CMPU 366 · Natural Language Processing

# Natural Language Processing

1 December 2025

## Briefly, in the time remaining:

Some project information

Course wrap-up and reflection

# Project information

| | | |
|---|---|---|
| Mon. | Dec. 1 | You are here |
| Wed. | Dec. 3 | **Project presentations** |
| Mon. | Dec. 8 | **Project presentations** |
| Wed. | Dec. 10 | Last day of classes – but not for us; it’s a “Friday” |
| | | 11:59 p.m. **Project draft due** |
| Sun. | Dec. 14 | End of study period |
| | | 11:59 p.m. **Project due** |

# Presentations

Describing work in progress

Think of it as a chance to tell me and your classmates the cool things you’re working on – and to get some feedback and questions while you still have time to address them!

| | |
|---|---|
| Team of 1–2: | 5–6 minutes |
| Team of 3: | Up to 8 minutes |

We want to have maximum time for presentations and minimize the time switching between groups, so I’ll have all the slides ready to go on one computer.

Please send me your slides (Google Slides, PDF, PowerPoint, Keynote) before the start of the class where you’re presenting.

# What have we learned?

“The idea of giving computers the ability to process human language is as old as the idea of computers themselves.”

Jurafsky & Martin, 2nd ed.

*Tasks*

*Techniques*

Tasks

Techniques

*Tokenization*

| *Tasks* | *Techniques* |
|---|---|
| *Tokenization* | *Regular expressions* |

| *Tasks* | *Techniques* |
|---|---|
| *Tokenization* | *Regular expressions* |
|  | *Rule-based systems* |

## *Tasks*

*Tokenization*

## *Techniques*

*Regular expressions*

*BPE*

*Rule-based systems*

## *Tasks*

*Tokenization*

*Sentence segmentation*

## *Techniques*

*Regular expressions*

*BPE*

*Rule-based systems*

## *Tasks*

*Tokenization*

*Sentence segmentation*

*Representation learning*

## *Techniques*

*Regular expressions*

*BPE*

*Rule-based systems*

## *Tasks*

*Tokenization*

*Sentence segmentation*

*Representation learning*

## *Techniques*

*Regular expressions*

*Rule-based systems*

*Word count embeddings*

*BPE*

| Tasks | Techniques | |
|---|---|---|
| Tokenization | Regular expressions | BPE |
| Sentence segmentation | Rule-based systems | |
| Representation learning | Word count embeddings | Word2vec |

| *Tasks* | *Techniques* | |
|---|---|---|
| *Tokenization* | *Regular expressions* | *BPE* |
| *Sentence segmentation* | *Rule-based systems* | |
| *Representation learning* | *Word count embeddings* | *Word2vec* |
| *Classification*<br>*Sentiment analysis, author identification, etc.* | | |

## *Tasks*

*Tokenization*

*Sentence segmentation*

*Representation learning*

*Classification*

*Sentiment analysis, author identification, etc.*

## *Techniques*

*Regular expressions* *BPE*

*Rule-based systems*

*Word count embeddings* *Word2vec*

*Logistic regression*

## Tasks

Tokenization

Sentence segmentation

Representation learning

Classification

Sentiment analysis, author identification, etc.

## Techniques

Regular expressions

BPE

Rule-based systems

Word count embeddings

Word2vec

Logistic regression

Neural networks

| *Tasks* | *Techniques* | |
|---|---|---|
| *Tokenization* | *Regular expressions* | *BPE* |
| *Sentence segmentation* | *Rule-based systems* | |
| *Representation learning* | *Word count embeddings* | *Word2vec* |
| *Classification*<br>*Sentiment analysis, author identification, etc.* | *Logistic regression* | *Neural networks* |
| *Language modeling* | | |

| Tasks | Techniques | |
|---|---|---|
| Tokenization | Regular expressions | BPE |
| Sentence segmentation | Rule-based systems | |
| Representation learning | Word count embeddings | Word2vec |
| Classification<br>Sentiment analysis, author identification, etc. | Logistic regression | Neural networks |
| Language modeling | N-grams | |

| Tasks | Techniques | |
|---|---|---|
| Tokenization | Regular expressions | BPE |
| Sentence segmentation | Rule-based systems | |
| Representation learning | Word count embeddings | Word2vec |
| Classification<br>Sentiment analysis, author identification, etc. | Logistic regression | Neural networks |
| Language modeling | N-grams | Transformers |

## Tasks

Tokenization

Sentence segmentation

Representation learning

Classification

Sentiment analysis, author identification, etc.

Language modeling

Text generation

## Techniques

Regular expressions

Rule-based systems

Word count embeddings

Logistic regression

N-grams

BPE

Word2vec

Neural networks

Transformers

## Tasks

*Tokenization*

*Sentence segmentation*

*Representation learning*

*Classification*

*Sentiment analysis, author identification, etc.*

*Language modeling*

*Text generation*

## Techniques

*Regular expressions*

*Rule-based systems*

*Word count embeddings*

*Logistic regression*

*N-grams*

*Chaining*

*BPE*

*Word2vec*

*Neural networks*

*Transformers*

## Tasks

*Tokenization*

*Sentence segmentation*

*Representation learning*

*Classification*

*Sentiment analysis, author identification, etc.*

*Language modeling*

*Text generation*

## Techniques

*Regular expressions*

*BPE*

*Rule-based systems*

*Word count embeddings*

*Word2vec*

*Logistic regression*

*Neural networks*

*N-grams*

*Transformers*

*Chaining*

*Autoregressive models*

# Skills

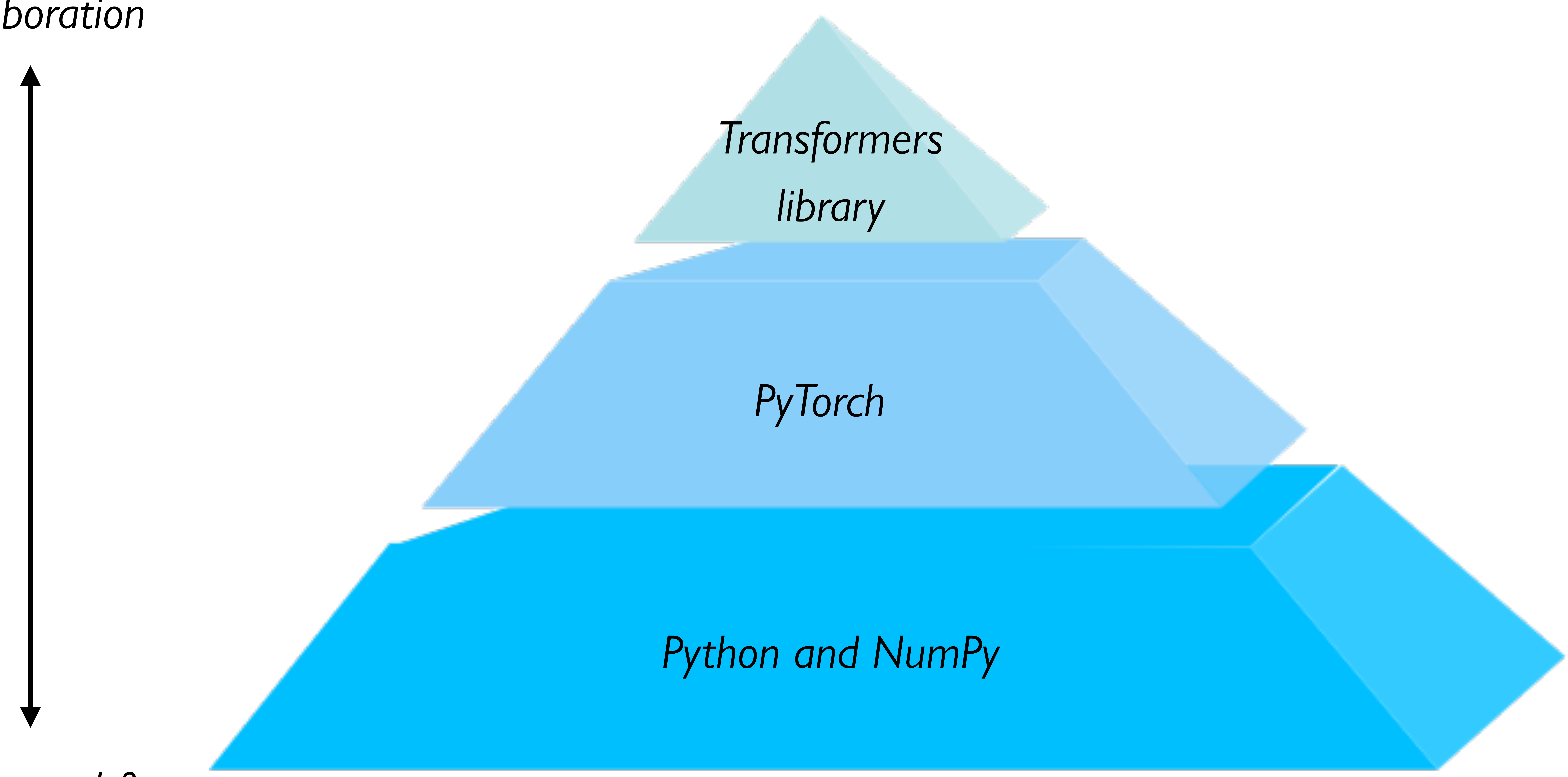
*Text processing*: Working with strings, reading in text datasets, tokenizing, cleaning text, regular expressions

