AACL 2026

The 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing

Tutorial Abstracts

November 6-10, 2026

©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 979-8-89176-443-9

Introduction

Welcome to the tutorial program of AACL-IJCNLP 2026!

This year, we are delighted to accommodate six tutorials that reflect both the rapid evolution and the expanding scope of natural language processing. A prominent theme across the program is the transition from language models as standalone components toward more persistent and interactive systems. Tutorials on agentic deep research and self-improving language agents explore how language models can retrieve information, use tools, maintain memory, verify results, and adapt over extended periods of interaction. Together, they highlight an emerging direction in natural language processing.

At the same time, the program examines some of the foundations and broader implications of this progress. The tutorials address how language model evaluation can be treated as a rigorous measurement problem, how diffusion models offer an alternative paradigm for language generation, and how multilingual language models can better account for code-switching as a natural form of human communication. The program also includes an interactive tutorial on ethical considerations in NLP, emphasizing the importance of reflecting on the societal consequences of our research as NLP technologies become increasingly integrated into real-world applications.

This year, we did a joint call for tutorial proposals with EMNLP 2026 alongside their tutorial co-chairs: Ana Marasović and Goran Glavaš. We received 57 submissions, of which 8 were accepted to AACL (two tutorials were canceled) and 6 to EMNLP. All of the proposals were reviewed by multiple of the tutorial co-chairs before a broader discussion among all. When a tutorial chair has a COI with a tutorial proposal, reviews and decisions are made by other tutorial chairs. As in previous years, it was a highly competitive selection process with many excellent submissions. While the tutorials by no means capture all potentially interesting topics relevant to our research community, we anticipate that they will engage attendees from a wide range of backgrounds and levels of expertise after the main conference closes.

We would like to thank the tutorial instructors for their contributions. Our thanks go to the conference organizers for effective collaboration, and in particular to Ana Marasović and Goran Glavaš (tutorial chairs of EMNLP 2026), Yue Zhang (general chair), Trang Vu and Junxian He (publication chairs), Leyang Cui (handbook chair), Derek F. Wong and Haofen Wang (local organizing chair), and Runzhe Zhan and Zhiwen Xie (local organizing committees).

We hope you enjoy the tutorials.

AACL-IJCNLP 2026 Tutorial Co-chairs

Naoaki Okazaki

Zhongqing Wang

Organizing Committee

General Chair

Yue Zhang, Westlake University, China

Program Chairs

Ruifeng Xu, Harbin Institute of Technology, Shenzhen, China
Ayu Purwarianti, Institut Teknologi Bandung, Indonesia
Tong Rong, Singapore Institute of Technology, Singapore
Jun Suzuki, Tohoku University, Japan

Tutorial Chairs

Naoaki Okazaki, Institute of Science Tokyo, Japan
Zhongqing Wang, Soochow University, China

Program Committee

Program Committee

Naoaki Okazaki, Institute of Science Tokyo, Japan
Zhongqing Wang, Soochow University, China
Goran Glavaš, University of Würzburg, Germany
Ana Marasović, University of Utah, USA

Table of Contents

Code-Switching for Multilingual LLMs
Haneul Yoo, Juhyun Oh, Genta Indra Winata, Ruochen Zhang, Thamar Solorio and Alice Oh . . 1

The Future of NLP Needs Ethical Considerations: An Interactive Tutorial for Researchers and Students
Xi Ai, Vamshi Krishna Bonagiri, Monojit Choudhury, Abdoul Jalil Djiberou Mahamadou, Esther Gan, Zhijing Jin, Min-Yen Kan, Seunghun J. Lee, Yisong Miao, Margot Mieskes, Aurélie Névéol and Yajing Yang 11

A Tutorial on Measurement Science for Language Model Evaluation
Sang T. Truong, Serena Lutong Wang, Rodolfo Corona and Sanmi Koyejo 20

Post-Deployment Self-Improving Language Agents: Learning, Memory, and Adaptation
Estevam Hruschka 27

From Retrieval-Augmented Generation (RAG) to Agentic Deep Research
Xin Luna Dong, Sanat Sharma, Kai Sun, Jiaqi Wang, Yinglong Xia, Xiao Yang, Scott Wen-tau Yih, Nan Tang and Franklin Zhang 32

Diffusion Language Models: Foundations and Frontiers of Non-Autoregressive Language Generation
Daisuke Oba and Danushka Bollegala 44

Code-Switching for Multilingual LLMs

Haneul Yoo[1] **Juhyun Oh**[2] **Genta Indra Winata**[3]
Ruochen Zhang[4] **Thamar Solorio**[5] **Alice Oh**[2]
[1]New York University [2]KAIST [3]Capital One [4]Microsoft [5]MBZUAI
haneul.yoo@nyu.edu 411juhyun@kaist.ac.kr genta.winata@capitalone.com
ruochenzhang@microsoft.com thamar.solorio@mbzuai.ac.ae alice.oh@kaist.edu

Abstract

This tutorial provides a structured overview of code-switching for multilingual large language models (LLMs), connecting linguistic theory, NLP methodology, and human-centered perspectives. Code-switching is a natural communicative practice among multilingual speakers and is increasingly central to LLM interactions, where users bring mixed-language practices to prompts and models themselves may produce mixed-language outputs. The tutorial will cover theoretical foundations from sociolinguistics, code-switching behavior in human-AI interaction, datasets and evaluation benchmarks, synthetic data generation, multilingual LLM training and inference, mechanistic interpretability of unintended language switching, speech and multimodal code-switching, and societal impacts. By synthesizing work across NLP, linguistics, HCI, speech processing, and multilingual communities, the tutorial aims to give participants a coherent view of the field and identify open challenges for building multilingual LLMs that better reflect how multilingual users actually communicate.

1 Introduction

Code-switching, the practice of alternating between two or more languages within a single conversation or utterance, is a natural and widespread phenomenon among multilingual speakers worldwide (Myers-Scotton, 1993; Poplack, 1980). Far from reflecting linguistic deficiency, it demonstrates sophisticated communicative competence, serving pragmatic, social, and cultural functions that monolingual communication cannot always capture. Therefore, it has been extensively studied in (socio)linguistics, with theoretical frameworks developed to explain when, why, and how speakers switch between languages (Kamwangamalu, 1989; Dewaele and Li, 2014; Halpin and Melzi, 2021).

While multilingual large language models (LLMs) have advanced rapidly, realizing their promise for the world's multilingual users requires more than expanding the languages they support. It also requires engaging with how multilingual people actually communicate. Chief among these communicative practices is code-switching, which multilingual users naturally bring to LLM interactions (Oh et al., 2026; Bawa et al., 2020; Choi et al., 2023), and which LLMs also produce in their outputs (Li et al., 2025). As such, code-switching has emerged as a critical phenomenon at the intersection of natural language processing (NLP), linguistics, and human-computer interaction (HCI) (Doğruöz et al., 2021; Winata et al., 2023a; Dedema et al., 2026), spurring growing research from dataset construction and model evaluation, to training strategies, mechanistic interpretability, speech processing, and societal impact.

Despite this rapidly expanding landscape, research on code-switching for multilingual LLMs remains scattered across communities. This tutorial bridges this gap by providing a structured overview of theoretical foundations, empirical findings, open questions, and future directions.

2 Tutorial Outline

This tutorial will focus on multilingual LLMs through the lens of code-switching. For each session, we will present a systematic overview of relevant topics, followed by an in-depth exploration of recent advancements. The tutorial will conclude with an audience-engaged discussion of open challenges and emerging research opportunities.

2.1 Background and Preliminaries [30 min]

Linguistic Theories on Code-switching. Decades of research in sociolinguistics have produced rich theoretical frameworks for understanding code-switching. Structural accounts—ranging from the Matrix Language Frame (Joshi, 1982; Myers-Scotton, 1993) to the Equivalence

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 1–10
November 6-10, 2026 ©2026 Association for Computational Linguistics

Constraint (Poplack, 1980) and Functional Head Constraint (Belazi et al., 1994)—attempt to characterize the syntactic conditions under which switching occurs, while constraint-free theories argue that switching is governed solely by the grammars of the contributing languages (MacSwan, 2005). More recent approaches, such as *translanguaging*, challenge strict language boundaries, viewing multilingual practice as a unified, fluid repertoire rather than alternation between discrete systems (Wei, 2018). Sociolinguists have also identified varied motivations for code-switching—from marking social identity and in-group solidarity to navigating power asymmetries and achieving pragmatic precision (Muysken, 2020; Poplack, 1980; Heller, 1992). We will survey these frameworks and examine how, and to what extent, they have been operationalized in NLP research on multilingual LLMs.

Code-switching in Human-AI Interaction. While code-switching in human-human interaction is well-documented, its manifestation in human-AI interaction remains far less understood (Dedema et al., 2026). Existing studies on LLM usage patterns suggest that multilingual users do switch languages when interacting with AI systems (Oh et al., 2026). Yet, the motivations and dynamics differ in important ways. For example, users interact with a text-based, non-human interlocutor whose language capabilities are often unpredictable, which may shape when and how they choose to code-switch (Choi et al., 2023; Bawa et al., 2020). We will review what is currently known about these patterns and highlight the key open questions around user expectations and behavior in multilingual human-AI interaction.

2.2 Evaluation and Datasets [30 min]

Code-switching Data. We will survey the current landscape of code-switching corpora and evaluation benchmarks, focusing on how dataset scarcity shapes model assessment and data creation practices. We will review broad multilingual NLP benchmarks (Yang and Chai, 2025; Abdaljalil et al., 2025; Mohamed et al., 2026; Farhansyah et al., 2026) and task-specific evaluations for translation (Huzaifah et al., 2024; Zhang et al., 2025b), dialogue summarization (Suresh et al., 2025), red-teaming (Upadhayay and Behzadan, 2024; Song et al., 2025; Yoo et al., 2025b), and output-language confusion (Oh et al., 2026). Despite recent progress, we will emphasize that code-switching remains a challenging setting for multilingual LLM evaluation, not a solved proxy task (Zhang et al., 2023).

Synthetic Data Generation. We will then examine the data bottleneck. Most available corpora are concentrated in a small set of English-involving language pairs, such as Hindi–English (Khanuja et al., 2020; Singh et al., 2018) and Mandarin Chinese–English (Li et al., 2022; Lovenia et al., 2022; Xiang et al., 2020; Li et al., 2012; Lyu et al., 2010), while naturally occurring non-English or multilingual code-switching resources remain scarce (Winata et al., 2023a). This scarcity has made synthetic generation central to scaling evaluation and training resources. We will compare rule- or theory-based methods (Kuwanto et al., 2026; Winata et al., 2026; Yan et al., 2025), toolkit-driven generation for specific language pairs (Jayanthi et al., 2021; Rizvi et al., 2021), prompting-based generation with frontier LLMs (Yong et al., 2023; Kim et al., 2025; Yang and Chai, 2025), and multi-agent frameworks (Xie et al., 2025).

Implications of LLM-based Synthetic Data. Finally, we will discuss the methodological implications of using LLMs to generate code-switching data. While LLM-based generation improves scalability, coverage, and controllability (Winata et al., 2026), it also raises concerns about distributional mismatch with human bilingual speech, contamination between generation and evaluation, and the risk of encoding model-preferred rather than community-authentic switching patterns (Yoo et al., 2025a). We will frame these trade-offs as a central challenge for future code-switching evaluation.

2.3 Multilingual Language Modeling [30 min]

Code-Switching in LLM Training. We will examine how code-switching can be leveraged throughout the full lifecycle of multilingual LLMs, from pre-training to inference. We will first introduce recent work that incorporates code-switching into model training, including pre-training (Wang et al., 2025b) and continual pre-training (Yoo et al., 2025a) that expose models to mixed-language distributions and improve cross-lingual generalization. We will then cover downstream adaptation methods, including supervised fine-tuning (Son et al., 2026; Li et al., 2024), instruction tuning (Lee et al., 2024), and reinforcement learning (Li et al., 2025), highlighting how code-switching signals can improve multilingual capabilities across tasks.

Code-Switching in LLM Inference. Beyond training, we will discuss the role of code-switching in inference time, including in-context learning (Yoo et al., 2026) and reasoning (Wang et al., 2025a), where code-switching can serve as a mechanism to elicit latent multilingual knowledge. We will also introduce recent findings on cultural knowledge activation, where code-switching can help models access culturally grounded knowledge that may not surface in monolingual settings (Kim et al., 2025). Taken together, we will emphasize code-switching as a mechanism that interacts with multilingual representation learning and reasoning.

2.4 Mechanistic Interpretability [30 min]

Code-Switching from Interpretability Lens. We will examine code-switching in multilingual LLMs through the lens of mechanistic interpretability, focusing on recent work that identifies and controls the internal causes of unintended switching. We will discuss recent findings that language confusion emerges in final layers where models transition from shared to language specific representation (Nie et al., 2025), and complementary sparse-autoencoder analyses suggesting that unexpected code-switching can result from overactivated target-language features (Deng et al., 2026). Finally, we will highlight how these mechanistic insights enable targeted interventions, including neuron editing, feature suppression, and partial circuit tuning, which can reduce language confusion while preserving or even improving multilingual performance (Xiao et al., 2026). Taken together, code-switching is governed by identifiable internal mechanisms rather than diffuse model behavior, making it a tractable target for interpretability-driven control.

2.5 Speech and Multimodality [30 min]

Code-switching in Speech Processing. We will briefly review research on code-switching in speech processing, from early bilingual recognition to multilingual speech foundation models and speech-aware LLMs. Early work treated mixed-language speech primarily as a boundary-detection and decoding problem, focusing on language-boundary identification and bilingual recognition (Shia et al., 2004; Lyu et al., 2006). Subsequent progress was enabled by shared Mandarin–English and multilingual code-switching corpora (Lyu et al., 2010; Li et al., 2012, 2022; Lovenia et al., 2022). We will then cover the recent shift toward multilingual LLM-assisted speech models, including language alignment losses (Liu et al., 2025a), speech-conditioned LLMs with mixture-of-experts routing (Zhang et al., 2025a), and parameter-efficient adaptation for intra-sentential switching (Zhao et al., 2025; Yang et al., 2025). We will connect these methods to newer benchmark-style resources, which broaden the evaluation beyond spontaneous bilingual speech to multilingual and domain-specific settings (Ugan et al., 2024; Zhou et al., 2025; Yan et al., 2025; Xie et al., 2025). We will also discuss speech-to-speech interaction models, where recent results show that language alignment remains fragile even for large speech LLMs (Liu et al., 2025b). In addition, we will survey textless or zero-resource approaches for spoken settings where one or more languages lack standardized writing systems (Huang et al., 2024; Chen et al., 2023; Zhang et al., 2025c).

Multimodal Code-Switching. We will expand from speech processing to multimodal settings where speech is jointly modeled with text or vision. For translation, early work established code-switching speech translation as a distinct task and compared cascaded and end-to-end approaches (Weller et al., 2022). More recent systems explicitly interleave speech and text representations, with CoSTA (Shankar et al., 2025) showing that cross-modal alignment helps in low-resource code-switching speech translation. For video, we will discuss audiovisual settings, such as code-switching lip reading, highlighting the need to align language choice with visual cues rather than audio alone (Zhang et al., 2024). We will position code-switching as a genuinely multimodal phenomenon relevant to meeting video, social media clips, assistive communication, and embodied agents.

2.6 Impacts and Future Directions [30 min]

Societal Impacts. We will finally discuss the impacts of code-switching research in LLMs on the global society and the major research questions that remain. We begin with the research in the human-computer interaction field, where early studies show that multilingual users prefer chatbots that code-mix in a reciprocal way (Bawa et al., 2020; Choi et al., 2023). More specific challenges in multilingual LLM interaction have been discussed in Dedema et al. (2026); Oh et al. (2026). Users who belong in linguistic or cultural minorities are more disadvantaged by the prevalence of LLMs, and

code-switching can alleviate this problem, as we can see in domain-specific applications such as English education as a foreign language (Park et al., 2025) and health care (Harrington et al., 2022). Along with these papers, research shows that code-switching can help reduce cultural and linguistic challenges of immigrants (Lee et al., 2025).

Future Directions. We will turn this discussion into an interactive program that engages the audience as multilingual users, researchers, and practitioners. Participants will be invited to reflect on their own experiences with code-switching in LLM interactions, identify examples from their linguistic communities, discuss why such switching occurs, and surface edge cases that remain difficult for current models. Building on these observations, the tutorial will conclude by synthesizing open challenges and future research directions in code-switching for multilingual LLMs, emphasizing collaboration among NLP, linguistics, HCI, speech processing, and multilingual communities.

3 Tutorial Information

Type & Length. 3-hour, cutting-edge tutorial.

Target Audience. The tutorial will be accessible to anyone with basic knowledge of machine learning (ML) and natural language processing (NLP), and will be of interest to researchers, students, and practitioners in both academia and industry.

Prerequisites. Following knowledge is assumed:
- ML: supervised learning, transformer models
- NLP: language modeling, data and evaluation

Past Tutorials. Related tutorials include:
- EMNLP 2019 tutorial on *Processing and Understanding Mixed Language Data* (Choudhury et al., 2019) offered a comprehensive overview of the linguistic and computational aspects of code-mixing, at a time when the field lacked a coherent entry point. However, because it was developed before the rise of multilingual LLMs, it does not address recent advances and open questions surrounding LLM-based evaluation, generation, and use of synthetic code-switching data at scale.
- ACL 2023 tutorial on *Everything you need to know about Multilingual LLMs: Towards fair, performant and reliable models for languages of the world* (Sitaram et al., 2023) reviewed data, training, evaluation, responsible AI, and deployment for multilingual LLMs at a broad level, with limited space for an in-depth treatment of code-switching. Our tutorial complements these prior gaps by revisiting code-switching as a central key for multilingual LLMs.

4 Reading List

- A survey of Code-Switching: Linguistic and Social Perspectives for Language Technologies (Doğruöz et al., 2021)
- The decades progress on code-switching research in NLP: A systematic survey on trends and challenges (Winata et al., 2023b)
- Multilingual Large Language Models Are Not (Yet) Code-Switchers (Zhang et al., 2023)
- SwitchLingua: The First Large-Scale Multilingual and Multi-Ethnic Code-Switching Dataset (Xie et al., 2025)
- Investigating and Scaling up Code-Switching for Multilingual Language Model Pre-Training (Wang et al., 2025b)
- Language Mixing in Reasoning Language Models: Patterns, Impact, and Internal Causes (Wang et al., 2025a)
- Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models (Nie et al., 2025)
- CoSTA: Code-Switched Speech Translation using Aligned Speech-Text Interleaving (Shankar et al., 2025)
- Toward a Multilingual Conversational Agent: Challenges and Expectations of Code-mixing Multilingual Users (Choi et al., 2023)

5 Ethics Statement

This tutorial will present current research on the use of code-switching for training, evaluation, interpretation, and societal impact of multilingual LLMs in the field. Although we aim to promote linguistic diversity by discussing issues related to multilingual models trained in 100+ languages, many languages worldwide are yet to be supported by these models.

6 Tutorial Instructors

Haneul Yoo is an Assistant Professor/Faculty Fellow in the Department of Computer Science at the NYU Courant Institute School of Mathematics, Computing, and Data Science. She received her PhD in the School of Computing at KAIST. With a background in machine learning (ML) and natural language processing (NLP), she has worked on developing language models and methods to improve

generalization across languages and designing resources and evaluations for underrepresented languages and cultures. She has published papers in major NLP conferences, including ACL, EMNLP, and NAACL, and has been named a Rising Star at KAIST exploreCSR supported by Google. She has served as a Social Chair at NeurIPS 2026 and co-organized the Multilingual and Equitable Language Technologies (MELT) workshop at COLM 2025, the Women in Machine Learning (WiML) symposium at ICML 2026, and the Generative AI for the World (GenAI4World) workshop at COLM 2026. She delivered invited talks on code-switching and multilingual/multicultural LLMs at Princeton, UMD, Mila, and NYU, and served as a instructor for Natural Language Processing at NYU.

Juhyun Oh is a PhD student in the School of Computing at KAIST and is currently a visiting scholar at Mila. She previously studied Linguistics at Seoul National University, where she received both her BA and MA. Her research focuses on developing evaluation methodologies for language models that better reflect real-world user experiences, particularly across diverse linguistic and cultural contexts. Her work on improving support for underrepresented languages in NLP has been published at top NLP and ML conferences. She has served as a Guest Lecturer for the LLM Seminar at Georgia Tech, delivering a talk titled "Rethinking Evaluation for Language Models," and as a Teaching Assistant for the AI Tech Boostcamp at NAVER Connect Foundation. She has co-organized the Generative AI for the World (GenAI4World) workshop at COLM 2026.

