NLP for Social Good: A Survey of Challenges, Opportunities, and Responsible Deployment

Antonia Karamolegkou[1], Angana Borah[2], Eunjung Cho[3], Sagnik Ray Choudhury[4], Martina Galletti[5,6], Rajarshi Ghosh[7], Ishani Mondal[8], Pranav Gupta[9], Oana Ignat[10], Priyanka Kargupta[11], Neema Kotonya[12], Hemank Lamba[12], Sun-Joo Lee[13], Arushi Mangla[10], Deniz Nazarova[14], Poli Nemkova[4,13], Dina Pisarevskaya[15], Naquee Rizwan[16], Nazanin Sabri[17], Dominik Stammbach[18], Anna Steinberg[19,20], David Tomás[21], Steven R Wilson[22], Bowen Yi[2,23], Jessica Zhu[8], Arkaitz Zubiaga[15], Anders Søgaard[1], Alexander Fraser[20,27], Zhijing Jin[3,24,25,26], Rada Mihalcea[2], Joel R. Tetreault[12], Daryna Dementieva[27]

[1]University of Copenhagen [2]University of Michigan-Ann Arbor [3]ETH Zurich [4] University of North Texas [5]Sony Computer Science Laboratories - Paris [6]University of Rome "La Sapienza" [7]Algoverse [8]University of Maryland, College Park [9]Lowe's [10]Santa Clara University [11]University of Illinois Urbana-Champaign [12]Dataminr [13] United Nations Development Programme (UNDP) [14]University of Washington [15]Queen Mary University of London [16]IIT Kharagpur [17]University of California San Diego [18] Princeton University [19]LMU Munich [20]Munich Center for Machine Learning (MCML) [21]University of Alicante [22]University of Michigan-Flint [23]University of Southern California [24]Max Planck Institute for Intelligent Systems, Tübingen [25]Vector Institute [26]University of Toronto [27]Technical University of Munich

Correspondence: antka@di.ku.dk, daryna.dementieva@tum.de

Abstract

Recent advancements in large language models (LLMs) have unlocked unprecedented possibilities across a range of applications. However, as a community, we believe that the field of Natural Language Processing (NLP) has a growing need to approach deployment with greater intentionality and responsibility. In alignment with the broader vision of AI for Social Good (Tomašev et al., 2020), this paper examines the role of NLP in addressing pressing societal challenges. Through a cross-disciplinary analysis of social goals and emerging risks, we highlight promising research directions and outline challenges that must be addressed to ensure responsible and equitable progress in NLP4SG research.

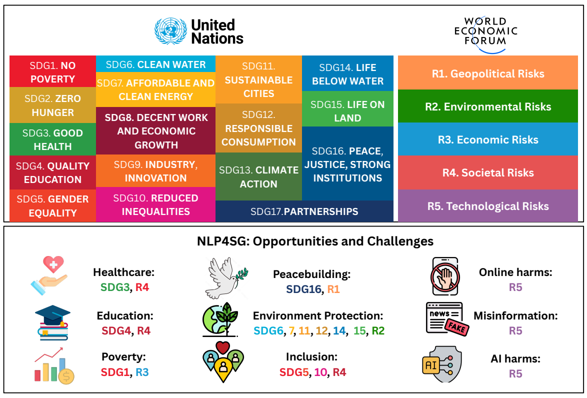

Figure 1: Mapping NLP applications for Social Good (**NLP4SG**) with global goals and risks.

1 Introduction

"Understanding the problem is half the solution."
— Charles Kettering

To fully realize the potential of NLP, it is essential to look beyond technical achievements and reframe tasks around pressing societal needs. We draw on insights from the United Nations Sustainable Development Goals[1] (UN SDGs) and the 2025 Global Economic Risks Report[2] (GR) to provide a foundation for an interdisciplinary recontextualization of NLP, encouraging reflection on how language technologies intersect with today's most pressing challenges. We selected these two agendas as, from a social good perspective, UN SDGs offer a global framework for fostering peace and prosperity for people and the planet. However, while highly influential, these goals were established in 2015—prior to the rapid advancements in artificial intelligence. To contextualize them within today's technological landscape, we also draw on insights from the 2025 GR Report, which highlights both the transformative potential and the emerging global risks (GRs) associated with technology and information processing. Our effort builds on prior research that assesses the role of NLP through positive impact (Hovy and Spruit, 2016; Jin et al., 2021), maps NLP4SG work to the SDGs (Adauto et al., 2023; Gosselink et al., 2024), outlines open questions in modern NLP (Ignat et al., 2024b), and limitations in NLP and AI pipelines (Mihalcea et al., 2025).

[1]https://sdgs.un.org/goals

[2]https://www.weforum.org/publications/global-risks-report-2025/digest

Thus, our research goal in this work is threefold:

RQ1: What NLP-based solutions already support positive social impact?

RQ2: What are the main challenges in developing NLP for positive social impact?

RQ3: What promising directions are currently overlooked?

To address these questions, we identify nine key NLP research directions, as illustrated in Figure 1, using the relevant SDGs and GRs. The following sections summarize research tasks,[3] the open challenges and opportunities within each domain, concluding with overarching open research directions and recommended community actions for proactive NLP4SG efforts.

2 Healthcare

NLP in healthcare is a space that targets problems associated with SDG3: Good Health and Well-being by improving healthcare delivery and outcomes. It also targets critical global risks (GR4), including the decline in health and well-being, health workforce shortages, and infectious disease outbreaks.

Mental Health Given that the need for mental health support increasingly remains unmet, NLP systems, specifically LLMs, address an access gap for users who are limited in time and resources (Tong et al., 2023; Hua et al., 2024). As *counselors*, they could assist in *detecting* conditions like depression and addiction from clinical or social media data (Giuntini et al., 2020; Yang et al., 2023a); *responding* empathetically by interpreting emotion and generating therapeutic dialogues (Shen et al., 2020; Grandi et al., 2024); and *tracking* user mood or crises over time (Čosić et al., 2024; Gong et al., 2019). As *clients*, NLP tools simulate diverse personas to train and evaluate counselors (Louie et al., 2024; Liu et al., 2025a).

Physical Health Prior work has focused on physical well being, using social media mining for tracking physical activity, sleep patterns (Sakib et al., 2021; Shakeri Hossein Abad et al., 2022), diet habits (van Erp et al., 2021; Hu et al., 2023), gauging public health attitudes such as mask-wearing during COVID-19 (He et al., 2021), and detecting risky behaviors like substance use (Hu et al., 2021; Lin et al., 2023). In clinical settings, NLP aids in medical record analysis via classification for treatment decisions, named entity recognition for patient-trial matching, relation extraction linking symptoms and treatments, predictive modeling of treatment responses (Jerfy et al., 2024), and information extraction to organize text reports (Sheikhalishahi et al., 2019; Landolsi et al., 2023).

Challenges: Key challenges include *health-related data*, which is often scarce, sensitive, and affected by systemic biases, with language limitations and underrepresentation of marginalized groups raising ethical and privacy concerns (Ford et al., 2019; Shakeri Hossein Abad et al., 2022; Gunal et al., 2025); *evaluation*, which must go beyond accuracy to reflect fairness, contextual understanding, human-centered values (e.g. empathy), and ensure reproducibility; and *long-term user impact*, as reliance on LLMs for sensitive tasks risks sycophancy, ELIZA effects (Ekbia, 2008), and overdependence. There is also a *lack of causal frameworks* and *interpretable* models, which complicates the understanding in the outcomes of NLP systems (Zhang et al., 2022). Remote care with LLM-based tools risks diverting individuals from essential in-person treatment (Khawaja and Bélisle-Pipon, 2023; Sweeney et al., 2021). This highlights a broader challenge: ensuring such technologies *support—not replace*—human professionals, with responsible use as a guiding principle(Shakeri Hossein Abad et al., 2022; Brown and Halpern, 2021).

Opportunities: Future work should extend beyond text to *multimodal approaches* that integrate speech prosody, facial expressions, physiological signals, and other data sources like sensors, wearables, or images for richer context and personalization (Puce et al., 2025). *Multi-agent* and *adaptive dialogue systems* that consider user history, emotions, and culture could boost performance. Designing *holistic evaluation frameworks* in real-world simulations to ensure privacy, explainability, fairness, and accessibility can help address AI risks (Lawrence et al., 2024; Yao et al., 2024). At the policy level, NLP can analyze public sentiment on AI used for healthcare applications from actual users, and health guidelines to inform better regulations and health campaigns (Lindquist et al., 2021). Crucially, *interdisciplinary research* is needed to build AI-augmented therapeutic frame-

[3] Details about methodology and motivation of paper selection can be found in the Appendix A.

works that complement human care—especially for underserved communities.

3 Education

The integration of LLMs into education offers tools that support SDG4: Quality Education, and also addresses a variety of risks by fostering informed and critical individuals (GR4, GR5). NLP is used for automated feedback (Jurenka et al., 2024; Bauer et al., 2023; Gao et al., 2024; Stamper et al., 2024; Ramesh and Sanampudi, 2022), tailored support (Kazemitabaar et al., 2024; Daheim et al., 2024), and self-paced learning (Kazemitabaar et al., 2023). NLP tools can expand access in underserved regions (Yu et al., 2024), bridge language gaps (Molina et al., 2024; Kwak and Pardos, 2024), assist learners with disabilities (Cheng et al., 2024), and ease teacher workload (Lan and Chen, 2024; Wang et al., 2024b; Shridhar et al., 2022).

AI Literacy As AI systems become more embedded in everyday life, AI literacy is essential. A recent review (Yang et al., 2025) draws connections with digital, data, and algorithmic literacies. However, NLP-driven efforts for AI literacy remain scattered (Long and Magerko, 2020; Congress, 2023; Moorkens et al., 2024; Tapo et al., 2025; Korea Education and Research Information Service, 2025)

Challenges: *Model limitations*—such as lack of pedagogical reasoning (Wang and Demszky, 2023; Macina et al., 2023, 2025), misaligned explainability (Okolo and Lin, 2024), and accuracy (Stamper et al., 2024; Kargupta et al., 2024)—hinder the effective integration of NLP into education. *Mixed perceptions and mistrust* toward AI also remain a barrier (Nader et al., 2022; Laupichler et al., 2024). Broader issues like *language, cultural differences, curriculum gaps, weak policy support, and limited infrastructure*—especially in developing regions—further restrict equitable access to AI/NLP in education (Kathala and Palakurthi, 2025).

Opportunities: Further research should be encouraged for *specific groups*, such as teachers (Du et al., 2024), students (Shen and Cui, 2024), and other professionals (Lo, 2024). Another important opportunity is the *alignment* with expert-annotated pedagogical traces and grounding evaluations in curriculum-linked outcomes, to bridge the gap between subject expertise and pedagogical effectiveness (Macina et al., 2025; Lucy et al., 2024). *Human-in-the-loop approaches* have also shown promise in scaling expert strategies while preserving teacher agency, particularly benefiting underserved communities (Wang et al., 2024c). *Multi-agent simulations* offer privacy-preserving environments to test classroom policies and assess equity metrics before real-world deployment (Zhang et al., 2025). For example, Socratic planning agents encourage critical thinking over rote memorization, fostering deeper learning experiences (Kargupta et al., 2024). Finally, *community-driven* initiatives are crucial for consolidating best practices for the development of safe and accountable educational AI systems (Chu et al., 2025; Wen et al., 2024).

4 Poverty

Economic downturn (GR3) is the sixth highest-ranked GRs, with serious implications for poverty—one of the world's most pressing challenges (Lister, 2021) and the focus of SDG1. Nearly 700 million people continue to live on less than $2.15 per day (Hasell et al., 2024; World Bank Group, 2024). NLP methods have been used to extract socioeconomic patterns from news (Lampos et al., 2014), analyze text (Paterson and Gregory, 2018; Hoeschle et al., 2025), classify poverty status (Muñetón-Santa et al., 2022), extract poverty-related dimensions from interviews (Muñetón-Santa and Orozco-Arroyave, 2023) or analyze global narratives around poverty (Curto et al., 2024), and to identify performance disparities among different socioeconomic groups (Cercas Curry et al., 2024; Nwatu et al., 2023).

Challenges: A major challenge is the fact that *poverty data are often scarce or incomplete* (Tingzon et al., 2019; Fatehkia et al., 2020). Indicators like income or poverty status are rarely shared, making it hard to infer them from text. As a result, most socio-economic NLP studies rely on proxies such as mean income (Hasanuzzaman et al., 2017; Abraham et al., 2020), education (Cercas Curry et al., 2024), or (un)employment (Preoţiuc-Pietro et al., 2015b,a). These proxies may not accurately reflect true poverty levels and often vary across studies, resulting in *limited comparability and impact*.

Opportunities: To enable *large-scale, global analysis* of poverty, high-quality datasets labeled with income, socioeconomic status, or poverty indicators are essential, potentially gathered via data donations, user surveys, or social media statistics.[4] Additionally, *use-case-specific NLP applications* can advance poverty research (Adauto et al., 2023). For example, *model analysis* can track the performance of current systems across socio-economic levels, *information extraction* can monitor government funding for poverty alleviation, and *machine translation* can improve resource access for non-English speakers. Finally, future work should develop *clearer guidelines and taxonomies* for categorizing poverty-related data and NLP methods to ensure consistency and comparability across studies.

5 Peacebuilding

Peacebuilding is essential to achieving SDG16, and despite escalating threats like state-based armed conflict (R1); NLP's application in key tools for peacebuilding like conflict prediction, human rights monitoring, and physical safety remains limited and underexplored.

Human Rights Violations NLP in human rights mainly involves classification and detection of violations across languages, such as Arabic Twitter (Alhelbawy et al., 2016), Jordanian Arabic social media (Khalafat et al., 2021), Russian and Ukrainian Telegram (Nemkova et al., 2023), and English Twitter (Pilankar et al., 2022). Other tasks include threat detection for human rights defenders (Ran et al., 2023) and interpretable forced labor identification (Guzman et al., 2024). LLMs show potential for zero- and few-shot annotation but require human oversight (Nemkova et al., 2025b).

Conflict Prediction NLP in conflict prediction mainly involves extracting and forecasting conflict dynamics using structured datasets like UCDP and ACLED (Hegre et al., 2019). Tasks include topic modeling for early signals (Mueller et al., 2024), modeling conflict stages with LLMs (Croicu and von der Maase, 2025; Nemkova et al., 2025a), automated report generation via retrieval-augmented generation (Nemkova et al., 2025c), and conflict content extraction and classification with domain-adapted models like ConfliBERT (Hu et al., 2022).

Physical Safety NLP systems support understanding and intervention in threats to physical safety. They can be used to analyze social media, police reports, surveillance footage, health records, and news articles. Techniques like topic modeling (More and Francis, 2021), sentiment analysis (Blevins et al., 2016), entity extraction (Pavlick et al., 2016), and document classification (Chang et al., 2018) help identify patterns of violence, perpetrators, and victims, informing interventions and policy decisions.

Challenges: First, *Data and Domain Complexity*: references to violations and threats are often indirect or euphemistic, complicating detection. Labeled multilingual datasets—especially for low-resource or sensitive regions—are scarce (AL-Saif and Alotaibi, 2019; MacPhaul et al., 2023; Parker, 2020; Chang et al., 2018), and causes are often multifaceted and extend beyond textual signals. Annotation is labor-intensive with expert disagreement common (Blandfort et al., 2019; Levy et al., 2022; Botelle et al., 2022; Chew et al., 2023), while evaluation suffers from the absence of standardized benchmarks and reliable ground truth, limiting model validation and comparison. Second, *Model Limitations*: NLP models struggle with code-switching, generalization across languages, regions, and out-of-distribution data, and often reflect training biases. The black-box nature of LLMs reduces transparency, hindering their deployment in high-stakes contexts, with real-world application still in early stages and lacking validated best practices. Lastly, *Ethical and Evaluation Challenges*: misclassifications can lead to legal or advocacy risks, highlighting the need for rigorous evaluation and ethical care.

Opportunities: *On a computational level*, tools such as conflict prediction models and systems for detecting human rights violations can support evidence-based policymaking, inform the development and targeting of interventions, and enable rapid programmatic adjustments in response to evolving conditions. Developing multilingual and cross-regional models will enable more comprehensive monitoring. Further, NLP methods can contribute to *operational efficiency* for implementing organizations. For example, LLMs offer capabilities for rapid document review and synthesis, facilitating timely and contextually informed analysis. RAG systems can further enhance information access by efficiently extracting relevant content from document repositories. Additionally, LLMs can support scenario generation, situational analysis, and strategic planning. Lastly, a dedicated *peacebuilding workshop* could offer a crucial forum for

[4] https://datareportal.com/social-media-users

interdisciplinary collaboration on NLP tools for humanitarian aid and crisis response.

6 Environment Protection

NLP offers scalable tools for climate mitigation and adaptation—supporting key SDGs (SDG6, SDG11, SDG12, SDG14, SDG15) and addressing critical global environmental risks (R2)—by extracting insights from unstructured text like scientific papers, policy reports, and assessments (on Climate Change, 2022). Techniques such as topic modeling (Sietsma et al., 2023), summarization (Ghinassi et al., 2024), and classification (Varini et al., 2020; Stammbach et al., 2023; Bingler et al., 2022; Schimanski et al., 2023) enable analysis of diverse datasets. NLP can also help to detect misinformation and greenwashing by verifying claims (Diggelmann et al., 2020; Hsu et al., 2024). While fine-tuning or pretraining LLMs on climate text is common (Leippold et al., 2022; Thulke et al., 2024), RAG-based chatbots (Vaghefi et al., 2023) and automated fact-checkers (Leippold et al., 2025) are emerging rapidly.

Challenges: Extracting quantitative information, particularly from sustainability reports, remains challenging due to *the complex, multi-modal and non-standardized nature of the data sources*, leading to pipelines specifically designed for information extraction from tables (Mishra et al., 2024; Dimmelmeier et al., 2024). Furthermore, LLMs are prone to *hallucination* (Vaghefi et al., 2023) or are contaminated with false, conflicting or outdated climate information (Fore et al., 2024). Bulian et al. (2024) propose an evaluation framework based on presentational and epistemological qualities to evaluate quality and factual accuracy of LLMs.

Opportunities: NLP is enabling *cross-disciplinary collaboration*, bridging domain experts with the fields of computer science, social science, and economics to analyze climate policies (Gandhi et al., 2024) or bringing the scientific community and non-governmental organization together to uncover narratives in public climate discourse (Gehring and Grigoletto, 2023; Rowlands et al., 2024). This is particularly important in expanding research to include *cross-cultural and multilingual perspectives* (Zhou et al., 2024; Bird et al., 2024).

