Natural Language Processing (2026), 1–7
doi:10.1017/nlp.2026.10030

RESPONSIBLE NLP

Responsible NLP: Why performance is no longer enough

Salima Lamsiyah

University of Luxembourg, Luxembourg
Email: salima.lamsiyah@uni.lu

(Received 1 June 2026 UTC; revised 1 June 2026 UTC; accepted 2 June 2026 UTC)

Abstract

Natural language processing (NLP) has moved from a specialized research field into the everyday infrastructure of writing, search, translation, education, journalism, public administration, and scientific work. This transition changes what counts as progress. Accuracy, fluency, and benchmark performance remain important, but they are no longer sufficient when language technologies shape knowledge, decisions, identities, and public trust. This column introduces Responsible NLP as a research orientation that integrates fairness, transparency, privacy, safety, cultural diversity, environmental awareness, and human agency across the full life cycle of language technologies. It argues that responsibility is not an external constraint on innovation, but a condition for meaningful and trustworthy innovation. Future research must therefore ask not only whether an NLP system works but also for whom it works, under which assumptions, with what risks, and with what forms of accountability.

Keywords: Responsible NLP; large language models; fairness; accountability; human-centered AI

1. Introduction

> "All models are wrong, but some are useful."
> George E. P. Box (Box 1976)

Box's well-known warning has become newly urgent in the age of large language models (LLMs). In natural language processing (NLP), models are not only used to classify documents, translate sentences, retrieve information, or summarize texts but also to participate in the production of language itself. They draft emails, generate explanations, support education, moderate online spaces, assist journalists, translate between communities, and help institutions communicate with citizens. In doing so, they do not merely process language; they influence what becomes visible, credible, fluent, searchable, and actionable.

This is why Responsible NLP is no longer optional. For many years, progress in NLP was largely measured through better scores: higher accuracy, stronger benchmark results, larger datasets, larger models, and more fluent outputs. These achievements have transformed the field. Yet the same progress has also made the limits of performance-centered evaluation increasingly difficult to overlook. A model can be accurate on average and still fail systematically for particular communities. It can be fluent and still be misleading. It can be helpful in one context and harmful in another. It can produce elegant language while erasing the social, cultural, and political conditions from which language derives its meaning.

The central question for contemporary NLP is therefore not only *Can the system perform the task?* but also *Who defines the task? Whose language is represented? Who benefits? Who may be*

harmed? Who can contest the output? and *Who is accountable when language technologies fail?* These questions do not sit outside NLP research. They now belong to its scientific core.

The first piece in this column offers a general introduction to Responsible NLP. Future articles can examine specific topics in depth, such as bias and fairness, multilingual and low-resource NLP, responsible evaluation, documentation, explainability, misinformation, privacy, environmental costs, and the governance of LLMs. The aim here is to establish a simple starting point: ***performance is necessary, but it is not enough.***

2. From tools to infrastructure

Responsible NLP has become urgent because language technologies have undergone a change in status. They are no longer isolated tools used by specialists. They are becoming infrastructure for communication, knowledge production, and decision support. LLMs and conversational systems such as ChatGPT, Claude, Gemini, LLaMA-based systems, and Copilot-style assistants have made NLP visible to a broad public. However, the deeper transformation is not the chatbot interface itself. It is the increasing reliance of social, educational, professional, and institutional life on systems that generate, rank, translate, summarize, and interpret language at scale.

This transformation matters because language is not a neutral container for information. Language carries cultural memory, social identity, power relations, emotion, ambiguity, and context. A translation system does not only move words between languages but may also move concepts between cultural worlds. A summarization system does not only shorten text but decides what is central and what can disappear. A moderation system does not only classify content but shapes the boundaries of public speech. A question-answering system does not only retrieve knowledge but frames what counts as an answer.

