



PARTNERSHIP ON AI

# 2026 Transparency Report on Foundation Model Impacts

**A Progress Report on Post-Deployment  
Governance Practices**

Jacob Pratt  
Albert Tanjaya

This report is an update  
to PAI's 2025 report,  
"[Documenting the Impacts  
of Foundation Models](#)"



# Contents

<b>Executive Summary</b>	<b>3</b>
Progress analysis	4
Key Insights	5
<b>Introduction</b>	<b>7</b>
Reaffirming the value of public post-deployment impact documentation	7
Four categories of information, four practices	8
Updates to how we evaluate progress for 2026	10
<b>Has there been progress in how providers document the impact of foundation models?</b>	<b>11</b>
Practice 1: Share usage information	11
Practice 2: Enable and share research on post-deployment societal impact indicators	14
Practice 3: Report incidents and policy violations	19
Practice 4: Share user feedback	23
<b>Key Insights</b>	<b>25</b>
<b>Conclusion</b>	<b>28</b>
<b>Appendices</b>	
1. Summary of Practices	29
2. Methodology	30
<b>Acknowledgements</b>	<b>34</b>
<b>Works Cited</b>	<b>35</b>

# Executive Summary

Realizing the benefits of AI requires an ecosystem of trust and accountability with transparency as its foundation. [Academics](#) and [governments](#) recognize the role of transparency, and at the India AI Impact Summit, frontier AI developers [committed](#) to sharing real-world AI usage information with governments. However, voluntary commitments, recent [legislation](#), and [voluntary codes](#) are increasingly focused on disclosing information behind closed doors. As foundation models shape our society, private disclosures alone cannot foster the trust and accountability that public transparency enables.






This report measures the progress of 13 organizations in publicly documenting the impacts of their foundation models. Building on our 2023 [Guidance for Safe Foundation Model Deployment](#) and 2025 [progress-tracking methodology](#), this report explores how well practice leaders are adopting four practices that encapsulate 19 processes for public impact reporting, and how the rest of the field is keeping pace. This analysis is based on the review of more than 150 papers, articles, websites, and reports.

## OVERALL FINDING 1

**Several leading organizations are defining what information to share and how.** However, the rest of the field often lags in adopting information-sharing practices, raising questions about how to ensure that society benefits from whole-of-field impact information.

## OVERALL FINDING 2

**Our analysis reveals a scattered and incomplete landscape of public impact data: we know more about usage and potential labor market impacts than ever, but we don't know the prevalence of harms, serious incidents, or broader societal impacts.** We see progress in how organizations share usage information, but significant gaps remain in reporting incidents, policy violations, and broader societal impact indicators. Companies participate inconsistently and scatter information across blog posts, earnings calls, and research papers, in line with [recent PAI research](#) on the formal reporting practices of 50 major companies. This shows that the field hasn't overcome the five core challenges listed in our [2025 report](#):

-  a lack of standardized documentation norms
-  barriers to data sharing and coordination
-  misaligned incentives
-  limited information-sharing infrastructure
-  the decentralized nature of open model deployment

## Progress analysis

The foundation of this work is an analysis of more than 150 publicly available sources of information shared by [13 foundation model providers](#). Focusing on the value this information provides other actors, we evaluated the progress of organizations pioneering information sharing in each practice, whom we call “practice leaders,” and how the rest of the field is keeping pace.

PRACTICE	PROGRESS MADE BY:	
	PRACTICE LEADERS	THE REST OF THE FIELD
1 Share usage information	↑ SIGNIFICANT	– LIMITED
2 Enable and share research on post-deployment societal impact indicators	↗ MODERATE	↗ MODERATE
3 Report incidents and policy violations	↗ MODERATE	– LIMITED
4 Share user feedback	↗ MODERATE	↗ MODERATE

The four practices are broken down into 19 processes – activities that support how foundation model providers adopt practices – and assessed using an evaluation rubric. The full methodology is in [Appendix 2](#).



### PRACTICE 1

#### Share usage information

PROCESS	2025 LEVEL OF ADOPTION	2026 DEGREE OF PROGRESS
1.1 Conduct surveys or user research to understand downstream usage	N/A	N/A
1.2 Create tools to support the sharing of activity logs with trusted third parties for analysis	NONE	↗ MODERATE
1.3 Implement and track watermarking or identifiers	NONE	– LIMITED
1.4 Report aggregate usage statistics, across geographies, sectors, or use cases, including usage in high-risk use cases	LOW	↑ SIGNIFICANT
1.5 Share information on downstream applications of the model	LOW	– LIMITED



### PRACTICE 2

#### Enable and share research on post-deployment societal impact indicators

PROCESS	2025 LEVEL OF ADOPTION	2026 DEGREE OF PROGRESS
2.1 Report on labor impact indicators	LOW	↑ SIGNIFICANT
2.2 Report on environmental impact indicators	LOW	↗ MODERATE
2.3 Report on synthetic content impact indicators	LOW	↗ MODERATE
2.4 Disclose third-party research access	MEDIUM	↗ MODERATE
2.5 Disclose organizational resourcing commitments and dedicate funding commitments towards post-deployment societal impacts	HIGH	↗ MODERATE



### PRACTICE 3

## Report incidents and disclose policy violations

PROCESS	2025 LEVEL OF ADOPTION	2026 DEGREE OF PROGRESS
3.1 Monitor for incidents	MEDIUM	↗ MODERATE
3.2 Monitor for policy violations	MEDIUM	↗ MODERATE
3.3A Share summaries of internal incident reports	LOW	↗ MODERATE
3.3B Share summaries of internal policy violation reports	LOW	– LIMITED
3.4 Systematically report AI incidents to a third party	NONE	– LIMITED



### PRACTICE 4






## Share user feedback

PROCESS	2025 LEVEL OF ADOPTION	2026 DEGREE OF PROGRESS
4.1 Disclose the process of having a feedback mechanism for stakeholders	HIGH	↗ MODERATE
4.2 Aggregate individual user feedback records to have as summaries	MEDIUM	– LIMITED
4.3 Disclose the feedback follow-up process or, if warranted, the redress mechanism process	LOW	– LIMITED
4.4 Create incentive structures to invite stakeholders to participate in the feedback process proactively	LOW	↗ MODERATE

## Key Insights

Analyzing the underlying data provided the following key insights. Discussion of the [data](#) (p. 11–24) and further discussion of each [insight](#) (p. 25–27) is provided later in the report.

INSIGHT	RELATED TO PRACTICE
<b>INSIGHT 1</b> Foundation model providers that host their own services can track why and how people use foundation models without compromising privacy. This can inform public policy.	1. SHARE USAGE INFORMATION
<b>INSIGHT 2</b> Providers are signaling risks to entry-level workers and changing workforce skill requirements, but have not published research on risks related to worker abuse and sourcing.	1. SHARE USAGE INFORMATION 2. SHARE SOCIETAL RESEARCH (LABOR)
<b>INSIGHT 3</b> Providers report environmental impacts differently, so standards are crucial.	2. SHARE SOCIETAL RESEARCH (ENVIRONMENTAL)
<b>INSIGHT 4</b> Providers report AI-generated or uploaded Child Sexual Abuse Material (CSAM), but the prevalence of other foundation model misuses remains unknown.	2. SHARE SOCIETAL RESEARCH (SYNTHETIC MEDIA) 3. REPORT INCIDENTS & VIOLATIONS

<b>INSIGHT 5</b>	Providers report some security incidents voluntarily – both privately and publicly, where appropriate. This accelerates threat detection and attribution, but overlooks broader harms to users and society.	 <a href="#">3. REPORT INCIDENTS &amp; VIOLATIONS</a>
<b>INSIGHT 6</b>	Providers prioritize generating feedback on security risks over broader safety concerns.	 <a href="#">4. SHARE USER FEEDBACK</a>
<b>INSIGHT 7</b>	Scientific and policy guidance for open model impact reporting faces significant implementation challenges that reduce actual reporting.	 <a href="#">1. SHARE USAGE INFORMATION</a>  <a href="#">3. REPORT INCIDENTS &amp; VIOLATIONS</a>  <a href="#">4. SHARE USER FEEDBACK</a>

By publishing progress in reporting practices on a yearly basis, we aim to encourage good practices from leading providers, realign incentives, promote accountability, and catalyze a “race to the top.” However, achieving consistency requires action beyond any single organization, so we will work with our partner community to review and formalize our recommendations to overcome these challenges.

This report is one part of PAI’s work to develop trust in the AI value chain. It also supports our work on [Strengthening the AI Assurance Ecosystem](#) by analyzing information disclosure, a key aspect of transparency, which itself enables assurance. It also complements our [research](#) on the disclosure of AI-related impacts, risks, and opportunities, as well as our Enterprise Steering Committee’s [documentation work](#) by analyzing disclosures outside formal channels and documentation created for purposes beyond managing impacts and risks across the value chain.

To discuss this work, please contact [jacob@partnershiponai.org](mailto:jacob@partnershiponai.org).

# Introduction

There is widespread agreement on the need for transparency. Transparency has been central to AI governance since the field's inception, and was codified in the [OECD AI Principles](#) in 2019. Since then, [academics](#), [governments](#), and [industry organizations](#) have repeatedly emphasized the importance of sharing information about AI systems and their impacts: it enables accountability, helps customers make informed choices, allows developers to build safer downstream products, and empowers policymakers to govern effectively. This contributes to a world where businesses and other entities can innovate and deploy trustworthy AI systems with the [assurance](#) that they are helping our society rather than harming it.

Despite this consensus, public transparency is taking a back seat to partial, government-only disclosures. Concerns that public transparency — publishing select information openly for anyone to access — could expose commercial trade secrets or reveal security vulnerabilities have led [recent legislation](#), [voluntary codes](#), and frontier AI developer [commitments](#) to prioritize partial disclosures to governments and downstream partners. While these considerations are important, this shift away from sharing information publicly risks leaving society in the dark about how these systems are transforming our lives, threatening public confidence, and undermining innovation.

With AI embedded in our lives, public transparency is more critical than ever. AI investments are [booming](#), engineers self-report that 26.9% of new code [is now AI-authored](#),<sup>1</sup> and companies are rapidly deploying agents that can [fail](#) in consequential ways. Foundation models are driving this agentic-driven change, but we are still in the dark about many of their impacts. As regulators scale up their monitoring and enforcement capabilities, the public must play a vital role in holding companies accountable for the ways these models transform our lives. This requires public access to information about the impacts of foundation models.

<sup>1</sup> AI-authored code is defined as “code merged upstream, or in a customer-facing environment that is written by AI and merged without significant human intervention.”  
Note on data: n=42,644.  
Timeframe is Nov 1, 2025 - Feb 1, 2026.

Last year, we measured how widely foundation model providers documented their models' impacts. Now we are updating our work. Point-in-time analyses provide useful insights, but tracking progress over time is crucial for understanding how challenges, information gaps, and reporting practices evolve. These data point us toward how the ecosystem may develop, and inform how we might shape it for the better.

## Reaffirming the value of public post-deployment impact documentation

In 2023, PAI worked with its expert community across the AI ecosystem to create its [Guidance for Safe Foundation Model Deployment](#), a framework for responsible development and deployment. Monitoring the impact of models after release or deployment<sup>2</sup> is a core part of this framework, which we further explored in the first progress report, [Documenting the Impacts of Foundation Models](#).