7 Inclusion and Inequalities

Inclusive NLP that accounts for demographic, linguistic, and accessibility diversity is key to reducing systemic inequalities. These efforts support SDG5 and SDG10, and address inequality—a top global concern ranked among the most severe short- and long-term risks (GR4).

Gender Bias Gender bias in NLP—especially in hiring (De-Arteaga et al., 2019), healthcare, and moderation—can reinforce stereotypes and harm marginalized groups. Initial work focused on static embeddings using gender direction (Bolukbasi et al., 2016) and WEAT (Caliskan et al., 2017), later extending to contextual tasks like coreference (Zhao et al., 2018) and occupation classification (De-Arteaga et al., 2019). Recent studies examine LLM-agent interactions (Borah and Mihalcea, 2024) to track how biases spread. Mitigation includes post-processing (e.g., fine-tuning, DPO), interpretability, causal inference (Attanasio et al., 2023; Cai et al., 2024), and counterfactual or intervention-based prompting (Plyler and Chi, 2025). Bias is commonly assessed via fairness gaps (e.g., F1, error disparities) or associations (e.g., WEAT, Bias-Score).

Cultural and Linguistic Representation NLP systems often fail to capture the full diversity of human cultures and languages. Culture's complexity - including evolving norms, values, and worldviews - varies widely across communities (Saha et al., 2025). Despite culture's crucial role in communication, AI models frequently reflect incomplete cultural perspectives, overrepresenting dominant or cultural linguistic narratives typically from high-resource languages or communities (Hershcovich et al., 2022a; Mihalcea et al., 2025).

Underrepresented Communities Users with diverse accessibility needs remain underserved, continuing to face significant communication barriers (Khanuja et al., 2023a). NLP has been used in a wide range of assistive applications, including as augmentative communication (Park et al., 2022), text simplification (Espinosa-Zaragoza et al., 2023), text-to-speech (Kumar et al., 2023), speech recognition (Li et al., 2022), braille processing (Tejesh et al., 2025), image captioning and subtitling (Stefanini et al., 2021), question answering (Gurari et al., 2018), assistive chatbots (Grassini et al., 2024), reading aids (Wang et al., 2024d), and sign language translation (Rust et al., 2024).

Challenges: Many issues revolve around *binary gender assumptions* and limited *cross-linguistic/cultural generalizability* (Stanovsky et al., 2019; Adilazuarda, 2024). Western-centric resources and lack of intersectional frameworks reinforce biases and marginalization (Kleinberg and Raghavan, 2021; Sewunetie et al., 2024). Implicit biases and compounded discrimination, e.g., at the intersection of race, gender and disability, are underexplored (Stanczak and Augenstein, 2021; Guo and Caliskan, 2021; Wald, 2021). *Data* challenges persist due to scarce multimodal datasets, especially for Braille and sign language (Hutchinson and Prabhakaran, 2020; De Sisto et al., 2022). *Limited interdisciplinary collaboration* slows progress (Kusters et al., 2020). Reliance on synthetic or prompt-generated data hampers real-world validity and hides systemic biases (Venkit et al., 2025; Morales et al., 2024), while opaque data sourcing blocks accountability (Bender and Friedman, 2018).

Opportunities: Inclusivity in NLP advances through participatory approaches centering marginalized voices, including co-designed fairness goals in gender bias mitigation (Borah and Mihalcea, 2024; Ma et al., 2023; Lauscher et al., 2022) and collaborative data collection with cultural experts and underrepresented groups (Bird and Yibarbuk, 2024; Newman-Griffis et al., 2024; UNICEF, 2020; Hirmer et al., 2021). Solutions such as dynamic audit pipelines (Park et al., 2023), lightweight model editing (Park et al., 2023; Cai et al., 2024) and counterfactual data augmentation (Zmigrod et al., 2019) help adapt models to sociocultural shifts. Fine-tuning LLMs can improve fairness and relevance (Mai and Carson-Berndsen, 2024; Bartl and Leavy, 2024). Future work should consider identity compositionality (Welch et al., 2020) and pluralistic alignment reflecting complex social affiliations (Sorensen et al., 2024). Personalized, multimodal interaction design is vital for adaptive, accessible systems (Paice et al., 2025; Wang et al., 2024d).

8 Online Harms

The recent ease of access to digital devices (like smartphones and those based on IoT) has fueled the spread of digital violence globally (Bjelajac and Filipović, 2021). This is central to Technological Risks, one of the most critical risks identified in the 2025 Global Risks Report (GR5). There has been a large body of work on abusive/offensive/toxic/harmful speech classification (Diaz-Garcia and Carvalho, 2025), generation of counter speech (Bonaldi et al., 2024; Saha et al., 2024; Wang et al., 2024a) and text detoxification (Dementieva et al., 2025; Dale et al., 2021), including several languages (Aluru et al., 2020).

Challenges: Recent social media platform statements[5] highlight the limitations of AI-based content moderation at scale. Challenges include the *subjective nature* of moderation, *regional regulations*, and *the culturally diverse* and implicit nature of content (Ocampo et al., 2023). LLMs struggle with volatile topics without *frequent fine-tuning* (Roy et al., 2023), and efforts are hindered by *unclear label taxonomies*, *biases* in pre-trained models, and limited *collaboration between lawmakers, platforms, and researchers* (Yimam et al., 2024).

Opportunities: Opportunities for future work include building low-resource, robust, and generalizable LLM moderation frameworks that are deployable in real time, requiring *stronger collaboration between researchers, moderators, and policymakers* (Munzert et al., 2025; Bui et al., 2025). A clearer, *more widely accepted taxonomy of digital violence* is needed to reduce ambiguity in label definitions (Yimam et al., 2024). The current focus on text-based hate overlooks multimodal content like video and audio, calling for LLMs with *multimodal moderation capabilities*. To address culturally diverse and implicit content, future systems should integrate *explainability*, *fact-checking*, and real-time tools such as RAGs and search APIs. Finally, *moderation efforts must expand beyond hate speech* to cover the full spectrum of digital violence, while also embracing more inclusive approaches that *consider author and receiver demographics*, and broader cultural and linguistic diversity.

9 Misinformation

Ranked among the top five global risks in both short- and long-term scenarios (GR5), misinformation has a direct impact on the success of many of the SDGs.[6] NLP approaches to misinformation in-

[5] https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes

[6] https://www.un.org/sites/un2.un.org/files/information-integrity-and-sdgs-en.pdf

clude tasks such as fake news detection, rumor classification, stance detection, and fact checking (Oshikawa et al., 2020; Nakashole and Mitchell, 2014). Early methods used stylometric features, while modern systems use neural models, pre-trained LMs, and techniques like retrieval-augmented generation and prompt-based learning in multimodal and multilingual settings (Akhtar et al., 2023; Chen and Shu, 2024b).

Challenges: There is a *data scarcity*, especially in low-resource languages, emerging domains, and shifting distributions (Guo et al., 2022). *Multilingual and conflicting evidence* complicates verification (Schlichtkrull, 2024; Zhang et al., 2024), while *low-visibility claims* targeting marginalized groups often go unchecked (Guo et al., 2022). In *high-stakes domains* like healthcare and politics, reliable detection is critical to prevent real-world harm (Abdul-Mageed et al., 2021; Zhao et al., 2023b), yet systems still lack *robustness, fairness, and explainability*. Lastly, LLMs pose *misuse risks* by generating convincing falsehoods (Buchanan et al., 2021; Gabriel et al., 2024).

Opportunities: A key opportunity lies in developing *human-centered evaluation* methods tailored to real-world misinformation detection tasks (Das et al., 2023a). This would also benefit from *interdisciplinary collaboration*—for example, by integrating fundamental theories from social sciences and economics (Zhou and Zafarani, 2020), or by partnering with moderators, policymakers, and professional fact-checkers (Warren et al., 2025). Additionally, advancing the *timeliness of ground truth* through LLM-agents with access to web evidence and external knowledge bases can improve adaptability to fast-evolving misinformation. Developing *domain-agnostic features* that generalize across topics, languages, modalities, and time would help track shifting deceptive styles (Chen and Shu, 2024b) and identify check-worthy claims.

10 AI Harms

AI harms are among the top global threats (GR5). NLP is part of the problem—but also part of the solution. We start with the black-box problem and follow the harm taxonomy by Weidinger et al. (2022) to explore NLP mitigation strategies.

The Black Box Problem LLMs often operate as black boxes, offering limited interpretability and transparency (Hassija et al., 2024). With little insight into their design or training, the public and AI community remain reliant on what creators choose to disclose. To address this, NLP research has explored textual explanations, such as summaries (Atanasova et al., 2020; Kotonya and Toni, 2020b), contrastive (Schuster et al., 2021), and counterfactual forms (Yang et al., 2020; Tolkachev et al., 2022). Visual methods like LIME (Ribeiro et al., 2016), ACE (Ghorbani et al., 2019), and heatmaps (Arras et al., 2017) highlight input relevance, while structural tools like SVCCA (Raghu et al., 2017), t-SNE, and TCAV (Kim et al., 2018) uncover concept-level insights.

Representation and Toxicity NLP techniques are actively developed to address biases in training data and AI models (Gallegos et al., 2024), to reduce bias and toxicity. Further discussion is provided in §7 on inequality and §8 on digital toxicity.

Privacy, Safety and Malicous Uses NLP approaches to harm reduction include learning to refuse or prevent memorization to protect personal data (Carlini et al., 2021; Liu et al., 2025b), adversarial training (Goyal et al., 2023), safe decoding (Xu et al., 2024b), authorship verification (Huang et al., 2024a), safety alignment (Bhardwaj et al., 2024), and red-teaming to expose vulnerabilities (Purpura et al., 2025).

Ungrounded Knowledge As pointed in §9 LLM outputs can include hallucinations and spread misinformation (Pan et al., 2023; Huang et al., 2024b; Chen and Shu, 2024b). Mitigation strategies include grounding with external knowledge via RAG (Lewis et al., 2020; Peng et al., 2023; Asai et al., 2024), expressing uncertainty (Yang et al., 2023c; Feng et al., 2024; Xiong et al., 2024; Deng et al., 2024), and self-verification methods (Weng et al., 2023; Hong et al., 2024).

Socioeconomic and Environmental Harms In NLP, researchers have called for greater climate awareness and transparency through reporting frameworks (Hershcovich et al., 2022b), while others focus on reducing energy use with model optimization techniques like pruning, quantization, and distillation (Schwartz et al., 2020; Jin et al., 2024; Zhu et al., 2024). Tackling socioeconomic harms (§4, §7) further requires closer collaboration with sociology and HCI (Blodgett et al., 2024; Card et al., 2024) to further enhance the limited work in this direction.

Challenges: Key barriers to prevent AI harms include transitioning from prototype to deployment which entails coping with *data distributional shifts* between "lab" and "field" data, and managing *bias or subjective labels*. Many current benchmarks often overlook *low-resource* or politically *sensitive domains* (Kim et al., 2025; Cho et al., 2025). Additionally, the lack of standardized practices for measuring *real-world impact* complicates evaluation beyond standard metrics. The *effectiveness* of the proposed tasks is often questioned (Chen and Shu, 2024a; Kotonya and Toni, 2024), and there are very *few rigorous and holistic evaluation frameworks* (Atanasova et al., 2023; Liang et al., 2023). There are also several governance problems and establishing *equitable partnerships* with nonprofits, where power imbalances or misaligned goals can hinder collaboration.

Opportunities: *Field trials* of NLP systems under real-world conditions are necessary, especially in low-resource or use-case-specific settings, while encouraging participatory design and critical user engagement (Pataranutaporn et al., 2025). On the modeling side, advances in *retrieval-augmented generation (RAG)* and *knowledge augmentation* methods can improve reliability by enhancing credibility (Xu et al., 2024a; Chen et al., 2024, 2025). *Multidisciplinary approaches* such as logical relationships between inputs (Ayoobi et al., 2025; Freedman et al., 2025) and mechanistic interpretations of model behavior (Hou et al., 2023; Yu and Ananiadou, 2024) can further enhance transparency. *New research directions* could be implemented by studying LLM overreliance, manipulation and unfair distribution of benefits from model access. Embracing *process-aware NLP* enables alignment with domain logic and ethical goals through explainability and fairness-by-design (Bernardi et al., 2024; Zhuang et al., 2025). Finally, *mapping existing frameworks* for AI4SG and NLP4PI (Fairness, Accountability, and Transparency in Machine Learning (FAT/ML), 2020; Floridi et al., 2021), suggested actions (NIST, 2024) and principles for trustworthy AI (OECD, 2024), can support the development of adaptable, unified guidelines for responsible NLP deployment.

11 Summary and Call to Action

To achieve the full potential of NLP4SG, the community should move beyond task-centric innovation toward impact-driven systems that are safe, inclusive, and globally relevant. This work reviewed nine key application areas, addressing **RQ1** on existing impactful solutions with concrete examples of how modern NLP technologies have already contributed. At the same time, in addressing **RQ2** on key open challenges, we find that areas like misinformation, online harms, and education have seen sustained attention in NLP, while domains such as poverty alleviation, environmental protection, peacebuilding are only starting to gain traction in response to real-world crises. In response to **RQ3** concerning promising directions for future work, we outline a set of universal challenges, opportunities, and actionable recommendations to guide future research in NLP4SG:

Recurring Challenges: Across these diverse domains, we identified five systemic challenges that continue to hinder progress: (1) *Data scarcity and representational bias*, particularly affecting low-resource languages and marginalized groups; (2) *Misaligned evaluation metrics* that fail to capture human-centered qualities like empathy or cultural sensitivity; (3) *Safety, privacy, and ethical concerns*, which are magnified in high-stakes settings such as mental health or misinformation detection; (4) *Limited generalization* across cultures, modalities, and real-world conditions; and (5) *Infrastructural gaps* and institutional fragmentation, especially in underrepresented scenarios.

Emerging Opportunities: Despite these barriers, we highlight several promising cross-cutting directions: (1) *Multimodal and multilingual learning* can help systems better reflect real-world diversity; (2) *Human-AI collaboration* enables more aligned, adaptive and interpretable NLP pipelines; (3) *Participatory design and evaluation* ensure that benchmarks and systems are co-developed with affected communities; (4) *Retrieval-augmented and policy-aware methods* offer new tools for building verifiable and context-sensitive applications; and (5) *Explainability and AI literacy* foster critical engagement and equitable access.

Call to Action: To advance NLP4SG, we call on the community to: (1) develop joint benchmarks featuring multilingual, culturally diverse, and socially grounded data; (2) collaborate closely with domain experts, such as educators, health practitioners, and civil society organizations, to co-design evaluation frameworks that reflect end-user needs; (3) pursue human-centered methodologies instead of one-size-fits-all solutions. Progress depends on pluralistic, context-aware roadmaps that align with both local realities and global development goals. Finally, while modern LLMs offer significant potential, it is crucial to ensure their *affordability* and *accessibility* so they serve the public good rather than exacerbate existing inequalities.

NLP has the tools to move beyond abstract benchmarking and toward socially responsive technologies designed with—and for—impacted communities. Realizing this vision requires not just technical innovation but also sustained interdisciplinary collaboration, inclusive practices, and a commitment to long-term global equity. We hope our findings can help earlier career researchers find their research niche and that more advanced researchers will have a fresh overview of the field to foster NLP4SG applications with a more interdisciplinary paradigm.

Limitations

This paper offers a high-level interdisciplinary perspective on aligning NLP research with societal needs, grounded in the UN Sustainable Development Goals (SDGs) and the World Economic Forum's Global Risks Report. While our proposed framework maps NLP research directions to these agendas, the related works list discussed is not exhaustive. While we aimed to cover the most impactful topics based on the authors' expertise, we acknowledge that ongoing advancements in NLP may enable new tasks and uncover deeper layers of impact beyond what could be envisioned at the time of writing.

Furthermore, while we highlight areas of overlap between the SDGs and global risks, the mappings remain somewhat subjective. Another key assumption we made is that positive impact is highly aligned with the UN Sustainable Development Goals - but this might not be true i.e. positive impact could still be derived from NLP tools without there being an explicit alignment towards SDGs. At the same time, we believe that our work will open more cross-disciplinary discussions and more ideas for NLP4SG applications.

Ethics Statement

This work is grounded in the belief that NLP research should be aligned with broader societal priorities and developed with care for its downstream impact. In proposing mappings between NLP directions, global goals, and risks, we are mindful of the potential for unintended consequences and overgeneralization. The open areas of work mentioned throughout the paper are just suggestive, and the practitioners, when working on them, should carefully evaluate the setting and downstream impacts.

We emphasize the importance of interdisciplinary collaboration and the inclusion of marginalized perspectives in shaping responsible NLP research. Where possible, we cite and build on existing initiatives in NLP for social good, and we support efforts to foreground equity, inclusivity, and transparency in both research framing and methodology. Finally, we acknowledge that AI assistants were used for proofreading this paper, a practice we consider ethically acceptable under appropriate usage.

Acknowledgments

This work was supported by the Novo Nordisk Foundation (grant NNF20SA0066568), which funded the research of Anders Søgaard and Antonia Karamolegkou. Antonia Karamolegkou was also supported by the Onassis Foundation - Scholarship ID: F ZP 017-2/2022-2023. Zhijing Jin was supported partly by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; and by Schmidt Sciences SAFE-AI Grant. Daryna Dementieva's work was additionally supported by the Friedrich Schiedel TUM Fellowship.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, El Moatez Billah Nagoudi, Dinesh Pabbi, Kunal Verma, and Rannie Lin. 2021. Mega-COV: A billion-scale dataset of 100+ languages for COVID-19. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3402–3420, Online. Association for Computational Linguistics.

Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2819–2826.

Amani S. Abumansour and Arkaitz Zubiaga. 2023. Check-worthy claim detection across topics for automated fact-checking. *PeerJ Comput. Sci.*, 9:e1365.

Fernando Adauto, Zhijing Jin, Bernhard Schölkopf, Tom Hope, Mrinmaya Sachan, and Rada Mihalcea. 2023. Beyond good intentions: Reporting the research landscape of NLP for social good. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 415–438, Singapore. Association for Computational Linguistics.

Muhammad Adilazuarda. 2024. Beyond Turing: A comparative analysis of approaches for detecting machine-generated text. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 1–12, Mexico City, Mexico. Association for Computational Linguistics.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Mohammed Ali Al-Garadi, Sangmi Kim, Yuting Guo, Elise Warren, Yuan-Chi Yang, Sahithi Lakamana, and Abeed Sarker. 2022a. Natural language model for automatic identification of intimate partner violence reports from twitter. *Array*, 15:100217.

Mohammed Ali Al-Garadi, Yuan-Chi Yang, and Abeed Sarker. 2022b. The role of natural language processing during the COVID-19 pandemic: Health applications, opportunities, and challenges. *Healthcare (Basel)*, 10(11).