Genta Indra Winata is a Senior Research Manager in LLM Core and Agentic AI team at Capital One. His research interests include language modeling, multilingual NLP, and conversational agents. He is actively involved in large-scale dataset collection and the development of multimodal and multilingual reward models and language models, with a particular focus on advancing research in low-resource languages. He has served as organizer of Sixth Workshop on Computational Approaches to Linguistic Code-Switching (CALCS) at EMNLP 2023, the Multilingual Representation Learning (MRL) at EMNLP 2023 and EMNLP 2026, and the South-East Asian Language Processing (SEALP) workshops collocated with AACL 2024 and COLING 2025, and Financial Technology and Natural Language Processing (FinNLP) at EMNLP 2025. He also delivered a tutorial on the Current Status of NLP in South East Asia with Insights from Multilingualism and Language Diversity at AACL 2023. He was a keynote speaker at SPELLL 2023, the Fields Matter Workshop at ACL 2024, and the Workshop of Language Models for Underserved Communities (LM4UC) at NAACL 2025.

Ruochen Zhang is a Senior Reseacher at Microsoft. She received her PhD from Brown University. Her research focuses on understanding the mechanisms underlying model weaknesses and brittleness, particularly through the lens of multilinguality. She has published at major NLP and ML venues, including EMNLP, ICLR, and NeurIPS. She has served as an Area Chair at LREC-COLING 2024 and as a panelist at the 7th Workshop on Computational Approaches to Linguistic Code-Switching (CALCS). She also co-organized the tutorial "Current Status of NLP in Southeast Asia: Insights from Multilingualism and Language Diversity" at AACL 2023. She is one of the initiators of the grassroots community for Southeast Asian researchers, SEACrowd, and has consistently co-organized Birds of a Feather events at major *CL conferences.

Thamar Solorio is Professor of NLP at MBZUAI, where she also serves as Vice Provost for Faculty Excellence and Advancement. She has been working on code-switching for over two decades, with seminal papers dating back to the pre-LLM era on topics such as prediction of code-switching points and enabling technology for the processing of code-switching data. She has served as General Chair of EMNLP 2024, Program Chair for NAACL 2019, and has served in numerous other organization roles including D&I co-chair, demonstration co-chair, and faculty advisor to the student research workshop for different *CL conferences. She has also led, and collaborated in the organization of multiple shared tasks on different topics, including code-switching. Thamar has delivered keynote talks at *CL and/or associated events such as NAACL 2021, CoNLL 2024, and WNUT on topics related to code-switching, and artificial social intelligence. Through her career she has prepared and delivered multiple courses at the undergraduate and graduate level.

Alice Oh is a KAIST ICT Endowed Professor in the School of Computing at KAIST. She received her PhD in Computer Science from MIT in 2008.

Her major research area is at the intersection of natural language processing (NLP) and computational social science (CSS), with a recent focus on multilingual and multicultural aspects of LLMs. She has served as Tutorial Chair for NeurIPS 2019, and DEI Chair for ICLR 2019 and COLM 2024, Program Chair for ICLR 2021 and NeurIPS 2022, General Chair for ACM FAccT 2022, NeurIPS 2023, and Festival of Learning 2026 which co-locates AIED, EDM, and ACM L@S. She presented on the topics of code-switching and multilingual/multicultural LLMs as an invited speaker at Stanford, Cornell Tech, CMU, and UCLA in 2025. She also gave invited keynotes at various international conferences including ACML, ICWSM, and RSS.

References

Samir Abdaljalil, Erchin Serpedin, Khalid Qaraqe, and Hasan Kurban. 2025. Evaluating multilingual and code-switched alignment in LLMs via synthetic natural language inference. *arXiv preprint arXiv:2508.14735*.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Hedi M. Belazi, Edward J. Rubin, and Almeida Jacqueline Toribio. 1994. Code switching and X-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2):221–237.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. 2023. Speech-to-speech translation for a real-world unwritten language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4969–4983, Toronto, Canada. Association for Computational Linguistics.

Yunjae J. Choi, Minha Lee, and Sangsu Lee. 2023. Toward a multilingual conversational agent: Challenges and expectations of code-mixing multilingual users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Monojit Choudhury, Anirudh Srinivasan, and Sandipan Dandapat. 2019. Processing and understanding mixed language data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Meredith Dedema, Chiao Xuen Goh, and Pengyi Zhang. 2026. When generative AI mixes languages: Multilingual users' code-switching behavior in human-LLM interaction. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, New York, NY, USA. Association for Computing Machinery.

Boyi Deng, Yu Wan, Baosong Yang, Fei Huang, Wenjie Wang, and Fuli Feng. 2026. SASFT: Sparse autoencoder-guided supervised finetuning to mitigate unexpected code-switching in LLMs. In *The Fourteenth International Conference on Learning Representations*.

Jean-Marc Dewaele and Wei Li. 2014. Intra- and inter-individual variation in self-reported code-switching patterns of adult multilinguals. *International Journal of Multilingualism*, 11(2):225–246.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online. Association for Computational Linguistics.

Mohammad Rifqi Farhansyah, Hanif Muhammad Zhafran, Farid Adilazuarda, Shamsuddeen Hassan Muhammad, Maryam Ibrahim Mukhtar, Nedjma Ousidhoum, Genta Indra Winata, Ayu Purwarianti, and Alham Fikri Aji. 2026. PingPong: A natural benchmark for multi-turn code-switching dialogues. *arXiv preprint arXiv:2601.17277*.

Emily Halpin and Gigliana Melzi. 2021. Code-switching in the narratives of dual-language latino preschoolers. *International Journal of Bilingual Education and Bilingualism*, 24(9):1271–1287.

Christina N. Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "it's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Monica Heller. 1992. The politics of codeswitching and language choice. *Journal of Multilingual and Multicultural Development*, 13(1-2):123–142.

Kuan-Po Huang, Chih-Kai Yang, Yu-Kuan Fu, Ewan Dunbar, and Hung-Yi Lee. 2024. Zero resource code-switched speech benchmark using speech utterance pairs for multiple spoken languages. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10006–10010.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia. ELRA and ICCL.

Sai Muralidhar Jayanthi, Kavya Nerella, Khyathi Raghavi Chandu, and Alan W Black. 2021. CodemixedNLP: An extensible and open NLP toolkit for code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 113–118, Online. Association for Computational Linguistics.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Nkonko Mudipanu Kamwangamalu. 1989. Code-mixing and modernization. *World Englishes*, 8(3):321–332.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the 4th Workshop on Computational Approaches to Code Switching*, pages 9–16, Marseille, France. European Language Resources Association.

Seoyeon Kim, Huiseo Kim, Chanjun Park, Jinyoung Yeo, and Dongha Lee. 2025. Can code-switched texts activate a knowledge switch in LLMs? a case study on English-Korean code-switching. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22303–22327, Suzhou, China. Association for Computational Linguistics.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2026. Linguistics theory meets LLM: Code-switched text generation via equivalence constrained large language models. In *Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL)*, pages 1–14, Grand Hyatt Manchester San Diego, 1 Market Pl, San Diego, CA 92101. Association for Computational Linguistics.

Jaeseong Lee, YeonJoon Jung, and Seung-won Hwang. 2024. COMMIT: Code-mixing English-centric large language model for multilingual instruction tuning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3130–3137, Mexico City, Mexico. Association for Computational Linguistics.

Sunok Lee, Dasom Choi, Lucy Truong, Nitin Sawhney, and Henna Paakki. 2025. Into the unknown: Leveraging conversational AI in supporting young migrants' journeys towards cultural adaptation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Chengfei Li, Shuhao Deng, Yaoping Wang, Guangjing Wang, Yaguang Gong, Changbin Chen, and Jinfeng Bai. 2022. TALCS: An open-source Mandarin-English code-switching corpus and a speech recognition baseline. In *Proc. Interspeech 2022*, pages 1741–1745.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. 2025. The impact of language mixing on bilingual LLM reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32531–32548, Suzhou, China. Association for Computational Linguistics.

Ying Li, Yue Yu, and Pascale Fung. 2012. A Mandarin-English code-switching corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2515–2519, Istanbul, Turkey. European Language Resources Association (ELRA).

Zhuoran Li, Chunming Hu, Junfan Chen, Zhijun Chen, Xiaohui Guo, and Richong Zhang. 2024. Improving zero-shot cross-lingual transfer via progressive code-switching. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6388–6396. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Hexin Liu, Xiangyu Zhang, Haoyang Zhang, Leibny Paola Garcia-Perera, Andy W. H. Khong, Eng Siong Chng, and Shinji Watanabe. 2025a. Aligning speech to languages to enhance code-switching speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4712–4725.

Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. 2025b. Cs3-bench: Evaluating and enhancing speech-to-speech LLMs for Mandarin–English code-switching. *arXiv preprint arXiv:2510.07881*.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Yan Xu, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram E. Shi, and Pascale Fung. 2022. ASCEND: A spontaneous Chinese-English dataset for code-switching in multi-turn conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7259–7268, Marseille, France. European Language Resources Association.

Dau-cheng Lyu, Ren-yuan Lyu, Yuang-chin Chiang, and Chun-nan Hsu. 2006. Speech recognition on code-switching among the chinese dialects. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010. SEAME: a Mandarin-English code-switching speech corpus in South-east Asia. In *Proc. Interspeech 2010*, pages 1986–1989.

Jeff MacSwan. 2005. Codeswitching and generative grammar: A critique of the MLF model and some remarks on "modified minimalism". *Bilingualism: language and cognition*, 8(1):1–22.

Amr Mohamed, Yang Zhang, Michalis Vazirgiannis, and Guokan Shang. 2026. Lost in the mix: Evaluating LLM understanding of code-switched text. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44922–44938, San Diego, California, United States. Association for Computational Linguistics.

Pieter Muysken. 2020. Code-switching and grammatical theory. In *The bilingualism reader*, pages 280–297. Routledge.

Carol Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press.

Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2025. Mechanistic understanding and mitigation of language confusion in English-centric large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 690–706, Suzhou, China. Association for Computational Linguistics.

Juhyun Oh, Haneul Yoo, Faiz Ghifari Haznitrama, and Alice Oh. 2026. OLA: Output language alignment in code-switched LLM interactions. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46603–46624, San Diego, California, United States. Association for Computational Linguistics.

Junyeong Park, Jieun Han, Yeon Su Park, Youngbin Lee, Suin Kim, Juho Kim, Alice Oh, and So-Yeon Ahn. 2025. Understanding EFL learners' code-switching and teachers' pedagogical approaches in LLM-supported speaking practice. *arXiv preprint arXiv:2512.23136*.

Shana Poplack. 1980. Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Online. Association for Computational Linguistics.

Bhavani Shankar, Preethi Jyothi, and Pushpak Bhattacharyya. 2025. CoSTA: Code-switched speech translation using aligned speech-text interleaving. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9194–9208, Abu Dhabi, UAE. Association for Computational Linguistics.

Chi-Jiun Shia, Yu-Hsien Chiu, Jia-Hsin Hsieh, and Chung-Hsien Wu. 2004. Language boundary detection and identification of mixed-language speech based on map estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–381.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. A Twitter corpus for Hindi-English code mixed POS tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 12–17, Melbourne, Australia. Association for Computational Linguistics.

Sunayana Sitaram, Monojit Choudhury, Barun Patra, Vishrav Chaudhary, Kabir Ahuja, and Kalika Bali. 2023. Everything you need to know about multilingual LLMs: Towards fair, performant and reliable models for languages of the world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 21–26, Toronto, Canada. Association for Computational Linguistics.

Guijin Son, Donghun Yang, Hitesh Laxmichand Patel, Amit Agarwal, Hyunwoo Ko, Chanuk Lim, Srikant Panda, Minhyuk Kim, Nikunj Drolia, Dasol Choi, Kyong-Ha Lee, and Youngjae Yu. 2026. Pushing on multilingual reasoning models with language-mixed chain-of-thought. In *The Fourteenth International Conference on Learning Representations*.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2025. Multilingual blending: Large language model safety alignment evaluation with language mixture. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3433–3449, Albuquerque, New Mexico. Association for Computational Linguistics.

Sathya Krishnan Suresh, Tanmay Surana, Lim Zhi Hao, and Eng Siong Chng. 2025. CS-sum: A benchmark for code-switching dialogue summarization and the limits of large language models. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pages 31–47, Hybrid. Association for Computational Linguistics.

Enes Yavuz Ugan, Ngoc-Quan Pham, and Alexander Waibel. 2024. DECM: Evaluating bilingual ASR performance on a code-switching/mixing benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4468–4475, Torino, Italia. ELRA and ICCL.

Bibek Upadhayay and Vahid Behzadan. 2024. Sandwich attack: Multi-language mixture adaptive attack on LLMs. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 208–226, Mexico City, Mexico. Association for Computational Linguistics.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schuetze. 2025a. Lan-

guage mixing in reasoning language models: Patterns, impact, and internal causes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2637–2665, Suzhou, China. Association for Computational Linguistics.

Zhijun Wang, Jiahuan Li, Hao Zhou, Rongxiang Weng, Jingang Wang, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025b. Investigating and scaling up code-switching for multilingual language model pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11032–11046, Vienna, Austria. Association for Computational Linguistics.

Li Wei. 2018. Translanguaging as a practical theory of language. *Applied Linguistics*, 39(1):9–30.

Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. End-to-end speech translation for code switched speech. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1435–1448, Dublin, Ireland. Association for Computational Linguistics.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023a. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2023b. The decades progress on code-switching research in nlp: A systematic survey on trends and challenges. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978.

Genta Indra Winata, David Anugraha, Patrick Amadeus Irawan, Anirban Das, Haneul Yoo, Paresh Dashore, Shreyas Kulkarni, Ruochen Zhang, Haruki Sakajo, Frederikus Hudi, Anaelia Ovalle, Syrielle Montariol, Felix Gaschi, Michael Anugraha, Rutuj Ravindra Puranik, Zawad Hayat Ahmed, Adril Putra Merin, and Emmanuele Chersoni. 2026. Can large language models understand, reason about, and generate code-switched text? *arXiv preprint arXiv:2601.07153*.

Rong Xiang, Mingyu Wan, Qi Su, Chu-Ren Huang, and Qin Lu. 2020. Sina Mandarin alphabetical words:a web-driven code-mixing lexical resource. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 833–842, Suzhou, China. Association for Computational Linguistics.

Yuxin Xiao, Zhen Huang, Wenxiao Wang, Binbin Lin, Xiaofei He, Xu Shen, and Jieping Ye. 2026. How do language models speak languages? a case study on unintended code-switching.

Peng Xie, Xingyuan Liu, Yequan Bie, Tsz Wai Chan, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. 2025. SwitchLingua: The first large-scale multilingual and multi-ethnic code-switching dataset. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Sai Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, Dominik Klement, Sara Althubaiti, Puyuan Peng, Matthew Wiesner, Thamar Solorio, Ahmed Ali, Sanjeev Khudanpur, and Shinji Watanabe. 2025. CS-FLEURS: A massively multilingual and code-switched speech dataset. In *Interspeech 2025*, pages 743–747.

Hongli Yang, Yizhou Peng, Hao Huang, and Sheng Li. 2025. Adapting Whisper for parameter-efficient code-switching speech recognition via soft prompt tuning. In *Interspeech 2025*, pages 5203–5207.

Yilun Yang and Yekun Chai. 2025. CodeMixBench: Evaluating code-mixing capabilities of LLMs across 18 languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2139–2169, Suzhou, China. Association for Computational Linguistics.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Haneul Yoo, Jiho Jin, Kyunghyun Cho, and Alice Oh. 2026. Gradual code-switching as inference-time cross-lingual representational alignment for LLMs. In *Third Conference on Language Modeling*.

Haneul Yoo, Cheonbok Park, Sangdoo Yun, Alice Oh, and Hwaran Lee. 2025a. Code-switching curriculum learning for multilingual transfer in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7816–7836, Vienna, Austria. Association for Computational Linguistics.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2025b. Code-switching red-teaming: LLM evaluation for safety and multilingual understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13392–13413, Vienna, Austria. Association for Computational Linguistics.

Fengrun Zhang, Wang Geng, Hukai Huang, Yahui Shan, Cheng Yi, and He Qu. 2025a. Boosting code-switching ASR with mixture of experts enhanced

speech-conditioned LLM. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Wenbo Zhang, Aditya Majumdar, and Amulya Yadav. 2025b. CHAI for LLMs: Improving code-mixed translation in large language models through reinforcement learning with AI feedback. *arXiv preprint arXiv:2411.09073*.

Xueyi Zhang, Chengwei Zhang, Mingrui Lao, Peng Zhao, Jun Tang, Yanming Guo, Siqi Cai, Xianghu Yue, and Haizhou Li. 2024. Language without borders: A dataset and benchmark for code-switching lip reading. In *Advances in Neural Information Processing Systems*, volume 37, pages 30727–30739. Curran Associates, Inc.

Yuhao Zhang, Xiangnan Ma, Kaiqi Kou, Peizhuo Liu, Weiqiao Shan, Benyou Wang, Tong Xiao, Yuxin Huang, Zhengtao Yu, and JingBo Zhu. 2025c. Leveraging unit language guidance to advance speech modeling in textless speech-to-speech translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1448–1460, Vienna, Austria. Association for Computational Linguistics.

Jiahui Zhao, Hao Shi, Chenrui Cui, Tianrui Wang, Hexin Liu, Zhaoheng Ni, Lingxuan Ye, and Longbiao Wang. 2025. Adapting Whisper for code-switching through encoding refining and language-aware decoding. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jiaming Zhou, Yujie Guo, Shiwan Zhao, Haoqin Sun, Hui Wang, Jiabei He, Aobo Kong, Shiyao Wang, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. 2025. CS-Dialogue: A 104-hour dataset of spontaneous Mandarin–English code-switching dialogues for speech recognition. *arXiv preprint arXiv:2502.18913*.

The Future of NLP Needs Ethical Considerations: An Interactive Tutorial for Researchers and Students

Xi Ai and **Vamshi Bonagiri** and **Monojit Choudhury** and **Abdoul J. Djiberou Mahamadou**

Esther Gan and **Zhijing Jin** and **Min-Yen Kan** and **Seunghun J. Lee**

Yisong Miao and **Margot Mieskes** and **Aurélie Névéol** and **Yajing Yang**

acl-ethics-committee@inria.fr

Abstract

With NLP research being rapidly productionized into real-world applications, it is important to be aware of and think through the consequences of our work. Such ethical considerations are important in both senior and junior roles of researchers and students (e.g. privacy, consent, fairness, among others).

This tutorial will equip participants with basic guidelines for thinking deeply about ethical issues and review common considerations that recur in NLP research. The methodology is interactive and participatory, including discussion and group work related to case studies that involve bias and broader issues (e.g., environmental impact). Participants will gain practical experience on when to flag a paper for ethics review and how to write an ethical consideration section to be shared with the broader community. Most importantly, the participants will be co-creating the tutorial outcomes and extending tutorial materials to share as public outcomes.

1 Motivation and Structure

In late 2021, the Association for Computational Linguistics' executive committee appointed an ACL Ethics Committee (hereafter, AEC) to oversee long-term policy on ethical issues, and to charge it with training its constituent membership on ethical issues. The AEC carried out a 2023 survey of the constituency, finding that many respondents felt that clear guidelines on acceptable practices regarding authoring and reviewing were needed (Benotti et al., 2023a): half highlighted the need for more resources and forums to raise community awareness on ethical issues in research and to clarify ethical review policies; 36% mentioned the creation of dedicated training materials, and 26% mentioned outreach to discuss ethical research.

This tutorial proposal follows from the 2023 survey mandates, to provide interactive opportunities to communicate and train our membership on ethical guidelines and research practices. **Our proposal is also endorsed by the previous and current ACL presidents.** Both have described ethics as a cornerstone of the synchronous development of CL/NLP: "*we need to place ethics in the curriculum for all NLP students — not as an elective, but as a core part of their education*".

This proposal also draws on successful 2023 and 2025 past tutorials on NLP reviewing and socially responsible NLP (Cohen et al., 2021; Tsvetkov et al., 2018; Benotti et al., 2023b, 2025) that have attracted 50–80 attendees. Past members of our team have been involved in these iterations.