<sup>2</sup> Releasing a model refers to making its weights or code available to download and run. Deploying a model refers to making an instance of the model available for use through an interface, such as an API or product.

This 2025 report identified how different deployment configurations, such as integrated services and building on open source models, have different collaboration requirements for collecting, collating, and sharing information about the impacts of foundation models. At the same time, a recent study of 272 experts suggested that the foundation model providers' duty to publish information about the impacts their models have on society remains consistent.<sup>3</sup> This responsibility sits with their core role in shaping societal impact through design, development, and distribution.

### Public impact reporting as one component of transparency

We recognize that it is not appropriate to share all information publicly, and that poorly executed transparency can divert resources from more valuable work, raise privacy and competition concerns, and create perverse incentives to avoid identifying issues in the first place.<sup>3</sup> We also acknowledge the value of sharing some information only with specific actors, such as government or trusted researchers, or of sharing anonymously for discussion in private forums.<sup>4</sup> Our process definitions account for many of these valuable activities, and we assess progress against them in this report.

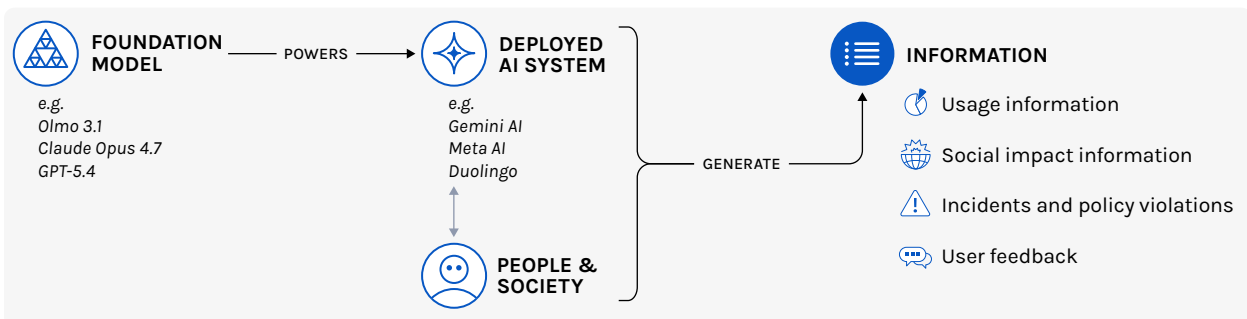
This is why our judgments on good reporting are not based on the expectation that all information should be accessible to everyone: for example, certain security vulnerability incident reports may require restricted reporting channels with strict confidentiality protections, and detailed usage telemetry requires aggregation or anonymization to protect privacy. However, this report argues sharing information publicly is a key part of building a governance ecosystem founded on trust and accountability.

<sup>3</sup> "What Should Companies Share about Risks from Frontier AI Models?" METR, 27 June 2025.

<sup>4</sup> Kolt et al. describe the value of differential disclosure and anonymized reporting for mitigating frontier AI risks in "Responsible Reporting for Frontier AI Development"

## Four categories of information, four practices

The broad scope of impact information reflects society's complexity. Our 2025 report focused on four broad categories of information generated by integrating foundation models into systems used by people and society.



After finding that collecting and collating relevant information was a key barrier to publication, we identified 19 supporting processes<sup>5</sup> that facilitate four information-sharing practices. We summarize these practices and their supporting processes below.

<sup>5</sup> The practices and processes are described in more detail in the subsequent analysis and in the Appendix.



**PRACTICE 1**

**Share usage information**

DESCRIPTION	EXAMPLE
Document usage information by conducting user research, supporting trusted third-party usage analysis, implementing and tracking identifiers, and reporting aggregate usage statistics, including downstream applications.	Anthropic’s <a href="#">Economic Index</a>



**PRACTICE 2**

**Enable and share research on post-deployment societal impact indicators**

DESCRIPTION	EXAMPLE
Document labor, environmental, and synthetic content impact indicators, and support trusted third-party access and research.	The Allen Institute for AI’s <a href="#">Holistically Evaluating the Environmental Impact of Creating Language Models</a>



**PRACTICE 3**

**Report incidents and disclose policy violations**

DESCRIPTION	EXAMPLE
Document incidents of harm and policy violations by conducting monitoring, publishing internal summary reports, and systematically reporting incidents through the appropriate channels.	Google’s <a href="#">GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools</a>



**PRACTICE 4**

**Share user feedback**

DESCRIPTION	EXAMPLE
Document user feedback by disclosing collection mechanisms, sharing aggregated feedback analysis, and incentivizing proactive participation in security and safety feedback.	OpenAI’s <a href="#">Safety Bug Bounty program</a>

## Updates to how we evaluate progress for 2026

In [2025](#), we measured the progress among foundation model providers in adopting information-sharing practices and found mixed results. Coordinating and sharing between organizations was challenging, standardized reporting was rare, and foundation model providers had few incentives to share information publicly. However, systems powered by foundation models have become more widespread over the last year, and understanding their impacts has become increasingly important.

For this 2026 evaluation, we broadened our focus to include how well practice leaders are implementing information-sharing practices. As a proxy for quality, we assessed whether the information shared by foundation model providers was sufficient for other actors to amplify the societal benefits from these models, manage and mitigate risks, and develop evidence-based, proportionate policy.<sup>6</sup> We used over 150 public information sources to rate the level of progress across the four practice areas, establishing a scale from minimal to significant. We hope that this sets the stage for more readers to look beyond superficial reporting to the actual documentation from foundation model providers.

<sup>6</sup> These benefits were described in our previous report.

The rest of the report focuses on how well organizations document the impacts of foundation models. For each post-deployment documentation practice, we answer two questions:

- **What might good impact reporting look like?**

We describe how companies can best conduct impact reporting by comparing their activities against our [Guidance for Safe Foundation Model Deployment](#), our [2025 report](#), and the 150 examples analyzed in this work.

- **Has there been progress for 2026?**

This year, we updated our methodology for assessing the progress of practice leaders and the rest of the field for all four practices, and to account for how well specific organizations document their impacts.<sup>7</sup> We also updated our approach to assessing the level of progress for each of the 19 processes, and share an overview of key reporting from organizations in this report.

<sup>7</sup> Our full methodology can be found in [Appendix 2](#).

# Has there been progress in how providers document the impact of foundation models?



## PRACTICE 1

### Share usage information

The use of foundation models continues to surge: 79% of businesses regularly [use generative AI](#),<sup>8</sup> and 92.6% of developers [use AI coding assistants](#).<sup>9</sup> As recognized in The India AI Impact Summit's 2026 [Frontier AI Commitments](#), effectively governing this expanding deployment will require actors to know where, how, and why these systems are used, yet this information remains limited to targeted surveys or remains within individual companies.

This practice focuses on documenting how downstream stakeholders use foundation models. Our research since 2023 highlights that foundation model providers can:

- Conduct user surveys across geographies, sectors, and use cases with documented methodologies and open data.
- Provide trusted third parties with structured, privacy-preserving access to usage datasets collected with user consent.
- Implement watermarks or identifiers for AI-generated content and track their usage online.
- Share aggregate usage statistics across geographies, sectors, and use cases through open data. Usage can include information on who uses the models, for what tasks and purposes (including high-risk use cases), and how they are being used.

<sup>8</sup> McKinsey reports that 33% of businesses adopted generative AI in at least one function in 2023, and that has increased at every reassessment, up to 79% of businesses regularly using generative AI in at least one function in 2025.

<sup>9</sup> Usage involves using an AI coding assistant at least once a month for work. n=121,502. Timeframe is Nov 1 2025 - Feb 1 2026.

#### PERSPECTIVES FROM OUR 2026 SURVEY

“Sharing internal telemetry externally (to select parties) can advance oversight (auditing) and policy over time.”

— ILAN STRAUSS, AI DISCLOSURES PROJECT

“Pre-deployment evaluations are becoming less reliable due to evaluation awareness of AI models, and diverse agentic scaffolding. Post-deployment evidence from staged release and usage data is becoming more important over time.”

— ANONYMOUS

## Has there been progress for 2026?

Our 2025 report highlighted major barriers to sharing usage data, such as the complexities of collaborating across the open model value chain and the privacy risks associated with sharing personal data. However, as hosted foundation model providers began to advance privacy-preserving analysis methods, we saw early examples of usage reporting that could be further developed.

For 2026, we see one organization improving transparency practices and inspiring other organizations to follow suit and innovate, though widespread adoption will require a much more concerted, ecosystem-wide effort.

Practice 1: Progress made since 2025		
MADE BY	DEGREE OF PROGRESS	
Practice leaders (Anthropic, Microsoft, OpenAI)	↑ SIGNIFICANT	
The rest of the field	– LIMITED	

Progress made in each process		
PROCESS	2025 LEVEL OF ADOPTION	2026 DEGREE OF PROGRESS
1.1 Conduct surveys or user research to understand downstream usage <sup>10</sup>	N/A	N/A
1.2 Create tools to support the sharing of activity logs with trusted third parties for analysis	NONE	↗ MODERATE
1.3 Implement and track watermarking or identifiers	NONE	– LIMITED
1.4 Report aggregate usage statistics, across geographies, sectors, or use cases, including usage in high-risk use cases	LOW	↑ SIGNIFICANT
1.5 Share information on downstream applications of the model	LOW	– LIMITED

<sup>10</sup> As in our last report, we state that model providers are not expected to play a key role in conducting surveys, so we do not measure the level of adoption and progress. However, we highlight it in this report to encourage adoption, and discuss how model providers have adopted the process in the discussion section

**Discussion**

Advances in privacy-preserving, large-scale dataset analysis have significantly expanded the scope and utility of the usage information shared by leading organizations.<sup>11</sup> Anthropic leads the field by sharing reams of aggregated data through its [Economic Index work](#),<sup>12</sup> analyzing interactions from 115 countries and providing an [interactive dashboard](#) to improve accessibility. OpenAI built on this methodology to [analyze](#) millions of messages to identify how users use ChatGPT, and Microsoft [analyzed](#) 200,000 messages from Microsoft Copilot to understand patterns in AI-assisted work. One-off and repeated sharing of this information can inform labor market analysis (see Practice 2) and policymaking.

<sup>11</sup> OpenAI state that the key elements of their privacy-preserving approach are “Automated classification of messages” – where analysis of user messages is done using automated classifiers on de-identified and PII-scrubbed usage data – and “Aggregated employment data via a data clean room.”

<sup>12</sup> A full list of relevant research from Anthropic can be seen in [Supplement 1](#).

**Table 1.** Policy-relevant information about AI usage identified in data published by [Anthropic](#), [Microsoft](#), and [OpenAI](#).