Erfan Al-Hossami, Razvan Bunescu, Ryan Teehan, Laurel Powell, Khyati Mahajan, and Mohsen Dorodchi. 2023. Socratic questioning of novice debuggers: A benchmark dataset and preliminary evaluations. In *Proceedings of the 18th workshop on innovative use of nlp for building educational applications (bea 2023)*, pages 709–726.

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. 2021. Automatic speech recognition: Systematic literature review. *Ieee Access*, 9:131858–131876.

Ayman Alhelbawy, Poesio Massimo, and Udo Kruschwitz. 2016. Towards a corpus of violence acts in Arabic social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1627–1631, Portorož, Slovenia. European Language Resources Association (ELRA).

Stephen R. Ali, Thomas D. Dobbs, Hayley A. Hutchings, and Iain S. Whitaker. 2023. Using chatgpt to write patient clinic letters. *The Lancet Digital Health*, 5(4):e179–e181.

Syed Asif Ali. 2023. Artificial intelligence techniques to understand braille: a language for visually impaired individuals. In *Handbook of Research on Artificial Intelligence Applications in Literary Works and Social Media*, pages 254–276. IGI Global.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *Preprint*, arXiv:2402.13231.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Hissah ALSaif and Taghreed Alotaibi. 2019. Arabic text classification using feature-reduction techniques for detecting violence on social media. *International Journal of Advanced Computer Science and Applications*, 10(4).

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *Preprint*, arXiv:2004.06465.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. "what is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12(8):1–23.

Alina Arseniev-Koehler, Susan D. Cochran, Vickie M. Mays, Kai-Wei Chang, and Jacob G. Foster. 2022. Integrating topic modeling and word embedding to characterize violent deaths. *Proceedings of the National Academy of Sciences*, 119(10):e2108801119.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Giuseppe Attanasio, Flor Miriam Plaza-del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. *arXiv preprint arXiv:2310.12127*.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Hamed Ayoobi, Nico Potyka, and Francesca Toni. 2025. Protoargnet: Interpretable image classification with super-prototypes and argumentation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 1791–1799. AAAI Press.

Kumar Ayush, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. 2020. Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612*.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 215–236, Cham. Springer International Publishing.

Marion Bartl and Susan Leavy. 2024. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, Bangkok, Thailand. Association for Computational Linguistics.

Elisabeth Bauer, Martin Greisel, Ilia Kuznetsov, Markus Berndt, Ingo Kollar, Markus Dresel, Martin R Fischer, and Frank Fischer. 2023. Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda. *British Journal of Educational Technology*, 54(5):1222–1245.

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.

M. L. Bernardi, A. Casciani, M. Cimitile, and 1 others. 2024. Conversing with business process-aware large language models: the bpllm framework. *Journal of Intelligent Information Systems*, 62:1607–1629.

Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. 2024. Language models are Homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14138–14149, Bangkok, Thailand. Association for Computational Linguistics.

Savita Bhat and Vasudeva Varma. 2023. Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 100–107, Bali, Indonesia. Association for Computational Linguistics.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Steven Bird, Angelina Aquino, and Ian Gumbula. 2024. Envisioning NLP for intercultural climate communication. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 111–122, Bangkok, Thailand. Association for Computational Linguistics.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839.

Željko Bjelajac and Aleksandar Filipović. 2021. Specific characteristics of digital violence and digital crime. *Pravo-teorija i praksa*, 38(4):16–32.

Philipp Blandfort, Desmond U. Patton, William R. Frey, Svebor Karaman, Surabhi Bhargava, Fei-Tzin Lee, Siddharth Varia, Chris Kedzie, Michael B. Gaskell, Rossano Schifanella, Kathleen McKeown, and Shih-Fu Chang. 2019. Multimodal social media analysis for gang violence prevention. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):114–124.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *Preprint*, arXiv:2110.06733.

Terra Blevins, Robert Kwiatkowski, Jamie MacBeth, Kathleen McKeown, Desmond Patton, and Owen Rambow. 2016. Automatically processing tweets from gang-involved youth: Towards detecting loss and aggression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2196–2206, Osaka, Japan. The COLING 2016 Organizing Committee.

Su Lin Blodgett, Amanda Cercas Curry, Sunipa Dev, Michael Madaio, Ani Nenkova, Diyi Yang, and Ziang Xiao, editors. 2024. *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Mexico City, Mexico.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angana Borah, Aparna Garimella, and Rada Mihalcea. 2025. Towards region-aware bias evaluation metrics. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 108–131, Albuquerque, New Mexico. Association for Computational Linguistics.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Borger, Pablo Mosteiro, Heysem Kaya, Emil Rijcken, Albert Ali Salah, Floortje Scheepers, and Marco Spruit. 2022. Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting. *Expert Systems with Applications*, 199:116720.

Riley Botelle, Vishal Bhavsar, Giouliana Kadra-Scalzo, Aurelie Mascio, Marcus V Williams, Angus Roberts, Sumithra Velupillai, and Robert Stewart. 2022. Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study. *BMJ Open*, 12(2):e052911.

Julia EH Brown and Jodi Halpern. 2021. Ai chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM-Mental Health*, 1:100017.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. Truth, lies, and automation: How language models could change disinformation.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Jannis Bulian, Mike S Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Hübscher, Christian Buck, Niels G Mede, Markus Leippold, and Nadine Strauß. 2024. Assessing large language models on climate information. In *41st International Conference on Machine Learning*, Proceedings of Machine Learning Research (PMLR). MLResearch Press.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2025. High-dimension human value representation in large language models. *Preprint*, arXiv:2404.07900.

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. 2024. Locating and mitigating gender bias in large language models. In *International Conference on Intelligent Computing*, pages 471–482. Springer.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Dallas Card, Anjalie Field, Dirk Hovy, and Katherine Keith, editors. 2024. *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*. Association for Computational Linguistics, Mexico City, Mexico.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *J Med Syst*, 47(1):33.

Amanda Cercas Curry, Zeerak Talat, and Dirk Hovy. 2024. Impoverished language technology: The lack of (social) class in NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8675–8682, Torino, Italia. ELRA and ICCL.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

Serina Chang, Ruiqi Zhong, Ethan Adams, Fei-Tzin Lee, Siddharth Varia, Desmond Patton, William Frey, Chris Kedzie, and Kathy McKeown. 2018. Detecting gang-involved escalation on social media using context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Brussels, Belgium. Association for Computational Linguistics.

Alicja Chaszczewicz, Raj Shah, Ryan Louie, Bruce Arnow, Robert Kraut, and Diyi Yang. 2024. Multi-level feedback generation with large language models for empowering novice peer counselors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4130–4161, Bangkok, Thailand. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2024a. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Canyu Chen and Kai Shu. 2024b. Combating misinformation in the age of llms: Opportunities and challenges. *AI Mag.*, 45(3):354–368.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11674–11684.

Guoxin Chen, Kexin Tang, Chao Yang, Fuying Ye, Yu Qiao, and Yiming Qian. 2024. SEER: Facilitating structured reasoning and explanation via reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5901–5921, Bangkok, Thailand. Association for Computational Linguistics.

Lihu Chen, Adam Dejl, and Francesca Toni. 2025. Identifying query-relevant neurons in large language models for long-form texts. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 23595–23604. AAAI Press.

Haocong Cheng, Si Chen, Christopher Perdriau, and Yun Huang. 2024. Llm-powered ai tutors with personas for d/deaf and hard-of-hearing online learners. *arXiv preprint arXiv:2411.09873*.

Robert F Chew, Kirsty J Weitzel, Peter Baumgartner, Caroline W Oppenheimer, Brianna D'Arcangelo, Autumn Barnes, Shirley Liu, Adam Bryant Miller, Ashley Lowe, and Anna C Yaros. 2023. Improving text classification with boolean retrieval for rare categories: A case study identifying firearm violence conversations in the crisis text line database. Technical report, Research Triangle Park (NC).

Eunjung Cho, Won Ik Cho, and Soomin Seo. 2025. Hermit kingdom through the lens of multiple perspectives: A case study of LLM hallucination on North Korea. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3353–3371, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, and 1 others. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification. Kaggle.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge. Kaggle.

Ben Cohen, Moreah Zisquit, Stav Yosef, Doron Friedman, and Kfir Bar. 2024. Motivational interviewing transcripts annotated with global scores. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11642–11657, Torino, Italia. ELRA and ICCL.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020a. STANDER: An expert-annotated dataset for news stance detection and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4086–4101, Online. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020b. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

U.S. Congress. 2023. H.R. 6791 - Artificial Intelligence Literacy Act of 2023. Accessed: 2025-05-15.

Krešimir Čosić, Vanja Kopilaš, and Tanja Jovanovic. 2024. War, emotions, mental health, and artificial intelligence. *Frontiers in Psychology*, 15.

Mihai Croicu and Simon Polichinel von der Maase. 2025. From newswire to nexus: Using text-based actor embeddings and transformer networks to forecast conflict dynamics. *arXiv preprint arXiv:2501.03928*.

Georgina Curto, Svetlana Kiritchenko, Kathleen Fraser, and Isar Nejadgholi. 2024. The crime of being poor: Associations between crime and poverty on social media in eight countries. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 32–45, Mexico City, Mexico. Association for Computational Linguistics.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8386–8411.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023a. The state of human-centered nlp technology for fact-checking. *Information Processing & Management*, 60(2):103219.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023b. Hatemm: A multi-modal dataset for hate video classification. *Preprint*, arXiv:2305.03915.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France. European Language Resources Association.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, and 1 others. 2024. Overview of the multilingual text detoxification task at pan 2024. *Working Notes of CLEF*.

Dorottya Demszky and Heather Hill. 2023. The ncte transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Don't just say "I don't know"! self-aligning large language models for responding to unknown questions with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13652–13673, Miami, Florida, USA. Association for Computational Linguistics.

Jose A. Diaz-Garcia and Joao Paulo Carvalho. 2025. A survey of textual cyber abuse detection using cutting-edge language models and large language models. *Preprint*, arXiv:2501.05443.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*. NeurIPS.

Andreas Dimmelmeier, Hendrik Doll, Malte Schierholz, Emily Kormanyos, Maurice Fehr, Bolei Ma, Jacob Beck, Alexander Fraser, and Frauke Kreuter. 2024. Informing climate risk analysis using textual information - a research agenda. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 12–26, Bangkok, Thailand. Association for Computational Linguistics.

Hua Du, Yanchao Sun, Haozhe Jiang, A. Y. M. Atiquil Islam, and Xiaoqing Gu. 2024. Exploring the effects of AI literacy in teacher learning: an empirical study. *Humanities and Social Sciences Communications*, 11:1–10.

Hamid Reza Ekbia. 2008. *Artificial dreams: The quest for non-biological intelligence*, volume 200. Cambridge University Press Cambridge.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Paloma Moreda, and Manuel Palomar. 2023. Automatic text simplification for people with cognitive disabilities: Resource creation within the ClearText project. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 68–77, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Fairness, Accountability, and Transparency in Machine Learning (FAT/ML). 2020. Principles for Accountable Algorithms. `https://www.fatml.org/resources/principles-for-accountable-algorithms`. Accessed: 2025-05-16.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. 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 3226–3240, Online. Association for Computational Linguistics.

Masoomali Fatehkia, Benjamin Coles, Ferda Ofli, and Ingmar Weber. 2020. The relative value of facebook advertising data for poverty mapping. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 934–938.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Luciano Floridi, Josh Cowls, Thomas C King, and Mariarosaria Taddeo. 2021. How to design ai for social good: Seven essential factors. *Ethics, Governance, and Policies in Artificial Intelligence*, pages 125–151.

Elizabeth Ford, Keegan Curlewis, Akkapon Wongkoblap, and Vasa Curcin. 2019. Public opinions on using social media content to identify users with depression and target mental health care advertising: Mixed methods survey. *JMIR Ment Health*, 6(11):e12942.

Michael Fore, Simranjit Singh, Chaehong Lee, Amritanshu Pandey, Antonios Anastasopoulos, and Dimitrios Stamoulis. 2024. Unlearning climate misinformation in large language models. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 178–192, Bangkok, Thailand. Association for Computational Linguistics.

Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. Argumentative large language models for explainable and contestable claim verification. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 14930–14939. AAAI Press.

Yi R. Fung, Tuhin Chakraborty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2024. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. *Preprint*, arXiv:2210.08604.

Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asuman E. Ozdaglar. 2024. MisinfoEval: Generative AI in the era of "alternative facts". In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8566–8578, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Nupoor Gandhi, Tom Corringham, and Emma Strubell. 2024. Challenges in end-to-end policy extraction from climate action plans. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 156–167, Bangkok, Thailand. Association for Computational Linguistics.

Rujun Gao, Hillary E Merzdorf, Saira Anwar, M Cynthia Hipwell, and Arun R Srinivasa. 2024. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6:100206.

Muskan Garg, Chandni Saxena, V. Gokula Krishnan, Ruchi Joshi, Sriparna Saha, Vijay K. Mago, and B. Dorr. 2022. Cams: An annotated corpus for causal analysis of mental health issues in social media posts. *ArXiv*, abs/2207.04674.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113.

Kai Gehring and Matteo Grigoletto. 2023. Analyzing climate change policy narratives with the character-role narrative framework. Technical report, CESifo Working Paper No. 10429. 85 Pages.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39.

Iacopo Ghinassi, Leonardo Catalano, and Tommaso Colella. 2024. Efficient aspect-based summarization of climate change reports with small language models. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 123–139, Miami, Florida, USA. Association for Computational Linguistics.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Felipe T Giuntini, Mirela T Cazzolato, Maria de Jesus Dutra dos Reis, Andrew T Campbell, Agma JM Traina, and Jo Ueyama. 2020. A review on recognizing depression in social networks: challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing*, 11:4713–4729.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. 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 1596–1611, Online. Association for Computational Linguistics.

Jue Gong, Gregory E. Simon, and Shan Liu. 2019. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PLoS ONE*, 14.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Brigitte Hoyer Gosselink, Kate Brandt, Marian Croak, Karen DeSalvo, Ben Gomes, Lila Ibrahim, Maggie Johnson, Yossi Matias, Ruth Porat, Kent Walker, and James Manyika. 2024. Ai in action: Accelerating progress towards the sustainable development goals. *Preprint*, arXiv:2407.02711.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in nlp. *ACM Comput. Surv.*, 55(14s).

Alessandro De Grandi, Federico Ravenda, Andrea Raballo, and Fabio Crestani. 2024. The emotional spectrum of llms: Leveraging empathy and emotion-based markers for mental health support. *ArXiv*, abs/2412.20068.

Elia Grassini, Marina Buzzi, Barbara Leporini, and Alina Vozna. 2024. A systematic review of chatbots in inclusive healthcare: insights from the last 5 years. *Universal Access in the Information Society*, pages 1–9.

Aynur Guluzade, Naguib Heiba, Zeyd Boukhers, Florim Hamiti, Jahid Hasan Polash, Yehya Mohamad, and Carlos A. Velasco. 2025. ELMTEX: fine-tuning large language models for structured clinical information extraction. A case study on clinical reports. *CoRR*, abs/2502.05638.

Aylin Ece Gunal, Bowen Yi, John D. Piette, Rada Mihalcea, and Veronica Perez-Rosas. 2025. Examining Spanish counseling with MIDAS: a motivational interviewing dataset in Spanish. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 866–872, Albuquerque, New Mexico. Association for Computational Linguistics.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin

Hu. 2024. An investigation of large language models for real-world hate speech detection. *Preprint*, arXiv:2401.03346.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. Association for Computational Linguistics.

Danna Gurari, Qing Li, Andrew Stangl, Chen Guo, Yunfei Lin, Kristen Grauman, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617.

Erick Mendez Guzman, Viktor Schlegel, and Riza Batista-Navarro. 2024. Towards explainable multi-label text classification: A multi-task rationalisation framework for identifying indicators of forced labour. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 98–112, Miami, Florida, USA. Association for Computational Linguistics.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Fatima Haouari, Tamer Elsayed, and Reem Suwaileh. 2024. Overview of the CLEF-2024 checkthat! lab task 5 on rumor verification using evidence from authorities. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 311–320. CEUR-WS.org.

Mohammed Hasanuzzaman, Sabyasachi Kamila, Mandeep Kaur, Sriparna Saha, and Asif Ekbal. 2017. Temporal orientation of tweets for predicting income of users. In *55th Annual Meeting of the Association for Computational Linguistics 2017*, pages 659–665. Association for Computational Linguistics.

Joe Hasell, Bertha Rohenkohl, Pablo Arriagada, Esteban Ortiz-Ospina, and Max Roser. 2024. Poverty. `https://ourworldindata.org/poverty`. Online; accessed 14 May 2025.

Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. 2024. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16:45–74.

Lu He, Changyang He, Tera L Reynolds, Qiushi Bai, Yicong Huang, Chen Li, Kai Zheng, and Yunan Chen. 2021. Why do people oppose mask wearing? a comprehensive analysis of U.S. tweets during the COVID-19 pandemic. *J. Am. Med. Inform. Assoc.*, 28(7):1564–1573.

Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayyed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, and 2 others. 2019. Views: A political violence early-warning system. *Journal of Peace Research*, 56(2):155–174.

Philipp Heinrich, Andreas Blombach, Bao Minh Doan Dang, Leonardo Zilio, Linda Havenstein, Nathan Dykes, Stephanie Evert, and Fabian Schäfer. 2024. Automatic identification of COVID-19-related conspiracy narratives in German telegram channels and chats. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1932–1943, Torino, Italia. ELRA and ICCL.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022a. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022b. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephanie Hirmer, Alycia Leonard, Josephine Tumwesige, and Costanza Conforti. 2021. Building representative corpora from illiterate communities: A review of challenges and mitigation strategies for developing countries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2176–2189, Online. Association for Computational Linguistics.

Lisa Hoeschle, Shuang Liu, and Xiaohua Yu. 2025. Let the poor talk about "poverty": Revisiting poverty alleviation in rural china with machine learning. *Poverty & Public Policy*, 17(2):e70005.

Faye Holder, Sanober Mirza, Namson-Ngo Lee, Jake Carbone, and Ruth E. McKie. 2023. Climate obstruction and facebook advertising: how a sample of climate obstruction organizations use social media to disseminate discourses of delay. *Climatic Change*, 176(2):1–21.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2024. A closer look at the self-verification abilities of large language models in logical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 900–925, Mexico City, Mexico. Association for Computational Linguistics.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. 2023. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4919, Singapore. Association for Computational Linguistics.

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.

