When Model Outputs Become Social Evidence: Challenges for LLM-Mediated Web and Social Media Analysis

Antonela Tommasel
Johannes Kepler University Linz
Linz, Austria
CONICET-UNCPBA, ISISTAN
Tandil, Argentina
antonela.tommasel@jku.at

Markus Schedl
Johannes Kepler University Linz and
Linz Institute of Technology
Linz, Austria
markus.schedl@jku.at

Abstract

Large language models (LLMs) are increasingly used to support Web and social media analysis through model-generated outputs such as labels, scores, summaries, and natural-language explanations or rationales. In many cases, these outputs are used to support broader conclusions about online content, communities, platforms, harm, credibility and social meaning. However, using model-generated outputs as evidence introduces a methodological challenge. These systems can support analysis at scale, but they are not neutral observers, and their judgments may reflect Western-centric norms, culturally specific assumptions, translation artifacts, platform biases and learned stereotypes. In this position paper, we argue that conclusions drawn from LLM-mediated analysis should not be treated as neutral or universal readings of online content, but as outcomes of an analytical pipeline whose outputs depend on language, culture, platform context, modality and interpretation. To support this position, we examine how model outputs are shaped across the analysis pipeline, including task and category definition, data selection, input representation, prompting, model judgment, evaluation, interpretation and use. We discuss these issues through a synthetic multimodal hate speech example. We conclude by outlining implications for more context-sensitive, transparent, and accountable uses of LLMs in Web and social media analysis.

CCS Concepts

• **Computing methodologies** → **Natural language generation**;
• **Human-centered computing** → **Social media**.

Keywords

large language models, social media analysis, model evaluation, multimodal analysis, responsible AI

ACM Reference Format:
Antonela Tommasel and Markus Schedl. 2026. When Model Outputs Become Social Evidence: Challenges for LLM-Mediated Web and Social Media Analysis. In *The 5th International Workshop on Multimodal Human Understanding for the Web and Social Media (MUWS '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3841457.3841516

MUWS '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2946-1/2026/11
https://doi.org/10.1145/3841457.3841516

1 Introduction

Large language models (LLMs) are increasingly used to analyse Web and social media data at scale[1]. Beyond content generation, they are being adopted as instruments for annotation, classification, moderation and interpretation. They can produce labels, scores, summaries and natural-language rationales, and these outputs are increasingly used to support conclusions about online harm, credibility, public discourse, communities and social meaning [8, 9, 11, 18]. Recent work [8, 9] has shown the promise of LLMs as scalable annotators and evaluators, while extensive research has documented that LLMs can learn, reproduce and amplify social biases [6, 10, 13]. As a result, LLMs are becoming attractive but problematic tools for socially sensitive analysis of online environments.

Using these models to support Web and social media analysis introduces a methodological paradox. The same models whose outputs are used to draw conclusions about online content may themselves encode Western-centric norms, culturally specific assumptions, platform biases, translation artifacts and learned stereotypes. Prior work has argued that natural language processing (NLP) analyses pertaining to harm, representation, and social groups are not merely technical assessments, but inherently normative judgments that require explicit attention to affected communities, social power relations and the harms that systems may produce or amplify [3]. This concern becomes especially pressing in multimodal Web contexts, where meaning is often informal, ironic, platform-specific and culturally situated. Cultural values and assumptions can be reflected in LLM outputs [1, 17], dialectal and racial biases can shape hate speech and toxicity detection [5, 15], and harmful or biased meanings may emerge from image-text combinations rather than from either modality alone [10, 13, 14]. A phrase, meme, image, emoji, hashtag or visual symbol may carry different meanings across communities, and these meanings may not be preserved through translation or captured by universal label sets.

In this position paper, we argue that conclusions drawn from LLM-mediated analysis *should not be treated as neutral or universal readings of online content*. Instead, LLM outputs should be understood as one component of a broader context-sensitive analysis pipeline shaped by language, culture, modality and interpretation. This pipeline includes decisions about what data is selected, what task is defined, how input content is represented, how prompts frame the model interaction, what learned patterns the model encodes, how outputs are evaluated, and how they are interpreted and

[1] In this paper, we use LLMs broadly to refer to contemporary generative models, many of which are now multimodal and can process not only text, but also images and other forms of content.

used. This view builds on prior work showing that bias can arise from multiple sources in NLP systems, including data collection, annotation, representation, modelling, and evaluation [12, 16, 18]. It also builds on research showing that LLMs may encode, reproduce, or amplify biases learned from training data, alignment procedures, language coverage and dominant cultural assumptions [10, 13], as well as calls for more transparent documentation of datasets, their composition, intended uses and social context [2, 7].