CATEGORY	DATA POINTS
<b>Overall usage</b>	<ul style="list-style-type: none"> <li>Total daily message volume</li> <li>Weekly active users (WAU)</li> <li>Messages per weekly active user</li> </ul>
<b>Who</b> uses foundation models	<ul style="list-style-type: none"> <li>Age</li> <li>Gender</li> <li>Country of registration</li> <li>Sector</li> <li>Work function</li> <li>Occupation</li> <li>Education level</li> <li>Other demographic information</li> </ul>
<b>What</b> are foundation models used to do	<ul style="list-style-type: none"> <li>Work vs. non-work</li> <li>Occupation classification (e.g. Computer and Mathematical)</li> <li>Task/work activity (e.g. debugging, read documents or materials to inform work processes)</li> <li>Complex/long tasks vs. simple/short</li> </ul>
<b>Why</b> are foundation models used	<ul style="list-style-type: none"> <li>Automation vs. augmentation</li> <li>User intent (asking, doing, expressing)</li> </ul>
<b>How</b> are foundation models being used	<ul style="list-style-type: none"> <li>API vs. browser</li> <li>Model selection</li> </ul>

**Tracking information on open model usage remains difficult**, as [Stability AI](#) recognizes, though simple analysis can be conducted using public data from open model hosting platforms, such as Hugging Face,<sup>13</sup> or in collaboration with inference providers, such as OpenRouter.<sup>14</sup> Tracking watermarks is also challenging, as best practices for labeling content are still being developed, presenting a significant barrier for broader usage tracking. Evidence of organizations developing tools to share activity logs with third parties for analysis beyond OpenAI’s work, detailed above, is limited.

**Academic institutions, government organizations, and increasingly model providers, share survey-based research that provides usage information.** For example, the US Census Bureau releases [data](#) on actual and planned AI usage by industry sector every two weeks, while Stanford HAI’s [AI Index](#) details AI and Generative AI usage by industry, function, region, and use case. Anthropic also developed a tool to survey Claude users, published [initial findings](#) from 1,250 users in 2025, and [updated](#) their findings in March 2026.

**Companies broadly report on usage in marketing materials and earnings calls, demonstrating that additional information can be compiled and shared when needed.** For example, in its Q3 earnings call, Google [reported](#) that “over 10 billion pieces of content have already been watermarked with SynthID” and [shared](#) that Gemini has 650 million monthly active users.<sup>15</sup>

<sup>13</sup> For example, one researcher analyzed the most downloaded models from Hugging Face.

<sup>14</sup> OpenRouter analyzed anonymized request-level metadata for billions of prompt-completion pairs as part of their [State of AI](#) report.

<sup>15</sup> Note that Google also highlighted that they had 400 million active users when responding to the [Foundation Model Transparency Index](#), which was published in December 2025, so numbers can vary.



## PRACTICE 2

### Enable and share research on post-deployment societal impact indicators

Societal and individual well-being depend on a strong labor market, a healthy environment, and trust in information, but warnings about [AI therapists](#) and [reductions in employment](#) suggest these may be at risk. Reporting against impact indicators can ensure societal resilience, and an Ada Lovelace Institute [survey](#) found that 85% of people in the UK believe AI companies should publicly disclose the societal and environmental costs of their AI systems. However, tracking these impacts with precision is a fraught and complex challenge that will require a variety of indicators to capture what matters in a vibrant, diverse, and healthy society.

Practice 2 focuses on sharing and analyzing measurable indicators of impact within our society.<sup>16</sup> Our research since 2023 highlights that foundation model providers can:

- Report against labor impact indicators, including new evaluations of economic impacts, usage data for organizational productivity metrics, and usage data for impacted workforce skills and sectors.
- Report against environmental impact indicators, including independently audited usage of energy, water, compute, and greenhouse gas emission estimates.
- Report against synthetic content impact indicators, including emerging technical research on synthetic media authentication methodologies, CSAM reporting metrics, UX/UI improvement reporting related to trust and labeling, and other emerging risks and harms enabled by synthetic content.
- Provide access to models and usage data for research purposes, with dedicated, ongoing funding programs for societal, economic, and environmental impact research.

<sup>16</sup> We recognize that while direct attribution of impacts to specific models may not be possible, tracking key indicators can help understand emerging patterns and potential effects.

#### PERSPECTIVES FROM OUR 2026 SURVEY

“It is important to understand the large-scale benefits and risks that emerge only when the technology is widely deployed and its cumulative effects are felt, in addition to warning shots of individual instances of misuse or failure.”

— ANONYMOUS

### Has there been progress for 2026?

Our previous report highlighted that foundation model providers reported limited information about environmental, economic, and other societal indicators that reflect impacts, though many had disclosed resourcing and funding committed to researching post-deployment societal impacts.

For 2026, we see uneven progress among practice leaders, with no organization tracking impacts across all indicators.

## Practice 2: Progress made since 2025

MADE BY	DEGREE OF PROGRESS
Practice leaders ( <i>Allen Institute for AI, Anthropic, Google, Meta, Microsoft</i> )	↗ MODERATE
The rest of the field	↗ MODERATE

### Progress made in each process

PROCESS	2025 LEVEL OF ADOPTION	2026 DEGREE OF PROGRESS
2.1 Report on labor impact indicators	LOW	↑ SIGNIFICANT
2.2 Report on environmental impact indicators	LOW	↗ MODERATE
2.3 Report on synthetic content impact indicators	LOW	↗ MODERATE
2.4 Disclose third-party research access	MEDIUM	↗ MODERATE
2.5 Disclose organizational resourcing commitments and dedicate funding commitments towards post-deployment societal impacts	HIGH	↗ MODERATE

## Discussion

We look at the three impact areas highlighted in turn, then summarize information on collaboration and funding efforts.

### LABOR IMPACTS

AI systems are increasingly capable of performing economically valuable activities once exclusive to humans. They do so in ways that affect well-being, reshape labor markets, and raise questions about the distribution of economic gains. The [International AI Safety Report](#) highlights that “General-purpose AI systems can automate or help with tasks that are relevant to many jobs worldwide”, but also states that “predicting labor market impacts is difficult”. This underscores the need for additional research to improve our predictive capabilities.

**Research on labor impacts has evolved from relying on external observations and inferences to using internal telemetry.** Work research has primarily been led by civil society, unions, and academia using direct employee surveys and public government datasets. However, model developers are now sharing unprecedented usage data from generative AI models, as discussed in Practice 1. This provides novel, real-world data points that allow researchers to move beyond theoretical job risk to track actual patterns of task displacement and augmentation over time. This empirical foundation will serve as the basis for developing new methodologies to assess the long-term effects of AI on global labor and shared prosperity.

**Model providers are measuring AI’s impact by developing indicators related to tasks, skills, and the workforce.** They are doing so by analyzing [conversation logs](#), conducting [surveys](#), and defining new [benchmarks](#) to evaluate a model’s performance on real-world tasks. This information may be fueling much of the [recent drive](#) for AI adoption and use

among enterprise-level organizations, with a focus on providing tangible ROI data such as productivity gains. It can also help mitigate the risks of displacement and inequity for vulnerable labor communities and demographics by preparing the workforce for changes in the [labor market](#). Emerging analysis shows that the younger generation of workers is experiencing a relative decline in employment since AI's integration into enterprises. The question that remains is: what should actors such as unions, enterprises, and policymakers do now with this information?

**Other labor indicators related to human rights and policymaking are less frequently reported.** A [report](#) from the International Labour Organization (ILO) reviewed 245 global AI ethics frameworks addressing work, labor rights, and fundamental rights, and found that notable policy gaps remain in areas such as algorithmic management, worker surveillance, and automation-related displacement. Ongoing research by model providers and their [private-public partnerships](#) with [civil society and academia](#), informed by workers and their representatives, can provide the empirical foundation for evidence-based regulation and organizational governance. This progress in reporting labor indicators represents one step toward accountability mechanisms that ensure AI development aligns with the values of [shared prosperity](#) that PAI believes in.

## ENVIRONMENTAL IMPACTS

The environmental impact of the development and use of foundation models is a well-known fact. The International Energy Agency [projects](#) that data centers could consume up to 3% of the global energy supply by 2030, while the mining of rare earth minerals, the use of water resources for cooling, and the excess heat from underwater data centers will [affect](#) environmental sustainability. However, there is also optimism that [AI can help](#) reduce overall emissions. Optimizing resource use, nudging people towards more sustainable behaviors, and modeling climate systems are all ways AI can reduce environmental impact.<sup>17</sup> Given this backdrop, it's crucial to track the information on how the use of foundation models benefits and harms the environment.

**Environmental reporting is progressing.** The Allen Institute for AI published reproducible environmental impact [data](#) for training its OLMO models, as well as simulated operational impact data. Mistral [worked with](#) external auditors to understand the greenhouse gas emissions, water, and materials consumption of the Mistral Large 2 model over its lifecycle, as well as the impacts of a page-long chat response. Google [released estimates](#) on the impact of a median Gemini Apps text prompt, and Sam Altman [blogged](#) about the average energy use of a ChatGPT query (though didn't provide a methodology).

**However, there are significant discrepancies among these published findings due to differences in methodology.** The reported water usage for an "average" Mistral LeChat query was 173x higher than for an "average" Google Gemini app query (45 vs. 0.26 mL). To confront discrepancies, third-party evaluation initiatives, such as the [ML.ENERGY Leaderboard](#) and [AI Energy Score Leaderboard](#), compare models using their own measurement methods, but the results are not a proxy for impact because only foundation model providers know the

<sup>17</sup> For example, the International Energy Agency (IEA) highlighted that widespread adoption of AI could lead to energy savings of 8% by 2035 in areas such as the manufacturing of electronics.

internal systems, hardware, and configurations that contribute to the total impact and the total scale of usage across services, making private analysis unreproducible.

Other companies have published sustainability reports without providing AI-specific figures,<sup>18</sup> or highlighting initiatives that help reduce AI-related energy and water use.<sup>20</sup>

### SYNTHETIC MEDIA IMPACTS

New media generation models, including Google's [Nano Banana](#), Meta's [Lama 4](#), and OpenAI's [Sora 2](#), now produce more realistic and sophisticated outputs than those initially included in the scope of this series of reports. The accessibility and sophistication of these image generation models enable the creation of advanced AI tools that contribute to a growing number of [real-world harms](#) related to deepfakes, while the lack of a universal approach in providing users with a standardized content provenance mechanism has contributed to a growing [societal inability](#) to discern authentic from synthetic. Provenance serves as the necessary infrastructure to restore public confidence in digital artifacts by providing a verifiable trail of a media's origin and history. In an effort to lead from the front, model providers have led and supported research on content provenance [good practices](#), and advanced [technical research](#) and [innovation](#).

**Since our last report, model providers have continued research on watermarking for AI-generated and altered content.** Last year's report highlighted Meta, Google, Microsoft, and OpenAI as model providers<sup>20</sup> that released public documentation on their synthetic media impact indicators, including research on invisible watermarking and disclosure mechanisms. This year, additional research led to products such as Google's [SynthID-O](#), Meta's [Chunky Seal](#), and Microsoft's [VeriTrail](#). Similarly, many model providers<sup>21</sup> adopted the [C2PA standard](#) as baseline best practices for contextual disclosures, although an October 2025 [third-party audit](#) found that five major platforms were failing in these commitments.<sup>22</sup>

**While there has been some industry alignment on user-facing transparency mechanisms, model providers differ in their technical approaches to model-level transparency.** An attempt to standardize model-level marking in the [EU Code of Practice on Transparency of AI-generated Content](#) (something we call for in the PAI [Synthetic Media Framework](#)) was removed between the [first](#) and [second](#) drafts. This omission could lead to a fragmented approach to disclosure mechanisms among downstream providers. Nonetheless, there has been some general progress on disclosure at the output level as the second draft requires signatories to provide at least two layers of [markings](#) (digitally signed metadata, invisible watermark, or fingerprinting).

