The Foundation Model Transparency Index v1.1 May 2024

Anonymous authors
Paper under double-blind review

Abstract

Foundation models are increasingly consequential yet extremely opaque. To characterize the status quo, the Foundation Model Transparency Index was launched in October 2023 to measure the transparency of leading foundation model developers. The October 2023 Index (v1.0) assessed 10 major foundation model developers (e.g. OpenAI, Google) on 100 transparency indicators (e.g. does the developer disclose the wages it pays for data labor?). At the time, developers publicly disclosed very limited information with the average score being 37 out of 100. To understand how the status quo has changed, we conduct a follow-up study (v1.1) after 6 months: we score 14 developers against the same 100 indicators. While in v1.0 we searched for publicly available information, in v1.1 developers submit reports on the 100 transparency indicators, potentially including information that was not previously public. We find that developers now score 58 out of 100 on average, a 21 point improvement over v1.0. Much of this increase is driven by developers disclosing information during the v1.1 process: on average, developers disclosed information related to 16.6 indicators that was not previously public. We observe regions of sustained (i.e. across v1.0 and v1.1) and systemic (i.e. across most or all developers) opacity such as on copyright status, data access, data labor, and downstream impact. We publish *transparency reports* for each developer that consolidate information disclosures: these reports are based on the information disclosed to us via developers. Our findings demonstrate that transparency can be improved in this nascent ecosystem, the Foundation Model Transparency Index likely contributes to these improvements, and policymakers should consider interventions in areas where transparency has not improved.

1 Introduction

Foundation models are the epicenter of artificial intelligence (AI) as AI begins to shape how the economy and society function (Bommasani et al., 2021). For such a high-impact technology, transparency is vital to facilitate accountability, competition, and collective understanding. As an illustrative example, the current lack of transparency regarding the data used to build foundation models makes it difficult to assess what copyrighted information is used to train foundation models. Governments around the world are intervening to increase transparency: for example, the EU AI Act and the US's proposed AI Foundation Model Transparency Act take major strides by mandating a number of disclosure requirements (Bommasani et al., 2024, Appendix A).

To characterize the transparency of the foundation model ecosystem, Bommasani et al. (2023a) introduced the Foundation Model Transparency Index (FMTI). Launched in October 2023, the first iteration of the index (FMTI v1.0) scored 10 major foundation developers (e.g. OpenAI, Google, Meta) based on publicly available information regarding 100 transparency indicators. These 100 indicators span matters such as the data, labor, and compute used to build models; the capabilities, limitations, and risks associated with models; and the distribution of models as well as the impact of their use. FMTI v1.0 established that, in the status quo, the foundation model ecosystem was opaque: the average score was 37 points out of 100. Yet FMTI v1.0 also identified heterogeneity in public disclosures: while the top score was a 54, for 82 indicators at least one developer scored a point.

To understand how the landscape has evolved in the last 6 months, we conduct a follow-up study (FMTI v1.1).[1] To enable direct comparison, we retain the 100 transparency indicators and the associated threshold for awarding a point from FMTI v1.0. However, instead of searching for public information as was done in FMTI v1.0, we request that developers report the relevant information for each indicator. We implemented this change for three reasons: (i) **completeness:** we obviate the concern that information was missed when searching the Internet; (ii) **clarity:** we reduce uncertainty by having developers affirmatively disclose information; and (iii) **scalability:** we remove the effort required for researchers to conduct an open-ended search for decentralized public information.

We contacted 19 foundation model developers, and 14 provided reports related to the 100 transparency indicators (Adept, AI21 Labs, Aleph Alpha, Amazon, Anthropic, BigCode/Hugging Face/ServiceNow, Google, IBM, Meta, Microsoft, Mistral, OpenAI, Stability AI, Writer).[2] Given each developer's initial report, we provided scores based on whether each disclosure satisfied the associated indicator. Developers responded to these initial scores, engaging in dialogue via email and virtual meetings, and clarifying matters in many cases. Following this iterative process, for each developer we publish a *transparency report* that consolidates the information it discloses. These reports contain new information, which developers had not disclosed publicly prior to the start of FMTI v1.1. On average, developers disclosed information related to 16.6 indicators that was not previously public.