We propose a **hybrid tutorial** to best allow equitable access to the topic of this tutorial, especially to familiarize new community members and those who cannot afford access to attend physically. We plan to have dedicated presenters that can coordinate activities for the expected online participants and plan to use e-resources that facilitate virtual group discussion (e.g., Google Docs).

Our tutorial aligns with the EMNLP 2026 theme of **New Missions for NLP Research**, in calling for a re-think of the integral incorporation of ethics in NLP evaluation. Our tutorials will also help to fulfill the strong mandate given by our constituency.

2 Tutorial Content

Type: 1/2 day, Introductory
Expected Attendees: 50–100
Audience: Researchers, students, and other interested parties
Desired Location: Preferably in **both** locations.
Prerequisites: Introductory background in natural language processing and deep learning. Familiarity with text mining and generation and standard NLP tasks. Fluent command of English.

Note about location: We have tutorial organizers ready to be be physically and virtually present for *both* AACL-IJCNLP (Hangqin, China) and

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 11–19
November 6-10, 2026 ©2026 Association for Computational Linguistics

Segment Topic	Led by
1. Introduction and Foundations for Ethics	Presenters
2. Case Studies: Research that requires ethics review — First reading	Participants
3. Structured Interaction / Dialogue	Presenters, Participants
4. Case studies — Second reading (Rotation)	Participants
5. Group Presentations	Group Leads
6. Summary and Common Issues	Presenters
7. Discussing and Troubleshooting Ethics and Further Resources	Presenters

Table 1: Tutorial Outline. Each segments' duration is 30 minutes, a total of 3 hours (3 1/2 w/ break). Segments 2–6 will be conducted in small-group interaction.

EMNLP (Budapest, Hungary). We strongly hope that the tutorial chairs can consider allowing this tutorial to be delivered in both locations as:

· This tutorial has not been fielded at EMNLP, which reaches a different group of researchers;
· This tutorial has not been fielded in Asia, which is burgeoning seat of future NLP research.

Ethical considerations overarch our duties as researchers and scientists, and we need to consider the ramifications of our work. The need for a better understanding of ethics is reflected in both researcher and student roles.

Unintended and harmful ethical lapses and consequences can be largely avoided through continuous communication. Rather than assume that research is purely an intellectual pursuit, our tutorial invites participants to consider ethics as an integral component of the holistic framework of impactful research work. Table 1 presents our proposed tutorial's outline. Our aim is to provide hands-on experience with ethical issues through a small-group activity, both at the physical conference and in breakout rooms for online participants.

Ethics requires healthy debate and deep thought, and for these reasons, our structure incorporates a Socratic exercise, where participants spend a large part of the session discussing a concrete case of research which requires deep ethics review. A Community of Inquiryapproach will be taken such that participants engage in role-playing and discussing ethical issues through reading 1–2 problematic hypothetical research abstracts from a curated set (§ 2.2). Using Socratic-style questioning, presenters guide the participants to engender discussion and realize ethical issues in the works.

To encapsulate the exercise, the presenters will first introduce the key ways that ethics impacts authoring and reviewing (Segment 1), summarise the group discussions' key points (Segment 6) and conclude with pointers to references and other training materials (Segment 7). Importantly, the participants will be co-building the tutorial outcomes and will be working to create further tutorial materials to share as public outcomes of the exercises. For many issues in ethics, the evolving discussion creates more value than the actual conclusions. This is why we propose such a dialectic approach.

Due to the necessary interactivity of the session, we plan to limit the registrations for the tutorial to 100. This is to cater to having approximately a 25:1 ratio for presenters to participants. A larger volume than this jeopardizes the necessary interactive nature of the tutorial, which requires input from all participants.

2.1 Topics (Segments 1 & 6)

Depending on the location(s) where the tutorial will be executed, we will run versions of the tutorial that best fit the expertise of the participating instructors. We detail some possible topics below.

Multilingual and Multicultural Bias. *Sample Presenters: Xi Ai, Vamshi Bonagiri, Abdoul J. Djiberou Mahamadou, Monojit Chowdhury, Zhijing Jin and Seunghun J. Lee.*

This section of the tutorial will take a multilingual and multicultural perspective on bias and equity in NLP, with particular attention to low-resource languages and under-represented communities. It will cover fairness-aware data practices and evaluation beyond capability scores, including robustness and reliability metrics and auditing frameworks, along with mitigation strategies that promote more equal model quality and reduced harms across languages and cultures.

Environmental and Societal Impact. *Sample Presenters: Margot Mieskes and Aurélie Névéol.* This section of the tutorial will take a broad approach at NLP system evaluation (especially with respect to Large Language Models — LLMs) by considering impact on users and on the environment.

Our community's experiments have a cost: The machines on which the experiments are run have to be built and require resources to run (e.g., electricity, water for cooling). Participants will learn

how to assess impact incurred by complex NLP systems at various stages of the software lifecycle, including training, distilling, inference.

Furthermore, the experiments have social implications because the language data being processed often raises questions related to user privacy, speaker representation, and bias related harm.

Ethical Considerations for Author and Reviewer roles.[1] *Presenters: Esther Gan, Min-Yen Kan, Yisong Miao, Margot Mieskes and Yajing Yang.*

Ethics review is a critical step in the review process, where flagging is handled by reviewers during their assessments. While essential for research integrity, flagging must be applied judiciously with appropriate justification, to avoid hindering the process and stalling valuable research.

Our tutorial will train potential scientific reviewers on how to make this decision: We will highlight identified patterns (based on previous ARR cycles) leading to unnecessary flagging, helping reviewers to more accurately determine when a paper truly needs a deeper ethics review. We will review common mistakes to avoid such as *flagging for missing section(s)*, and *flagging for reasons that could be addressed with feedback of the technical review*.

2.2 Case studies (Segments 2–5)

Our tutorial also covers the core ethics reviewing process itself. In the interactive portion of the tutorial, we will discuss research abstracts and will facilitate group discussions guided by critical questions about the proposed technology. Participants will be encouraged to discuss the following questions that will relate closely to the topics presented during Segment 1:

- Conflict of Interests (COIs): What constitutes a COI? How does one report it? What interactions could result members with different roles (reviewer and author)?

- Ethics of the research question: Would answering this research question advance science without violating social contracts? What are potentials for misuse?

- Social impact of the proposed technology and its potential dual use: Who could benefit from such a technology? Who can be harmed by such a technology? Could sharing data and models have major effects on people's lives?

- Privacy: Who owns the data? Understanding the differences between published versus publicized data, the concept of user consent, and implicit assumptions of users on how their data will be used.

- Bias in data: What are possible artifacts in data, given population-specific distributions? How representative is this data to address the target task?

- Social bias and unfairness in models: Is there sufficient control for confounding variables and edge cases? Is the system optimized for misaligned objectives or amplify data bias?

- Is the proposed evaluation sufficient? Is there a utility-based evaluation beyond accuracy; e.g., measurements of false positive and false negative rates as measurements of fairness? What is "the cost" of misclassification and fault (in)tolerance?

> **Abstract:** Faces contain more information about sexual orientation than can be perceived by the human brain. We used deep neural networks to extract features from over 35 thousand facial images. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 80% of cases, and in 70% of cases for women. Accuracy increased to 90% and 80%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Prediction models aimed at gender alone detected with 55% and 53% accuracy for gay males and gay females, respectively. Such findings advance our understanding of the origins of sexual orientation and the limits of human perception. Given that organizations are using computer vision algorithms to detect people's intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

Our case studies will be hypothetical; i.e., we will not use abstracts from existing studies but will create abstracts that will allow us to highlight potential ethical issues covering multiple, diverse ethics-related topics, including human subjects research and institutional review board (IRB) approval, bias and fairness, privacy, misinformation, toxicity/content moderation, energy considerations/green AI. We will develop several representative case studies for participants to choose from; we show an example below that illustrates multiple

[1] We note that there is complementary tutorial proposal from the ARR Chairs that will touch on ethics flagging. We believe that our tutorial section will complement theirs.

problematic aspects within one study, which was adapted from an actual problematic recent study.

The abstract above introduces an unethical research question, a demographically biased data set, a data collection procedure that violates user privacy, a problematic evaluation procedure, and claims/potential applications that can lead to significant harms to individuals.

2.3 Readings & Resources (Segment 7)

We will cover a diversity of primary research on ethics, sourced beyond the presenters' own works, in the final plenary sessions of the tutorial. Also, due to the abbreviated length of the 1/2-day format, our tutorial will cross reference sources from the list, rather than specifically require participants to do readings before the tutorial.

A full reading list[2] of over 200 works has been cross-compiled by the AEC, sourced from university courses on NLP Ethics and related topics. Topics and readings relate directly to the topics to be presented, which include: **data usage** (Mieskes, 2017; Bender and Friedman, 2018; Gebru et al., 2021), **crowdsourcing** (Fort et al., 2011; Callison-Burch, 2014; Fort et al., 2014; Hara et al., 2018; Toxtli et al., 2021), **biases** (Blodgett et al., 2020; Fort et al., 2024; Seth et al., 2025), **language diversity** (Tatman, 2017; Jurgens et al., 2017; Zmigrod et al., 2019; Tan et al., 2020; Koenecke et al., 2020; Bird, 2020), **rigorous and meaningful evaluation** (Caglayan et al., 2020; Tan et al., 2021), **environmental impact** (Strubell et al., 2019; Zhou et al., 2020; Henderson et al., 2020; Bannour et al., 2021) and **human harms and values** (Hovy and Spruit, 2016; Leidner and Plachouras, 2017).

Our tutorial also benefits from its prior iterations. We will extend our publicly-available tutorial presentation materials[3] available on the ACL Ethics website. As an example, annotated presentation slides (with presenter notes) will be modified and made available, such that tutorial participants can bring exercises of different lengths into classroom settings for research groups as well as undergraduate and graduate classes. While we plan to record the plenary, lecture-styled sessions (Sessions 1, 6 and 7), due to the sensitive and formative nature of the small-group discussions, we will not record the small-group discussions so that participants can speak freely and off-the-record.

[2] https://github.com/acl-org/ethics-reading-list

[3] https://ethics.aclweb.org/tutorials/

3 Diversity considerations

The instructors of this tutorial are affiliated in different geographic regions. Zhijing Jin in North America, Aurélie Névéol and Margot Mieskes in Europe; Vamshi Bonagiri, Monojit Choudhury and Abdoul Jalil Djiberou Mahamadou in the Middle East; Xi Ai, Esther Gan, Min-Yen Kan, Seunghun J. Lee, Yisong Miao and Yajing Yang in Asia. They also represent different career stages: PhD candidates, postdoctoral researchers, assistant, associate and full professors; Three are co-chairs of the AEC, two others are members of the AEC and one other is a student volunteer with the AEC. Five identify with the female gender and six with the male gender. We note that there are potentially more junior presenters that will be identified by the core investigators in due course.

We will promote this tutorial to all the ACL members but particularly to affinity groups such as Masakhane, LatinX, disabled in AI, indigenous in AI, and similar groups with the help of EquiCL Big Interest Group. It is crucial to reach a diverse audience for this tutorial to best foster rich discussions. We also aim to accommodate attendees needing special attention for registration or attendance.

4 Ethical considerations

We aware that we do not compose a perfectly diverse committee and commit to pay close attention to ensure all participants' points of views are faithfully acknowledged.

We decided to use synthetic case studies in the form of abstracts, rather than real and complete articles, in order to preserve the anonymity of the authors, to refrain from personal criticism, and to allow participants to focus on the discussion than on the reading. We will create a variety of abstracts (c.f., Appendix), with different forms, exemplifying different ethical issues; however, they will not cover all the possible ethical issues in the domain.

5 Logistics

Our tutorial requires standard conference presentation equipment, including a projector, screen, and both presenter and audience microphones (where feasible), and reliable internet access (for remote participation).

6 Limitations

We are requesting to check whether there are restrictions on content that should be avoided to be compliant to the local venue (especially within China, the venue for AACL-IJCNLP).

7 Presenters (listed in alphabetical order)

Note: We have annotated each presenter with two-letter codes to indicate which venue ((E)MNLP or (A)ACL-IJCNLP) and modality of participation (**(P)hysical** or *(V)irtual*), where already known.

Xi Ai [E*V*, A**P**] (*baridxai@u.nus.edu*, he/him) is a research fellow at WING@NUS and NUSRI-CQ. His research focuses on multilingual and cross-lingual NLP and machine learning for NLP. He has co-organized Workshop on Computational Approaches to Linguistic Code-Switching (CALCS) at *CL conferences.

Vamshi Bonagiri [E**P**, A*V*] (*vamshi.bonagiri@mbzuai.ac.ae*, he/him) is a PhD student at MBZUAI. His research focuses on AI safety and the societal impacts of AI technologies. He has also taught and mentored various NLP and responsible and safe AI courses.

Monojit Choudhury [E**P**] (*monojit.choudhury@mbzuai.ac.ae*, he/him) is a Professor of NLP at MBZUAI, Chief Scientist of the Institute of Agriculture and AI, and a member of AEC. His work focuses on equitable creation and deployment of language technology.

Abdoul Jalil Djiberou Mahamadou [E*V*, A**P**] (*abdoul.mahamadou@mbzuai.ac.ae*, he/him) is a postdoc at MBZUAI working on personalization, cultural sensitivity, and inclusion in AI. He did a postdoc in AI ethics at the Stanford Centre for Biomedical Ethics, where he taught bioethics and research ethics. He researched algorithm bias, fairness, and participatory AI governance.

Esther Gan [E*V*, A**P**] (*esther.gan@u.nus.edu*, she/her) is a doctoral student with NUS and an active volunteer with the AEC. She has taught topics related to ethics as NUS teaching staff.

Min-Yen Kan [E*V*, A**P**] (*knmnyn@nus.edu.sg*, he/him) is an Associate Professor at NUS and a co-chair of the AEC. He was a member of the ACL executive committee, the inaugural Information Officer and a previous ACL Anthology Director.

Seunghun J. Lee [A**P**] (*seunghun@icu.ac.jp*, he/him) Senior Associate Professor of Linguistics at the International Christian University. He is also an Adjunct Professor of African Languages at the University of Venda in South Africa, and an Honorary Associate Professor in the Centre for Linguistic Science & Technology at the Indian Institute of Technology Guwahati, India.

Yisong Miao [E*V*, A**P**] (*yisong@comp.nus.edu.sg*, he/him) is a doctoral student at NUS. He has been an active member of the discourse area in ACL. He has been a teaching assistant at NUS for its NLP course, covering ethical topics.

Margot Mieskes [E**P**, A*V*] (*margot.mieskes@h-da.de*, she/her) Professor for Information Science at the University of Applied Sciences, Darmstadt, Germany. She works in both the CL and NLP areas, focusing on summarization, evaluation, data quality, ethics, experiment reproducibility, and biases in LLMs. She is also part of the ACL Rolling Review Editor in Chief team.

Aurélie Névéol [E**P**, A*V*] (*aurelie.neveol@lisn.fr*, she/her) is a Senior Researcher at CNRS and a co-chair of the AEC She has been a member of SIG BioMed since 2006. Her work addresses NLP applied to the biomedical domain guided by responsible science principles.

Yajing Yang [E*V*, A**P**] (*yajing.yang@u.nus.edu*, she/her) is a doctoral student at NUS where her research focuses on financial NLP and data-to-text generation. She was a teaching assistant for NUS' NLP course.

Zhijing Jin [E**P**, A*V*] (*zjin@cs.toronto.edu*, she/her) is an Assistant Professor at the University of Toronto. Her research focuses on LLMs, multi-agent system, causal inference in NLP with an emphasis on social good applications. She is a co-chair of the AEC and co-organized NLP for Positive Impact workshops at *CL conferences and the ACL Year-Round Mentorship.

References

Nesrine Bannour, Sahar Ghannay, Aurélie Névéol, and Anne-Laure Ligozat. 2021. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 11–21, Virtual. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Luciana Benotti, Mark Dredze, Karën Fort, Pascale Fung, Dirk Hovy, Jin-Dong Kim, Min-Yen Kan, Malvina Nissim, and Yulia Tsvetkov. 2023a. Yes, we care (more)! results of the 2021 ethics and natural language processing survey. Technical Report v230714, ACL Ethics Committee. Report on the 2021 Ethics and Natural Language Processing Survey, released 14 July 2023.

Luciana Benotti, Fanny Ducel, Karën Fort, Guido Ivetta, Zhijing Jin, Min-Yen Kan, Seunghun J. Lee, Minzhi Li, Margot Mieskes, and Adriana Pagano. 2025. Navigating ethical challenges in NLP: Hands-on strategies for students and researchers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 7–8, Vienna, Austria. Association for Computational Linguistics.

Luciana Benotti, Karën Fort, Min-Yen Kan, and Yulia Tsvetkov. 2023b. Understanding ethics in NLP authoring and reviewing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–24, Dubrovnik, Croatia. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chris Callison-Burch. 2014. Crowd-workers: Aggregating information across turkers to help them find higher paying work. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2(1):8–9.

Kevin Cohen, Karën Fort, Margot Mieskes, Aurélie Névéol, and Anna Rogers. 2021. Reviewing natural language processing research. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 14–16, online. Association for Computational Linguistics.

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Karën Fort, Gilles Adda, Benoît Sagot, Joseph Mariani, and Alain Couillault. 2014. Crowdsourcing for language resource development: Criticisms about amazon mechanical turk overpowering use. In *Human Language Technology Challenges for Computer Science and Linguistics*, pages 303–314, Cham. Springer International Publishing.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, Javier Torroba Marchante, Shilin Xie, Sergio E. Zanotto, and Aurélie Névéol. 2024. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association*

for Computational Linguistics (Volume 2: Short Papers), pages 591–598, Berlin, Germany. Association for Computational Linguistics.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Margot Mieskes. 2017. A quantitative study of data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain. Association for Computational Linguistics.

Agrima Seth, Monojit Choudhury, Sunayana Sitaram, Kentaro Toyama, Aditya Vashistha, and Kalika Bali. 2025. How deep is representational bias in llms? the cases of caste and religion. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2319–2330.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4153–4169, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020. Mind your inflections! Improving NLP for non-standard Englishes with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5647–5663, Online. Association for Computational Linguistics.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the invisible labor in crowd work. *ACM Human Computer Interaction*, 5:1–26.

Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. 2018. Socially responsible NLP. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 24–26, New Orleans, Louisiana. Association for Computational Linguistics.

Sharon Zhou, Alexandra Luccioni, Gautier Cosne, Michael S Bernstein, and Yoshua Bengio. 2020. Establishing an evaluation metric to quantify climate change image realism. *Machine Learning: Science and Technology*, 1(2):025005.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

8 Appendix

We are unsure whether appendices are allowed, but we have included examples of other abstracts Here are examples of other abstracts our team is developing and may field during the actual tutorial. Please disregard if not allowed.

Abstract: Do people feel better on when on vacation? We study the social media sentiment of families when on vacation and during normal periods. Using a snowball sampling method, we solicited participants active on various social media platforms ($N = 1,337$) to voluntarily disclose when and where they went on vacation. We then crawled our participants social media posts to determine whether specific time periods and destinations for vacations were correlated with higher levels of enjoyment. Initial analysis showed that such features do yield significant prediction performance improvement: for example, vacations taken close-by tended to have markedly higher satisfaction. Our model, VACAY-OK, accounts for such factors in novel gating network architecture, improving joint vacation period and sentiment detection by over 10% F_1. Our analysis validates patterns of interest: vacations with immediate family exhibit a bimodal distribution, and that ones without in-laws have a significantly higher satisfaction and detection rate. For robustness in identification, we collect and study the pattern and text of social media posts by parents, children, extended family members and friends of vacation-goers. To ensure both the accessibility and the reproducibility of our experiments, we have gathered the raw social media data, culled with permission from the original posters, and released it as a zip archive, available without moderation, in the footnote.

The above abstract illustrates challenges with minors giving consent and with individuals releasing confidential data unknowingly (details shared by friends).

Abstract: In this paper, we present the largest existing language model as of today (989 trillion parameters, dataset of 968 Pb), `BigBlue`. We trained it using not only freely available Web content, but also Web archives and publicly-accessible EPub versions of all the books of the commericial Amazonia bookshop web storefront. The best results were attained using reinforcement learning with human feedback (RLHF) coming from crowdworkers. The model improves the state-of-the-art in Natural Language Processing (NLP) in most tasks, including sentiment analysis (+0.02 F-measure), dependency syntax (+0.015 UAS), named-entity recognition (+0.05 F-measure), etc. `BigBlue` is available through an API on our company's website.