Angel Hsu, Mason Laney, Ji Zhang, Diego Manya, and Linda Farczadi. 2024. Evaluating ChatNetZero, an LLM-chatbot to demystify climate pledges. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 82–92, Bangkok, Thailand. Association for Computational Linguistics.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2025. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Guanlan Hu, Mavra Ahmed, and Mary R. L'Abbé. 2023. Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared to traditional methods. *American Journal of Clinical Nutrition*, 117(3):553–563.

Mengke Hu, Ryzen Benson, Annie T. Chen, Shu-Hong Zhu, and Mike Conway. 2021. Determining the prevalence of cannabis, tobacco, and vaping device mentions in online communities using natural language processing. *Drug and Alcohol Dependence*, 228:109016.

Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. 2022. ConfliBERT: A pre-trained language model for political conflict and violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5469–5482, Seattle, United States. Association for Computational Linguistics.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, Andrew Beam, and 1 others. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024a. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.

Hu Huang, Bowen Zhang, Yangyang Li, Baoquan Zhang, Yuxi Sun, Chuyao Luo, and Cheng Peng. 2023. Knowledge-enhanced prompt-tuning for stance detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Tianyi Huang, Jingyuan Yi, Peiyang Yu, and Xiaochuan Xu. 2025. Unmasking digital falsehoods: A comparative analysis of llm-based misinformation detection strategies. *Preprint*, arXiv:2503.00724.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, and 52 others. 2024b. Position: TrustLLM: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20166–20270. PMLR.

Mina Huh, Fangyuan Xu, Yi-Hao Peng, Chongyan Chen, Hansika Murugu, Danna Gurari, Eunsol Choi, and Amy Pavel. 2024. Long-form answers to visual questions from blind and low vision people. *arXiv preprint arXiv:2408.06303*.

Ben Hutchinson and Vinodkumar Prabhakaran. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Oana Ignat, Longju Bai, Joan Nwatu, and Rada Mihalcea. 2024a. Annotations on a budget: Leveraging geo-data similarity to balance model performance and annotation cost. *Preprint*, arXiv:2403.07687.

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Ece Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan C. Nwatu, Veronica Perez-Rosas, Siqi Shen, and 3 others. 2024b. Has it all been solved? open NLP research questions not solved by large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia. ELRA and ICCL.

Andrew Jerfy, Olivia Selden, and Rajesh Balkrishnan. 2024. The growing impact of natural language processing in healthcare and public health. *Inquiry: A Journal of Medical Care Organization, Provision and Financing*, 61:469580241290095.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12186–12215, Bangkok, Thailand. Association for Computational Linguistics.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, and 1 others. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.

Onno P Kampman, Michael Xing, Charmaine Lim, Ahmad Ishqi Jabir, Ryan Louie, Jimmy Lee, and Robert JT Morris. 2025. Conversational self-play for discovering and understanding psychotherapy approaches. *Preprint*, arXiv:2503.16521.

Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani Tur, and Jiawei Han. 2024. Instruct, not assist: LLM-based multi-turn planning and hierarchical questioning for socratic code debugging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9475–9495, Miami, Florida, USA. Association for Computational Linguistics.

George Karystianis, Armita Adily, Peter W Schofield, David Greenberg, Louisa Jorm, Goran Nenadic, and Tony Butler. 2019. Automated analysis of domestic violence police reports to explore abuse types and victim injuries: Text mining study. *J. Med. Internet Res.*, 21(3):e13067.

Krishna Chaitanya Rao Kathala and Shashank Palakurthi. 2025. AI Literacy Framework and Strategies for Implementation in Developing Nations. In *Proceedings of the 2024 16th International Conference on Education Technology and Computers*, ICETC '24, page 418–422, New York, NY, USA. Association for Computing Machinery.

Ashkan Kazemi, Zehua Li, Veronica Perez-Rosas, Scott A. Hale, and Rada Mihalcea. 2022a. Matching tweets with applicable fact-checks across languages. *Preprint*, arXiv:2202.07094.

Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, Scott A. Hale, and Rada Mihalcea. 2022b. Matching tweets with applicable fact-checks across languages. *Preprint*, arXiv:2202.07094.

Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara Jane Ericson, David Weintrop, and Tovi Grossman. 2023. How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli calling international conference on computing education research*, pages 1–12.

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the 2024 chi conference on human factors in computing systems*, pages 1–20.

Monther Khalafat, Ja'far S. Alqatawna, Rizik M. H. Al-Sayyed, Mohammad Eshtay, and Thaeer Kobbaey. 2021. Violence detection over online social networks: An arabic sentiment analysis approach. *International Journal of Interactive Mobile Technologies (iJIM)*, 15(14):pp. 90–110.

Ashar Khan, Mohd Shahid Husain, and Anam Khan. 2018. Analysis of mental state of users using social media to predict depression! a survey. *International Journal of Advanced Research in Computer Science*, 9(2):100–106.

Fahima Khanam, Farha Akhter Munmun, Nadia Afrin Ritu, Aloke Kumar Saha, and Muhammad Firoz. 2022. Text to speech synthesis: A systematic review, deep learning based architecture and future research direction. *Journal of Advances in Information Technology*, 13(5).

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. *Preprint*, arXiv:2404.01247.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023a. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023b. Evaluating the diversity, equity and inclusion of nlp technology: A case study for indian languages. *Preprint*, arXiv:2205.12676.

Zoha Khawaja and Jean-Christophe Bélisle-Pipon. 2023. Your robot therapist is not your therapist: understanding the role of ai-powered mental health chatbots. *Frontiers in Digital Health*, 5.

Parisa Jamadi Khiabani and Arkaitz Zubiaga. 2024. Cross-target stance detection: A survey of techniques, datasets, and challenges. *Preprint*, arXiv:2409.13594.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR.

Dahyun Kim, Sukyung Lee, Yungi Kim, Attapol Rutherford, and Chanjun Park. 2025. Representing the under-represented: Cultural and core capability benchmarks for developing Thai large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4114–4129, Abu Dhabi, UAE. Association for Computational Linguistics.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. MEGAnno+: A human-LLM collaborative annotation system. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–176, St. Julians, Malta. Association for Computational Linguistics.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw multilingual toxic comment classification. `https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification`. Kaggle.

Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats*, 2(2).

Korea Education and Research Information Service. 2025. Welcome message. Accessed: 2025-05-15.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2024. Towards a framework for evaluating explanations in automated fact verification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16364–16377, Torino, Italia. ELRA and ICCL.

Anuj Kumar. 2022. A study: Hate speech and offensive language detection in textual data by using rnn, cnn, lstm and bert model. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1–6.

Yogesh Kumar, Apeksha Koul, and Chamkaur Singh. 2023. A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. *Multimedia Tools and Applications*, 82(10):15171–15197.

Versha Kumari, Khuhed Memon, Burhan Aslam, and Bhawani Shankar Chowdhry. 2023. An effective approach for violence detection using deep learning and natural language processing. In *2023 7th International Multi-Topic ICT Conference (IMTIC)*, pages 1–8.

Remy Kusters, Dusan Misevic, Hugues Berry, Antoine Cully, Yann Le Cunff, Loic Dandoy, Natalia Díaz-Rodríguez, Marion Ficher, Jonathan Grizou, Alice Othmani, and 1 others. 2020. Interdisciplinary research in artificial intelligence: challenges and opportunities. *Frontiers in big data*, 3:577974.

Yerin Kwak and Zachary A Pardos. 2024. Bridging large language model disparities: Skill tagging of multilingual educational content. *British Journal of Educational Technology*, 55(5):2039–2057.

Sunjae Kwon, Xun Wang, Weisong Liu, Emily Druhl, Minhee L Sung, Joel Reisman, Wenjun Li, Robert D. Kerns, William Becker, and Hongfeng Yu. 2023. Odd: A benchmark dataset for the natural language processing based opioid related aberrant behavior detection. *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, 2024:4338–4359.

Vasileios Lampos, Daniel Preoţiuc-Pietro, Sina Samangooei, Douwe Gelling, and Trevor Cohn. 2014. Extracting socioeconomic patterns from the news: Modelling text and outlet importance jointly. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 13–17, Baltimore, MD, USA. Association for Computational Linguistics.

Yu-Ju Lan and Nian-Shing Chen. 2024. Teachers' agency in the era of llm and generative ai. *Educational Technology & Society*, 27(1):I–XVIII.

Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. 2023. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516.

Matthias Carl Laupichler, Alexandra Aster, Marcel Meyerheim, Tobias Raupach, and Marvin Mergen. 2024. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. *BMC Medical Education*, 24.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Markus Leippold, Julia Anna Bingler, Mathias Kraus, and Nicolas Webersinke. 2022. Climatebert: A pretrained language model for climate-related text. In *AAAI Fall Symposium 2022*, Arlington, Virginia.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2025. Automated fact-checking of climate claims with large language models. *npj Climate Action*, 4(1):17.

Sharon Levy, Emily Allaway, Melanie Subbiah, Lydia Chilton, Desmond Patton, Kathleen McKeown, and William Yang Wang. 2022. SafeText: A benchmark for exploring physical safety in language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2407–2421, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jinyu Li and 1 others. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

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 Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification, Outstanding Certification.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 208–221, Turku, Finland. Linköping University Electronic Press.

Shuo-Yu Lin, Xiaolu Cheng, Jun Zhang, Jaya Sindhu Yannam, Andrew J Barnes, J Randy Koch, Rashelle Hayes, Gilbert Gimm, Xiaoquan Zhao, Hemant Purohit, and Hong Xue. 2023. Social media data mining of antitobacco campaign messages: Machine learning analysis of facebook posts. *J Med Internet Res*, 25:e42863.

Joseph Lindquist, Diana M. Thomas, Dusty Turner, Jeanne Blankenship, and Theodore K. Kyle. 2021. Food for thought: A natural language processing analysis of the 2020 dietary guidelines publice comments. *The American Journal of Clinical Nutrition*, 114(2):713–720.

Ruth Lister. 2021. *Poverty*. John Wiley & Sons.

Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee, James Pennebaker, and Rada Mihalcea. 2025a. Eeyore: Realistic depression simulation via supervised and preference optimization. *Preprint*, arXiv:2503.00018.

Zhenhua Liu, Tong Zhu, Chuanyuan Tan, and Wenliang Chen. 2025b. Learning to refuse: Towards mitigating privacy risks in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1683–1698, Abu Dhabi, UAE. Association for Computational Linguistics.

Leo Lo. 2024. Evaluating AI Literacy in Academic Libraries: A Survey Study with a Focus on U.S. Employees. *Coll. Res. Libr.*, 85.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.*

Meir Lotan, Joav Merrick, and Eli Carmeli. 2005. A review of physical activity and well-being. *Int J Adolesc Med Health*, 17(1):23–31.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA. Association for Computational Linguistics.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

Li Lucy, Tal August, Rose E Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. 2024. MathFish: Evaluating language model math reasoning via grounding in educational curricula. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5644–5673, Miami, Florida, USA. Association for Computational Linguistics.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023. Intersectional stereotypes in large language models: Dataset and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *arXiv preprint arXiv:2502.18940*.

Erin MacPhaul, Li Zhou, Stephen J. Mooney, Deborah Azrael, Andrew Bowen, Ali Rowhani-Rahbar, Ravali Yenduri, Catherine Barber, Eric Goralnick, and Matthew Miller. 2023. Classifying firearm injury intent in electronic hospital records using natural language processing. *JAMA Network Open*, 6(4):e235870–e235870.

Nourane Mahdy, Dalia A Magdi, Ahmed Dahroug, and Mohammed Abo Rizka. 2020. Comparative study: different techniques to detect depression using social media. In *Internet of Things—Applications and Future: Proceedings of ITAF 2019*, pages 441–452. Springer.

Long Mai and Julie Carson-Berndsen. 2024. Improving linguistic diversity of large language models with possibility exploration fine-tuning. *arXiv preprint arXiv:2412.03343*.

Matteo Malgaroli, Thomas D. Hull, Jamie M. Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13.

Niklas Mannhardt, Elizabeth Bondi-Kelly, Barbara Lam, Chloe O'Connell, M. Asiedu, Hussein Mozannar, Monica Agrawal, Alejandro Buendia, Tatiana Urman, I. Riaz, Catherine E. Ricciardi, Marzyeh Ghassemi, and David Sontag. 2024. Impact of large language model assistance on patients reading clinical notes: A mixed-methods study. *ArXiv*, abs/2401.09637.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection. *Preprint*, arXiv:2012.10289.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: Towards ai for everyone, with everyone, by everyone. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28657–28670.

Do June Min, Veronica Perez-Rosas, Ken Resnicow, and Rada Mihalcea. 2023. VERVE: Template-based ReflectiVE rewriting for MotiVational IntErviewing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10289–10302, Singapore. Association for Computational Linguistics.

Lokesh Mishra, Sohayl Dhibi, Yusik Kim, Cesar Berrospi Ramis, Shubham Gupta, Michele Dolfi, and Peter Staar. 2024. Statements: Universal information extraction from tables with large language models for ESG KPIs. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 193–214, Bangkok, Thailand. Association for Computational Linguistics.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. 2022. CoVERT: A corpus of fact-checked biomedical COVID-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 244–257, Marseille, France. European Language Resources Association.

Ismael Villegas Molina, Audria Montalvo, Benjamin Ochoa, Paul Denny, and Leo Porter. 2024. Leveraging llm tutoring systems for non-native english speakers in introductory cs courses. *arXiv preprint arXiv:2411.02725*.

Joss Moorkens, Pilar Sánchez-Gijón, Esther Simon, Mireia Urpí, Nora Aranberri, Dragoș Ciobanu, Ana Guerberof-Arenas, Janiça Hackenbuchner, Dorothy Kenny, Ralph Krüger, Miguel Rios, Isabel Ginel, Caroline Rossi, Alina Secară, and Antonio Toral. 2024. Literacy in digital environments and resources (LT-LiDER). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 55–56, Sheffield, UK. European Association for Machine Translation (EAMT).

Chihab Morales, Stefan Klemmer, and Thibault Sellam. 2024. Identifying and improving disability bias in gpt-based resume screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Krishna More and Frason Francis. 2021. Analyzing the impact of domestic violence on social media using natural language processing. In *2021 IEEE Pune Section International Conference (PuneCon)*, pages 1–5.

Gaku Morio and Christopher D Manning. 2023. An nlp benchmark dataset for assessing corporate climate policy engagement. In *Advances in Neural Information Processing Systems*, volume 36, pages 39678–39702. Curran Associates, Inc.

Hannes Mueller, Christopher Rauh, and Ben Seimon. 2024. Introducing a global dataset on conflict forecasts and news topics. *Data & Policy*, 6:e17.

Guberney Muñetón-Santa, Daniel Escobar-Grisales, Felipe Orlando López-Pabón, Paula Andrea Pérez-Toro, and Juan Rafael Orozco-Arroyave. 2022. Classification of poverty condition using natural language processing. *Social Indicators Research*, 162(3):1413–1435.

Guberney Muñetón-Santa and Juan Rafael Orozco-Arroyave. 2023. Identifying dimensions and weighting for poverty and well-being measurements through natural language of people. *Research Square, p. 1-40*.

Simon Munzert, Richard Traunmüller, Pablo Barberá, Andrew Guess, and Junghwan Yang. 2025. Citizen preferences for online hate speech regulation. 4(2):gaf032.

Karim Nader, Paul Toprac, Suzanne Scott, and Samuel Baker. 2022. Public understanding of artificial intelligence through entertainment media. *AI Soc.*, 39(2):1–14.

Ndapandula Nakashole and Tom M. Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore, Maryland. Association for Computational Linguistics.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II*, page 639–649, Berlin, Heidelberg. Springer-Verlag.

Preslav Nakov, Giovanni Da San Martino, Firoj Alam, Shaden Shaar, Hamdy Mubarak, and Nikolay Babulkov. 2022. Overview of the CLEF-2022 checkthat! lab task 2 on detecting previously fact-checked claims. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 393–403. CEUR-WS.org.

Ahmad Nasir, Aadish Sharma, Kokil Jaidka, and Saifuddin Ahmed. 2025. Llms and finetuning: Benchmarking cross-domain performance for hate speech detection. *Preprint*, arXiv:2310.18964.

Deniz Nazarova. 2023. Application of artificial intelligence in mental healthcare: generative pre-trained transformer 3 (gpt-3) and cognitive distortions. In *Proceedings of the Future Technologies Conference*, pages 204–219. Springer.

Apollinaire Poli Nemkova, Sarath Chandra Lingareddy, Sagnik Ray Choudhury, and Mark V Albert. 2025a. Do large language models know conflict? investigating parametric vs. non-parametric knowledge of llms for conflict forecasting. *arXiv preprint arXiv:2505.09852*.

P. A. Nemkova, S. Ubani, and M. V. Albert. 2025b. Comparing llm text annotation skills: A study on human rights violations in social media data. Presented at the AAAI Workshop on AI for Social Impact, 2025 (non-archival).

Poli Nemkova, Solomon Ubani, Suleyman Olcay Polat, Nayeon Kim, and Rodney D. Nielsen. 2023.

Detecting human rights violations on social media during russia-ukraine war. *arXiv preprint arXiv:2306.05370*.

Poli Apollinaire Nemkova, Suleyman Olcay Polat, Rafid Ishrak Jahan, Sagnik Ray Choudhury, Sun joo Lee, Shouryadipta Sarkar, and Mark V. Albert. 2025c. Towards automated situation awareness: A rag-based framework for peacebuilding reports.

Denis Newman-Griffis, Bonnielin Swenor, Rupa Valdez, and Geoffrey Mason. 2024. Disability data futures: Achievable imaginaries for ai and disability data justice. *arXiv preprint arXiv:2411.03885*.

Huy Nghiem and Hal Daumé Iii. 2024. HateCOT: An explanation-enhanced dataset for generalizable offensive speech detection via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5938–5956, Miami, Florida, USA. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917. ACM.

Jingwei Ni, Julia Bingler, Chiara Colesanti-Senni, Mathias Kraus, Glen Gostlow, Tobias Schimanski, Dominik Stammbach, Saeid Ashraf Vaghefi, Qian Wang, Nicolas Webersinke, Tobias Wekhof, Tingyu Yu, and Markus Leippold. 2023. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 21–51, Singapore. Association for Computational Linguistics.

Yizhao Ni, Alycia Bachtel, Katie Nause, and Sarah J. Beal. 2021. Automated detection of substance use information from electronic health records for a pediatric population. *Journal of the American Medical Informatics Association : JAMIA*, 28:2116 – 2127.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.

NIST. 2024. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology (NIST). Available free of charge from the NIST website.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024. A challenge dataset and effective models for conversational stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 122–132, Torino, Italia. ELRA and ICCL.

