THE UNIVERSITY of EDINBURGH
informatics

THE UNIVERSITY of EDINBURGH
Edinburgh Futures Institute | Centre for Technomoral Futures

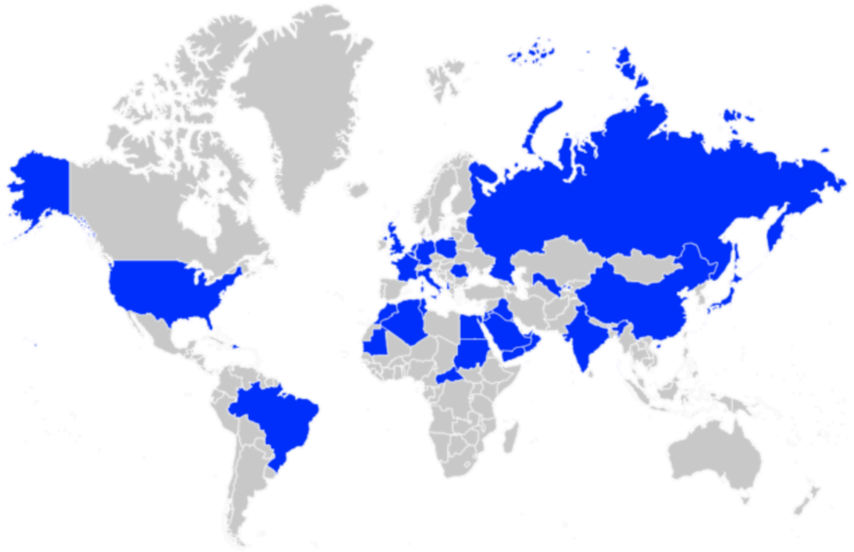
On the Futility of Evaluation

17-Nov-2025

Zeerak Talat
z@zeerak.org | @zeeraktalat
www: zeerak.org

SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models

May 12, 2025



SHADES: Towards a Multilingual Assessment of Stereotypes in Large Language Models

Margaret Mitchell[1], Hamdan Al-Ali[2] Giuseppe Attanasio[3], Ioana Baldini[4], Miruna Clinciu[5,6], Jordan Clive[7], Pieter Delobelle[8,45], Manan Dey[9], Kaustubh Dhole[10], Timm Dill[11], Amirbek Djanibekov[2], Tair Djanibekov[12], Jad Doughman[2], Ritam Dutt[13], Jessica Zosa Forde[14], Jay Gala[2], Avijit Ghosh[1], Sil Hamilton[15], Carolin Holtermann[11], Jerry Huang[16,17], Lucie-Aimée Kaffee[1], Janavi Kasera[18], Tanmay Laud[19,20], Anne Lauscher[11], Roberto Luis López[21], Jonibek Mansurov[2], Maraim Masoud[22], Sagnik Mukherjee[23], Nurdaulet Mukhituly[2], Nikita Nangia[24], Shangrui Nie[25], Anaelia Ovalle[26], Giada Pistilli[1], Esther Ploeger[27], Jeremy Qin[16,17,28], Dragomir Radev[29], Vipul Raheja[30], Beatrice Savoldi[31], Shanya Sharma[32], Xudong Shen[33], Karolina Stańczak[16,34], Arjun Subramonian[26], Kaiser Sun[35], Eliza Szczechla[36], Tiago Timponi Torrent[37,38], Deepak Tunuguntla[39], Emilio Villa-Cueva[2], Marcelo Viridiano[40], Oskar van der Wal[41], Adina Yakefu[1], Kayo Yin[42], Mike Zhang[27], Sydney Zink[43], Aurélie Névéol[44], Zeerak Talat[6]

You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings

Zeerak Talat[1], **Aurélie Névéol**[2], **Stella Biderman**[3,4], **Miruna Clinciu**[5,6,7], **Manan Dey**[8], **Shayne Longpre**[9], **Alexandra Sasha Luccioni**[10], **Maraim Masoud**[11], **Margaret Mitchell**[10], **Dragomir Radev**[12], **Shanya Sharma**[13], **Arjun Subramonian**[14,15], **Jaesung Tae**[10,12], **Samson Tan**[16,17], **Deepak Tunuguntla**[18], **Oskar van der Wal**[19]

[1]Digital Democracies Institute, Simon Fraser University [2]Université Paris-Saclay, CNRS, LISN [3]Booz Allen Hamilton [4]EleutherAI [5]Edinburgh Centre for Robotics [6]Heriot-Watt University [7]University of Edinburgh [8]SAP [9]MIT [10]Hugging Face [11]Adapt Centre, Trinity College Dublin [12]Yale University [13]Walmart Labs, India [14]University of California, Los Angeles [15]Queer-in-AI [16]AWS AI Research and Education [17]National University of Singapore [18]Independent Researcher [19]University of Amsterdam

