 - 15:30     *Closing*