Joan Nwatu, Oana Ignat, and Rada Mihalcea. 2023. Bridging the digital divide: Performance variation across socio-economic factors in vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10686–10702, Singapore. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

OECD. 2024. Oecd principles on artificial intelligence. `https://www.oecd.org/en/topics/ai-principles.html`. Accessed May 2025.

Chinasa T. Okolo and Hongjin Lin. 2024. "You can't build what you don't understand": Practitioner Perspectives on Explainable AI in the Global South. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

Intergovernmental Panel on Climate Change. 2022. *Climate Change 2022: Impacts, Adaptation, and Vulnerability*. Cambridge University Press, Cambridge, UK and New York, NY, USA.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

A. Paice, M. Biallas, and A. Andrushevich. 2025. Assistive and inclusive technology design for people with disabilities (special needs). In *Human-Technology Interaction: Interdisciplinary Approaches and Perspectives*, pages 329–347. Springer Nature Switzerland, Cham.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Vanessa Panaite, Andrew R. Devendorf, Dezon K. Finch, Lina Bouayad, Stephen L Luther, and Susan K. Schultz. 2022. The value of extracting clinician-recorded affect for advancing clinical research on depression: Proof-of-concept study applying natural language processing to electronic health records. *JMIR Formative Research*, 6.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Chanjun Park, Yoonna Jang, Seolhwa Lee, Jaehyung Seo, Kisu Yang, and Heuiseok Lim. 2022. PicTalky: Augmentative and alternative communication for language developmental disabilities. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 17–27, Taipei, Taiwan. Association for Computational Linguistics.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*.

Ye-Jean Park, Abhinav Pillai, Jiawen Deng, Eddie Guo, Mehul Gupta, Mike Paget, and Christopher Naugler. 2024. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making*, 24(1):72.

Susan T. Parker. 2020. Estimating nonfatal gunshot injury locations with natural language processing and machine learning models. *JAMA Network Open*, 3(10):e2020664–e2020664.

Pat Pataranutaporn, Chayapatr Archiwaranguprok, Samantha W. T. Chan, Elizabeth Loftus, and Pattie Maes. 2025. Slip through the chat: Subtle injection of false information in llm chatbot conversations increases false memory formation. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1297–1313, New York, NY, USA. Association for Computing Machinery.

Sajan B. Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108.

Laura L Paterson and Ian N Gregory. 2018. *Representations of poverty and place: Using geographical text analysis to understand discourse*. Springer.

Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The gun violence database: A new task and data set for NLP. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1024, Austin, Texas. Association for Computational Linguistics.

Sagi Pendzel, Tomer Wullach, Amir Adler, and Einat Minkov. 2023. Generative ai for hate speech detection: Evaluation and findings. *Preprint*, arXiv:2311.09993.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and 1 others. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500, Singapore. Association for Computational Linguistics.

Yash Pilankar, Rejwanul Haque, Mohammed Hasanuzzaman, Paul Stynes, and Pramod Pathak. 2022. Detecting violation of human rights via social media. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 40–45, Marseille, France. European Language Resources Association.

Mitchell Plyler and Min Chi. 2025. Iterative counterfactual data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):19931–19938.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9):e0138717.

Luca Puce, Nicola Luigi Bragazzi, Antonio Currà, and Carlo Trompetto. 2025. Harnessing generative artificial intelligence for exercise and training prescription: Applications and implications in sports and physical activity—a systematic literature review. *Applied Sciences*, 15(7).

Alberto Purpura, Sahil Wadhwa, Jesse Zymet, Akshay Gupta, Andy Luo, Melissa Kazemi Rad, Swapnil Shinde, and Mohammad Shahed Sorower. 2025. Building safe GenAI applications: An end-to-end overview of red teaming for large language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 335–350, Albuquerque, New Mexico. Association for Computational Linguistics.

Vidya Puthenpura, Siddhi Nadkarni, Michael DiLuna, Kimberly Hieftje, and Asher Marks. 2023. Personality changes and staring spells in a 12-Year-Old child: A case report incorporating ChatGPT, a natural language processing tool driven by artificial intelligence (AI). *Cureus*, 15(3).

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Shihao Ran, Di Lu, Aoife Cahill, Joel Tetreault, and Alejandro Jaimes. 2023. A new task and dataset on detecting attacks on human rights defenders. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7089–7113, Toronto, Canada. Association for Computational Linguistics.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Esteban A Ríssola, David E Losada, and Fabio Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, 2(2):1–31.

Naquee Rizwan, Paramananda Bhaskar, Mithun Das, Swadhin Satyaprakash Majhi, Punyajoy Saha, and Animesh Mukherjee. 2025. Exploring the limits of zero shot vision language models for hate meme detection: The vulnerabilities and their interpretations. *Preprint*, arXiv:2402.12198.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Harri Rowlands, Gaku Morio, Dylan Tanner, and Christopher Manning. 2024. Predicting narratives of climate obstruction in social media advertising. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5547–5558, Bangkok, Thailand. Association for Computational Linguistics.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. 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 2116–2129, Online. Association for Computational Linguistics.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M.C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism.

In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024. On zero-shot counterspeech generation by LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12443–12454, Torino, Italia. ELRA and ICCL.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers : Detecting hate speech against women. *Preprint*, arXiv:1812.06700.

Sougata Saha, Saurabh Kumar Pandey, and Monojit Choudhury. 2025. Meta-cultural competence: Climbing the right hill of cultural awareness. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8025–8042, Albuquerque, New Mexico. Association for Computational Linguistics.

Ahmed Shahriar Sakib, Md Saddam Hossain Mukta, Fariha Rowshan Huda, A K M Najmul Islam, Tohedul Islam, and Mohammed Eunus Ali. 2021. Identifying insomnia from social media posts: Psycholinguistic analyses of user tweets. *J Med Internet Res*, 23(12):e27613.

Hind Saleh, Areej Alhothali, and Kawthar Moria. 2021. Detection of hate speech using bert and hate speech word embedding with deep model. *Preprint*, arXiv:2111.01515.

SAMHSA. 2023. What is mental health. `https://www.samhsa.gov/mental-health`. Accessed: 2024-04-10.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. Climatebert-netzero: Detecting and assessing net zero and reduction targets. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 15745–15756. Association for Computational Linguistics.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.

Michael Sejr Schlichtkrull. 2024. Generating media background checks for automated source critical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4927–4947, Miami, Florida, USA. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM*, 63(12):54–63.

Walelign Tewabe Sewunetie, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Hellina Hailu Nigatu, Gashaw Kidanu, Zewdie Mossie, Hussien Seid, Eshete Derb, and Seid Muhie Yimam. 2024. Evaluating gender bias in machine translation for low-resource languages. In *5th Workshop on African Natural Language Processing*.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 82–92, Online. Association for Computational Linguistics.

Zahra Shakeri Hossein Abad, Gregory P Butler, Wendy Thompson, and Joon Lee. 2022. Physical activity, sedentary behavior, and sleep on twitter: Multicountry and fully labeled public data set for digital public health surveillance research. *JMIR Public Health Surveill*, 8(2):e32355.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *Preprint*, arXiv:1711.08536.

Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9977–10000, Toronto, Canada. Association for Computational Linguistics.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Med Inform*, 7(2):e12239.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. *Preprint*, arXiv:2405.04655.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Yanyan Shen and Wencheng Cui. 2024. Perceived support and AI literacy: the mediating role of psychological needs satisfaction. *Frontiers in Psychology*, 15.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 395–405, New York, NY, USA. Association for Computing Machinery.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *CoRR*, abs/1809.01286.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Anne J. Sietsma, Rick W. Groenendijk, and Robbert Biesbroek. 2023. Progress on climate action: a multilingual machine learning analysis of the global stocktake. *Climatic Change*, 176(12):173.

Ruba Skaik and Diana Inkpen. 2020. Using social media for mental health surveillance: a review. *ACM Computing Surveys (CSUR)*, 53(6):1–31.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C. Park. 2018. Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. 2021. Classification aware neural topic model for covid-19 disinformation categorisation. *PLOS ONE*, 16:e0247086.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *International Conference on Artificial Intelligence in Education*, pages 32–43. Springer.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *arXiv preprint arXiv:1906.00591*.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. From show to tell: A survey on deep learning-based image captioning. *arXiv preprint arXiv:2107.06912*.

Colm Sweeney, Courtney Potts, Edel Ennis, Raymond Bond, Maurice D. Mulvenna, Siobhan O'neill, Martin Malcolm, Lauri Kuosmanen, Catrine Kostenius, Alex Vakaloudis, Gavin Mcconvey, Robin Turkington, David Hanna, Heidi Nieminen, Anna-Kaisa Vartiainen, Alison Robertson, and Michael F. Mctear. 2021. Can chatbots help support a person's mental health? perceptions and views from mental healthcare professionals and experts. *ACM Trans. Comput. Healthcare*, 2(3).

Tangfei Tao, Yizhe Zhao, Tianyu Liu, and Jieli Zhu. 2024. Sign language recognition: A comprehensive review of traditional and deep learning approaches, datasets, and challenges. *IEEE Access*, PP:1–1.

Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K. Keita, and Michael Leventhal. 2025. GAIfE: Using GenAI to improve literacy in low-resourced settings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7914–7929, Albuquerque, New Mexico. Association for Computational Linguistics.

Matthew Tassava, Cameron Kolodjski, Jordan Milbrath, Adorah Bishop, Nathan Flanders, Robbie Fetsch, Danielle Hanson, and Jeremy Straub. 2024. Development of an ai anti-bullying system using large language model key topic detection. *Preprint*, arXiv:2408.10417.

G Tejesh, RM Pruthvi, Pavitra Gonal, Rashmi Karchi, and 1 others. 2025. Multilingual braille and voice translation model for visually impaired learners. In *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, volume 3, pages 1–6. IEEE.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, and 7 others. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *Preprint*, arXiv:2401.09646.

Isabelle Tingzon, Ardie Orden, Kevin Thomas Go, Stephanie Sy, Vedran Sekara, Ingmar Weber, Masoomali Fatehkia, Manuel García-Herranz, and D Kim. 2019. Mapping poverty in the philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:425–431.

George Tolkachev, Stephen Mell, Stephan Zdancewic, and Osbert Bastani. 2022. Counterfactual explanations for natural language interfaces. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 113–118, Dublin, Ireland. Association for Computational Linguistics.

Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, and 1 others. 2020. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1):2468.

Fangziyun Tong, Reeva Lederman, Simon D'Alfonso, Katherine Berry, and Sandra Bucci. 2023. Conceptualizing the digital therapeutic alliance in the context of fully automated mental health apps: A thematic analysis. *Clinical Psychology & Psychotherapy*, 30(5):998–1012.

UNICEF. 2020. Producing disability-inclusive data: Why it matters and what it takes. Technical report, United Nations Children's Fund, New York, NY, USA. Accessed: 2025-05-16.

Aziza Usmanova, Ahmed Aziz, Dilshodjon Rakhmonov, and Walid Osamy. 2022. Utilities of artificial intelligence in poverty prediction: a review. *Sustainability*, 14(21):14238.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4(1):480.

Marieke van Erp, Christian Reynolds, Diana Maynard, Alain Starke, Rebeca Ibáñez Martín, Frederic Andres, Maria C. A. Leite, Damien Alvarez de Toledo, Ximena Schmidt Rivera, Christoph Trattner, Steven Brewer, Carla Adriano Martins, Alana Kluczkovski, Angelina Frankowska, Sarah Bridle, Renata Bertazzi Levy, Fernanda Rauber, Jacqueline Tereza da Silva, and Ulbe Bosma. 2021. Using natural language processing and artificial intelligence to explore the nutrition and sustainability of recipes and food. *Frontiers in Artificial Intelligence*, Volume 3 - 2020.

Francesco Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. Climatext: A dataset for climate change topic detection. In *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*. NeurIPS.

P. N. Venkit, M. Srinath, and S. Wilson. 2025. A study of implicit language model bias against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Mike Wald. 2021. Ai data-driven personalisation and disability inclusion. *Frontiers in artificial intelligence*, 3:571955.

Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024a. Intent-aware and hate-mitigating counterspeech generation via dual-discriminator guided LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9131–9142, Torino, Italia. ELRA and ICCL.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

Rose Wang and Dorottya Demszky. 2024. Edu-ConvoKit: An open-source library for education conversation data. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 61–69, Mexico City, Mexico. Association for Computational Linguistics.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024b. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.

Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, and Dora Demszky. 2024c. Tutor CoPilot: A human-AI approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.

Ru Wang, Zach Potter, Yun Ho, Daniel Killough, Linxiu Zeng, Sanbrita Mondal, and Yuhang Zhao. 2024d. Gazeprompt: Enhancing low vision people's reading experience with gaze-aware augmentations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Neha Warikoo, Tobias Mayer, Dana Atzil-Slonim, Amir Eliassaf, Shira Haimovitz, and Iryna Gurevych. 2022. Nlp meets psychotherapy: Using predicted client emotions and self-reported client emotions to measure emotional coherence. *ArXiv*, abs/2211.12512.

Greta Warren, Irina Shklovski, and Isabelle Augenstein. 2025. Show me the work: Fact-checkers' requirements for explainable automated fact-checking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. 2024. AI for education (AI4EDU): Advancing personalized education with LLM and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6743–6744.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *Preprint*, arXiv:2410.12705.

World Bank Group. 2024. Poverty, Prosperity, and Planet Report: Pathways Out of the Polycrisis. https://www.worldbank.org/en/publication/poverty-prosperity-and-planet. Online; accessed 14 May 2025.

Jinge Wu, Rowena Smith, and Honghan Wu. 2022. Ontology-driven self-supervision for adverse childhood experiences identification using social media datasets. *ArXiv*, abs/2208.11701.

Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. 2016. Transfer learning from deep features for remote sensing and poverty mapping. *Preprint*, arXiv:1510.00098.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024a. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024b. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605, Bangkok, Thailand. Association for Computational Linguistics.

Chenghao Yang, Tuhin Chakrabarty, Karli R. Hochstatter, Melissa N Slavin, Nabila El-Bassel, and Smaranda Muresan. 2023a. Identifying self-disclosures of use, misuse and addiction in community-based social media posts. *ArXiv*, abs/2311.09066.

Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Y. Yang, Y. Zhang, D. Sun, and 1 others. 2025. Navigating the landscape of ai literacy education: insights from a decade of research (2014–2024). *Humanities and Social Sciences Communications*, 12(1):374.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023b. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023c. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

Seid Muhie Yimam, Daryna Dementieva, Tim Fischer, Daniil Moskovskiy, Naquee Rizwan, Punyajoy Saha, Sarthak Roy, Martin Semmann, Alexander Panchenko, Chris Biemann, and Animesh Mukherjee. 2024. Demarked: A strategy for enhanced abusive speech moderation through counterspeech, detoxification, and message management. *Preprint*, arXiv:2406.19543.

Kayo Yin, Amir Moryossef, Jami Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. 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 7347–7360. Association for Computational Linguistics.

Renzhe Yu, Zhen Xu, Sky CH-Wang, and Richard Arum. 2024. Whose chatgpt? unveiling real-world educational inequalities introduced by large language models. *arXiv preprint arXiv:2410.22282*.

Zeping Yu and Sophia Ananiadou. 2024. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3293–3306, Miami, Florida, USA. Association for Computational Linguistics.

Yunhao Yuan, Koustuv Saha, Barbara Keller, Erkki Isometsä, and Talayeh Aledavood. 2023. Mental health coping stories on social media: A causal-inference study of papageno effect. In *Proceedings of the ACM Web Conference*.

Youngsik Yun and Jihie Kim. 2024. Cic: A framework for culturally-aware image captioning. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page

1625–1633. International Joint Conferences on Artificial Intelligence Organization.

Anwar Hossain Zahid, Monoshi Kumar Roy, and Swarna Das. 2025. Evaluation of hate speech detection using large language models and geographical contextualization. *Preprint*, arXiv:2502.19612.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024. Do we need language-specific fact-checking models? the case of Chinese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1899–1914, Miami, Florida, USA. Association for Computational Linguistics.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2025. Simulating classroom education with LLM-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023a. C-STANCE: A large dataset for Chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385, Toronto, Canada. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Runcong Zhao, Miguel Arana-catania, Lixing Zhu, Elena Kochkina, Lin Gui, Arkaitz Zubiaga, Rob Procter, Maria Liakata, and Yulan He. 2023b. PANACEA: An automated misinformation detection system on COVID-19. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 67–74, Dubrovnik, Croatia. Association for Computational Linguistics.

Haiqi Zhou, David Hobson, Derek Ruths, and Andrew Piper. 2024. Large scale narrative messaging around climate change: A cross-cultural comparison. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 143–155, Bangkok, Thailand. Association for Computational Linguistics.

Jinfeng Zhou, Yuxuan Chen, Jianing Yin, Yongkang Huang, Yihan Shi, Xikun Zhang, Libiao Peng, Rongsheng Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. Crisp: Cognitive restructuring of negative thoughts through multi-turn supportive dialogues. *Preprint*, arXiv:2504.17238.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

Zhengyuan Zhu, Zeyu Zhang, Haiqi Zhang, and Chengkai Li. 2025. RATSD: Retrieval augmented truthfulness stance detection from social media posts toward factual claims. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3366–3381, Albuquerque, New Mexico. Association for Computational Linguistics.

Tianyi Zhuang, Chuqiao Kuang, Xiaoguang Li, Yihua Teng, Jihao Wu, Yasheng Wang, and Lifeng Shang. 2025. Docpuzzle: A process-aware benchmark for evaluating realistic long-context reasoning capabilities. *Preprint*, arXiv:2502.17807.

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.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2021. Explainable depression detection with multi-modalities using a hybrid deep learning model on social media. *Preprint*, arXiv:2007.02847.

A Paper Selection: Thematic Tables of Datasets, NLP Tasks, Evaluation Metrics, and References

A.1 Health and Well-being

Mental health is a key component of human health, encompassing our emotional, psychological, and social well-being (SAMHSA, 2023). Currently, mental health issues are a multifactorial global crisis complicated by individual risk factors and various socioeconomic and clinical factors, but NLP methods show promising potential to enhance mental healthcare (Zhang et al., 2022). In this context, we define NLP tasks as activities performed by NLP techniques in roles similar to counselors and clients. In the role of **counselors**, NLP tools engage in several core tasks: (1) **Detection and classification** of mental health conditions, such as depression (Giuntini et al., 2020; Mahdy et al., 2020; Khan et al., 2018) and addiction (Yang et al., 2023a; Kwon et al., 2023; Ni et al., 2021), using data sources like clinical notes (Panaite et al., 2022; Calvo et al., 2017) and social media posts (Skaik and Inkpen, 2020; Chancellor and De Choudhury, 2020; Ríssola et al., 2021); (2) **Responding** to users by interpreting their emotional states (Warikoo et al., 2022; Sabour et al., 2024; Grandi et al., 2024), generating therapeutic and empathetic responses (Shen et al., 2020; Grandi et al., 2024; Sharma et al., 2023; Nazarova, 2023; Zhou et al., 2025), and providing actionable feedback for support quality (Min et al., 2023; Chaszczewicz et al., 2024; Althoff et al., 2016); (3) **Tracking** emotion and mood via time-series data analysis (Čosić et al., 2024) and detecting mental health crises over time (Gong et al., 2019; Yuan et al., 2023). Conversely, when NLP tools serve as **clients**, they typically simulate client personas from diverse backgrounds to train counselors (Louie et al., 2024; Liu et al., 2025a; Hsu et al., 2025; Kampman et al., 2025). The literature for this section was selected based on the existing survey papers such as (Zhang et al., 2022; Malgaroli et al., 2023) as well as keyword search (e.g., "AI for mental health", "LLM-based mental health applications", "mental health chatbots", "AI therapy").