**Efforts to combat AI-generated CSAM content and deepfakes have also become more prominent in the private and public sectors.** Model providers such as [Anthropic](#), [Stability AI](#), and [OpenAI](#) published transparency reports on CSAM detection processes, while policymakers are developing legislation, including the US' [TAKE IT DOWN Act](#), California's [SB 942](#), and the EU's [Code of Practice on Transparency of AI-Generated Content](#).

<sup>18</sup> For example, Google notes that their "total data center electricity consumption grew by 27% in 2024, [though this isn't] solely driven by AI."

<sup>19</sup> For example, [Meta](#) and [Microsoft](#).

<sup>20</sup> PAI's [Safeguarding Trust and Dignity in the Age of AI-Generated Media](#) is set of recommendations driven by a large set of organizations not included in the scope of this report.

<sup>21</sup> Model providers in our scope that adopt C2PA standards include Microsoft, Google, Meta, OpenAI, Anthropic, and Stability AI.

<sup>22</sup> Amidst the progress, it is important to point out that, for the C2PA standard to be a success, universal adoption by all actors in the synthetic media ecosystem is essential, as the standard's transparency and disclosure do not carry over when content moves from a C2PA-adopting platform to one that is not.

**There is significant collaboration among actors to prevent the misuse of synthetic media.**

These efforts are driven by collaboration among model providers and other value chain actors, as evidenced by PAI's [Synthetic Media Framework case studies](#) and Thorn and All Tech Is Human's [Safety by Design for Generative AI: Preventing Child Sexual Abuse](#) commitments.

#### **PROVIDING MODEL ACCESS, RESOURCES AND FUNDING FOR EXTERNAL RESEARCH**

Research driven by non-industry entities is vital to any research ecosystem. However, inference costs can be a significant barrier to research, with one team [estimating](#) it would cost USD \$20,000 to run the Online Mind2Web evaluation on Claude Opus 4.1. In other cases, access can be a barrier, especially for frontier models, and for societal research it's crucial that researchers avoid conflicts of interest that could bias methods and results. This is why foundation model providers must create structured access programs and fund research without stipulations.

**Most companies provide some level of access and funding for researchers and scholars.**

[Anthropic](#), [Cohere](#), [Google](#), [IBM](#), [Microsoft](#), and [OpenAI](#) provide scholarship opportunities, funding, or access credits to their systems to support research related to societal impacts. [Amazon](#) funds a challenge related to developing trustworthy AI. Also, open models are inherently accessible to researchers, though the level of "openness" can vary.<sup>23</sup>

**Making the level of access and funding more transparent and accessible would support**

**accountability.** The exact amounts spent on access support or funding is not easily identified given the fragmented nature of reporting. This makes comparison with stated commitments and other organizations difficult. While the level of investment will vary by organizational mission and founding values, business model, structure, and other factors, making this more transparent and accessible will enable greater accountability.

<sup>23</sup> See I. Solaiman's [The Gradient of Generative AI Release: Methods and Considerations](#) and the [European Open Source AI Index](#) for more information on levels of "openness."



### PRACTICE 3

## Report incidents and disclose policy violations

The public is reporting more AI incidents, with the [AI Incident Database](#) recording a 35% year-over-year increase in reports.<sup>24</sup> However, these reports lack the technical specificity and causal analysis needed to be actionable, resulting in information that is more noise than signal. While a few high-profile cases have begun to crystallize public concern about serious failures, such as alleged AI-influenced [suicides](#) and [hallucinations](#), there is currently a lack of systematic reporting to move beyond anecdotes toward true accountability for serious failures.

Additionally, the collective impact of platform misuse can be underestimated because it occurs below the threshold for these incidents. However, the numbers are substantial: in one six-month period, Google [removed](#) over 2.4 billion URLs from Search due to Intellectual Property (IP) infringements, and OpenAI proactively [banned](#) 1,847 accounts for child exploitation.<sup>25</sup> Transparency about the scale and distribution of such violations is an important mechanism for enabling public scrutiny and providing policymakers with the evidence required to combat digital harms effectively.

Practice 3 focuses on documenting information on incidents and policy violations. Our research since 2023 highlights that foundation model providers can:

- Document and enact a comprehensive, continuous coverage incident monitoring framework that specifies risks, thresholds, and escalation procedures, with [real-time failure detection](#) for hosted access products
- Publish detailed summaries, where appropriate,<sup>26</sup> of security incidents – including actor identifiers, misuse patterns, and remediation actions taken – and of safety incidents – including failure details, impacts, and remediation actions taken.<sup>27</sup>
- Systematically report incidents to a third party through formal industry or regulatory incident reporting channels, building on established reporting practices such as CSIRT usage and NIS frameworks.
- Monitor for usage policy violations, including child sexual abuse material (CSAM).
- Share aggregate data on violations by category, enforcement actions, appeals and overturns, and reporting to external bodies, such as NCMEC.

<sup>24</sup> There was a 35% rise in incidents reported between October 1, 2023 – September 30, 2024 (249 incidents) and October 1, 2024 – September 30, 2025 (336 incidents).

<sup>25</sup> We note that these are mandated as part of the EU's DSA reporting requirements for very large online platforms and search engines.

<sup>26</sup> We note that public disclosure of security incidents is not always appropriate. For example, certain vulnerabilities may require restricted reporting channels with strict confidentiality protections, or may warrant a separate reporting system from safety incidents altogether.

<sup>27</sup> We use definitions shared by Microsoft: Security incidents involve failures of confidentiality, integrity, or availability – for example, a threat actor altering the intent of a system. Safety incidents involve harm to users or society without necessarily compromising system security – for example, a system providing inconsistent quality of service across users without explicit instruction to do so. Alternate definitions for safety, rights, and security incidents are shared in [Designing Incident Reporting Systems for Harms from General-Purpose AI](#) (p42).

## PERSPECTIVES FROM OUR 2026 SURVEY

“It’s really helpful for policymakers to know about acute incidents—like the recent cyber-attacks reported by Anthropic. However these incidents can be anecdotal and difficult to develop evidence based policy with—because you don’t know if you’ve captured everything, and some harms aren’t exactly ‘incidents’ (especially systemic ones).” — ANONYMOUS

“Understanding incidents of AI harms occurring inside and outside of model developers [...] will be critical for ensuring that we trust in AI systems and have genuine confidence that they are doing more good than harm. Also, that we have early warning signs and good forecasts or models for predicting and managing threats, failures, and setting related policy.”

— PETER SLATTERY, MIT AI RISK INITIATIVE

“I place increased significance on bio/cyber misuse risk mitigation and reporting, and believe that research on that front (and on AI on the defence side) can help us proactively avoid bad outcomes.”

— VASILIOS MAVROUDIS, THE ALAN TURING INSTITUTE

### Has there been progress for 2026?

Our 2025 report showed that foundation model providers conducted abuse monitoring when integrating models into their products and had public reporting channels for policy violations. However, there were significant challenges in monitoring open models. Companies had released tools, such as [LlamaGuard](#) and [Granite Guardian](#), to help downstream deployers monitor for specific risks, though there had been little coordination to identify model-related incidents. When companies identified incidents or policy violations, they rarely published any data, and information-sharing infrastructures were not developed for third-party disclosures.

For the 2026 report, we still see barriers to establishing common definitions, incident thresholds, and information-sharing infrastructures, and we still find incentives that block widespread reporting of incidents and policy violations.<sup>28</sup>

**28** Incidents are strongly shaped by deployment context and downstream modifications that deployers can observe and remediate, which provides a barrier to foundation model providers reporting without the right coordination channels (see Challenge 2 on [Data Sharing and Coordination Barriers from Documenting the Impacts of Foundation Models](#)).

### Practice 3: Progress made since 2025

MADE BY	DEGREE OF PROGRESS
Practice leaders (Anthropic, OpenAI)	↗ MODERATE
The rest of the field	– LIMITED

#### Progress made in each process

PROCESS	2025 LEVEL OF ADOPTION	2026 DEGREE OF PROGRESS
3.1 Monitor for incidents	MEDIUM	↗ MODERATE
3.2 Monitor for policy violations	MEDIUM	↗ MODERATE
3.3A Share summaries of internal incident reports <sup>29</sup>	LOW	↗ MODERATE
3.3B Share summaries of internal policy violation reports <sup>29</sup>	LOW	– LIMITED
3.4 Systematically report AI incidents to a third party	NONE	– LIMITED

<sup>29</sup> We have separated sharing summaries of internal incident and policy violation reports to enable more granular analysis.

## Discussion

Below, we examine how organizations report incidents (3.1, 3.3A, 3.4) and policy violations (3.2, 3.3B).

### REPORT INCIDENTS

**Governments and policymakers have provided more information on and embedded in legislation what constitutes an incident, including reporting instructions.** However, thresholds for harm are not aligned. California will [establish](#) a “mechanism to be used by a frontier developer or a member of the public to report a critical safety incident,” to which frontier model developers must submit incidents, and New York will [require](#) large providers to submit incidents to the Attorney General and Division of Homeland Security and Emergency Services. The EU General-Purpose AI (GPAI) Code of Practice [describes](#) how GPAI providers must report serious incidents to the EU AI Office with [supporting guidance](#), and the OECD [published](#) a common reporting framework for reporting AI incidents. However, the thresholds for what constitutes an incident vary by jurisdiction.<sup>30</sup>

<sup>30</sup> For example, the EU defines a “serious AI incident” by “the death of a person, or serious harm to a person’s health,” “serious harm to property” and other criteria. However, California defines a “critical safety incident” by specific activities that lead to a death or serious injury, “[a material contribution] to the death of, or serious injury to, more than 50 people” and “more than one billion dollars (\$1,000,000,000) in damage to, or loss of, property.”

**There is still no coordinated and systematic incident reporting.** Organizations are not reporting incidents through any channels, nor are they publishing information about the process, which would educate other actors on threats and the efficacy of remediation actions.<sup>31</sup>

<sup>31</sup> We note that incidents are strongly shaped by deployment context and downstream modifications that deployers can observe and remediate, which provides a barrier to foundation model providers reporting without the right coordination channels (see Challenge 2 on Data Sharing and Coordination Barriers from [Documenting the Impacts of Foundation Models](#)).

**Foundation model providers continue to conduct significant monitoring,** as documented in the responsible scaling policies and frontier model frameworks

of [Amazon](#), [Anthropic](#), [Cohere](#), [Google](#), [Meta](#), [Microsoft](#), [OpenAI](#), and [xAI](#).<sup>32, 33</sup> The Frontier Model Forum [summarizes](#)<sup>34</sup> these approaches. Amazon [points to](#) “incident escalation and response pathways” that enable rapid remediation of reported AI safety incidents, while Google [describes](#) a “hierarchical supervision” procedure in which the “most suspicious or unclear cases are escalated to more capable, expensive monitors.” Both underscore the importance of automated monitoring, complemented by robust governance protocols for responding to identified incidents.