Humanistic and social theories of language have long emphasized that meaning is interpretative and situated. Eco argued that texts invite interpretation rather than containing a single fixed meaning (Eco 1979). Hall showed that communication is shaped by encoding, decoding, and dominant cultural frameworks (Hall 1980). For NLP, these insights are not abstract philosophical decoration. They remind us that language models operate in a domain where meaning, context, and power cannot be fully reduced to tokens, labels, and probabilities.

This does not make NLP impossible. It makes responsibility necessary. If language is socially situated, then responsible language technology must be evaluated not only by internal technical measures but also by its effects on people, communities, institutions, and cultures.

3. What Responsible NLP means

Responsible NLP can be understood as the design, development, evaluation, deployment, and governance of language technologies in ways that reduce harm, support human agency, and promote social benefit (Behera *et al.* 2023; Blodgett *et al.* 2022; Cheng, Varshney, and Liu 2021). It is not a single method, metric, or checklist. It is an approach to research that treats NLP systems as sociotechnical systems: technical artifacts embedded in social contexts.

This means responsibility begins before model training. It starts when researchers define a problem. Some NLP tasks are framed as if they were purely technical, although they involve contested human categories: toxicity, misinformation, hate speech, sentiment, emotion, credibility, fluency, quality, and even "helpfulness." These categories often depend on culture, context, speaker identity, institutional norms, and historical power relations. A responsible approach asks whether the task formulation itself is appropriate, whose definitions are being used, and whether affected communities have been considered.

Responsibility continues through data curation. NLP datasets are not raw mirrors of language. They are collected, filtered, labeled, cleaned, licensed, and interpreted through human choices. Data statements (Bender and Friedman 2018), datasheets for datasets (Gebru *et al.* 2021), and

careful documentation practices help make these choices visible. Without such documentation, models can inherit biases, exclusions, and privacy risks while appearing technically neutral.

Responsibility also shapes modeling and alignment. Methods such as reinforcement learning from human feedback (Ouyang *et al.* 2022) and constitutional AI (Bai *et al.* 2022) attempt to align model behavior with explicit preferences or principles. These approaches are important, but they also raise difficult questions: Which values are encoded? Who writes the principles? and How are conflicts between helpfulness, harmlessness, truthfulness, and freedom of expression resolved? Alignment is not only an optimization problem but also a normative and institutional problem.

Finally, responsibility must extend to evaluation and deployment. Aggregate benchmark scores are too narrow for systems that operate across languages, communities, and real-world contexts. Responsible evaluation should examine subgroup performance, failure modes, robustness, uncertainty, potential misuse, environmental costs, labor conditions, and the ability of users to understand and contest system outputs. Model cards (Mitchell *et al.* 2019), audits (Raji *et al.* 2020), and risk management frameworks (National Institute of Standards and Technology 2023) offer partial tools for this broader view.

4. Why performance alone fails

Performance-centered NLP has produced impressive systems, but it can hide the very questions that matter most once systems leave the laboratory. The problem is not that benchmarks are useless. The problem is that benchmarks can become a substitute for judgment.

First, average performance hides unequal failure:
An NLP system may perform well overall while failing for particular dialects, minority languages, social groups, or communicative styles. Bias in language technologies has been documented in word embeddings (Bolukbasi *et al.* 2016), language models (Bender *et al.* 2021), and stereotype-focused evaluations (Mostafazadeh Davani *et al.* 2025). These failures are not merely technical imperfections. They can affect access to information, visibility, reputation, safety, and opportunity.

Second, fluency can be mistaken for understanding:
LLMs produce text that often appears confident, coherent, and meaningful. This fluency can create an illusion of understanding, especially when users are under time pressure or lack domain expertise. Ethical and social risk analyses of language models have warned that such systems can generate misinformation, enable manipulation, expose private information, and create new forms of over-reliance (Bommasani *et al.* 2021; Weidinger *et al.* 2021).

Third, language categories are never innocent:
Many NLP tasks require classification: toxic or not toxic, relevant or not relevant, credible or not credible, biased or unbiased. Classification can be useful, but it can also make incomplete categories appear complete. Bowker and Star's work on classification reminds us that categories organize social life and can make some people visible while making others difficult to name (Bowker and Star 2000). In Responsible NLP, labels should be treated as situated decisions, not as universal truths.