The FMTI v1.1 results demonstrate ample room for improvement, as well as tangible improvements in transparency over half a year. On average, developers disclose information that satisfies 58 of the 100 transparency indicators. Developers are least transparent with respect to the upstream resources required to build their foundation models, scoring 46%, in comparison to 65% on downstream indicators and 61% on model-related indicators. Developers can become more transparent by drawing on the transparency practices of other developers—at least one developer scores a point on 96 of the 100 indicators, and multiple developers score a point on 89 indicators.

We find that developers' scores improved significantly over FMTI v1.0, with a 21 point improvement in the mean overall score. Scores improved across every domain, with upstream, model, and downstream scores improving by 6–7 points. Each of the 8 developers that were evaluated in both v1.0 and v1.1 improved their scores, with an average increase of 19 points. While some developers disclosed substantially more (e.g. AI21

[1] URL to project website is anonymized for submission.

[2] The change from FMTI v1.0 is the inclusion of 6 developers (Adept, Aleph Alpha, IBM, Microsoft, Mistral, Writer) and the exclusion of 2 developers (Cohere, Inflection).

Labs's score increased by 50 points), others made fairly marginal changes (e.g. OpenAI's score increased by just 1 point).

These findings affirm that greater transparency is feasible in the foundation model ecosystem. Further, they suggest that the Foundation Model Transparency Index in tandem with other interventions drives improvements in transparency. However, given that there has been very little progress on specific indicators (e.g. external data access, mitigations evaluations), we encourage policymakers to assess what level of minimum transparency is needed and to pursue policy interventions accordingly. We publish the transparency reports for all developers to enable further research.[3]

2 Background

To contextualize our effort, we describe FMTI v1.0 and prior work on multi-iteration indices.

2.1 The Foundation Model Transparency Index

Bommasani et al. (2023a) launched the Foundation Model Transparency Index in October 2023. To conceptualize transparency for foundation models, they introduced a hierarchical taxonomy aligned with the foundation model supply chain (Bommasani et al., 2023b). This taxonomy featured three top-level *domains*: the *upstream* resources involved in developing a foundation model, the foundation *model* itself and its properties, and the *downstream* use of the foundation model. These domains aggregate 23 subdomains (e.g. the upstream domain contains data, labor, data access, and compute as subdomains) and 100 binary transparency *indicators*. Using public information identified through a systematic search protocol, FMTI v1.0 scored 10 companies (AI21 Labs, Amazon, Anthropic, Cohere, Google, BigScience/Hugging Face, Inflection, Meta, OpenAI, Stability AI) from 0–100 on the 100 indicators. Companies were sent initial scores and allowed to contest them before FMTI v1.0 was released.

FMTI v1.0 showed a pervasive lack of transparency in the foundation model ecosystem, with the highest-performing developer scoring just 54 out of 100. Developers disclosed very little information about the labor or compute used to build their foundation models (scoring just 17% on these subdomains), or their real world impact (scoring 11% on this subdomain). Open foundation model developers—which refers to developers who released their flagship foundation model openly (i.e. with widely available model weights; Kapoor et al., 2024)—outperformed closed developers by a wide margin: all three developers with open flagship foundation models were among the top four scoring developers. Several companies disclosed almost no information about their flagship foundation models, with three companies scoring 25% or less.

2.2 Indices over time

A central objective of an index is to track a concept over time to characterize changes. In doing so, many notable indices have evolved over time to reflect changing circumstances and priorities. For instance, the Human Development Index (HDI) was changed multiple times in the 1990s and 2000s, largely in response to academic criticism (Klasen, 2018; Stanton, 2007). Despite these changes, which complicate direct comparisons across index iterations, the HDI remains one of the most trustworthy and popular indices for human development.