**Companies are sharing more details about specific incidents of malicious use.** In 2025, Google’s Threat Intelligence Group published [two reports](#) detailing incidents of misuse. Anthropic published a [Threat Intelligence Report](#) in August 2025, identifying incidents of misuse and the steps taken to detect and mitigate them. OpenAI published a similar [report](#) in October. Microsoft also published cybersecurity-focused [information](#) on AI threats and responses. These reports provide case studies to show how foundation models are used for end-to-end fraudulent operations, cyber operations, and malware development, along with the steps organizations took to respond to these threats (see the Appendix for more information).

**Foundation model providers coordinate efforts and are increasing private information sharing.** [Microsoft](#) has internal and external coordination mechanisms, including “nation-state actor analyses reported by Microsoft’s Threat Analysis Center.” All foundation model providers about vulnerabilities, threats, and concerning capabilities through the Frontier Model Forum.

## REPORT POLICY VIOLATIONS

**A minority of companies report statistics on Child Sexual Abuse Material (CSAM).** [Anthropic](#) and [OpenAI](#) both report the number of banned accounts and the total number of CSAM-related content items reported to the National Center for Missing & Exploited Children. OpenAI also [reported](#) categorized notices received and proactive responses to policy violations, as required by EU regulations. However, most providers do not provide summaries of the scale and type of policy violations being reported.

**Organizations are not reporting on categories of policy violations, whether identified through monitoring or public reporting.** Organizations are conducting monitoring, and there continue to be mechanisms for the public to report policy violations by open- and restricted-access model providers.<sup>35</sup> However, we are not seeing aggregated information on these violations, in line with what organizations share as part of their EU Digital Service Act reporting obligations.

**32** Mistral committed to provide public transparency on how they develop and deploy their frontier AI models and systems responsibly, but have not published a public framework. IBM shared details on “how IBM’s AI governance framework and our organizational culture supports the responsible development and use of AI and aligns with the Seoul commitments’ core objectives” instead of providing a framework document. Alibaba and Deepseek have not released frameworks publicly.

**33** These continue to be updated past the Jan 1, 2026 analysis limit for this paper, such as [Anthropic’s Version 3.0](#).

**34** This is a technical report focused on risks and mitigations related to chemical, biological, radiological and nuclear (CBRN) information threats, advanced cyber threats, and advanced autonomous behavior threats. While these do not cover the entire range of “incidents,” they are an important subset.

**35** For example, Cohere, Meta, and Mistral provide email addresses for users to report usage policy violations.



## PRACTICE 4

### Share user feedback

User feedback is a crucial tool for developing secure, high-quality digital products. However, with the rise of foundation models, these feedback loops are now an important input into governance: how users interact with these models serves as a vital proxy for real world impact. The use of this information is currently underexplored, but could be important to how we govern these systems in the future.

Practice 4 focuses on documenting and sharing feedback received on models through a provider's feedback mechanism. Our research since 2023 highlights that foundation model providers can:

- Provide multiple feedback channels, document how feedback informs improvements, and share feedback summaries
- Provide open forums for feedback documentation and tracking
- Develop and disclose a follow-up or redress process after feedback, with reporting on aggregate statistics
- Create AI-specific, safety and security focused incentive structures to generate feedback

#### PERSPECTIVES FROM OUR 2026 SURVEY

“Models and products are built for people to use. It is important to understand user needs.”

— REENA JANA, GOOGLE

### Has there been progress for 2026?

Last year, foundation model deployers offered a myriad of ways for users to provide feedback. Thumbs-up/thumbs-down feedback in UIs was common, alongside email or web forms, and forums, mainly for developers. However, feedback aggregation was limited, and incentives were rare.

So far in 2026, organizations have shared feedback on issues affecting their users and provided opportunities for in-app feedback. Rising security risks are prompting organizations to gather feedback on security issues, but they need to do more to encourage feedback on a broader range of problems.

## Practice 4: Progress made since 2025

MADE BY	DEGREE OF PROGRESS
Practice leaders (Amazon, Anthropic, Google)	↗ MODERATE
The rest of the field	– LIMITED

### Progress made in each process

PROCESS	2025 LEVEL OF ADOPTION	2026 DEGREE OF PROGRESS
4.1 Disclose the process of having a feedback mechanism for stakeholders	HIGH	↗ MODERATE
4.2 Aggregate individual user feedback records to have as summaries	MEDIUM	– LIMITED
4.3 Disclose the feedback follow-up process or, if warranted, the redress mechanism process	LOW	– LIMITED
4.4 Create incentive structures to invite stakeholders to participate in the feedback process proactively	LOW	↗ MODERATE

## Discussion

**In-product feedback remains prevalent**, including thumbs-up/thumbs-down functionality, emails, forms, developer forums, and community boards for open models that aggregate and rank user-driven feedback. Foundation model providers say they use feedback to improve their products. [Google](#) and [Anthropic](#) stated that specific policy and product updates are based on user feedback.

**Providers have also disclosed feedback on major user experience issues.** Anthropic provided [summaries](#) and responses to three infrastructure issues reported by users, one of which affected 16% of Sonnet 4 requests. However, aggregation remains predominantly community-driven via open feedback boards. There are also examples in which providers summarized internal feedback using foundation model tools.<sup>36</sup>

**There are many examples of how to incentivize feedback on security issues, but few for safety issues.** [Google](#) and [Meta](#) build on existing practices by offering bug bounties to security researchers. However, these do not apply to the wide range of failures or safety risks users may encounter. The broadest examples come from [Anthropic](#) and [OpenAI](#), which slightly widen the security-focused scope to cover universal jailbreaks and biochemical risks. Whether model providers follow up on feedback beyond what's described on community boards remains unknown.

<sup>36</sup> For example, IBM highlight how Key Point Summarization achieves this for internal feedback summarization.

# Key Insights

INSIGHT

RELATED TO PRACTICE

**INSIGHT 1** Foundation model providers that host their own services can track why and how people use foundation models without compromising privacy. This can inform public policy.

 1. SHARE USAGE INFORMATION

Anthropic, Microsoft, and OpenAI have demonstrated that analyzing usage patterns at a massive scale is possible without compromising individual privacy by using foundation models to categorize and aggregate user interactions. By measuring usage factors over time, policymakers can examine how we are using foundation models now, how we might use them in the future, and the potential impacts on labor and the economy. However, anonymized datasets across organizations will need to be accessible and comparable to be more representative of our society and more valuable to policymakers. Emerging initiatives, such as the UK's [AI Economics Institute](#), should consider how to facilitate this.

**INSIGHT 2** Providers are signaling risks to entry-level workers and changing workforce skill requirements, but have not published research on risks related to worker abuse and sourcing.

 1. SHARE USAGE INFORMATION  
 2. SHARE SOCIETAL RESEARCH (LABOR)

Many model providers have published extensive studies on how their AI technologies are affecting the labor market, using a range of methodologies based on model usage data (see Insight 1). Most of the reporting focuses on [risks in entry-level hiring](#) among the younger generation, challenges around “skill transformations” across different occupations (which address both the loss of tasks to technology and upskilling), and the productivity indicators that capture the benefits and trade-offs of deployment for enterprise use. Notably, there is less reporting on other labor issues, such as [worker abuse-related risks](#) (e.g., worker well-being and worker surveillance) and [sourcing-related risks](#) (e.g., compensation for data labelers). Since labor impact indicators are numerous and varied, it will be difficult to standardize methodologies or agree on which indicator is most pressing; however, other actors in the value chain, including workers, can help set these agendas.

---


**INSIGHT 3**      **Providers report environmental impacts differently, so standards are crucial.**

 2. SHARE SOCIETAL RESEARCH (ENVIRONMENTAL)

Last year, we highlighted that the lack of standards for environmental reporting was blocking adoption. This year, some foundation model providers have shared information that initially appears comparable but differs enough to warrant further exploration. For example, Google Gemini’s reported “average” water usage per prompt (0.26 mL) is 173x lower than Mistral LeChat’s (45 mL), likely reflecting methodological differences rather than actual efficiency differences. Methods for measuring environmental impact can and should be comparable, so formal or de facto standardization can build on Allen Institute for AI’s detailed [methodology](#), ISO’s [technical report](#) on environmental sustainability, or existing environmental impact [leaderboards](#).

---

**INSIGHT 4**      **Providers report AI-generated or uploaded Child Sexual Abuse Material (CSAM), but the prevalence of other foundation model misuses remains unknown.**

 2. SHARE SOCIETAL RESEARCH (SYNTHETIC MEDIA)

 3. REPORT INCIDENTS & VIOLATIONS

Anthropic and OpenAI reported 5,618 and 182,226 instances of CSAM in 2025, respectively. However, in line with 2024 research highlighting a “lack of transparency in enforcement”, foundation model providers do not report on broader violations of usage policies. Without data on how organizations respond to policy violations, and what these violations are, it is impossible to understand the risks involved with the use of these systems and to audit the effectiveness of safety policies.

---

**INSIGHT 5**      **Voluntary incident reporting – both privately and publicly, where appropriate – focuses on security incidents. This accelerates threat detection and attribution, but overlooks broader harms to users and society.**

 3. REPORT INCIDENTS & VIOLATIONS

Anthropic, Google, Microsoft, and OpenAI’s incident reports show that public disclosure of threat patterns enabled peer organizations to identify connected malicious activity, while private threat intelligence partnerships enabled rapid, detailed information exchange. This combination of public reporting and private coordination appears to be more effective than either approach alone. However, there is limited public reporting on impactful model failures or “safety” incidents, such as hallucinations or reliability issues (as seen in the Anthropic example in Practice 4), despite these falling within EU and Californian definitions at certain impact thresholds.

---

**INSIGHT 6****Providers prioritize generating feedback on security risks over broader safety concerns.** 4. SHARE USER FEEDBACK

Building on a strong history of bug bounties, foundation model providers incentivize security-focused feedback. However, this leaves a significant gap for other harms – there’s no financial incentive to report model issues related to CSAM, hallucinations or other failures. The current structure does not incentivize this feedback and mirrors the system that encourages security over safety incident reporting in Insight 5.

---

**INSIGHT 7****Scientific and policy guidance for open model impact reporting faces significant implementation challenges that reduce actual reporting.** 1. SHARE USAGE INFORMATION  
 3. REPORT INCIDENTS & VIOLATIONS  
 4. SHARE USER FEEDBACK

The International AI Safety Report’s [Second Key Update](#) and the EU’s [GPAI Code of Practice](#) detail how model providers could track the diffusion and use of their open models – including watermarking the text, image, video, or audio output, or even the model weights themselves. They also highlight that tampering and watermark stripping reduce the efficacy of these methods. The challenges of collecting and collating post-deployment impact information on open models will require additional research and coordination.

---

# Conclusion

After reviewing over 150 sources and interviewing community members, this report builds on our earlier work and presents three key contributions to improve information sharing and to document the post-deployment impact of foundation models:

- Updated descriptions of good practices for documenting post-deployment impacts.
- An evaluation of progress by 13 organizations in implementing those practices, with highlighted examples.
- An analysis of key gaps and trends, with seven key insights.