Fourth, scale changes the meaning of harm:
A smallerror in a research prototype may have limited consequences. The same error in a widely deployed language system can affect millions of interactions. Harms can also be cumulative: repeated misgendering, repeated mistranslation, repeated erasure of minority perspectives,

repeated exposure to toxic content by annotators, or repeated privileging of dominant languages and worldviews.

Fifth, success for users may conflict with success for institutions:
A system optimized for engagement, productivity, or cost reduction may not optimize for dignity, fairness, learning, democratic deliberation, or care. Responsible NLP therefore asks what kind of progress a system supports. ***Efficiency is valuable, but it is not the only human value.***

5. Language, power, and cultural diversity

Responsible NLP must pay particular attention to linguistic and cultural diversity. The current wave of generative AI is often described as universal, but its universality is uneven. High-resource languages, especially English, are disproportionately represented in training data, evaluation benchmarks, commercial products, and research attention. Low-resource, Indigenous, regional, and minoritized languages often remain under-supported or are represented through data extracted without adequate consent, governance, or cultural understanding.

This imbalance matters because language is tied to identity and belonging. When NLP systems work poorly for a language, they do not merely fail a market segment; they may reinforce the idea that some communities are less worthy of technological investment. When systems translate culturally specific expressions into dominant conceptual frameworks, they may flatten differences that should be preserved. When models generate polished text in standardized registers, they may quietly reward linguistic conformity over expressive plurality.

The risk is not only exclusion but also homogenization. Language technologies can normalize dominant ways of speaking, reasoning, and classifying the world. This is why multilingual NLP should not be reduced to increasing the number of supported languages. It should also ask how languages are represented, who controls the data, whether communities benefit, and whether technological intervention supports language vitality rather than extraction.

In this sense, Responsible NLP is aligned with NLP for social good, but it must remain careful about the phrase "good." Recent work on NLP for social good emphasizes cross-disciplinary partnerships, human-centered methods, and responsible deployment (Karamolegkou *et al.* 2026). These are essential because social benefit cannot be declared by researchers alone. It must be negotiated with the people and communities affected by the system.

6. The human in human-centered NLP

Human-centered NLP should mean more than adding a user study after a model is built. It requires a commitment to preserving human agency, interpretation, and empathy. This is especially important in domains such as education, health, migration, journalism, public services, and crisis response, where language technologies may influence vulnerable people.

Consider education. NLP systems can help detect early signs of student disengagement, summarize feedback, support multilingual learners, and assist teachers with administrative work. Used responsibly, such systems can help ensure that students are not lost in large-scale data systems. But the same systems can also reinforce inequality if they rely on biased indicators, ignore infrastructural disparities, or replace human judgment with automated risk scores. An alert that a student is at risk is not an explanation of why the student is struggling. It is an invitation for human care, inquiry, and support.

This distinction is central. Responsible NLP should augment human capabilities, not displace human responsibility. The goal is not to build systems that remove humans from difficult decisions, but systems that help humans make better, more informed, and more equitable decisions. In high-stakes settings, human oversight should be meaningful, not symbolic. Users should know

Table 1. A life cycle view of Responsible NLP

Research dimension	Guiding question
Problem formulation	Who defines the task, and whose interests does the task serve?
Data	Whose language is included, excluded, labeled, anonymized, or extracted? Is data biased?
Modeling	What assumptions, values, and trade-offs are encoded in the system?
Evaluation	Does the system work across contexts, communities, and realistic failure modes? Is evaluation fair?
Deployment	Who can understand, contest, monitor, and repair the system?
Governance	What forms of accountability exist when harms occur?
Sustainability	What environmental, economic, and labor costs are created by the system?

when they are interacting with AI-generated language, understand the system's limitations, and have routes to contest decisions that affect them.