The above abstract illustrates some of the many issues around LLM, including carbon footprint.

Abstract: In this paper, we present the first language model developed for Fridonian, a regional language spoken in the South-East of Austrafrancia by more than 500,000 people. We detail the collection of the corpus, the creation of the model and its evaluation on basic NLP tasks. The corpus creation itself is an achievement, as it is well-known and documented (Birdoff, 2020) that Fridonian speakers are reluctant to allow the recording of their language, for fear it will be used against them by the country's Sidonian majority. However, we managed to record 100 hours of speech in different settings: family discussions, local authority meetings, ceremonies and traditional story telling contests. We used this dataset to train a language model which we evaluated on sentiment analysis and named-entity recognition tasks, with impressive results (resp. 0.56 and 0.75 in F-measure). The created language model (`FridoBERT`) and the dataset (`FridoSPEECH`) are freely available on GitHub.

The above abstract illustrates the lack of consideration of the speakers and their needs and context.

Abstract: Is sentiment analysis dependent on culture or only on language? In this paper we describe the collection and annotation of a billion-word dataset over 7 different Spanish dialects spoken in Latin America, annotated with fine-grained sentiment analysis cues. To the best of our knowledge, this is the largest Latin American dataset currently available. We used data from public groups in Telegram where locals talk about products and services. We evaluate the state-of-the-art sentiment analysis models on it, showing that their performance is significantly lower than the state-of-the-art results on well-known benchmark for Spanish. The dataset annotation took three months. We selected crowdworkers after a careful exam to ensure they were fluent in the target dialect. We payed them 4 US dollars per hour, a fair wage considering it was two times the minimum wage in these countries. Our sentiment classifier fine-tuned per dialect outperforms previous models on Spanish by large margins. Our model has been used for 12 months by hundreds of companies in Latin America. To serve the model, we implemented a costing model to provide this service at a uniform price to all regions of the world. We conducted three usability studies in collaboration with three companies, which demonstrated that the integration of our sentiment analysis detector markedly increased users' engagement with their marketing advertisements.

The abstract presents several potential ethical challenges in data collection, model performance, and commercialization. First, the dataset was collected from public Telegram groups, raising concerns about privacy and informed consent, as users may not have been fully aware their data was being harvested for sentiment analysis. Additionally, sentiment analysis models can be culturally dependent, and there is an ethical risk that the fine-grained cues might not accurately capture the nuances of different dialects, leading to biases or misrepresentation. Economic fairness is another issue, as the crowdworkers were paid 4 USD per hour, which,

while considered twice the minimum wage in some regions, may still be viewed as exploitative given the global profits likely generated from their labor. Furthermore, the model's use by companies for marketing could raise concerns about manipulation of consumer behavior, especially if the model tailors advertising based on dialect and exploits specific cultural tendencies. Lastly, while the uniform global pricing model for the sentiment analysis service may appear equitable, it could disadvantage regions with lower purchasing power, perpetuating technological and economic disparities. Addressing these ethical challenges requires transparency in data use, fair compensation for labor, and minimizing biases in AI models across cultural contexts.

Abstract: With the rise of large language models (LLMs), they have become useful in critical settings such as healthcare support, helping reduce administrative burden and improve predictive analytics. However, there is an emerging concern that LLMs encode gender and racial biases, which propagate to healthcare decisions. In this work, we propose a simple and effective approach to controlling for biases in LLMs through data anonymization, specifically focusing on patient names that often reveal gender and race. We introduce a novel and highly granular dataset of over 9.9 million patients' electronic health records annotated with patient demographics, including gender and race. We first fine-tune state-of-the-art LLMs on the raw healthcare data, and establish the presence of harmful biases in standard NLP benchmarks such as coreference resolution (10% lower F1 score for women compared to men) and named entity recognition (7% lower for White compared to Black patients). We also show that naively fine-tuned language models can be used to predict gender and race of the patients in held-out health records with high accuracy. Then, we show that a simple anonymization of the health data, by replacing patient names with generic placeholders, reduces the prediction gap in finetuned LLMs by up to 4% for gender, and up to 2.5% absolute for race. Our findings address important questions for fairness in NLP and algorithmic decision-making. Our code and data are publicly available to facilitate reproducibility.

The above abstract illustrates privacy concerns with using healthcare data, releasing data publicly, the problem with assumptions about privacy – that obfuscating names does not anonymize data since many other features correlate with demographics, annotation issued with binary gender and race, US-centric assumptions about conceptualization of race, evaluation of fairness on intrinsic benchmarks that might not reflect the true use case scenarios. Additional potential discussion is on what constitues "high accuracy" when the harm from misclasification is high (the weight of wrong predictions should be high).

A Tutorial on Measurement Science for Language Model Evaluation

Sang Truong[1], **Serena Wang**[2], **Rodolfo Corona**[3], **Sanmi Koyejo**[1]
[1]Stanford University, [2]University of British Columbia, [3]Harvard University
sttruong@cs.stanford.edu

Abstract

Benchmark scores are used to support claims about language model capabilities, comparisons, and suitability for deployment. The evidence needed for these claims depends on what is measured, how observations are collected, and how evaluation shapes model developer behavior. This tutorial brings together measurement theory, statistical modeling, and mechanism design to examine six common claims in language model evaluation. Part I introduces measurement principles for interpreting benchmark results and assessing the scope of the claims they support. Part II examines how evaluations shape model developer incentives and how their design can encourage model improvements that serve diverse user needs. For each claim, we explain when it can hold, where it breaks, and how to formulate a better-supported interpretation. Worked examples and short discussions connect these principles to practical evaluation decisions. Participants leave with practical tools for stating evaluation targets, selecting relevant evidence, and reporting the limits of conclusions drawn from benchmark scores.

1 Description and Learning Outcomes

Benchmarks are a central instrument through which the NLP community compares language models and assesses their suitability for use. Yet an observed score and the inference drawn from it are different objects. "The model gets 87% on this test" describes performance under a particular protocol. Claims that a model reasons well, outperforms another model, or is suitable for a deployment setting require additional evidence. Broad benchmark coverage alone does not establish that the evaluation measures the intended construct or supports the intended use (Messick, 1995; Raji et al., 2021; Liang et al., 2023).

This tutorial draws on measurement theory, statistical inference, and mechanism design to examine what benchmark results justify and what evaluations reward. We ask when evidence from an evaluation supports its intended use, and how evaluation design shapes model development. The emphasis is on drawing defensible conclusions and designing evaluations that encourage improvements serving diverse user needs.

The tutorial is organized around six recurring claims in two parts. **Part I: Measurement Theory** examines construct validity, comparable and efficient measurement, variation across evaluation conditions, and deployment relevance. **Part II: Incentive-aware Robust Evaluation** examines how evaluation design can reward progress that serves users. Across both parts, each claim follows the same progression: *what holds, where it breaks, and a better-supported claim*. We distinguish sampling uncertainty, which can sometimes be reduced by collecting more observations, from specification problems, for which more observations of the wrong target do not supply the missing validity evidence.

Learning outcomes. Participants will learn to:

- State the construct, target population, protocol, and deployment decision behind an evaluation claim.
- Assess whether benchmark aggregation supports the named capability.
- Explain how linked calibration and adaptive item selection support efficient comparisons between models evaluated on different questions.
- Use response patterns to prioritize item review and distinguish shared, model-specific, and item-specific language effects.
- Check whether the evaluated outcome and system match the intended deployment, and identify incentives that can improve scores without improving that outcome.

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 20–26
November 6-10, 2026 ©2026 Association for Computational Linguistics

These skills are relevant to benchmark authors, researchers comparing models, and practitioners selecting systems. The tutorial emphasizes the conditions under which an evaluation is informative and the evidence needed to extend its interpretation. It connects foundational work in psychometrics with contemporary NLP case studies and economic models of evaluation.

2 Related Tutorials

The tutorial builds on complementary treatments of evaluation metrics (Khapra and Sai, 2021), evaluation practice and reproducibility (Sutawika and Schoelkopf, 2024), item response theory for NLP (Lalor et al., 2024), human-centered evaluation (Blodgett et al., 2024), and the science of benchmarking (Ma et al., 2025). Our organizing focus is the inference a practitioner wants to draw from a score. Six concrete claims connect construct validity, common-scale comparisons, evaluation conditions, deployment decisions, and incentives, with worked examples that make their assumptions and limitations explicit.

3 Target Audience and Prerequisites

The intended audience is CL/NLP researchers, students, and practitioners who build benchmarks, report model comparisons, or use evaluation results to select systems. We assume familiarity with language model evaluation and basic probability and statistics, including averages, variance, and uncertainty. Familiarity with supervised learning and maximum-likelihood estimation is helpful for the statistical examples. No prior background in psychometrics or mechanism design is required; item response theory is introduced through worked examples. Discussion prompts require no programming setup. Python familiarity is useful only for optional notebook material.

4 Tutorial Outline

The tutorial follows the sequence in Table 1: 180 minutes of instruction and discussion, with a 30-minute break between C3 and C4. Corona presents C1 and C4, Truong presents C2 and C3, and Wang presents C5 and C6, with 50 minutes per instructor. Koyejo leads the 15-minute opening and 15-minute closing. Each claim combines a motivating example, the relevant measurement or incentive concepts, and a more defensible interpretation.

Segment	Minutes
Opening: evaluation as measurement	15
C1–C3: constructs, scales, language effects	75
Break	30
C4–C6: deployment and incentives	75
Closing synthesis and discussion	15
Total	210

Table 1: Three-and-a-half-hour tutorial schedule: 180 minutes of instruction and discussion and a 30-minute break.

Opening: Evaluation as Measurement. We motivate the distinction between an observed score and a claim about capability or deployment. Examples of benchmark defects, incomplete reporting, and cultural assumptions introduce the need to state what is being measured (Burnell et al., 2023; Singh et al., 2025). The opening distinguishes sampling error from specification error and introduces the six questions below.

4.1 Part I: Measurement Theory

C1: Does averaging benchmark scores robustly establish a capability? Averaging can reduce benchmark-specific variation while preserving what the benchmarks share. That shared component need not be the capability named by their labels. We introduce convergent and discriminant validity and examine evidence about benchmarks grouped under capability and safety concepts (Messick, 1995; Desai et al., 2026; Ren et al., 2024). An application example considers how coverage and weights determine what an aggregate healthcare score represents (Bedi et al., 2026; Arora et al., 2025). Participants learn to ask both whether benchmarks agree and whether that agreement supports the intended construct, rather than relying on a shared label alone.

C2: Can we compare models reliably with fewer evaluations? Taking a coherent construct as the starting point, we ask how to compare models evaluated on different question subsets. Expensive agent trajectories motivate collecting fewer responses per model. We introduce the Rasch model as a common scale for model ability and item difficulty, emphasizing linked calibration as the basis for comparison. Amortized calibration and adaptive selection using Fisher information illustrate how to obtain informative measurements with fewer evaluations (Lord, 1980; Truong et al., 2025a). We then show how expected response associations can guide item inspection: correlation-

based diagnostics in *Fantastic Bugs* help prioritize review and uncover benchmark or grading problems (Truong et al., 2025b). A statistical flag calls for inspection; it does not establish a flaw on its own. The takeaway is that a response model can support efficient measurement and identify questions worth inspecting, subject to adequate calibration and model fit.

C3: After translation, are scores still comparable? Observed jailbreak rates vary across languages, and English is not always the least vulnerable language (Zhang et al., 2026). This empirical pattern motivates adding language as an evaluation facet. For exposition, we extend C2's Rasch model with language effects to distinguish a shared language shift from model–language and prompt–language effects. The case study shows that model families can shift in opposite directions and individual prompts can depart substantially from the shared shift; large differences can persist even for accurate translations. Participants learn to examine which aspects of the evaluation contribute to a score gap before interpreting it as a change in model safety. Language provides a concrete example of how averages can obscure systematic variation across evaluation conditions.

C4: Does a leaderboard winner fit my product? A stable ranking does not by itself establish that a model meets a deployment requirement, and model's ranking can change depending on the components it's paired with in a system. We examine whether the scored outcome matches the deployment outcome and whether the evaluated system matches the system being deployed. Examples consider differences between benchmark success and task completion (Whitfill et al., 2026; Tonja et al., 2026), as well as changes to the system through prompts, tools, or interaction protocols (Feng et al., 2025; Mizrahi et al., 2024; Liu et al., 2026). Assessment of benchmark ontology, content, and form helps participants articulate which aspects of the deployment are represented and which are missing (Corona et al., 2026). Participants learn to specify a deployment outcome and seek evidence for that outcome rather than assuming a leaderboard position answers the product question.

4.2 Part II: Incentive-aware Robust Evaluation

C5: Is evaluation gaming mostly cheating or data leakage? We broaden the discussion beyond contamination to repeated model releases, optimization for evaluated tasks, tailoring to particular user populations, and selective disclosure of evaluations. These behaviors can arise from following the incentives an evaluation creates, even without breaking its rules. Goodhart's law and mechanism design provide a language for separating developer objectives from the outcomes an evaluator wants to encourage (Goodhart, 1984; Manheim and Garrabrant, 2018; Wang et al., 2025). Participants learn to describe how a score could improve while a relevant capability, population, or deployment objective is neglected.

C6: Do private, uncontaminated test sets solve gaming? Private holdouts address some failure modes, but evaluation design must also account for adaptation, information asymmetry, and heterogeneous preferences. We introduce an evaluator–developer game and explain how randomized evaluation can change incentives for capability coverage under explicit assumptions (Dwork et al., 2015; Truong et al., 2026). Preference leaderboards illustrate a second problem: optimizing average user preference may neglect substantial minority groups. We discuss how pluralistic rankings can represent these preferences, with a brief overview of robust evaluation and counterfactual metrics (Haghtalab et al., 2026; Khalaf et al., 2026; Wang et al., 2024). The emphasis is on whether an evaluation rewards useful model development and supports decisions across user groups, under the assumptions each design requires.

Closing Synthesis and Discussion. We combine the six claims into an evaluation checklist: define the claim and target population; inspect the items and scoring; check calibration and variation across evaluation conditions; seek deployment evidence; examine incentives; and report item-level responses and the full evaluation protocol. We briefly discuss open questions about changing model behavior, validating automated scorers, and testing generalization to future tasks. The closing asks participants what evidence would make a benchmark score useful for their own setting and how they would report the supported interpretation and its limits.

Reading List. Recommended starting points are Messick (1995); Desai et al. (2026) for validity; Lord (1980); Truong et al. (2025a,b) for measurement and item diagnostics; Zhang et al. (2026); Corona et al. (2026) for evaluation conditions and

deployment; and Dwork et al. (2015); Wang et al. (2025); Haghtalab et al. (2026) for incentives and robust evaluation. The reading list combines foundational work, research by the instructors, and work by other authors. An expanded bibliography accompanies the tutorial materials.

5 Diversity and Accessibility

Linguistic, cultural, and population diversity are technical concerns throughout the tutorial. C3 examines how language effects vary across models and prompts in multilingual safety evaluation; C4 asks whose tasks and deployment contexts an evaluation represents; and C6 examines how aggregate preferences can overlook minority user groups. Benchmarks developed in English-speaking, Western contexts provide motivating examples of a broader problem: evidence from one setting need not support the same interpretation in another. The instructors bring expertise in computer science, statistics, policy, economics, and measurement. Discussion prompts require no local software installation, and publicly available slides and readings support participation beyond the conference.

6 Tutorial Materials

The tutorial website, `https://aimslab.stanford.edu/tutorial`, provides the bibliography and a central location for slides and supplementary resources. The instructional materials comprise worked examples, discussion prompts, and the closing evaluation checklist. Any optional notebooks will be linked from the same site; they are not required for participation. We will make the slides and bibliography publicly available.

7 Ethics Statement

Validity judgments depend on the intended use and on the perspectives of affected populations. Statistical diagnostics and automated audits can inform human review but cannot determine cultural appropriateness or justify deployment on their own. We discuss gaming to motivate defenses and make their assumptions explicit; the same insights could also inform strategic behavior. Better measurement can support harmful applications as well as beneficial ones, so measurement quality does not substitute for evaluating the acceptability of an application. Throughout the tutorial, we emphasize uncertainty, representation, and limits on the conclusions an evaluation supports.

Instructor Biographies

Sang Truong (Stanford University, US) is a Ph.D. candidate in Computer Science at the Stanford Artificial Intelligence Laboratory. His research develops foundational and applied methods for AI measurement sciences, drawing on tools from measurement theory, probabilistic machine learning, and psychometrics to support the scientific development of safe AI systems that serve people across diverse backgrounds and needs. He has co-organized the 1st workshop on AI Measurement Science at the Conference on Language Modeling, the 1st and 2nd Workshops on Language Models for Underserved Communities (LM4UC) at NAACL 2025 and AAAI 2026, as well as the Data-centric Machine Learning Research Workshops at ICML 2024 and ICLR 2024.

Serena Wang (University of British Columbia; serenalw@cs.ubc.ca) is an Assistant Professor of Computer Science at the University of British Columbia. She previously held a postdoctoral fellowship at Harvard University. Wang's research focuses on incentives in the AI evaluation ecosystem, including mechanism design and information design approaches to robust evaluation design. Her recently published work includes strategic evaluation to mitigate gaming behavior (Wang et al., 2024), robust leaderboard design under preference heterogeneity (Procaccia et al., 2025; Khalaf et al., 2026; Haghtalab et al., 2026), and incentives for information sharing between model developers and evaluators (Wang et al., 2025). Wang is a Canada CIFAR AI Chair, and a recipient of the Harvard Center for Research on Computation and Society Postdoctoral Fellowship, the Apple Scholars in AI/ML PhD Fellowship and the NSF Graduate Research Fellowship. She has previously organized workshops at AAAI and Harvard, and given tutorials at the Open Data Science Conference and Grace Hopper Conference.

Rodolfo Corona (Harvard University; rcorona@hks.harvard.edu) is a Postdoctoral Fellow in Technology and Geopolitics at the Harvard Kennedy School Belfer Center. He received his Ph.D. in Computer Science from UC Berkeley, where he was advised by Dan Klein and Trevor Darrell. His research spans grounded language learning, evaluations, and AI governance, having presented his work at ACL, NAACL, and NeurIPS. Rodolfo's current research is focused on AI benchmarks, studying how the cultural, regional, and deployment contexts of AI systems impact the validity of their evaluations. Rodolfo has taught international audiences, particularly in Latin American conferences. He has served as Hackathon and Practicals Chair for SALA 2026, as an organizer and instructor for the KHIPU 2025 practicals, and on the advisory board for RIIAA 6.0. He has served as a graduate student instructor for UC Berkeley's CS 288 (graduate Natural Language Processing) and CS 188 (Artificial Intelligence) courses. Rodolfo is a recipient of the NSF Graduate Research Fellowship, the Ford Foundation Fellowship, the UC Berkeley Chancellor's Fellowship, and a Fulbright Student Research Grant.

Sanmi Koyejo (Stanford University; sanmi@cs.stanford.edu) is an associate professor in the Department of Computer Science at Stanford University. He leads the Stanford Trustworthy Artificial Intelligence Research (STAIR) lab, focusing on the principles and practice of trustworthy machine learning. His research on emergent behaviors, scaling laws, and novel evaluation methodologies has been recognized with several awards, including the NeurIPS, ACL, and UAI Outstanding Paper Awards, the Presidential Early Career Award for Scientists and Engineers, a Skip Ellis Early Career Award, a Sloan Fellowship, and an NSF CAREER Award. He served as co-General Chair of NeurIPS 2022, Workshop Chair at NeurIPS 2021 and ICLR 2021, and has co-organized workshops including ICML 2024 Workshop on Theoretical Foundations of Foundation Models (TF2M) and ICLR 2024 Workshop on Reliable and Responsible Foundation Models (RRFM).

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, and 1 others. 2026. Holistic evaluation of large language models for medical tasks with medhelm. *Nature medicine*, 32(3):943–951.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts (EMNLP)*, pages 39–43.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. 2023. Rethink reporting of evaluation results in AI. *Science*, 380:136–138.