Physical well-being refers to maintaining one's bodily health through behaviors such as regular physical activity, adequate sleep, balanced nutrition, good hygiene, and avoidance of harmful substances (Lotan et al., 2005). To stimulate discussion on the role of NLP in physical well-being, the works included in the paper were selected based on a targeted keyword search including "physical activity", "sleep", "nutrition", "hygiene", "substance use", and "clinical report analysis", which also plays a central role in physical health. NLP techniques have been applied to various aspects of physical well-being, leveraging unstructured text data to monitor behaviors and inform interventions. For instance, physical activity and sedentary behavior can be tracked through analysis of social media using NLP-driven health surveillance systems that mine Twitter posts to estimate physical activity levels, sedentary behavior, and sleep patterns in populations (Sakib et al., 2021; Shakeri Hossein Abad et al., 2022). For nutrition, NLP has been employed to assess dietary habits by using models that can automatically classify foods and meals from descriptions, even providing personalized diet advice (van Erp et al., 2021; Hu et al., 2023). Regarding hygiene, during the COVID-19 pandemic NLP was used to estimate public perceptions of mask-wearing and other hygiene practices by mining social media posts (Al-Garadi et al., 2022b). Likewise, in harmful habit avoidance, NLP methods help identify substance use patterns and risky behaviors from online text (Hu et al., 2021; Lin et al., 2023). The analysis of clinical reports also plays a central role in physical health. These documents capture critical patient information that is often not represented in structured data fields. Consequently, clinical report analysis has emerged as a core subtask within NLP for physical health, enabling the extraction, classification, and summarization of medically relevant information directly from unstructured clinical narratives (Landolsi et al., 2023). In this research area there is a growing role of LLMs in extracting information from clinical reports. The work by (Mannhardt et al., 2024) shows how GPT-4 can support patients by simplifying clinical notes, improving comprehension and confidence, though with some factual inaccuracies. In (Guluzade et al., 2025), the authors introduce a large annotated dataset (ELMTEX) and find that fine-tuned small LLMs outperform larger ones in extracting structured information efficiently. These findings are confirmed in pathology reports, where fine-tuned models achieve higher accuracy and fewer hallucinations than prompt-based methods (Park et al., 2024). LLMs can also be used in clinical reports to generate and summarize documentation such as patient notes, discharge summaries, and case reports, offering improvements in efficiency, organization, and standardization of medical writing (Park et al., 2024; Ali et al., 2023; Patel and Lam, 2023; Cascella et al., 2023). They can help identify grammar errors and inconsistencies in extracted data (e.g., lab values), thereby potentially reducing documentation errors (Ali et al., 2023). These applications may alleviate administrative burdens on healthcare professionals, allowing more time for direct patient care (Lee et al., 2023). Nonetheless, the performance of LLMs is limited by variability in accuracy depending on case complexity, the risk of generating incorrect or fabricated content (hallucinations), and susceptibility to user framing, emphasizing the need for careful prompt design and human oversight (Puthenpura et al., 2023).

Datasets	**NLP Task(s)**	**Evaluation Metrics**	**Reference**
Expert-annotated opioid-related posts on Reddit	Detecting addiction	Accuracy, macro-F1	Yang et al. (2023a)
Motivational Interviewing (MI) dataset	Generating therapeutic dialogues	ROUGE, embedding-based metrics (greedy matching, embedding average, and vector extrema), ratio of distinct n-grams, human annotator evaluation	Shen et al. (2020)
Empathetic Dialogues Dataset, Reddit Mental Health Dataset, DailyDialog Dataset	Generating empathetic dialogues for mental health support	BERTSCORE, accuracy, precision, recall, F-1	Grandi et al. (2024)
MI-TAGS	Publicly available mental health dataset	Accuracy, macro F-1, ROC AUC	(Cohen et al., 2024)
Reddit Self-reported Depression Diagno- sis dataset	Interpretable NLP models in mental health	Precision, recall, F-1	(Song et al., 2018)
Depression dataset, Non-depression dataset, Depression-candidate dataset	Interpretable NLP models in mental health	Precision, recall, F1, accuracy	(Zogan et al., 2021)
CAMS	Causal Analysis of Mental health issue	Accuracy	(Garg et al., 2022)
Dataset of principles	Simulating patient personas	Consistency with Context, speech style, Principle Adherence	Louie et al. (2024)
Publicly available depression-related conversations (RED, HOPE, ESC, AnnoMI-Full), expert-annotated preferences	Simulating patient personas	Expert evaluation on contrast with AI-like responses, linguistic authenticity, cognitive pattern authenticity, subtle emotional expression, profile adherence and personalization. Automatic evaluation on symptom severity, cognitive distortion, and overall depression Severity	Liu et al. (2025a)
MIDAS	Publicly available dataset in mental health counseling	Expert evaluation, reflection to question ratio, accuracy, F-1	Gunal et al. (2025)
Reddit Mental Health Dataset	Publicly available mental health dataset on social media	Recall, precision, F-1	Wu et al. (2022)
Longitudinal Patient Health Questionnaire	Tracking user mood or mental health crises	Spearman's rank-order correlation, mean squared error	Gong et al. (2019)
FeedbackESConv	Providing feedback to counselors	Automatically-computed quality scores, domain experts	(Chaszczewicz et al., 2024)
PAIR, AnnoMI	Providing feedback to counselors	Edit effect (reflection score), content preservation, perplexity, coherence, specificity	(Min et al., 2023)
Anonymized counseling conversations from a NGO	Providing feedback to counselors	Adaptability, dealing with ambiguity, creativity, making progress, change in perspective	(Althoff et al., 2016)
Annotated clinical notes	Predicting and understanding mental health outcomes	Accuracy	(Panaite et al., 2022)
Thought Records Dataset, Mental Health America	Responding to users' negative thoughts	Automatic (BLEU, ROUGE-1, ROUGE-L, BertScore); Human (Relatability, Helpfulness)	

Insomnia data set consisting from Twitter	Text classification, Correlation analysis betwen language use and insomnia, Topic modeling	True-positive rate, False-positive rate, AUC	(Sakib et al., 2021)
Twitter corpus (LPHEADA) labeled for relevance to physical activity, sedentary behavior, and sleep	Text classification, Semantic consistency evaluation, location inference	Precision, Recall, F1 Score, AUC-ROC, Average precision	(Shakeri Hossein Abad et al., 2022)
Various online recipe databases, both structured and unstructured: recipe websites, historical recipe archives, nutritional databases, sustainability data	NER, Information extraction, semantic linking, recommender system	Qualitative analysis	(van Erp et al., 2021)
Food Label Information and Price (FLIP) Database	Text classification, Regression, t-SNE visualization	Accuracy, Precision, Recall, F1-score, MSE	(Hu et al., 2023)
Electronic health records, social media platforms (Twitter, Reddit, Facebook, YouTube), scientific literature, news and web sources	Information Extraction, Health Behavior Analysis, Early outbreak detection, Misinformation detection, Question Answering	Accuracy, Precision, Recall, F1-score (for classification), AUC-ROC, MSE (for regression)	(Al-Garadi et al., 2022b)
Posts manually labeled with six annotation categories from Reddit	NER, WSD, Sequence labeling, Social media text analysis	Precision, Recall, F1-score	(Hu et al., 2021)
Facebook posts from antitobacco campaigns	Sentiment analysis, Topic modeling, Text classification	Odds ratios, Agreement rate	(Lin et al., 2023)
Hypothetical clinical scenarios related to skin cancer	Text generation, Readability assessment	Readability score, Likert ratings	(Ali et al., 2023)
Simulation of real-world clinical and research use cases	Clinical note generation, Detection of potential misuse, Language style adaptation	Qualitative analysis	(Cascella et al., 2023)
Real-world pathology reports	Information extraction, Hallucination detection	Accuracy	(Park et al., 2024)
12 clinical notes: 4 synthetic and 8 real	Text simplification, Definition extraction, FAQ generation, Information extraction, Prompt engineering	Quantitative evaluation of a survey, Readability score, Qualitative interviews	(Mannhardt et al., 2024)
ELMTEX Dataset (clinical summaries)	Information extraction, Entity normalization	ROGUE, BERTScore, Precision, Recall, F1	(Guluzade et al., 2025)

Table 1: Overview of datasets, NLP tasks, evaluation metrics, and references from healthcare-related studies.

A.2 Education

NLP for Education Tools & Systems. The integration of NLP systems into educational settings has garnered significant attention, particularly with the widespread adoption of LLMs by students. These NLP-based educational applications, such as intelligent tutoring systems, offer the potential to deliver personalized, high-quality education to underserved regions and populations. Specifically, these systems aim to enhance learning experiences by providing timely and personalized support to both teachers and students, including the following tasks: personalized and/or curriculum-aligned question generation (Kargupta et al., 2024; Lucy et al., 2024), scaffolded dialogue tutoring(Kazemitabaar et al., 2024), adaptive knowledge tracing (Kargupta et al., 2024), automated feedback (Jurenka et al., 2024), teacher coaching (Wang and Demszky, 2023), and student simulation for testing classroom policies/activities (Zhang et al., 2025).

Methodologies. Intelligent tutoring systems primarily focused on (1) modeling teacher-student and student-student interactions using transcripts, (2) devising knowledge state spaces for specific domains/problems to trace student knowledge throughout an interaction, and (3) generating single-turn responses to students. With the emergence of LLMs, recent methodologies have expanded upon these tasks to include:

1. **Multi-Turn Socratic Dialogue and Planning.** Recent methodologies leverage large language models (LLMs) to engage students in multi-turn Socratic dialogues, promoting critical thinking and problem-solving without directly providing answers. For instance, *TreeInstruct* employs a state space-based planning algorithm to dynamically construct question trees based on student responses, effectively guiding learners through multi-turn code debugging tasks (Kargupta et al., 2024). Similarly, the *Socratic Questioning of Novice Debuggers* dataset benchmarks LLMs' abilities to employ Socratic methods in assisting novice programmers through single-turn interactions (Al-Hossami et al., 2023).

2. **Expert Decision Modeling.** To emulate expert tutoring behaviors, some works model the decision-making processes of experienced educators. *Bridging the Novice-Expert Gap* utilizes cognitive task analysis to capture experts' identification of student errors, remediation strategies, and instructional intentions, informing LLM responses in math tutoring scenarios (Wang et al., 2024b).

3. **Curriculum-Aligned Evaluation.** Evaluating LLMs' mathematical reasoning has shifted towards alignment with educational curricula. *MathFish* assesses whether models can identify and apply specific math skills and concepts as outlined in standardized curricula, using publisher-labeled data from open educational resources (Lucy et al., 2024).

4. **Open-Ended Pedagogical Benchmarking.** To assess LLMs' instructional capabilities beyond problem-solving, *MathTutorBench* introduces a benchmark evaluating open-ended pedagogical skills. It measures models' abilities across various educational tasks, emphasizing the quality of instructional interactions (Macina et al., 2025).

5. **Simulated Student Interactions.** Datasets like *MathDial* are created by pairing human teachers with LLMs simulating student behavior, generating rich pedagogical dialogues. This approach aids in training and evaluating models on realistic tutoring scenarios (Macina et al., 2023).

6. **Classroom Discourse Analysis.** Large-scale datasets such as the *NCTE Transcripts* provide insights into teacher-student interactions. These transcripts, annotated for dialogic discourse moves, help in analyzing effective instructional practices and inform the development of NLP tools for education (Demszky and Hill, 2023).

7. **Educational Conversation Toolkits.** Open-source frameworks like *Edu-ConvoKit* facilitate the analysis of educational conversations by offering tools for preprocessing, annotation, and analysis tailored to educational research needs (Wang and Demszky, 2024).

We have included an overview of the various dataset resources and their corresponding tasks and evaluation metrics in Table 2. These papers have been collected based on recent prominence and relevance to different challenges and opportunities present within the NLP Education space.

Datasets	NLP Task(s)	Evaluation Metrics	Reference
MathDial	Text-to-Text Generation (Tutoring Response Generation)	sBLEU, BERTScore, KF1, Uptake, Success@k, Telling@k, Human Evaluation (Coherence, Correctness, Equitable tutoring)	(Macina et al., 2023)
MULTI-DEBUG	Text-to-Text Generation, Text Classification (Socratic Question Generation, Multi-Turn Planning)	Relevance, Indirectness, Logical Flow, Overall Success Rate, Average # of Turns	(Kargupta et al., 2024)
Bridge	Text Classification, Text-to-Text Generation (Remediation of Math Mistakes, Decision-Making Modeling)	Human Evaluation (usefulness, care, human-soundingness, preference), Log Odds Ratio	(Wang et al., 2024b)
MathFish	Text Classification, Topic Modelling, Text-to-Text Generation (Math Reasoning Evaluation, Curriculum Alignment)	Weak Accuracy, Exact Accuracy	(Lucy et al., 2024)
MathTutorBench	Text-to-Text Generation (Evaluation of Pedagogical Capabilities in LLM Tutors)	Accuracy, BLEU, F1, Win Rate (Pedagogical skill metrics)	(Macina et al., 2025)
National Center for Teacher Effectiveness (NCTE)	Text Classification (Evaluation of Classifying Educational Discourse Features)	Accuracy, Precision, Recall, F1	(Demszky and Hill, 2023)

Table 2: Overview of datasets, NLP tasks, evaluation metrics, and references from NLP for Education studies.

A.2.1 AI Literacy

To identify relevant literature on AI literacy, we used Semantic Scholar and Google Scholar search terms such as “ai literacy” and (“ai literacy” + “social impact” + “nlp”). We found most of the papers selected for this topic to be on classroom-based studies and measurements and metrics for AI literacy, along with some interdisciplinary papers connecting AI literacy to other disciplines, for example, psychology.

A.3 Peace Building

We have organized highlighted papers found in our review of papers at the intersection of peace building and NLP in Table 3.

To identify the most relevant literature on human rights violation detection using NLP and on conflict prediction, we employed three search strategies: querying the ACL Anthology, conducting searches on Google Scholar, and utilizing the Consensus research discovery platform. The keywords used included: “human rights,” “human rights violations detection,” “armed conflict prediction,” and “conflict forecasting.”

The selected works on physical safety were identified through a keyword search in ACL and Google Scholar for papers on “physical safety”, “domestic violence”, “gun violence” and “firearm injury”. The Google Scholar search also included the keyword “nlp”. We did not include papers that were more focused on the mental health implications of physical safety (e.g. suicide via firearms), or on larger organizational peace-building efforts (e.g. terrorism and police brutality).

We find most of these tasks in this topic to be focused on document classification of rare events. As a result, primary evaluation metrics are precision, recall, and F1. There are a few papers that apply unsupervised tasks like topic modeling and generation that use coherence and similarity, respectively, as their primary metrics.

A.4 Poverty

We began our review of papers focused on the study of poverty by through keyword searches such as “poverty detection" on Google Scholar. However, the majority of these studies addressed this issue through the use of satellite imagery (Tingzon et al., 2019; Ayush et al., 2020), or audience estimates from advertising platforms (Fatehkia et al., 2020). From these sets of papers, we included a review of the use of AI for poverty prediction (Usmanova et al., 2022). We then modified our keyword searches on Google

Datasets	NLP Task(s)	Evaluation Metrics	Reference
Crisis Text Line Database	Document Classification (detect firearm injury or violence)	Precision, Recall, Accuracy	(Chew et al., 2023)
National Violent Death Report System	Topic Modeling (characterize trends in violent deaths)	Coherence, Topic diversity, Coverage	(Arseniev-Koehler et al., 2022)
National Electronic Injury Surveillance System Series	Document Classification (classify location of nonfatal gunshot injuries)	Accuracy, Precision, Recall, AUC	(Parker, 2020)
SafeText (Reddit)	Generation (generating advice)	Similarity, Confidence, Perplexity, Accuracy	(Levy et al., 2022)
Surveillance Cameras	Speech to Text, Document Classification (Detect violence)	Precision, Recall, Accuracy, Loss	(Kumari et al., 2023)
Twitter	Document Classification (detect violence related tweets)	Precision, Recall, Accuracy	(ALSaif and Alotaibi, 2019)
Police Reports	Document Classification (Classify abuse type and victim injuries)	Precision, Recall, F1	(Karystianis et al., 2019)
Twitter	Document Classification (detect intimate partner violence)	Accuracy, F1	(Al-Garadi et al., 2022a)
Electronic Health Records	Document Classification (classify firearm injury intent)	Precision, Recall, F1 (MacPhaul et al., 2023)	
Gun Violence Database (deprecated compilation of news articles)	Entity Extraction (identify event details like participants, roles, location, time)	Precision, Recall	(Pavlick et al., 2016)
Twitter	POS tagging, Machine translation, Sentiment Analysis, Document Classification (detect aggression and loss)	Precision, Recall, F1 (Blevins et al., 2016)	
Twitter	Document Classification (detect aggression, loss, and substance use)	Precision, Recall, F1, Average Precision	(Blandfort et al., 2019)
Twitter	Document Classification (detect aggression and loss)	Precision, Recall, F1 (Chang et al., 2018)	
Electronic Health Records	Document Classification (If a patient will become violent)	F1, Confusion Matrices	(Borger et al., 2022)
Electronic Health Record	Document Classification (type of violence and patient status)	Precision, Recall, F1	(Botelle et al., 2022)

Table 3: Overview of Peace Building and Physical Safety related NLP studies

Scholar to “nlp income", “nlp poverty", and “text poverty classification" to ensure a focus on studies from the NLP domain. For papers related to our task, we would investigate studies that have cited them as well. We have highlighted a number of the most prominent papers found in our review in Table 4.