As an index is conducted repeatedly, and the world changes as measured by the index, a natural question is how the index contributes to this change. Attributing why corporate behavior (such as disclosure practices) changes is notoriously difficult. Companies generally do not reveal why they make changes and changes generally reflect a confluence of multiple factors. Kogen (2022) provides a unique demonstration of an index's impact, analyzing the 2018 Ranking Digital Rights Index (RDR), which ranked the freedom of expression and privacy policies of 26 of the world's largest ICT companies. By reviewing internal RDR documents and interviewing relevant stakeholders (e.g. representatives from 11 companies and 14 civil society groups), Kogen concluded that RDR had clear influence and that indexes can be useful resources for social movements. In the case of FMTI, we expect that the Index brings attention to the disclosure practices of companies,

[3]URL to data is anonymized for submission.

making it easier for media, policymakers, investors, customers and the public to apply additional pressure that engenders greater transparency. Additionally, the Index provides clarity to companies by setting concrete targets and empowers employees within companies to push for greater transparency.

To reason about an index's impact over time, we also draw inspiration from Raji & Buolamwini (2019). In 2017, Buolamwini & Gebru (2018) demonstrated significant performance disparities across demographic groups in 3 face recognition systems (from IBM, Microsoft, and Megvii). A year later, Raji & Buolamwini (2019) audited five systems (IBM, Microsoft, Megvii, Amazon, Kairos): they found that the original 3 systems had reduced the performance disparities considerably, whereas the 2 new systems in 2018 showed large disparities comparable to those seen in the 3 systems from 2017. In §4.2, we similarly explore how foundation model developers fare in FMTI v1.1 when stratified by whether they were assessed in FMTI v1.0.

3 Methods

FMTI v1.1 involves four steps: indicator selection, developer selection, information gathering, and scoring. We describe these steps and how they relate to their implementation in FMTI v1.0 below.

3.1 Indicator selection

To concretize transparency, we use the 100 indicators from FMTI v1.0: Bommasani et al. (2023a) defined these indicators based on the literature regarding foundation models and AI. The 100 indicator are listed by name in Figure 8.[4] These indicators span three domains. First, 32 upstream indicators address transparency related to the ingredients and processes of model development, including data, compute, and labor. Second, 33 model indicators address transparency related to the properties and function of the model, including model access, capabilities, risks, and safety mitigations. Third, 35 downstream indicators address transparency related to the release and deployment of models, including usage policies, distribution, privacy protections, and impact. Prior work has strongly motivated the importance of each area of evaluated transparency, from labor (Gray & Suri, 2019a; Crawford, 2021; Hao & Seetharaman, 2023), data (Bender & Friedman, 2018; Gebru et al., 2018; Longpre et al., 2023b;a), compute (Lacoste et al., 2019; Schwartz et al., 2020; Patterson et al., 2021; Luccioni & Hernández-García, 2023), evaluation (Liang et al., 2023), safety (Cammarota et al., 2020; Longpre et al., 2024a), privacy (EU, 2016; Brown et al., 2022; Vipra & Myers West, 2023; Winograd, 2023), policies (Kumar et al., 2022; Weidinger et al., 2021; Brundage et al., 2020), and impact (Tabassi, 2023; Weidinger et al., 2023).

3.2 Developer selection

In FMTI v1.0, Bommasani et al. (2023a) selected 10 foundation model developers: all 10 were companies developing salient foundation models with consideration given for diversity (e.g. type of company, type of foundation model). Further, for each foundation model developer, Bommasani et al. (2023a) designated a *flagship* foundation model that was used as the basis for scoring the developer. In FMTI v1.1, we require companies to submit *transparency reports*[5]: we reached out to leadership at 19 companies: 01.AI, Adept, AI21 Labs, Aleph Alpha, Amazon, Anthropic, BigCode/Hugging Face/ ServiceNow, Cohere, Databricks, Google, IBM, Inflection, Meta, Microsoft, Mistral, OpenAI, Stability AI, Writer, and xAI.[6] 14 developers agreed to prepare reports and designated their flagship foundation model.[7]

[4]For full definitions, see Bommasani et al. (2023a, Appendix B).

[5]As we describe later, companies submitted an initial report that was modified through the FMTI v1.1 process. The final report that we publish is validated by the company but, therefore, different from this initial report. For brevity, we refer to both as transparency reports.