Rodolfo Corona, Sang T. Truong, Nhi N. Truong, Ritwik Gupta, Fred Philippy, Atnafu Lambebo Tonja, Mena Attia, Fahim Faisal, Kaushal Kumar Maurya, Belu Ticona, Sumaya Nur Adan, Fazl Barez, Omar U. Florez, Supheakmungkol Sarin, Aseem Srivastava, Xiaoyuan Yi, Nick Haber, Dan Klein, Thamar Solorio, and 3 others. 2026. Uplifting human decision making in AI evaluation by automating benchmark validity analysis. In *Workshop on Combining Theory and Benchmarks at ICML*.

Meera Desai, Sang T. Truong, Hanna Wallach, Alex Chouldechova, A. Feder Cooper, Jean Garcia-Gathright, Daniel E. Ho, Abigail Z. Jacobs, Sanmi Koyejo, Nicholas Pangakis, and Angelina Wang. 2026. What AI benchmarks actually measure: Adapting convergent and discriminant validity to interrogate fifty-six AI benchmarks. In *Conference on Language Modeling*.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science*.

Haiwen Feng, Long Lian, Lisa Dunlap, Jiahao Shu, XuDong Wang, Renhao Wang, Trevor Darrell, Alane Suhr, and Angjoo Kanazawa. 2025. Visually prompted benchmarks are surprisingly fragile. *arXiv preprint arXiv:2512.17875*.

Charles A. E. Goodhart. 1984. Problems of monetary management: The UK experience. In *Monetary Theory and Practice*, pages 91–121. Macmillan.

Nika Haghtalab, Ariel D. Procaccia, Han Shao, Serena Lutong Wang, and Kunhe Yang. 2026. Pluralistic leaderboards. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 39356–39375. PMLR.

Hadi Khalaf, Serena Lutong Wang, Daniel Halpern, Itai Shapira, Flavio Calmon, and Ariel D. Procaccia. 2026. Robust AI evaluation through maximal lotteries. In *Proceedings of the 43rd International Conference on Machine Learning*.

Mitesh M. Khapra and Ananya B. Sai. 2021. A tutorial on evaluation metrics used in natural language generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials (NAACL-HLT)*, pages 15–19.

John P. Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts (EACL)*, pages 9–13.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*.

Fengshuo Liu, Ying Liu, Ruize Sun, Lie Luo, and Siyuan Guo. 2026. Coding agents have converged: Why the swe-bench leaderboard can no longer order its top entries, and what to measure instead. *arXiv preprint arXiv:2609.17394*.

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.

Ziqiao Ma, Michael Saxon, and Xiang Yue. 2025. The science of benchmarking: What's measured, what's missing, what's next. Tutorial at the 39th Conference on Neural Information Processing Systems (NeurIPS).

David Manheim and Scott Garrabrant. 2018. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation.

Transactions of the Association for Computational Linguistics, 12:933–949.

Ariel Procaccia, Ben Schiffer, Serena Wang, and Shirley Zhang. 2025. Metritocracy: representative metrics for lite benchmarks. *Advances in Neural Information Processing Systems*, 38.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *NeurIPS Datasets and Benchmarks Track*.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. 2024. Safetywashing: Do AI safety benchmarks actually measure safety progress? In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Lintang Sutawika and Hailey Schoelkopf. 2024. Challenges in language model evaluations. Tutorial at the 41st International Conference on Machine Learning (ICML).

Atnafu Lambebo Tonja, Srija Anand, Emilio Villa-Cueva, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Muhidin A. Mohamed, Debela Desalegn Yadeta, Negasi Haile Abadi, Abigail Oppong, Nnaemeka Casmir Obiefuna, Idris Abdulmumin, Naome A Etori, Eric Peter Wairagala, Kanda Patrick Tshinu, Imanigirimbabazi Emmanuel, Gabofetswe Malema, Alham Fikri Aji, David Ifeoluwa Adelani, and Thamar Solorio. 2026. Afri-MCQA: Multimodal cultural question answering for African languages. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40249–40282, San Diego, California, United States. Association for Computational Linguistics.

Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. 2025a. Reliable and efficient amortized model-based evaluation. *Preprint*, arXiv:2503.13335.

Sang Truong, Serena Wang, Nick Haber, and Sanmi Koyejo. 2026. Strategic evaluation: Incentivizing AI capability coverage with private benchmarks.

Sang T. Truong, Yuheng Tu, Michael Hardy, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Benjamin W. Domingue, Nick Haber, and Sanmi Koyejo. 2025b. Fantastic bugs and where to find them in AI benchmarks. In *Conference on Neural Information Processing Systems, Datasets and Benchmarks Track*.

Serena Wang, Stephen Bates, P Aronow, and Michael Jordan. 2024. On counterfactual metrics for social welfare: Incentives, ranking, and information asymmetry. In *International Conference on Artificial Intelligence and Statistics*, pages 1522–1530. PMLR.

Serena Wang, Michael Jordan, Katrina Ligett, and R Preston McAfee. 2025. Relying on the metrics of evaluated agents. In *Proceedings of the ACM on Web Conference 2025*, pages 1468–1487.

Parker Whitfill, Cheryl Wu, Joel Becker, and Nate Rush. 2026. Many swe-bench-passing prs would not be merged into main. METR research note.

Max Zhang, Ameen Patel, Sang T. Truong, and Sanmi Koyejo. 2026. Why do safety guardrails degrade across languages? In *Conference on Language Modeling*.

Post-Deployment Self-Improving Language Agents: Learning, Memory, and Adaptation

Estevam Hruschka
Megagon Labs
Mountain View, CA, USA
estevam@megagon.ai

Abstract

Large language models are increasingly deployed as long-running language agents that retrieve information, use tools, maintain memory, and operate in environments that change after deployment. Yet post-deployment learning is still often handled as a loose collection of prompt edits, retrieval refreshes, memory writes, corrections, or occasional fine-tuning. This tutorial develops a unified view of **post-deployment self-improving language agents**: systems that use deployment experience to produce persistent changes that improve subsequent performance while remaining empirically grounded, auditable, and controllable.. We connect modern agent research with never-ending, lifelong, and continual learning while shifting the unit of analysis from model parameters only to the complete deployed system. We explore the literature, from classic papers to very recent approaches. As running example, we use a information-extraction agent that continuously extracts typed entities and relations from an evolving document stream into a persistent knowledge base, encounters new terminology and relations, receives corrections and conflicting evidence, and uses retained knowledge to influence later extraction. This example grounds important questions: what should be learned, what evidence should justify an update, where the change should reside, and how it should be validated, governed, and evaluated over time. Particular emphasis is placed on persistent state, learning substrates, provenance, feedback loops, rollback, regression detection, forgetting, temporal generalization, and trajectory-based evaluation. The material for this tool is available at: github.com/megagonlabs/pdsi-tutorial .

1 Motivation and Scope

NLP is moving from single-turn prediction toward persistent language agents that retrieve documents, call tools, maintain memory, interact with users, and act over extended time horizons. Their environments are inherently non-stationary: knowledge sources and terminology change, APIs evolve, and new failure modes emerge after release. Work on retrieval, tool use, memory, reflection, post-training, and agent evaluation has consequently made adaptation a central problem for NLP (Lewis et al., 2020; Schick et al., 2023; Yao et al., 2023; Shinn et al., 2024; Salama et al., 2025).

The tutorial goes beyond models that can adapt within their current context: it explores whether a deployed system can use experience to improve its subsequent performance, and retain what it has learned across interactions. A system may retrieve an updated document without retaining anything from the episode; conversely, it may keep its weights frozen while learning a reusable extraction pattern, concept, procedure, or routing rule. The scientific object is therefore the evolving compound agent rather than only a checkpoint.

To keep this argument concrete, we use a fictional information-extraction agent, A_1, throughout. $A1$ reads scientific and technical documents and extracts typed entities and relations into an evolving knowledge base (KB_1). It is inspired by the closed-loop pattern exemplified by DySECT, where extraction populates a KB and accumulated knowledge feeds back to improve later extraction (Amin-Naseri et al., 2026). Our example extends that pattern to corrections, provenance, ontology revision, conflicting evidence, update gating, regression testing, and rollback, and by conneting these aspects to recent approaches described in the literature.

This perspective connects modern agents with lifelong learning's emphasis on transfer (Thrun, 1998), continual learning's concern with change and preservation (De Lange et al., 2022), and never-ending learning's sustained acquisition in systems whose current knowledge shapes what they learn next (Carlson et al., 2010; Mitchell et al., 2018).

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 27–31
November 6-10, 2026 ©2026 Association for Computational Linguistics

2 What Counts as Learning?

Self-improving systems are often described through actions, such as generating a critique, retaining a memory, updating a KB, revising a prompt, training or fine-tuning a model. None alone establishes improvement. We distinguish four elements. **Experience** includes documents, extracted candidates, feedback, failures, corrections, and observations. **Persistent state** is information retained across the declared episode or deployment boundary. **Learning substrates** are components that may change, such as memory, retrieval, structured knowledge, instructions, tools, policies, evaluators, or parameters. **Evaluation** determines whether the retained change improves a declared capability without unacceptable regressions.

For A_1, correcting *(Model-X, evaluated-on, Benchmark-Y)* could mean fixing one triple, learning an alias, adding a negative example, revising an ontology relation, changing the extraction prompt, altering retrieval, or updating parameters. These are different learning decisions with different scope and risk.

Following Mitchell (1997), we adopt a performance-based view of learning: a system learns from experience E, with respect to a class of tasks T and performance measure P, when its performance on T, as measured by P, improves with E. For post-deployment language agents, we additionally make persistence explicit: the effect of experience must be retained across the relevant interaction or deployment boundary so that it can influence subsequent behavior. A persistent update alone is therefore not evidence of learning. A KB can accumulate falsehoods, memory can amplify stale patterns, and parameter updates can degrade previously reliable behavior; the learning claim requires evidence of improved subsequent performance.

Never-ending learning further motivates a second question: does accumulated experience improve not only current performance, but also the system's ability to learn subsequent tasks? We therefore consider adaptation speed, evidence efficiency, transfer, preservation, and recovery in addition to endpoint performance.

3 Persistent State and Learning Substrates

Language agents provide more learning substrates than parameter-centric formulations suggest. A_1 makes these alternatives visible.

Memory and structured knowledge. The agent can retain individual corrections and failure cases, consolidate repeated evidence into concepts or rules, and update document indices, entity catalogs, ontologies, and KB facts. Long-term memory work illustrates the general problem of what should be remembered, consolidated, forgotten, or scoped (Maharana et al., 2024; Salama et al., 2025); RAG illustrates the separation between model parameters and external knowledge access (Lewis et al., 2020).

Prompts, examples, policies, and tools. A validated correction can become a few-shot example; an ontology constraint can become a validation rule; repeated ambiguity can trigger abstention or clarification; and the system can acquire procedures for normalization, entity resolution, contradiction checking, or source verification. Tool-use and reflection methods provide mechanisms that can participate in such loops (Schick et al., 2023; Yao et al., 2023; Madaan et al., 2023; Shinn et al., 2024).

Model parameters. Model training (FST, RL, continuous pre-training, etc.) remains important when repeated evidence suggests a broad pattern that should generalize beyond explicit memory or prompting. But parameter updates are one substrate among many and are often harder to attribute, reverse, and audit than localized changes. The recurring question is: *where should this experience be learned?* We compare substrates by persistence, scope, reversibility, generalization, cost, provenance, and the risk that an error propagates into future decisions.

The tutorial also distinguishes **what** is being learned. For A_1, *knowledge* includes facts, entity identities, taxonomic relations, and temporal scope; *skills* include extraction, normalization, entity resolution, and verification procedures; *policies* determine retrieval, routing, abstention, promotion, and recovery; and *self-monitoring* includes uncertainty, contradiction detection, and failure prediction. A single incident may bear on several targets, but the evidence needed to justify each can differ. A corrected triple may justify a local fact update without justifying a new global extraction rule. This separation helps prevent a common failure mode in self-improving systems: generalizing an update beyond the scope actually supported by the evidence. Special attention should be given to the forgetting ability, that is key for the parametric knowledge, but not necessarily the same in a post-deployment system. (where forgetting can prevent evidence accumulation)

4 Running Example: A_1

A_1 begins with a fixed extractor, relation schema, document retriever, and sparse KB_1. It processes documents and produces records such as *(method,*

evaluated-on, benchmark), *(system, uses, model)*, or *(organization, develops, technology)*. Across the tutorial, A_1 encounters a sequence of incidents. Several documents first use a new term for an existing concept: should the agent learn an alias, create a concept, or leave the ambiguity unresolved? Repeated evidence then suggests a relation absent from the ontology: when is schema expansion justified, and should older extractions be revisited? A curator corrects a mistaken relation: is this an instance repair or evidence for a reusable rule? Later, documents support incompatible triples, raising questions about provenance, source authority, temporal validity, and uncertainty.

The key incident comes when KB_1 feeds accumulated knowledge back into extraction. Validated concepts and relations can improve later recognition, as in closed-loop extraction systems such as DySECT (Amin-Naseri et al., 2026). But the same mechanism can create a self-reinforcing error: an incorrect concept biases future extraction, which then produces apparently supporting evidence for that concept. Persistent state has changed not only predictions but also the evidence the system is likely to encounter next.

Finally, authoritative evidence invalidates an earlier belief or ontology decision. A_1 must revise rather than merely append/forget. Successful learning may therefore require making an old belief inactive while retaining provenance needed for historical questions or diagnosis. But, the old belief should not be forgotten, as it might happen that new future evidence can make it valid again. The same incidents are revisited through the lenses of substrates, evidence, governance, and evaluation so that the tutorial develops one cumulative argument.

5 Evidence and Governed Improvement

Persistent adaptation alone is insufficient. An extraction system that promotes every candidate can learn from copied misinformation, correlated documents, noisy feedback, adversarial content, or self-generated evidence that merely echoes its current beliefs. Post-deployment learning therefore requires an explicit transition from experience to evidence to confidence update to an authorized state change.

For A_1, a candidate triple is not automatically knowledge. Relevant evidence includes the source passage, provenance, temporal scope, relation constraints, corroborating or contradicting sources, and review status/confidence. Several pages quoting one original source do not constitute independent corroboration. A curator may be authoritative about a schema decision but not every factual claim. A model-generated explanation may help locate an error without being sufficient evidence for a persistent update.

This framing separates **autonomous update generation** from **autonomous update deployment**. A_1 may propose a relation type, rule, memory, or training example without being authorized to promote it directly. Validation gates, staged release, provenance, rollback, and runtime constraints support bounded autonomy (Wang et al., 2025; Zhu et al., 2025). Closed-loop feedback makes independence especially important: if a mistaken ontology category influences retrieval, prompting, and judging, agreement among those components may reflect a shared error. The correct learning decision can therefore be to quarantine, defer, or reject an update.

Contradiction is likewise not always an error to eliminate. Two sources may describe different time periods, domains, or definitions. A_1 should first align scope before choosing among them, and it may preserve both records with explicit validity conditions. This makes revision different from indiscriminate forgetting: an obsolete active belief can become inactive for current extraction while remaining available for historical reasoning and audit. Evidence can therefore justify retention, revision, rejection, or deliberate non-update.

6 Evaluation Through Time

Post-deployment learning evaluation focuses on whether an evolving system becomes more capable while preserving capabilities that still matter. Snapshot precision, recall, or F1 are necessary but insufficient.

Update effects: whether one retained change improves behavior relative to a matched starting configuration: does learning an alias improve entity linking on unseen documents beyond the corrected instance? **Update-policy effects**: whether rules for proposing and committing changes produce useful trajectories across many incidents: do high-confidence KB_1 additions improve later extraction without amplifying false relations? **Deployment effects**: whether the evolving system remains beneficial under the changing document stream and task definition.

Metrics therefore span extraction quality (precision, recall, calibration, error type), knowledge-state quality (contradictions, provenance, stale facts, ontology coherence), and learning quality (gain, durability, transfer, forgetting, evidence efficiency, regression, and recovery). A larger KB is not itself evidence of better competence, and success on previously cor-

rected cases is not evidence of generalization.

The preservation target must also be explicit. If a relation type has genuinely become obsolete, changing current behavior is not forgetting; if the task still requires both old and new usages, collapsing them is a regression. Likewise, learning speed and cumulative benefit matter: an update policy that eventually reaches the same quality as a frozen baseline may still be valuable if it needs less supervision on later concepts, while one that repeatedly relearns the same distinction has not demonstrated effective cumulative learning.

Temporal evaluation matters when world knowledge or task definitions change (Tang et al., 2025; Jiang et al., 2026). Evaluation data must also be protected from repeated optimization: if documents used to choose ontology updates are recycled as evidence of success, adaptive overfitting can masquerade as continual improvement. Thus, we emphasize held-out or delayed evaluation channels, regression suites, exposure tracking, and explicit accounting for distributional change. And also, which capabilities improved, which remained reliable, at what cost, over what horizon, and under what comparison.

7 Tutorial Structure and Learning Outcomes

Part I – From train-once/deploy-many to learning after deployment (20 min). System boundaries and the difference between transient adaptation and persistent learning, introduced through A_1's static extraction baseline.

Part II – Foundations for learning under change (25 min). Lifelong, continual, and never-ending learning; transfer, stability, plasticity, forgetting, drift, consolidation, and cumulative acquisition.

Part III – Persistent state and learning substrates (30 min). Memory, retrieval, structured knowledge, ontologies, prompts, tools, policies, evaluators, and parameter updates; A_1 illustrates where a correction or new concept should reside.

Part IV – Experience and evidence (25 min). Candidate extractions, corrections, retrieved and synthetic evidence, contradiction, provenance, dependence, and temporal scope.

Part V – Governed self-improvement (30 min). Update proposals, schema changes, validation, authorization, staged deployment, rollback, selective forgetting, and autonomy.

Part VI – Longitudinal evaluation (30 min). Extraction and knowledge-state quality, regression detection, transfer, forgetting, temporal generalization, evaluation exposure, and trajectory-based comparison.

Part VII – A_1 exercise and open problems (10 min). Participants diagnose an extraction failure, choose among an instance correction, memory update, ontology change, policy revision, or model update, specify admissible evidence and validation, and decide whether the proposal should be promoted. A 10-minute break is included separately.

Participants should leave with: (1) a framework for distinguishing learning from transient adaptation and improvement from mere state growth; (2) a vocabulary for deciding what an agent should learn, where it should reside, and how it should be governed; and (3) an evaluation framework for testing whether an evolving agent genuinely improves over time.

8 Breadth, Ethics, and Materials

The tutorial is not about a single IE architecture. A_1 is a pedagogical thread connecting language agents, retrieval and memory, post-training, never-ending and continual learning, knowledge representation, and evaluation under non-stationarity. The principles apply equally when persistent state is stored in memories, documents, code, tools, policies, or parameters.

Post-deployment learning also creates risks: persistent stores can expose private or proprietary information; feedback loops can amplify bias or misinformation; schema changes can erase useful distinctions; and temporary attacks or extraction errors can become durable state. We emphasize data minimization, provenance, access control, temporal validity, validation, rollback, appropriate human or institutional authority, and monitoring for uneven regressions. The objective is inspectable, empirically evaluated adaptation rather than unconstrained self-modification.

Slides, a reading list, conceptual and evaluation maps, and worked exercises will be made available (github.com/megagonlabs/pdsi-tutorial). The tutorial targets NLP/LLM researchers, practitioners working on adaptive or agentic NLP systems, and graduate students familiar with modern neural NLP and LLMs; prior background in continual or never-ending learning is not required.

9 Conclusion

As language models become persistent agents, learning cannot be identified only with parameter updates. A corrected triple or newly discovered concept matters scientifically only when we can explain what persisted, why the evidence justified the change, how it affected later behavior, and whether benefits survived without unacceptable regressions. The tutorial develops this system-level perspective as a bridge between lifelong, continual, and never-ending learning and the challenge of building language agents that can improve responsibly after deployment.

References

Moin Amin-Naseri, Hannah Kim, and Estevam Hruschka. 2026. A Dynamic Self-Evolving Extraction System. arXiv:2603.06915.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2022. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*, 44(7):3366–3385.

Xunyi Jiang, Dingyi Chang, Julian McAuley, and Xin Xu. 2026. When benchmarks age: Temporal misalignment through large language model factuality evaluation. In *EACL 2026, Volume 2*, pages 500–512.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 33:9459–9474.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Swabha Swayamdipta, Yiming Yang, and Shrimai Prabhumoye. 2023. Self-Refine: Iterative refinement with self-feedback. In *NeurIPS*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *ACL 2024*, pages 13851–13870.

Tom M. Mitchell. 1997. *Machine Learning*. McGraw-Hill.

Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Bishan Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi, Matt Gardner, Bryan Kisiel, et al. 2018. Never-ending learning. *Communications of the ACM*, 61(5):103–115.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. MemInsight: Autonomous memory augmentation for LLM agents. In *EMNLP 2025*, pages 33124–33140.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *NeurIPS*, 36:68539–68551.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuan-Jing Huang, Yu-Gang Jiang, et al. 2025. EvoWiki: Evaluating LLMs on evolving knowledge. In *ACL 2025*, pages 948–964.

Sebastian Thrun. 1998. Lifelong learning algorithms. In *Learning to Learn*, pages 181–209. Springer.

Haoyu Wang, Christopher M. Poskitt, and Jun Sun. 2025. AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents. arXiv:2503.18666.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *ICLR*.

He Zhu, Tianrui Qin, King Zhu, Heyuan Huang, Yeyi Guan, Jinxiang Xia, Hanhao Li, Yi Yao, Ningning Wang, Pai Liu, et al. 2025. OAgents: An empirical study of building effective agents. In *Findings of EMNLP 2025*, pages 13354–13369.

From Retrieval-Augmented Generation (RAG) to Agentic Deep Research

Xin Luna Dong, Sanat Sharma, Kai Sun, Jiaqi Wang, Yinglong Xia, Xiao Yang, Scott Wen-tau Yih
Meta, USA
{lunadong, sanatsharma, sunkaicn, jqwang, yxia, xiaoyangfb, scottyih}@meta.com

Nan Tang
HKUST (Guangzhou), Guangzhou, China
nantang@hkust-gz.edu.cn

Franklin Zhang
UW CSE, Seattle, USA
frankyz@uw.edu

Abstract

Despite well-known hallucination issues, LLMs have become an increasingly indispensable source of information, and the underlying technology has advanced rapidly to make their answers far more reliable. Early on, *Retrieval-Augmented Generation (RAG)* emerged as the dominant remedy, grounding LLM responses in external knowledge and evolving from simple retrieve-then-read pipelines into modular, graph-enhanced, and agentic systems. More recently, *Agentic Deep Research* has pushed the frontier further, equipping LLMs with autonomous planning, multi-hop investigation, and iterative synthesis to tackle open-ended questions that no single retrieval pass can answer. This tutorial offers an in-depth treatment of modern RAG and Deep Research, grounded in an AI-assisted systematic analysis of 2,000+ recent papers (2020–2026). Attendees will leave with a structured roadmap, evidence-backed practical recommendations, and a clear map of open research opportunities.

1 Motivation and Scope

The way people seek information is being rebuilt around LLMs: users now turn to AI assistants for everything from quick lookups to in-depth investigations. Yet LLMs remain prone to hallucinations and stale knowledge (Tonmoy et al., 2024; Yang et al., 2024a), making trustworthy, current, well-grounded answers one of the field's central engineering challenges.

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by retrieving relevant external knowledge at inference time rather than relying solely on parametric memory (Huang et al., 2023; Ji et al., 2023). RAG reduces hallucinations with grounded generation, and offloads long-tail factual knowledge to external corpora, enabling smaller, more efficient models without sacrificing coverage (Gao et al., 2024). However, RAG itself introduces new challenges: noisy retrieval, knowledge conflicts between parametric and retrieved information, and wasted computation from unnecessary retrieval. The past three years have witnessed a burst of research aimed at improving RAG's factuality, robustness, generalizability, and latency.

The next leap is already underway. Products such as *Gemini Deep Research*, *OpenAI Deep Research*, and *Claude Research* go beyond single-pass retrieval, equipping LLMs with *Agentic Deep Research* capabilities—planning multi-step investigations, issuing iterative searches and tool calls, and synthesizing comprehensive reports. Deep Research opens the door to open-ended queries while raising new challenges in planning, long-horizon credit assignment, faithful synthesis, and exploration cost.

This tutorial offers a comprehensive treatment of both. For RAG, we organize the landscape into three paradigms—*Modular Pipeline*, *Graph-Enhanced*, and *Agentic* RAG. For Deep Research, we cover system architectures, planning and tool-use strategies, and evaluation. We also situate both within *context engineering*, *agentic systems*, and *long-term memory*, and discuss multi-modal extensions across text, images, and structured data.

This tutorial is itself an instance of AI-facilitated Deep Research in action. The authors first read ∼150 classical papers to develop an initial taxonomy, then employ a research agent to analyze ∼2,000 additional papers on Factuality, RAG, and Agent (2020–2026) to generate comprehensive summaries[1]. The presenters curate and enrich this synthesis with first-hand practical insights that may not be documented elsewhere. New papers will continue to be incorporated before and after the tutorial to support ongoing learning. This combination of AI-driven synthesis with expert curation in

[1] papers.lunadong.com/area/factuality; papers.lunadong.com/area/rag; papers.lunadong.com/area/agent.

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 32–43
November 6-10, 2026 ©2026 Association for Computational Linguistics

itself is a contribution, setting a replicable model for future tutorials and surveys.

1.1 Tutorial Type

This is a **cutting-edge** tutorial. It synthesizes the rapid advances in RAG from 2020–2026, with particular emphasis on emerging paradigms (graph-enhanced and agentic RAG) and techniques (RL-driven retrieval policies, deep research agents) from 2025–2026 that have not yet been covered in prior tutorials. While we provide sufficient background for newcomers with basic LLM familiarity, the primary focus is on presenting the latest research frontier and open problems.

1.2 Target Audience and Prerequisites

- **Researchers** in NLP, data mining, and knowledge management seeking a structured map of the RAG landscape with clear identification of open problems.
- **Applied scientists and engineers** building production LLM systems that need evidence-backed guidelines for choosing among RAG architectures and retrieval strategies.
- **Graduate students** entering the field who wish for a comprehensive entry point to 2K+ papers organized by topic, paradigm, and timeline.

Prerequisites: Familiarity with basic LLM concepts (pre-training, fine-tuning, prompting). No prior knowledge of RAG or hallucination detection is assumed.

2 Related Work

No prior tutorial covers the complete 2023–2026 RAG landscape. Ours introduces a three-paradigm framework (Modular, Graph-Enhanced, Agentic), incorporates 18+ months of advances beyond the most recent tutorial—including RL-driven retrieval policies, GNN-LLM integration, multi-hop reasoning, and autonomous agentic systems—and provides evidence-backed deployment recommendations synthesized from ∼2000 papers.

3 Tentative Schedule (3 Hours)

3.1 Introduction & Factuality Overview (30 min)

We open by demonstrating failure modes of standalone LLMs on knowledge-intensive tasks: outdated facts, entity confusion, confident fabrication, and multi-hop reasoning breakdowns. We provide a high-level overview of the factuality landscape, preview the two-pronged solutions as two sides of the coin: *knowledge internalization and hallucination suppression*—improving what models know and how reliably they express it, and *retrieval-augmented generation*—grounding in external evidence. We show RAG's evolution timeline from simple retrieve-and-read pipelines (2023) through adaptive and graph-enhanced architectures (2024) to autonomous RL-driven agentic systems (2025–2026), based on our systematic analysis of ∼2000 papers. The remainder of the tutorial deep-dives into RAG.

3.2 RAG Deep Dive (60 min)

The next part of the tutorial provides a comprehensive survey of RAG, across *Modular Pipeline RAG* and *Graph-Enhanced RAG*.

3.2.1 Modularized RAG Pipeline (40 min)

We introduce the five-stage modular architecture—triggering, query rewriting, retrieval, post-processing, and answer generation—as independently optimizable components (Gao et al., 2024; Lewis et al., 2020). We trace the evolution from monolithic retrieve-and-read (2023) to modular, swappable pipelines (2024—2025) and show how each stage can be upgraded independently.

1. *RAG Triggering*—When *not* to retrieve: adaptive triggering is the first and most impactful efficiency decision. We analyze the cost—quality trade-off and provide guidelines for choosing a triggering strategy based on deployment constraints.

2. *Query Rewriting*—Transforming user queries for effective retrieval. We describe various strategies and identify which strategies work best for single-hop vs. multi-hop questions.

3. *Retrieval*—The retrieval stage itself: We focus on the surprising finding that retrieval relevance correlates weakly with downstream generation quality—what matters is *informativeness*, not *similarity*.

4. *Post-processing & Context Compression*—What to do with retrieved passages before feeding them to the generator: We describe various methods for reranking, context compression, and position bias mitigation.

5. *Answer Generation & Knowledge Conflicts.*—The final generation stage: We focus on

noise-resilient generation that maintains quality even with imperfect retrieval, and the critical problem of knowledge conflicts:

There are a lot of studies on each component, and we focus on explaining how the most prominent problem is solved. Additionally, we highlight how each component can be extended to answer complex questions, where assembling and reasoning is essential.

3.2.2 Graph-Enhanced RAG (20 min)

Graph-enhanced RAG pipelines construct *Knowledge Graphs (KGs)* from corpora and leverage graph structure to enable structured reasoning. We explain how KG construction and hybrid retrieval combine structured and unstructured knowledge for RAG, how we integrate Graph Neural Networks (GNN) with LLM reasoning, and how we tackle multi-hop questions that require chaining evidence across graph nodes for Complex QA.

3.3 Agentic Deep Research Deep Dive (60 min)

Building on the traditional RAG foundations, we then examine how Agentic Deep Research extends retrieval into autonomous, multi-step investigation.

3.3.1 Agentic RAG (30 min)

Static single-pass retrieval fails on complex multi-hop questions, wastes compute on simple queries, and cannot adapt to varying information needs during generation. Recent Agentic AI trend leads to the Agentic RAG pipelines that iteratively decide whether to retrieve, what to retrieve, and how to integrate external knowledge during language model decoding, treating RAG as a dynamic decision-making process. We describe the paradigm shift from hand-crafted pipelines to learned retrieval policies, and how to scale agentic RAG to reasoning-heavy queries.

3.3.2 Deep Research (30 min)

While agentic RAG enables dynamic retrieval decisions during generation, Deep Research extends this paradigm to autonomous, long-horizon research workflows. Motivated by complex information needs (e.g., literature surveys, competitive analyses) that exceed single-query RAG, deep research agents operate over minutes to hours, iteratively browsing the web, reasoning over gathered evidence, and producing structured long-form reports grounded in cited sources. We describe key techniques at each stage of the deep research pipeline (intent clarification and planning, web exploration, reasoning and synthesis, and report generation) and examine how modern systems execute this pipeline.

3.4 Synthesis & Future Directions (30 min)

We conclude the tutorial with **key takeaways** organized by practitioner profile (researcher, engineer, student), with evidence-backed guidelines for choosing the right RAG architecture. We review the **emerging trends** that shape the future of RAG: RL transforming RAG into autonomous agentic systems with process-level supervision; multimodal RAG extending retrieval and generation to images, video, and tables; the convergence of modular, graph-based, and agentic paradigms into hybrid architectures; domain-specific RAG deployment in healthcare, finance, and legal domains with specialized benchmarks; and long-context trade-offs as context windows scale to millions of tokens. Finally, we discuss **open research opportunities** to inspire more research in this field: unified retrieval-generation architectures, adversarially robust retrieval, long-form and multi-turn RAG for dialogue systems, scalable graph construction for knowledge-intensive domains, compute-optimal inference balancing retrieval cost vs. quality, and efficient agentic RAG for latency-sensitive production deployments.

4 Audience Engagement Strategies

We employ three strategies to encourage participation throughout:

Running example: A single compound question evolves across all sessions, with each technique showing concrete improvement over the previous solution, providing a narrative thread that builds intuition incrementally.

Live polls: At key decision points, attendees predict outcomes before we reveal benchmark results, encouraging active reasoning about trade-offs.

Online companion: The interactive survey we maintain over time will let attendees explore papers, methods, timelines, and benchmarks during and after the tutorial, extending engagement beyond the session.

5 Estimated Audience Size

We estimate an audience of **200–300 attendees**. RAG has become one of the most actively re-

searched areas in NLP, and interest has grown steadily over the past three years. Related tutorials have drawn increasing audiences:

- *Retrieval-Augmented Text Generation*, IJCAI 2022 (∼50 attendees).
- *Retrieval-based Language Models and Applications*, ACL 2023 (∼150 attendees).
- *RAG Meets LLMs*, KDD 2024 (∼100 attendees).
- *RAG for AI-Generated Content*, ACL 2024 (∼200 attendees).

The tutorial team also organized the KDD Cup 2024 (CRAG (Yang et al., 2024a)) and KDD Cup 2025 (CRAG-MM (Wang et al., 2025a)) competitions on RAG, which collectively attracted ∼3K registered participants across ∼600 teams, further demonstrating strong community interest. Our tutorial offers the most comprehensive coverage to date—spanning modular, graph-enhanced, and agentic paradigms with 18+ months of advances beyond the most recent tutorial—and we expect continued growth in audience size.

6 Reading List

The following papers provide essential background and representative coverage across the three RAG paradigms. Asterisked entries (*) are recommended pre-reading.

* Lewis et al. (2020). The original paper introducing RAG for knowledge-intensive NLP tasks.
* Gao et al. (2024). A comprehensive survey of RAG for LLMs, covering architecture taxonomy and evaluation.
* Huang et al. (2023). A survey on hallucination in LLMs, motivating the need for RAG.
- Guu et al. (2020). REALM: retrieval-augmented language model pre-training, an early foundational approach.
- Asai et al. (2023). Self-RAG: learning to retrieve, generate, and critique through self-reflection.
- Jiang et al. (2023b). FLARE: active retrieval augmented generation with adaptive triggering.
- Edge et al. (2024). GraphRAG: a graph-based approach to query-focused summarization.
- Jin et al. (2025). Search-R1: training LLMs to reason and leverage search engines with reinforcement learning.
- Yang et al. (2024a). CRAG: a comprehensive benchmark for end-to-end RAG systems.
- Ni et al. (2025). A recent survey on trustworthy RAG, covering robustness and reliability.

A complete reading list of ∼2000 papers, organized by topic and continuously updated, is available at the tutorial's online companion[1].

7 Societal Impact

Positive impacts: More reliable LLM systems in high-stakes domains; better-informed deployment decisions grounded in evidence from ∼2000 papers and the tutors' real practice; acceleration of research toward robust retrieval-augmented architectures; democratized access through openly available interactive survey materials.

Potential concerns: Improved RAG performance could create false confidence if systems are deployed without rigorous evaluation—we explicitly discuss current limitations. Adversarial robustness remains an open challenge, and no existing defense is universally robust. We emphasize that current RAG techniques significantly reduce but do not eliminate hallucination risk.

8 Ethics Statement

This tutorial covers RAG techniques whose primary goal is to *reduce* hallucinations and improve LLM factual reliability–an inherently ethics-positive objective. Still, several considerations arise. First, RAG reduces but does not eliminate hallucination risk, and improved performance may encourage deployment with insufficient human oversight; we address current limitations throughout and caution against treating RAG as a complete factuality solution. Second, agentic RAG and deep research systems that autonomously browse and synthesize web content pose dual-use concerns: the same capabilities enabling trustworthy information gathering could produce more convincing source-grounded disinformation. We discuss adversarial robustness and responsible deployment as open challenges. Third, retrieval over personal or sensitive data–especially for the wearable device applications covered–raises privacy concerns requiring appropriate access controls and data governance. Finally, this tutorial uses AI-assisted analysis of ∼2000 papers; all AI-generated summaries are curated and verified by the presenters for accuracy and proper attribution.

9 Limitations

This tutorial has no venue-specific limitations: the topic is broadly relevant, not region-specific, and requires only standard presentation equipment.

10 Tutors

We have a strong set of tutors, including early inventors of RAG, active researchers, industry practitioners (with extra insights from wearables devices), and strong system builders. The same team also hosted KDDCups on RAG (CRAG (Yang et al., 2024a), CRAG-MM (Wang et al., 2025a)) in the past two years.

10.1 In-Person Presenters

Xin Luna Dong (Email: lunadong@meta.com. Phone: 1-201-650-3494) **(Corresponding Tutor)** is a Principal Scientist at Meta Wearables AI, where she leads the RAG and Agentic AI efforts for building trustworthy and personalized assistants on wearable devices. Previously, she spent over a decade advancing knowledge graph technology, including the Amazon Product Graph and the Google Knowledge Graph. She is co-author of Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases and Big Data Integration. She was named an ACM Fellow and an IEEE Fellow for "significant contributions to knowledge graph construction and data integration", awarded the VLDB Women in Database Research Award and VLDB Early Career Research Contribution Award, and invited as an ACM Distinguished Speaker. She has presented many tutorials in the past at KDD, WebConf, WSDM, ACL, VLDB, SIGMOD, etc. She serves in the PVLDB advisory committee, was a member of the VLDB endowment, a PC co-chair for KDD'2022 ADS track, WSDM'2022, VLDB'2021, and Sigmod'2018.

Sanat Sharma is a Research Engineer at Meta Reality Labs, working on LLMs, RAG, post training, and NLP. Sanat has contributed to multiple conference papers in venues like ICLR, CVPR, SIGIR, AAAI, ICDM, ECIR. Sanat was an organizer for the KDD 2025 Cup.

Jiaqi Wang is a Machine Learning Engineer at Meta Reality Labs. Her recent work focuses on RAG and Multimodal LLMs. She built the CRAG-MM benchmark and organized the associated KDD Cup competition and workshop. Prior to Meta, she developed personalized ranking models for e-commerce recommendation systems.

Yinglong Xia is an Applied Research Scientist at Meta Recommendation Systems (MRS), where he focuses on large-scale recommendation systems with knowledge-driven AI technology for personalization. Prior to that, he was a Chief Architect on Enterprise AI at Futurewei / Huawei US, and a Research Staff Member and a Postdoc at IBM T.J. Watson Research Center. He published 100+ technical papers and 40+ patents. He is active in professional activities, such as serving as an AE for IEEE TBD/TKDE, an ADS area chair at KDD 2025 and 2026, an industry co-chair at CIKM 2024 and WSDM 2026, etc.

Xiao Yang is a Research Scientist at Meta Reality Labs. Her recent research focuses on RAG, factuality, and agentic AI. Prior to Meta, Xiao focused on Natural Language Understanding at Siri, Apple Inc. Xiao published multiple papers in major AI venues like NeurIPS, EMNLP, NAACL. She also served as the organizer of 2024 and 2025 KDD Cups.

10.2 Contributors

Kai Sun is a Research Scientist at Meta Reality Labs. His recent research focuses on RAG, LLM factuality, and knowledge retrieval, particularly for applications on wearable devices. He served as an organizer for the 2024 and 2025 KDD Cups, both focused on RAG, and has been an area chair for major NLP venues since 2022.

Scott Wen-tau Yih is a Research Scientist at Meta Fundamental AI Research (FAIR) and an affiliate professor at University of Washington. Prior to this, he held notable positions as Principal Research Scientist at the Allen Institute for Artificial Intelligence (2017-2019) and Senior Researcher at Microsoft Research (2005-2017), after receiving his PhD from the University of Illinois at Urbana-Champaign in 2005. Dr. Yih has published numerous influential papers, including WikiQA, DPR, and RAG, which have garnered significant recognition. He received the best paper award from CoNLL'11 and an outstanding paper award from ACL'15. Additionally, he has served in various leadership roles within the NLP and ML communities, including program co-chairs (CEAS'09, CoNLL'14, EMNLP'21) and senior area chairs for NLP (ACL, NAACL, EMNLP, EACL) and ML (ICLR, NeurIPS) conferences. In 2024, Dr. Yih was selected as an ACL Fellow for "significant contributions to information extraction and question answering, neural retrieval and retrieval-augmented generation."

Franklin Zhang is a student at UW CSE, responsible for the infrastructure and maintenance of the Online AI Companion, and actively shapes the tutorial content with fresh learning perspectives.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, Andr´as Gy¨orgy, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesv´ari, Ali Taylan Cemgil, and Nenad Tomasev. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint*.

Uri Alon, Frank F. Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval. In *ICML*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint*.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke S. Zettlemoyer, Hanna Hajishirzi, and Wen-tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*.

Jiaxin Bai and Yangqiu Song. 2025. Autoschemakg: Autonomous kg construction through dynamic schema induction from web-scae corpora. *arXiv preprint*.

Vincent-Pierre Berges, BArlas Oguz, Daniel Haziza, Wen-tau Yih, Luke Zettlemoyer, and Gargi Ghosh. 2024. Memory layers at scale. *arXiv preprint*.