*Evaluation*: Computing accuracy, precision, recall, F-score, perplexity, and cosine similarity

*Implementing classification models*

*Building* and *running* neural network models

More abstraction & collaboration



Transformers library

PyTorch

Python and NumPy

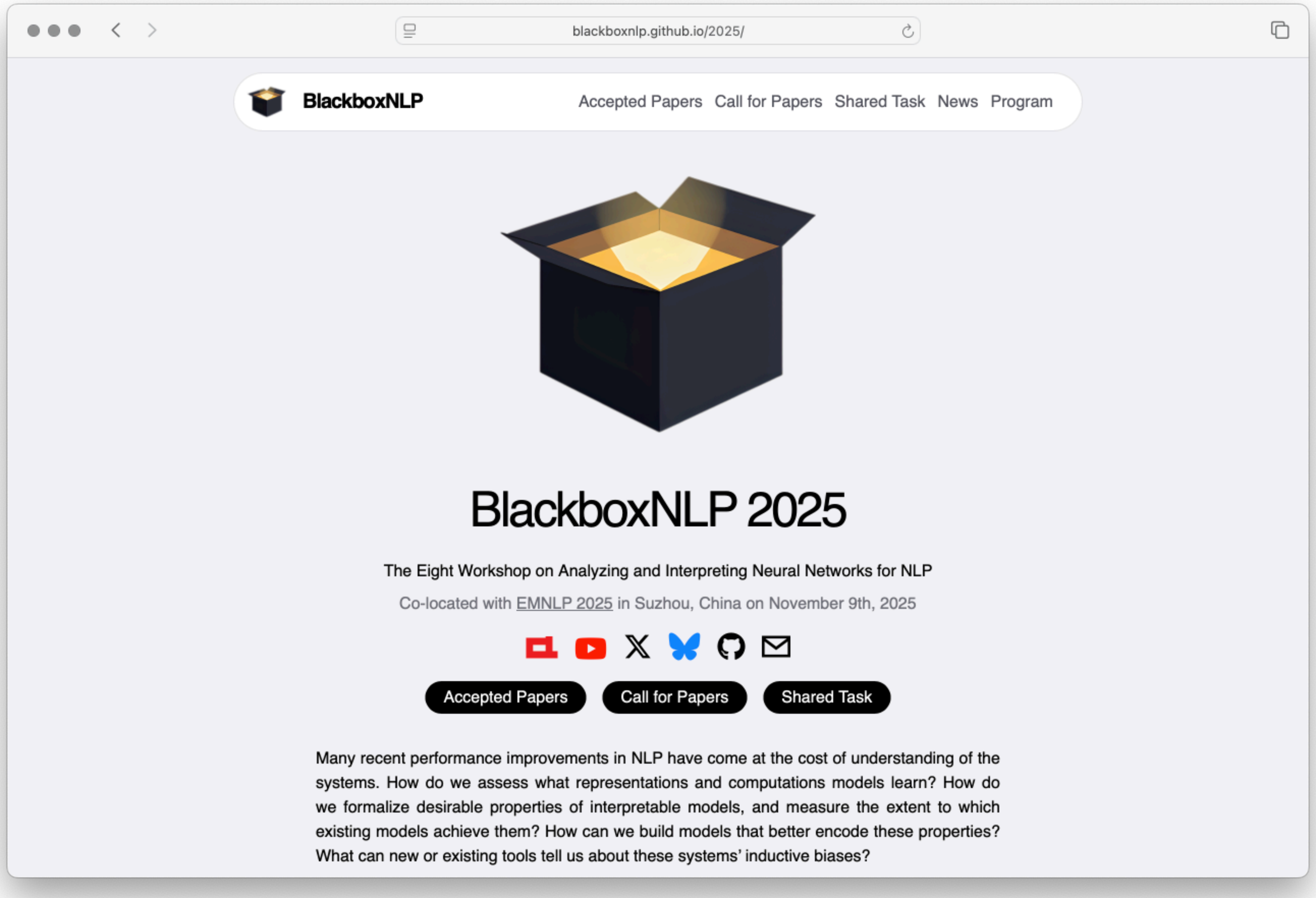
More control & understanding

# What is there still to learn?

# Interpretability and explainability

How do blackbox models make decisions?

What are large language models learning, and how do they use this information?