Rather than proposing another benchmark, this position paper provides a framework for examining how methodological choices across this pipeline can shape model outputs and the conclusions drawn from them. We use this framework to outline guidelines for treating model outputs as fallible evidence within context-sensitive, transparent and accountable analysis. While our running example concerns multimodal hate speech/toxicity detection, the same principles apply to other Web and social media analysis tasks where model outputs are used as evidence. By reframing LLM-mediated analysis as a process rather than an automated judgment, we highlight the need for approaches that preserve cultural context, evaluate cross-modal meaning, report uncertainty and disagreement.

2 Beyond model outputs: The pipeline behind LLM-mediated analysis

LLM-mediated Web and social media analysis is often presented as a simple operation in which content is given to a model and the model returns an output such as a label, score, summary or natural-language rationale. Figure 1 uses this simplified view as a starting point (upper part). This view is operationally attractive because it makes model-mediated analysis appear scalable, flexible, and portable across tasks. However, it also risks treating the model output as if it directly captured the relevant social phenomenon. A model-generated label may appear to describe a post directly, and a rationale may appear to explain why the label is appropriate. Nonetheless, such outputs are shaped by decisions about what problem is being studied, what content and context are included, how the model is prompted, and how the resulting output is evaluated and used.

The lower part of Figure 1 expands the simplified view into a *situated analysis pipeline*. In this pipeline, input content does not move directly from content to model output. Rather, the analysis is shaped by a sequence of decisions concerning what task is defined, what data is selected, how input is represented, how the model is prompted, what the model contributes through its learned patterns and capabilities and how the resulting output is evaluated and used. This contrast is not meant simply to add more processing steps. Rather, it shifts attention from the final model output to the sequence of choices through which that output is produced, interpreted, and connected to broader conclusions. Before the model produces an output, researchers have already decided which platforms, languages, modalities, and communities are represented; how task-relevant concepts such as harm, toxicity, or credibility are defined; and whether local expressions, images, or contextual cues are translated, removed, summarised, or preserved. Once the model produces an output, further decisions determine whether that output is treated as a provisional label, a rationale to be inspected, evidence to be triangulated, or a basis for broader conclusions.

This matters because model outputs are rarely the endpoint of Web and social media analysis. They are often used to support broader conclusions, e.g., that a post contains hate speech, that a set of posts reflects toxic discourse, that a topic is associated with harmful framing, that a platform amplifies particular narratives or that a community is represented in particular ways. Such conclusions are especially difficult to interpret in cultural, social and multimodal contexts, where online meaning is rarely contained in text alone and rarely stable across contexts. A phrase, meme, emoji, hashtag, image, or visual symbol may be neutral in one community and harmful, ironic, reclaimed or politically charged in another[2]. Similarly, a post may not contain an explicitly harmful statement, but may still support an exclusionary or socially consequential interpretation through the relation between image and caption, the use of local symbols, or the framing of a social group as suspicious, threatening, inferior or out of place. In such cases, the relevant object of analysis is not simply the text, the image or the model output, but the relation among content, context, category, model and interpretation.

Viewing LLM-mediated analysis as a pipeline therefore shifts the analytical question. Instead of treating a model output as a direct reading of content, we must ask *what choices made that output possible, what meanings were preserved or lost, whose interpretation was privileged, and how the resulting output is used to support broader conclusions.* This is the methodological challenge at the core of our argument. LLMs can support large-scale analysis of Web and social media, but their outputs should not be treated as neutral observations. They are fallible evidence produced within a chain of methodological choices. Even when a model output is treated as usable for analysis, choices can shape what it appears to reveal about harm, credibility, representation or social meaning.

3 From pipeline choices to model outputs

We now apply our situated pipeline perspective to the synthetic multimodal hate speech post in Figure 1. The post is deliberately simplified and AI-generated. It combines a potentially exclusionary caption with an image of a crowd in a neighbourhood setting. The caption refers to *people like them* and states that they should *go back to where they came from*, but the target group, surrounding thread, platform context and local meaning of the image are not specified[3]. Although the running example is synthetic, similar issues arise in established multimodal hate speech benchmarks such as the *Hateful Memes Challenge* [14], where interpretation depends on how image and text are considered together rather than as isolated modalities. The goal is not to assign responsibility to a single component, but to show how decisions across the pipeline accumulate and shape the outputs that later inform broader claims. Table 1 follows the post across the analysis pipeline.