Sebastian Borgeaud. 2022. Retro: Improving language models by retrieving from trillions of tokens. *arXiv preprint*.

Sindhuja Chaduvula, Ahmed Y. Radwan, Azib Farooq, Yani Ioannou, and Shaina Raza. 2026. Reducing hallucinations in llms via factuality-aware preference learning. *arXiv preprint*.

Manish Chandra, Debasis Ganguly, and I. Ounis. 2026. Lure-rag: Lightweight utility-driven reranking for efficient rag. *arXiv preprint arXiv:2601.19535*.

Haotian Chen, Qingqing Long, Siyu Pu, Xiao Luo, Wei Ju, Meng Xiao, Yuanchun Zhou, Jianghua Zhao, and Xuezhi Wang. 2026. Deepera: A deep evidence reranking agent for scientific rag qa. *arXiv preprint arXiv:2601.16478*.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2025a. Coke: Teaching large language models to express knowledge boundary from their own signals. In *KnowFM workshop*.

Mingda Chen, Yang LI, Karthik Padthe, Rulin Shao, Alicia Sun, Luke Zettlemoyer, Gargi Ghosh, and Wen-tau Yih. 2025b. Ewe: Improving factuality w. explicit working memory. In *ACL*.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025c. Research: Learning to reason with seach for llms via rl. In *NeurIPS*.

Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. 2025d. Co-marl: Improving retrieval-augmented generation through multi-agent reinforcement learning. In *NeurIPS*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? In *ICML*.

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. *arXiv preprint*.

Daniel Christoph, Max Ploner, Patrick Haller, and Alan Akbik. 2025. Learning to study for better factual recall. *arXiv preprint*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in llms. In *ICLR*.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I don't know: Explicit modeling of uncertainty with an [idk] token. In *NeurIPS*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Cove: Chain-of-verification reduces hallucination in llms. *arXiv preprint*.

Hanxing Ding, Shuchang Tao, Liang Pang, Zihao Wei, Liwei Chen, Kun Xu, Huawei Shen, and Xueqi Cheng. 2025. On the diminishing returns of complex robust rag training in the era of powerful llms. *SigIR-AP*.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of llms: A survey. *arXiv preprint*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form llms. In *ACL*.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *NAACL*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Seven Truitt, and Jonothan Larson. 2024. From local to global: A

graph rag approach to query-focused summarization. *arXiv preprint*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak llms. *arXiv preprint*.

Wenqi Fan, Yujuang Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards ra llm. *arXiv preprint*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024a. Maat: Enhancing noise robustness of retrieval-augmented lms with adaptive adversarial training. In *Annual Meeting of the Association for Computational Linguistics*.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2024b. Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. *arXiv preprint*.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2025. Kirag: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. *arXiv preprint*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. In *ACL*.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for rag systems. *arXiv preprint*.

L. Gao, X. Ma, J. Lin, Callan, and J. 2023. Precise zero-shot dense retrieval without relevance labels. In *In ACL*.

Y. Gao and Y. Xiong. 2024. Retrieval-augmented generation for ai-generated content: A survey and tutorial. In *Tutorial at ACL, Bangkok, Thailand*.

Y. Gao, Y. Xiong, and X. Gao. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *EMNLP*.

Yuntao Gui and James Cheng. 2025. Search-r3: Unifying reasoning and embedding generation in llms. *arXiv preprint*.

Aman Gupta and Vinamra Benara. 2024. Rag vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint*.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of rag: Evolution, current landscape, and future directions. *arXiv preprint*.

Bernal Jimenez Guti´errez´, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for llms. In *NeurIPS*.

Bernal Jimenez Guti´errez´, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. Hipporag2: From rag to memory: Non-parametric continual learning for llms. *arXiv preprint*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint*.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *EMNLP*.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2024. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In *North American Chapter of the Association for Computational Linguistics*.

Yucheng Hu and Yuxing Lu. 2024. Rag and rau: A survey on ra lm in natural language processing. *arXiv preprint*.

L. Huang, W. Yu, and W. Ma. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Gautier Izacard and Edouard Grave. 2021. Fid: Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt ra llm through question complexity. In *North American Chapter of the Association for Computational Linguistics*.

Z. Ji, N. Lee, and R. Frieske. 2023. Survey of hallucination in natural language generation. In *ACM Computing Surveys, 55(12):1–38*.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, Han, and Jiawei. 2025. Deepretrieval: Hacking real search engines and retrievers w. llms via rl. In *COLM*.

Z. Jiang, F. Xu, and L. Gao. 2023a. Active retrieval augmented generation. In *In EMNLP*.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Flare: Active retrieval augmented generation. In *EMNLP*.

Bowen Jin, hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arik, Dong Wang, Hamed Azmani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engies with rl. In *COLM*.

Saurav Kadavath, Tom Conerly, and ... Jared Kaplan. 2022. Language models (mostly) know what they know. *arXiv preprint*.

Nikhil Kandpal, haikang Deng, Adam Orberts, Eric Wallace, and Colin Raffel. 2023. Llms struggle to learn long-tail knowledge. In *ICML*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah GOldblum, and Andrew Gordon Wilson. 2024. Llms must be taught to know what they don't know. In *NeurIPS*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. knn-lm: Generalization through memorization: Nearest neighbor language models. In *ICLR*.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. 2025. Knowledge entropy decay during lm pre-training hinders new knowledge acqusition. In *ICLR*.

Abdullatif Koksal, Renat Aksitov, and Chung-Ching Chang. 2023. Har: Hallucination augmented recitations for language models. *arXiv preprint*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in nlg. In *ICLR*.

Michael Lan, Philip Torr, Austin Meek, David Krueger, and Fazl Barez. 2024. Sparse autoencoders reveal universal feature spaces for hallucination detection. *arXiv preprint*.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. Conli: Chain of natural language inference for reducing llm ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*.

Yongqi Leng, Yikun Lei, Xikai Liu, Meizhi Zhong, Bojian Xiong, Yurong Zhang, Yan Gao, Yao Hu, and Deyi Xiong. 2025. Decision-execution separation for retrieval-augmented generation as mdp. *arXiv preprint*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Rag: Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeuIPS*.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. 2025a. Entropy-trend constraint for proactive retrieval triggering. *arXiv preprint*.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. 2025b. Etc: Modeling uncertainty trends for timely retrieval in dynamic rag. *arXiv preprint*.

Jiaqi Li, Yixuan Tang, and Yi Yang. 2025c. Ustuning: Know the unknown: An uncertainty-sensitive method for llm instruction tuning. In *ACL*.

Kenneth Li, Oam Patel, Fernanda Viegas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention (iti): Eliciting truthful answers from a lm. In *NeurIPS*.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025d. [c] search-o1: Agentical search-enhanced large reasoning models. *arXiv preprint*.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2024a. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. In *International Conference on Learning Representations*.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiouzhu Mei, and Michael Bendersky. 2024b. Self-route: Rag or long-context llms? a comprehensive study and hybrid approach. In *EMNLP*.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for llms. *arXiv preprint*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint*.

Xiaoqiang Lin, Aritra Ghosh, K. H. Low, Anshumali Shrivastava, and Vijai Mohan. 2025. Refrag: Rethinking rag based decoding. *arXiv preprint arXiv:2509.01092*.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. 2024a. Think-on-graph 2.0: Deep and faithful llm reasoning w. knowledge-guided rag. In *International Conference on Learning Representations*.

Zexiong Ma, Shengnan An, Zeqi Lin, Yanzhen Zou, Jian-Guang Lou, and Bing Xie. 2024b. Depac: Enhancing retrieval-augmented generation with dehallucinating parallel context extension. In *Long-Context Fnd workshop*.

E. Magesh and D. Kvasyuk. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Stanford HAI Working Paper*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust lms: Investigating effectiveness of parametric and non-parametric memories. In *ACL*.

K. Meng, D. Bau, A. Andonian, Belinkov, and Y. 2022. Locating and editing factual associations in gpt. In *In NeurIPS*.

Dezhuang Miao, Yibin Du, Xiang Li, Xiaoming Zhang, Jiahe Li, Bo Zhang, Bingyu Yan, Lian Zhang, and Litian Zhang. 2025. R-cot: Reverse cot and causal path verification: A modular plugin for aligning llms with kgs. In *CIKM*.

Sabrina J. Mielke, Anthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *arXiv preprint*.

Dehai Min, Kailin Zhang, Tongtong Wu, and Lu Cheng. 2025. Quantifying uncertainty from the pre-training corpus for retrieval-augmented generation. *arXiv preprint*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *COLM*.

N. Muennighoff, H. Su, and L. Wang. 2024. Generative representational instruction tuning. In *In ICLR*.

D. Muhlgay, O. Ram, and I. Magar. 2024. Generating benchmarks for factuality evaluation of language models. In *In EACL*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint*.

Bo Ni, Zheyuan Liu, Leyao Wang, ... Luna Dong, Yinglong Xia, ... Wenqi Fan, Erik Blasch, Yu Wang, Meng Jiang, and Tyler Derr. 2025. Towards trustworthy rag for llms: A survey. *arXiv preprint*.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When do llms need ra? mitigating llms' overconfidence helps ra. In *ACL*.

Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint*.

Xu Pan, Ely Hahami, Zechen Zhang, and Haim Sompolinsky. 2025. Mega: Memorization and knowledge injection in gated llms. *arXiv preprint*.

Core Francisco Park, Zechen Zhang, and Hidenori Tanaka. 2025. Sys-2-ft: New news: System-2 fine-tuning for robust integration of new knowledge. *arXiv preprint*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving llms with external knowledge and automated feedback. *arXiv preprint*.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph rag: A survey. *arXiv preprint*.

F. Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktaschel, and Sebastian Riedel. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *NAACL*.

Yifu Qiu, Varun Embar, Shay B. Cohen, and Benjamin Han. 2023a. Tweak: Think while you write: Hypothesis verfication promotes faithful knowledge-to-text generation. *arXiv preprint*.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023b. Detecting and mitigating hallucinations in multilingual summarization. *arXiv preprint*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context ralm: In-context retrieval-augmented language models. *TACL*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M. Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in llms – an extensive definition, quantification, and prescriptive remediations. In *EMNLP*.

Evgeniia Razumovskaia, Ivan Vulic, Pavle Markovic, Tomasz Cichy, Qian Zheng, Tsung-Hsien Wen, Budzianowski, and Pawel. 2023. Dial beinfo for faithfulness: Improving factuality of information-seeking dialogue via behavioural fine-tuning. *arXiv preprint*.

Revanth Gangi Reddy and Chengxiang Zhai. 2024. Charmbana: Progressive responses with real-time internet searchfor knowledge-powered conversations. In *WSDM*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. Investigating the factual knowledge boundary of llms with retrieval augmentation. In *Coling*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? by roberts et al. In *EMNLP*.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint*.

H. Shi and X. Qiu. 2024. Adaptive context-aware decoding for faithful rag. *arXiv preprint*.

Weijia Shi, Xiaochuang Han, M. Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and S. Yih. 2023a. Why lms hallucinate? In *North American Chapter of the Association for Computational Linguistics*.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries. In *ICLR*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint*.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Seeker: Language models that seek for knowledge: Modular search generation for dialogue and prompt completion. In *EMNLP*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *EMNLP*.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *arXiv preprint*.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic rag based on the information needs of llms. In *ACL*.

J. Sun, C. Xu, and L. Tang. 2024a. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *In ICLR*.

Jiashuo Sun, Chengjin Xun, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024b. Think-on-graph: Deep and responsible reasoning of llm on kg. In *ICLR*.

Kai Sun, Y. Xu, Hanwen Zha, Yue Liu, and Xinhsuai Dong. 2023a. Head-to-tail: How knowledgeable are large language models? a.k.a. will llms replace knowledge graphs? In *North American Chapter of the Association for Computational Linguistics*.

Yushi Sun, Kai Sun, Yifan Ethan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2025. Kerag: Knowledge-enhanced retrieval-augmented generation for advanced question answering. In *EMNLP*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Dnny Zhou. 2023b. Recite: Recitation-augmented language models. In *ICLR*.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023a. Fine-tuning language models for factuality. In *NeurIPS workshop*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023b. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint*.

Zhiliang Tian, Wei Bi, Xiaopeng Li, and N. Zhang. 2019. Learning to abstract for memory-augmented conversational response generation. In *Annual Meeting of the Association for Computational Linguistics*.

S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinjia Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. "a comprehensive survey of hallucintion mitigation techniques in llms". *arXiv preprint*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Ircot: Iterleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint*.

Lynn Vonderhaar, Daniel A. Machado, and Omar Ochoa. 2025. Rag for llms: A survey. In *International Conference on Artificial Intelligence Testing*.

Tu Vu, Mohit Ivyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, quoc Le, and Thang Luong. 2023. Freshllms: Refreshing llms with search engine augmentation. *arXiv preprint*.

Jiaqi Wang, Xiao Yang, Kai Sun, Parth Suresh, Sanat Sharma, Adam Czyzewski, Derek Andersen, Surya Appini, Arkav Banerjee, Sajal Choudhary, Shervin Ghasemlou, Ziqiang Guan, Akil Iyer, Haidar Khan, Lingkun Kong, Roy Luo, Tiffany Ma, Zhen Qiao,

David Tran, and 22 others. 2025a. Crag-mm: Multi-modal multi-turn comprehensive rag benchmark. In *arXiv*.

Ke Wang, Yiming Qin, Nikolaos Dimitriadis, Alessandro Favero, and Pascal Frossard. 2025b. Memoir: Lifelong model editing with minimal overwrite and informed retention for llms. *arXiv preprint arXiv:2506.07899*.

X. Wang, J. Wei, and D. Schuurmans. 2023a. Self-consistency improves chain of thought reasoning in language models. In *In ICLR*.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang LI, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian LI, Bing Yin, JIngbo Shang, and Julian McAuley. 2024. Memoryllm: Towards self-updatable llms. In *ICML*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Filco: Learning to filter context for rag. *arXiv preprint*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Simpleqa: Measuring short-form factuality in llms. *arXiv preprint*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024b. Safe: Long-form factuality in llms. *arXiv preprint*.

Zhepei Wei, Wei-lin Chen, and Yu Meng. 2025. Instructrag: Instructing retrieval augmented generation via self-synthesized rationals. In *ICLR*.

Di Wu, Jia-Chen Gu, Kai-Wei Chang, and Nanyun Peng. 2025. Self-routing rag: Binding selective retrieval with knowledge verbalization. *arXiv preprint arXiv:2504.01018*.

Yiqing Xie, Wenxuan Zhou, Pradyot Prakash, Di Jin, Yuning Mao, Quintin Fettes, Arya Talebzadeh, Sinong Wang, Han Fang, Carolyn Rosé, Daniel Fried, and Hejia Zhang. 2025. Fence: Improving model factuality w. fine-grained critic-based evaluator. In *ACL*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*.

Hao-Xiang Xu, Jun-Yu Ma, Ziqi Peng, Yuhao Sun, Zhen-Hua Ling, and Jia-Chen Gu. 2026. Multiplicative orthogonal sequential editing for scalable knowledge updates. *arXiv preprint*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Crag: Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024a. Crag–comprehensive rag benchmark. In *NeurIPS*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. Alignment for honesty. In *NeurIPS*.

S. Yao, D. Yu, and J. Zhao. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *In NeurIPS*.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in llm geneations w. local intrinsic dimension. *arXiv preprint*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024a. Evaluation of rag: A survey. *arXiv preprint*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024b. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *EMNLP*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *arXiv preprint*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng JI, and Tong Zhang. 2024a. R-tuning: Instructint llms to say "i don't know". In *NAACL*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *EMNLP*.

Jie Zhang, Bo Tang, Wanzi Shao, Wenqiang Wei, Jihao Zhao, Jianqing Zhu, Zhiyu Li, Wen Xi, Zehao Lin, Feiyu Xiong, and Yanchao Tan. 2025. Tadarag: Task adaptive rag via on-the-fly kg construction. *arXiv preprint*.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, M. Zaharia, Ion Stoica, and Joseph Gonzalez. 2024b. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

Penghao Zhao, Hailin Zhang, Jie Jiang, and Bin Cui. 2024. Rag for ai-generated content: A survey. *arXiv preprint*.

Yi Zhao, Yanxiang He, Xiaotong Zhang, and Weidong Wen. 2025. Combining retrieved with generated contexts via a listwise reranker for open-domain question answering. *Neurocomputing*.

H. Zheng, S. Mishra, and X. Chen. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *In ICLR*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Trime: Training with in-batch memory. In *EMNLP*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *EMNLP*.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S. Yu. 2024. Trustworthiness in rag systems: A survey. *arXiv preprint*.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. 2025. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint*.

Yijie Zhu, Haojie Zhou, Wanting Hong, Tailin Liu, and Ning Wang. 2025. Reap: Enhancing rag with recursive evaluation and adaptive planning for multi-hop qa. *arXiv preprint*.

Diffusion Language Models: Foundations and Frontiers of Non-Autoregressive Language Generation

Daisuke Oba
Institute of Science Tokyo, Japan
daisuke.oba@nlp.comp.isct.ac.jp

Danushka Bollegala
University of Liverpool, United Kingdom
danushka@liverpool.ac.uk

Abstract

Diffusion language models (DLMs) have recently emerged as a non-autoregressive approach to language generation, enabling parallel, iterative, and refinement-based decoding beyond the conventional left-to-right paradigm. This tutorial offers the NLP community a technical map of this emerging field: starting from diffusion formulations for text, we clarify their connections to autoregressive language modeling, masked language modeling, and earlier iterative generation methods. We then survey recent progress in large-scale DLMs, efficient inference, training objectives, and post-training, including reinforcement learning for reasoning and alignment. Throughout the tutorial, we treat DLMs not as a solved replacement for autoregressive language models, but as a rapidly developing paradigm whose efficiency claims, controllability, practical assessment, and deployment implications require careful examination. We close by discussing representative open directions, from broader formulations such as flow-based language modeling to reliability, safety, and social bias as key requirements for the social acceptability of DLMs.

1 Motivation and Relevance

Autoregressive language models have become the dominant paradigm for open-ended text generation, driven by Transformer architectures and their successful scaling to modern LLMs (Vaswani et al., 2017; Grattafiori et al., 2024). Yet their left-to-right factorization imposes a fundamental sequential structure: each token is generated after the previous one, even when many parts of an output could in principle be predicted, revised, or refined in parallel. Diffusion language models (DLMs) revisit this limitation in the LLM era by replacing single-pass next-token continuation with iterative denoising and refinement, allowing multiple tokens to be updated within the same generation step. This makes DLMs attractive not only for controllable or infilling-style generation, but also as a possible route toward few-step, parallel, and ultra-fast language generation.

The topic is timely because DLM research has rapidly moved from early text diffusion and discrete generative modeling toward scalable language modeling. Early work introduced continuous diffusion for controllable text generation and discrete diffusion in structured spaces (Li et al., 2022; Austin et al., 2021). More recent work has clarified how diffusion can be applied to discrete text by studying different corruption processes, denoising objectives, and interpretations of masked or discrete diffusion models (Lou et al., 2024; Ou et al., 2024; Shi et al., 2024; Sahoo et al., 2024, 2025). Together, these developments connect DLMs to masked language modeling and iterative refinement while raising broader questions about the right formulation for text generation (Devlin et al., 2019; Ghazvininejad et al., 2019).

Recent scaling results have further raised the stakes. Masked DLMs have been scaled on text, large language diffusion models such as LLaDA and Dream have attracted broad attention, and follow-up work has explored long-context generation, code generation, and hybrid autoregressive–diffusion architectures (Nie et al., 2025a,b; Ye et al., 2025; Liu et al., 2025a; Gong et al., 2025; Arriola et al., 2025; Wang et al., 2025b). In parallel, the literature has rapidly expanded along two practical axes: making DLM generation faster through improved decoding, caching, early stopping, and distillation (Ma et al., 2025; Liu et al., 2025b; Wu et al., 2025; Israel et al., 2025; Wei et al., 2025; Luxembourg et al., 2025; Deschenaux and Gulcehre, 2025; Zheng et al., 2026; Zhang et al., 2026; Oba et al., 2026a), and adapting training or post-training methods to diffusion trajectories (He et al., 2025; Zhao et al., 2025; Tang et al., 2026; Wang et al., 2026; Oba et al., 2026b). This rapid growth makes DLMs an important but difficult area for

Proceedings of the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 15th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 44–50
November 6-10, 2026 ©2026 Association for Computational Linguistics

NLP researchers to enter.

