RAG4Reports 2026

The First Workshop on Multilingual Report Generation via Retrieval Augmented Generation (RAG4Reports)

Proceedings of the Workshop

July 4, 2026

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 979-8-89176-417-0

Preface

Welcome to the First Workshop on Multilingual Report Generation via Retrieval Augmented Generation (RAG4Reports), a half-day workshop co-located with ACL 2026 in San Diego, CA, USA.

> *A workshop on multilingual long-form report generation focusing on evaluating faithfulness and information coverage, tying together retrieval, generation, and evaluation research.*

Incorporating external source information with parametric knowledge from large language models to provide a comprehensive and grounded response to the users has become a central component in modern AI applications. *Report generation* is a long-form retrieval-augmented generation (RAG) task with strict attestation requirements that makes it well-suited to explore questions of RAG evaluation and generation: a long-form report summarizing the relevant information in a corpus is produced in response to a report request, and the generated report should provide proper attribution to source documents to establish trust. RAG4Reports 2026 focuses on two urgent problems in report generation that require a community effort to tackle: *evaluation* and *multilinguality*. While numerous research papers on RAG have been published in recent years, including proposed evaluation approaches, the community has yet to reach a consensus on benchmarks and evaluation measures for long-form, citation-grounded outputs. At the same time, as generation models are increasingly multilingual and capable of incorporating source information in different languages, report generation systems should be fair to different languages and produce reports that the user can consume from relevant sources regardless of languages. The aim of this workshop is to bring together researchers from information retrieval, natural language processing, and applied domains to establish common ground for developing and evaluating multilingual report generation systems.

The workshop solicited contributions through two tracks. The *research track* received 11 submissions through OpenReview, of which 9 were accepted; of these, one was subsequently withdrawn and two were accepted as non-archival submissions. The *shared task track* attracted 21 registered teams, from which we received 8 system description papers. The proceedings therefore contain 14 papers (6 research and 8 shared task), and 16 accepted papers are presented at the workshop as posters, with a subset selected for oral presentation.

A central component of this workshop is the RAG4Reports shared task, which builds on the 2024 NeuCLIR[1] and 2025 RAGTIME[2] shared tasks at the Text Retrieval Conference (TREC) and extends them to the natural language understanding community, with an explicit focus on non-English languages. The shared task was hosted on TIRA[3], which enables reproducible system evaluation through containerized submissions.

The shared task consists of two sub-tasks. *Task A: Automatic Report Evaluation* asks participants to rank system-generated reports from the 2025 TREC RAGTIME submissions against human annotations, accepting both fully automatic and semi-automatic submissions (the latter using organizer-provided essential facts). *Task B: Multilingual Report Generation* asks participants to generate long-form, citation-grounded reports from a corpus of four million English, Chinese, Russian, and Arabic documents sampled from Common Crawl News (2021–2024), where each generated sentence must be attributable to a cited source and the report must match the language of the request. Submissions to Task B were scored automatically with the ARGUE framework and its automatic implementation, Auto-ARGUE[4], which combines nugget coverage (whether information units expected of a good report are present) with sentence support (whether each generated sentence is faithful to its cited evidence).

We are delighted to host a keynote talk by Professor Chris Callison-Burch (University of Pennsylvania) on rubric-based evaluation of LLM-generated text, which speaks to the central theme of the RAG4Reports Workshop. Prof Callison-Burch has more than 200 publications, which have been cited

[1] https://neuclir.github.io/
[2] https://trec-ragtime.github.io/
[3] https://www.tira.io/
[4] https://github.com/hltcoe/auto-argue

over 36,000 times. He is a Sloan Research Fellow, and he has received faculty research awards from Google, Microsoft, Amazon, Facebook, and Roblox, in addition to funding from DARPA, IARPA, and the NSF.

This workshop would not have been possible without the contributions of many people. We thank the program committee and external reviewers for their careful and timely feedback; the shared task participants for engaging with the benchmark and pushing its limits; and the TIRA team, particularly Maik Fröbe, for hosting and supporting our evaluation infrastructure. We thank the ACL 2026 workshop chairs and publication chairs for their guidance throughout the organization process. Finally, we thank all of the authors and shared task participants who submitted their work.

We hope you enjoy the workshop and the discussions it sparks.

The RAG4Reports 2026 Workshop Organizers
Eugene Yang, Dawn Lawrie, Sean MacAvaney, James Mayfield,
Luca Soldaini, and Andrew Yates

Organizing Committee

Workshop Organizers

Eugene Yang, Human Language Technology Center of Excellence, Johns Hopkins University
Dawn Lawrie, Human Language Technology Center of Excellence, Johns Hopkins University
Sean MacAvaney, University of Glasgow
James Mayfield, Human Language Technology Center of Excellence, Johns Hopkins University, and Johns Hopkins University Applied Physics Laboratory
Luca Soldaini, Microsoft AI
Andrew Yates, Human Language Technology Center of Excellence, Johns Hopkins University

