HUMAN RIGHTS CENTER
UC Berkeley School of Law

AN INTERNATIONAL ANALYSIS OF THE HUMAN RIGHTS IMPACTS OF LARGE LANGUAGE MODELS

In Law, Journalism, and Education

Front cover: Students learn math on iPads at the secondary school Gehrden in Hannover, Germany on March 7, 2019. Gehrden is one of the leading digital learning schools in Germany. Photo by Julian Stratenschulte/picture alliance via Getty Images.

HUMAN
RIGHTS
CENTER
UC Berkeley School of Law

AN INTERNATIONAL ANALYSIS OF THE HUMAN RIGHTS IMPACTS OF LARGE LANGUAGE MODELS

In Law, Journalism, and Education

August 2026

Betsy Popken
Vyoma Raman
Camille Chabot

HUMAN RIGHTS CENTER

The Human Rights Center pursues justice through science, technology, and law. Based at the world's top public university, we investigate human rights violations and international crimes; set standards for practitioners; and train the next generation of human rights defenders.

2224 Piedmont Avenue, Berkeley, CA 94720
Telephone: 510.642.0965 | Email: hrc@berkeley.edu | humanrights.berkeley.edu

CONTENTS

1. Introduction 3

2. Executive Summary 5

3. Methodology 7

3.1 Information Sources 7

3.1.1 Sector Selection 7

3.1.2 International Scope 7

3.2 Interdisciplinary Approach 8

3.2.1 Business and Human Rights 8

3.2.2 Qualitative Interviews 10

3.2.3 Model-Level Analysis of Human Rights Risks 11

3.3 Limitations 11

4. Actual and Potential Human Rights Impacts 13

4.1.1 Computational Foundations of Large Language Models 13

4.1.2 Social Impacts of Large Language Models 15

4.1.3 Governance of Generative AI 16

4.2.1 Background 18

4.2.2 International Stakeholder Interview Analysis 20

4.3.1 Background 21

4.3.2 International Stakeholder Interview Analysis 24

4.4.1 Background 26

4.4.2 International Stakeholder Interview Analysis 28

4.5.1 Geographic Differentiation 30

4.5.2 Sources of Variation in LLM Adoption 31

4.5.3 Cultural and Ideological Homogenization 32

4.5.4 Process and Purpose Disruption 33

4.5.5 Capitalist Incentives 34

4.5.6 Changes in Labor Practices 35

4.5.7 Human-LLM Interaction 36

5. Risk Assessment and Prioritization 37

5.1 Output-Level Rights Analysis 37

5.1.1 Approach 38

5.1.2 Hallucination in Legal Research 39

5.1.3 Election Reporting Quality 41
5.1.4 Adherence to Educational Standards 42
5.1.5 Toward Rights-Aligned Evaluation 44
5.2 Use-Level Rights Analysis 46
5.2.1 Risks 46
Most Salient Risks in Law 48
Most Salient Risks in Education 49
Most Salient Risks in Journalism 51
Most Salient Cross-Cutting Risk 52
5.2.2 Opportunities 52
Most Salient Opportunities in Law 54
Most Salient Opportunities in Education 55
Most Salient Opportunities in Journalism 56
Most Salient Cross-Cutting Opportunities 57
5.2.3 Takeaways 57
6. Identification of Appropriate Action for Remediation 58
1. Assess Human Rights Impacts of LLMs 60
2. Define Remediation Stages 62
1. Identifying Harms 62
2. Assigning Responsibility and Defining Remedy 63
3. Implementing Remedies 63
3. Regulate General-Purpose AI 63
4. Set Guidelines for the Use of LLMs in Professional Contexts 64
5. Diversify Training Data 64
6. Create User Feedback Mechanisms 65
7. Design LLM Interfaces Responsibly 65
8. Tailor Tools to Professional Tasks 66
9. Cultivate Critical AI Literacy 66
7. Conclusion 68
8. Acknowledgements 69
Appendix 72
A. Interview Format 72
B. Glossary 75

1. INTRODUCTION

Technological advancements have continually reshaped our understanding and application of human rights. Digital technologies have transformed personal communications, public discourse, and societal interactions globally. Generative artificial intelligence (GenAI) marks the next wave of technology-based transformation of personal, professional, and public life, with growing adoption across professional sectors and everyday activities. However, public-facing research on the specific human rights impacts of GenAI remains sparse, leaving critical gaps in our understanding of its societal implications.

The release of ChatGPT in late 2022 signaled a pivotal moment in the access to and adoption of large language models (LLMs), a prominent category of GenAI that generates text outputs. Subsequently, tools such as Anthropic's Claude and Google's Gemini further expanded access to these sophisticated models, transforming how people across various sectors engage with synthetic content. This accessibility has rapidly accelerated the deployment of LLMs in professional contexts, including in law, journalism, and education.

This report examines how the use of LLMs by legal professionals, journalists, and educators affects people's human rights, including the rights of clients, audiences, and students. To guide this assessment, we ask three central questions:

- How might the use of LLMs in law, journalism, and education negatively and positively impact human rights?
- How do the human rights risks and opportunities associated with the use of LLMs vary based on the user's location in the world and the social and economic conditions in which people in their communities live and work?
- What measures can be taken, and by which entities, to effectively mitigate the human rights risks and leverage the opportunities presented by LLMs?

Our analysis is guided by Sayash Kapoor's and Arvind Narayanan's idea of AI as "normal technology."[1] In other words, we view AI as a tool whose impacts depend on how people and organizations design, use, and oversee it. Its benefits and harms are shaped less by the technology itself than by the decisions people make about where and how it is used. Accordingly, we examine how LLMs affect human rights in the institutions where they are adopted—such as schools, news organizations, courts, and workplaces—and through the decisions of the people who build, deploy, govern, and use them. This differs from approaches that focus mainly on hypothetical future AI systems or assume that AI itself is the primary driver of change.

We use a mixed-methods approach that combines a human rights assessment with a human rights analysis of existing model evaluations. While current model evaluations provide valuable insights into technical performance, they often lack the granularity and alignment needed to substantiate claims about human rights impacts. In response, we release a companion piece that proposes a rights-aligned evaluation framework to better assess these impacts.[2] This methodological approach allows us to draw connections between model behaviors and real-world consequences, grounding both qualitative insights and quantitative metrics in a human rights framework.

This report is structured as follows: after this introduction and the executive summary (Section 2), we present our detailed methodology (Section 3). Then, we identify sector-specific and cross-cutting impacts of LLMs from a literature review and global stakeholder interviews (Section 4). We then synthesize our findings to articulate the human rights risks and opportunities that emerge from the use of LLMs (Section 5), culminating in tailored recommendations for different actors within the AI ecosystem (Section 6). We conclude by reflecting on our assessment (Section 7). We also include appendices with further details on our interview structure (Appendix A) and a glossary of terminology (Appendix B). If the terminology we use is unfamiliar to you, we suggest you begin by reading the glossary in Appendix B.

1 Narayanan, Arvind, and Sayash Kapoor. "AI as Normal Technology." *Knight First Amendment Institute*. April 15, 2025. https://knightcolumbia.org/content/ai-as-normal-technology.

2 Raman, Vyoma, Camille Chabot, and Betsy Popken. 2025. "Assessing Human Rights Risks in AI: A Framework for Model Evaluation". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8 (3):2107-20. https://doi.org/10.1609/aies.v8i3.36699.

2. EXECUTIVE SUMMARY

Objectives. Since the release of ChatGPT in late 2022, people around the world have increasingly used large language models (LLMs) in their work. This is an international and interdisciplinary analysis of the human rights risks and opportunities of LLMs being used in the fields of law, journalism, and education. The intended audience of this white paper is policymakers; industry groups, organizations, and entities in law, journalism, and education; and executives and practitioners within AI companies creating and distributing LLMs.

Methods. The assessment follows a mixed-methods approach. We conducted a literature review of the impacts LLMs have on human rights, and their specific impacts in the fields of law, journalism, and education. We analyzed regulations and guidance on artificial intelligence by states across the world. Drawing on the human rights impact assessment framework used as a due diligence tool under the UN Guiding Principles for Business and Human Rights (UNGPs), we blended stakeholder interviews with a literature review and our own analysis to identify rights-impacting trends, allocating scores to each human rights risk based on its saliency. Although not required in human rights impact assessments, we also included human rights opportunities we identified that resulted from peoples' use of LLMs in their work. We interviewed 56 people from 24 countries, including at least one expert and one practitioner from each sector on each continent, and representatives from corporations developing LLMs. We also analyzed existing sector-specific model evaluations to connect performance metrics to human rights impacts.

Findings. In the legal field, LLMs can improve access to justice by assisting litigants who do not have counsel, such as in small claims settings. However, reliance on LLMs by legal practitioners in high-risk settings can risk violating clients' due process and fair trial rights, as well as the right to remedy, if the responses generated by LLMs are incomplete or wrong. The use of LLMs by journalists can affect peoples' right of access to information. While providing assistance with writing and research can improve the productivity and efficiency of journalists operating in resource-constrained settings, LLM hallucinations can also lead to the spread of misinformation. In education, personalized learning systems make it easier

for educators to adequately assist students with individual challenges, but LLMs are not typically aligned with various educational standards. Across sectors, freedom of thought is at risk because peoples' overreliance on LLMs can stifle higher-level critical thinking. The rights to equality and nondiscrimination are also at risk. Because the data that models are trained on mostly originates from dominant, widely available sources, LLMs can promote cultural and ideological homogenization, privileging certain worldviews while marginalizing others.

Recommendations. To effectively mitigate the human rights risks posed by LLMs, different actors must jointly implement measures. Key recommendations include establishing clear remediation processes to identify, assign responsibility for, and remedy harms; requiring comprehensive human rights risk assessments across the AI lifecycle; and creating regulations that target specific high-risk use cases rather than banning LLMs altogether. We also recommend diversifying datasets to better represent global communities, designing refusal mechanisms and response tones based on the severity of human rights risks, and developing profession-specific LLMs to address field-specific challenges. Finally, LLM developers and deployers as well as industry groups and policymakers should collectively support the development of users' AI literacy.

3. METHODOLOGY

3.1 Information Sources

This human rights assessment is based on a variety of data, including publicly-available information on large language models (LLMs), such as journalistic articles describing uses of LLMs and related incidents; interviews with individuals around the world who work in law, journalism, and education; interviews with experts studying technology applied to these three fields; interviews with experts studying the societal impacts of generative AI; interviews with employees at companies that produce LLMs; and analysis of algorithmic audits used to evaluate the performance of LLMs in journalism, law, and education.

3.1.1 Sector Selection

We focus on the human rights impacts of LLMs when used by professionals in three fields: law, journalism, and education. All three professions have different roles (e.g., investigative reporter, sports writer, weather forecaster) and levels (e.g., editor, staff writer, producer) that leverage a diversity of skills to serve a broad community (e.g., readers and listeners). These sectors are linked to the work of the UC Berkeley Human Rights Center, a multidisciplinary institution whose research team has intimate knowledge of practice in each sector. Selecting three specific fields allowed us to conduct a deeper study of how the use of LLMs in a professional context can put certain human rights at risk, or create opportunities to advance human rights.

3.1.2 International Scope

We are conscious of our positionality at UC Berkeley, which is an elite, public university in the United States, adjacent to Silicon Valley. Given our location and the connection of the tech sector to UC Berkeley, we chose to decenter the Bay Area by engaging with several stakeholders in each habitable continent and analyzing the ways in which the applications and impacts of LLMs vary in different communities and across geopolitical contexts. Importantly, while LLMs may be more widely-used in some places than in others, the

volume of usage is not necessarily associated with the types of usage or the magnitude of human rights impacts. A limitation we face is that we recruited interviewees through email and conducted all interviews in English, resulting in a sample of people who are digitally engaged and have been trained in the English language.

3.2 Interdisciplinary Approach

UC Berkeley's Human Rights Center is a multidisciplinary research center that uses science, technology, and law in the service of human rights. In addition to a shared background in human rights, our research team members contributed different areas of expertise to this project. Betsy Popken is a lawyer and educator, and has worked at the intersection of the field of Business and Human Rights and technology; Vyoma Raman is a computer scientist who specializes in algorithmic justice; and Camille Chabot is a social scientist with a background in international relations. Building on this interdisciplinary background, we model how human rights practitioners can work alongside scientists and engineers building machine-learning-based models to better integrate human rights into LLMs. Our methodology reflects this multidisciplinary approach, blending traditional elements of a human rights assessment with using grounded theory to code interviews and analyze model evaluations through a human rights lens. To enhance the quality of this assessment, we assembled an advisory group of experts in computer science, Business and Human Rights, journalism, education, and law to help guide our work.

3.2.1 Business and Human Rights

A human rights-based approach provides a well-defined framework based on existing laws and internationally agreed-upon principles. International human rights law provides valuable guidance on the definition and protection of rights relevant to LLMs, such as non-discrimination and privacy, and is adept at balancing complex situations involving competing rights.[3] Moreover, human rights law strikes a balance between universal rights and consideration for local contexts and culture, establishing a global standard for safeguarding fundamental rights while allowing governments some margin of appreciation in implementing these rights.[4]

3 Fjeld, Jessica et al. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." SSRN Electronic Journal (2020). Available: http://dx.doi.org/10.2139/ssrn.3518482.

4 The 'margin of appreciation' is enshrined in the preamble to the ECHR and has comparable impacts in the UN human rights system. Jones, Kate. 2023. "AI Governance and Human Rights Resetting the Relationship." https://www.chathamhouse.org/sites/default/files/2023-01/2023-01-10-AI-governance-human-rights-jones.pdf.

Several international human rights instruments provide the foundation for this framework. These include the Universal Declaration of Human Rights (UDHR),[5] which sets out fundamental rights and freedoms; the International Covenant on Civil and Political Rights (ICCPR),[6] which protects civil and political rights such as freedom of expression and privacy; the International Covenant on Economic, Social and Cultural Rights (ICESCR),[7] which addresses rights related to work, education, and an adequate standard of living; the Convention on the Rights of Persons with Disabilities (CRPD),[8] which safeguards the rights of persons with disabilities; and the Convention on the Rights of the Child (CRC),[9] which ensures the protection and participation of children. Together with others, these instruments establish a globally-recognized framework for identifying, assessing, and protecting human rights in contexts such as the deployment of LLMs.

The second pillar of the UN Guiding Principles on Business and Human Rights (UNGPs) sets a standard for companies to conduct human rights due diligence to identify and address the adverse human rights impacts of their business operations. The UN B-Tech Project has provided guidance for technology and AI companies on how to translate human rights due diligence in the UNGPs to mitigate the human rights risks of their products and services. A human rights impact assessment is one due diligence tool and serves as a basis of our methodology. The UNGPs call for analyzing the salience of human rights risks, which is assessed across three dimensions: scale (the gravity or seriousness of the potential harm to human rights), scope (the number of individuals who could be affected by the impact), and remediability (how easily the harm can be remedied or reversed, from the perspective of those affected).[10] We combined this with an assessment of likelihood (the probability that a particular human rights impact will actually occur).[11]

---

5 United Nations. *Universal Declaration of Human Rights*. Paris: United Nations, 1948. https://www.un.org/en/about-us/universal-declaration-of-human-rights.

6 United Nations. *International Covenant on Civil and Political Rights*. New York: United Nations, 1966. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

7 United Nations. *International Covenant on Economic, Social and Cultural Rights*. New York: United Nations, 1966. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights.

8 United Nations. *Convention on the Rights of Persons with Disabilities*. New York: United Nations, 2006. https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html.

9 United Nations. *Convention on the Rights of the Child*. New York: United Nations, 1989. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child.

10 Salience refers to the significance or prominence of a human rights risk. The most salient risks are those most likely to cause serious negative impacts on individuals or communities, either because of their scope, scale, likelihood, or lack of remediability.

11 BSR (Business for Social Responsibility), Human Rights Assessments: Identifying Risks, Informing Strategy (San Francisco: BSR, 2013), https://www.bsr.org/en/reports/human-rights-assessments-identifying-risks-informing-strategy.

Human rights can be affected by adversarial actors who exploit AI systems for malicious purposes such as surveillance, disinformation, or cyberattacks, as well as benign actors, whose application of AI toward harmless goals may inadvertently cause harm. We focused on the latter, taking a human-centric approach that identifies pervasive but often unnoticed impacts across social groups and professional contexts. We used the four saliency factors to identify relevant risks and opportunities for LLM. While an analysis of opportunities is not common in human rights assessments, the inclusion of opportunities more faithfully illustrates how the perceptions and realities of the benefits and drawbacks of LLMs were uneven across regions in the world.

3.2.2 Qualitative Interviews

We conducted interviews with legal professionals, journalists, and educators to understand how people view or apply LLMs in their work. Over the span of 15 months, we interviewed 56 people from 24 countries, including at least one expert and one practitioner in each sector on each inhabited continent. We recognize that much has changed in the LLM landscape during this time and have endeavored to minimize the impact of this on our findings by blending several data types. We used a combination of purposive and snowball sampling to recruit interviewees. First, we identified potential interviewees based on the authors of or people described in journalistic and academic articles we read during our literature review. We reached out to them over email and LinkedIn. Then, we asked respondents to refer us to others in their network whose backgrounds were relevant to our focus. We made a concerted effort to interview professionals at various career stages with differing levels of experience.

Professional Field	Role[12]	Region
Journalism: 14	Field Expert: 19	Asia: 5
Law: 15	Field Practitioner: 22	Europe: 7
Education: 12	AI Expert: 7	Middle East and North Africa (MENA): 5
	AI Practitioner: 8	North America: 18
		Oceania: 8
		South America: 6
		Sub-Saharan Africa: 7

Interviewee composition by professional field, role, and region.

Upon recruiting participants, we took a systematic approach to understanding how each sector is affected by LLMs. First, we asked each type of interviewee a standardized set of questions concerning use cases, opportunities, risks, and practices for mitigating risks; workplace approaches to LLMs; labor market changes; and AI governance (Appendix

12 A field expert is a person who researches, writes, and speaks about one of the three fields, whereas a field practitioner is a person who practices in one of the three fields. An AI expert is a person who researches, writes, and speaks about AI, whereas an AI practitioner is a person who works in the field of AI.

A). Most of the interviews were recorded for internal purposes during the analysis. We applied grounded theory to code the interviews. Our research team used open coding to identify various themes based on our transcripts and recordings. We then grouped these open codes into axial codes[13] that represent different recurring concepts. Finally, we aggregated thematically similar axial codes into larger categories that reflected key patterns and contrasts in the data. We reviewed and refined these categories multiple times before integrating them into a cohesive interpretation and analysis.

3.2.3 Model-Level Analysis of Human Rights Risks

In addition to adopting law and social science approaches to our research, which are popular in human rights impact assessments, we incorporated computer science perspectives into our analysis. This involved applying knowledge of the limitations and capabilities of LLMs to inform our understanding of how they might create certain human rights risks through factors like bias, stochasticity, and hallucination. It also shaped our recommendations to different actors to mitigate human rights risks and maximize opportunities by ensuring that our recommendations are technically feasible.

Furthermore, we used our knowledge of machine-learning evaluation techniques to analyze findings from existing approaches and connect them to human rights. First, we selected one academic paper describing an evaluation of LLMs used in each of journalism, education, and law. We selected these based on their relevance to rights-affecting activities, which were identified through our interviews. Then, we established a link between the natural language processing (NLP) task being evaluated and a human rights risk, and analyzed the relationship of the evaluation metrics to the principles of scope, scale, and likelihood.[14] Finally, we used these connections and the findings to determine the level of risk each NLP task poses to relevant human rights.

3.3 Limitations

Given the large scope of this research area and the capacities and resources of our team, we had to make a number of decisions that impact the representativeness and timeliness of our findings.

First, our research is not fully representative. We intentionally depict our assessment as international rather than global due to the impossibility of representing the entire world in our data and conclusions. To ensure some level of geographic diversity among our

13 A type of code specified within grounded theory to systematically identify relationships between open codes.

14 We omitted remediability from our analysis of model performance metrics because the ability to remediate a problem is a capacity of the organizational system in which the model operates, not a property of model outputs.

participants, we interviewed 56 people living in 24 countries across seven regions of the world. We chose to do this so that our results reflect a range of human rights impacts that occur across various regional and professional contexts. However, we do not claim that these individuals represent all of the perspectives of people in their country with their expertise, or even the majority perspective of that population. We also conducted all interviews in English, meaning that our interviewees were those who had sufficient exposure to English to participate in interviews.

Second, the timeframe of this work may impact its findings. We started conducting interviews in the summer of 2023, nearly a lifetime ago when considering the pace of advancements in AI development and research since that time. As such, the experiences and observations reported by our interviewees at the beginning of this timeframe may no longer reflect their perspectives. To counterbalance this possibility, we have included more recent literature and discourse when identifying impacted rights. We also asked interviewees future-facing questions to understand their perceptions of potential impact, and we saw some of these predictions manifest. When we started this work, LLMs had recently gained prominence in mainstream public awareness. While a variety of other general-purpose models have developed since then, we retained our focus on LLMs because we found them to be continually relevant across a variety of rights-impacting professions.