Evaluating the Social Impact of Generative AI Systems in Systems and Society

Irene Solaiman[1*] **Zeerak Talat**[2*] **William Agnew**[3] **Lama Ahmad**[4] **Dylan Baker**[5] **Su Lin Blodgett**[6] **Canyu Chen**[7] **Hal Daumé III**[8] **Jesse Dodge**[9] **Isabella Duan**[10] **Felix Friedrich**[11,12] **Avijit Ghosh**[1] **Usman Gohar**[13] **Sara Hooker**[14] **Yacine Jernite**[1] **Ria Kalluri**[15] **Alberto Lusoli**[16] **Alina Leidinger**[17] **Michelle Lin**[18,19] **Xiuzhu Lin**[11] **Sasha Luccioni**[1] **Jennifer Mickel**[21] **Margaret Mitchell**[1] **Jessica Newman**[22] **Anaelia Ovalle**[22] **Marie-Therese Png**[23] **Shubham Singh**[24] **Andrew Strait**[25] **Lukas Struppek**[11,26] **Arjun Subramonian**[22]

[1]Hugging Face, [2]Mohamed Bin Zayed University of Artificial Intelligence, [3]Carnegie Mellon University, [4]OpenAI, [5]DAIR, [6]Microsoft Research, [7]Illinois Institute of Technology, [8]University of Maryland, [9]Allen Institute for AI, [10]University of Chicago, [11]TU Darmstadt, [12]hessian.AI, [13]Iowa State University, [14]Cohere for AI, [15]Stanford University, [16]Simon Fraser University, [17]University of Amsterdam, [18]Mila - Quebec AI Institute, [19]McGill University, [21]University of Texas at Austin, [21]University of California, Berkeley, [22]University of California, Los Angeles, [23]Oxford University, [24]University of Illinois Chicago, [25]Ada Lovelace Institute, [26]DFKI

You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings

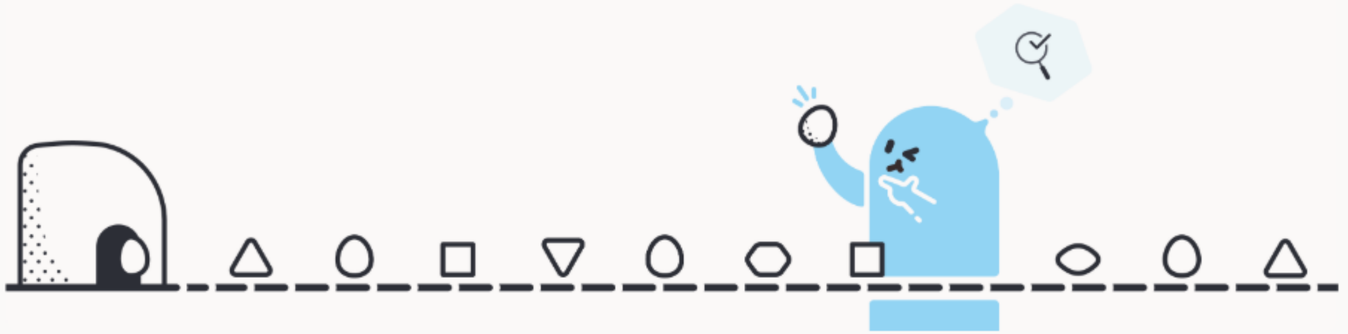
Zeerak Talat[1], **Aurélie Névéol**[2], **Stella Biderman**[3,4], **Miruna Clinciu**[5,6,7], **Manan Dey**[8], **Shayne Longpre**[9], **Alexandra Sasha Luccioni**[10], **Maraim Masoud**[11], **Margaret Mitchell**[10], **Dragomir Radev**[12], **Shanya Sharma**[13], **Arjun Subramonian**[14,15], **Jaesung Tae**[10,12], **Samson Tan**[16,17], **Deepak Tunuguntla**[18], **Oskar van der Wal**[19]

[1]Digital Democracies Institute, Simon Fraser University [2]Université Paris-Saclay, CNRS, LISN [3]Booz Allen Hamilton [4]EleutherAI [5]Edinburgh Centre for Robotics [6]Heriot-Watt University [7]University of Edinburgh [8]SAP [9]MIT [10]Hugging Face [11]Adapt Centre, Trinity College Dublin [12]Yale University [13]Walmart Labs, India [14]University of California, Los Angeles [15]Queer-in-AI [16]AWS AI Research and Education [17]National University of Singapore [18]Independent Researcher [19]University of Amsterdam