[6]FMTI v1.0 evaluated BigScience/Hugging Face, which together developed the BLOOMZ model; in FMTI v1.1 we evaluate BigCode/Hugging Face/ServiceNow, which together developed the StarCoder model. Throughout this paper we refer to BigCode/Hugging Face/ServiceNow as a single entity (i.e. the developer of StarCoder), though Hugging Face and ServiceNow are companies while BigCode is "an open scientific collaboration working on the responsible development and use of large language models for code" supported by ServiceNow and Hugging Face. See `https://www.bigcode-project.org/docs/about/mission/`.

[7]We provided guidance that the flagship foundation model should be "based on a combination of the following factors: greatest resource expenditure, most advanced capabilities, and greatest societal impact."

Name	Flagship Model	Release	Input	Output	Status	Headquarters	WH1	WH2	FMF	AIA
Adept	Fuyu-8B	Open weights	T, I	T	Startup	USA				
AI21 Labs †	Jurassic-2	API	T	T	Startup	Israel				
Aleph Alpha	Luminous Supreme	API	T, I	T	Startup	Germany				
Amazon †	Titan Text Express	API	T	T	Big Tech	USA	✓			
Anthropic †	Claude 3	API	T, I	T	Startup	USA	✓		✓	
BC/HF†/SN	StarCoder	Open weights	T	T	Startup	USA				✓
Google †	Gemini 1.0 Ultra API	API	T, I, A, V	T, I	Big Tech	USA	✓		✓	
IBM	Granite	API	T	T	Big Tech	USA		✓		✓
Meta †	Llama 2 70B	Open weights	T	T	Big Tech	USA	✓			✓
Microsoft	Phi-2	Open weights	T	T	Big Tech	USA	✓		✓	
Mistral	Mistral 7B	Open weights	T	T	Startup	France				
OpenAI †	GPT-4	API	T, I	T	Startup	USA	✓		✓	
Stability AI†	Stable Video Diffusion	Open weights	T	V	Startup	UK		✓		✓
Writer	Palmyra-X	API	T	T	Startup	USA				

Table 1: **Selected Foundation Model Developers.** Information on the 14 selected foundation model developers: the developer name, its flagship model, the release strategy for the model, the model's input and output modalities, the developer's corporate status, and the developer's headquarters. BC/HF/SN abbreviates BigCode/Hugging Face/ServiceNow. T, I, A, and V abbreviate text, image, audio, and video as modalities, respectively. † indicates the developer was evaluated in FMTI v1.0. We also indicate if developers were involved in the White House's voluntary commitments for the management of risks posed by AI announced in July 2023 (WH1), and commitments by additional organizations in the same areas announced in September 2023 (WH2) as well as if they are founding member of the Frontier Model Forum (FMF) or AI Alliance (AIA). Concurrent with the release of FMTI v1.1, as of May 20, 2024, Amazon and Meta have joined the FMF.

Table 1 describes the developers and their flagship foundation models. 8 of the 14 developers are hold-overs from FMTI v1.0.[8] Three developers are assessed for the same models as v1.0 (Jurassic-2 for AI21 Labs, Llama 2 for Meta, GPT-4 for OpenAI), whereas five are assessed for new models (Titan Text Express for Amazon, Claude 3 for Anthropic, StarCoder for BigCode/HuggingFace/ServiceNow, Gemini 1.0 Ultra API for Google, and Stable Video Diffusion for Stability AI). The six new developers and their models are Fuyu-8B for Adept, Luminous Supreme for Aleph Alpha, Granite for IBM, Phi-2 for Microsoft, Mistral 7B for Mistral, and Palmyra-X for Writer. In aggregate, the FMTI v1.1 composition has broader geographic coverage (e.g. from 0 to 2 companies headquartered in the European Union), broader modality coverage (e.g. Gemini takes audio and video as input), and a more even balance of open and closed foundation models (i.e. from 3 of 10 open models in v1.0 to 6 of 14 open models in v1.1).