4. ACTUAL AND POTENTIAL HUMAN RIGHTS IMPACTS

4.1 Literature Review

4.1.1 Computational Foundations of Large Language Models

Recent advances in natural language processing (NLP) have resulted in large language models (LLMs), which generate remarkably fluent and context-aware text while also raising new challenges around reliability, safety, and ethical use. These models are considered "large" because they contain billions of parameters[15] that help them do well at relevant tasks, which we define as activities such as translating text, answering questions, summarizing documents, or generating conversational responses. LLMs are made possible by transformers,[16] a machine-learning architecture that is extremely effective at processing enormous amounts of text in parallel, making them more efficient to train and better at handling context. This has resulted in emergent capabilities of LLMs to perform tasks they were not explicitly trained for with sufficient prompt engineering.

LLMs are built through a multi-step process that begins with the collection of text data from sources like Common Crawl,[17] which contain petabytes of web data (one petabyte is equivalent to one million gigabytes). The first step is pretraining, where the model learns to predict words or phrases by learning language patterns, grammar, and context. After pretraining, models undergo fine-tuning, where they are trained further on a specific dataset to specialize their abilities. One example of this is instruction tuning, where models are fine-tuned to follow specific prompts or instructions.

15 Internal weights in the neural network.

16 Ashish Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems* 30 (2017): 5998–6008, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

17 *Common Crawl (nonprofit organization)*, https://commoncrawl.org/.

In November 2022, ChatGPT was released,[18] allowing the general public to interact with a powerful LLM for the first time. A variety of interfaces have been developed for LLMs: ChatGPT is an example of a chatbot that users can prompt with queries to receive natural language responses. This environment can sometimes cause people to anthropomorphize[19] the models due to ChatGPT's strong grasp of natural language and command of rhetoric through tone, persuasiveness, confidence, and hedging. Individuals may heavily rely on LLMs, especially if the confidence expressed by LLMs is mistaken for their correctness,[20] due to the convincing language LLMs use in their responses.[21,22]

The models underlying chatbots can also be integrated into larger systems through API calls (requests that one piece of software makes to another piece of software through an Application Programming Interface or API). In NLP contexts, a "system" refers to an integrated collection of software components, such as databases, model outputs, and user interfaces, that work together to perform complex functions. These enable developers to extend capabilities further and create LLM agents,[23] which are able to perform complex tasks by dynamically interacting with data and systems to automate workflows, interact with databases, and analyze large volumes of text. For example, as productivity assistants, LLM-powered agents can coordinate across email, calendars, and task managers to draft messages and schedule meetings.

Despite the demonstrated abilities of LLMs to produce convincing text, they tend to be "black boxes" with internal workings that are not fully understood.[24] This means it can be challenging to explicitly control LLM outputs and prevent them from producing false, offensive, or otherwise harmful outputs. AI fairness and safety are concerned with measuring harm in model responses and attempting to mitigate problematic outputs. This is done using model evaluations, which take multiple forms: techniques like red teaming[25] draw

18 OpenAI, "Introducing ChatGPT," *OpenAI*, https://openai.com/index/chatgpt/.

19 Murray Shanahan, "Talking about Large Language Models," Communications of the ACM 67, no. 2 (February 2024): 68–79, https://doi.org/10.1145/3624724.

20 Sanyam Kapoor et al., Large Language Models Must Be Taught to Know What They Don't Know, NeurIPS 2024, https://proceedings.neurips.cc/paper_files/paper/2024/file/9c20f16b05f5e5e70fa07e2a4364b80e-Paper-Conference.pdf.

21 Sunnie S. Y. Kim et al., "'I'm Not Sure, But...': Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust," FAccT 2024, https://doi.org/10.1145/3630106.3658941.

22 Kaitlyn Zhou et al., *Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty*, ACL 2024, https://doi.org/10.18653/v1/2024.acl-long.198.

23 Z. Xi, W. Chen, X. Guo, et al., "The Rise and Potential of Large Language Model Based Agents: A Survey," Science China Information Sciences 68 (2025): 121101, https://doi.org/10.1007/s11432-024-4222-0.

24 Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys* 51, no. 5, Article 93 (September 2019): 1–42, https://doi.org/10.1145/3236009.

25 Deep Ganguli et al., *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, arXiv preprint, September 2022, https://arxiv.org/abs/2209.07858.

on cybersecurity approaches to stress-test models for dangerous outputs, and algorithmic audits[26] analyze model weights, outputs, and confidence scores in a systematic way.

Technical solutions have been developed to address certain types of problems. LLMs have a tendency to "hallucinate" by sharing or generating false information, since they are not optimized for truth. To combat this, retrieval-augmented generation (RAG)[27] can be used to ground the model's responses in verified external data. To reduce the tendency of LLMs to generate sensitive content, including sexually explicit material, hate speech, or self-harm, LLMs can be further modified using alignment techniques like preference tuning to generate text reflecting human values and preferences. Content filters (e.g., in Azure AI)[28] are also applied in systems using LLMs to block the model from receiving input or producing output that contains material that violates certain guidelines, such as hate speech or the promotion of violence.

4.1.2 Social Impacts of Large Language Models

Researchers have studied the impacts of generative AI on society from the initial stages of development.[29] For instance, scholars have argued that LLM development can manifest data colonialism, in which the extraction of data and the concentration of economic value reproduce historical patterns of unequal power. This dynamic may be reinforced when LLM developers—many of whom are headquartered in the United States—depend on low-cost data annotation and other AI labor in lower-income countries, raising concerns about labor rights, equitable benefit-sharing, and accountability. In 2021, OpenAI began outsourcing the task of labeling unsafe content to Kenyan workers working for the equivalent of $1.32 to $2 per hour.[30] LLMs have also been shown to take a significant environmental toll. Training a single BERT model, which is much smaller than more recent LLMs, is estimated to consume as much energy as a trans-American flight, and data centers that power LLMs

26 Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes, "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)* (New York: Association for Computing Machinery, 2020), 33–44, https://doi.org/10.1145/3351095.3372873.

27 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)* (Red Hook, NY: Curran Associates Inc., 2020), Article 793, 9459–9474. https://dl.acm.org/doi/abs/10.5555/3495724.3496517.

28 Microsoft, *"Content Filtering Overview," Azure AI Services – OpenAI*, https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter?tabs=warning%2Cuser-prompt%2Cpython-new.

29 Bender, Emily M. et al. "On the Dangers of Stochastic Parrots." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021). Available: http://dx.doi.org/10.1145/3442188.3445922.

30 Perrigo, Billy. "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic." *Time*, January 18, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/.

consume significant amounts of electricity and water while increasing carbon emissions.[31, 32] Harms to the environment impact the human rights to health, livelihood, and more, and are unevenly externalized to historically disadvantaged communities.[33]

The consequences of LLMs' inequities also extend into their deployment. These models rely on datasets with inherent biases that propagate through the training process, resulting in models perpetuating stereotypical and discriminatory associations based on factors such as race, ethnicity, gender, and ability. For example, widely used LLMs have been found to exhibit caste bias toward Dalit job applicants, questioning their qualifications and acculturation.[34] When deployed in professional settings, such biases can cause concrete harm. This can occur in contexts such as criminal decision-making, where LLMs have been found to treat speakers of African American Vernacular English more harshly than speakers of Standard American English.[35]

At the same time, LLMs hold significant potential for social empowerment. For instance, in New Zealand, LLMs have been instrumental in helping to preserve and revitalize the Māori language. This approach aligns with the principle of data sovereignty, where communities gain control over their data to prevent unauthorized use. In this case, Māori-run media organization Te Hiku Media collaborated with the local Māori community to transcribe hours of audio from elders, and train a language tool based on the recordings.[36]

4.1.3 Governance of Generative AI

The development of AI technology has been accompanied by a rise in AI governance.[37] Between 2023 and 2024, references to artificial intelligence in legislative proceedings grew

---

31 Bender, Emily M. et al. "On the Dangers of Stochastic Parrots." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021). http://dx.doi.org/10.1145/3442188.3445922.

32 Regilme, Salvador Santino F. "Artificial Intelligence Colonialism: Environmental Damage, Labor Exploitation, and Human Rights Crises in the Global South." *SAIS Review of International Affairs* 44, no. 2 (2024): 75-92. https://dx.doi.org/10.1353/sais.2024.a950958.

33 *Ibid.*

34 Dammu, Preetam Prabhu Srikar, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. "'They Are Uncultured': Unveiling Covert Harms and Social Threats in LLM-Generated Conversations." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 20339–20369. Miami, FL: Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.1134.

35 Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. "AI Generates Covertly Racist Decisions About People Based on Their Dialect." *Nature* 633, no. 8028 (2024): 147–154.

36 Hao, Karen. 2022. "A New Vision of Artificial Intelligence for the People." MIT Technology Review. April 22, 2022. https://www.technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people/.

37 We conducted a thorough review of AI governance for each region of the world, covering national strategies, legislation, and guidelines, and then distilled this information to highlight key trends and notable gaps.

by 21.3% across 75 countries, representing a ninefold surge compared to 2016.[38] Increased regulation has been accompanied by recent government investment in resources to gain or grow AI leadership.[39] In 2017, Canada launched the first national AI strategy, and dozens more have done so since.[40,41] In 2023, half of the AI strategies that were announced or published came from low and lower-middle-income countries,[42] a significant change from previous years that were dominated by high and upper-middle-income countries.[43] Despite this diversification, there is still a large AI governance gap between low and middle-income countries and high-income countries, which continue to lead the development of AI legislation both in the number of laws passed and in their degree of influence.[44]

The past few years have also seen a rise in international collaboration on AI governance, evidenced by the release of proposed frameworks such as the G7's *International Guiding Principles*[45] and the *International Code of Conduct for Organizations Developing Advanced AI Systems*.[46] These documents add on to already established multilateral frameworks such as the Organisation for Economic Co-operation and Development's (OECD) 2019 *AI Principles* and UNESCO's 2021 *Recommendations on the Ethics of Artificial Intelligence*.[47]

38 Stanford University. *AI Index Report 2025*. Stanford: Stanford HAI, 2025.

39 Stanford University. *AI Index Report 2026*. Stanford: Stanford HAI, 2026.

40 Government of Canada. *Pan-Canadian Artificial Intelligence Strategy*. 2017. https://ised-isde.canada.ca/site/ai-strategy/en.

41 Oxford Insights. "GovernmentAI Readiness Index 2024," December 20, 2024. https://oxfordinsights.com/ai-readiness/ai-readiness-index/.

42 Our income classification is based on the World Bank's Group country classifications by income level for 2024–2025.

43 Oxford Insights. "GovernmentAI Readiness Index 2024," December 20, 2024. https://oxfordinsights.com/ai-readiness/ai-readiness-index/.

44 *Ibid.*

45 G7. *Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems*. Hiroshima: G7, 2023. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system. The Group of Seven (G7) is an intergovernmental organization composed of seven of the world's most advanced economies: Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States. All G7 members are classified as high-income countries by the World Bank.

46 G7. *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*. Hiroshima: G7, 2023. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems.

47 UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. SHS/BIO/PI/2021/1. Paris: UNESCO, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137. The United Nations Educational, Scientific and Cultural Organization (UNESCO) is a specialized agency of the United Nations with 194 member states and 12 associate members as of 2024. Its membership includes countries from all income levels and regions, making it one of the most globally representative intergovernmental organizations. Member states work together to promote international collaboration in education, science, culture, and communication.

Regional collaboration has also expanded, as illustrated by agreements such as the Santiago Declaration[48] in Latin America.

Despite these efforts, there is a notable lack of binding legislation specifically addressing generative AI technology. China adopted such rules for public-facing applications in 2023; the European Union adopted the AI Act in 2024, which regulates generative AI through a tier-based system; South Korea's AI Framework Act, enforced in 2026, focuses on high-impact AI systems; and Kyrgyzstan, Peru, and the Philippines also have binding legislation. Most countries, however, opt instead for non-binding guidance consisting of frameworks on how AI can be used responsibly by companies and individuals.

4.2 Law

4.2.1 Background

Legal practice plays a pivotal role in creating and maintaining a just and orderly society. It encompasses litigation, transactions, and investigation work, as well as various fields such as employment, criminal, and intellectual property law, and its practice differs across jurisdictions.

Legal practice intersects with human rights in several significant ways. Laws can either promote or endanger human rights, while lawyers play a crucial role in upholding equality before the law, the right to remedy,[49] and access to due process.[50] Inadequate legal representation can lead to arbitrary arrests or concerns over the impartiality of judgments, negatively impacting human rights. Access to justice, another human right, can be challenging due to legal costs and other factors worldwide.

The legal profession's relationship with technology has long been characterized by a mix of enthusiasm and caution. Although technology has transformed every segment of the legal

48 UNESCO Office Santiago and Regional Bureau for Education in Latin America and the Caribbean. *Declaration of Santiago 2024: Extraordinary Meeting of Ministers of Education from Latin America and the Caribbean – Towards a Regional Framework for Educational Reactivation, Recovery, and Transformation, January 25–26, 2024, Santiago, Chile*. Santiago: UNESCO, 2024. https://unesdoc.unesco.org/ark:/48223/pf0000388789_eng.

49 The right to remedy refers to the entitlement of individuals to seek and obtain redress when their rights are violated, including access to effective judicial or administrative procedures. It is enshrined in international human rights law, notably in Article 2(3) of the International Covenant on Civil and Political Rights (ICCPR), which obliges states to ensure that individuals have accessible, effective remedies when their rights are infringed.

50 The right to due process and fair trial guarantees that every individual is entitled to a fair and public hearing by an independent and impartial tribunal in the determination of their rights, obligations, and any criminal charges. See United Nations, *Universal Declaration of Human Rights*, Article 10, 1948. https://www.un.org/en/about-us/universal-declaration-of-human-rights.

ecosystem, lawyers have often been slow to adopt cutting-edge technologies due to concerns about potential conflicts with ethical requirements.[51] Recent evidence suggests that AI adoption by U.S. federal judges may be similarly slow: in 2025, only 22.4% of them used AI tools at least weekly, and over a third had never used them in their work.[52] However, there are cases of the legal industry transitioning from a lawyer-centric field rooted in precedent into an interdisciplinary, tech-enabled competitive marketplace powered by "legal tech."[53] In a 2024 UNESCO survey, 44% of legal professionals from 96 countries said they used LLMs for professional purposes including searching (44%), document drafting (28%), and brainstorming (14%).[54] A 2025 survey of law firm employees in the U.S. found that 31% of legal professionals used LLMs in their work, despite low firm-wide adoption.[55] A Reuters survey also indicates that attorneys working in law firms are most optimistic about using LLMs for professional tasks, while lawyers working for the government tend to be most pessimistic about professional use of LLMs.[56]

The UN Special Rapporteur on the Independence of Judges and Lawyers has highlighted the need for dedicated AI governance tools to ensure that the deployment of AI in judicial and legal contexts aligns with human rights, due process, and the rule of law.[57] Recommendations include that legal professionals be wary of techno-solutionism in the use of AI in judicial systems, that AI adoption remain under the responsibility and oversight of the judicial branch, and that AI does not prevent access to a human judge and lawyer.[58]

---

51 Fisher, Keith Robert. "ABA Ethics Opinion on Generative AI Offers Useful Framework." *Business Law Today*, American Bar Association, October 3, 2024. https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-october/aba-ethics-opinion-generative-ai-offers-useful-framework/.

52 Jaitley, Anika, Daniel W. Linna Jr., Hon. Xavier Rodriguez, V. S. Subrahmanian, and Siyu Tao. "Artificial Intelligence in Federal Courts: A Random-Sample Survey of Judges." *The Sedona Conference Journal* 27, forthcoming 2026. https://www.lawnext.com/wp-content/uploads/2026/03/Artificial_Intelligence_in_Federal_Courts_preprint.pdf.

53 O'Leary, Chris. "Legal Technology's Authoritative Guide on Use Cases and Benefits." *Thomson Reuters Legal*, September 9, 2024. https://legal.thomsonreuters.com/blog/technology-in-law-is-the-new-norm/.

54 Gutiérrez, Juan David. *UNESCO Global Judges' Initiative: Survey on the Use of AI Systems by Judicial Operators.* Paris: UNESCO, 2024. CI/DIT/2024/JI/01 Rev. 14 pp. https://unesdoc.unesco.org/ark:/48223/pf0000389786.

55 MyCase. "The Legal Industry Report 2025." *Law Technology Today*, American Bar Association, May 6, 2025. https://www.americanbar.org/groups/law_practice/resources/law-technology-today/2025/the-legal-industry-report-2025/.

56 Thomson Reuters Institute. *2025 Generative AI in Professional Services Report*. May 15, 2025, 6, 7. https://www.thomsonreuters.com/en/reports/2025-generative-ai-in-professional-services-report.

57 Gutiérrez, Juan David. *Guidelines for the Use of AI Systems in Courts and Tribunals*. Paris: UNESCO, 2025. https://doi.org/10.58338/LIEY8089.

58 United Nations. *AI in Judicial Systems: Promises, Pitfalls, and People-Centred Approaches*. Report of the Special Rapporteur on the Independence of Judges and Lawyers, Margaret Satterthwaite. UN Doc. A/80/169. New York: United Nations, 2025. https://docs.un.org/en/A/80/169.

4.2.2 International Stakeholder Interview Analysis

The 15 legal professionals we interviewed were dispersed across private practice, academia, regulatory and policy bodies, advisory roles, and the judiciary, working on technology, media and telecommunications, AI, data governance, and regulatory compliance across 10 countries, with levels of seniority ranging from junior professionals to senior practitioners and judges. Our interviewees most frequently reported using general-purpose tools, especially ChatGPT and Copilot, as well as tools designed specifically for legal purposes, including Harvey and CoCounsel. Some interviewees also used LLMs created by their law firms based on Microsoft or OpenAI platforms.

Our interviews revealed a key tension regarding the use of LLMs by legal professionals. While the legal field tends to be more risk-averse and legal professionals are often more cautious about using LLMs than journalists and educators, firms are experiencing pressure from clients to lower the cost of their legal services, accelerating the use of LLMs in this field. LLMs are used rather extensively in the legal profession, but their typical use tends to be limited to certain tasks that are repetitive and simple, such as summarizing documents, drafting contracts and emails, and translating documents into multiple languages. Several interviewees pointed out the limited capacity of LLMs to perform higher-level tasks such as legal analysis, which they noted are better performed by lawyers:

> *I use LLMs often, but for limited tasks. I use ChatGPT as a language assistant since I need to read legal documents from different countries, it helps me get a sense of the content. I also use it to summarize documents and sometimes to draft templates. I don't use it for legal analysis; its performance is too weak.*

Others noted that the use of LLMs in law has the potential to increase litigants' access to justice. LLMs can be used as a preliminary option for individuals to assess whether or not they need legal counsel. LLMs can also be used by clients to help simplify and translate legal documents, and to assist with simple cases like small claims, thus democratizing access to justice for people with limited financial means. One expert highlighted these advantages:

> *Here in Singapore, the judiciary has announced a trial project where LLMs are used in court to provide legal information to litigants in small claims civil tribunals. No lawyers are involved in these tribunals, and civilians may be unclear about the process of filing documents.*

However, LLMs can also result in substantial human risks when used in the legal field. For instance, the ability of well-resourced international firms to build their own proprietary AI and better protect their clients' data creates a divide with low-resourced firms. The latter usually rely on open-source models, which are less expensive but also less private. Further, because LLMs can perform simple and repetitive tasks quickly and efficiently, they may reduce professional training opportunities. A practitioner from Oceania raised concerns that junior lawyers may miss out on learning experiences like drafting contracts or summarizing legal documents:

Before, junior lawyers would repeatedly do tasks like reading legal documents and drafting clauses. [I would ask them why they would perform each task to make sure they understood the purpose of what they were doing.] Now, LLMs are replacing that repetition and taking this knowledge away from lawyers. The question is, how to train a team then? I cannot ask my team to repeatedly draft the same clause because now, a tool can do it, so how should I train them to work alongside the tool, and make sure they understand the "why"? Part of the challenge is that clients want us to use the tool because it is more efficient, but at the same time, we have a moral duty to think about how to train the next generation of lawyers.

Additional risks come when high-impact legal professionals, particularly judges, use LLMs to help them make decisions in cases. Several interviewees reported that some judges use LLMs to assess legal questions, draft rulings, and otherwise lessen their workload. This use could have a substantial impact on litigants' right to a fair trial by influencing judges' decisions or leading them to overlook certain elements of a case. This risk is particularly acute when judges' technical understanding of LLMs is low and when judges do not need to report whether they used LLMs in their decision-making.

4.3 Journalism

4.3.1 Background

Journalists research and produce stories to inform the public with accurate, up-to-date information about events; investigate and report on injustices and misconduct; and educate the public on topics and issues, thus fostering a more informed public discourse. The stages of the journalistic process are sometimes divided into newsgathering, which involves collecting information about a particular event or topic; news production, which involves converting the information into a cohesive story; verification, which involves independently fact-checking the story; and distribution, which involves delivering the story to the audience.[59]

Journalism plays a vital role in upholding human rights. By keeping the public informed about global events and individuals, the industry is a key facilitator of the right of access to information.[60] Journalism also enables the human rights of free expression, opinion, and belief and serves as an instrument to uncover human rights violations of various kinds

59 Hannes Cools and Nicholas Diakopoulos, "Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities," *Journalism Practice*, August 26, 2024. https://doi.org/10.1080/17512786.2024.2394558.