7. A research agenda for Responsible NLP

If Responsible NLP is to become a serious research agenda, it must be operationalized. The following questions, summarised in Table Table1, can guide future work and future pieces in this column.

Several practical directions follow from this agenda. Responsible NLP needs better documentation standards; richer evaluations beyond leaderboard performance; participatory design methods involving affected communities; stronger multilingual and low-resource benchmarks; clearer reporting of uncertainty and limitations; better privacy protection; careful treatment of harmful text in research; and more attention to the workers who collect, annotate, moderate, and evaluate language data. Application areas such as text summarization, fact-checking, and hate-speech detection show why these questions must be addressed at task level rather than only at the level of general AI principles (Liu *et al.* 2023; Vargas, Benevenuto, and Pardo 2026).

It also needs institutional support. Responsibility cannot depend only on individual researchers making heroic choices. Conferences, journals, funders, universities, companies, and public institutions shape what work is rewarded. This includes what costs are counted: training and deploying large models can create substantial environmental burdens, which is why Green AI should be treated as part of Responsible NLP rather than as a separate concern (Schwartz *et al.* 2020; Strubell, Ganesh, and McCallum 2019). The emergence of policy frameworks, including the European Union's AI Act and broader AI risk management efforts, signals that responsible AI is becoming part of the governance environment in which NLP research will operate (Commission 2024; European National Institute of Standards and Technology 2023). The research community should not wait for regulation to define responsibility from the outside. It should help build the concepts, methods, evidence, and practices that make responsible innovation possible.

8. Conclusion

Responsible NLP is sometimes presented as a constraint: a set of warnings that slows down technical progress. This framing is misleading. Responsibility is not the opposite of innovation. It is what allows innovation to become trustworthy, socially meaningful, and scientifically honest.

The future of NLP will not be judged only by whether models become larger, faster, or more fluent, but also by whether language technologies help people understand each other, preserve cultural diversity, support fair decisions, protect vulnerable communities, reduce misinformation, respect privacy, and strengthen rather than weaken human agency. In a field built around language, this is not a secondary concern. It is the heart of the matter.

Performance asks whether a model can produce an answer. Responsible NLP asks whether the answer should be trusted, used, shared, challenged, or refused. It is this distinction that renders Responsible NLP no longer optional.

Competing interests. The author declare that the manuscript complies with the ethical standards of the journal and that there are no competing interests to disclose.

References

Bai Y., **Kadavath S.**, **Kundu S.**, **Askell A.**, **Kernion J.**, **Jones Andy**, **Chen A.**, **Goldie A.**, **Mirhoseini A.**, **Cameron M.** , et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv: 2212.08073.

Behera R.K., **Bala P.K.**, **Rana N.P.** and **Irani Z.** (2023). Responsible natural language processing: A principlist framework for social benefits. *Technological Forecasting and Social Change* **188**, 122306. https://doi.org/10.1016/j.techfore.2022.122306

Bender E.M. and **Friedman B.** (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. https://doi.org/10.1162/tacl_a_00041

Bender E.M., **Gebru T.**, **McMillan-Major A.** and **Shmitchell S.** (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Blodgett S.L., **Liao Q.V.**, **Olteanu A.**, **Mihalcea R.**, **Muller M.**, **Scheuerman M.K.**, **Tan C.** and **Yang Q.** (2022). Responsible language technologies: Foreseeing and mitigating harms. *Chi Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–3.

Bolukbasi T., **Chang K.-W.**, **Zou J.Y.**, **Saligrama V.** and **Kalai A.** (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*.

Bommasani R., **Hudson D.A.**, **Adeli E.**, **Altman R.**, **Arora S.**, **von Arx S.**, **Bernstein M.S.**, **Bohg J.**, **Bosselut A.**, **Emma B.**, et al. (2021). On the opportunities and risks of foundation models, arXiv preprint arXiv: 2108.07258.

Bowker G.C. and **Star S.L.** (2000). *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.