3.3 Information gathering

In FMTI v1.0, Bommasani et al. (2023a) identified publicly-available sources of information for each developer through a systematic protocol for searching the Internet, which provided the information for all scoring decisions. This approach has four potentially undesirable properties. First, given that information is decentralized across the Internet, the researchers may have missed information.[9] Second, the relationship between a piece of public information and an indicator may be indirect and oblique, leading to greater subjectivity in scoring. Third, focusing on public information aligns with a developer's current level of transparency but does not provide developers with an opportunity to disclose further information. Finally, and most fundamentally, this search significantly adds to the cost of executing the index.

In FMTI v1.1, we request transparency reports from each developer that directly address each of the 100 indicators. This change in the information gathering process alters the dynamics for the four aforementioned considerations. First, if we assume developers are strongly incentivized to be their own best advocates and are certainly the most knowledgeable entities about their models, then the information they compile should be complete. Second, by having developers directly clarify information on indicators affirmatively,

[8] Cohere and Inflection declined to participate in FMTI v1.1.

[9] However, this does beg the question of whether the information being public is truly constitutive of transparency if it is not discovered through a systematic and high-effort search.

uncertainties that contributed to more subjective scoring are addressed. Third, by allowing developers to include information that was not-previously public, which is made public through this process, opportunities arise for greater transparency. Finally, by having developers gather information, the cost we bear is reduced.

3.4 Scoring

In FMTI v1.0, once information was identified, two researchers independently scored each of the 1000 (indicator, developer) pairs. The agreement rate was 85.2% (148 disagreements): in the event of disagreement, the researchers discussed and came to agreement. These initial scores were sent to developers to permit rebuttal: following a two-week rebuttal process, final scores were published in October 2023.

For FMTI v1.1, using the information identified through the developer-submitted transparency reports, two researchers independently scored each of the 1400 (indicator, developer) pairs. The standard of each indicator is the same as in FMTI v1.0: see Bommasani et al. (2023a, Appendix B) for the per-indicator scoring standard. The agreement rate was 85.3% (206 disagreements): in the event of disagreement, the researchers discussed and came to agreement. These initial scores were sent to developers to permit rebuttal: in contrast to FMTI v1.0, the rebuttal process was a more iterative multi-week process that involved email exchanges and video meetings. Through this correspondence, companies clarified existing information and disclosed new information, which is reflected in the final form of the transparency reports we publish. Ultimately, companies validated their transparency reports and approved their release. In doing so, unlike FMTI v1.0, these reports make explicit instances where (i) developers disclose useful information yet (ii) we felt the disclosure was insufficient to award a point.

3.5 Timeline

We summarize the execution of FMTI v1.1 as follows:

1. *Developer solicitation (December 2023 – January 2024).* We contacted leadership at 19 companies developing foundation models, requesting they submit transparency reports.
2. *Developer reporting (February 2024).* 14 developers designated their flagship foundation model and submitted transparency reports in relation to each of the 100 transparency indicators for their flagship model.
3. *Initial scoring (March 2024).* We reviewed the developers' reports, ensuring a consistent horizontal standard across all developers in terms of how each indicator was scored.
4. *Developer response (April 2024).* We returned the scored reports to developers, who then contested scores on specific indicators (potentially including the disclosure of additional information to justify a different score). Following this process, we finalized the transparency reports, which were validated by the developers prior to public release.

4 Results

We analyze this iteration of the Index on its own (§4.1), in relation to the first iteration (§4.2), and specifically in terms of new disclosures (§4.3).

4.1 Standalone results of FMTI v1.1

While the average score on the Index has significant room for improvement, there is high variability among developers. Based on the overall scores (right of Figure 1), 11 of the 14 developers score below 65, showing that there is a significant lack of transparency in the foundation model ecosystem and substantial room for improvement across developers. The mean and median are 57.93 and 57 respectively, with a standard deviation of 13.98. The highest-scoring developer scores points for 85 of the 100 indicators, while the lowest-scoring developer scores 33. The 3 top-scoring developers (BigCode/Hugging Face/ServiceNow,