60 The right of access to information guarantees that individuals can seek, receive, and impart information held by public authorities. It is essential for transparency, accountability, and meaningful participation in public life. See United Nations General Assembly. 1948. Universal Declaration of Human Rights, Article 19. See https://www.un.org/en/about-us/universal-declaration-of-human-rights. Office of the United Nations High Commissioner for Human Rights. 1966. *International Covenant on Civil and Political Rights*, Article 19. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights.

and broadly disseminate information about them. Various bodies within the journalism industry, such as the Society for Professional Journalists, establish journalistic ethics, which prioritize truth, fairness, and independence. The human right to access information is at risk when the local press is defunded, stories are siloed and filtered based on the political or other leanings of the reader, and when young audiences consume most of their news through social media influencers. When governments intervene in the operations of newsrooms, journalistic ethics and the rights of freedom of expression and access to information can be threatened by censorship, disinformation, and propaganda.

The journalism industry is often characterized as fast-paced and competitive, which makes rapid research and content production essential, and incentivizes the adoption of new tools that promote efficiency, even as newsrooms prioritize accuracy and impartiality and struggle to invest in expertise in emerging technologies.[61,62,63] Journalists have a long history of using AI tools across the stages of news production even before the advent of LLMs, including monitoring and identifying trends that could turn into new stories (e.g., Reuters News Tracer[64]), filtering user-generated content to find sources, grammar and spell checking, translating and tagging content before publishing, and moderating comments on articles.[65,66,67,68,69] Due to their utility for writing and related tasks, LLMs are now used to write content, summarize news events, answer questions about documents, identify relevant content, promote articles on social media, and curate newsletters, freeing journalists to focus more time on research and story generation. However, some journalists report that LLMs

61 Amy Slaffert, "Why the 'Scoop' Mentality Is Bad for News," *Knight Digital Media Center* (blog), August 2011. http://www.knightdigitalmediacenter.org/blogs/slaffert/2011/08/why-scoop-mentality-bad-news.html.

62 Sadia Jamil, "Artificial Intelligence and Journalistic Practice: The Crossroads of Obstacles and Opportunities for the Pakistani Journalists," *Journalism Practice* 15, no. 1 (July 7, 2020): 1–23. https://doi.org/10.1080/17512786.2020.1788412.

63 *AI, Journalism, and Public Interest Media in Africa*, International Media Support, June 2023. https://www.mediasupport.org/publication/ai-journalism-and-public-interest-media-in-africa/.

64 "Is It News or Noise?" *Thomson Reuters 2016 Annual Report*, Thomson Reuters, 2016. https://archive.annual-report.thomsonreuters.com/2016/is-it-news-or-noise.html.https://archive.annual-report.thomsonreuters.com/2016/is-it-news-or-noise.html.

65 Sylvia M. Chan-Olmsted, "A Review of Artificial Intelligence Adoptions in the Media Industry," *International Journal on Media Management* 21, no. 3 (November 25, 2019): 193–215. https://doi.org/10.1080/14241277.2019.1695619.

66 Jonathan Stray, "Making Artificial Intelligence Work for Investigative Journalism," *Digital Journalism* 7, no. 8 (2019): 1076–1097, https://doi.org/10.1080/21670811.2019.1630289.

67 Charlie Beckett and Mira Yaseen, *Generating Change: A Global Survey of What News Organisations Are Doing with Artificial Intelligence* (London: JournalismAI at LSE, 2023). https://www.journalismai.info/research/2023-generating-change.

68 Hannes Cools and Nicholas Diakopoulos, "Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities," *Journalism Practice*, August 26, 2024. https://doi.org/10.1080/17512786.2024.2394558.

69 Charlie Beckett and Mira Yaseen, *Generating Change: A Global Survey of What News Organisations Are Doing with Artificial Intelligence* (London: JournalismAI at LSE, 2023). https://www.journalismai.info/research/2023-generating-change.

are less useful for producing content than they expected due to issues such as inaccuracies, references to improper source material, and editorializing.[70,71]

Despite these challenges, 77% of journalists worldwide reported using LLMs in 2024,[72] and media organizations have attempted to modify internal practices to integrate them further. *CNET*[73] and *News Corp*[74] use LLMs to write articles that are then reviewed by human editors. *The India Today*,[75] *Doordarshan*,[76] and *The Garden Island newspaper*[77] have experimented with AI avatars that deliver news broadcasts. However, these approaches have had a human cost in some parts of the industry. *Buzzfeed* laid off 12% of its staff in December 2022[78] and *Business Insider* laid off 21% of theirs in May 2025,[79] both with the intention of using LLMs to replace a subset of their labor. A number of news organizations have additionally negotiated deals with LLM companies (e.g., OpenAI with *AP*,[80] *The Atlantic*,[81] and *Axios*;[82]

70 Jon Keegan, "I Used ChatGPT as a Reporting Assistant. It Didn't Go Well," *The Markup*, March 16, 2024, https://themarkup.org/hello-world/2024/03/16/i-used-chatgpt-as-a-reporting-assistant-it-didnt-go-well.

71 BBC, *Research into AI Assistants*, BBC Research & Development. https://www.bbc.co.uk/aboutthebbc/documents/bbc-research-into-ai-assistants.pdf.

72 Muck Rack, *The State of Journalism 2025*.

73 Frank Landymore, "CNET Is Quietly Publishing Entire Articles Generated by AI," *Futurism*, updated January 15, 2023. https://futurism.com/the-byte/cnet-publishing-articles-by-ai.

74 Amanda Meade, "News Corp Using AI to Produce 3,000 Australian Local News Stories a Week," *The Guardian*, July 31, 2023. https://www.theguardian.com/media/2023/aug/01/news-corp-ai-chat-gpt-stories.

75 M. Stanly, "AI Journalism in India: Did India Enter a New Era of Broadcasting?" *INDIAai*. https://indiaai.gov.in/article/ai-journalism-in-india-did-india-enter-a-new-era-of-broadcasting.

76 "India's Farmers Are Now Getting Their News From AI Anchors," *Bloomberg News*, June 18, 2024. https://www.bloomberg.com/news/newsletters/2024-06-18/india-s-farmers-are-now-getting-their-news-from-ai-anchors.

77 Guthrie Scrimgeour, "The AI Reporter That Took My Old Job Just Got Fired," *Wired*, November 21, 2024. https://www.wired.com/story/the-ai-reporter-who-took-my-old-job-just-got-fired/.

78 Phil Rosen, "ChatGPT Is about to Write BuzzFeed Content — and Investors Are Loving It," *Business Insider*, January 27, 2023. https://www.businessinsider.com/chatgpt-stock-buzzfeed-layoffs-jobs-investors-economy-recession-markets-meta-2023-1.

79 "Business Insider Cuts 21 % of Workforce, Memo Shows," *Reuters*, May 29, 2025. https://www.reuters.com/technology/business-insider-cuts-21-workforce-memo-shows-2025-05-29/.

80 Matt O'Brien, "ChatGPT-maker OpenAI Signs Deal with AP to License News Stories," *Associated Press*, July 13, 2023, https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

81 *The Atlantic announces product and content partnership with OpenAI*, press release, *The Atlantic*, May 29, 2024. https://www.theatlantic.com/press-releases/archive/2024/05/atlantic-product-content-partnership-openai/678529/.

82 *Axios, OpenAI Partner to Expand Local News Coverage*, press release, Axios, January 15, 2025. https://www.axios.com/press-past-releases/axios-openai-partner-to-expand-local-news-coverage.

Amazon with *The New York Times*;[83] Meta with *Reuters*[84]). In response to these actions, journalism unions around the world are seeking job protections related to AI use.[85]

4.3.2 International Stakeholder Interview Analysis

Our 14 journalism interviewees spanned 11 countries and worked at think tanks, academia, and digital, print, and broadcast-first newsrooms. They reported using several generative AI tools, including Claude, Llama, ChatGPT, and WordPress Chatbot. Interviewees noted that while LLMs bring some new capabilities, they generally build on the functionality of existing newsroom technologies rather than introduce a major shift in how journalism is practiced. LLM-based tools, which have been used for tasks like content tagging, headline generation, search engine optimization, and database querying, add to existing technologies that provide support with event monitoring, translation, data visualization, and audience engagement. As one expert from Sub-Saharan Africa commented, journalists feel pressure to learn new skills to survive in the profession:

> *Journalists have been taught to adapt since [the field is] threatened by new technologies. Like with social media—we have become the leaders in social media content creation. We don't want to miss out and be made obsolete.*

Interviewees tended to attribute the tech-forward culture within journalism to two pressures. First, the advent of digital media significantly disrupted the traditional, long-standing systems that newsrooms have used to raise revenue and maintain their operations. One expert from Oceania anticipated that such disruptions might be further exacerbated by LLMs:

> *Around half the newsrooms in Australia are half the size they were a decade ago. The sustainability of journalism is a debate across the world because traditional commercial advertising outlets have disappeared. News publishers and digital platforms have a "frenemy" relationship today—newsrooms are reliant [on platforms] to reach their audience but also vulnerable to their policies.*

Second, there is intense competition to be the first to "break" a story, and journalists, especially those working in news and related beats, must meet very tight deadlines to do so. LLM tooling makes it easier to meet and even tighten the timeline of producing an article,

83 *The Times and Amazon Announce an A.I. Licensing Deal*, *The New York Times*, May 29, 2025. https://www.nytimes.com/2025/05/29/business/media/new-york-times-amazon-ai-licensing.html.

84 "Meta Platforms to Use Reuters News Content in AI Chatbot," *Reuters*, October 25, 2024. https://www.reuters.com/technology/artificial-intelligence/meta-platforms-use-reuters-news-content-ai-chatbot-2024-10-25/.

85 Gretel Kahn, "'Like Nailing Jell-O to a Wall': Why Unions Are Struggling to Protect Journalists' Rights in the Age of AI," *Reuters Institute for the Study of Journalism*, April 23, 2026. http://reutersinstitute.politics.ox.ac.uk/news/nailing-jell-o-wall-why-unions-are-struggling-protect-journalists-rights-age-ai.

sometimes shaving minutes or more off various tasks on both the editorial and business sides of news production. As one practitioner from Europe noted:

> *For a lot of journalists, LLMs are a huge opportunity for bridging the digital gap when English is not their first language. Investigative journalists also need to read lots of reports…. News reporters need to quickly write [articles] in English…. Across journalists, we see significant productivity gains and a relaxation of financial and time pressures.*

This improvement in efficiency and productivity can allow journalists to focus their time on more complex analytical tasks and those requiring human intervention. Many newsrooms have started to provide training to their employees to help reap these benefits, with an emphasis on how to use AI responsibly and avoid losing the "human" component of journalism. A practitioner from the MENA region commented:

> *There is a fear that AI is going to take away the irreplaceable human element from journalism. So part of the training is about what cannot be replaced by AI: the moral stance of the journalist [and] real human interviews.*

Despite the benefits that LLMs can bring, some journalists we spoke with are concerned about the risks LLMs might pose to their profession. While not all newsrooms allow journalists to use LLMs to help them write, interviewees whose employers do allow it generally describe using LLMs to write formulaic articles, such as those about sporting events, weather, and market conditions, rather than more narrative-driven features.

Given journalists' professional mandate to find facts and synthesize them into articles, journalists worry that reliance on biased LLMs may increase or introduce new biases in reporting, influencing how issues are represented. One interviewee expressed concern for the welfare of communities with low internet access since LLMs are trained on fully digitized data and are therefore not exposed to perspectives and cultures that are not "online."

The public's access to information may also be at risk if journalists use LLMs for professional tasks. For instance, due to the possibility they will hallucinate, LLMs may produce misinformation when generating outputs. One interviewee described how this resulted in the rampant spread of political misinformation in India, threatening election integrity. Several interviewees emphasized that journalists have a dual responsibility: not only to convey accurate information, but also to actively contain the spread of falsehoods. As one expert from Asia explained,

> *Today, we have access to more information than ever before for the first time in history. But we don't have the capacity to process all of it, or to distinguish what is truthful and what is not. LLMs are trained on diverse sources of information, and that makes them easy to manipulate.*

When people feel they cannot trust the news, they may turn to other, often less reliable, sources for information.

4.4 Education

4.4.1 Background

Education systems around the world vary significantly, reflecting the cultural, economic, and societal contexts of each region. Different approaches significantly shape learners' experience and filter societal values. When evaluating human rights in education, it is important to consider both the right to education[86] and the other human rights of students and educators within an educational environment. These include the rights to non-discrimination, dignity, self-empowerment, access to information, and labor rights, among others. Teacher training programs often focus on the needs of vulnerable populations and diversity training,[87] and many schools strive to provide equitable access for all students, including minorities, refugees, and disabled students. Despite these efforts, children often do not experience parity in education. They may be denied education as a result of disciplinary action;[88] female and gender-nonconforming students, racially or ethnically minoritized students,[89] and students with disabilities can face systemic inequities;[90] and students not fluent in the language of instruction may not be able to access the content.[91]

---

86 The right to education is a cornerstone of international human rights law, recognized as essential for the full development of the human personality and the promotion of understanding, tolerance, and peace. Article 13 of the International Covenant on Economic, Social and Cultural Rights (ICESCR) affirms that education shall be directed to the full development of the human personality and dignity and shall enable all persons to participate effectively in society. It obliges states to ensure free and compulsory primary education, accessible secondary education, and progressively free higher education. This right is further elaborated in various international instruments, including the Universal Declaration of Human Rights. See United Nations General Assembly. 1948. *Universal Declaration of Human Rights*, Article 26. United Nations General Assembly. 1966. *International Covenant on Economic, Social and Cultural Rights*, Article 13. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights.

87 Todd Jennings, "Addressing Diversity in US Teacher Preparation Programs: A Survey of Elementary and Secondary Programs' Priorities and Challenges from across the United States of America," *Teaching and Teacher Education* 23, no. 8 (November 2007): 1258–1271. https://www.sciencedirect.com/science/article/abs/pii/S0742051X06000965.

88 Kate Taylor and Elizabeth A. Harris, "When Students With Disabilities Are Removed From Classrooms as 'Informal Banishments,'" *The New York Times*, February 9, 2023, https://www.nytimes.com/2023/02/09/us/students-disabilities-informal-removal.html.

89 Eesha Pendharkar, "Students of Color Disproportionately Suffer From Police Assaults at School, Says Report," *Education Week*, January 13, 2023. https://www.edweek.org/leadership/students-of-color-disproportionately-suffer-from-police-assaults-at-school-says-report/2023/01.

90 Sula Kim, "Lawsuit Claims Children with Autism Were Abused at St. Tammany School," *WDSU*, April 12, 2023, https://www.wdsu.com/article/lawsuit-claims-children-with-autism-were-abused-at-st-tammany-school/43571412.

91 Grace Beahm Alford, "English-Language Learners, Immigrant Families Struggle in Savannah-Chatham Schools," *Savannah Morning News*. May 8, 2023, https://www.savannahnow.com/story/news/2023/05/08/english-language-learners-immigrant-families-savannah-chatham-schools/70184982007/.

Whether or not educators have access to emerging technologies can shape their approach to teaching. Some schools and countries may face challenges with limited access to and training in new technologies. UNESCO's AI Competency Framework for Teachers seeks to close this gap by laying the groundwork for training programs.[92] New technologies have often positively impacted students with disabilities and non-heritage language speakers, while online learning and massive open online courses (MOOCs) have broadened access to higher education. Within this context, LLMs have both shown potential for application in teaching and intensified a variety of concerns in the pedagogical process, such as depersonalized education,[93] assessment fairness,[94] exacerbation of inequalities between students,[95] homogenization of knowledge,[96] and more broadly, the delegation of pedagogical responsibilities to machines. Generative AI tutoring risks degrading learning quality through weak or inaccurate instructional support and undermining student-teacher relationships through overreliance, despite its potential to provide individualized and immediate support to students.[97] Other risks of using LLMs in education include misinformation, bias,[98] surveillance,[99] data breaches,[100] and overreliance,[101] all of which are worsened when such tools are used with children.

---

92 Miao, Fengchun, and Mutlu Cukurova. *AI Competency Framework for Teachers*. Paris: UNESCO, 2024. https://www.unesco.org/en/articles/ai-competency-framework-teachers.

93 Kashmir Hill, "Professors Grapple With ChatGPT's Rise on Campus," *The New York Times*, May 14, 2025, https://www.nytimes.com/2025/05/14/technology/chatgpt-college-professors.html.

94 Johanna Fleckenstein, Jennifer Meyer, Thorben Jansen, and Stefan Daniel Keller, "Do Teachers Spot AI? Evaluating the Detectability of AI-Generated Texts among Student Essays," *Computers and Education: Artificial Intelligence* 6, no. 5 (June 2024): 100209. https://doi.org/10.1016/j.caeai.2024.100209.

95 Abd-alrazaq, Alaa et al. "Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions." JMIR Medical Education 9 (2023): e48291. Available: http://dx.doi.org/10.2196/48291.

96 Wright, Dustin, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. "Epistemic Diversity and Knowledge Collapse in Large Language Models." *arXiv*, October 5 2025. https://doi.org/10.48550/arXiv.2510.04226.

97 Burns, Mary. "What the Research Shows about Generative AI in Tutoring." Brookings Institution, January 27, 2026. https://www.brookings.edu/articles/what-the-research-shows-about-generative-ai-in-tutoring/.

98 Milano, Silvia, Joshua A. McGrane, and Sabina Leonelli. "Large language models challenge the future of higher education." Nature Machine Intelligence 5.4 (2023): 333–334. Available: http://dx.doi.org/10.1038/s42256-023-00644-2.

99 Sharon Lurye and Claire Bryan, "Schools Use AI to Monitor Kids, Hoping to Prevent Violence. Our Investigation Found Security Risks," *The Associated Press*. March 12, 2025, https://apnews.com/article/ai-school-chromebook-gaggle-goguardian-securly-25a3946727397951fd42324139aaf70f.

100 Kasneci, Enkelejda, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." EdArXiv. January 30. https://doi.org/10.35542/osf.io/5er8f.

101 Kos'myna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. *Your Brain on ChatGPT: Accumulation of*

4.4.2 International Stakeholder Interview Analysis

The 12 educators we interviewed were dispersed in consulting, edtech, research, and teaching, and worked across primary, secondary, and higher education in eight countries. They reported using a wide variety of tools, including ChatGPT, Gemini, Llama, and Claude. Some interviewees described using LLMs to design curricula and lessons, translate text, assess student work product, personalize tutoring, and grade assignments. However, one practitioner from sub-Saharan Africa described significant variation among educators in their comfort and openness to using LLM-based tools:

> *Some of my colleagues are excited by AI and risk-takers; others are in-between because they see some benefits but do not want to break the status quo.… Others are avoidant because they find it irrelevant to their work and are averse to the risks that may come with switching.*

Some educators reported that educational guidelines determined the extent to which LLMs should be used in their work. Once some schools implemented a plan and training for integrating LLMs into the classroom, educators picked it up quickly. However, some educators reported feeling coerced into using them.

Our interviews showed that one of the key benefits of using LLMs in education is personalized learning. Given the variations in learning styles of students and the limited capacity of a teacher to provide one-on-one instruction, a practitioner from the MENA region described how LLMs can help fill in the gaps:

> *LLMs can personalize learning with videos and tutoring that meet a child where they are at. This can also affect adaptive assessment, [which] changes difficulty and approach based on how a child has answered information in the past.… LLMs also help student feedback be given as soon as possible—no need to wait for teachers or peers to review work.*

In addition to helping students individually, LLMs have been used to support students who face additional challenges in the classroom. These include language learners, who may need some material to be in their first language to assist with comprehension, and disabled students, who may benefit from content simplification and alternative media to support learning disabilities. The practitioner from the MENA region also discussed language modification capabilities:

> *Some lessons are taught in Arabic but most are in English. Our materials all start in English, but students can translate them into Arabic. The AI tutor can also simplify explanations and… read text aloud.*

However, our interviews showed that the benefits of LLMs are not evenly distributed. As an expert from Latin America pointed out, access to these tools varies widely: some public

Cognitive Debt when Using an AI Assistant for Essay Writing Task. MIT Media Lab, June 10, 2025. Accessed August 10, 2026. https://www.media.mit.edu/publications/your-brain-on-chatgpt/.

primary schools may have only a single computer, or none at all, limiting student exposure to LLMs entirely:

> *There is already a digital divide [between] people who have access to the internet and those who do not. If this is not solved, LLMs are doomed to deepen differences in productivity, education, and the labor market. If there is access to the internet but no education on LLMs, the split will be even more important…. It will be very difficult to maintain equal access to education, jobs, and justice.*

Furthermore, there are concerns about the ramifications of using LLMs to support the education of under-resourced communities. Our interviews suggest that such communities may provide students with LLM tooling before there is a clear understanding of their impacts because LLMs can help serve their student population at a lower cost. Our interviewees also point out that material communicated through LLMs tends to be less diverse, exposing students to a narrower range of language, perspectives, and content. This risks narrowing how students express themselves and what they learn, reducing opportunities for creativity, critical thinking, and cultural relevance. An expert from North America describes the risks:

> *Tech is…not better than a person, but it's better than nothing….In academia, there's a collapsing of linguistic and subject diversity and a persistent degree of homogenization.*

Some interviewees called for educational institutions to more deeply assess the actual impact of LLM tooling on students before adopting it more widely. They pointed to studies showing poor learning recall and mixed results on exam improvements, noting that any benefits observed were often small and difficult to attribute directly to the use of LLMs. One practitioner from North America also raised concerns about more significant long-term effects:

> *If you're younger right now, learning a skill with a chatbot, it's not clear you're really learning it…. Basic learning tasks are filled with cognitive load, which is reduced through AI reliance…. My colleagues and I are concerned that ChatGPT could undermine critical thinking.*

Interviewees mentioned a variety of different risks to students, including a reduction of critical thinking, a lack of awareness about the retention of private data, and exposure to biased content. These correspond to risks to the rights to privacy and freedom of thought. Some educators emphasized that human rights risks are exacerbated for minors, whose brains are not fully mature and who are more susceptible to the harms that LLMs pose. They cited critical AI literacy as a skill that can be a first step towards protecting students if it is taught in schools alongside exposure to these technologies.