BlackboxNLP 2025

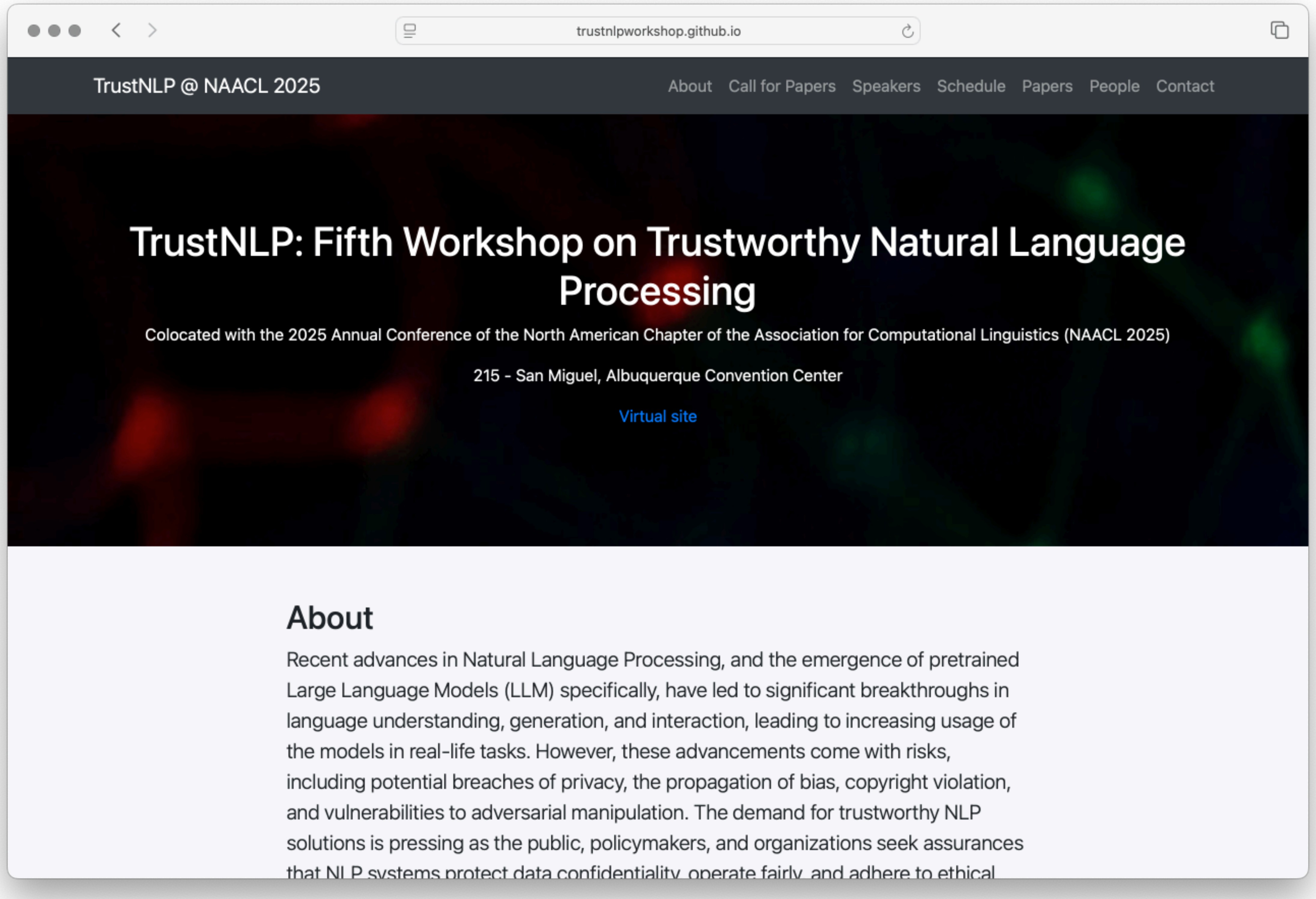
The Eight Workshop on Analyzing and Interpreting Neural Networks for NLP

Co-located with EMNLP 2025 in Suzhou, China on November 9th, 2025

Many recent performance improvements in NLP have come at the cost of understanding of the systems. How do we assess what representations and computations models learn? How do we formalize desirable properties of interpretable models, and measure the extent to which existing models achieve them? How can we build models that better encode these properties? What can new or existing tools tell us about these systems' inductive biases?

What are the limits of what large language models can do?

How can we intervene when models are incorrect?



## TrustNLP: Fifth Workshop on Trustworthy Natural Language Processing

Colocated with the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025)

215 - San Miguel, Albuquerque Convention Center

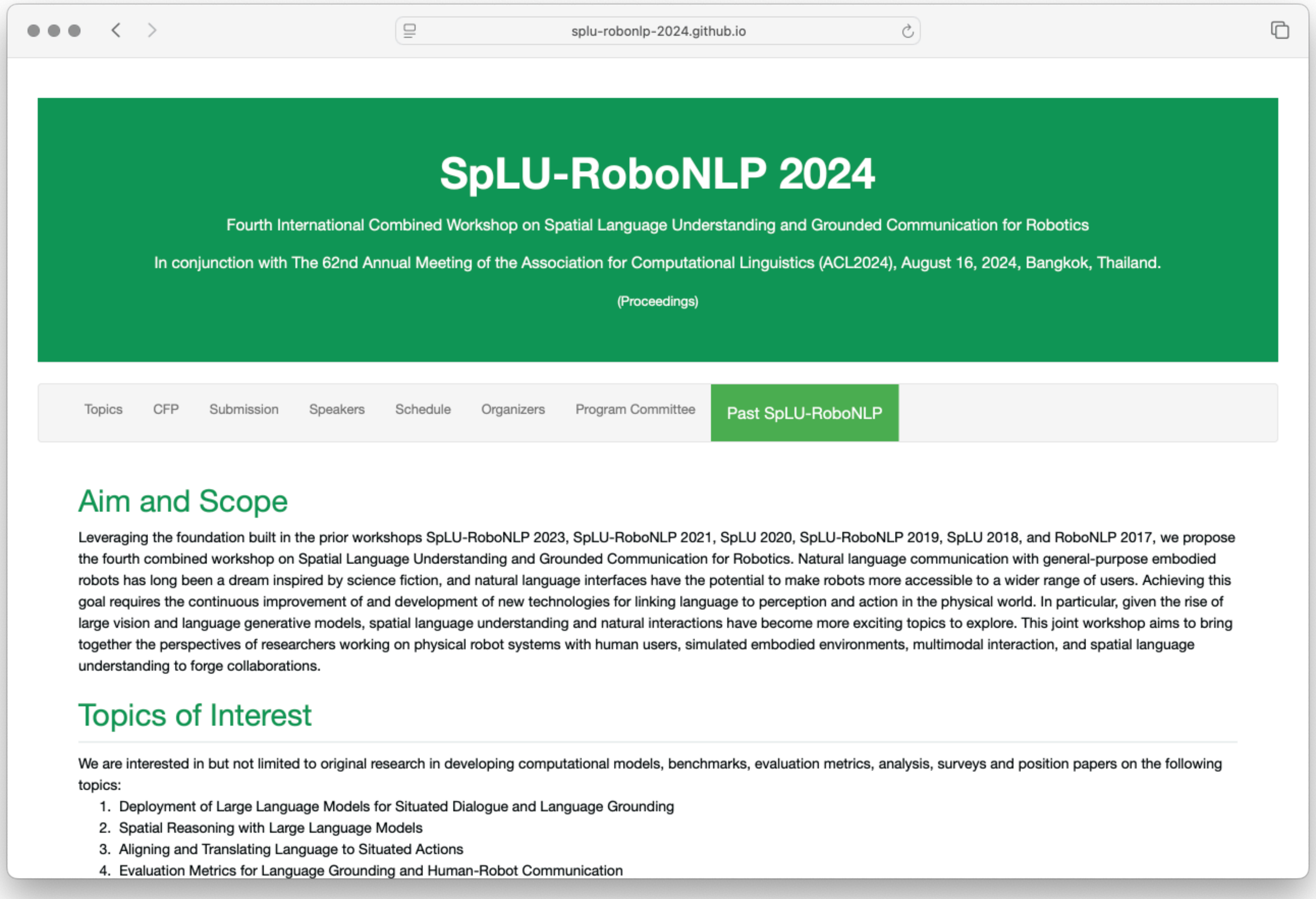
### About

Recent advances in Natural Language Processing, and the emergence of pretrained Large Language Models (LLM) specifically, have led to significant breakthroughs in language understanding, generation, and interaction, leading to increasing usage of the models in real-life tasks. However, these advancements come with risks, including potential breaches of privacy, the propagation of bias, copyright violation, and vulnerabilities to adversarial manipulation. The demand for trustworthy NLP solutions is pressing as the public, policymakers, and organizations seek assurances that NLP systems protect data confidentiality, operate fairly, and adhere to ethical

# Grounding and multimodality

How do we combine text-based models with other modalities, like vision and perceptual information?

How can we help models relate to the physical world?



# SpLU-RoboNLP 2024

Fourth International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics

In conjunction with The 62nd Annual Meeting of the Association for Computational Linguistics (ACL2024), August 16, 2024, Bangkok, Thailand.

(Proceedings)

## Aim and Scope

Leveraging the foundation built in the prior workshops SpLU-RoboNLP 2023, SpLU-RoboNLP 2021, SpLU 2020, SpLU-RoboNLP 2019, SpLU 2018, and RoboNLP 2017, we propose the fourth combined workshop on Spatial Language Understanding and Grounded Communication for Robotics. Natural language communication with general-purpose embodied robots has long been a dream inspired by science fiction, and natural language interfaces have the potential to make robots more accessible to a wider range of users. Achieving this goal requires the continuous improvement of and development of new technologies for linking language to perception and action in the physical world. In particular, given the rise of large vision and language generative models, spatial language understanding and natural interactions have become more exciting topics to explore. This joint workshop aims to bring together the perspectives of researchers working on physical robot systems with human users, simulated embodied environments, multimodal interaction, and spatial language understanding to forge collaborations.