[2]For example, *Pepe the Frog* originated as a cartoon character and circulated widely as a humorous meme, but was later appropriated in some online communities as an alt-right and white supremacist symbol. Importantly, the symbol is not hate-related in every use, as its meaning depends on the accompanying imagery, language, platform, and community context. See https://www.adl.org/resources/hate-symbol/pepe-frog.

[3]This underspecification is intentional. It allows us to foreground common challenges in multimodal social media analysis, such as implied targets, image-text relations, platform cues and culturally situated interpretation, without reproducing a real post or exposing affected communities. The example is therefore used illustratively, to clarify the pipeline's methodological logic, rather than as empirical validation of the framework.

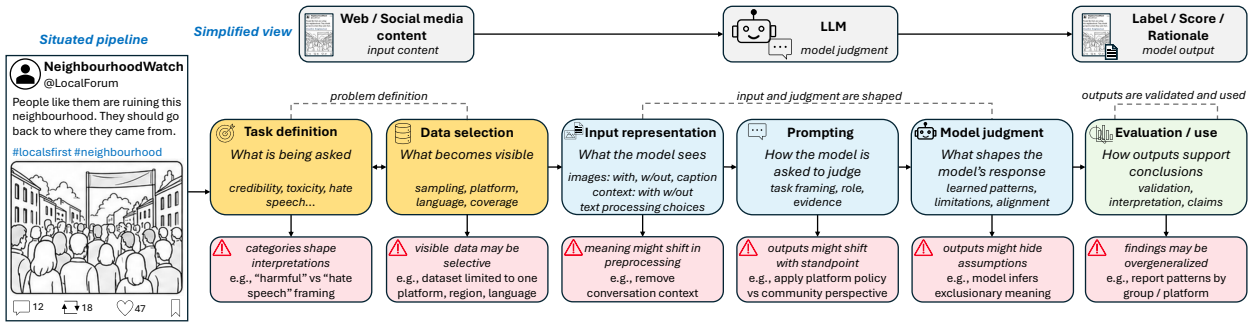

Figure 1: The upper row shows a simplified view of the commonly adopted pipeline in which content is mapped to a label, score or rationale through an LLM. The lower row expands this process into our proposed pipeline in which model outputs are shaped before, during, and after output generation. The synthetic post on the left is illustrative and AI-generated.

Table 1: Example-guided view of how the pipeline stages can shape LLM-mediated analysis of a synthetic multimodal hate speech post.

Pipeline stage	Guiding question	Possible choice	Implication if overlooked
Task definition	What phenomenon is being analysed?	The post may be framed as hate speech, toxicity, xenophobic framing, harmful stereotyping, political complaint or ambiguous rhetoric.	Outputs may appear comparable across tasks or contexts even when categories shape interpretations of harm, evidence or social meaning.
Data selection	What content becomes available for analysis?	The analysis may include only posts from certain platforms, languages, time periods, communities, reports or moderation histories.	Aggregate conclusions may reflect platform access, moderation history or language coverage rather than broader online discourse.
Input representation	What version of the content reaches the model?	The model may receive only the caption, the image and caption together, a translated version, an automatic image caption, or additional thread and platform context.	The model may judge a transformed version of the multimodal content rather than the socially meaningful content users encountered.
Prompting and standpoint	How is the model asked to judge?	Prompts may ask for a neutral label, a platform-policy decision, a credibility explanation or the perspective of a potentially affected community.	Outputs may shift depending on whether ambiguity, implied targets and contextual uncertainty are made visible in the prompt, making prompt choices appear as model findings rather than methodological decisions.
Model judgment	What shapes the model's response?	The model may infer a target group from "people like them," treat "go back to where they came from" as exclusionary, ignore the image or overinterpret the image.	A fluent output may conceal uncertain or culturally mismatched inferences about identity, locality, migration, protest or neighbourhood conflict.
Evaluation, interpretation, and use	How do outputs support conclusions?	The output may be compared with benchmark labels, platform policy, expert review, general annotator judgments or judgments from people familiar with the local context. It may then be interpreted or aggregated across posts, topics, platforms, languages or communities.	Partial or unstable outputs may be generalized into apparently objective claims about platforms, languages, groups or events.