4.5 Cross-Cutting Themes from International Stakeholder Interviews

4.5.1 Geographic Differentiation

Our interviews revealed that professionals use LLMs differently in their work across geographic contexts, reflecting differing perceptions of LLMs, uneven levels of awareness, and distinct sources of pressure to use these technologies.

Varying Perceptions of LLMs. In our interviews, it appeared that professionals in the Global South[102] tended to be more enthusiastic about the potential for LLMs to improve their work, noting "a clear appetite for the technology" despite some skepticism over IP and data protection. Other interviewees noted LLMs' potential to level the playing field across countries with different levels of power, economic strength, and influence at a relatively limited cost. By contrast, interviewees using LLMs in Global North countries tended to focus more on the risks associated with their use of LLMs. A journalism expert from Oceania remarked "extreme caution about using LLMs in [journalistic] reports" and the willingness of Australian journalists to "keep their product clean [from AI] as an effective marketing strategy."

Differentiated Awareness. Overall, our interviews indicate a potential gap in professionals' awareness of the risks of LLMs between users in the Global North and in the Global South. One of our interviewees, a journalism expert based in Sub-Saharan Africa, highlighted the lack of training for journalists interested in applying LLMs to their work. "Journalists understand what ChatGPT means but they don't know what LLMs are," he noted. By contrast, a legal practitioner from Asia highlighted "wide awareness of the existence of LLMs and generally how they may have an impact" among his peers.

Sources of Pressure to Use LLMs. Some interviewees remarked that clients and partnering organizations exerted pressure on professionals to use LLMs for efficiency and cost reduction purposes. For example, a legal expert in North America noted, "Law firms are very exposed from the purchaser perspective and they face client pressure to use high-tech methods like LLMs." Other interviewees described the pressure to use LLMs as coming mostly from within the workplace. An education practitioner in Latin America noted: "Educators are encouraged to use LLMs in their work but they lack the training to do so. They need to take the initiative to learn about it and play with it themselves."

---

102 We acknowledge that this term has complex, varied, and sometimes homogenizing uses, but we apply it anyway to recharacterize prevailing notions about who supports perspectives deemed as critical. See here for more of its limitations: Radiya-Dixit, Evani, and Angèle Christin. 2025. "Same Stereotypes, Different Term? Understanding the 'Global South' in AI Ethics". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8 (3):2081-93. https://doi.org/10.1609/aies.v8i3.36697.

4.5.2 Sources of Variation in LLM Adoption

LLMs can affect the human rights of individuals by posing risks and creating opportunities specific to professions and geographies. The difference in peoples' exposure to the risks and opportunities of LLMs can be further exacerbated by the digital divide, lack of resources, and language accessibility.

The Digital Divide. Several interviewees highlighted the digital divide as a key factor inhibiting the widespread adoption of LLMs and understanding of its risks. Access to digital devices and the internet, which are precursors to LLM awareness and access, varies based on a country's wealth and level of urbanization. For example, a journalism expert in Sub-Saharan Africa described how access to technology impacts users' ability to use LLMs:

> *Most people only have smartphones and they only use WhatsApp [to access online information] because it's cheaper [than a computer]…. Using [an LLM chatbot] through WhatsApp gives them access to information… but it's a trickle down effect, people in remote areas have few relatives who are professionals, so they get content that is cascaded through WhatsApp… the awareness is not there about how information and data is being used, and there are no mechanisms to test accuracy.*

Due to lack of direct and informed exposure, users may be less aware of issues like hallucinations and data privacy vulnerabilities. As a result, they may be especially vulnerable to the risks LLM-based tools pose.

Fewer Resources. The use of LLMs in professional settings varies based on the resources that organizations, such as law firms, schools, and newsrooms, can dedicate to understanding and using LLMs in their work and field. For instance, a journalist at a multinational newsroom based in North America describes how their organization is "pretty far ahead" of other newsrooms, investing heavily in exploratory teams that build proprietary software for specific journalistic purposes. They were also "required to go through internal training" before applying LLMs to relevant tasks, and their newsroom created early internal guidelines and tools for using LLMs in journalistic work.

Geographic Restrictions. Both geopolitical and linguistic factors affect access to and the utility of using LLMs in professional contexts. LLM developers and deployers have restricted LLM access to people in certain countries due to political and regulatory hurdles. For example, OpenAI does not offer ChatGPT in certain jurisdictions, such as China, Russia, and previously Zimbabwe. However, some individuals in these regions have circumvented these bans using a Virtual Private Network (VPN). Residents of these countries have also created their own LLMs, such as the popular platform DeepSeek in China.

Poor Language Understanding. While LLMs can help facilitate multilingual communication, disparities in the language capabilities of LLMs exist. One legal practitioner in the MENA region described his experience:

I use it for limited purposes but very often, like assistance in translation and other language support… English is better… also very accurate French and less in Spanish… [LLMs] are used in arbitration, which is mostly in English…. The [Modern Standard Arabic] it produces is a bit distinct, [and it is] not good for other dialects.

Moreover, LLMs often struggle with languages and dialects that are underrepresented in their training data, leading to disparities in performance. Both DeepSeek and ChatGPT, for example, predominantly use Mandarin from Mainland China which is likely due to their training data in the language.[103] This bias may reinforce the dominance of Mandarin as used in that context while underrepresenting linguistic variations from Taiwan, Hong Kong, and other Chinese-speaking regions. Additionally, the limited inclusion of Cantonese and other dialects raises concerns about whose voices are amplified and whose are excluded.

4.5.3 Cultural and Ideological Homogenization

While LLMs offer vast knowledge and linguistic capabilities, their outputs often reflect ideologies that do not always account for the diverse identities of their users. The data used to train these models is not neutral; it is shaped by the dominant social, political, and linguistic frameworks of the United States, reinforcing the perspectives of those who create and control them. For instance, OpenAI reports that 93% of ChatGPT-3's training data was written in English, leading to concerns that LLM-generated content is filtered through a majoritarian lens.[104] Some interviewees described this as a form of ideological homogenization. An expert from Sub-Saharan Africa shared:

Some information and languages are left out [of datasets] during [LLM] training. LLMs reproduce dominant narratives, [have elite ways of looking at society... [and focus] on promoting people who are nondisabled so the views of people with disabilities are not represented. The training datasets are also Eurocentric so what LLMs write about certain [regions] doesn't make sense.

Some legal professionals also expressed skepticism about LLMs' ability to accurately reflect local legal systems. For example, a practitioner from Latin America described her distrust of LLMs:

I don't use LLMs for text generation and drafting. I just don't trust [generated] information regarding local laws and jurisprudence.

This dynamic reinforces existing power imbalances, as the overrepresentation of dominant ideologies in LLM outputs can contribute to minimizing non-dominant perspectives.

---

103 Siu, Sai Cheong. 2023. "Revolutionizing Translation with AI: Unravelling Neural Machine Translation and Generative Pre-Trained Large Language Models." SSRN Electronic Journal. http://dx.doi.org/10.2139/ssrn.4499768.

104 OpenAI. 2023. "Explosion in the Number of Tokens / Words Generated." OpenAI Developer Community. August 7, 2023. https://community.openai.com/t/explosion-in-the-number-of-tokens-words-generated/317702.

To mitigate such risks, some interviewees advocated for localized adaptation of LLMs that take into account cultural and linguistic diversity. A journalism expert from Sub-Saharan Africa noted:

> *If LLMs were improved to include local languages in their dataset, people from remote areas who cannot access services such as health could benefit immensely.… There should also be more partnerships with local universities, it can help tweak LLMs and make them more accurate.*

However, some communities, particularly some Indigenous groups, resist the inclusion of their languages and cultural knowledge in AI training datasets, citing concerns over exploitation and data sovereignty. When asked about the human rights risks of using LLMs, a journalist from Oceania answered:

> *Some concerns include data usage without permission, particularly for first nation people and cultures. There is also a problem of data colonialism and Indigenous data sovereignty.…*

These tensions highlight a fundamental challenge in LLM development: ensuring inclusivity without reinforcing extractive practices that perpetuate historical inequities.

4.5.4 Process and Purpose Disruption

While LLMs can make a number of tasks easier and more efficient across professions, the process of completing a task without an LLM often fulfills a social role or cognitive function. For example, when writing an investigative article, a journalist must exercise judgement on what warrants a story, conduct research and interview sources to shed light on what is happening, and craft an article that covers the relevant facts and ideas. While the ultimate written output may be an article, one expert in Europe explained that this is just one part of the field of journalism's larger mandate:

> *Journalists and others need to see themselves as an informational health service.… [Many have] faced persecution and legal, political, or financial constraints and see [LLMs] as an opportunity to help them bear witness and use their morality as humans.*

Given that LLMs are increasingly adopted in newsrooms, some interviewees identified a risk that too much emphasis will be placed on creating the article itself and not enough on the process of bearing witness and maintaining a healthy information ecosystem. This "disruption" may erode long-standing accountability mechanisms for governments and other actors, potentially resulting in corruption, discrimination, and other abuses of power.

This is particularly salient in the context of education, where the broader goal is to equip students with the skills to contribute meaningfully to society. Several educators expressed concerns about academic integrity when students have access to LLMs to complete assignments: while they might be able to write a paper with an LLM, they would not go through the same process of using critical and structured thinking to develop an argument.

These examples illustrate how LLMs can reduce the purpose of a task to its tangible end product. Automating outputs, when done without care, may displace the human practices that give end products meaning and threaten the rights that those practices protect.

4.5.5 Capitalist Incentives

As LLMs have grown in popularity, industry professionals feel pressure to adopt them in their day-to-day practice. Clients of law firms request that attorneys use LLMs to reduce billable hours, while schools and newsrooms view LLMs as a way to sustainably operate with limited funds. Interviewees in all three fields acknowledged increases in productivity as a result of LLMs. Professional interest in these tools has created opportunities for LLM developers and deployers to strategically ally with industry organizations. As an Oceania-based journalism expert described:

> *Looming on the horizon is… when individual [news] organizations sign deals with certain AI companies, they exclude other tools…. Strategic alliances are going to concentrate the monopolies more and there will be a scramble of big companies to stay at the top.*

Such dependence on AI companies has concerning implications. Some interviewees describe how rapid advancements in technology have previously negatively impacted the journalism industry, making it vulnerable to platforms that profit from journalistic content without fairly compensating the newsrooms that produce it.[105] As AI-generated news summaries grow in popularity, newsrooms must once again reinvent their revenue streams.

As described by an expert in education in North America, lower income communities are particularly likely to experiment with emerging technologies like LLMs earlier on:

> *It's been very celebrated, but tech is a stop gap in communities lacking resources…. Since [lower income school districts] need affordable options, they're where [educational technology] gets tested first, and [students] are more vulnerable to adverse effects earlier.*

Furthermore, such tooling gives technology companies the power to make implicit decisions about what education means and what is important to learn. To mitigate these vulnerabilities resulting from capitalist incentives while enabling the benefits that LLMs can provide, an education practitioner in North America called for LLM-based educational technology to be removed from the purview of technology companies entirely:

> *I would like a university consortium of Google doc-esque apps with minimal functionality and no generation…. Learning management systems should be non-surveillant without learning analytics. Maybe it should be developed by a textbook publisher.*

In general, interviewees further called for greater competition among LLM developers and deployers to incentivize the prioritization of ethical principles as a selling point for new tools.

---

105 See NYT v. OpenAI.

4.5.6 Changes in Labor Practices

Our research reveals that while LLM-driven job losses in journalism, education, and law are limited in scope, the impact of LLMs on labor markets is significant. Employers are sometimes disincentivized from hiring new workers and may opt not to replace retiring employees, creating new barriers to entry for job seekers. Similar to past technological advancements, LLMs reduce the number of workers needed to perform certain tasks, thereby decreasing the number of positions available on the job market. A GenAI expert from Latin America summarized these concerns:

> *LLMs are seen as a threat for employment because they have the capability to produce news without a journalist. Newsrooms only need one person to check [AI-generated] content.*

More commonly, LLMs are changing the way work is performed by augmenting human labor rather than replacing it entirely. Many workers now integrate LLMs into their workflow to enhance efficiency, allowing them to shift focus from simple and repetitive tasks to higher-level tasks. A journalism professional in Oceania described the shift he observed in local newsrooms:

> *LLMs are used mostly for back-end efficiency gains, that's where the news organizations see the most value. Many of them have stated that they did not intend to replace journalists but to make them more efficient.*

As a result, the skillsets required in various professions may evolve, with greater emphasis on critical reasoning and complex problem-solving rather than technical execution. This shift in labor dynamics carries mixed implications for workers' well-being. Some interviewees reported positive effects, including reduced workloads and the ability to dedicate more time to meaningful aspects of their jobs, such as developing more effective connections with service beneficiaries. For example, an educator in Oceania described how LLMs free up her time to create deeper connections with students, enhancing their learning experience:

> *LLMs cut down on preparation time for teachers. Before LLMs, it took me 9 minutes to prepare a lesson's content and worksheets, but now, it only takes me 3 minutes. This gives me more time to build relationships with students and help them one-on-one.*

Additionally, the efficiency gains from LLMs can improve work-life balance, allowing employees to complete tasks more quickly and reclaim personal time. However, some interviewees expressed concerns about rising productivity expectations, with some workers feeling pressured to produce more in shorter time frames. Another consequence is the reduced availability of entry-level training opportunities. Since LLMs can efficiently handle repetitive tasks, junior employees—who traditionally learn by completing such assignments—may struggle to develop foundational skills, potentially weakening long-term professional development. A lawyer and educator based in Europe described this risk in the legal field:

> *Law firms are hierarchical pyramids. Because work efficiency goes up, the number of entry-level positions may decrease because AI can assist people.*

4.5.7 Human-LLM Interaction

Risk Exacerbation. Like other emerging technologies, LLMs can foster both overreliance and misuse among users. Our interviews indicate that the widespread enthusiasm surrounding AI has contributed to an "AI hype," encouraging workers to delegate tasks to LLMs beyond their intended capabilities. This overreliance can have significant consequences, particularly when LLMs are used for high-stakes decision-making, leading to potential human rights implications. As highlighted by a North American educator, at the organizational level, this enthusiasm can manifest as technosolutionism—the belief that LLMs can resolve a wide range of challenges, including those they were not designed to address:

> *Tech solutionists want to use LLMs as tutors, but there are a lot of risks in doing that. Synthetic media is not necessarily an appropriate teaching tool. This idea is pushed by people who are selling tech or who don't understand it.*

Risk Differentiation through Anthropomorphisation. Unlike past technological advancements, LLMs introduce unique risks due to their conversational interface, which can create the illusion of interacting with a human being. This is particularly concerning for users with limited technical understanding, who may mistakenly attribute reasoning capabilities to LLM-generated responses. Additionally, some workers use LLMs as search engines—despite their lack of source verification—leaving them vulnerable to misinformation, hallucinations, and biases that can compromise the quality and accuracy of their work. A lawyer in North America summarized these concerns:

> *Most people don't understand how LLMs work, their training, development, and what they're supposed to do. People are conditioned to believe [LLMs] are like search engines, that they prioritize accuracy and specificity, but actually they are all about next-word prediction. LLMs are optimized for human preferences, they are not aligned to accuracy or factuality.*

Risk Exacerbation through Power. In addition to low technical understanding, a prominent factor exacerbating risk is the level of power that is delegated to LLMs to perform tasks. LLMs are substantially more likely to harm human rights when used by high-impact users, individuals who are in decision-making positions. A law professor in Colombia raised concerns over high-impact usage of LLMs in the legal field:

> *Some judges in Colombia have used ChatGPT to draft their decision.... If LLMs are used in a judiciary process where an individual's rights are at stake, it raises concern because we don't know what judges use to rule. This is the case when judges hide that they use ChatGPT, but also when they disclose [their use] because the tool doesn't provide sources and is not always reliable.*

Through our literature review and international stakeholder engagement, we have identified human rights risks relevant to the use of LLMs in law, journalism, and education. We identify how risks vary across professions, geographies, cultures, and levels of access to resources. The following parts of this paper focus on assessing the severity of these risks and identifying appropriate actions to address them.

5. RISK ASSESSMENT AND PRIORITIZATION

Drawing on the evidence from our literature review and global stakeholder interviews, we assessed the potential human rights risks arising from the use of LLMs by legal professionals, educators, and journalists in their work. To evaluate the salience, or significance, of each risk, we applied four criteria[106] commonly used in human rights due diligence:

- **Scope:** How widespread is the impact on individuals and/or communities?
- **Scale:** How serious is the impact?
- **Remediability:** To what degree is it possible to restore the rightsholder(s) or make amends to those who have suffered a human rights harm?
- **Likelihood:** How likely is the impact?

We categorized each right as scoring low, medium, or high on each of these criteria. We conducted our analysis of risks in two stages. First, we identified papers that audit LLM outputs with respect to how they are used in these three fields and categorized output-level risks using the criteria described above. Second, we drew on our literature review and stakeholder interviews, conducting a more traditional human rights assessment of the use of LLMs in each of the three professions. Although human rights assessments do not typically identify human rights opportunities, we did so in this paper because doing so best represented the learnings from our interviews of people in the Global South.

5.1 Output-Level Rights Analysis

As our international stakeholder interviews surfaced, the human rights impacts of LLMs are most apparent once they are embedded in real-world use. However, by the time these harms become evident in real-world contexts, it can be too late for effective mitigation. Thus, in addition to identification and remedy after harms appear, human rights due

106 United Nations Office of the High Commissioner for Human Rights. 2011. Guiding Principles on Business and Human Rights. https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf. For a similar approach, see BSR 2022 Human Rights Assessment. https://www.bsr.org/reports/BSR-FAQ-on-Human-Rights-Assessments.pdf.

diligence necessitates a preventative approach for identifying risks earlier, at the model level, before deployment and integration into real-world processes. Given expertise and resource constraints that may prevent human rights practitioners from conducting output-level model evaluations explicitly centered on human rights risks,[107] we offer three examples of how to reinterpret the results of existing model evaluations to surface insights on human rights risks, selecting those focused on tasks that may be done by professionals in our fields of interest.[108]

5.1.1 Approach

We adopt a model-centric lens, focusing on how patterns in model outputs themselves may give rise to human rights risks and supplementing our stakeholder interview analysis with a synthesis of quantitative model evaluations. Since LLMs are widely understood to be prone to consistent patterns of error, output-level rights analysis requires identifying and characterizing recurring responses that may pose risks. When situated in concrete applications, these patterns allow us to anticipate and analyze potential human rights harms along four axes of risk saliency used in human rights assessments:

- **Scope:** How widespread is the impact on individuals and/or communities?
- **Scale:** How serious is the impact?
- **Likelihood:** How likely is the impact?

Remediability, which involves action taken to mitigate harm, involves the influence of individuals and organizations. While model design decisions such as output explainability and transparency can make it easier to understand and remediate issues, remediability cannot be measured at the output level. While scope, scale, and remediability are inherently prospective and reflect potential future risks, likelihood is retrospective and analyzes how frequently the risk appeared in the past.

In the output-level rights analysis, for each human right identified as being at risk, our research team assigned a rating of "Low," "Medium," or "High" for each of scale, scope, and likelihood. We defined them as follows:

107 See this example: Raman, Vyoma, Camille Chabot, and Betsy Popken. 2025. "Assessing Human Rights Risks in AI: A Framework for Model Evaluation". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8 (3):2107-20. https://doi.org/10.1609/aies.v8i3.36699.

108 This section is intended to illustrate how the methodologies of AI evaluation and human rights assessments might be bridged, not to stand alone as a comprehensive analysis of human rights risks in these fields.

Factor	Low	Medium	High
Scope	A small number of the relevant population is at risk.	Approximately half of the relevant population is at risk.	A significant majority or all of the relevant population is at risk.
Scale	The risk is associated with indirect adverse impacts on the rightsholder(s).	The risk is associated with direct adverse impacts on the rightsholder(s).	The risk is associated with lasting or irreversible adverse impacts on the rightsholder(s).
Likelihood	0-5% chance of the problematic output being produced.	5-20% chance of the problematic output being produced.	20-100% chance of the problematic output being produced.

To develop a fine-grained understanding of these dynamics, we examine three case studies evaluating LLM outputs relevant to law, journalism, and education. These case studies were selected because their scope aligns with ours by assessing a real-world use case of LLMs by professionals in these fields, and because the metrics they report relate to problems we also identify in the findings of our stakeholder interviews and literature, enabling us to extrapolate human rights risks from those scores. Our analysis thus translates output-level evaluation metrics into concrete human rights risks, which makes model assessments actionable in real-world contexts and in preventing harm. By mapping model behaviors to human rights dimensions, we can identify how certain model outputs may manifest as rights harms. However, the scope of this analysis is necessarily limited to the tasks and evaluation metrics selected by the original researchers.