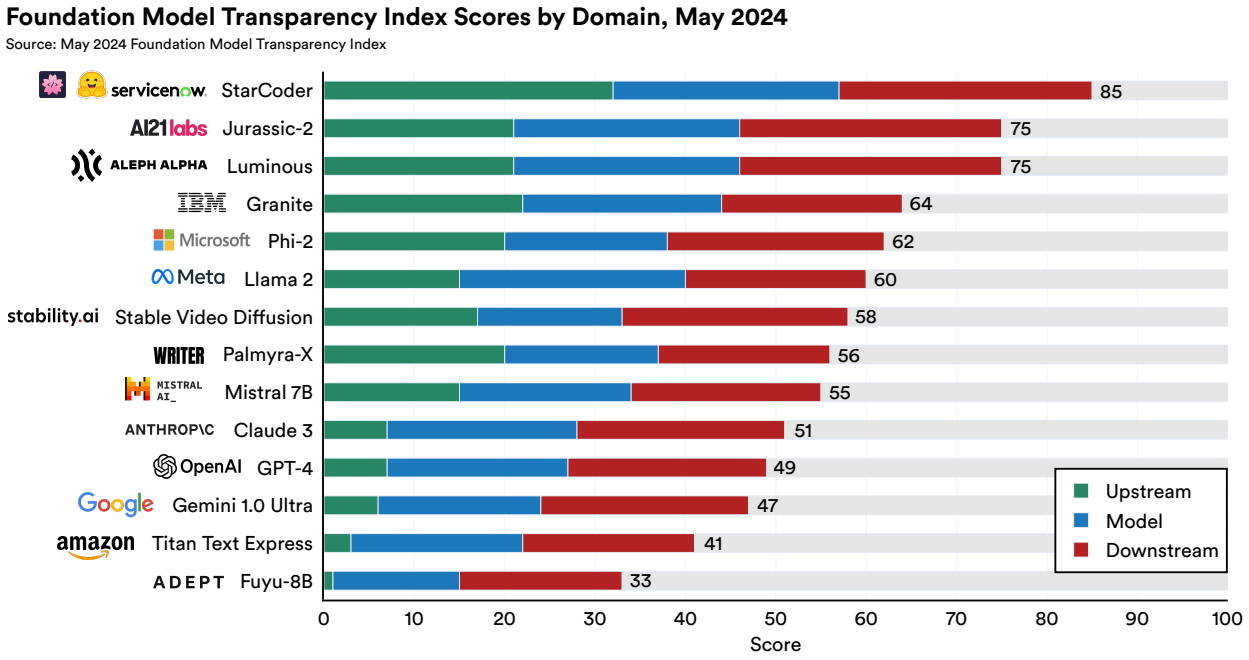

Figure 1: **Scores by Domain.** The overall scores disaggregated into the three domains: upstream, model, and downstream.

Aleph Alpha, and AI21 Labs) are more than one standard deviation above the mean, the next 9 are near the mean (IBM, Microsoft, Meta, Stability AI, Writer, Anthropic, OpenAI, Mistral, Google), and the 2 lowest-scoring developers are more than one standard deviation below the mean (Amazon, Adept).

Improvement is feasible for each developer. In spite of significant opacity, for 96 of the 100 indicators there exists some developer that scores points, and of these there are 89 where multiple developers score points. The disclosures that developers make to satisfy these indicators provide a concrete example of how all developers can be more transparent. If developers emulate the most-transparent developer for each indicator, overall transparency would improve sharply.

Developers disclose significant new information, which contributes to their scores. A developer's total score on the Foundation Model Transparency Index reflects the information that it discloses about its flagship foundation model in relation to each of the 100 indicators in the Index. In this version of the Index, we note where this information is disclosed as part of the process of conducting the Index (i.e. where developers disclosed the information for the first time via their report, or where developers updated their documentation in response to the Index process). This new information contributes to developer's overall scores: developers on average release new information related to 16.6 indicators, which improves the average score by 14.2 points. For example, AI21 Labs and Writer newly disclosed the carbon emissions associated with building their flagship foundation models (200-300 tCO2eq and 207 tCO2eq respectively). See §4.3 for further details.