Task definition. What is being asked. Task definition shapes the phenomenon the analysis is intended to capture. The same content may be analysed for credibility, toxicity, hate speech, sentiment or harmful framing. Each framing creates a different object of analysis because it directs attention to different forms of evidence and makes different outputs appear appropriate. For instance, a toxicity task may focus on overt abusive language, while a harmful-framing task may focus on how the post constructs an opposition between an implied *us* and *them*. Categories therefore do not simply organize content. They carry methodological and normative judgments about, for example, what counts as harm, which targets matter, what evidence is sufficient, how intent is inferred, and how platform or policy contexts are understood. In the example, a label such as *toxic*, *hateful*, or *harmful* may not refer to the same construct across platforms, languages or communities. The same post may then be classified as non-toxic, offensive, xenophobic, hateful or ambiguous depending on the task definition.

Data selection. What becomes visible. Data selection determines which platforms, languages, communities, modalities, and interaction contexts become available for model-mediated analysis. In the example, inclusion depends on platform access, sampling strategy, language coverage, and whether both text and image are collected. For instance, if a dataset is built from reported or moderated posts, the analysis may overrepresent content already recognized as problematic by a platform or community, while leaving less visible forms of exclusion underrepresented. In a benchmark setting such as the *Hateful Memes Challenge*, the available data reflects a curated set of image-text pairs designed for multimodal hate speech detection. Such data is useful for controlled evaluation, but conclusions drawn from it should not be automatically generalized to all meme cultures, platforms, languages or communities. This stage may not explain why a model judges a particular post as harmful, credible, exclusionary or ambiguous, but it shapes the aggregate patterns that can later be reported. As a result, model outputs may appear to reveal broad tendencies in online discourse while reflecting the subset of data selected for analysis.

Input representation. What the model really sees. Input processing determines what version of the content actually reaches the

model. This includes decisions about translation, image processing, optical character recognition (OCR), automatic image captioning, text cleaning, and whether threads, hashtags, emojis, usernames, timestamps or platform metadata are preserved. The implication is that the model may judge a transformed version of the content rather than the content encountered by users. In the example, showing the comment in isolation may leave unclear who *them* refers to, while preserving thread context may make the target group explicit. Removing the image, hashtags, emojis, quoted speech, or surrounding posts may also change whether the post appears to be a generic complaint, ambiguous political rhetoric or exclusionary abuse. Similarly, an automatic image caption may preserve that there is a crowd, but not whether the scene signals a protest, public concern or an ordinary gathering. Input processing is therefore not only a technical preparation step. It shapes the object of analysis by determining which linguistic, visual, conversational and platform-specific cues remain available to the model.

Prompting and standpoint. How the model is asked to judge. Prompting determines how the defined task is applied in a specific model interaction. It specifies what criteria it should use, what role it should assume, what kind of evidence it should prioritize, whether it should provide a label or a score, and whether uncertainty or ambiguity should be reported. In the example, the output may change depending on whether the prompt asks for a binary label, a platform-policy decision or an uncertainty-aware assessment. For example, a binary prompt may force the post into a hate/non-hate decision, while an uncertainty-aware prompt may instead surface the missing target, context and image-text relation as reasons for caution. Also, prompt variants do not merely change the wording of the task. They might alter the standpoint from which the content is evaluated and the type of reasoning the model is encouraged to perform. For example, persona-based prompting may make certain harms more visible, but it can also create the misleading impression that the model can authentically represent a group's lived experiences.

Model judgment. What shapes the model's response. Model judgment refers to what the model contributes once the input representation and prompt have been fixed. A model does not approach labels such as *toxic*, *hateful*, *xenophobic* or *harmful* as empty categories. It interprets them through associations learned from training data, alignment procedures, instruction tuning and prior uses of similar terms. As a result, model outputs may reflect not only the content and prompt, but also pre-learned concepts of harm, credibility, offensiveness and social identity. In the example, the model may treat the comment as ambiguous because the target group is not explicit, or it may infer exclusionary meaning from the phrase *go back to where they came from*. These judgments are shaped by models' training data, language coverage, cultural assumptions, alignment norms, and multimodal capabilities. In multimodal hate speech settings, a model may also rely on learned associations between visual objects, social groups, and textual cues, rather than correctly interpreting the image-text relation. The key issue is not simply that the model may be incorrect. Rather, the model can produce fluent and plausible explanations while relying on shallow, dominant or culturally inappropriate assumptions. In multimodal settings, this is particularly difficult to detect. The model may appear confident even when it has misread a local expression, dialectal cue, visual symbol or image-text relation.