5.1.2 Hallucination in Legal Research

Many interviewees in law mentioned LLMs' current and potential utility for legal content analysis and document drafting. Interviewees noted that while these tools could increase efficiency, the risks are especially pronounced given the high-stakes nature of legal work. In many cases, lawyers have received or been threatened with sanctions after submitting LLM-generated court documents that included hallucinated case citations.[109] Given such concerns, many law firms have opted for more specialized AI systems that are tailored to legal use and subject to internal ethical and privacy constraints. These concerns underscore the importance of evaluating legal LLMs for accuracy and reliability before deployment.

To assess these issues, researchers in the study "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools"[110] conducted a technical evaluation of several proprietary legal LLMs that use retrieval-augmented generation (RAG). Rather than relying solely on information encoded in an LLM during training, RAG systems first retrieve relevant documents from an external database and provide them to the model as

109 Anna Tong, "AI 'Hallucinations' in Court Papers Spell Trouble for Lawyers," *Reuters*, February 18, 2025.

110 Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools." *Journal of Empirical Legal Studies* 22 (2025): 216–42. https://doi.org/10.1111/jels.12413.

source material for generating its response. For example, the LLM systems studied in this paper first retrieve relevant case law or statutes from proprietary legal databases and then use that information to respond to legal queries. The researchers designed a framework to classify responses to legal queries by correctness (is the answer right?) and groundedness (does the response refer to legal precedent?). Despite using RAG, at least 17% of model responses were both incorrect and ungrounded, and at least 8% were one or the other. No system provided correct and grounded responses for more than 65% of queries. These results varied across the four types of questions: general research, fact-finding, correcting false premises, and jurisdiction/time-specific questions.

Even when used by trained legal professionals, incorrect or ungrounded information can threaten the right to due process and access to a fair trial.[111] While these tools are often framed as time-saving aids, systematic flaws in their outputs can mislead practitioners and erode the quality of legal reasoning. Yet there is no uniform standard for how legal professionals should review and verify AI-generated content, leaving the quality and rigor of human oversight to vary across practitioners and institutions. This unevenness in oversight can amplify disparities in legal outcomes, particularly where time, resources, or expertise to scrutinize AI-generated work are limited. In legal practice, where arguments must be grounded in precedent and factual accuracy, the inclusion of fabricated citations or misapplied legal concepts can materially affect outcomes or delay the course of justice. These risks are especially pronounced when professionals rely on AI-generated drafts with limited time capacity for thorough review.

These findings give insight into the saliency of risk to the rights to due process and fair trials, which are endangered when LLMs hallucinate when providing assistance with answering legal questions. For this and following subsections, our analysis of each of the three key criteria for each salient risk is represented in tables, with the scope, scale, and likelihood of each right color-coded as High, Medium, and Low.

111 The right to due process and fair trial guarantees that every individual is entitled to a fair and public hearing by an independent and impartial tribunal in the determination of their rights, obligations, and any criminal charges. United Nations, *Universal Declaration of Human Rights*, Article 10, 1948. https://www.un.org/en/about-us/universal-declaration-of-human-rights.

Scope: Medium	Scale: Medium	Likelihood: High
Inexperienced lawyers or self-represented litigants may be less likely to identify errors in model responses than experienced legal professionals, making them more susceptible to risk. The scope of risk also varies depending on the type of legal task. Answers to questions with false premises, such as about a nonexistent statute, that fail to discredit the premise may mislead people who have less experience with a legal system, who are more likely to hold flawed assumptions that would then be confirmed by models.	The scale of harm is high in cases where LLM systems provide fully incorrect answers with hallucinated legal citations in formal documents, because such errors can mislead judges or opposing counsel and undermine the credibility of the attorney or firm submitting error-ridden documents. When answers are correct but improperly grounded, the scale is a bit lower because the legal argument may still hold even if it is not well supported.	Magesh et al. find that no system achieved more than 65% accuracy. This means that at least 35% of outputs were ungrounded or incorrect, which represents the likelihood of encountering errors.

5.1.3 Election Reporting Quality

In the journalism sector, many interviewees described an increasing reliance on LLMs to produce content in under-resourced environments, particularly in regions with few local newsrooms or limited access to traditional reporting infrastructure. While expanding coverage can potentially enhance access to information, several interviewees raised concerns about freedom of opinion and the right to participate in public affairs if LLMs are applied to sensitive contexts like elections.

This tradeoff is examined in "A Methodological Proposal to Evaluate Journalism Texts Created for Depopulated Areas Using AI,"[112] where researchers assessed quality control measures for reporting on AI-generated election information in a sparsely populated region of Spain. The study developed a rubric to assess the quality of AI-authored articles across three dimensions: adherence to journalistic norms and style, inclusion of relevant media (e.g. links, audiovisual imagery), and compliance with standards for public information (e.g. accuracy, nondiscriminatory language). Journalism experts were recruited to manually review this criterion. The results of the evaluation showed that while surface-level features, such as grammar and spelling, were largely adequate, the articles often lacked sufficient context, failed to interpret information accurately, and sometimes lacked a cohesive narrative flow. AI-generated content also exhibited deficiencies in using inclusive language and the selection of appropriate images, videos, and other media.

These issues reveal the human rights risks associated with the use of LLMs by newsrooms. Incomplete or poorly contextualized news content can distort public understanding, violating the right to access information and impairing the public's ability to engage meaningfully in democratic processes. When reporting covers elections or public policy without sufficient context or accuracy, it threatens the right to participate in public affairs.

112 Luis Mauricio Calvo Rubio, María José Ufarte Ruiz, and Francisco José Murcia Verdú, "A Methodological Proposal to Evaluate Journalism Texts Created for Depopulated Areas Using AI," *Journal. Media* 5, no. 2 (2024): 671–87. https://doi.org/10.3390/journalmedia5020044.

Moreover, content that omits inclusive language or overlooks marginalized perspectives may contribute to the exclusion of vulnerable groups, infringing on principles of non-discrimination and equality in information access. Even if not deliberately misleading, the systemic shortcomings of these tools in generating high-quality journalism pose real risks in information-poor environments.

The results covered a range of factors that helped us evaluate the saliency of the risk to access to information. This right is at risk when there are errors in LLM-generated reporting, as identified by the study.

Scope: Medium	Scale: High	Likelihood: Medium
LLMs are more likely to be used for reporting in regions that lack alternative sources of reliable news, which are less populated. However, the use of similar tools across multiple local outlets may replicate these harms across communities.	Errors primarily occur in politically sensitive or time-critical reporting, such as the stated context of election coverage, where misinformation or misinterpretation can directly affect public decision-making. This is particularly a problem since alternate sources of information are unavailable in the region of application. The scale is higher for content-related errors and is lower for stylistic or surface-level errors that do not compromise core information.	In 10–20% of articles, reviewers noted issues with context, interpretation, and spelling or grammar, which affected coherence and cohesiveness.

5.1.4 Adherence to Educational Standards

Several interviewees in the education sector reported the increasing application of LLMs to curriculum development for classes and personalized learning for individual students. Both of these involve determining what is important for students to learn and incorporating it into individual or group instruction. Educational "standards" are the skills and concepts in a curriculum. Alignment between LLM tutoring and educational standards can provide insight into the extent to which LLMs are able to support (or harm) the right to education as a whole.

"MathFish: Evaluating Language Model Math Reasoning via Grounding in Educational Curricula"[113] evaluates LLMs' ability to determine how well math problems follow K-12 mathematics skills and concepts ("standards"). After presenting two datasets that label problems with learning goals identified by pedagogical experts, such as comparing fraction

113 Li Lucy, Tal August, Rose E. Wang, Luca Soldaini, Courtney Allison, and Kyle Lo, "MathFish: Evaluating Language Model Math Reasoning via Grounding in Educational Curricula," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Miami, FL: Association for Computational Linguistics, 2024), 5644–5673. https://doi.org/10.18653/v1/2024.findings-emnlp.323.

sizes or multiplying multi-digit numbers, researchers evaluated the ability of LLMs to connect math problems with relevant learning goals. The researchers did so by testing LLM performance on two tasks: First, LLMs were prompted to verify the level of alignment between an LLM-generated math problem and a given educational standard. Second, LLMs were prompted to tag a given math problem with the relevant standards. For the first task, the researchers found that GPT-4 overestimates the alignment of math problems to standards by over 40% compared to curriculum specialist teachers. For the second task, they found that GPT-4 perfectly labeled math problems with standards, problem topic, and other attributes only 4.8% of the time, while there was imperfect overlap between predicted and actual attributes 50.2% of the time.

These findings raise concerns that the right to education may be at risk if educators use GPT-4 to provide additional support because students may receive content that does not meet their needs. Educational standards serve as expert-designed reference points that define essential learning goals and ensure consistency and quality across instructional materials. When LLM-generated content does not align with these standards, students may be exposed to math problems that are irrelevant, misleading, or poorly sequenced, undermining their ability to acquire the knowledge and skills to which they are entitled. While most immediate is the risk to core educational outcomes, there are also secondary concerns around instructional coherence: misaligned content may disrupt lesson planning and reduce the pedagogical value of classroom or personalized instruction. In the longer term, if such tools are used disproportionately in under-resourced or unsupervised settings, they may also contribute to inequality by providing a lower-quality education to students with fewer support mechanisms, introducing downstream risks of discrimination and unequal outcomes.

We assess the risk of using LLMs to the right to education, which is at risk if an LLM's understanding of pedagogical goals does not reflect the actual curriculum.

Scope: Medium	**Scale: High**	**Likelihood: High**
Any student who uses LLMs to study math can experience this risk. The finding of curriculum misalignment, which threatens the right to education, is specific to GPT-4, but it is unclear how it translates to other LLMs. Even if there are more specialized LLMs that can do better than GPT-4, students may receive more exposure to models in the GPT family than more specialized models due to the cost and availability of ChatGPT.	The scale of harm is high in verification tasks, where models significantly overestimate a problem's alignment with standards. This can result in students spending substantial time learning content that does not support their intended pedagogical goals, detracting from their overall growth. Errors in tagging tasks pose a medium risk, since partial overlaps between predicted and actual educational standards still offer some understanding of what a model adds. The scale is particularly high for students in personalized learning contexts where teacher oversight may be limited.	The likelihood of encountering erroneous outputs is very high in verification tasks (40% overestimation) and high in tagging tasks (with exact match accuracy of only 4.8%).

5.1.5 Toward Rights-Aligned Evaluation

While the findings in this section are informative, they represent only a subset of possible risks. Despite growing awareness of human rights risks posed by LLMs, most technical evaluations remain focused on general performance metrics without explicitly linking those metrics to human rights outcomes. As a result, potential harms must often be inferred rather than directly measured. In the studies we reviewed here, we extrapolated risks to the rights to due process, access to information, and education from task-agnostic benchmarks rather than measuring them directly. This indirect approach limits evaluative clarity and comprehensiveness. Critical dimensions such as severity of harm, frequency of use, and the ease of correction are often inadequately captured by standard model benchmarks, currently limiting the ability of companies to identify and prevent the risks to human rights before deployment.

To address this gap, model evaluations can be explicitly designed around human rights impacts, beginning at the level of task and metric selection. In a companion piece[114] focused on the use of LLMs in political news journalism, we propose and demonstrate a rights-aligned evaluation paradigm that centers rights from the outset and links model behavior to real-world consequences through a structured, preventative framework.

This approach involves three key steps:

1. Align evaluation tasks with specific human rights risks.
Rather than relying on general benchmarks, select tasks that reflect real-world activities where specific rights are at stake. For example, legal citation accuracy implicates the right to due process and a fair trial; framing in election coverage relates to the right to access reliable information and participate in public affairs; and the alignment of math content with educational standards affects the right to education. These tasks provide more meaningful insights than abstract performance tests because they allow us to evaluate how model errors may lead to material rights harms to humans.

2. Develop metrics that capture salient forms of risk and their frequency.
Metrics should not merely assess correctness or fluency, but distinguish between degrees and types of harmful behavior—such as outright hallucinations vs. subtle misrepresentations, or stylistic errors versus substantive omissions. In addition to identifying whether risk is present, metrics should also capture how often these issues occur in practice. Without

114 Raman, Vyoma, Camille Chabot, and Betsy Popken. 2025. "Assessing Human Rights Risks in AI: A Framework for Model Evaluation". *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8 (3):2107-20. https://doi.org/10.1609/aies.v8i3.36699.

attention to salience[115] and frequency, evaluations risk underestimating the significance of recurring failure modes in high-impact domains.

3. Interpret results through the four axes of human rights risk assessment.
Once rights-relevant tasks and metrics are established, results should be analyzed using the four established dimensions of salience in human rights due diligence.[116] This step ensures that evaluation results are actionable and aligned with the broader accountability frameworks in Business and Human Rights practice. This analysis allows developers, regulators, and impacted communities to understand which risks may matter most, how often they might occur, and what steps might be necessary to mitigate them.

While this three-step approach does not cover the full spectrum of model risks, it offers a replicable structure for conducting evaluations that are directly tied to human rights impacts. Broader adoption will require expanded task coverage, cross-context evaluations, and collaborative design involving affected stakeholders, domain experts, and model developers. The utility of this framework also depends on institutions that develop LLMs adopting these practices as part of their due diligence responsibilities—not just for mitigating harm after the fact, but for guiding safer model development from the outset.

Embedding an analysis of human rights risk directly into model evaluation design is essential to shift from reactive mitigation to preventative due diligence. Traditional LLM evaluation frameworks often prioritize surface-level performance metrics—such as fluency or benchmark scores—while overlooking the real-world implications of model behavior. By contrast, rights-aligned evaluation foregrounds human rights salience from the outset: selecting tasks with meaningful human impact, designing metrics that capture the severity and probability of potential harm, and interpreting results through established human rights criteria. This approach allows us to anticipate where and how models might systematically generate outputs that lead to the violation of peoples' rights, offering early insight into risks before deployment in sensitive contexts.

Yet model-level evaluations, while necessary, are not sufficient. Their scope is constrained by the particular tasks and assumptions embedded in technical experiments. To capture the broader social and institutional dynamics that shape the impact of LLMs, a second, complementary layer of analysis is required, focusing on the system and society levels.[117] This

115 Salience refers to the significance or prominence of a human rights risk. The most salient risks are those most likely to cause serious negative impacts on individuals or communities, either because of their scope, scale, likelihood, or lack of remediability.

116 Scale: How serious is the impact? Scope: How widespread is the impact on individuals and/or communities? Remediability: To what degree is it possible to restore the situation or make amends to those who have suffered a human rights harm? Likelihood: How likely is the impact?

117 The system level refers to how LLM outputs interact with the institutions, applications, and processes in which they are embedded to create some material effect (e.g., hiring decisions), while the society level refers to their broader impacts on communities and societal norms (e.g. employment patterns).

broader assessment draws on stakeholder interviews and contextual expertise to evaluate the salience of risks and opportunities across key domains where LLMs are already being deployed. While model-level analysis identifies concretely how certain failure modes arise, holistic assessment further reveals which rights are most at stake, who is most vulnerable, and under what conditions harms or benefits are likely to materialize. Taken together, these two steps form a more comprehensive framework for human rights due diligence, one that links model design and evaluation to the lived realities of those most affected by these technologies.

5.2 Use-Level Rights Analysis

Drawing on qualitative data from our stakeholder interviews, our analysis of technical metrics from model evaluations, and our literature review, we holistically assessed the salience[118] of human rights risks and opportunities associated with using LLMs in the fields of law, journalism, and education. Although not typically included in traditional human rights assessments, we included opportunities because we found them uniquely represented in the Global South and believe it important to represent such perspectives. While LLMs may impact the same human right both negatively (risk) and positively (opportunity), an opportunity does not justify a risk. The human rights we identify at risk are defined in the Glossary in Appendix B.

5.2.1 Risks

When evaluating the human rights at risk when journalists, educators, and legal professionals use LLMs in their work, we applied four criteria[119] commonly used in human rights due diligence to evaluate the salience of human rights risks:

- **Scope:** How widespread is the impact on individuals and/or communities?
- **Scale:** How serious is the impact?
- **Remediability:** To what degree is it possible to restore the rightsholder(s) or make amends to those who have suffered a human rights harm?
- **Likelihood:** How likely is the impact?

118 Salience refers to the significance or prominence of a human rights risk. In this context, the most salient risks are those most likely to cause serious negative impacts on individuals or communities, either because of their scope, scale, likelihood, or lack of remediability.

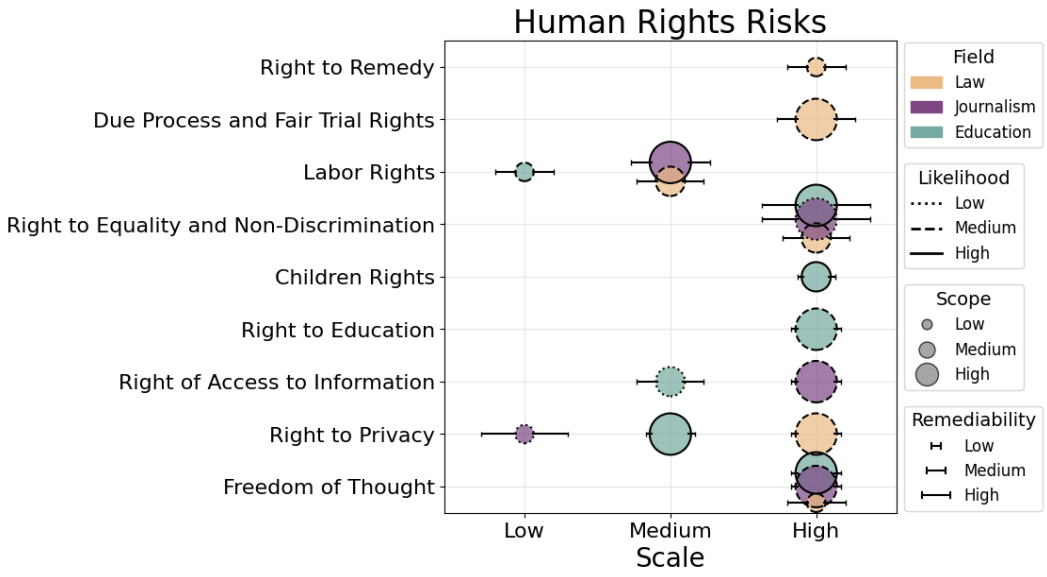
119 United Nations Office of the High Commissioner for Human Rights. 2011. *Guiding Principles on Business and Human Rights*. https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf. For a similar approach, see BSR. 2022. *Human Rights Assessment*. https://www.bsr.org/reports/BSR-FAQ-on-Human-Rights-Assessments.pdf.

For each human right identified as being at risk, our research team assigned a rating of "Low," "Medium," or "High" for each of scale, scope, remediability, and likelihood. We defined them as follows:

Factor	Low	Medium	High
Scope	A small number of the relevant population is at risk.	Approximately half of the relevant population is at risk.	A significant majority or all of the relevant population is at risk.
Scale	The risk is associated with indirect adverse impacts on the rightsholder(s).	The risk is associated with direct adverse impacts on the rightsholder(s).	The risk is associated with lasting or irreversible adverse impacts on the rightsholder(s).
Remediability	Remedy can rarely or will not return those affected to the same or equivalent condition before the adverse impact occurred.	Remedy may or may not help return those affected to the same or equivalent condition before the adverse impact occurred.	Remedy would return or is likely to return those affected to the same or equivalent position before the adverse impact occurred.
Likelihood	0-33% chance of the risk occurring.	34-66% chance of the risk occurring.	67-100% chance of the risk occurring.

Our assessment, graphs, and tables account for the fact that High is worst and Low is best for scope, scale, and likelihood, while Low is worst and High is best for remediability.

Our analysis of the salience of human rights risks is visualized as follows:



This graph illustrates the salience of human rights risks based on four key criteria: scale, scope, remediability, and likelihood. For example, the risk to privacy from using LLMs in the legal field scores high in both scale and scope, has low remediability, and a moderate likelihood.

Below, we identify and describe each of the most salient risks in law, education, and journalism. Our analysis of each of the four key criteria for each salient risk is represented in tables, with the scope, scale, and likelihood of each right color-coded as High, Medium,

and Low, and with remediability color-coded as Low, Medium, and High, because the risk is higher the less possible it is to restore the rightsholder(s) or make amends to those who have suffered a human rights harm.

Most Salient Risks in Law

Right to Privacy.[120] This right is at risk when lawyers, litigants, or judges input confidential case information, including personal identifying information of plaintiffs or defendants, into LLMs. Our analysis focuses on cases in court, but LLMs can also threaten privacy when used for other legal tasks like drafting contracts.

Scope: High The risk to privacy of all plaintiffs and defendants is high in scope if their personal identifying information is entered into LLMs because one or more persons not associated with the case would have access to it. However, the number of impacted individuals is dependent on whether or not such confidential information is entered, which is addressed below under likelihood. Because legal systems usually involve two or more parties with concurring interests, these parties usually have stakes in keeping certain information private and making other information public to best support their claim.	**Remediability: Low** If confidential information is disclosed to the public or irrelevant parties, the harm, particularly to litigants, can be irreversible.
Scale: High If privacy is compromised and critical information becomes available to a third party, this information can affect the outcome of the case and therefore the life of the plaintiff or defendant. This is particularly true if identifying information becomes available to the public,[121] and not only employees of the company developing or deploying LLMs.[122] In high-stakes cases, including criminal cases, consequences could include imprisonment or release of individuals.	**Likelihood: Moderate** While better-resourced firms may use proprietary AI tools with stronger safeguards, many others lack access to such protections. This means that some firms will be less likely to disclose private information than others.