Our work shows that publicly accessible impact data is incomplete and often hard to find. Company participation is inconsistent, and important information is scattered across blog posts, earnings calls, and research papers. These findings are in line with PAI's recent research on the [formal reporting practices](#) of 50 major companies.

The field has yet to overcome the five core challenges listed in our [2025 report](#): a lack of standardized documentation norms; barriers to data sharing and coordination; misaligned incentives; limited information-sharing infrastructure; and the decentralized nature of open model deployment. However, by surfacing good practices from leading providers, we hope to encourage greater consistency, to some extent realign incentives, and catalyze a “race to the top.” However, consistency will require action from all organizations.

Next, we will work with our community of partners to develop recommendations for overcoming these challenges. From there, we will publish guidance on what foundation model providers, policymakers, and other actors should do to drive the adoption of good practices.

This report is part of PAI's work to develop trust in the AI value chain. It supports our work on [Strengthening the AI Assurance Ecosystem](#) by providing an analysis of one aspect of transparency, a key enabler of assurance. It also complements our research on the [disclosure of AI-related impacts, risks, and opportunities](#) and our [documentation work](#) through our Enterprise Steering Committee by analyzing disclosures outside of formal channels and documentation with a purpose beyond managing impacts and risks across the value chain.

To discuss this work, please contact [jacob@partnershiponai.org](mailto:jacob@partnershiponai.org).

## APPENDIX 1

# Summary of Practices

PRACTICE	PROCESSES	EXAMPLES
<b>Share usage information</b>	1.1 Conduct surveys or user research to understand downstream usage	Anthropic’s <a href="#">Economic Index</a> shows that Singapore has the highest proportion of Claude usage amongst its working age population globally, with the greatest proportion of usage on the topic of “Assist[ing] with academic assignments and coursework across multiple disciplines”
	1.2 Create tools to support the sharing of activity logs with trusted third parties for analysis	
	1.3 Implement and track watermarking or identifiers	
	1.4 Report aggregate usage statistics, across geographies, sectors, or use cases, including usage in high-risk use cases	
	1.5 Share information on downstream applications of the model	
<b>Enable and share research on post-deployment societal impact indicators</b>	2.1 Report on labor impact indicators	The Allen Institute for AI’s <a href="#">Holistically Evaluating the Environmental Impact of Creating Language Models</a> estimates power usage, carbon emissions, and water consumption from training their dense transformers and compares them to other open models.
	2.2 Report on environmental impact indicators	
	2.3 Report on synthetic content impact indicators	
	2.4 Disclose third-party research access	
	2.5 Disclose organizational resourcing commitments and dedicate funding commitments towards post-deployment societal impacts	
<b>Report incidents and disclose policy violations</b>	3.1 Monitor for incidents	Google’s <a href="#">GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools</a> details PUKCHONG’s (aka UNC4899) misuse of Gemini across the attack lifecycle.
	3.2 Monitor for policy violations	
	3.3A Share summaries of internal incident reports <sup>37</sup>	
	3.3B Share summaries of internal policy violation reports <sup>37</sup>	
	3.4 Systematically report AI incidents to a third party	
<b>Share user feedback</b>	4.1 Disclose the process of having a feedback mechanism for stakeholders	OpenAI’s <a href="#">Safety Bug Bounty program</a> provides incentives for safety feedback.
	4.2 Aggregate individual user feedback records to have as summaries	
	4.3 Disclose the feedback follow-up process or, if warranted, the redress mechanism process	
	4.4 Create incentive structures to invite stakeholders to participate in the feedback process proactively	

<sup>37</sup> For this year’s report, we separate out reporting incidents and reporting policy violations to provide deeper analysis on each area.

## APPENDIX 2

# Methodology

In 2025, we reviewed whether 12 foundation model providers had adopted 18 processes and four practices, and shared an assessment of the “level of adoption” in the field.<sup>38</sup> This highlighted how much, and how little, foundation model providers are adopting the activities described in PAI’s 2023 [Guidance for Safe Foundation Model Deployment](#).

However, we recognized limitations to the approach taken.<sup>39</sup> Since then, foundation model providers have changed how they release their models, focusing on multiple, tailored model releases, organizations have been created, gained influence, and been bought out, while others have changed their business model entirely. We’ve also seen other researchers highlight high-level gaps in transparency,<sup>40</sup> so we are adapting our approach to complement this evolving field and avoid duplicating efforts.

Our overall approach can be summarized as:

1. Identify the progress each organization has made against each of the 19 processes.<sup>41</sup> The information sources and judgments are documented in [Supplement 1](#).
2. Use information from step 1 to judge the progress of each process, using the impact rubric in [Supplement 2](#) to guide decision-making. Our judgments are documented in [Supplement 3](#).
3. Use information from steps 1 and 2 to judge the progress by “practice leaders” for each of the four practices. Our judgments are documented in [Supplement 3](#).
4. Use information from steps 1 and 2 to judge the progress of the “rest of the field.” Our judgments are documented in [Supplement 3](#).

<sup>38</sup> More information on our methodology can be found in PAI’s [Documenting the Impacts of Foundation Models](#) report.

<sup>39</sup> For example, norms for these practices – such as what constitutes a good incident report summary – are not developed enough to measure the quality of transparency with simple “yes” and “no” questions. And while we may want foundation providers who share their model weights openly to disclose these impacts, there will be a different route to measurement compared to companies which host their models and provide API or browser access, so simpler evaluation methods will miss significant nuances.

<sup>40</sup> For example, the [2025 Foundation Model Transparency Index](#) provides pass/fail evaluations for 100 transparency indicators. However, this approach covers upstream, model, and downstream indicators, so is unable to provide detailed analysis into each indicator. We aim to complement this work by sharing more detailed analysis on subsets of this work, including usage and impact subdomains.

<sup>41</sup> To enable more detailed analysis, this year we split process 3.3 into two separate processes: 3.3A: *Share summaries of internal incident reports*; and 3.3B: *Share summaries of internal policy violation reports*.

## Defining “the field” of foundation model providers

There are many foundation model providers in the field, so we focused our review on the following 13 organizations. We do not claim that this represents the entire field, but we chose organizations based on their impact and to provide a representative selection by release type, geography, and release strategy (i.e., B2B or B2C). We do not link documentation to a specific model, as providers increasingly distribute multiple base models with different variants and versions, and hosted systems may dynamically select a model based on the query’s needs. The models provided are examples of what documentation could relate to.

ORGANIZATION	EXAMPLE MODELS	HOSTED OR DOWNLOADABLE/“OPEN” MODELS
<b>Alibaba</b>	Qwen 3.5 family (e.g. Qwen3.5-397B-A17B)	<a href="#">Downloadable</a> / “open”
<b>Allen Institute for AI</b>	<a href="#">Olmo 3 family</a> (e.g. Olmo 3-Think (32B), Olmo 3-RL Zero (7B))	<a href="#">Downloadable</a> / “open”
<b>Amazon</b>	<a href="#">Amazon Nova family</a> (e.g. Nova Premier, Nova Act)	Hosted
<b>Anthropic</b>	Claude 4 family (e.g. <a href="#">Opus 4.5</a> , <a href="#">Opus 4.1</a> , <a href="#">Sonnet 4.5</a> )	Hosted
<b>Cohere</b>	Command family (e.g. <a href="#">Command A</a> , <a href="#">Command A Translate</a> , <a href="#">Command R+</a> )	Hosted
<b>DeepSeek</b>	DeepSeek family (e.g. Deepseek V3, R1, V3.2)	<a href="#">Downloadable</a> / “open”
<b>Google</b>	<a href="#">Gemini 3 family</a> (e.g. Gemini 3 Pro, Gemini 3 Deep Think)	Hosted
<b>IBM</b>	Granite 4.0 family (e.g. Granite 4.0 Nano, Granite 4.0 Small)	<a href="#">Downloadable</a> / “open”
<b>Meta</b>	<a href="#">Llama 4 family</a> (e.g. Llama 4 Scout, Llama 4 Maverick)	<a href="#">Downloadable</a> / “open”
<b>Mistral</b>	<a href="#">Mistral Medium 3</a> , <a href="#">Mistral Large 2</a> , Mistral Small	Hosted (Mistral Medium 3) and <a href="#">downloadable</a> / “open” (Mistral Small)
<b>Microsoft</b>	<a href="#">Phi family</a> (e.g. Phi-4, Phi-4-mini, Phi-4-multimodal)	<a href="#">Downloadable</a> / “open”
<b>OpenAI</b>	<a href="#">GPT-5.1 family</a> (GPT-5.1 Instant, GPT-5.1 Thinking)	Hosted
<b>Stability AI</b>	<a href="#">Stable Diffusion 3.5 family</a> (Large, Turbo, Medium)	Hosted

## Defining “progress”

This year, we focus more on assessing the quality of documentation and providing space for discussion. While this is more challenging to quantify and summarize, and may be open to greater interpretation, this approach accounts for the nuances of documentation and creates a strong foundation for developing valuable recommendations. We cover progress at both the practice and process levels.

Our reporting will cover two dimensions:

### 1. Progress in sharing more useful information by “practice leaders.”

Ultimately, we want organizations to share information so that *other actors can use that information to achieve a benefit*. This might be to amplify the societal benefits of AI (e.g., by promoting valuable use cases), to manage and mitigate the risks of AI (e.g., by creating safer downstream products), or to develop evidence-based policies (e.g., by analyzing data to assess and prioritize governance interventions). Doing better might look like:

- Entirely new information: Data on a topic that has not been shared before.
- Higher quality information: Data that covers an existing topic, but is more complete, unique, accurate, accessible, or covers a longer time period.<sup>42</sup>

For each process, some organizations share more useful information than others; our evaluations are presented in [Supplement 3](#). Some organizations then excel at groups of processes at the practice level, and we define these as “practice leaders.” We evaluate their progress in improving the quality of a specific practice to identify how well the best in the field are doing.

### 2. Progress in the number of foundation model providers adopting an activity.

Once one or more organizations share useful information, we want other organizations to follow suit to ensure the field can be held to account evenly and that the benefits for policymakers and users are widely shared. Therefore, we evaluate the progress of “the field” in adopting specific processes and practices to a level near that of practice leaders.

To operationalize our principle of measuring the quality of documentation by *how well it enables other actors to achieve a benefit*, we provide guidance on what constitutes significant progress for each process in our impact rubric, which is available in [Supplement 2](#). We then use our judgment to evaluate the overall progress made in each process by the field, and document our rationale. Our approach is summarized below.

<sup>42</sup> These criteria are based on Quality Assurance Framework of the European Statistical System and the UK’s Government Data Quality Framework.

## Progress in practice adoption

DEGREE OF PROGRESS	BY PRACTICE LEADERS	BY THE REST OF THE FIELD
↑ SIGNIFICANT	Improvements since November 2024 have been <b>major</b> and have <b>had a strong impact</b> on how other actors can use this information to amplify societal benefits, manage and mitigate risks, or develop evidence-based, proportionate policy.	The <b>majority</b> of foundation model providers have improved their implementation of this practice to near the level of the practice leaders.
↗ MODERATE	Improvements since November 2024 have been <b>moderate</b> and have <b>had a material impact</b> on how other actors can use this information to amplify societal benefits, manage and mitigate risks, or develop evidence-based, proportionate policy.	A <b>minority</b> of foundation model providers have improved their implementation of this practice to near the level of the practice leaders.
– LIMITED	Improvements since November 2024 have been <b>minor</b> and have <b>not had a material impact</b> on how other actors can use this information to amplify societal benefits, manage and mitigate risks, or develop evidence-based, proportionate policy.	<b>Few</b> other foundation model providers have improved their implementation of this practice to near the level of the practice leaders.