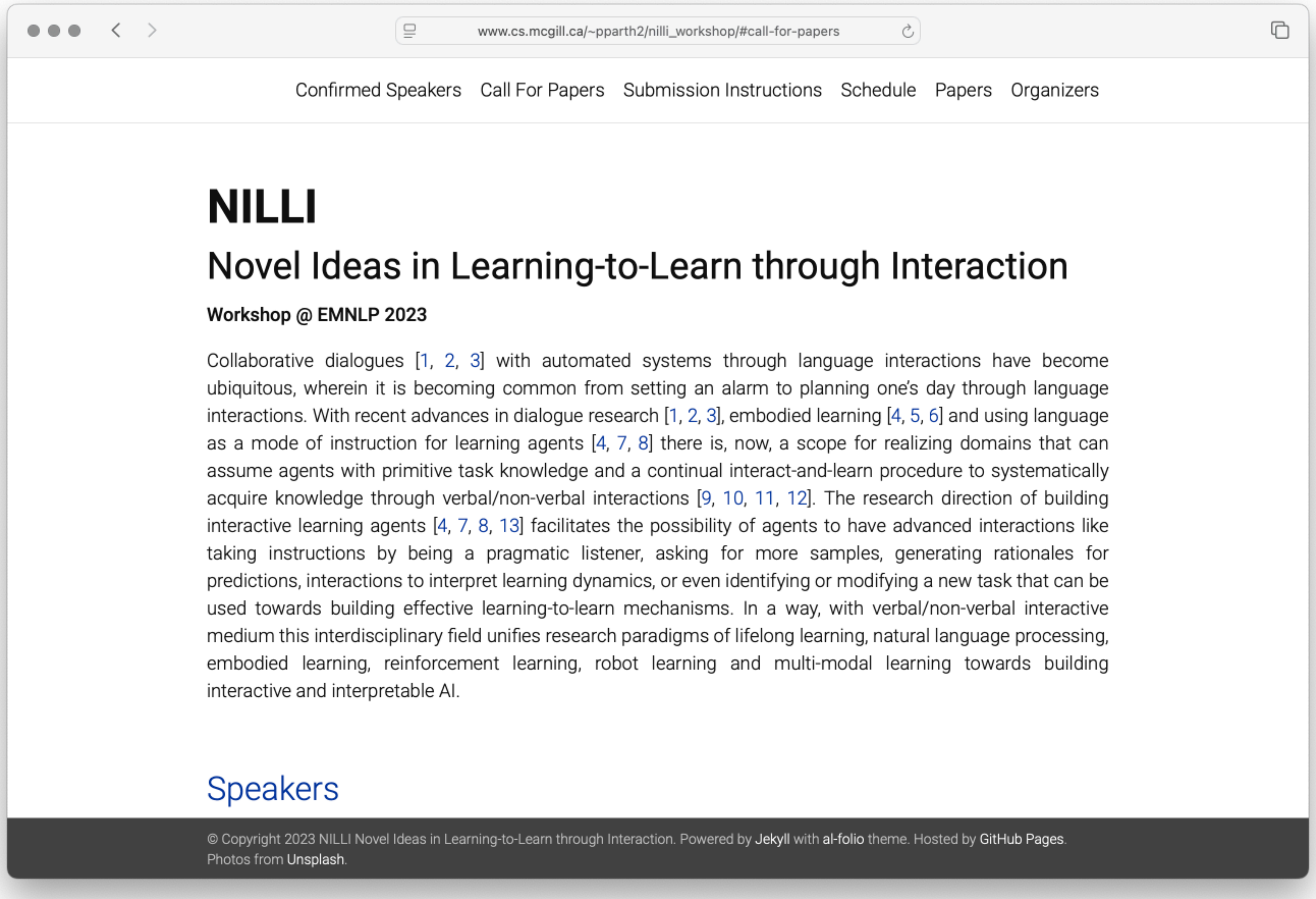
## Topics of Interest

We are interested in but not limited to original research in developing computational models, benchmarks, evaluation metrics, analysis, surveys and position papers on the following topics:

1. Deployment of Large Language Models for Situated Dialogue and Language Grounding
2. Spatial Reasoning with Large Language Models
3. Aligning and Translating Language to Situated Actions
4. Evaluation Metrics for Language Grounding and Human-Robot Communication

How do we help models understand social interaction?

How much of a role does access to perceptual information play in learning?



# NILLI

## Novel Ideas in Learning-to-Learn through Interaction

**Workshop @ EMNLP 2023**

Collaborative dialogues [1, 2, 3] with automated systems through language interactions have become ubiquitous, wherein it is becoming common from setting an alarm to planning one's day through language interactions. With recent advances in dialogue research [1, 2, 3], embodied learning [4, 5, 6] and using language as a mode of instruction for learning agents [4, 7, 8] there is, now, a scope for realizing domains that can assume agents with primitive task knowledge and a continual interact-and-learn procedure to systematically acquire knowledge through verbal/non-verbal interactions [9, 10, 11, 12]. The research direction of building interactive learning agents [4, 7, 8, 13] facilitates the possibility of agents to have advanced interactions like taking instructions by being a pragmatic listener, asking for more samples, generating rationales for predictions, interactions to interpret learning dynamics, or even identifying or modifying a new task that can be used towards building effective learning-to-learn mechanisms. In a way, with verbal/non-verbal interactive medium this interdisciplinary field unifies research paradigms of lifelong learning, natural language processing, embodied learning, reinforcement learning, robot learning and multi-modal learning towards building interactive and interpretable AI.