Due Process and Fair Trial Rights. These rights are at risk for plaintiffs and defendants in two main circumstances: when courtrooms use LLMs for decision-making and when legal professionals representing plaintiffs or defendants in court use LLMs to draft their

120 This encompasses both what data requires restricted access and how that access must be restricted.

121 Mireshghallah, Niloofar, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. "Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild." In *Proceedings of the Conference on Language Modeling (COLM 2024)*.

122 For example, tech companies in the United States can be summoned to provide users' information by the Department of Justice. In 2018, the U.S. Department of Justice demanded metadata on 73 phone numbers and 36 emails from Apple. Fung, Brian, and Evan Perez. "Justice Department Requested Data on 73 Phone Numbers and 36 Email Addresses from Apple." *CNN Politics*, June 11, 2021. https://edition.cnn.com/2021/06/11/politics/justice-department-apple-congress-leak-investigation.

submissions to the court. If courtrooms use LLMs for decision-making purposes, the right to due process and a fair trial for plaintiffs and defendants is at risk. If legal professionals representing plaintiffs or defendants in court use LLMs in their submissions to the court, that also puts their due process and fair trial rights at risk.

Scope: High Any litigant can be affected if lawyers or judges use LLMs at any stage of a trial, arbitration, or other proceeding.	**Remediability: Medium** If it is known that LLMs were used by courts or litigants, the possibility of appeal—except in the instance of the highest court—means this risk has the possibility of remediation, with higher remediability in first instance cases and among parties who can afford the costs of an appeal.
Scale: High The gravity of the risk is particularly high where the stakes of outcomes are high for plaintiffs and defendants, including criminal cases. Where case outcomes may set a precedent for future cases, particularly in courts where you appeal prior decisions, the use of LLMs presents risk to the rights of future plaintiffs and/or defendants.	**Likelihood: Medium** Some legal professionals are cautious in their use of publicly available LLMs, but the risk may be higher when professionals are overburdened and adopt these technologies to boost efficiency without adequate oversight or are experimental and use LLMs to develop new case theories. Currently, LLMs can hallucinate caselaw, which is a clear indicator of AI-generated legal documents and usually identified.

Other Relevant Risks: The use of LLMs in the legal field can also affect the right to remedy when, for instance, lawyers overly rely on LLMs for legal research and overlook certain legal nuances or precedent; freedom of thought[123] when legal professionals lean too heavily on LLM-generated reasoning and neglect their own critical thinking capabilities; and labor rights,[124] particularly when junior lawyers face barriers to employment because senior lawyers automate routine tasks such as writing citations using LLMs.

Most Salient Risks in Education

Freedom of Thought. Freedom of thought is at risk when educators or students use LLMs to regularly generate content, including essays or code, and when students receive

123 Freedom of thought is a core, non-derogable human right that protects the inner forum of an individual's mind from external interference. It encompasses the right to hold thoughts privately, to resist coercion or manipulation of one's mental processes, and to be free from punishment based solely on one's thoughts. Recognized under Article 18 of the International Covenant on Civil and Political Rights (ICCPR), this right has been affirmed as absolute and inviolable. See Alegre, Susie. 2022. *Freedom to Think: The Long Struggle to Liberate Our Minds*. London: Atlantic Books.

124 Labor rights affirm the freedom of association and effective recognition of the right to collective bargaining; the elimination of all forms of forced or compulsory labour; the effective abolition of child labour; the elimination of discrimination in respect of employment and occupation; and a safe and healthy working environment. See International Labour Organization. *ILO Declaration on Fundamental Principles and Rights at Work, as Amended in 2022*. Geneva: ILO, 2022. https://www.ilo.org/sites/default/files/2024-04/ILO_1998_Declaration_EN.pdf.

LLM-generated feedback in response to their ideas rather than instructor feedback. The use of LLMs by educators and students can make their minds more vulnerable to outside pressure from LLM responses, other digital platforms and tools, and individuals.

Scope: High Students have high exposure to LLM-generated content from their own use as well as materials from their educators. This means they will widely be subject to limited freedom of thought, particularly if students use them in their coursework.	**Remediability: Low** Once students' thinking skills have atrophied, they are hard to recover. Further, for students who have not developed critical thinking skills as children, opportunities to acquire those later in life may be fewer.
Scale: High Scholars have connected overreliance on LLMs to neural, linguistic, and behavioral underperformance and raised concerns over accumulation of a "cognitive debt" when students rely on LLMs for essay writing.[125] These can impact students' mental development and learning throughout their lives.	**Likelihood: High** The likelihood of this harm occurring in the field of education is high given the wide accessibility and low cost of LLMs for any student or educator with a computer or smartphone and an internet connection.

Right to Education. LLMs can be used in classrooms but are not usually calibrated to align learning content with the instructor's, school's, or system's pedagogical methods. This mismatch can undermine students' ability to process, retain, and deeply understand material, which can impact the development of the human personality and dignity of students and affect their ability to participate effectively in society. While students' formal access to education may remain, the quality of their learning and the depth of their comprehension are at stake.

Scope: High Due to the growing adoption of LLMs by educators, an increasing number of students are exposed to LLM-generated content and feedback. Because general-purpose LLMs are free and widely accessible, it is very easy for students to use them in their coursework.	**Remediability: Low** If foundational knowledge is not acquired early on, remediation later becomes significantly more difficult, making this risk particularly salient for younger learners and individuals learning new subjects and skills.
Scale: High Because foundational knowledge is necessary to the acquisition of more complex concepts, the scale of this risk is particularly high for younger students, whose basic skills are still developing. Inadequate grounding at this stage can have lasting educational consequences.	**Likelihood: Medium** The right to education is primarily impacted in settings where LLMs are heavily relied upon without sufficient human oversight to bridge the gap between content delivery and learning methods.

125 Nataliya Kosmyna et al., *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task*, arXiv, June 10, 2025. https://doi.org/10.48550/arXiv.2506.08872.

Other Relevant Risks: The use of LLMs in education can also impact children's rights if overreliance on LLMs impact minors' education. This can include impacts to freedom of access to information when educators improperly use LLMs as search engines and rely on inaccurate or incomplete information as a result; right to privacy when educators input personal and potentially identifying information into LLMs; labor rights if educators are expected to perform more tasks or teach more students with the assistance of LLMs; and the right to equality and non-discrimination if some students have access to LLMs and others do not.

Most Salient Risks in Journalism

Freedom of Thought. Reliance by journalists on LLMs to write, source, and fact-check articles can erode their capacity to select, verify, and analyze information, thus limiting their freedom of thought.

Scope: High If journalists use LLMs to write, source, and fact-check articles, the number of journalists whose freedom of thought would be at risk is high because the journalists may produce material without doing the background work themselves.	**Remediability: Low** Similar to the severity of impact to journalists, remediability will depend on the journalists' number of years of experience. For journalists with years of experience, the remediability of the risk is quite high, whereas for journalists with little or no experience, it may take more time to remediate the impact to their freedom of thought.
Scale: High The scale of impact to journalists' freedom of thought if they use LLMs to write, source, and fact-check articles depends on their years of experience reporting on their own. While more senior journalists' freedom of thought won't likely be impacted much at all, more junior journalists will be unlikely to develop the mental skills necessary to do this work on their own.	**Likelihood: Medium** To the degree that journalists abide by the high level of human oversight that is required by journalistic standards, they will still exercise critical thinking, fixing any inaccuracies, lack of nuance, or bias. However, market-driven incentives, time pressures, and temptations to outsource thinking to LLMs can incentivize journalists to rely more on LLMs for additional aspects of their work, which may degrade their ability to exercise oversight and think independently over time.

Right of Access to Information. The right of a journalist and a reader to access information may be threatened when journalists use LLMs for newsgathering, research, and article writing. If an LLM misses a critical piece of evidence or fails to recognize relevant patterns in a body of evidence, journalists' articles may not identify important information.

Scope: High Both journalists and their readers may be affected. The number of journalists and readers at risk if a journalist uses LLMs for newsgathering, researching, and article writing can be quite high. Our research reflects that journalists often use LLMs for other purposes, and that often they use LLMs specifically crafted for journalistic reasons. Nonetheless, we also understand that LLMs are used by some journalists for these purposes.	**Remediability: Low** Even if journalists wish to amend an article to include information that was missed, they may have difficulty identifying what information was missed. This problem is exacerbated by the fact that LLMs tend to overindex on mainstream perspectives, ignoring marginalized perspectives that are also harder for people to find.
Scale: High When LLMs are used to perform research, the omission of a particular piece of information in an article may not be an intentional editorial choice made by a professional journalist. This means that readers may not be exposed to potentially critical information in shaping their understanding of the topic. Routine exposure to inaccurate information that the reader expects to be accurate can result in a failure to critically analyze claims and the formation of opinions based on misinformation. Surface-level information might provide readers with a simplistic picture while inducing them to think that they have a complete and thorough understanding of the topic. This may erode readers' ability to perceive and grapple with nuance over time. Biased information may pervasively change a reader's worldview by normalizing a certain interpretation of the world.	**Likelihood: Medium** Due to cultural norms around impartiality and accuracy, journalists may be cautious in their approach to minimize the odds of missing information. However, in fast-paced settings, there may be a strong incentive to use LLMs to speed up the newsgathering process and reduce time-consuming human oversight. Errors may be more likely to occur when applying LLMs to research than to writing because of the vast datasets typically produced by LLMs when used for research purposes.

Other Relevant Risks: The use of LLMs in journalism can also impact the right to equality and non-discrimination if the outputs contain misinformation about or bias towards minority communities or other marginalized groups; and labor rights, particularly when newsrooms rely on human-LLM collaboration to achieve the same work with fewer journalists.

Most Salient Cross-Cutting Risk

Freedom of Thought. Freedom of thought is most critically at risk when people over-rely on LLMs, introducing potentially inaccurate information and bias that is hard to detect. This affects both professionals who use LLMs to complete their tasks and the people who consume the content the professionals create because of any biases or inaccuracies introduced by the LLM. Freedom of thought also has a risk of becoming more serious over time because over-reliance may reduce professionals' ability to detect inaccuracies and biases, and reliance on LLMs may become an obligation rather than a choice. Given that LLMs produce text, they are particularly applied within the knowledge economy, which is an area where critical thinking is particularly essential to jobs.

5.2.2 Opportunities

Opportunities are not required as part of human rights due diligence processes, and therefore, no set criteria for how to evaluate them exists. We decided to include opportunities

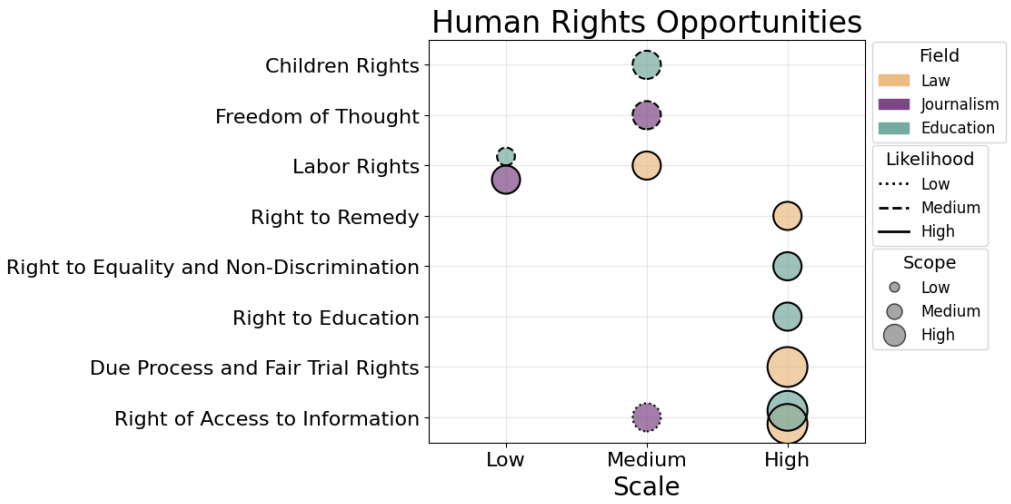
because our research showed they were most experienced by professionals in Global South communities. The human rights we identify as opportunities are defined in the Glossary in Appendix B. We borrowed from the criteria to evaluate the saliency of human rights risks, and applied similar considerations to assess the saliency of human rights opportunities, considering the scale, scope, and likelihood of potential benefits:

- **Scope:** How widespread is the benefit to individuals and/or communities?
- **Scale:** How valuable is the benefit?
- **Likelihood:** How likely is the benefit to occur?

We excluded remediability because opportunities do not require corrective action. Our research team scored a rating of "Low," "Medium," or "High" for each of the three criteria for each human right we identified as potentially benefiting from the introduction of LLMs. We defined them as follows:

Factor	Low	Medium	High
Scope	The benefit will reach few people in the relevant population.	Approximately half of the relevant population will benefit.	The benefit will reach a significant majority or all of the relevant population.
Scale	The benefit has fairly low value to the rightsholder(s).	The benefit has medium value to the rightsholder(s).	The benefit has high value to the rightsholder(s).
Likelihood	0-33% chance of the benefit occurring.	34-66% chance of the benefit occurring.	67-100% chance of the benefit occurring.

Our analysis is presented in the graph below. While LLMs may impact the same human right positively (opportunity) as it does negatively (risk), an opportunity does not justify a risk.



This graph represents the salience of human rights opportunities in journalism, law, and education based on three criteria: scale, scope, and likelihood. For example, the labor rights opportunity from using LLMs in the legal field scores medium in scale, medium in scope, and has a high likelihood.

We identify and describe each of the most salient human rights opportunities in law, education, and journalism. Our analysis of each of the three criteria for each salient opportunity is represented in tables, with the scope, scale, and likelihood of each right color-coded as High, Medium, and Low.

Most Salient Opportunities in Law

Right of Access to Information. LLMs can explain legal procedures or ideas, simplify complex legal documents, translate them into multiple languages, and help interpret case law and technical documents.

Scope: High	Scale: High	Likelihood: High
Legal professionals can use LLMs broadly to interpret case law and technical documents outside of their area of expertise. While specialized systems have been developed to assist in identifying relevant case law, such tools are not accessible widely, unlike general LLMs. Litigants and other parties can similarly use general-purpose LLMs to learn information relevant to their cases, and other information-seekers more broadly can use them to learn legal information applicable to their daily lives even if they are not engaged in litigation.	The specialized technical language in legal documents requires advanced training to comprehend and interpret. LLMs can bridge this gap for those without legal training, making the opportunity to access information significant in scale.	Given the wide availability of general-purpose LLMs in multiple languages and their ease of use, it is very likely that litigants and other parties as well as users will benefit from improved access to information in the legal field. Legal professionals can also benefit from public tools by growing their knowledge of law outside their area of expertise.

Due Process and Fair Trial Rights. LLMs can enhance access to justice by improving litigants' and other parties' understanding of legal procedures, including those that typically do not involve representation, like small claims. Furthermore, they can expand access to a variety of tasks that may be unaffordable because they typically require legal assistance, such as drafting contracts.

Scope: High	Scale: High	Likelihood: High
LLMs can expand the scope of people who are able to benefit from legal proceedings by making them more accessible to the public that may not know where to start. They can also assist both self-represented litigants and litigants with legal counsel by improving their understanding of legal procedures and potentially preventing negligence or malpractice by legal representatives.	Improved understanding can help litigants make informed decisions about legal representation, predict potential costs, and better engage with their cases, all of which are critical components of due process and a fair trial. For self-represented litigants with limited financial means, this is particularly impactful because LLMs may be the only affordable option to understand their legal case.	Growing efforts to enhance access to justice with public LLM-based tools, as well as the broad availability of general-purpose models, increase the likelihood that prospective litigants and other parties will benefit from improved due process.

Other Relevant Opportunities: The use of LLMs in the legal field can positively impact other human rights. These include the right to remedy, for example, when litigants use LLMs to better anticipate and allocate resources to the various stages of their legal cases;

and labor rights – when legal professionals delegate repetitive tasks to LLMs to focus on higher-level tasks and developing connections with their clients.

Most Salient Opportunities in Education

Right of Access to Information. LLMs represent an opportunity to access more complex and diverse information in the context of education. In particular, they can simplify and translate material and provide tailored explanations in a variety of styles to accommodate different needs for accessing information.

Scope: High	Scale: High	Likelihood: High
Both students and educators can benefit from better access to information using LLMs. For multilingual learners and learners with disabilities, LLMs are an opportunity to adapt existing learning materials so that they can better access the information contained in them. In under-resourced educational institutions where educators are unable to meet the full needs of their students, LLMs can increase the number of students who can access information in a format that they can understand. For educators, LLMs represent an opportunity to access materials in foreign languages.	LLMs not only enable access to information that was previously unavailable due to its complicated or unfamiliar nature, but also support deeper understanding of content by generating tailored exercises and individualized feedback. These abilities collectively support both the availability and legibility of information, increasing access on multiple levels.	Even as access to devices and the internet varies in educational settings, LLMs are being increasingly integrated into smartphone applications, making them more broadly accessible. Given the wide array of LLM capabilities, the likelihood of this opportunity is high across educational levels, subject areas, and geographies.

Right to Equality and Non-Discrimination. LLMs in classrooms offer an opportunity to improve learning for some students, including those with disabilities or studying in a second language. The personalization features that LLMs enable in educational contexts can create equitable opportunities for these groups to learn educational settings.

Scope: Medium	Scale: High	Likelihood: High
LLMs can impact the right to equality and non-discrimination for many students, including those with disabilities, learners who require a specific style of teaching, and multilingual students. However, the number of students who benefit is limited by the digital divide: only students with access to necessary resources, such as a computer and reliable internet, can take advantage of LLMs. These students may also already be more likely to receive individualized support, including tutoring or other external resources.	The potential impact of LLMs is substantial because they can enable educators to adapt content to diverse learning styles, particularly in resource-constrained environments where such personalization would otherwise be difficult or impossible.	Given wide access to LLMs for students and educators with a computer or smartphone and internet connection, and the large need for tailored teaching methods, students across educational levels, subject areas, and geographies are likely to benefit from this opportunity.

Other Relevant Opportunities: The use of LLMs by educators can also benefit the right to education when LLMs are used as a tool by educators to further and adapt their teaching depending on the specific needs of students; children's rights, when children benefit from a better and more individualized education due to their own use of LLMs or the use of LLMs in the classroom.

Most Salient Opportunities in Journalism

Freedom of Thought. In journalism, LLMs can enhance freedom of thought by reducing the time spent by journalists on repetitive tasks, such as summarization and editing, thereby freeing journalists to focus on investigative work and in-depth reporting. Some journalists report using their extra time to conduct additional fact-checking and bias analysis. This can also improve the quality and diversity of content that readers have exposure to, expanding the knowledge base from which they form their opinions.

Scope: Medium	Scale: Medium	Likelihood: Medium
Journalists who work on beats involving repetitive tasks, like sports, may have more time to pursue intellectually stimulating stories, while journalists who produce specialized or in-depth content may not see a change, as their tasks are not easily delegated to LLMs. If more in-depth articles are published as a result, readers may also benefit from more reporting to use when forming opinions.	LLMs free up time in fast-paced journalistic environments. However, this simply creates more time to focus on intellectually challenging tasks, instead of transforming the way journalists execute them by introducing new tools or training. For readers, access to articles that come out of these tasks can help them engage with the subject matter at a deeper level.	Even when repetitive tasks can be offloaded, the extra time is not always reinvested in higher-level work and may be diverted to administrative tasks or offset by newsroom decisions to reduce staff.

Labor Rights. LLMs can help improve working conditions, access to language and coding skills, and long-term job security for journalists.

Scope: Medium	Scale: Low	Likelihood: High
Journalists working on time-sensitive beats, such as breaking news and severe weather affecting peoples' immediate safety, can benefit from faster or more extensive newsgathering. However, journalists that cover other beats that require relationship development, such as community reporting or investigative reporting, may not benefit in this way.	While LLMs may reduce some time pressure by completing repetitive tasks, some efficiency gains may be offset by heightened expectations to produce more stories with fewer resources.	Journalists have eagerly and rapidly adopted LLMs for tasks requiring minimal oversight, and increased industry-wide guidelines on appropriate LLM use emboldens journalists to use LLMs and receive the full benefits.

Other Relevant Opportunities: The use of LLMs in journalism can also impact the right of access to information, when time savings allow journalists to conduct deeper investigations or document on underreported topics.

Most Salient Cross-Cutting Opportunities

Access to Information. Across all sectors, enhanced access to information emerged as a salient opportunity associated with LLMs. LLMs can dynamically improve access to information across a wide variety of domains by adapting content to one's level of understanding and translating it to a preferred language. In addition, it can enhance comprehension by listing examples, creating questions to check understanding, and providing feedback, thus allowing users not only access to new technical or dense content but also to deepen their understanding of previously available information. Both professionals and recipients of services can benefit from these capabilities.

5.2.3 Takeaways

Our analysis suggests that the salience of both human rights risks and opportunities often stems from the same underlying uses of LLMs. For example, the use of LLMs by journalists can both impair and enhance freedom of thought depending on how and by whom they are used, making human rights outcomes more difficult to predict. Further, human rights opportunities are tempered by the risks associated with the use of LLMs. For example, even if LLMs may enhance access to information, such an opportunity may become a risk if hallucinations occur.