Evaluating the Social Impact of Generative AI Systems in Systems and Society

Irene Solaiman[1*] **Zeerak Talat**[2*] **William Agnew**[3] **Lama Ahmad**[4] **Dylan Baker**[5] **Su Lin Blodgett**[6] **Canyu Chen**[7] **Hal Daumé III**[8] **Jesse Dodge**[9] **Isabella Duan**[10] **Felix Friedrich**[11,12] **Avijit Ghosh**[1] **Usman Gohar**[13] **Sara Hooker**[14] **Yacine Jernite**[1] **Ria Kalluri**[15] **Alberto Lusoli**[16] **Alina Leidinger**[17] **Michelle Lin**[18,19] **Xiuzhu Lin**[11] **Sasha Luccioni**[1] **Jennifer Mickel**[21] **Margaret Mitchell**[1] **Jessica Newman**[22] **Anaelia Ovalle**[22] **Marie-Therese Png**[23] **Shubham Singh**[24] **Andrew Strait**[25] **Lukas Struppek**[11,26] **Arjun Subramonian**[22]

[1]Hugging Face, [2]Mohamed Bin Zayed University of Artificial Intelligence, [3]Carnegie Mellon University, [4]OpenAI, [5]DAIR, [6]Microsoft Research, [7]Illinois Institute of Technology, [8]University of Maryland, [9]Allen Institute for AI, [10]University of Chicago, [11]TU Darmstadt, [12]hessian.AI, [13]Iowa State University, [14]Cohere for AI, [15]Stanford University, [16]Simon Fraser University, [17]University of Amsterdam, [18]Mila - Quebec AI Institute, [19]McGill University, [21]University of Texas at Austin, [21]University of California, Berkeley, [22]University of California, Los Angeles, [23]Oxford University, [24]University of Illinois Chicago, [25]Ada Lovelace Institute, [26]DFKI



RESEARCH COMMUNITY

EvalEval Coalition

We are a research community developing scientifically grounded **research outputs** and robust **deployment infrastructure** for **broader impact evaluations**.

WHO EVALUATES AI'S SOCIAL IMPACTS? MAPPING COVERAGE AND GAPS IN FIRST AND THIRD PARTY EVALUATIONS

Anka Reuel[1,*], **Avijit Ghosh**[2,*], **Jenny Chim**[3,*]

Andrew Tran[4,◇], **Yanan Long**[5,◇], **Jennifer Mickel**[6,◇], **Usman Gohar**[7,◇],
Srishti Yadav[8,◇], **Pawan Sasanka Ammanamanchi**[4,◇]

Mowafak Allaham[9], **Hossein A. Rahmani**[10], **Mubashara Akhtar**[11], **Felix Friedrich**[12], **Robert Scholz**[13],
Michael Alexander Riegler[14], **Jan Batzner**[15,16], **Eliya Habba**[17], **Arushi Saxena**[18], **Anastassia Kornilova**[19],
Kevin Wei[20], **Prajna Soni**[21], **Yohan Mathew**[4], **Kevin Klyman**[1], **Jeba Sania**[20], **Subramanyam Sahoo**[22],
Olivia Beyer Bruvik[1], **Pouya Sadeghi**[23], **Sujata Goswami**[24], **Angelina Wang**[25],

Yacine Jernite[2,†], **Zeerak Talat**[26,†], **Stella Biderman**[6,†],
Mykel Kochenderfer[1,†], **Sanmi Koyejo**[1,†], **Irene Solaiman**[2,†]

ABSTRACT

Foundation models are increasingly central to high-stakes AI systems, and governance frameworks now depend on evaluations to assess their risks and capabilities. Although general capability evaluations are widespread, social impact assessments covering bias, fairness, privacy, environmental costs, and labor practices remain uneven across the AI ecosystem. To characterize this landscape, we conduct the first comprehensive analysis of both first-party and third-party social impact evaluation reporting across a wide range of model developers. Our study examines 186 first-party release reports and 183 post-release evaluation sources, and complements this quantitative analysis with interviews of model developers. We find a clear division of evaluation labor: first-party reporting is sparse, often superficial, and has declined over time in key areas such as environmental impact and bias, while third-party evaluators including academic researchers, nonprofits, and independent organizations provide broader and more rigorous coverage of bias, harmful content, and performance disparities. However, this complementarity has limits. Only model developers can authoritatively report on data provenance, content moderation labor, financial costs, and training infrastructure, yet interviews reveal that these disclosures are often deprioritized unless tied to product adoption or regulatory compliance. Our findings indicate that current evaluation practices leave major gaps in assessing AI's societal impacts, highlighting the urgent need for policies that promote developer transparency, strengthen independent evaluation ecosystems, and create shared infrastructure to aggregate and compare third-party evaluations in a consistent and accessible way.

🤗 Dataset Code