The upstream domain is the most opaque and the region where the most transparent developers distinguish themselves. Breaking the results down by domain (Figure 1), developers performed best on indicators in the downstream domain, scoring 65% of available points overall in comparison to 61% on the model domain and 46% in the upstream domain. Developers scored worse across upstream indicators: of the 20 indicators where developers score highest, just one indicator (model objectives) is in the upstream domain. High-scoring developers often differentiate themselves in terms of their transparency in the upstream domain; the top scorer overall, BigCode/Hugging Face/ServiceNow, scored all 32 points, whereas the lowest scorer overall, Adept, scored just one point. The standard deviation of scores on the upstream domain (8.8) is more than double that of the other two domains (3.6 each), reflecting the much greater spread across the

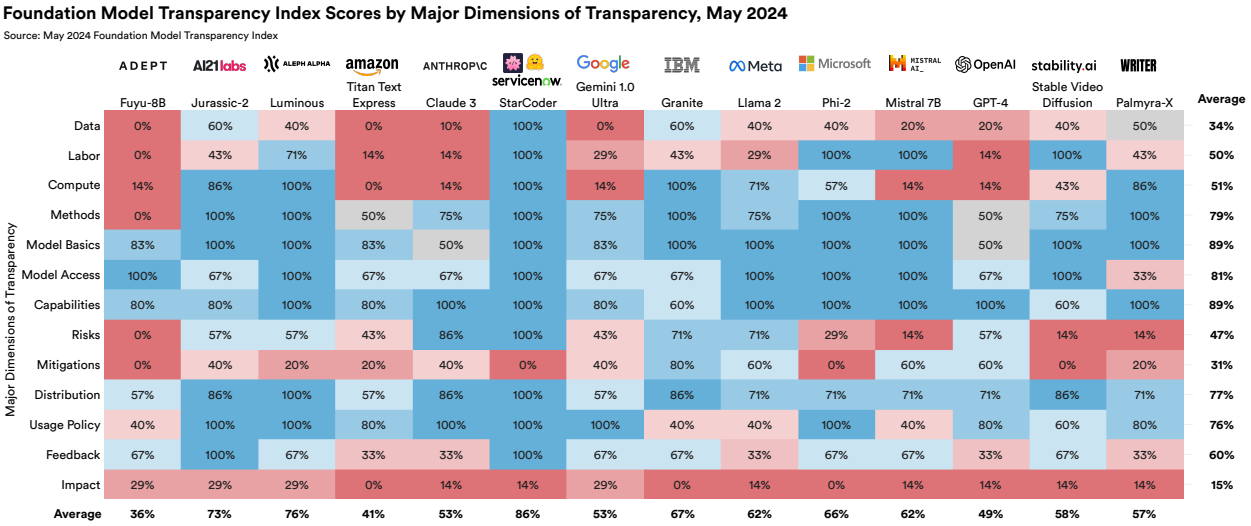

Figure 2: **Scores by Major Dimensions of Transparency.** The fraction of achieved indicators in each of the 13 major dimension of transparency. Major dimension of transparency are large subdomains within the 23 subdomains.

domain. On the whole, developers are less transparent about the data, labor, and compute used to build their models than how they evaluate or distribute their models. Prior work has emphasized the importance of these particularly opaque domains (Crawford, 2021; Gebru et al., 2018; Luccioni & Hernández-García, 2023).

Scores varied across subdomains, with developers scoring best on user interface, capabilities, and model basics. Disaggregating each domain, we consider the 23 subdomains with Figure 2 showing results for 13 major subdomains. Scores varied greatly across subdomains: 86 percentage points separate the average scores on the most transparent and least transparent subdomains. The subdomains with the highest scores are user interface (93%) capabilities (89%), model basics (89%), documentation for downstream deployers (89%), and user data protection (88%). Each of these subdomains is in the downstream domain, where developers scored near or above 70% on 6 of the 9 subdomains. These high scores were achieved in part through the release of new information. For instance, AI21 Labs released its first model card for Jurassic-2 during the FMTI v1.1 process, and Aleph Alpha and Stability AI made significant changes to their existing model cards. High scores on these subdomains reflects that disclosure in these areas is relatively less onerous for developers—disclosing documentation for downstream deployers as well as information about capabilities and model basics makes it easier and more appealing for customers to make use of a companies' flagship foundation model, meaning it is in developers' interest to do so.