Our analysis also reveals that while certain human rights risks are field-specific, many risks span several professional domains, even if they manifest in different ways. This calls for a coordinated response by various actors both at the cross-field and the field-specific levels. At the cross-field level, relevant actors include governments and those who develop and deploy LLMs. At the professional field level, relevant actors include industry groups, which are in the best position to understand the context-specific risks and opportunities of LLMs in their profession.

6. IDENTIFICATION OF APPROPRIATE ACTION FOR REMEDIATION

The numerous human rights risks posed by large language models (LLMs) require a coordinated response that spans the entire AI ecosystem, since risks emerge from the interplay of many actors who each wield distinct levers of action. Below, we identify and discuss the unique roles and responsibilities of four categories of actors—development actors, deployment actors, industry groups, and government actors.

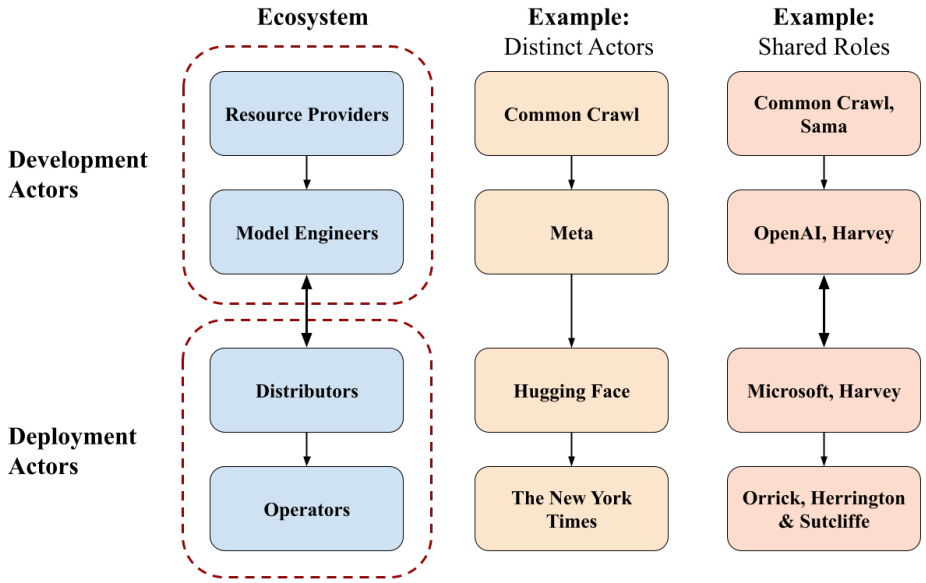
Two primary categories of actors are deeply involved in the model development lifecycle. Development actors, which contribute to the creation of models used in AI systems, include both resource providers and model engineers. Resource providers, such as Sama (which supplies labeled data) and Nvidia (whose hardware enables model training), set the availability, performance, and cost of model development resources, placing practical limits on what models can be trained and deployed. Model engineers, including OpenAI (developer and deployer of the GPT model family) and Harvey (which tailors LLMs for specialized legal use), make choices about how models are designed, trained, and improved. These models then reach the public through deployment actors, which expose AI systems to users, including distributors and operators. Distributors, such as Hugging Face (through its Model Hub and inference APIs) and Harvey (by integrating LLMs into legal research platforms), make the technology accessible across industries and geographies. Operators, like Orrick, Herrington, & Sutcliffe, LLP (which has purchased LLM-based tools for its legal practice)[126] and *The New York Times* (whose journalists use AI tools for reporting),[127] guide how LLMs are used in professional practice and have direct influence over users' experiences.

Professional norms and policy frameworks are shaped by industry groups and regulators, who play a crucial role in guiding the responsible use of LLMs. Industry groups, including the Society of Professional Journalists in journalism and National Bar Associations in law,

126 Orrick, Herrington & Sutcliffe LLP, "AI Adoption @ Orrick," Orrick. https://www.orrick.com/Innovation/AI-Adoption-at-Orrick.

127 *The New York Times,* "How The New York Times Uses AI in Journalism," October 7, 2024. https://www.nytimes.com/2024/10/07/reader-center/how-new-york-times-uses-ai-journalism.html.

set standards, offer training, and help practitioners navigate the changing landscape of LLM tools. Their recommendations can shape human rights-aligned practices. Regulators create the legal environment for LLMs at the local, national, and international levels. Examples include the New York City Council, Brazil's National Data Protection Authority, and the European Union. Regulatory approaches vary based on governments' priorities: some focus on data protection, while others prioritize innovation or national control of AI infrastructure.



Development and deployment actors in the AI lifecycle.

Our interviews show that weaknesses in one actor's sphere of influence can have ripple effects across the entire system. For instance, if news organizations as operators of LLMs provide limited training for the journalists they employ, the risk of spreading misinformation increases for the news makers and news readers. Accordingly, our recommendations address each category of actor and emphasize the need for coordination and shared responsibility across the ecosystem. The table below summarizes our recommendations and which actor(s) are responsible for undertaking them.

	LLM Developers		LLM Deployers			
	Resource providers *Ex: Scale AI*	**Model Engineers** *Ex: Meta*	**Distributors** *Ex: Hugging Face*	**Operators** *Ex: Orrick, Herrington & Sutcliffe LLP*	**Industry groups** *Ex: International Bar Association*	**Government** *Ex: Local, national, and international bodies*
1. Assess human rights impacts of LLMs	✓	✓	✓	✓	✓	✓
2. Define remediation stages	✓	✓	✓	✓	✓	✓
3. Regulate general-purpose AI						✓
4. Set guidelines for the use of LLMs in professional contexts				✓	✓	
5. Diversify training data	✓	✓				
6. Create user feedback mechanisms			✓	✓		
7. Design LLM interfaces responsibly			✓			
8. Tailor tools to professional tasks	✓	✓	✓	✓	✓	
9. Cultivate critical AI literacy	✓	✓	✓	✓	✓	✓

1. Assess Human Rights Impacts of LLMs

Resource Providers | Model Engineers | Distributors | Operators | Industry Groups | Government

We recommend that all stakeholders at different stages of the AI lifecycle conduct comprehensive human rights assessments.[128] We recommend that governments require these assessments of each of the named stakeholder groups.

128 Human rights impacts can refer to both risks and harms. Human rights risks refer to the potential for an activity, decision or tool to negatively impact the rights and freedoms of individuals or communities. Human rights harms, abuses, or violations occur when those rights and freedoms are actually infringed, through either action or inaction. In this sense, a human rights risk is the possibility or threat of a human rights harm occurring.

Human rights risk assessments should follow corporate due diligence principles described in Articles 17 and 18 of the UN Guiding Principles for Business and Human Rights (UNGPs)[129] and include the following steps, which can be tailored to the operations of different development and deployment actors, and scaled up or down depending on the size of the corporation:

1. **Identify risks and impacts** through a literature review and stakeholder interviews.
2. **Assess and prioritize risks** based on analysis of severity.
3. **Identify appropriate actions** to mitigate risks and remedy harm.

Development actors, who contribute to the creation of AI systems, should assess the human rights impacts of design decisions and model behavior:

1. **Resource providers** should assess the human rights impacts of data collection pipelines and hardware development practices.
2. **Model engineers** should conduct model-level rights evaluations, as described in Section 5.1, to assess the model's outputs before deployment.

Deployment actors, who interact more directly with users, should assess how LLMs interact and integrate with larger systems[130] across applications and use cases and design their evaluation procedures to center human rights:[131]

1. **Distributors** should assess the human rights impacts of the access and distribution of models by operators.
2. **Operators** should engage with users directly to understand how they are applying LLMs to their professional tasks and analyze the impacts that arise as a result.

We recommend that risk assessments include engagement with and analysis of the following:

1. **Regional Context:** This involves accounting for how geopolitical and cultural differences affect the use, effects, and perception of LLMs.
2. **Vulnerable Groups:** This involves identifying communities most vulnerable to potential harm for each context, which may include socially, economically, and politically marginalized communities.

---

129 United Nations Human Rights Council. 2011. *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework.* Geneva: United Nations.

130 Boming Xia, Qinghua Lu, Liming Zhu, and Zhenchang Xing, "An AI System Evaluation Framework for Advancing AI Safety: Terminology, Taxonomy, Lifecycle Mapping," in *AIware '24: Proceedings of the 1st ACM International Conference on AI-Powered Software* (New York: Association for Computing Machinery, 2024), 74–78. https://doi.org/10.1145/3664646.3664766.

131 Vyoma Raman, Camille Chabot, and Betsy Popken, "Assessing Human Rights Risks in AI: A Framework for Model Evaluation," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, no. 3 (2025): 2107–2120. https://doi.org/10.1609/aies.v8i3.36699.

2. Define Remediation Stages

Resource Providers | Model Engineers | Distributors | Operators | Industry Groups | Government

We further recommend that actors involved in the development, deployment, and use of AI systems collectively develop a process for each actor to engage in remediation.[132] As described in the UNGPs, remediation involves compensating for and correcting human rights harms.[133,134] There are five types of remedy: (1) restitution (restore, to the extent possible, whatever has been lost and return the rightsholder to the state before the abuses occurred); (2) compensation (money or other trade-offs for the cost of the harm); (3) rehabilitation (medical, psychological, legal, and social services to restore the victim); (4) satisfaction (cease the violation, acknowledge the harm, disclose the truth, provide an apology, and sanction those responsible); and (5) guarantee of non-repetition (change policies and procedures to prevent future harms and/or take disciplinary action).[135] The remediation process should be established at the outset to address potential harm to users as soon as possible after it occurs. The remediation process should include three stages: identifying, assigning responsibility for, and remedying harms.

1. Identifying Harms

Identifying harm as early as possible in the development and deployment of LLMs is essential to minimizing the scope and scale of human rights violations. Various actors can play a role in identifying harm at different stages of the process:

1. **Development stage:** Developers can identify and prevent some human rights risks before public release through rigorous evaluation of AI systems and human rights assessments.
2. **Deployment stage:** Deployers can uncover harms through human rights assessments and by monitoring usage in real time for subtle but pervasive harms.
3. **User feedback:** Users themselves may detect harms while interacting with LLMs. Therefore, it is necessary to establish clear channels for users to report harms and protocols, filtering and prioritizing user-reported issues, and notifying all relevant stakeholders that the harm has occurred.

---

132 See guidelines and examples here: Business for Social Responsibility. *Remedy for Generative AI-Related Harms*. February 2025. https://www.bsr.org/files/BSR-Remedy-for-Generative-AI-Related-Harms.pdf.

133 United Nations Human Rights Council, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework,* A/HRC/17/31, March 21, 2011. https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf.

134 Remedy can take five common forms: restitution, compensation, rehabilitation, satisfaction, and guarantees of non-repetition. See here for more: Business for Social Responsibility. *Access to Remedy*. August 2021. https://www.bsr.org/reports/bsr-access-to-remedy.pdf.

135 Business for Social Responsibility. Access to Remedy. August 2021. https://www.bsr.org/reports/bsr-access-to-remedy.pdf.

2. Assigning Responsibility and Defining Remedy

The UNGPs specify that actors that cause harm have the primary obligation to remedy that harm; those that contribute to harm share responsibility for remediation; and those directly linked to harm through their operations or services must use their leverage to prevent or mitigate future harms. We recommend that the six categories of actors work collectively to establish grievance mechanisms that interoperate across the AI supply chain so that information about harms flows to the party best positioned to address them.

Industry groups are most familiar with professional norms in their field and can coordinate across various organizations. For this reason, they should play a coordinating role by predefining responsibility-sharing frameworks and translating domain-specific risks, while developers and distributors must commit to embedding them in contracts and governance processes. Regulators should require and enforce their adoption rather than attempt to design them directly, ensuring feasibility. Such responsibility-sharing frameworks should ask:

1. Who is responsible for identifying the cause, contribution, and direct linkage to the harm?
2. Who decides how to address the harm to the affected user?
3. Who oversees systemic changes to prevent future occurrences?

3. Implementing Remedies

With responsibilities assigned, remedies must be timely, effective, and proportionate to the harm. Stakeholders should consider the following factors when deciding the remedy best suited to the case:

1. **Resources Needed:** These could include financial support, technical expertise, workforce capacity, direct communication with users, and database knowledge.
2. **Best-Suited Actors:** Based on the resources required, certain actors may be better positioned to address the harm, regardless of whether they were responsible for causing it.

3. Regulate General-Purpose AI

Government

While our analysis focused on sectoral impacts, and human rights concerns often materialize in concrete applications, it remains essential to regulate general-purpose AI as a whole. Many governing bodies have yet to establish regulations specifically addressing general-purpose AI (GPAI), with the exception of the European Union.[136] Several of the

136 European Commission. General-Purpose AI Code of Practice. Brussels: European Commission, 2025. https://digital-strategy.ec.europa.eu/en/policies/ai-code-practice.

most significant risks we identified, including threats to freedom of thought and equality and nondiscrimination, are intrinsic to how GPAI is used more broadly and therefore cut across sectors. Regulatory solutions must operate at the technology level in order to address the structural sources of harm, rather than only the downstream effects. At the same time, because human rights are especially salient in high-risk contexts such as healthcare, law, and education, domain-specific regulation is also necessary to provide more tailored protections. A dual approach combining general and sector-specific regulation of GPAI systems is needed to safeguard human rights comprehensively.

4. Set Guidelines for the Use of LLMs in Professional Contexts

Operators | Industry Groups

Because professions can impact peoples' access to rights, they often have formal and informal industry standards designed to enhance and protect these rights. Thus, we recommend that industry groups augment these by establishing guidelines on how LLMs can be applied within their respective professional fields. National and regional industry groups should be responsible for creating context-specific rules tailored to their jurisdiction, while international industry groups should collaborate to develop overarching guidelines that can be adapted across different global contexts. Operators should then operationalize these guidelines for internal organizational standards.

These guidelines should focus, to the extent possible, on maximizing opportunities from LLMs unique to each profession while mitigating risks to rights. For example, if the risk to the right to information when using LLMs to report on conflicts is deemed too high, how can journalists balance this with the necessity for timely and rapid reporting with limited resources in a conflict zone? While not an industry group, the *Associated Press* has released training on how journalists can integrate LLM tools into their professional workflows while balancing ethical commitments.[137] Guidance like this should be mainstreamed across the sector.

5. Diversify Training Data

Resource Providers | Model Engineers

Dataset diversity is a human rights imperative. Homogeneous training data entrenches existing power dynamics by privileging dominant cultures while marginalizing the knowledge and experiences of underrepresented groups. This lack of representation undermines the rights to equality, non-discrimination, access to information, and cultural participation and runs counter to the human rights ethos of transcending geopolitical boundaries. It also risks enforcing majoritarian values as defaults and suppressing pluralism in ways that contradict international human rights instruments such as the International

137 Associated Press Workflow Solutions. "AI in the Newsroom: AP's June 2025 Course on Practical, Ethical Adoption." *AP Workflow Solutions*. https://workflow.ap.org/news/ai-in-newsrooms-course-june-2025/.

Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social and Cultural Rights (ICESCR).

Responsibility for realizing these commitments falls on both resource providers and model developers. Data contractors and collection platforms must adopt rights-based practices that ensure diverse sourcing while providing labor protections for workers whose annotations shape model preferences. Model developers bear responsibility for auditing the representativeness of the datasets they use and prioritizing pluralism.

6. Create User Feedback Mechanisms

Distributors **Operators**

We encourage deployers to establish robust user feedback mechanisms in support of the rights to remedy, accountability, participation, and transparency. Without accessible channels for feedback, individuals are left without recourse, undermining their right to remedy, and eroding accountability for rights impacts. Feedback mechanisms also operationalize the right to participate in decisions that shape people's digital environments. Interfaces used to report feedback must be designed and tested for access and ease-of-use for people from varied backgrounds and with different abilities. Finally, while individual reports like incident databases[138] capture direct user harms, distributors and operators must also proactively assess aggregate risks such as stereotyping or cultural erasure, which may not be visible at the individual level but can be identified through audit-style assessment.

7. Design LLM Interfaces Responsibly

Distributors

Responsible interface design must center the human rights of access to information and privacy. In addition to model cards[139] and data sheets[140] that are often publicly available but primarily understandable only to experts, users need interfaces that clearly communicate the model's limits, including understandable indicators of uncertainty and data scarcity, so they can make informed judgments rather than passively accept outputs. Privacy protections should be built in as default configurations: no collection of personal data unless necessary for functionality, with opt-in consent and plain-language explanations of what is being collected and why. Design choices like discarding conversation data after the user finishes,

138 Responsible AI Collaborative. *AI Incident Database*. https://incidentdatabase.ai.

139 Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA: Association for Computing Machinery, 2019), 220–29. https://doi.org/10.1145/3287560.3287596.

140 Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford, "Datasheets for Datasets," *Communications of the ACM* 64, no. 12 (2021): 86–92. https://doi.org/10.1145/3458723.

allowing users to erase all their data or interact anonymously with one click, and providing plain-language explanations of privacy issues and controls can bridge the gap between the availability and legibility of information, mitigating common issues with transparency in practice.[141]

Equally important is ensuring cultural competency and inclusivity in design. Interfaces must be accessible and intuitive across diverse cultural, linguistic, and literacy contexts to prevent exclusion or indirect discrimination. This includes offering multilingual options and avoiding interface elements that may carry conflicting cultural meanings. These cumulatively require co-design testing with potential users across regions, socio-economic groups, and ability levels, ensuring systems are usable outside of the privileged contexts of their creators.

8. Tailor Tools to Professional Tasks

Resource Providers | Model Engineers | Distributors | Operators | Industry Groups

We encourage industry groups to create specifications for LLM tooling that mitigate relevant risks specific to their profession. Industry groups can then work with model engineers and distributors to implement these with available resources. For instance, most interviewees in the legal field highlighted data privacy as a significant concern. Harvey AI, a LLM company producing legal tools, was designed with stricter privacy controls[142] than general-purpose LLMs for handling personal data. This rights-based tailoring ensures that LLM deployment aligns with professional obligations and priorities and the most salient risks in each profession are preemptively mitigated.

9. Cultivate Critical AI Literacy

Resource Providers | Model Engineers | Distributors | Operators | Industry Groups | Government

All actors should advance critical AI literacy to protect users' dignity, agency, and freedom of thought. Industry groups and operators should lead efforts to equip professionals with the skills to question and contextualize LLM outputs in their fields. Developers and distributors should ensure that individuals using LLMs in personal or professional contexts are given accessible explanations of system limits and risks so they can exercise informed choice rather than passive reliance. Governments, working in partnership with developers and distributors, should integrate AI literacy into school curricula. Each of these components should be developed with accessibility, inclusivity, and cultural relevance in mind to ensure that AI literacy does not become yet another site of exclusion.

---

141 Ananny, Mike, and Kate Crawford. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20, no. 3 (2018). https://journals.sagepub.com/doi/10.1177/1461444816676645.

142 Counsel AI Corporation, *Privacy Policy*, last updated November 22, 2024. https://www.harvey.ai/legal/privacy-policy.

At a minimum, these training programs should cover the following key areas:

1. **Model limitations** (e.g., knowledge cutoffs, RAG[143] use): Equipping users to recognize contexts in which outputs may be hallucinated or incomplete, protecting their right to accurate and timely information.
2. **Assessing reliability:** Building the capacity to evaluate outputs critically, reinforcing agency and autonomy rather than passive reliance on the system.
3. **Recognizing bias:** Raising awareness of how outputs can perpetuate stereotypes or exclusion, supporting the right to equality and non-discrimination.
4. **Privacy practices:** Ensuring users understand data collection and consent options, safeguarding the right to privacy.

In the long term, fostering critical AI literacy in a population is the most influential way to mitigating risks to all rights: it can help users apply LLMs more strategically in underspecified contexts and may ultimately be the strongest line of defense against harms arising from new and emerging technologies in the future.

143 RAG (Retrieval-Augmented Generation) implementation refers to a technique used in natural language processing (NLP) models, where a retrieval mechanism is combined with a generative model to improve the quality and accuracy of responses. In RAG, relevant information is first retrieved from a large corpus or database and then fed into a generative model (like GPT) to create a response.

7. CONCLUSION

The strength of our assessment lies in our distinctive approach that draws on interdisciplinary expertise, international engagement, and field-specific case studies to evaluate the human rights impacts of LLMs. Combining legal, social, scientific, and technical perspectives, we apply the United Nations Guiding Principles on Business and Human Rights to connect the real-world applications of LLMs in law, education, and journalism to concrete human rights impacts. Our international approach surfaces disparities in access, infrastructure, linguistic inclusion, and geopolitics that mediate the effects of LLMs across different contexts and amplify potential risks and opportunities. These insights directly inform our recommendations, which involve both policy and model-level interventions that are context-aware and technically feasible.

We use our international stakeholder engagement and model-level rights evaluation to analyze the most salient human rights risks and opportunities across law, education, and journalism. Our analysis reveals that across these fields, freedom of thought may be most at risk, while the right of access to information is most likely to be enhanced by LLMs. We find that while there are some differentiated risks across law, education, and journalism, many risks span all three fields but manifest in different ways based on the context in which LLMs are used. Certain human rights are uniquely connected to the purpose of a particular professional field, highlighting the importance of considering the social contribution of specific tasks when integrating LLMs into workflows. We point out the contribution of several factors—including resource divides in geographic regions and places of employment, types of tasks, and career seniority—in exacerbating the risks associated with the use of LLMs. Based on these findings, we recommend a coordinated response by all relevant actors—governments, LLM development and deployment actors, and industry groups—to limit human rights risks and harness opportunities holistically across every stage of the life cycle of machine learning.