The current NLP literature on the topic of poverty and class is rather limited (Cercas Curry et al., 2024). A literature review published in 2024, only found 20 NLP papers that investigate socio-economic status in any capacity (Cercas Curry et al., 2024). Another review of artificial intelligence systems for the detection of poverty (Usmanova et al., 2022), conducted in 2022, identified 22 papers, only one of which used NLP models (Muñetón-Santa et al., 2022). Additionally, most of the datasets are not publicly available, hindering reproduction and progress in this domain.

A.5 Online Harms

Initial research primarily focused on developing classification frameworks across different spectrum of online harms based on existing AI models. Starting from traditional ML tools (Saha et al., 2018), further works exploited LSTMS/RNNs (Kumar, 2022), and with the advent of *Transformers* based models (like BERT (Saleh et al., 2021)), research rapidly unfolded with the increased development of frameworks

Datasets	NLP Task(s)	Evaluation Metrics	Reference
Twitter	Yearly Income Prediction	Pearson correlation, Mean Average Error	(Preoţiuc-Pietro et al., 2015b)
Twitter	Income Prediction, Temporal Orientation Classification	Accuracy, Precision, Recall, F1, MAE	(Hasanuzzaman et al., 2017)
Interview Transcripts	Poor or Extremely Poor Classification	Accuracy, Specificity, Sensitivity, F1	(Muñetón-Santa et al., 2022)
News	ESI Prediction, Unemployment Prediction	RMSE	(Lampos et al., 2014)

Table 4: Overview of Poverty related NLP studies

brewed on top of attention mechanism. Recent advancements of LLMs has lead to works using their generation capability (Guo et al., 2024; Tassava et al., 2024; Pendzel et al., 2023); where some works also incorporate infamous LLM strategies like zero-shot (Roy et al., 2023) and few-shot (Zahid et al., 2025) promptings, and fine-tuning (Nasir et al., 2025). For explainability, (Mathew et al., 2022) presented one of the first works that proposed one-hot vector representation to improve attention based models and recently, reasoning based explanation and interpretation frameworks (Yang et al., 2023b; Nirmal et al., 2024) to provide more contextual information to LLMs have garnered attention.

Apart from traditional classification based mitigation, a rapid shift towards proactive content moderation leveraging the generative capabilities of LLMs has been proposed. Numerous works, especially focused on two strategies– counter speech generation (Bonaldi et al., 2024; Saha et al., 2024; Wang et al., 2024a) and text detoxification (Dementieva et al., 2025; Dale et al., 2021) have been extensively explored. LLMs have proven to be sufficiently good at these tasks, but need further improvements for multilingual performance (Dementieva et al., 2024). Some recent works have further proposed the effectiveness of strong few-shot capabilities of LLMs for annotation of such complex datasets which can potentially reduce crowd-sourcing efforts (Bhat and Varma, 2023; Kim et al., 2024). Further, studies on curating multimodal datasets (Kiela et al., 2021) and understanding the strengths and limitations of multimodal LLMs have also garnered attention (Rizwan et al., 2025). We have highlighted significantly important datasets and studies on online harms in Table 5. These papers are shortlisted on the basis of the impact they drive thus aiding in the detection and mitigation of online harms through necessary content moderation strategies.

A.6 Misinformation

Methodology. Automated fact-checking process comprises the verification of claims - verifiable factual statements (Panchendrarajan and Zubiaga, 2024). Four major components of the fact-checking pipeline (Das et al., 2023a) are widely studied (Vlachos and Riedel, 2014; Thorne and Vlachos, 2018; Barrón-Cedeño et al., 2020; Guo et al., 2022) and include: (1) claim detection, checkworthiness, and prioritisation (based on their urgency or potential harm / impact) to identify claims from news and social media that should be processed given the limited human and automated fact-checking resources (Konstantinovskiy et al., 2021; Nakov et al., 2021; Abumansour and Zubiaga, 2023), often treated as a classification task; (2) evidence retrieval to collect trustworthy evidence for a claim (Thorne et al., 2018a; Augenstein et al., 2019); (3) veracity prediction based on this evidence; and, finally, (4) explanation of the outcome label for humans (Shu et al., 2019; Kotonya and Toni, 2020a; Atanasova et al., 2020; Lu and Li, 2020): summarizing the evidence, generating explanations and evaluating them. The existing datasets align with the fact-checking stages. They are included in Table 6. Other close tasks from the automated misinformation detection field, such as stance detection, rumor detection and fake news detection based on linguistic features are also included in this Table. In addition to this, there also more domain-specific datasets for misinformation detection - for example, multiple datasets were collected on COVID-19 topic (Abdul-Mageed et al., 2021; Song et al., 2021; Mohr et al., 2022; Heinrich et al., 2024), including a multilingual dataset for a shared task to predict fact-checking options for claims, including their verifiability and potential harm (Shaar et al., 2021).

Misinformation can be spread in various languages, but most datasets are still in English. Multilingual verification can use translation systems, but datasets in specific languages and multilingual datasets are still needed to train and evaluate monolingual and multilingual models - for example, comprehensive multilingual multitopic claim detection datasets (Panchendrarajan and Zubiaga, 2024) (despite the recent efforts e.g. in (Nakov et al., 2021; Kazemi et al., 2022a; Pikuliak et al., 2023)).

For our review, we firstly focused on surveys on automated misinformation detection (Guo et al., 2022; Das et al., 2023a; Panchendrarajan and Zubiaga, 2024; Khiabani and Zubiaga, 2024; Huang et al., 2025; Chen and Shu, 2024b; Oshikawa et al., 2020; Zhou and Zafarani, 2020), then studied papers and approaches mentioned.

A.7 Inequalities and Bias

To curate a representative set of datasets and evaluation strategies for gender bias in NLP, we selected papers that span a wide spectrum of model architectures (from static embeddings to transformer-based LMs) and task settings (e.g., coreference resolution, occupation classification, translation, multi-agent interactions). The selection emphasizes both foundational work and recent advancements that shaped current methodologies. Early studies such as those by (Bolukbasi et al., 2016) and (Caliskan et al., 2017) were included for their role in establishing intrinsic bias probing techniques like WEAT. We also incorporated task-specific evaluations such as pronoun-drop metrics (Stanovsky et al., 2019) and fairness gaps in classification tasks (De-Arteaga et al., 2019). Recent papers were chosen to highlight emerging directions in LLM-based analysis, including causal interventions (Cai et al., 2024), region-aware bias evaluations (Borah et al., 2025), and multi-agent propagation frameworks (Borah and Mihalcea, 2024). Collectively, these works offer a diverse yet focused lens into how gender bias manifests and is measured in modern NLP systems.

We present a subset of representative datasets and evaluation benchmarks for cultural bias in Table 8.

To identify relevant datasets and benchmark efforts addressing the needs of people in underrepresented communities, like people with accessibility needs, we reviewed domain-specific surveys and high-impact papers (e.g., based on citation count and venue of publication). To focus on specific disability dataset, the literature was retrieved using search queries such as *"NLP accessibility datasets"*, *"speech recognition for dysarthria"*, *"text simplification benchmark"*, and *"sign language translation dataset"* in Google Scholar and ACL Anthology. When disability-specific datasets were not available, we included in the table the most widely used datasets for the corresponding NLP task. We present a representative collection of works in Table 9.

A.8 Environmental Harms

Given the absence of a survey paper in this field, we began our review with research presented at the inaugural https://aclanthology.org/events/climatenlp-2024/. Additionally, we have included follow-up studies conducted by researchers within the ClimateNLP community. A significant portion of the work in this domain focuses on classifying climate-related claims, text, or stances. Detecting misinformation has emerged as a prominent topic, often tackled via external source verification and question-answering approaches. Information extraction, whether related to quantitative features or narrative insights, also plays a key role. The table 10 below highlights the key papers associated with these tasks.

Datasets	**NLP Task(s)**	**Evaluation Metrics**	**Reference**
Twitter	Multilingual and Multi-Aspect Hate speech detection spanning different target communities	Micro-F1 and Macro-F1 score	(Ousidhoum et al., 2019)
Twitter + Gab (HateXplain)	Hate, Offensive and Normal speech detection with rationales	**Classification** - Accuracy, Macro-F1 score, AUROC score; **Rationales** - IOU-F1 score, token-F1 score, AUPRC score *(Plausibility)* and comprehensiveness, sufficiency *(Faithfulness)*	(Mathew et al., 2022)
LLM generated explanations	Explainable hate speech detection with step-by-step reasoning generated by LLMs	Accuracy and F1 score	(Yang et al., 2023b)
Hate-COT	Offensive speech label explanation generated by GPT-3.5 Turbo	F1 score, Persuasiveness and Soundness	(Nghiem and Daumé Iii, 2024)
Latent Hatered	Detection of implicit hate speech	Precision, Recall, F1 score and Accuracy	(ElSherief et al., 2021)
Jigsaw toxicity datasets	Detection of different types of toxicity across multiple labels with corresponding severity score	Overall AUROC and Bias AUROC	(cjadams et al., 2017, 2019; Kivlichan et al., 2020)
Measuring hate speech (Comments from YouTube, Reddit and Twitter)	Rasch Measurement Theory (RMT) based continuous scoring of hate speech across multiple labels and targets	Hate speech score, difficulty of survey item and response, severity of rater	(Sachdeva et al., 2022)
CONAN and its variants	Generation of counter speech against hate speech through different NLP generation strategies	Semantic Similarity, Novelty, Diversity, Toxicity, Politeness, Intent Accuracy and Hate Mitigation	(Chung et al., 2019; Fanton et al., 2021; Bonaldi et al., 2022; Gupta et al., 2023)
Toxic instances from Jigsaw, Reddit and Twitter	Toxicity classifier and generation of detoxified speech for toxic instances	Accuracy, Fluency, Similarity and Joint score	(Logacheva et al., 2022)
Multilingual text detoxification dataset from multiple sources	Multilingual text detoxification with explanation	Style Transfer Accuracy (STA), Fluency (ChrF1), Content Similarity and Joint score	(Dementieva et al., 2025)
Facebook + Twitter memes	Hate, harm and misogyny detection in memes using different multimodal applications of NLP	Accuracy, Macro-F1 score and AUROC score	(Kiela et al., 2021; Pramanick et al., 2021; Fersini et al., 2022)
BitChute videos	Hateful videos classification at the intersection of NLP, vision and audio	Accuracy, Macro-F1 score, Precision and Recall	(Das et al., 2023b)

Table 5: Representative datasets, NLP task(s) and evaluation metrics from research on online harm.

Datasets	NLP Task(s)	Evaluation Metrics	Reference
Emergent	stance classification	accuracy, per-class precision and recall	(Ferreira and Vlachos, 2016)
Multi-Target Stance Dataset	stance classification	macro-averaged F1 score	(Sobhani et al., 2017)
PHEME	rumor detection and verification; stance classification	macro F1 score, accuracy	(Kochkina et al., 2018)
RumourEval 2019	stance towards a rumor: classification; veracity prediction: classification	macro F1 score; macro F1 score, RMSE	(Gorrell et al., 2019)
VAST	stance classification	macro-averaged F1 score	(Allaway and McKeown, 2020)
Will-They-Won't-They	stance classification for rumor verification	macro F1 score, unweighted avg F1, weighted avg F1	(Conforti et al., 2020b)
STANDER	stance classification; evidence retrieval	macro-averaged precision, recall and F1 score; precision@5 and recall@5	(Conforti et al., 2020a)
COVID-19-Stance	classification	accuracy, macro average precision, recall, F1 score	(Glandt et al., 2021)
P-Stance	stance classification	F avg, macro-average of F1 score	(Li et al., 2021)
ISD	stance detection classification	micro average F1 score	(Huang et al., 2023)
C-STANCE	stance classification	F1 scores for 3 classes and and F1 macro	(Zhao et al., 2023a)
MT-CSD	stance classification	F avg	(Niu et al., 2024)
TSD-CT	stance classification	F1 scores for each class and macro F1 score	(Zhu et al., 2025)
LIAR	fake news: classification	accuracy	(Wang, 2017)
FakeNewsAMT and Celebrity	fake news: classification	accuracy, precision, recall, and F1 score	(Pérez-Rosas et al., 2018)
Stance-annotated Reddit dataset	rumor stance and veracity prediction: classification	accuracy, F1 score	(Lillie et al., 2019)
Twitter-based dataset	classification	accuracy, average precision, ROC, F1 micro, F1 macro scores	(Volkova et al., 2017)
Claim detection dataset	claim detection: classification	precision, recall and F1 score	(Konstantinovskiy et al., 2021)
MultiFC	Claim verification: classification; evidence ranking	micro F1, macro F1	(Augenstein et al., 2019)
Snopes-based dataset	stance classification; evidence extraction: ranking; claim validation: classification	precision, recall and F1 macro; precision @5 and recall @5; macro presicion, recall and F1	(Hanselowski et al., 2019)
CLIMATE-FEVER	claim verification: retrieval, ranking and classification	accuracy	(Diggelmann et al., 2020)
SciFact	claim verification: retrieval and classification	precision, recall, F1 score	(Wadden et al., 2020)
PUBHEALTH	veracity prediction: classification; explanation generation	precision, recall, F1 macro, accuracy; ROUGE and coherence	(Kotonya and Toni, 2020b)
COVID-Fact	evidence retrieval, claim verification: classification	COVID-FEVER Score (similar to FEVER score)	(Saakyan et al., 2021)
X-Fact	claim verification: classification	F1 score	(Gupta and Srikumar, 2021)
FakeNewsNet	claim verification: classification	precision, recall, accuracy, F1 score	(Shu et al., 2018)
FEVER	claim verification: retrieval, ranking and classification	accuracy, F1 score; FEVER score (includes evidence retrieval and claim labels)	(Thorne et al., 2018a,b)
FEVEROUS	claim verification: retrieval, ranking and classification	FEVEROUS score (includes evidence retrieval and claim labels)	(Aly et al., 2021)
Multilingual claim matching dataset	claim matching: retrieval and classification	MAP@k, MRR and F1 score, accuracy	(Kazemi et al., 2022b)
CLEF-2022 CheckThat! Task 2	claim matching: ranking	MAP, reciprocal rank, Precision@k, MAP@5	(Nakov et al., 2022)
MultiClaim	claim matching: ranking	S@10	(Pikuliak et al., 2023)
NLP4IF-2021	claim detection: classification	precision, recall, F1 score	(Shaar et al., 2021)
CLEF-2022 CheckThat! Task 1	verifiable claim detection: classification	F1 score, accuracy, weighted F1	(Nakov et al., 2021)
CLEF-2024 CheckThat! Task 5	evidence retrieval; rumor classification	MAP and Recall@5; F1 macro and strict F1 macro	(Haouari et al., 2024)

Table 6: Datasets on misinformation detection.

Datasets	**NLP Task(s)**	**Key Evaluation Metrics**	**Reference**
GloVe, Word2Vec embeddings on Google corpus	Intrinsic bias probing	WEAT, gender–direction cosine	Bolukbasi et al. (2016); Caliskan et al. (2017)
WinoBias	Coreference resolution	F1 / precision gap (Female vs Male)	Zhao et al. (2018)
WinoMT	Machine translation (pronoun gender)	Gender accuracy, error rate	Stanovsky et al. (2019)
BiosBias (LinkedIn biographies)	Occupation classification	F1 diff., error disparity	De-Arteaga et al. (2019)
LLM-Agent Interaction Logs	Multi-agent bias propagation	Bias-Score, fairness gap	Borah and Mihalcea (2024)
BiosBias+GPT-J	Causal weight-editing for debiasing	Stereotype score, accuracy	Cai et al. (2024)
TED-Talk / IWSLTEn–It MT	Pronoun-drop attribution maps	Pronoun-drop rate, TER gap	Attanasio et al. (2023)
Synthetic counterfactual pairs (Iter CDA)	Bias-robust text classification	Bias amplification ratio, F1	Plyler and Chi (2025)
GeoWAC, Reddit, UN General Debates	Region-aware bias evaluation metric	Region-aware WEAT effect size, mismatch%	Borah et al. (2025)
CrowS-Pairs (intersectional ext.)	Intrinsic bias evaluation	Bias Score (Black vs White)	Guo and Caliskan (2021)

Table 7: Representative datasets and evaluation practices across gender-bias NLP tasks. These were some influencial datasets and papers in gender bias in NLP, covering the topics of bias detection and bias mitigation methods across static embeddings, dynamic embeddings, transformer-based LMs, and LLMs

Datasets	**NLP Task(s)**	**Key Evaluation Metrics**	**Reference**
GeoDE, GD-VCR, CVQA	Multimodal Captioning; Multi-Agent Collaboration	Alignment score; Completeness score; Cultural Info metric	(Bai et al., 2025)
XNLI, PAWS-X	Cross-lingual NLI; Paraphrase Detection	Per-culture accuracy gaps	(Hershcovich et al., 2022a)
World Values Survey (WVS-7)	Survey-Response Prediction; Alignment Measurement	Similarity scores; Alignment gap per group	(AlKhamissi et al., 2024)
CultureBank TikTok, CultureBank Reddit	Cultural QA; Zero-Shot QA; Fine-Tuning	QA accuracy improvements; Agreement levels	(Shi et al., 2024)
NormAd-Eti	Norm Classification (acceptable vs not)	Model accuracy vs human on explicit/abstract norms	(Rao et al., 2025)
GD-VCR	Culturally-Aware Captioning	Human eval (cultural descriptiveness ratings)	(Yun and Kim, 2024)
C4 web crawl	Commonsense Extraction; Classification; Clustering	Crowdsourced plausibility (PLA, COM, DIS); QA priming gains	(Nguyen et al., 2023)
Dollar Street	Zero-Shot Image–Text Alignment	Median CLIP score by income quartile; Spearman ρ	(Nwatu et al., 2023)
Universal Dependencies treebanks; SIGMORPHON; WMT news translation; XNLI cross-lingual NLI; TyDi QA/ SQuAD	Parsing; Inflection; MT; TTS; NLI; QA	Scaled performance utility; Global utility metrics	(Blasi et al., 2021)
ImageNet	Image Classification by Country	Accuracy gaps (US/EU vs developing regions)	(Shankar et al., 2017)
WikiAnn; Universal Dependencies treebanks; XNLI; TyDI QA/ChAII	NER; POS; NLI; QA	Utility $\times$ Demand; Gini coefficient; Throughput and memory	(Khanuja et al., 2023b)
Google Street View images; American Community Survey data; Voting precinct results	Vehicle Detection; Attribute Classification; Demographic Regression	Correlation vs ACS; Voting prediction accuracy	(Gebru et al., 2017)
ImageNet; NOAA Nighttime Lights; Google Static Maps satellite imagery	Proxy Task (Night-Light Prediction); Poverty Regression	Survey correlation vs LSMS; MAE	(Xie et al., 2016)
Gold annotations from Amazon Mechanical Turk for English, Hindi, Italian, Portuguese; Wikipedia corpora	Temporal Grounding	Hour-range accuracy vs gold annotations	(Shwartz, 2022)
FORK test set	Commonsense QA (Culinary)	Accuracy on US vs non-US probes; Statistical significance	(Palta and Rudinger, 2023)
TV show dialogues; Cross-culture shows; LDC conversational corpora	Norm Extraction; Self-Verification; Grounding	AUC for grounding; human-judged best-norm selection	(Fung et al., 2024)
Reddit corpus of 61,981 users; Word association benchmarks	Demographic-Conditioned LM; Word Association	Perplexity; word-association accuracy	(Welch et al., 2020)
GeoMLAMA; FORK; CANDLE; DLAMA	Cultural Commonsense QA; Country Prediction; Commonsense Verification	Accuracy gaps; uniformity analysis	(Shen et al., 2024)
Concept and Application dataset	Image Transcreation; Cultural Adaptation	Human evaluation (relevance; meaning preservation)	(Khanuja et al., 2024)
WorldCuisines	Multilingual VQA (Dish Prediction; Origin Prediction)	Accuracy; adversarial context drop	(Winata et al., 2025)
CVQA	Multilingual Visual QA	Accuracy; answer-matching metrics; performance drop analysis	(Romero et al., 2024)
LAION, GeoDE, DollarStreet	Annotation Suggestion; Data Selection	Annotation cost vs quality; coverage of cultural features	(Ignat et al., 2024a)
Value-relevant outputs from 8 LLMs; Reference human value distributions from surveys	Value Alignment Probing; Distribution Mapping	Correlation with survey ground truth	(Cahyawijaya et al., 2025)

Table 8: Representative datasets and evaluation practices across cultural bias NLP tasks.