Program Committee

Program Chairs

Eugene Yang, Human Language Technology Center of Excellence, Johns Hopkins University
Dawn Lawrie, Human Language Technology Center of Excellence, Johns Hopkins University
Sean MacAvaney, University of Glasgow
James Mayfield, Human Language Technology Center of Excellence, Johns Hopkins University, and Johns Hopkins University Applied Physics Laboratory
Luca Soldaini, Microsoft AI
Andrew Yates, Human Language Technology Center of Excellence, Johns Hopkins University

Reviewers

Andreas Chari, University of Glasgow
Susmita Das, University of Glasgow
Maxime Dassen, University of Amsterdam
Pooja Jhunjhunwala, Google
Kangheng Liang, University of Glasgow
Emmanouil Georgios Lionis, University of Glasgow
Andrew Parry, University of Glasgow
Saron Samuel, Johns Hopkins University
Siddharth AK Singh, University of Amsterdam

Keynote Talk

Autorubric: A Unified Framework for Rubric-Based LLM Evaluation

Chris Callison-Burch
University of Pennsylvania

Abstract: LLM-as-a-judge has become the default for evaluating open-ended generation, but the approach is riddled with silent failure modes, including position bias, verbosity bias, criterion conflation, sycophancy, and run-to-run inconsistency, that corrupt judgments without any visible signal. Mitigations exist, scattered across the LM-as-judge literature and decades of work in psychometrics and educational measurement, but every research group ends up paying a "Reinvention Tax," reimplementing option shuffling, ensemble voting, calibration, and reliability metrics from scratch.
I will present Autorubric, an open-source framework that consolidates these best practices into a single library with opinionated defaults: analytic per-criterion decomposition, mixed criterion types, ensemble judging, length penalties, and a full suite of psychometric reliability metrics. Beyond measurement, Autorubric's mandatory per-criterion explanations function as "textual gradients" for two downstream applications: rubric-guided prompt induction and RL with rubric rewards. Autorubric is available at https://autorubric.org.

Bio: Chris Callison-Burch is the Raj and Neera Singh Professor of Artificial Intelligence at the University of Pennsylvania, where he directs the online Master's in AI and teaches Penn Engineering's flagship AI course to more than 500 students each fall. In 2026 he received the Lindback Award for Distinguished Teaching, Penn's highest teaching honor. He chairs the advisory board for the Human Language Technology Center of Excellence at Johns Hopkins University. He testified before Congress in 2023 on generative AI and copyright law, and in 2026 participated in the Isaac Asimov Memorial Debate at the American Museum of Natural History, moderated by Neil deGrasse Tyson. He has authored more than 200 publications with over 36,000 citations, and is a Sloan Research Fellow with research support from DARPA, IARPA, NSF, and industry partners including Google, Microsoft, and Amazon.

Table of Contents

Program

Saturday, July 4, 2026

09:00 - 09:15 *Opening Remarks and Shared Task Introduction*

09:15 - 10:00 *Keynote: Chris Callison-Burch (University of Pennsylvania)*

10:00 - 10:30 *Research Paper Orals*

Decompose, Retrieve, Cite: A RAG Pipeline for Structured Report Generation from Technical Documentation
Himanshu Dhurve, Sreedath Panat, Rajat Dandekar and Raj Dandekar

REFSafE: A RAG-Enabled Framework for Predictive Risk Analysis and Automated Safety Report Generation in Mission-Critical Environments
Sanjay Das, Ran Elgedawy, Ethan Seefried, Ryan A. Burchfield, Gavin Wiggins, Dana Hewit, Sudarshan Srinivasan, Prasanna Balaprakash, Robert M. Patton, Todd Thomas and Tirthankar Ghosal

10:30 - 11:00 *Coffee Break*

11:00 - 11:30 *Shared Task Orals*

GenAIus at RAG4Reports 2026: Citation-Aware Compression for Multilingual Report Generation
Reyyan Yeniterzi and Suveyda Yeniterzi

UNH @ Rag4Reports: A Broad Exploration of LLM-Judges for RAG
Minna Tran, Ryan McCarthy, Aiden Parsons, Jaren Unzen and Laura Dietz

Crucible @ Rag4Reports: Generating Nuggets for Report Generation and Evaluation
Laura Dietz and Eugene Yang

11:40 - 12:30 *Poster Session*

StructSurvey: Structured Agentic Retrieval for Automated Survey Paper Generation
Paolo Pedinotti and Enrico Santus

DEO: Training-Free Direct Embedding Optimization for Negation-Aware Retrieval
Taegyeong Lee, Jiwon Park, Seunghyun Hwang and JooYoung Jang

Saturday, July 4, 2026 (continued)

EFSG: Evidence-First Structured Generation for Multilingual RAG Report Generation
Shaurya Gupta and Jatin Bedi

Adapting AutoARGUE for Automatic Report Evaluation under Missing Citation Annotations
Divrose Kaur, Jatin Bedi and Jasmeet Singh

JU-NLP-PG at RAG4Reports 2026: Memory-Efficient Multilingual Report Generation with 4-bit Quantized LLMs
Swayam Chatterjee and Dipankar Das