## AI usage disclosure

We used manual and agent-based search methods to find publicly available information for each organization, in line with research indicating findings that these methods are complementary.<sup>43</sup> We reviewed all agent outputs to reduce false positive matches.<sup>44</sup> We used AI to provide a two-sentence summary of each information source, to format citations, and to support report editing.

<sup>43</sup> The 2025 Foundation Model Transparency Index notes that in searching for relevant transparency information, “the agent demonstrated both complementary strengths and limitations compared to human evaluators.” The team suggests that “automated agents can meaningfully aid systematic search tasks.”

<sup>44</sup> The 2025 Foundation Model Transparency Index notes that agents provided more false positive matches than manual searching, finding that “automated agents are better suited for augmenting human transparency evaluation teams, as human judgment remains essential for verifying the relevance and accuracy of discovered information”

# Acknowledgements

This report was prepared with guidance from PAI's [Policy Steering Committee](#).

## Members of PAI's Policy Steering Committee in 2025-2026

Rumman Chowdhury, Humane Intelligence	Lisa Pearlman, Apple
Amanda Craig Deckard, Microsoft	Karine Perset, OECD
Alice Friend, Google	Benjamin Prudhomme, MILA
Alex Givens, Center for Democracy and Technology	Andrew Reiskind, Mastercard
Sam Gregory, WITNESS	Andrea Renda, CEPS
Sebastian Hallensleben, Resaro, CEN/CENELEC	Francesca Rossi, IBM
Antonia Kerle, BBC	Irene Solaiman, Hugging Face
Richard Mathenge, African Content Moderators Union	Elham Tabassi, Brookings Institution
Valeria Milanes, Asociación por los Derechos Civiles	David Wakeling, A&O Shearman
Alondra Nelson, Institute for Advanced Study	Deon Woods Bell, Gates Foundation
Marc Etienne Ouimette, Cardinal Policy	

## Special thanks to those who provided comments on the draft report:

Miranda Bogen, Center for Democracy and Technology  
Claire Dennis, Microsoft  
Ian Eisenberg, Credo AI  
Hector de Rivoire, Microsoft  
John Leo Tarver, Independent  
PAI team members (Christian Cardona, Rebecca Finlay, Stephanie Ifayemi, Claire Leibowicz, Eliza McCullough, Neil Uhl)

## We also appreciate the valuable contributions of the following people, and anonymous participants, who shared insights through our survey:

Reena Jana, Google  
Vasilios Mavroudis, Alan Turing Institute  
Peter Slattery, MIT AI Risk Initiative  
Merlin Stein, UK AI Security Institute  
Ilan Strauss, AI Disclosures Project

# Works Cited

This report uses two types of citations:

- Information sources from foundation model providers that informed our progress analysis. These are documented in [Supplement 1: Information Sources](#) and are not listed below.
- Other citations that informed the writing of the report but did not inform specific progress analysis judgments. These appear as hyperlinks throughout the report and are listed below by the first section they appear in.

## Executive Summary

1. Partnership on AI. “Strengthening the AI Assurance Ecosystem.” Partnership on AI, 18 Feb. 2026, [partnershiponai.org/resource/strengthening-the-ai-assurance-ecosystem/](https://partnershiponai.org/resource/strengthening-the-ai-assurance-ecosystem/).
2. Wan, Alexander, et al. “The 2025 Foundation Model Transparency Index.” Center for Research on Foundation Models, Stanford University, 2025, [crfm.stanford.edu/fmti/paper.pdf](https://crfm.stanford.edu/fmti/paper.pdf).
3. Office of the Governor of California. The California Report on Frontier AI Policy. State of California, 17 June 2025, [gov.ca.gov/wp-content/uploads/2025/06/June-17-2025-%E2%80%93-The-California-Report-on-Frontier-AI-Policy.pdf](https://gov.ca.gov/wp-content/uploads/2025/06/June-17-2025-%E2%80%93-The-California-Report-on-Frontier-AI-Policy.pdf).
4. Press Information Bureau, Government of India. “India AI Impact Summit: Frontier AI Commitments.” PIB India, 2025, [pib.gov.in/PressReleasePage.aspx?PRID=2230201&reg=3&lang=1](https://pib.gov.in/PressReleasePage.aspx?PRID=2230201&reg=3&lang=1).
5. California State Legislature. Senate Bill 53. California Legislative Information, 2025, [leginfo.ca.gov/faces/billTextClient.xhtml?bill\\_id=202520260SB53](https://leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53).
6. European Commission. “Code of Practice on General-Purpose AI (GPAI).” European Commission Digital Strategy, 2025, [digital-strategy.ec.europa.eu/en/policies/contents-code-gpai](https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai).
7. Partnership on AI. “Guidance for Safe Foundation Model Deployment.” Partnership on AI, 24 Oct. 2023, [partnershiponai.org/modeldeployment/#landing](https://partnershiponai.org/modeldeployment/#landing).
8. Partnership on AI. “Documenting the Impacts of Foundation Models.” Partnership on AI, 26 Feb. 2025, [partnershiponai.org/paper/documenting-the-impacts-of-foundation-models/](https://partnershiponai.org/paper/documenting-the-impacts-of-foundation-models/).
9. Partnership on AI. “Disclosure of AI-Related Impacts, Risks, and Opportunities.” Partnership on AI, 13 Nov. 2025, [partnershiponai.org/resource/disclosure-of-ai-related-impacts-risks-and-opportunities/](https://partnershiponai.org/resource/disclosure-of-ai-related-impacts-risks-and-opportunities/).
10. Partnership on AI. “Shaping AI Transparency Processes with NIST.” Partnership on AI, 17 March 2026, [partnershiponai.org/shaping-ai-transparency-processes-with-nist/](https://partnershiponai.org/shaping-ai-transparency-processes-with-nist/).

## Introduction

1. OECD. “OECD AI Principles: Transparency and Explainability.” OECD AI Policy Observatory, 2019, [oecd.ai/en/dashboards/ai-principles/P7](https://oecd.ai/en/dashboards/ai-principles/P7).
2. UK Government. “Frontier AI Safety Commitments, AI Seoul Summit 2024.” GOV.UK, 2024, [gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024](https://gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024).
3. Reuters. “U.S. AI Startups See Funding Surge While More VC Funds Struggle to Raise, Data Shows.” Reuters, 15 July 2025, [reuters.com/business/us-ai-startups-see-funding-surge-while-more-vc-funds-struggle-raise-data-shows-2025-07-15/](https://reuters.com/business/us-ai-startups-see-funding-surge-while-more-vc-funds-struggle-raise-data-shows-2025-07-15/).
4. Tacho, Laura. “Data vs Hype: How Orgs Actually Win with AI – The Pragmatic Summit.” YouTube, uploaded by The Pragmatic Engineer, 24 Feb. 2026, [youtu.be/LOHgRw43fFk?si=g5CKepYK6kqRL2r4&t=308](https://youtu.be/LOHgRw43fFk?si=g5CKepYK6kqRL2r4&t=308).
5. Partnership on AI. “Prioritizing Real-Time Failure Detection in AI Agents.” Partnership on AI, 11 Sept. 2025, [partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents/](https://partnershiponai.org/resource/prioritizing-real-time-failure-detection-in-ai-agents/).

6. METR. “What Should Companies Share about Risks from Frontier AI Models?” METR Blog, 27 June 2025, [metr.org/blog/2025-06-27-risk-transparency/](https://metr.org/blog/2025-06-27-risk-transparency/). (fn. 4)
7. Kolt, Noam, et al. “Responsible Reporting for Frontier AI Development.” Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 7, no. 1, 2024, pp. 768–783, [doi.org/10.1609/aies.v7i1.31678](https://doi.org/10.1609/aies.v7i1.31678) (fn. 5)

### Practice 1: Share Usage Information

1. McKinsey & Company. “The State of AI.” McKinsey & Company, 2025, [mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#](https://mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/).
2. Tacho, Laura. “Data vs Hype: How Orgs Actually Win with AI – The Pragmatic Summit.” YouTube, uploaded by The Pragmatic Engineer, 24 Feb. 2026, [youtu.be/LOHgRw43fFk?si=\\_89zDMsIDQtryRD&t=222](https://youtu.be/LOHgRw43fFk?si=_89zDMsIDQtryRD&t=222).
3. US Census Bureau. “Business Trends and Outlook Survey (BTOS).” United States Census Bureau, [census.gov/hfp/btos/about](https://census.gov/hfp/btos/about).
4. Stanford Human-Centered Artificial Intelligence. “2026 AI Index Report.” Stanford HAI, 2026, [hai.stanford.edu/ai-index/2026-ai-index-report](https://hai.stanford.edu/ai-index/2026-ai-index-report).
5. Bourdois, Loubna. “Hugging Face Models Stats.” Hugging Face Blog, 13 Oct. 2025, [huggingface.co/blog/lbourdois/huggingface-models-stats](https://huggingface.co/blog/lbourdois/huggingface-models-stats). (fn. 12)
6. OpenRouter. “State of AI.” OpenRouter, Dec. 2025, [openrouter.ai/state-of-ai](https://openrouter.ai/state-of-ai). (fn. 13)
7. Stanford CRFM. “Foundation Model Transparency Index, December 2025 Reports.” GitHub, Dec. 2025, [github.com/stanford-crfm/fmti/tree/main/Dec2025/reports](https://github.com/stanford-crfm/fmti/tree/main/Dec2025/reports). (fn. 14)