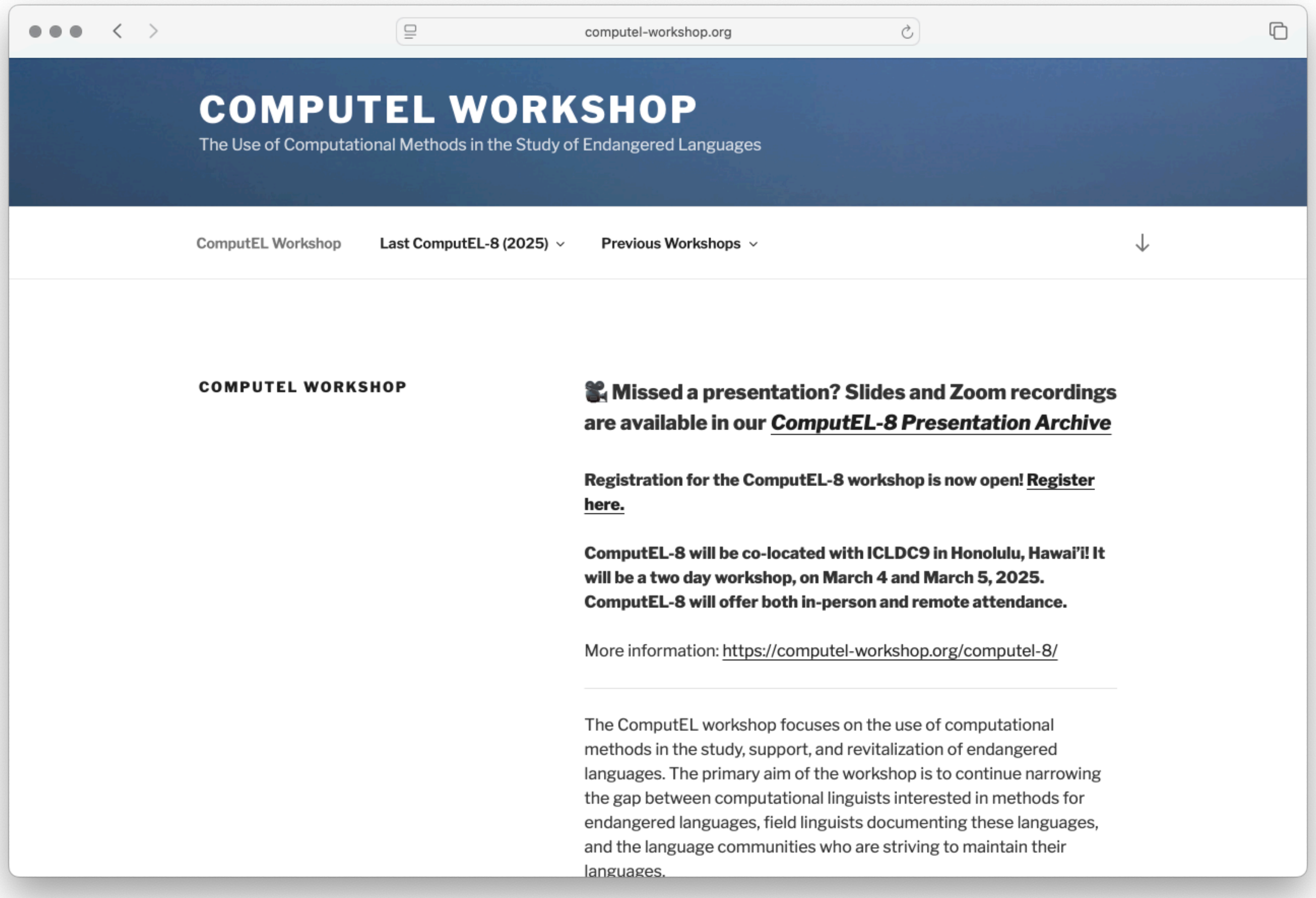
## Speakers

© Copyright 2023 NILLI Novel Ideas in Learning-to-Learn through Interaction. Powered by Jekyll with al-folio theme. Hosted by GitHub Pages.
Photos from Unsplash.

# Linguistic diversity

How can we improve NLP tools for low-resource languages?

How do models deal with different language varieties?



COMPUTEL WORKSHOP

The Use of Computational Methods in the Study of Endangered Languages

COMPUTEL WORKSHOP

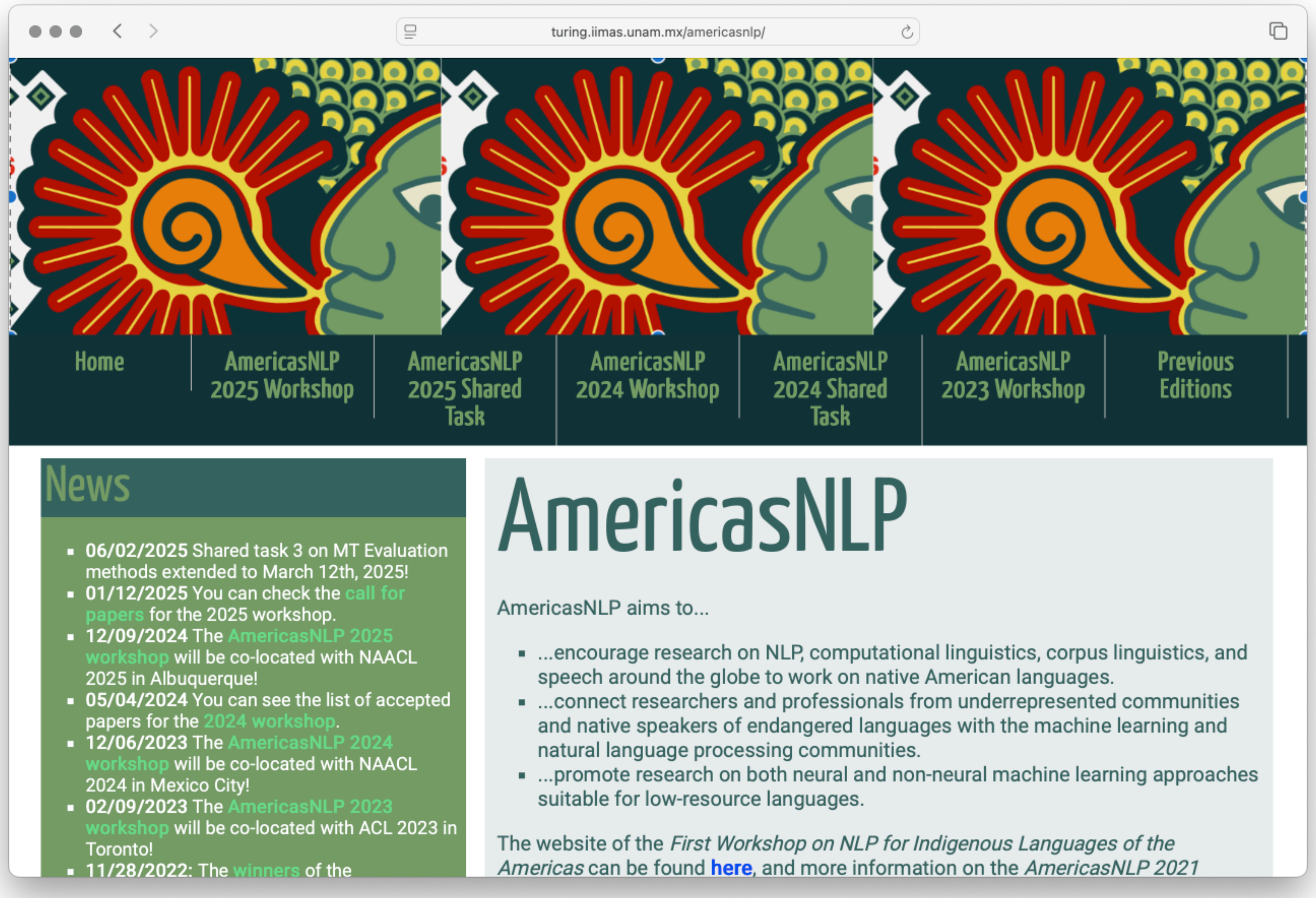
**🎥 Missed a presentation? Slides and Zoom recordings are available in our *ComputEL-8 Presentation Archive***

**Registration for the ComputEL-8 workshop is now open! Register here.**

**ComputEL-8 will be co-located with ICLDC9 in Honolulu, Hawai'i! It will be a two day workshop, on March 4 and March 5, 2025. ComputEL-8 will offer both in-person and remote attendance.**

More information: https://computel-workshop.org/computel-8/

The ComputEL workshop focuses on the use of computational methods in the study, support, and revitalization of endangered languages. The primary aim of the workshop is to continue narrowing the gap between computational linguists interested in methods for endangered languages, field linguists documenting these languages, and the language communities who are striving to maintain their languages.


turing.iimas.unam.mx/americasnlp/
Home
AmericasNLP 2025 Workshop
AmericasNLP 2025 Shared Task
AmericasNLP 2024 Workshop
AmericasNLP 2024 Shared Task
AmericasNLP 2023 Workshop
Previous Editions

News

- **06/02/2025** Shared task 3 on MT Evaluation methods extended to March 12th, 2025!
- **01/12/2025** You can check the call for papers for the 2025 workshop.
- **12/09/2024** The AmericasNLP 2025 workshop will be co-located with NAACL 2025 in Albuquerque!
- **05/04/2024** You can see the list of accepted papers for the 2024 workshop.
- **12/06/2023** The AmericasNLP 2024 workshop will be co-located with NAACL 2024 in Mexico City!
- **02/09/2023** The AmericasNLP 2023 workshop will be co-located with ACL 2023 in Toronto!
- **11/28/2022:** The winners of the