Datasets	NLP Task	Evaluation Metrics	Reference
Newsela, WikiLarge, WikiSmall, ASSET, MUSS, SimplicityDA, Simple Wiki	Text Simplification	SARI, FKGL, BLEU, BERTScore, Human Ratings	Al-Thanyyan and Azmi (2021)
LJSpeech, VCTK, LibriTTS, Blizzard Challenge, HiFi-TTS, M-AILABS, CSS10, AISHELL	Text-to-Speech	MOS (Mean Opinion Score), Intelligibility Score, Naturalness, Word Error Rate (WER), MCD (Mel-Cepstral Distortion)	Khanam et al. (2022); Kumar et al. (2023)
EasyCall, UASpeech, TORGO, CSLU Dysarthric, DEED, L2-ARCTIC, Google Project Euphonia	Speech Recognition	WER (Word Error Rate), CER (Character Error Rate), Accuracy, Intelligibility, Real-time Factor (RTF)	Alharbi et al. (2021); Li et al. (2022)
DSBI, Smart Braille Converter Corpus, Tamil-Braille Dataset	Braille Processing	Accuracy, Precision, Recall, BLEU (for translation), OCR Error Rate	Ali (2023)
MS COCO, VizWiz, Multi30K, Flickr8k, Flickr30k, STAIR Captions, TextCaps, OpenSubtitles	Image Captioning and Subtitling	BLEU, METEOR, ROUGE, CIDEr, SPICE, Human Ratings (fluency, adequacy), Caption Accuracy	Stefanini et al. (2021); Ghandi et al. (2023)
VizWiz VQA, TDIUC, VQA-Med, OK-VQA, GQA, TextVQA, DocVQA, OViQA, PathVQA,	Question Answering	Accuracy, VQA Score, BLEU, ANLS (Average Normalized Levenshtein Similarity), Human Ratings	Gurari et al. (2018); Chen et al. (2022); Huh et al. (2024)
PHOENIX14T, RWTH-PHOENIX-Weather, CSL-Daily, RWTH-BOSTON-104, ASLG-PC12, OpenASL, RWTH-SLT, Sign2Text, How2Sign, AUTSL	Sign Language	BLEU, ROUGE, METEOR, WER, Sign Error Rate (SER), Gloss Accuracy, Human Ratings	Yin et al. (2021); Tao et al. (2024).

Table 9: Representative datasets and evaluation practices across accessibility-related NLP tasks.

Datasets	NLP Task(s)	Evaluation Metrics	Reference
Global Stocktake Dataset from Climate Policy Radar	Topic modelling	Cosine similarity	(Sietsma et al., 2023)
SumIPCC: topic-annotated summaries and relative paragraphs from climate change reports	Aspect-based summarisation	Mean Reciprocal Rank (MRR), Carburacy-reweighted ROUGE score	(Ghinassi et al., 2024)
ClimateFever: real-world climate climate change claims with evidence sentences from Wikipedia	Text classification	Label-accuracy, F1, precision, recall	(Diggelmann et al., 2020)
TCFD-category labeled sentences from firms' annual reports, sustainability-, climate- or TCFD-reports and firms' webpage	Text classification	Accuracy	(Bingler et al., 2022)
CORP: paragraphs from common news, research articles and climate reporting of companies	Language modelling, text classification, sentiment analysis, fact-checking	Average cross-entropy loss, average validation loss, weighted F1 score	(Leippold et al., 2022)
Reduction target claims collected by Net Zero Tracker	Text classification	Accuracy, precision, recall, F1 score	(Schimanski et al., 2023)
HLEG reports, Net Zero Stocktake reports, Corporate Climate Responsibility Monitor Reports from NewClimate Institute	Q&A, text generation	Expert evaluation scores for quality, factual accuracy, relevance	(Hsu et al., 2024)
Corpus created from curated sources compiled by ERASMUS.AI (Pretraining), collection of scientific reports and papers (RAG), ClimaBench (Downstream tasks): collection of climate-related datasets for classification	Language modelling, Q&A, various downstream tasks (classification etc.)	Cross-entropy loss, average validation loss, weighted F1-score, BLEU scores, human evaluation	(Thulke et al., 2024)
Sixth Assessment Reports (AR6) of IPCC	Q&A, text generation	Accuracy score given by experts	(Vaghefi et al., 2023)
Text from IPCC, WMO, AbsCC (climate change abstracts), 1000S (abstracts by top 1000 climate scientists)	Text classification	Averaged micro-F1 score with different classification levels	(Leippold et al., 2025)
SemTabNet: tables from over 10K corporate ESG reports obtained using Deep Search toolkit	Information extraction	Tree Similarity Score	(Mishra et al., 2024)
Corporate annual and sustainability reports	Quantitative information extraction	Custom report-level metrics evaluating retrieval and accuracy of extractions	(Dimmelmeier et al., 2024)
ClimateQA: climate claims	Q&A	Manual inspection, ROUGE-L recall score, conditional probability, truth ratio, GPT-Match, GPT-Contradiction, AlignScore	(Fore et al., 2024)
Climate change-related questions from Google Trends, Skeptical Science, synthetic questions from English Wikipedia	Evaluating LLMs' responses	Presentational (style, clarity, correctness, tone) and epistemological (accuracy, specificity, completeness and uncertainty) properties	(Bulian et al., 2024)
Corporate sustainability reports	Text summarization, Text scoring, Q&A	Human evaluation for hallucinations, ROUGE precision score	(Ni et al., 2023)
Climate-related questions (question-source-answer pairs), sustainability report dataset (report-paragraph-question pairs)	Information extraction	Recall@K, Precision@K, F1@K for different top K values	(Schimanski et al., 2024)
Dataset of corporate climate policy engagement documents collected by LobbyMap	Information extraction (query, stance and evidence page indices), classification	Strict F-score, page overlap F-score, document F-score,	(Morio and Manning, 2023)
Facebook ads related to climate change (oil and gas sector), text and spend, impressions, demographic and regional distribution	Multi-label classification	Overall and sub-category specific F-score	(Holder et al., 2023; Rowlands et al., 2024)
News articles in English and Mandarin	Information extraction (narrative features)	Human evaluation, ROUGE-1, ROUGE-L, cosine similarity	(Zhou et al., 2024)

Table 10: Selection of influential datasets and papers in the ClimateNLP domain, ranging from topic modelling over text classification to question answering and information extraction

B Global Goals

We include an overview of the Sustainable Development Goals (SDGs) in Figure 2, which apply to many of the NLP applications discussed in this paper.

Figure 2: Overview of the SDG goals. Source: https://sdgs.un.org/goals

C Global Risks

We present the key global risks categorized by domain, as outlined in the *Global Risks Report 2025* by the World Economic Forum.[7] Each domain-specific table in Figure 3 highlights major threats along with their definitions. We also provide the global risks ranked by severity over the short and long term in Figure 4.

[7] https://www.weforum.org/publications/global-risks-report-2025

ENVIRONMENTAL	
Biodiversity loss and ecosystem collapse	Severe consequences for the environment, humankind and economic activity due to destruction of natural capital.
Critical change to Earth systems	Long-term, potentially irreversible changes to climate and ecological systems at regional or global level.
Extreme weather events (floods, heatwaves, etc.)	Loss of life and property due to events like wildfires, floods, or heatwaves exacerbated by climate change.
Natural resource shortages (food, water)	Supply shortages of food or water for humans, industries or ecosystems.
Non-weather-related natural disasters	Earthquakes, tsunamis, volcanoes, and space events like solar flares causing loss and disruption.
Pollution (air, soil, water, etc.)	Harmful materials introduced into air, water or soil due to human activity causing health and ecological damage.

SOCIETAL	
Decline in health and well-being	Regular or chronic impacts on physical and mental health and well-being that require substantive medical attention and/or limit activities of daily living.
Erosion of human rights and/or civic freedoms	Loss of protections for rights inherent to all human beings, regardless of individual status, and/or the freedoms that underpin civic space.
Inequality (wealth, income)	Present or perceived substantive disparities in the distribution of assets, wealth or income within or between countries.
Infectious diseases	Spread of viruses, parasites, fungi or bacteria leading to a widespread loss of life and economic disruption.
Insufficient public infrastructure and social protections	Non-existent, inadequate or inequitable public infrastructure, services and social protections.
Lack of economic opportunity or unemployment	Structural deterioration of work prospects or standards of work and/or persistent barriers to the realization of economic potential and security.
Involuntary migration or displacement	Forced movement or displacement across or within borders stemming from discrimination, disaster, conflict, or economic hardship.
Societal polarization	Ideological and cultural divisions within and across communities leading to social instability and economic or political disruption.

GEOPOLITICAL	
State-based armed conflict (proxy, civil wars, coups, terrorism, etc.)	Use of force between states or between state and non-state actors, manifesting as war and/or organized, sustained violence.
Biological, chemical or nuclear weapons or hazards	Intentional or accidental release of biological, chemical, nuclear or radiological hazards.
Geoeconomic confrontation (sanctions, tariffs, investment screening)	Use of economic tools by global powers to reshape interactions and constrain geopolitical rivals.
Intrastate violence (riots, mass shootings, gang violence, etc.)	Violence within a country or community that results in loss of life, injury or property damage, including gang violence and gender-based violence.

ECONOMIC	
Asset bubble burst	Prices of key assets become disconnected from the real economy and collapse.
Concentration of strategic resources and technologies	Control over critical resources or technologies by a few actors that manipulate access or pricing.
Crime and illicit economic activity (incl. cyber)	Global spread of illegal business activities undermining economies (e.g. trafficking, cybercrime, fraud).
Debt (public, corporate, household)	Unsustainable debt loads leading to bankruptcy, insolvency or sovereign crises.
Disruptions to a systemically important supply chain	Collapse of essential supply chains causing shocks to goods, markets or services.
Disruptions to critical infrastructure	Shut down of digital or physical infrastructure due to attacks or disasters.
Economic downturn (recession, stagnation)	Extended period of zero or negative economic growth.
Inflation	Sustained rise in prices eroding purchasing power.
Talent and/or labour shortages	Mismatch between labor demand and skilled supply across regions or industries.

TECHNOLOGICAL	
Adverse outcomes of AI technologies	Intended or unintended negative consequences of advances in AI and related technological capabilities (including Generative AI) on individuals, businesses, ecosystems and/or economies.
Adverse outcomes of frontier technologies (quantum, biotech, geoengineering)	Intended or unintended negative consequences of advances in frontier technologies on individuals, businesses, ecosystems and/or economies. Includes, but is not limited to: brain-computer interfaces, biotechnology, geoengineering and quantum computing.
Censorship and surveillance	Broad and pervasive observation of a place or person and/or suppression of communication, information and ideas, physically or digitally, to the extent that it significantly infringes on human and civil rights (e.g. privacy, freedom of speech and freedom of expression).
Cyber espionage and warfare	Use of cyber weapons and tools by state and non-state actors to gain control over a digital presence, cause operational disruption, and/or compromise or damage an entity's technological and information networks and infrastructure. Includes: defensive and offensive cyber operations that occur during or trigger armed conflict, and cyberattacks that steal classified, sensitive data or intellectual property to gain an advantage.
Misinformation and disinformation	Persistent false information (deliberate or otherwise) widely spread through media networks, shifting public opinion in a significant way towards distrust in facts and authority. Includes, but is not limited to: false, imposter, manipulated and fabricated content.
Online harms	Erosion of protection from and/or prevalence of harmful behaviour that poses a digital threat to the emotional or mental health and well-being of individuals. Includes, but is

Figure 3: Domain-specific global risks according to the *Global Risks Report 2025*. Each table illustrates key risks in societal, technological, geopolitical, environmental, and economic domains respectively.

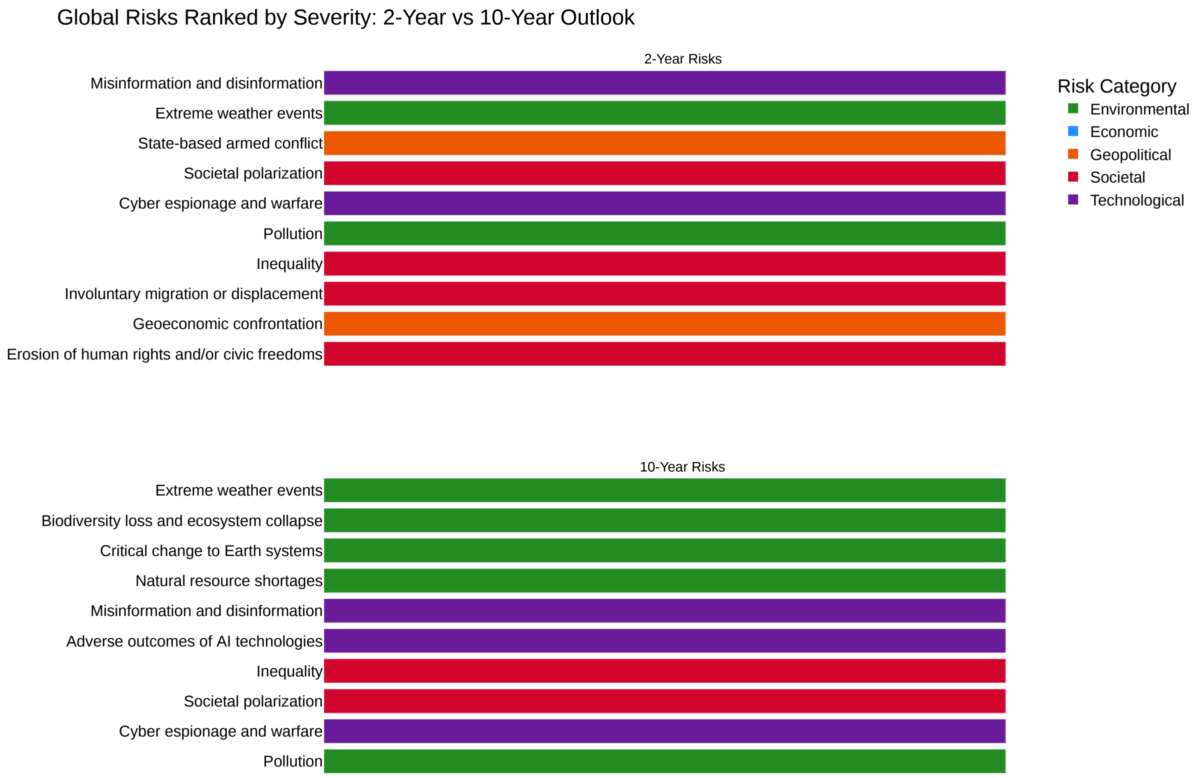

Figure 4: Global risks ranked by severity over the short and long term. Reproduction of Figure FIGURE C from: the Global Risks Report 2025

.

D Guidelines for authors

This work involved contributions from many collaborators with diverse backgrounds, who surveyed various topics and guided the writing process. After extensive discussions in the project's early phase, we developed the author guidelines shown in Figure 5. We share these here for transparency reasons and to assist other researchers undertaking similar multidisciplinary efforts.

Project Guidelines

Your task is to write a section (or subsection) based on a topic of your expertise. To ensure consistency across all sections, we will first collaboratively collect key papers, identify the needs they address, and outline the NLP methodologies they apply. This information, along with our notes, will be organised in a shared spreadsheet. Once the research is mapped out, we will proceed to draft the sections, following a consistent structure that highlights both current opportunities and existing challenges in the field.
Please work on your chosen topic tab and fill in the shared [link]spreadsheet. Below are the main suggestions to help guide your input:

- **Spreadsheet Columns:**
  - *Main Field Papers* – Select key papers in the area. We recommend starting from potential existing surveys in the field to identify key papers. If no surveys exist, we recommend using a keyword search and manual filtering of the most influential papers using the number of citations if needed.
  - *Social Needs Covered* – Societal challenges addressed. We need this information for Section 2. So, please update the column in your tab accordingly, even if you don't use it in your section.
  - *Popular Datasets* – Commonly used datasets
  - *NLP Task(s)* – How tasks are defined (generation, classification, etc.)
  - *NLP Methodology (Existing)* – Methods used in literature
  - *Evaluation* – Evaluation setup and metrics
  - *Limitations* – Limitations of current work
  - *Challenges / Open Questions* – Remaining gaps or issues
  - *Expected NLP Impact / Suggestions* – Potential contributions and ideas
  - *NLP Methodology Potential* – Methodological insights or improvements
- **Suggested structure for your section:**
  - *Methodology:* – Dataset, approach, evaluation
  - *Limitations:* – Challenges, Critical analysis, and Future Work
  - *Opportunities:* – Suggestions, Open Questions, Impact, Broader relevance and Impact in NLP
- **Suggested length for your topic-section:**
  - *We aim for three main discussions per topic: methodology, limitations, and opportunities.*
  - *Please use a table for the methodology section that will be added in the appendix.*
  - *We can aim for a maximum of 4 paragraphs for your chosen topic: general intro, methodology paragraph (going to the appendix), challenges, opportunities. In the main paper, you have one column each. Keep in mind that we might keep a shorter version in the final revisions.*

Figure 5: Guidelines for the authors of the paper. Please reach out for any clarification.