### Practice 2: Enable and Share Research on Post-Deployment Societal Impact Indicators

1. Stanford Human-Centered Artificial Intelligence. “Exploring the Dangers of AI in Mental Health Care.” Stanford HAI News, 11 June 2025, [hai.stanford.edu/news/exploring-the-dangers-of-ai-in-mental-health-care](https://hai.stanford.edu/news/exploring-the-dangers-of-ai-in-mental-health-care).
2. King’s College London. “New Study Reveals Early Impact of AI on Job Market in UK.” KCL News, 7 Oct. 2025, [kcl.ac.uk/news/new-study-reveals-early-impact-of-ai-on-job-market-in-uk](https://kcl.ac.uk/news/new-study-reveals-early-impact-of-ai-on-job-market-in-uk).
3. Ada Lovelace Institute. “Great Expectations: Public Attitudes Towards AI and Regulation.” Ada Lovelace Institute, 4 Dec. 2025, [adalovelaceinstitute.org/policy-briefing/great-expectations/](https://adalovelaceinstitute.org/policy-briefing/great-expectations/).
4. International AI Safety Report. International AI Safety Report 2026. 2026, [internationalaisafetyreport.org/publication/international-ai-safety-report-2026#2.3.1](https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026#2.3.1).
5. Partnership on AI. “Partnership on AI Launches New Initiative to Guide Enterprise Organizations in Responsible AI Adoption.” Partnership on AI, 24 April 2025, [partnershiponai.org/partnership-on-ai-launches-new-initiative-to-guide-enterprise-organizations-in-responsible-ai-adoption/](https://partnershiponai.org/partnership-on-ai-launches-new-initiative-to-guide-enterprise-organizations-in-responsible-ai-adoption/).
6. Brynjolfsson, Erik et al. “Canaries in the Coal Mine: Six Facts About the Recent Employment Effects of Artificial Intelligence.” Stanford Digital Economy Lab, Aug. 2025, [digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/](https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/).
7. International Labour Organization. “Governing AI in the World of Work: A Review of Global Ethics Guidelines.” ILO, 14 Nov. 2025, [ilo.org/resource/article/governing-ai-world-work-review-global-ethics-guidelines](https://ilo.org/resource/article/governing-ai-world-work-review-global-ethics-guidelines).
8. Partnership on AI. “Guidelines for AI and Shared Prosperity.” Partnership on AI, [partnershiponai.org/paper/shared-prosperity/](https://partnershiponai.org/paper/shared-prosperity/).
9. International Energy Agency. Energy and AI. IEA, 2025, [iea.blob.core.windows.net/assets/601eaec9-ba91-4623-819b-4ded331ec9e8/EnergyandAI.pdf](https://iea.blob.core.windows.net/assets/601eaec9-ba91-4623-819b-4ded331ec9e8/EnergyandAI.pdf).
10. Grantham Research Institute on Climate Change and the Environment. “What Direct Risks Does AI Pose to the Climate and Environment?” London School of Economics, 12 Sept. 2025, [lse.ac.uk/granthaminstitute/explainers/what-direct-risks-does-ai-pose-to-the-climate-and-environment/](https://lse.ac.uk/granthaminstitute/explainers/what-direct-risks-does-ai-pose-to-the-climate-and-environment/).
11. Lannelongue, Loïc, et al. “AI Can Help Reduce Overall Emissions.” Nature Sustainability, 23 June 2025, [nature.com/articles/s44168-025-00252-3](https://nature.com/articles/s44168-025-00252-3).
12. ML.ENERGY. “ML.ENERGY Leaderboard.” ML.ENERGY, [ml.energy/leaderboard/](https://ml.energy/leaderboard/).

13. AI Energy Score. “AI Energy Score Leaderboard.” Hugging Face Spaces, [huggingface.co/spaces/AIEnergyScore/Leaderboard](https://huggingface.co/spaces/AIEnergyScore/Leaderboard).
14. International Energy Agency. “AI for Energy Optimisation and Innovation.” IEA, [iea.org/reports/energy-and-ai/ai-for-energy-optimisation-and-innovation](https://iea.org/reports/energy-and-ai/ai-for-energy-optimisation-and-innovation). (fn. 16)
15. Geneva Internet Platform. “Meta Under Fire over AI Deepfake Celebrity Chatbots.” Dig.Watch, [dig.watch/updates/meta-under-fire-over-ai-deepfake-celebrity-chatbots](https://dig.watch/updates/meta-under-fire-over-ai-deepfake-celebrity-chatbots).
16. Cooke, Di et al. “As Good as a Coin Toss? Human Detection of AI-Generated Content.” Communications of the ACM, Sept. 2025, [cacm.acm.org/research/as-good-as-a-coin-toss-human-detection-of-ai-generated-content/](https://cacm.acm.org/research/as-good-as-a-coin-toss-human-detection-of-ai-generated-content/).
17. Partnership on AI. “Safeguarding Trust and Dignity in the Age of AI-Generated Media.” Partnership on AI, 17 June 2025, [partnershiponai.org/resource/safeguarding-trust-and-dignity-in-the-age-of-ai-generated-media/](https://partnershiponai.org/resource/safeguarding-trust-and-dignity-in-the-age-of-ai-generated-media/).
18. Coalition for Content Provenance and Authenticity. “C2PA: Content Credentials.” C2PA, [c2pa.org/](https://c2pa.org/).
19. Indicator Media. “Tech Platforms Fail to Label AI Content; C2PA Metadata Often Missing.” Indicator, Oct. 2025, [indicator.media/p/tech-platforms-fail-to-label-ai-content-c2pa-metadata](https://indicator.media/p/tech-platforms-fail-to-label-ai-content-c2pa-metadata).
20. National Telecommunications and Information Administration. “AI System Disclosures.” NTIA AI Accountability Policy Report, [ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/information-flow/ai-system-disclosures](https://ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/information-flow/ai-system-disclosures).
21. European Commission. “Code of Practice on Transparency of AI-Generated Content.” European Commission Digital Strategy, [digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content](https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content).
22. Partnership on AI. “PAI Synthetic Media Framework.” Partnership on AI, [syntheticmedia.partnershiponai.org/](https://syntheticmedia.partnershiponai.org/).
23. European Commission. “First Draft: Code of Practice on Transparency of AI-Generated Content.” European Commission Digital Strategy, [digital-strategy.ec.europa.eu/en/library/first-draft-code-practice-transparency-ai-generated-content](https://digital-strategy.ec.europa.eu/en/library/first-draft-code-practice-transparency-ai-generated-content).
24. European Commission. “Second Draft: Code of Practice on Marking and Labelling of AI-Generated Content.” European Commission Digital Strategy, [digital-strategy.ec.europa.eu/en/library/commission-publishes-second-draft-code-practice-marking-and-labelling-ai-generated-content](https://digital-strategy.ec.europa.eu/en/library/commission-publishes-second-draft-code-practice-marking-and-labelling-ai-generated-content).
25. Partnership on AI. “Glossary for Synthetic Media Transparency Methods, Part 1.” Partnership on AI, [partnershiponai.org/resource/glossary-for-synthetic-media-transparency-methods-part-1/](https://partnershiponai.org/resource/glossary-for-synthetic-media-transparency-methods-part-1/).
26. Tech Policy Press. “TAKE IT DOWN Act Tracker.” Tech Policy Press, [techpolicy.press/tracker/take-it-down-act-s146/](https://techpolicy.press/tracker/take-it-down-act-s146/).
27. Partnership on AI. “PAI Synthetic Media Framework: Case Studies.” Partnership on AI, [synthetic-media.partnershiponai.org/#case\\_studies](https://synthetic-media.partnershiponai.org/#case_studies).
28. Thorn. “Safety by Design for Generative AI: Preventing Child Sexual Abuse.” Thorn Blog, [thorn.org/blog/generative-ai-principles/](https://thorn.org/blog/generative-ai-principles/).
29. Kapoor, Sayash, et al. “Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation.” The Fourteenth International Conference on Learning Representations (ICLR 2026), 2026, [openreview.net/forum?id=vUaY1t64ZZ](https://openreview.net/forum?id=vUaY1t64ZZ).
30. Solaiman, Irene. “The Gradient of Generative AI Release: Methods and Considerations.” Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), 12–15 June 2023, Chicago, IL. ACM, 2023, pp. 111–122, [doi.org/10.1145/3593013.3593981](https://doi.org/10.1145/3593013.3593981). (fn. 22)
31. European Open Source AI Index. European Open Source AI Index. [osai-index.eu/](https://osai-index.eu/). (fn. 22)

### Practice 3: Report Incidents and Policy Violations

1. AI Incident Database. AI Incident Database. [incidentdatabase.ai/](https://incidentdatabase.ai/).
2. BBC News. “Teen Dies After Forming Emotional Relationship with AI Chatbot.” BBC News, 27 Aug. 2025, [bbc.co.uk/news/articles/cgerwp7rdlvo](https://bbc.co.uk/news/articles/cgerwp7rdlvo).
3. Reuters. “Conservative Activist Sues Google over AI-Generated Statements.” Reuters, 22 Oct. 2025, [reuters.com/legal/litigation/conservative-activist-sues-google-over-ai-generated-statements-2025-10-22/](https://reuters.com/legal/litigation/conservative-activist-sues-google-over-ai-generated-statements-2025-10-22/).

4. Google. Google Transparency Report: Search Removals for Intellectual Property, January–June 2025. Google, 2025, [storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27\\_2025-1-1\\_2025-6-30\\_en\\_v1.pdf](https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2025-1-1_2025-6-30_en_v1.pdf).
5. Wei, Kevin, and Lennart Heim. “Designing Incident Reporting Systems for Harms from General-Purpose AI.” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, no. 44, 2026, pp. 38016–38029, [doi.org/10.1609/aaai.v40i44.41139](https://doi.org/10.1609/aaai.v40i44.41139) (fn. 26)
6. New York State Legislature. Assembly Bill A6453, Amendment B. New York State Senate, 2025, [nysenate.gov/legislation/bills/2025/A6453/amendment/B](https://nysenate.gov/legislation/bills/2025/A6453/amendment/B).
7. European Commission. “AI Act: Draft Guidance and Reporting Template for Serious AI Incidents.” European Commission, [digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks](https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks).
8. OECD. Towards a Common Reporting Framework for AI Incidents. OECD Publishing, 28 Feb. 2025, [oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents\\_f326d4ac-en.html](https://oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html).
9. Frontier Model Forum. Technical Report on Mitigations for CBRN, Cyber, and Autonomous Threats. Frontier Model Forum, June 2025, [frontiermodelforum.org/uploads/2025/06/FMF-Technical-Report-on-Mitigations.pdf](https://frontiermodelforum.org/uploads/2025/06/FMF-Technical-Report-on-Mitigations.pdf).

## Key Insights

1. UK Government. “Chancellor’s Growth Plan Sets Key Principles for UK-EU Alignment.” GOV.UK, 17 March 2026, [gov.uk/government/news/chancellors-growth-plan-sets-key-principles-for-uk-eu-alignment](https://gov.uk/government/news/chancellors-growth-plan-sets-key-principles-for-uk-eu-alignment).
2. Partnership on AI. “Shared Prosperity: Worker Impacts.” Partnership on AI, [partnershiponai.org/paper/shared-prosperity/3/](https://partnershiponai.org/paper/shared-prosperity/3/).
3. International Organization for Standardization. “ISO/TR 86177: Information Technology – Artificial Intelligence – Environmental Sustainability.” ISO, [iso.org/standard/86177.html](https://iso.org/standard/86177.html).
4. International AI Safety Report. “Second Key Update.” International AI Safety Report, Dec. 2025, [internationalaisafetyreport.org/sites/default/files/2025-12/second-key-update-english.pdf](https://internationalaisafetyreport.org/sites/default/files/2025-12/second-key-update-english.pdf).
5. EU AI Office. “EU Code of Practice on General-Purpose AI: Summary.” Code of Practice for General Purpose AI, [code-of-practice.ai/?section=summary](https://code-of-practice.ai/?section=summary).

## Appendices

1. Bommasani, Rishi, et al. “The Foundation Model Transparency Index, December 2025.” Stanford CRFM, Dec. 2025, [crfm.stanford.edu/fmti/December-2025/paper.pdf](https://crfm.stanford.edu/fmti/December-2025/paper.pdf). (fn. 39, 43, 44)
2. European Statistical System. Quality Assurance Framework of the European Statistical System (ESS-QAF) v1.2. Eurostat, [ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646](https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646). (fn. 42)
3. UK Government. The Government Data Quality Framework. GOV.UK, [gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework](https://gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework). (fn. 42)