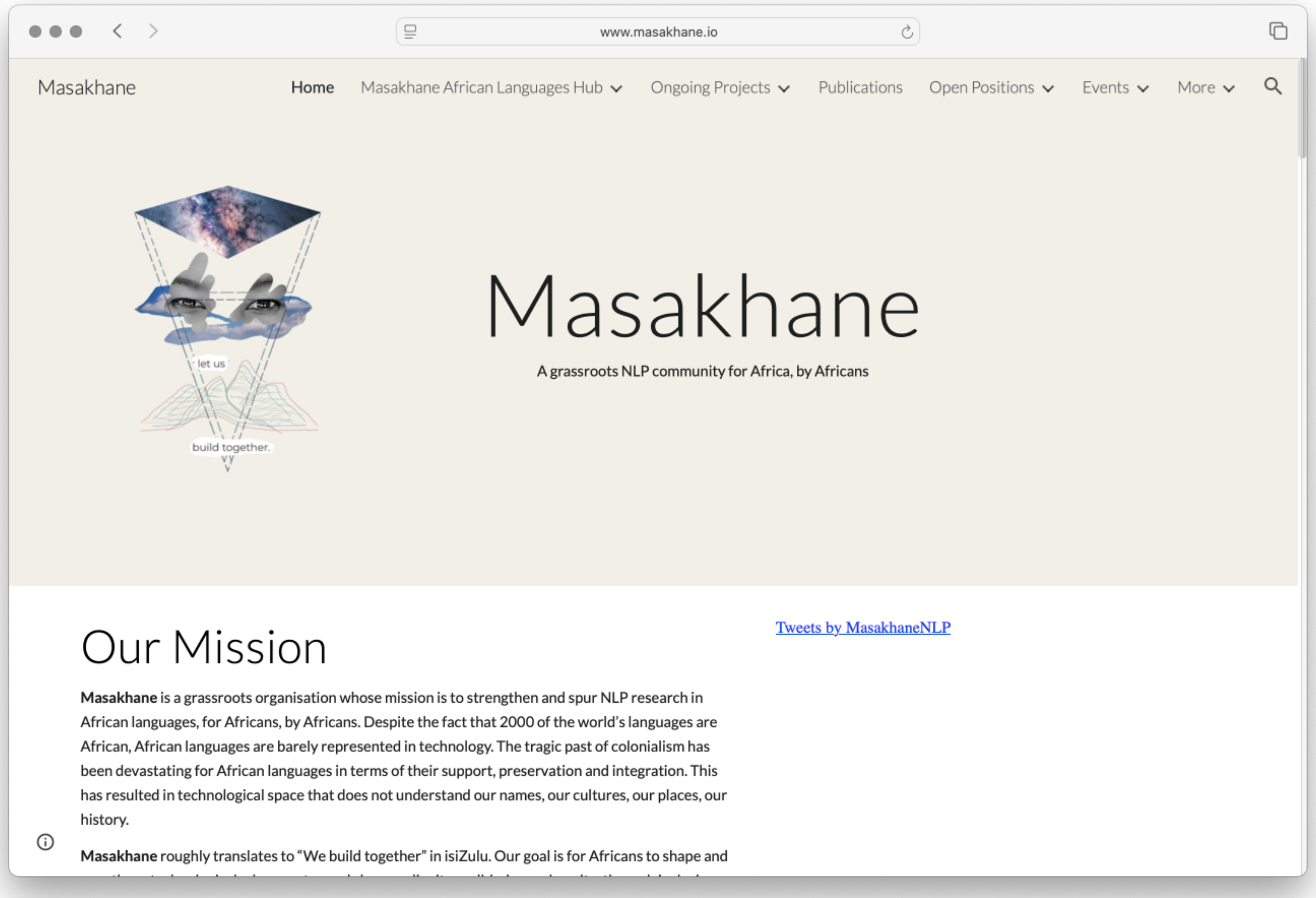
AmericasNLP

AmericasNLP aims to...

- ...encourage research on NLP, computational linguistics, corpus linguistics, and speech around the globe to work on native American languages.
- ...connect researchers and professionals from underrepresented communities and native speakers of endangered languages with the machine learning and natural language processing communities.
- ...promote research on both neural and non-neural machine learning approaches suitable for low-resource languages.

The website of the *First Workshop on NLP for Indigenous Languages of the Americas* can be found here, and more information on the *AmericasNLP 2021*

How can we achieve good performance with sparse data? With small models and limited compute?



# Masakhane

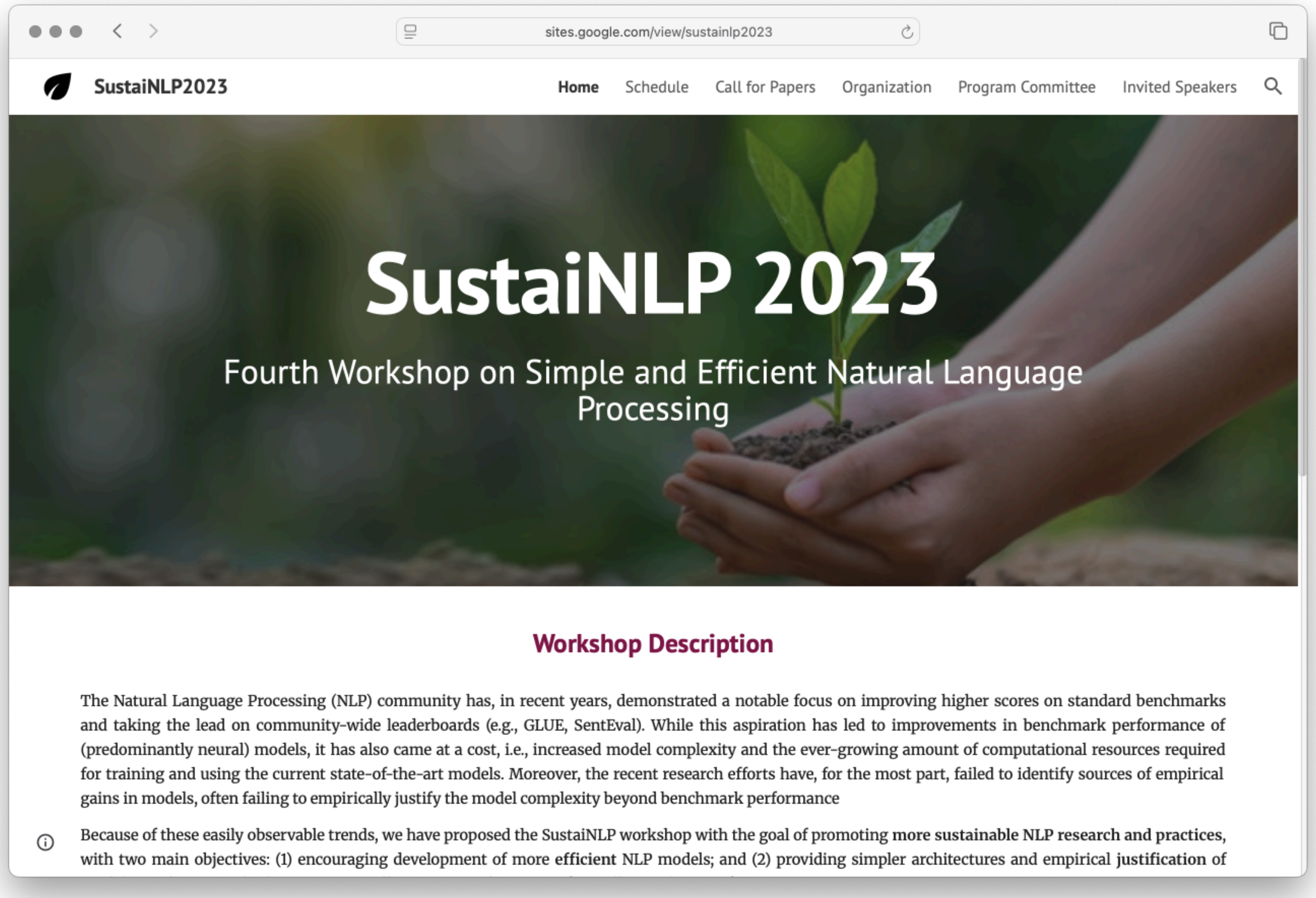
A grassroots NLP community for Africa, by Africans

## Our Mission

Tweets by MasakhaneNLP

**Masakhane** is a grassroots organisation whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans. Despite the fact that 2000 of the world's languages are African, African languages are barely represented in technology. The tragic past of colonialism has been devastating for African languages in terms of their support, preservation and integration. This has resulted in technological space that does not understand our names, our cultures, our places, our history.