Our approach offers a model for future human rights impact assessments of AI systems. As LLMs and other emerging technologies continue to shape rights-impacting domains, developers, deployers, industry groups, and governments alike must adopt methodologies that integrate human rights principles, computational realities, and local context. We urge researchers, practitioners, and corporations creating AI systems to adapt and apply this framework to other sectors to ensure that the development and behavior of AI upholds the rights and well-being of all individuals.

8. ACKNOWLEDGEMENTS

Many thanks to our Advisory Board members, all of whom helped guide the development of this report. They include:

- **Lindsey Andresen**, Director, Responsible Technology, BSR
- **Garance Burke**, Global Investigative Journalist, The Associated Press
- **Hannah Darnton**, Model Policy and Safety Systems, OpenAI (advised while working at BSR)
- **Eve Nuria Fleisig**, Postdoctoral Fellow, Princeton Center for Information Technology Policy
- **Juan David Gutiérrez Rodríguez**, Profesor asociado, Director de la Maestría en Derecho, Tecnología y Sociedad Digital
- **Ekaterina Kochmar**, Associate Professor, University of Victoria (advised while working at the Mohamed bin Zayed University of Artificial Intelligence, Natural Language Processing Department)
- **Alexa Koenig**, Faculty Director of UC Berkeley Human Rights Center; Director of HRC's Investigations Lab; Research Professor of Law at Berkeley Law
- **Lucy Li**, Assistant Professor, University of Wisconsin, Madison (advised while working on PhD at UC Berkeley)
- **Admire Mare**, Professor, Head of Department, Department of Communications and Media, Faculty of Humanities, University of Johannesburg

A warm thank you to the 56 individuals from around the world who agreed to be interviewed to inform our understanding of how LLMs are used in different professional contexts in different regions and countries as well as our analysis of the human rights risks and opportunities. A special thank you to:

- **Michael Agyemang Adarkwah**, Research Associate, Institute for Education and Culture, Chair of Adult Education, Friedrich Schiller University Jena
- **Idrees Ahmad**, Director of Journalism, Literature, Film, and Theatre Studies, University of Essex
- **Anchali Anandanayagam**, Partner, Hudson Gavin Martin
- **Carolina Botero**, Miembro de la Junta Directiva, Fundación Karisma
- **Charlie Beckett**, Professor in the Department of Media and Communications, London School of Economics

- **Garance Burke**, Global Investigative Journalist, The Associated Press
- **Flynn Coleman**, Visiting Scholar and Lecturer, Columbia University
- **Michael Davis**, Research Fellow, University of Technology Sydney
- **Giulia Gentile**, Lecturer in Law, Essex Law School
- **Pawel Gmyrek**, Senior Researcher, Research, International Labour Organization
- **Lauren Goodlad**, Distinguished Professor, English and Comparative Literature, Chair, Critical AI, Rutgers University
- **Simone Grassini**, Associate Professor, Department of Psychosocial Science, University of Bergen, Cognitive and Behavioral Neuroscience Lab, University of Stavanger
- **Neel Guha**, JD Candidate, Stanford Law School, PhD Candidate, Stanford Computer Science
- **Juan David Gutiérrez Rodríguez**, Profesor asociado, Director de la Maestría en Derecho, Tecnología y Sociedad Digital
- **Ekaterina Kochmar**, Associate Professor, University of Victoria (advised while working at the Mohamed bin Zayed University of Artificial Intelligence, Natural Language Processing Department)
- **Ammina Kothari**, Dean, Gwen Ifill School of Media, Humanities and Social Sciences, Simmons University
- **Josh Lee Kok Thong**, Managing Director, Asia-Pacific, Future of Privacy Forum
- **Brenda Leong**, Director, AI Division, ZwillGen
- **Haley Lepp**, PhD Candidate, Education, Stanford University
- **Alvaro Liuzzi**, Digital Media Consultant, Journalist, and Educator
- **Michael Legg**, Professor, Director, Centre for the Future of the Legal Profession, University of New South Wales Law & Justice
- **Nicolás Lozada Pimiento**, Docente, Departamento de Derecho de los Negocios, Universidad Externado de Colombia
- **Admire Mare**, Professor, Head of Department, Department of Communications and Media, Faculty of Humanities, University of Johannesburg
- **Christopher Musodza**, Managing Consultant, CNM Technologies
- **Tarek Nakkach**, General Counsel Middle East, Africa & Turkey, Kyndryl
- **Ridwan Oloyede**, Co-Founder, Technology Policy Lead, Tech Hive Advisory
- **Melanie Tarr**, Teacher Digital Technologies, Apple Teacher with Swift Playgrounds Recognition, Aveley Secondary College
- **Rajesh Sreenivasan**, Head, Technology, Media & Telecommunications, Rajah & Tann Singapore LLP
- **Kalliopi Terzidou**, PhD in Law, University of Luxembourg
- **Elodie Vialle**, Journalist, Press Freedom Defender
- **Carlos Gerardo Zermeño Vargas**, Profesor, Coordinador Nacional de Comunidades Académicas, Escuela de Humanidades y Educación, Campus Toluca, Tecnológico de Monterrey

A special thanks to **Alexander Kurt Preve**, Managing Associate at Dentons US LLP, for conducting a pro bono review for this report of AI regulations and frameworks in countries around the world (work completed while at Coblentz Patch Duffy & Bass LLP).

An additional thank you to representatives from **Anthropic**, **Google**, **Meta**, and **OpenAI** who helped us understand steps the companies take in the development and deployment stages to prevent, mitigate, and remediate the human rights risks of their LLMs.

APPENDIX

A. Interview Format

We interviewed four sets of stakeholders: AI experts, AI practitioners, field-specific experts, and field-specific practitioners. For each, we conducted semi-structured interviews with a different set of guiding questions.

AI experts

- What are the human rights risks of using LLMs?
  - » How do these differ across fields, such as journalism, education, and law?
  - » Do risks vary by geographic location?
  - » Are there inequities created along demographic lines (race, gender, social class, disabilities)?
  - » What makes the risks of LLMs unique from the risks of other types of AI?
- What are the human rights opportunities of using LLMs?
  - » How do these differ across fields, such as journalism, education, and law?
- How are LLMs impacting the workforce?
  - » How might there be changes to skill requirements as a result of LLMs?
  - » Might industries introduce new positions or eliminate existing positions in response to LLMs?
  - » How are LLMs impacting the valuation of labor across industries?
  - » How do you think the additional functionality of multimodal models and internet access affect uses in the workforce?
- What can companies that make LLMs do to minimize the harmful impacts and maximize the beneficial impacts of LLMs on human rights?
  - » How should companies test models before releasing them?
- How do you think LLMs should be regulated?
- What harm can be caused by malicious actors, and how can this be mitigated?

AI practitioners

- How, if at all, does your company incorporate responsible AI principles (safety, ethics, etc.) into the LLM product development pipeline?
  - » Are human rights embedded in your company's responsible AI principles?
- How has your company assessed the risks of using your LLM *before* and *since* releasing it to the public?
  - » Did these include human rights risks?
- What risks did you identify?
  - » Did they vary by geographic location?
- What are some common strategies you use to mitigate these risks? (e.g. content filters, preference tuning)
- What are the biggest challenges of risk assessment for individual LLM products?
  - » Examples?
  - » How did you work through them?
- How does your company communicate risks to users?
- If more resources were available, what more do you think your company could do to both assess and communicate the risks of using your LLM?
- How has your company assessed the opportunities of using your LLM before and since releasing it to the public?
  - » Did these include human rights opportunities?
- What opportunities did you identify?
  - » Did they vary by geographic location?
- What kind of data does your company collect on the way that people use your LLM?
- What steps does your company take to be transparent about the internal workings of your LLM?
  - » What barriers, if any, do you face to increasing transparency further?
- How do you think governments should regulate LLMs?

Field-specific experts

- In your experience, how aware are [field] professionals about LLMs?
  - » Does awareness of [field] professionals vary based on geographic location?
  - » Are [field] workplaces providing training to [field] professionals on LLMs?
- Do [field] professionals use LLMs for professional purposes?
  - » If so, how and how much?
  - » Does the use of LLMs by [field] professionals vary based on geographic location?
  - » Does the use of LLMs by [field] professionals vary based on their job description or subfield?
  - » How does the use of LLMs differ in [field] and other professional fields?
  - » How has the culture of [field] affected its adoption of new technologies like LLMs?
- What are the human rights risks of using LLMs in [field]?

- What are the human rights opportunities of using LLMs in [field]?
- How has the introduction of LLMs affected the employment of [field] professionals?
  - » Have there been changes to the skill requirements of [field] professionals as a result of LLMs?
  - » Are [field] workplaces and other [field] entities introducing new positions or eliminating existing positions in response to LLMs?
  - » How are LLMs impacting the valuation of the labor of [field] professionals?
- How do you think the use of LLMs will impact [field] in the next five years?
- What can companies that make LLMs do to minimize the harmful impacts and maximize the beneficial impacts of LLMs on human rights when used in [field]?

Field-specific practitioners

- On a scale of 1 to 10, how much would you say you know about LLMs?
  - » Do you know what they are?
  - » Do you know how they work?
- Do you use LLMs for professional purposes?
  - » If so, how often?
  - » For what tasks?
  - » How long have you been using LLMs for?
- Are your colleagues/professionals around you using LLMs for professional purposes?
  - » If so, how often?
  - » For what tasks?
  - » How long have they been using LLMs for?
  - » Does the use of LLMs vary based on their job description or subfield?
- What is your workplace's attitude towards LLMs?
  - » Is it encouraging/discouraging the use of LLMs?
  - » Are there any limitations/prohibitions in place?
  - » Is your workplace providing training on how to use LLMs?
- Have you seen your workplace and professional field change as a result of LLMs?
  - » Has it affected employment?
  - » Has it changed the tasks that are being performed?
  - » Do you anticipate it will in the next 5 years?
- What are the opportunities of using LLMs in your workplace?
- What are the risks of using LLMs in your workplace?
- How could LLMs be improved to help you better in your daily tasks?

B. Glossary

Agents (AI): Autonomous or semi-autonomous software systems that use natural language processing to perform tasks, such as answering questions or engaging in dialogue, often simulating human-like interaction or decision-making.

Algorithmic audits (AI): Systematic examination of the behavior or outputs of an algorithm with respect to certain inputs.

Alignment (AI): The practice of ensuring that model outputs align with the values, intentions, and preferences of humans. One technique is preference tuning, which fine-tunes (see below) a model to reflect data on human preferences or feedback.

Anthropomorphism: The tendency to attribute human characteristics, emotions, or intentions to non-human entities, including AI systems.

Application Programming Interface (API): A set of rules and protocols that allows different software applications to communicate. In the context of LLMs, APIs enable developers to integrate model outputs into other services or platforms.

Axial coding (Sociology): A task specified within grounded theory (see below) to systematically identify relationships between open codes.

Black box: A system whose internal workings are opaque or not easily interpretable. In LLMs, this refers to models where it is difficult to understand how specific inputs lead to specific outputs, complicating accountability and transparency.

Bug bashing: An informal process of identifying and fixing software bugs, often through focused, collaborative testing sessions. In AI contexts, it can involve stress-testing models for flaws or failures.

Children's rights: Refer to the specific human rights afforded to all individuals under the age of 18, recognizing their particular need for protection and development. These rights include the right to life, education, protection from abuse and exploitation, and participation in decisions affecting them. They are enshrined in the Convention on the Rights of the Child as well as other human rights instruments.

Civic participation: Refers to the right of individuals to take part in the conduct of public affairs, directly or through freely chosen representatives. It encompasses the rights to vote, run for public office, access public information, engage in peaceful assembly, and express opinions freely. It is enshrined in Article 25 of the *International Covenant on Civil and Political Rights* (ICCPR).

Common Crawl: A massive publicly available dataset consisting of web-crawled content from the internet. It is widely used to train LLMs.

Emergent capabilities: The abilities exhibited by general-purpose AI systems that were not explicitly programmed or anticipated during their training.

Fine-tuning (AI): A process by which a pretrained language model is further trained on a specific dataset to specialize it for a particular domain or task, such as legal reasoning or medical advice.

Freedom of thought: A core, non-derogable human right that protects an individual's mind from external interference. It encompasses the right to hold thoughts privately, to resist coercion or manipulation of one's mental processes, and to be free from punishment based solely on one's thoughts. Recognized under Article 18 of the International Covenant on Civil and Political Rights (ICCPR), this right has been affirmed as absolute and inviolable. See Alegre, Susie. 2022. *Freedom to Think: The Long Struggle to Liberate Our Minds*. London: Atlantic Books.

Grounded theory (Sociology): A qualitative research method to systematically collect and code data to build up to a theory.

Hallucinations (AI): Factually incorrect or unfaithful model outputs that are presented as plausible or true.

Human rights harm: When the rights and freedoms specified in human rights documents are infringed through action or inaction.

Human rights risk: The potential for an activity, decision, or tool to negatively impact the rights and freedoms of individuals or communities. A human rights risk is the possibility of a human rights harm occurring.

Instruction tuning (AI): A model post-training technique where models are optimized to better follow input instructions, improving usability.

International Covenant on Civil and Political Rights (ICCPR): A 1966 UN treaty articulating civil and political rights, such as the right to due process and fair trial and the right to access to information (see below).

International Covenant on Economic, Social and Cultural Rights (ICESCR): A 1966 UN treaty articulating economic, social, and cultural rights, such as the rights to education, health, housing, and social security (see below).

Labor rights: Affirm the freedom of association and effective recognition of the right to collective bargaining; the elimination of all forms of forced or compulsory labor; the effective abolition of child labor; the elimination of discrimination in respect of employment and occupation; and a safe and healthy working environment. See International Labour Organization. *ILO Declaration on Fundamental Principles and Rights at Work, as Amended in 2022*. Geneva: ILO, 2022. https://www.ilo.org/declaration/lang--en/index.htm.

Large language models (LLMs): Complex AI systems trained on vast amounts of language data to generate and predict human language.

Log probabilities: Numerical values representing the likelihood that a model assigns to specific outputs. These are used internally by LLMs to select the most probable next word or phrase and can be exposed externally to assist with interpreting model behavior.

Natural language processing (NLP): A subfield of computer science and linguistics focused on enabling machines to interpret and generate human language.

Neural networks: Mathematical models inspired by the structure of the human brain, composed of layers of interconnected nodes (neurons), and foundational to modern AI, including LLMs.

Open coding (Sociology): A task specified within grounded theory (see above) to label raw qualitative data with discrete concepts.

Pretraining (AI): The initial training phase where a language model learns language structure, facts, and reasoning patterns by processing massive amounts of general text data. It builds the model's foundational capabilities before any fine-tuning for specific tasks.

Protection from arbitrary detention: A fundamental human right that prohibits the arrest or detention of individuals without legal justification, due process, or judicial oversight. It is enshrined in Article 9 of the International Covenant on Civil and Political Rights.

Retrieval-augmented generation (RAG): A technique where an LLM retrieves relevant documents or information from external sources before generating an answer, aiming to improve factual accuracy and reduce hallucinations.

Red teaming: A testing process where experts adversarially challenge an AI system to uncover safety risks, biases, or harmful behaviors, particularly in high-stakes scenarios.

Representational harms: When content reinforces or amplifies harmful stereotypes, biases, or misrepresentations of people and groups.

Right of access to information: Guarantees that individuals can seek, receive, and impart information held by public authorities. It is essential for transparency, accountability, and meaningful participation in public life. It is enshrined in Article 19 of the Universal Declaration of Human Rights (UDHR) and in Article 19 of the International Covenant on Civil and Political Rights (ICCPR).

Right to asylum: The right to seek and enjoy asylum is a fundamental principle of international human rights law, recognizing that individuals fleeing persecution have the right to seek protection in another country. While not guaranteeing the right to be granted asylum, it ensures the right to apply for it without being penalized for illegal entry. It is enshrined in Article 14 of the Universal Declaration of Human Rights.

Right to due process and fair trial: Guarantees that every individual is entitled to a fair and public hearing by an independent and impartial tribunal in the determination of their rights and obligations. It is enshrined in Article 10 of the Universal Declaration of Human Rights. It is enshrined in Article 19 of the International Covenant on Civil and Political Rights (ICCPR).

Right to education: A cornerstone of international human rights law, recognized as essential for the full development of the human personality and the promotion of understanding, tolerance, and peace. Article 13 of the International Covenant on Economic, Social and Cultural Rights (ICESCR) affirms that education shall be directed to the full development of the human personality and dignity and shall enable all persons to participate effectively in society. It obliges states to ensure free and compulsory primary education, accessible secondary education, and a long-term commitment to gradually achieve free higher education. This right is also articulated in Article 26 of the Universal Declaration of Human Rights.

Rights to equality and non-discrimination: These rights guarantee that all individuals are treated equally before the law and protected from unjust differential treatment based on characteristics such as race, gender, religion, ethnicity, disability, or other protected status. They oblige states to prevent, prohibit, and remedy discrimination, both direct and indirect. These principles are enshrined in Articles 1 and 2 of the Universal Declaration of Human Rights, as well as in Articles 2, 24, and 26 of the International Covenant on Civil and Political Rights (ICCPR).

Right to freedom of movement: Guarantees that everyone lawfully within a state's territory has the right to move freely, choose their residence, and leave the country, as well as the right to return to one's own country. This right is articulated in Article 13 of the Universal Declaration of Human Rights and is enshrined in Article 12 of the ICCPR.

Right to health: Refers to the entitlement of every individual to the highest attainable standard of physical and mental well-being. It includes access to timely, acceptable, and affordable health care, as well as to underlying determinants of health such as clean water, adequate nutrition, housing, and a healthy environment. It is reflected in Article 25(1) of the Universal Declaration of Human Rights and enshrined in Article 12 of the International Covenant on Economic, Social and Cultural Rights (ICESCR).

Rights to life and security: Foundational principles of international human rights law. The right to life and the right to security of person are proclaimed in Article 3 of the Universal Declaration of Human Rights. The right to life, protected under Article 6 of the International Covenant on Civil and Political Rights (ICCPR), ensures that no one shall be arbitrarily deprived of life. The right to security of person, enshrined in Article 9 of the International Covenant on Civil and Political Rights (ICCPR), guarantees protection from unlawful violence, arbitrary arrest, and detention. Together, these rights uphold the fundamental dignity and physical integrity of every individual.

Right to housing: The guarantee that every individual has access to adequate, safe, and secure shelter that enables a life of dignity. It includes legal protection against forced eviction, access to basic services, affordability, habitability, and location. It is enshrined in Article 11 of the International Covenant on Economic, Social and Cultural Rights (ICESCR).

Right to nationality: Guarantees that every individual is entitled to legal recognition as a national of a state. It protects against statelessness and affirms that no one shall be arbitrarily deprived of their nationality or denied the right to change it. It is enshrined in Article 15 of the Universal Declaration of Human Rights.

Right to social security: Ensures that everyone has access to support systems that protect against economic and social risks such as unemployment, old age, disability, and illness. It includes the right to access social insurance and assistance necessary for an adequate standard of living. It is articulated in Article 22 of the Universal Declaration of Human Rights and enshrined in Article 9 of the International Covenant on Economic, Social and Cultural Rights.

Right to remedy: The entitlement of individuals to seek and obtain redress when their rights are violated, including access to effective judicial or administrative procedures. Article 8 of the Universal Declaration of Human Rights affirms the right to an effective remedy by competent national tribunals for acts violating fundamental rights granted by law. It is further enshrined in Article 2(3) of the ICCPR, which obliges states to ensure that individuals have accessible, effective remedies when their rights are infringed.

Parameters: Adjustable numerical values in a model that are learned during training. In LLMs, the number of parameters (often in billions) affects the model's capacity and complexity.

Salience: Refers to the significance or prominence of a human rights risk. In this context, the most salient risks are those most likely to cause serious negative impacts on individuals or communities, either because of their scope, scale, likelihood, or lack of remediability.

- **Scale:** How serious is the impact?
- **Scope:** How widespread is the impact on individuals and/or communities?
- **Remediability:** To what degree is it possible to restore the situation or make amends to those who have suffered a human rights harm?

- **Likelihood:** How likely is the impact?

Society level: Refers to the LLMs outputs' impact on communities and societal norms (e.g. employment patterns).

System (AI): The full setup through which an AI model operates, such as the model, prompt, user interface, retrieval tools, and any surrounding policies or governance layers.

System level: Refers to how LLM outputs interact with the institutions, applications, and processes in which they are embedded to create some material effect (e.g., hiring decisions).

Task (AI): A specific objective or function that an NLP system is designed to perform, such as translation, summarization, question answering, or sentiment analysis.

Transformers (AI): A neural network architecture that underpins most modern LLMs. It uses attention mechanisms to process text in parallel, allowing for greater efficiency and context awareness.

Universal Declaration of Human Rights (UDHR): A 1948 UN treaty articulating 30 fundamental human rights, including the rights to asylum and nationality (see above).

HUMAN RIGHTS CENTER
UC BERKELEY SCHOOL OF LAW
2224 PIEDMONT AVENUE
BERKELEY, CA 94720
510.642.0965

HRC@BERKELEY.EDU
HUMANRIGHTS.BERKELEY.EDU