Evaluation, interpretation and use. How outputs support conclusions. Evaluation determines how model outputs are assessed and whose judgments are treated as authoritative. In the example, accuracy against a single benchmark label may obscure disagreement between annotators who understand the local context and those who see only the post in isolation. Benchmarks can also reflect the definitions, policies, languages and cultural settings in which they were created. Such disagreement should not automatically be treated as noise. It may reveal differences in lived experience, familiarity with coded language or sensitivity to local histories of exclusion. Evaluation should therefore consider not only whether model labels match benchmark labels, but also whether rationales are appropriate, uncertainty is reported, and disagreement is treated as informative rather than noise. If such outputs are aggregated, a researcher might report high levels of exclusionary rhetoric around a topic or community, even though the result depends on how ambiguous cases, local knowledge and annotator disagreement were handled. Similarly, evaluation on Hateful Memes-style data can show whether a model performs well on a controlled benchmark, but it does not by itself establish that the same outputs support broader claims about real platform dynamics, community-specific meanings or the prevalence of hateful discourse online.

4 Toward context-sensitive LLM-mediated analysis

The pipeline view above shows that LLM-mediated analysis is not a simple process of applying a model to content. It is a chain of methodological choices through which model outputs are produced and made meaningful. Our aim in this position paper is not to produce a new benchmark or evaluation protocol, but to establish the following *principles* to help researchers use model outputs more cautiously when drawing conclusions about Web and social media.

Document the conditions of analysis. Researchers should report the main choices that structure model-mediated analysis, including task definition, data selection, input representation, prompt design, model version, output format, inference settings, evaluation procedure, treatment of uncertainty or disagreement, and the downstream claim the output is used to support. This documentation should also make explicit what content was available to the model, what transformations were applied, and what kind of output was produced. It should also include contingent conditions that may affect results, such as model updates, translation tools, image preprocessing, moderation filters, platform access, inference settings and post-processing decisions. The aim is not only reproducibility, but also interpretability, as readers should be able to understand what kind of evidence the model output represents.

Treat categories as analytical constructs. Labels should not be treated as self-evident or interchangeable. Researchers should define the constructs being analyzed, explain why the chosen categories are appropriate and clarify what kinds of evidence those categories require. Concretely, each category should be accompanied by a brief definition, inclusion and exclusion criteria and examples of borderline cases, especially when labels such as *toxic*, *hateful*, *offensive* or *harmful* are used across datasets, platforms or languages. Treating

categories as analytical constructs helps prevent model outputs from appearing more stable or comparable than the underlying definitions allow.

Inspect outputs beyond labels. Model-mediated analysis should not rely only on whether a label or score appears plausible. Researchers should also examine the evidential basis of model-generated rationales, the handling of uncertainty and the presence of disagreement. In practice, this can involve manually checking a sample of outputs for unsupported inferences, missing context, overconfident wording, and mismatches between the rationale and the available evidence. A fluent rationale may make an output appear well grounded even when it is based on missing context, weak evidence or an inappropriate interpretive frame. Evaluation should therefore ask not only whether the output matches a reference label, but also whether the rationale is warranted by the available content and whether uncertainty or ambiguity has been represented and not suppressed.

Evaluate context and modality together. For multimodal and socially sensitive content, evaluation should consider whether the model has preserved the relevant relation between text, image, platform context and social meaning. Access to multiple modalities should not be equated with understanding their interaction. A model may describe visual and textual elements separately while missing irony, contradiction or culturally specific symbolism that emerges only from their combination. A practical check is to compare outputs across input variants (such as text-only, image-only and multimodal), with or without contextual metadata, and to examine whether the model's judgment changes in ways that are justified by the added context. Researchers should therefore inspect whether cross-modal meaning is actually being evaluated, especially when model outputs are used to support claims about harm, credibility or public discourse.

Treat variability as part of the evidence. LLM-mediated outputs should be treated as generated responses rather than fixed measurements. They may vary across prompt formulations, input orderings, model versions, system instructions, inference settings and output formats. Researchers should report whether outputs remain stable across reasonable variations in the analysis setup, for example by repeating the analysis with alternative prompt phrasings, uncertainty options, input orderings or model versions. Instability does not always mean that a model is unusable, but it changes what kinds of claims the outputs can support. In some cases, variability may itself reveal ambiguity in the content, uncertainty in the task definition, or sensitivity to contextual framing.