At the same time, the central promise of DLMs is not yet settled. Practical speedups depend on the number of denoising steps, wall-clock latency, memory use, caching, sampling schedules, and quality degradation under few-step generation. These limitations are already motivating broader formulations of non-autoregressive generation, including flow-based language modeling, which suggests that current DLMs should be understood as part of a still-evolving design space rather than as a completed replacement for autoregressive LMs (Lipman et al., 2023; Lee et al., 2026; Hu et al., 2026; Deschenaux and Gulcehre, 2026). Thus, a tutorial is needed not only to explain how DLMs work, but also to clarify what they currently enable, where they remain weak, and what kinds of future formulations may address their limitations.

Finally, if DLMs become practical generation systems, their social acceptability must be studied as seriously as their efficiency. Fairness, safety, and reliability have become central concerns for autoregressive LMs, but they remain largely underexplored for diffusion-based generation. This is concerning because DLMs generate through different mechanisms—masking, denoising, confidence estimation, and iterative refinement—which may create different bias propagation and safety failure modes. This tutorial therefore positions social bias, safety, and reliability as key open problems for the NLP community, alongside technical questions about inference, training, post-training, and evaluation (Kaneko et al., 2022; Kaneko and Bollegala, 2022; Oba et al., 2024).

2 Tutorial Overview and Learning Goals

This tutorial focuses on diffusion language models as an emerging approach to non-autoregressive language generation. The core scope is DLMs themselves: their foundations, scalable model families, inference algorithms, training and post-training methods, and unresolved limitations. We discuss related frontiers such as flow-based language modeling, fairness, and safety as open directions motivated by the current limits of DLMs, rather than as separate primary topics (Lipman et al., 2023; Lee et al., 2026; Hu et al., 2026).

The tutorial has four learning goals. *First*, participants will place DLMs in context: how they depart from left-to-right language models, inherit ideas from masked language modeling and iterative refinement, and differ from earlier non-autoregressive decoding methods (Vaswani et al., 2017; Devlin et al., 2019; Ghazvininejad et al., 2019; Gu et al., 2019). *Second*, they will learn the main design choices behind diffusion over text, including absorbing-mask and uniform noising processes, denoising objectives, sampling procedures, and recent interpretations of discrete diffusion for language generation (Austin et al., 2021; Lou et al., 2024; Ou et al., 2024; Shi et al., 2024; Sahoo et al., 2024, 2025). *Third*, they will survey the current landscape: scaled DLMs, hybrid autoregressive–diffusion architectures, long-context and code models, efficient inference through caching, scheduling, and early stopping, and RL-based post-training for reasoning and alignment (Nie et al., 2025b; Ye et al., 2025; Arriola et al., 2025; Wang et al., 2025b; Liu et al., 2025a; Gong et al., 2025; Ma et al., 2025; Wu et al., 2025; Israel et al., 2025; Oba et al., 2026a; Zhao et al., 2025; Tang et al., 2026; Wang et al., 2026; Oba et al., 2026b). *Fourth*, they will learn to evaluate DLM claims critically, including speed–quality trade-offs, controllability, evaluation protocols, reliability, and broader formulations beyond current diffusion models (Li et al., 2022; Schiff et al., 2024; Zheng et al., 2023; Lipman et al., 2023; Lee et al., 2026; Hu et al., 2026). They will also examine social bias and safety as underexplored requirements for practical DLM deployment, building on prior work on bias evaluation and mitigation in language models (Kaneko and Bollegala, 2022; Kaneko et al., 2022; Oba et al., 2024).

By the end of the tutorial, participants will be able to navigate the DLM literature, assess claims about its advantages and limitations, and identify open problems in diffusion-based text generation.

3 Technical Coverage and Reading List

The citations below indicate the representative reading list for the tutorial. A small subset will be marked as recommended pre-tutorial readings in the final tutorial materials.

Foundations and Formulations. We first situate DLMs within the broader landscape of language generation. We contrast left-to-right autoregressive modeling (Vaswani et al., 2017) with earlier non-autoregressive and iterative refinement methods, such as masked parallel decoding and edit-based generation (Ghazvininejad et al., 2019; Gu et al., 2019). We then introduce diffusion-based generation for text as iterative denoising over discrete

token sequences or latent language representations. This part covers the main design choices behind DLMs: noising processes such as absorbing-mask and uniform corruption, denoising objectives, sampling procedures, and connections to masked language modeling and discrete diffusion theory (Devlin et al., 2019; Ho et al., 2020; Austin et al., 2021; Lou et al., 2024; Ou et al., 2024; Shi et al., 2024; Sahoo et al., 2024, 2025).

Scaled DLMs and Model Families. We next survey how diffusion-based text generation has evolved from early controllable generation and masked diffusion foundations into scalable DLM families (Li et al., 2022; Sahoo et al., 2024; Nie et al., 2025a). We cover large language diffusion models such as LLaDA (Nie et al., 2025b) and Dream (Ye et al., 2025), extensions to long-context (Liu et al., 2025a) and code generation (Gong et al., 2025), and hybrid autoregressive–diffusion approaches such as Block Diffusion (Arriola et al., 2025). The goal is to clarify what capabilities current DLMs demonstrate, where they are competitive with autoregressive baselines, and where important gaps remain.

Efficient Inference and Practical Speedups. A central focus of the tutorial is whether the parallelism of DLMs can translate into practical speedups. We organize recent inference work around reducing denoising steps, reusing computation through caching or early stopping, and selecting which tokens to update in parallel (Ma et al., 2025; Liu et al., 2025b; Wu et al., 2025; Israel et al., 2025; Wei et al., 2025; Luxembourg et al., 2025; Huang and Tang, 2025; Li et al., 2025; Deschenaux and Gulcehre, 2025; Zhang et al., 2026; Zheng et al., 2026; Oba et al., 2026a). We will emphasize that DLM efficiency should be assessed by wall-clock latency, memory use, throughput, and quality degradation, not only by the number of sampling steps.

Training and Post-training. We discuss how training and post-training methods must be adapted to the generation dynamics of DLMs. Topics include the training–inference gap and refinement-oriented objectives (He et al., 2025; Oba et al., 2026c), as well as reinforcement-learning methods for reasoning and alignment over diffusion trajectories (Zhao et al., 2025; Tang et al., 2026; Wang et al., 2025a; Zhu et al., 2025; Wang et al., 2026; Zekri and Boulle, 2025; Xie et al., 2025; Oba et al., 2026b). We highlight why DLMs cannot simply inherit post-training recipes from autoregressive LMs: intermediate denoising states, non-left-to-right trajectories, and iterative refinement introduce different credit-assignment and optimization challenges.

Open Problems and Emerging Frontiers. We conclude with open problems that will shape the next stage of DLM research. Technically, current DLMs still face unresolved challenges in reliable few-step generation, speed–quality tradeoffs, length control, and controllability, motivating broader non-autoregressive formulations such as flow-based and flow-map language modeling (Lipman et al., 2023; Stark et al., 2024; Lee et al., 2026; Hu et al., 2026; Potaptchik et al., 2026; Roos et al., 2026). Socially, DLMs remain largely unexplored with respect to safety and fairness, despite the central role of these issues in autoregressive and masked language models (Kaneko and Bollegala, 2022; Kaneko et al., 2022; Oba et al., 2024).

4 Tutorial Roadmap

Foundations and Landscape.

Why diffusion language models? Sequential bottlenecks in autoregressive decoding, parallel token updates, iterative refinement, and the promise of few-step generation.

Foundations. Diffusion over text: discrete and masked noising processes, denoising objectives, sampling procedures, and links to masked language modeling and iterative generation.

Model families and scaling. Masked diffusion LMs, large-scale DLMs, long-context and code-oriented models, and hybrid autoregressive–diffusion architectures.

Inference and Training.

Efficient inference. Turning DLM parallelism into practical speedups through adaptive decoding, caching, token updates, early stopping, and careful latency–quality measurement.

Training objectives. Denoising and refinement objectives, masking schedules, diffusion trajectories, and training–inference mismatch.

Post-training and alignment. Adapting RL and preference optimization to DLM trajectories, with emphasis on reasoning, intermediate denoising states, and credit assignment beyond left-to-right generation.

Open Frontiers and Responsible Deployment.

Beyond DLMs. Few-step degradation, speed–quality trade-offs, limitations of current diffusion formulations, and emerging flow-based formulations for language modeling.

Social acceptability and responsible deployment. Reliability, safety, social bias, fairness, multilinguality, and responsible deployment beyond autoregressive generation.

5 Diversity Considerations

The tutorial is designed to be accessible to a broad NLP audience, including participants without prior background in diffusion models. The instructor team combines junior and senior perspectives and spans institutions in Japan and the UK, bringing complementary expertise in diffusion language models, efficient generation, NLP, and fairness. In content, we will connect technical questions about DLMs to open problems in social bias, safety, and responsible deployment, rather than treating DLMs only as an efficiency technique. To support broad participation, we will make the tutorial slides and curated reading list publicly available.

6 Ethics Statement

DLMs aim to make text generation faster, more parallel, and potentially cheaper; these benefits may improve accessibility, but they may also lower the cost of misuse, spam, misinformation, and other harmful generation. In addition, the reliability, safety, and fairness properties of DLMs remain underexplored: their sampling dynamics differ from autoregressive generation, so existing findings and mitigation methods for AR LMs may not directly transfer. The tutorial itself does not introduce new datasets, user studies, or deployed systems, but it will cover research directions whose responsible development requires careful evaluation of safety, fairness and multilingual impact (Kaneko and Bollegala, 2022; Kaneko et al., 2022; Oba et al., 2024; Kaneko et al., 2025, 2026).

Tutor Biographies

Daisuke Oba is a Specially Appointed Assistant Professor at Institute of Science Tokyo, Japan. He brings hands-on experience in large-scale autoregressive LLM development in both academia and industry. He also has recent publications at top-tier ML venues, including ICLR, ICML, and NeurIPS (Oba et al., 2026a; Komiyama et al., 2026; Oi et al., 2026; Miyamoto et al., 2026; Moon et al., 2026; Oba et al., 2026b). Most relevant to this tutorial, he has contributed directly to the emerging DLM literature through work on early stopping for DLM decoding (Oba et al., 2026a), reinforcement learning over diffusion states (Oba et al., 2026b), and refinement objectives for DLMs (Oba et al., 2026c). He has also worked on fairness in LLMs (Oba et al., 2024; Hida et al., 2026), which informs the tutorial's discussion of social bias and responsible deployment. In this tutorial, he will lead the technical coverage of DLM foundations, efficient inference, and training/post-training methods, contrasting DLMs with autoregressive LLM pipelines.
Homepage: `https://daioba.github.io/`

Danushka Bollegala is a Professor of Natural Language Processing at the University of Liverpool, UK, and an Amazon Scholar at Amazon Search. He has published over 220 papers at leading NLP and ML venues including ACL, EMNLP, NAACL, EACL, AACL, NeurIPS, ICLR, AAAI, and IJCAI, and has extensive experience instructing international audiences through invited talks, keynotes, and tutorials at major conferences and summer schools. His research spans large language models, representation learning, and the safety, fairness, and social bias properties of LLMs, including widely-cited work on evaluating and mitigating social biases in masked and large language models (Kaneko and Bollegala, 2022; Kaneko et al., 2022; Oba et al., 2024), multilingual bias benchmarks, and ethical evaluation of human–LLM interactions (Kaneko et al., 2025, 2026). Most relevant to this tutorial, he has recently co-authored work with Daisuke Oba on diffusion language models, including an ICLR 2026 paper on efficient inference via early stopping for converged tokens in masked diffusion-LM decoding (Oba et al., 2026a), which directly informs the efficient-generation portion of this tutorial. Through his role as an Amazon Scholar at Amazon Search, he also brings an industrial perspective on how non-autoregressive paradigms such as DLMs may shape future large-scale deployment of language technologies, and is well-positioned to discuss practical implications, deployment trade-offs, and responsible-use considerations of diffusion-based generation alongside the technical content of the tutorial.
Homepage: `https://danushka.net/`

References

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. In *The Thirteenth International Conference on Learning Representations*.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993.

Justin Deschenaux and Caglar Gulcehre. 2025. Beyond autoregression: Fast LLMs via self-distillation through time. In *The Thirteenth International Conference on Learning Representations*.

Justin Deschenaux and Caglar Gulcehre. 2026. Language modeling with hyperspherical flows. *arXiv preprint arXiv:2605.11125*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*.

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. 2025. Diffucoder: Understanding and improving masked diffusion models for code generation. *arXiv preprint arXiv:2506.20639*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. *Advances in neural information processing systems*, 32.

Haoyu He, Katrin Renz, Yong Cao, and Andreas Geiger. 2025. Mdpo: Overcoming the training-inference divide of masked diffusion language models. *arXiv preprint arXiv:2508.13148*.

Rem Hida, Masahiro Kaneko, Daisuke Oba, Danushka Bollegala, and Naoaki Okazaki. 2026. DyMT-ESB: Dynamic multi-turn evaluation of social bias in user-llm interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2026*. To appear.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. 2026. Elf: Embedded language flows. *arXiv preprint arXiv:2605.10938*.

Chihan Huang and Hao Tang. 2025. Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation. *arXiv preprint arXiv:2505.14455*.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. 2025. Accelerating diffusion llms via adaptive parallel decoding. *arXiv preprint arXiv:2506.00413*.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask – evaluating social biases in masked language models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962, Vancouver, Canada. AAAI Press.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2025. An ethical dataset from real-world interactions between users and large language models. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI), AI and Social Good Track*, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2026. A multilingual social bias benchmark incorporating thinking processes. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, San Diego, USA. Association for Computational Linguistics. To appear.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Junpei Komiyama, Daisuke Oba, and Masafumi Oyamada. 2026. Best-of-infinity: Asymptotic performance of test-time LLM ensembling. In *The Fourteenth International Conference on Learning Representations*.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M Boffi, and Jinwoo Kim. 2026. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. 2025. Beyond fixed: Training-free variable-length denoising for diffusion large language models. *arXiv preprint arXiv:2508.00819*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.

Xiaoran Liu, Zhigeng Liu, Zengfeng Huang, Qipeng Guo, Ziwei He, and Xipeng Qiu. 2025a. Longllada: Unlocking long context capabilities in diffusion llms. *arXiv preprint arXiv:2506.14429*.

Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyuan Wei, Shaobo Wang, and Linfeng Zhang. 2025b. dllm-cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*.

Omer Luxembourg, Haim Permuter, and Eliya Nachmani. 2025. Plan for speed–dilated scheduling for masked diffusion language models. *arXiv preprint arXiv:2506.19037*.

Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2025. dkv-cache: The cache for diffusion language models. *arXiv preprint arXiv:2505.15781*.

Sora Miyamoto, Daisuke Oba, and Naoaki Okazaki. 2026. Aligning tree-search policies with fixed token budgets in test-time scaling of LLMs. In *Forty-third International Conference on Machine Learning*.

Sangwhan Moon, Daisuke Oba, Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. 2026. Beyond perplexity: UTF-8 validity in byte-aware language models. In *Forty-third International Conference on Machine Learning*.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. 2025a. Scaling up masked diffusion models on text. In *The Thirteenth International Conference on Learning Representations*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025b. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Daisuke Oba, Danushka Bollegala, Masahiro Kaneko, and Naoaki Okazaki. 2026a. Stopping computation for converged tokens in masked diffusion-LM decoding. In *The Fourteenth International Conference on Learning Representations*.

Daisuke Oba, Hiroki Furuta, and Naoaki Okazaki. 2026b. Diffusion-state policy optimization for masked diffusion language models. In *Advances in Neural Information Processing Systems*. To appear.

Daisuke Oba, Hiroki Furuta, and Naoaki Okazaki. 2026c. Drifting objectives for refining discrete diffusion language models. *arXiv preprint arXiv:2605.19470*.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2024. In-contextual gender bias suppression for large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1722–1742, St. Julian's, Malta. Association for Computational Linguistics.

Masanari Oi, Koki Maeda, Ryuto Koike, Daisuke Oba, Nakamasa Inoue, and Naoaki Okazaki. 2026. From correspondence to actions: Human-like multi-image spatial reasoning in multi-modal large language models. In *Forty-third International Conference on Machine Learning*.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. 2024. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*.

Peter Potaptchik, Jason Yim, Adhi Saravanan, Peter Holderrieth, Eric Vanden-Eijnden, and Michael S Albergo. 2026. Discrete flow maps. *arXiv preprint arXiv:2604.09784*.

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, Ismail Ilkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. 2026. Categorical flow maps. *arXiv preprint arXiv:2602.12233*.

Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T Chiu, and Volodymyr Kuleshov. 2025. The diffusion duality. In *Forty-second International Conference on Machine Learning*.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. 2024. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. 2024. Dirichlet flow matching with applications to dna sequence design. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46495–46513.

Xiaohang Tang, Rares Dolga, Sangwoong Yoon, and Ilija Bogunovic. 2026. wd1: Weighted policy optimization for reasoning in diffusion language models. In *The Fourteenth International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chenyu Wang, Paria Rashidinejad, DiJia Su, Song Jiang, Sid Wang, Siyan Zhao, Cai Zhou, Shannon Zejiang Shen, Feiyu Chen, Tommi Jaakkola, Yuandong Tian, and Bo Liu. 2026. SPG: Sandwiched policy gradient for masked diffusion language models. In *The Fourteenth International Conference on Learning Representations*.

Guanghan Wang, Gilad Turok, Yair Schiff, Marianne Arriola, and Volodymyr Kuleshov. 2025a. d2: Improved techniques for training reasoning diffusion language models. *arXiv preprint arXiv:2509.21474*.

Xu Wang, Chenkai Xu, Yijie Jin, Jiachun Jin, Hao Zhang, and Zhijie Deng. 2025b. Diffusion llms can do faster-than-ar inference via discrete diffusion forcing. *arXiv preprint arXiv:2508.09192*.

Qingyan Wei, Yaojie Zhang, Zhiyuan Liu, Dongrui Liu, and Linfeng Zhang. 2025. Accelerating diffusion large language models with slowfast: The three golden principles. *arXiv preprint arXiv:2506.10848*.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.

Shaoan Xie, Lingjing Kong, Xiangchen Song, Xinshuai Dong, Guangyi Chen, Eric P Xing, and Kun Zhang. 2025. Step-aware policy optimization for reasoning in diffusion large language models. *arXiv preprint arXiv:2510.01544*.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Oussama Zekri and Nicolas Boulle. 2025. Fine-tuning discrete diffusion models with policy gradient methods. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tunyu Zhang, Xinxi Zhang, Ligong Han, Haizhou Shi, Xiaoxiao He, Zhuowei Li, Hao Wang, Kai Xu, Akash Srivastava, Vladimir Pavlovic, and 1 others. 2026. T3d: Few-step diffusion language models via trajectory self-distillation with direct discriminative optimization. *arXiv preprint arXiv:2602.12262*.

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. 2025. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *arXiv preprint arXiv:2504.12216*.

Haoyang Zheng, Xinyang Liu, Xiangrui Kong, Nan Jiang, Zheyuan Hu, Weijian Luo, Wei Deng, and Guang Lin. 2026. Ultra-fast language generation via discrete diffusion divergence instruct. In *The Fourteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *Preprint*, arXiv:2505.19223.

Author Index

Ai, Xi, 11

Bollegala, Danushka, 44
Bonagiri, Vamshi Krishna, 11

Choudhury, Monojit, 11
Corona, Rodolfo, 20

Djiberou Mahamadou, Abdoul Jalil, 11
Dong, Xin Luna, 32

Gan, Esther, 11

Hruschka, Estevam, 27

Jin, Zhijing, 11

Kan, Min-Yen, 11
Koyejo, Sanmi, 20

Lee, Seunghun J., 11

Miao, Yisong, 11
Mieskes, Margot, 11

Névéol, Aurélie, 11

Oba, Daisuke, 44
Oh, Alice, 1
Oh, Juhyun, 1

Sharma, Sanat, 32
Solorio, Thamar, 1
Sun, Kai, 32

Tang, Nan, 32
Truong, Sang T., 20

Wang, Jiaqi, 32
Wang, Serena Lutong, 20
Winata, Genta Indra, 1

Xia, Yinglong, 32

Yang, Xiao, 32
Yang, Yajing, 11
Yih, Scott Wen-tau, 32
Yoo, Haneul, 1

Zhang, Franklin, 32
Zhang, Ruochen, 1