Box G.E.P. (1976). Science and statistics. *Journal of the American Statistical Association* **71**(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Cheng L., **Varshney K.R.** and **Liu H.** (2021). Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research* **71**, 1137–1181. https://doi.org/10.1613/jair.1.12814

Eco U. (1979). *The Role of the Reader: Explorations in the Semiotics of Texts*. Bloomington: Indiana University Press.

European Commission. (2024). Ai Act Enters into Force. Available at: https://commission.europa.eu/news/ai-act-enters-force-2024-08-01en (accessed 29 May 2026).

Gebru T., **Jamie M.**, **Briana V.**, **JenniferWortman V.**, **HannaWallach H.D.**III and **CrawfordK.** (2021). Datasheets for datasets. *Communications of the ACM* **64**(12), 86–92. https://doi.org/10.1145/3458723

Hall S. (1980). Encoding/decoding. In **Hall S.**, **Hobson D.**, **Lowe A.** and **Willisa P.** (eds), *Culture, Media, Language: Working Papers in Cultural Studies*. London: Hutchinson, pp. 128–138.

Karamolegkou A., **Borah A.**, **Cho E.**, **Choudhury S.R.**, **Galletti M.**, **Gupta P.**, **Ignat O.**, **Kargupta P.**, **Kotonya N.**, **Hemank L.**, et al. (2026). Nlp for social good: A survey and outlook of challenges, opportunities and responsible deployment. *Proceedings of the 19th conference of the european chapter of the association for computational linguistics*, pp. 5110–5170.

Liu Y.L., **Cao M.**, **Blodgett S.L.**, **Cheung J.C.K.**, **Olteanu A.** and **Trischler A.** (2023). Responsible ai considerations in text summarization research: A review of current practices. *Findings of the Association for Computational Linguistics: EMNLP*, pp. 6246–6261. https://doi.org/10.18653/v1/2023.findings-emnlp.417

Mitchell M., **SimoneWu A.Z.**, **Barnes P.**, **Vasserman L.**, **Hutchinson B.**, **Spitzer E.**, **Raji I.D.** and **Gebru T.** (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. https://doi.org/10.1145/3287560.3287596

Davani M., **Aida S.D.**, **Pérez-Urbina H.** and **Prabhakaran V.** (2025). A comprehensive framework to operationalize social stereotypes for responsible ai evaluations. In Proceedings of the 2025 conference on empirical methods in natural language processing, 30030-30043. Association for Computational Linguistics. 30030–30043. https://doi.org/10.18653/v1/2025.emnlp-main.1526

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). Technical report NIST AI 100-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Ouyang L., **JeffreyWu X.J.**, **Almeida D.**, **L.Wainwright C.**, **Mishkin P.**, **Zhang C.**, **Agarwal S.**, **Slama K.**, **Alex R.**, et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35**, 27730–27744.

Raji I.D., **Smart A.**, **White R.N.**, **Mitchell M.**, **Gebru T.**, **Hutchinson B.**, **Smith-Loud J.**, **Theron D.** and **Barnes P.** (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44. https://doi.org/10.1145/3351095.3372873

Schwartz R., **Dodge J.**, **Smith N.A.** and **Etzioni O.** (2020). Green ai. *Communications of the ACM* **63**(12), 54–63. https://doi.org/10.1145/3381831

Strubell E., **Ganesh A.** and **McCallum A.** (2019). Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650. https://doi.org/10.18653/v1/P19-1355

Vargas F., **Benevenuto F.** and **Pardo T.A.S.** (2026). Socially responsible and explainable automated fact-checking and hate speech detection. *Proceedings of the 17th international conference on computational processing of portuguese*, pp. 35–42.

Weidinger L., **Mellor J.**, **Rauh M.**, **Griffin C.**, **Uesato J.**, **Huang P.-S.**, **Cheng M.**, **Glaese M.**, **Balle B.**, **Atoosa K.**, et al. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv: 2112.04359.

Cite this article: Lamsiyah S. Responsible NLP: Why performance is no longer enough. *Natural Language Processing* https://doi.org/10.1017/nlp.2026.10030