Bound downstream conclusions. Conclusions drawn should remain tied to the limits of the pipeline that produced them. Model-mediated outputs may inform research findings, moderation decisions, fact-checking priorities, or public narratives, but they should not be treated as neutral observations of online discourse. As a practical rule, every aggregate claim should be traceable back to the data scope, category definition, model setup, evaluation procedure, and uncertainty or disagreement on which it depends. Claims about platforms, communities, languages, or harms should be framed in relation to the data, categories, model, evaluation procedure, and uncertainty and disagreement on which they depend. This is particularly important when outputs are aggregated or used to support broader claims about groups, events or online environments.

All in all, these guidelines suggest that model outputs should be treated as fallible analytical materials whose meaning depends on the pipeline through which they are produced and used. The goal is to make explicit how methodological choices, model mediation, and downstream consequences interact in the conclusions about online content and social meaning.

5 Conclusions

LLMs have become central tools for analysing Web and social media data, but their outputs should not be treated as neutral evidence. In this position paper, we argued that LLM-mediated analysis is a situated pipeline in which model outputs are shaped by choices about problem and task definition, input representation, prompting, model judgment, evaluation, interpretation and use. These choices affect what model outputs appear to reveal about the phenomena under analysis and therefore the conclusions that can be drawn.

The pipeline should not be understood as sufficient to fully explain the social, cultural and interpretive issues that may shape model-generated outputs. It makes the conditions of LLM-mediated analysis more explicit, but it does not replace community expertise, task-specific validation, interpretability methods or domain-specific theories of multimodal meaning. For analyses of complex multimodal content, it should therefore be complemented by approaches from explainable AI and computational social science, including taxonomies that distinguish author intent, image-text relations, domain-specific values and human interpretation (e.g., [4]).

The central challenge is therefore methodological as much as technical. As model-generated labels, scores, summaries and rationales become easier to produce at scale, researchers must be more explicit about the conditions under which those outputs are generated, interpreted and used. Treating LLMs as components of an analytical pipeline, rather than as neutral observers, helps clarify why model outputs should be treated as fallible evidence. The guidelines outlined in this paper aim to support more context-sensitive, transparent and accountable uses of LLMs.

Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF): 10.55776/COE12, 10.55776/DFH23, 10.55776/P36413; and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research: LIT-2024-13-SEE-111.

References

[1] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating Cultural Alignment of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12404–12422. doi:10.18653/v1/2024.acl-long.671

[2] 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 (2018), 587–604. doi:10.1162/tacl_a_00041

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 5454–5476.

[4] Gullal S. Cheema, Sherzod Hakimov, Eric Müller-Budack, Christian Otto, John A. Bateman, and Ralph Ewerth. 2023. Understanding image-text relations and news

values for multimodal news analysis. *Frontiers in Artificial Intelligence* Volume 6 - 2023 (2023). doi:10.3389/frai.2023.1125533

[5] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem (Eds.). Association for Computational Linguistics, Florence, Italy, 25–35. doi:10.18653/v1/W19-3504

[6] 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 (Sept. 2024), 1097–1179. doi:10.1162/coli_a_00524

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

[8] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120. arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2305016120 doi:10.1073/pnas.2305016120

[9] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on LLM-as-a-judge. *The Innovation* 7, 6 (2026), 101253. doi:10.1016/j.xinn.2025.101253

[10] Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 547–561.

[11] Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics* 11, 1 (2024), 20531680241236239. arXiv:https://doi.org/10.1177/20531680241236239 doi:10.1177/20531680241236239

[12] Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass* 15, 8 (2021), e12432.

[13] Jen-tse Huang, Jiantong Qin, Jianping Zhang, Youliang Yuan, Wenxuan Wang, and Jieyu Zhao. 2025. Visbias: Measuring explicit and implicit social biases in vision language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 17981–18004.

[14] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose, Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. 2021. The Hateful Memes Challenge: Competition Report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track (Proceedings of Machine Learning Research, Vol. 133)*, Hugo Jair Escalante and Katja Hofmann (Eds.). PMLR, 344–360. https://proceedings.mlr.press/v133/kiela21a.html

[15] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1668–1678. doi:10.18653/v1/P19-1163

[16] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5248–5264. doi:10.18653/v1/2020.acl-main.468

[17] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus* 3, 9 (2024), pgae346.

[18] Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, 82232–82251. https://proceedings.mlr.press/v267/wallach25a.html