**Masakhane** roughly translates to "We build together" in isiZulu. Our goal is for Africans to shape and



# SustaiNLP 2023

Fourth Workshop on Simple and Efficient Natural Language Processing

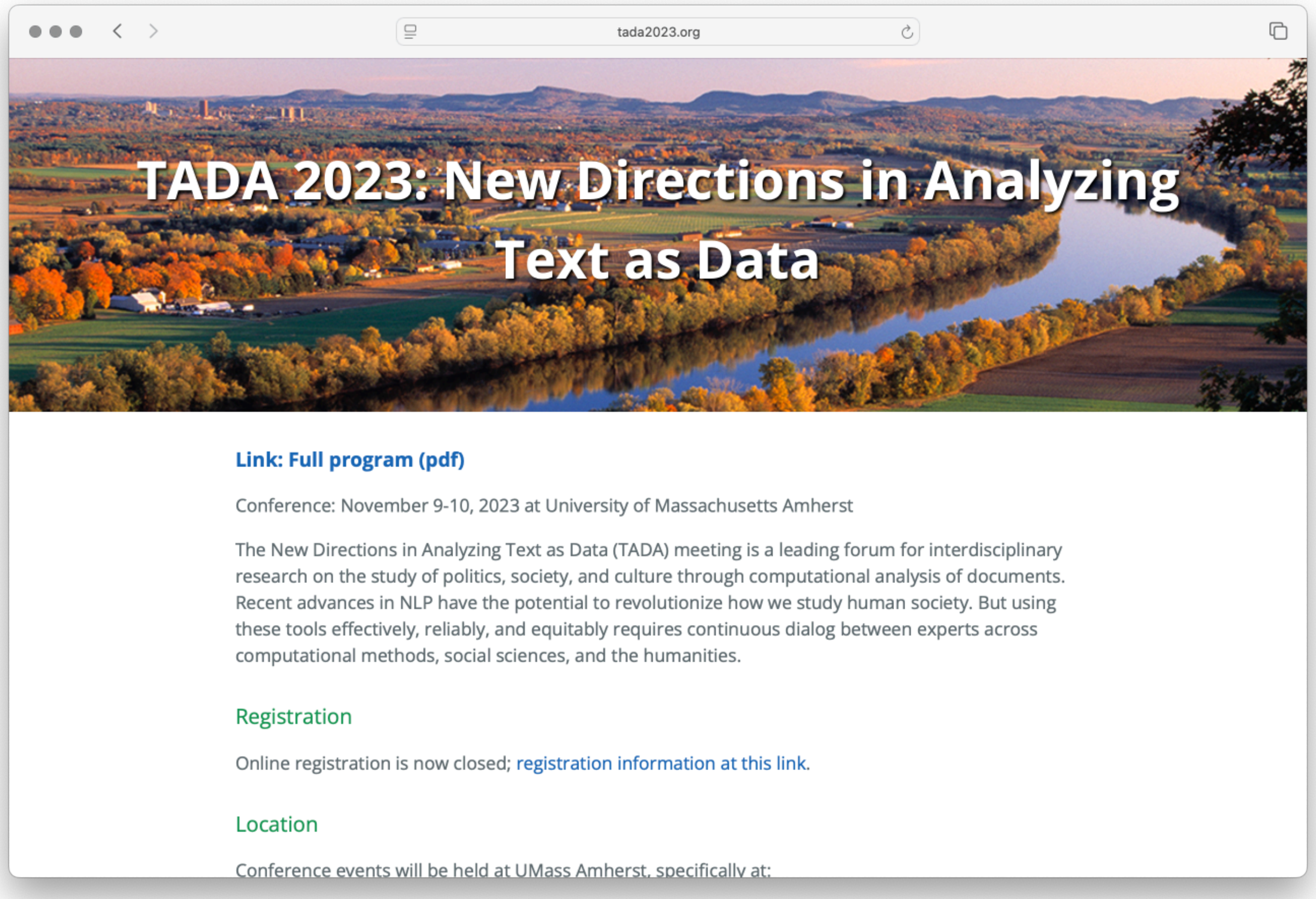
## Workshop Description

The Natural Language Processing (NLP) community has, in recent years, demonstrated a notable focus on improving higher scores on standard benchmarks and taking the lead on community-wide leaderboards (e.g., GLUE, SentEval). While this aspiration has led to improvements in benchmark performance of (predominantly neural) models, it has also came at a cost, i.e., increased model complexity and the ever-growing amount of computational resources required for training and using the current state-of-the-art models. Moreover, the recent research efforts have, for the most part, failed to identify sources of empirical gains in models, often failing to empirically justify the model complexity beyond benchmark performance

Because of these easily observable trends, we have proposed the SustaiNLP workshop with the goal of promoting **more sustainable NLP research and practices**, with two main objectives: (1) encouraging development of more **efficient** NLP models; and (2) providing simpler architectures and empirical **justification** of

# Computational social science

How can we share our best approaches for text processing with other fields?



TADA 2023: New Directions in Analyzing Text as Data

**Link: Full program (pdf)**

Conference: November 9-10, 2023 at University of Massachusetts Amherst

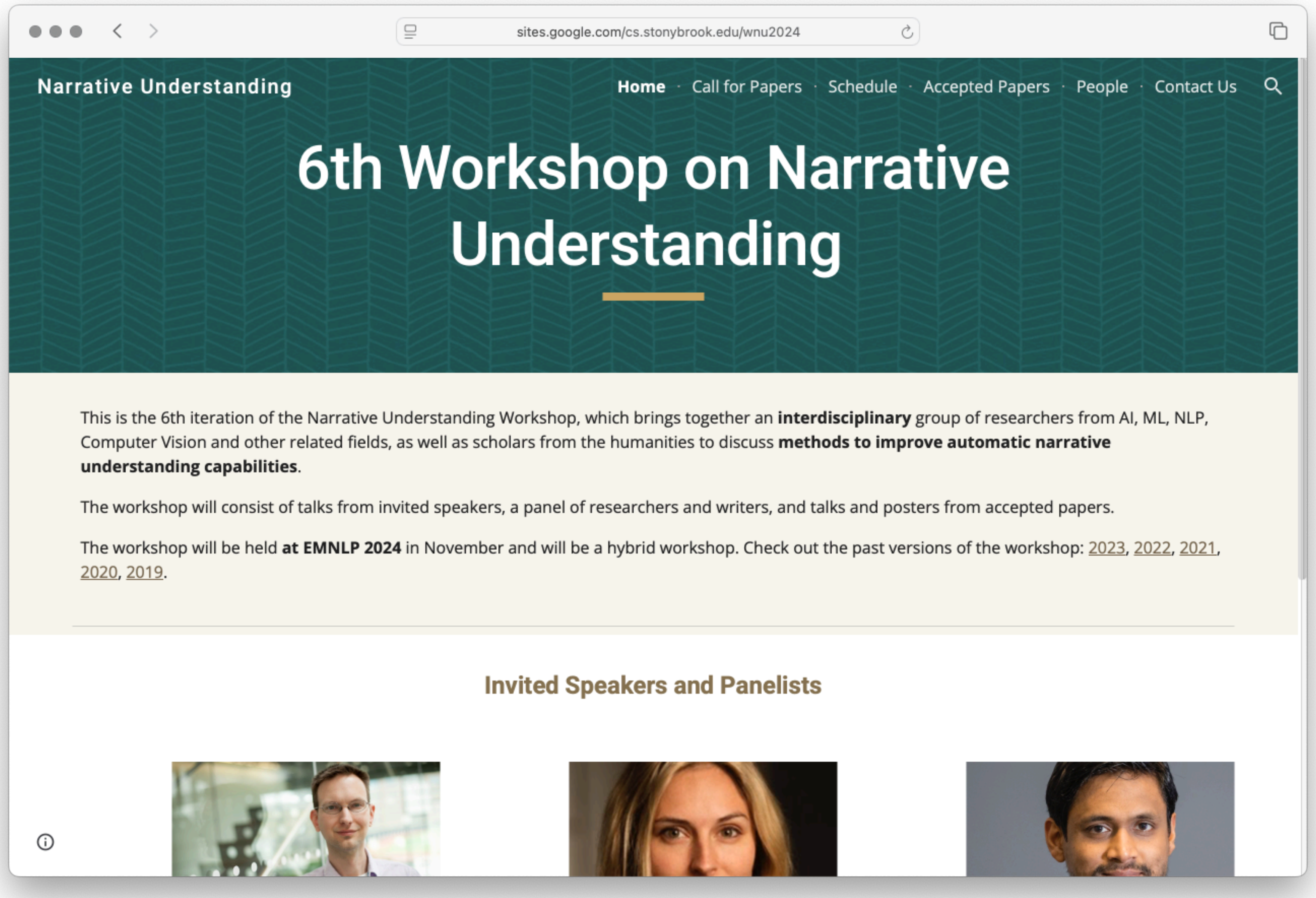
The New Directions in Analyzing Text as Data (TADA) meeting is a leading forum for interdisciplinary research on the study of politics, society, and culture through computational analysis of documents. Recent advances in NLP have the potential to revolutionize how we study human society. But using these tools effectively, reliably, and equitably requires continuous dialog between experts across computational methods, social sciences, and the humanities.

Registration

Online registration is now closed; registration information at this link.

Location

Conference events will be held at UMass Amherst, specifically at:



sites.google.com/cs.stonybrook.edu/wnu2024
Narrative Understanding
Home · Call for Papers · Schedule · Accepted Papers · People · Contact Us

6th Workshop on Narrative Understanding

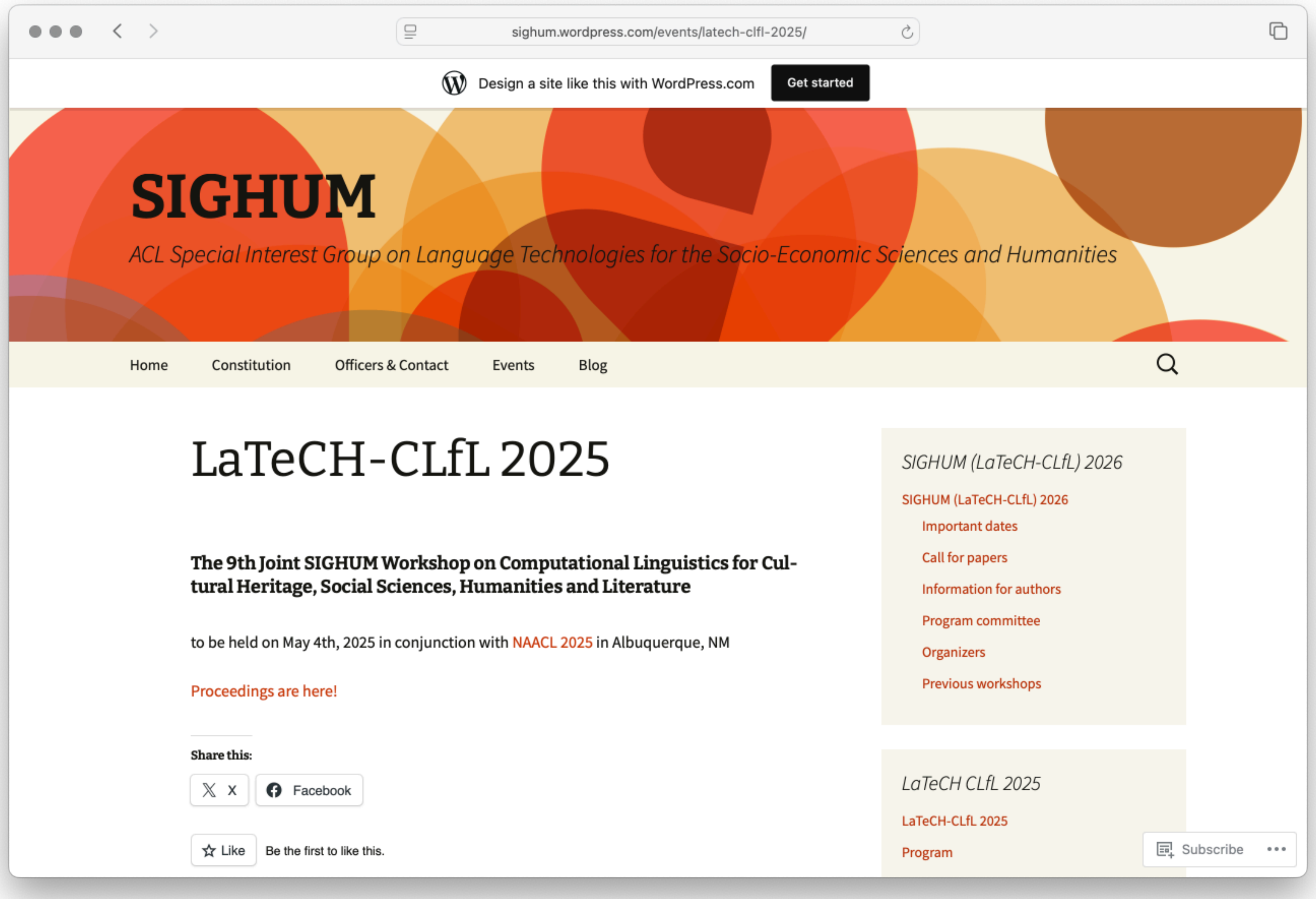
This is the 6th iteration of the Narrative Understanding Workshop, which brings together an **interdisciplinary** group of researchers from AI, ML, NLP, Computer Vision and other related fields, as well as scholars from the humanities to discuss **methods to improve automatic narrative understanding capabilities**.

The workshop will consist of talks from invited speakers, a panel of researchers and writers, and talks and posters from accepted papers.

The workshop will be held **at EMNLP 2024** in November and will be a hybrid workshop. Check out the past versions of the workshop: 2023, 2022, 2021, 2020, 2019.

Invited Speakers and Panelists

How well do NLP techniques work with limited data? In niche domains? In tasks that require expert judgment?


sighum.wordpress.com/events/latech-clfl-2025/
Design a site like this with WordPress.com
Get started
Home
Constitution
Officers & Contact
Events
Blog
Share this:
X
Facebook
Like
Be the first to like this.
Subscribe

SIGHUM

*ACL Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities*

LaTeCH-CLfL 2025

**The 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature**

to be held on May 4th, 2025 in conjunction with NAACL 2025 in Albuquerque, NM

Proceedings are here!

*SIGHUM (LaTeCH-CLfL) 2026*

- SIGHUM (LaTeCH-CLfL) 2026
  - Important dates
  - Call for papers
  - Information for authors
  - Program committee
  - Organizers
  - Previous workshops

*LaTeCH CLfL 2025*

- LaTeCH-CLfL 2025
- Program

go.vassar.edu/course/evals

# Acknowledgments

The class incorporates material from:

Carolyn Anderson, Wellesley College