Data access, impact, and trustworthiness are the least transparent subdomains. The subdomains with the lowest total scores are data access (7%), impact (15%), trustworthiness (29%), and model mitigations (31%). Developers score 50% or less on 10 of 23 subdomains in the index, including 3 of the 5 largest subdomains—impact (15%), data (34%), and data labor (50%). The lack of transparency in these subdomains shows that the foundation model ecosystem is still quite opaque—there is little information about how people use foundation models, what data is used to build foundation models, and whether foundation models are trustworthy.

Open developers match closed developers on downstream indicators, and exceed them on upstream indicators. Developers adopt different release strategies (Solaiman, 2023) for their flagship foundation models: six developers release open foundation models (Kapoor et al., 2024), meaning the model weights are widely available, whereas the other eight employ a more closed release strategy.[10] Open developers

[10]Aleph Alpha shares model weights for its flagship foundation model with some customers on premises, but this does not mean that the model weights are widely available and so it is considered a closed foundation model developer per the definition of (Kapoor et al., 2024).

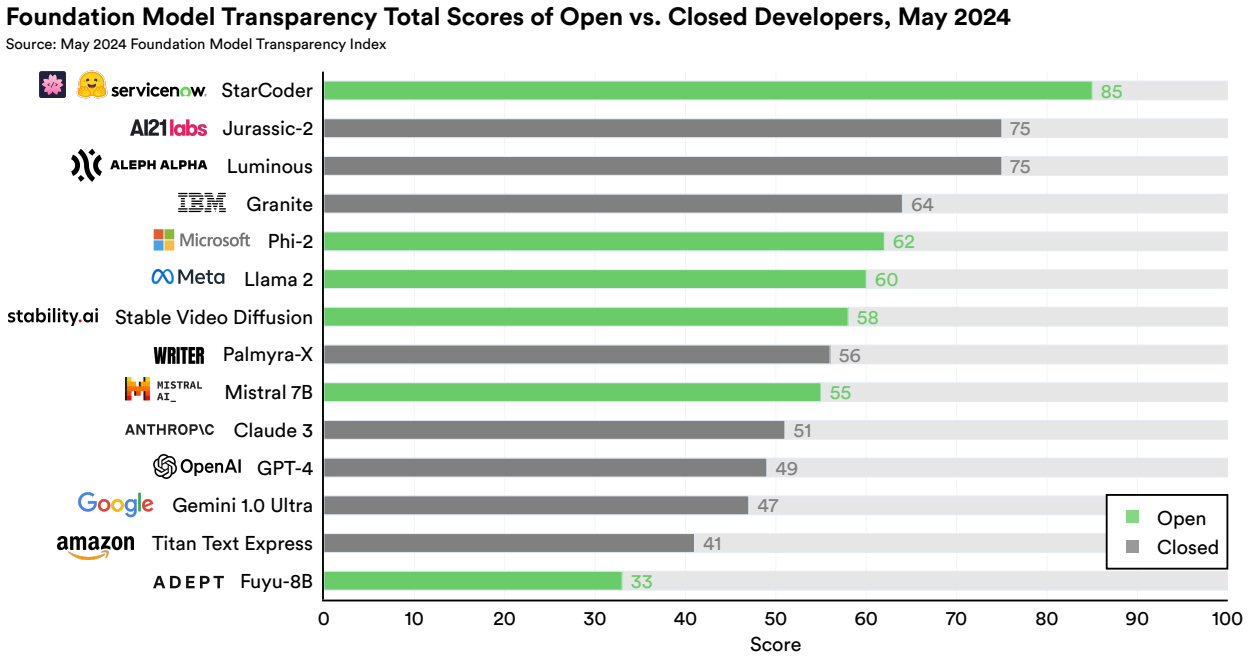

Figure 3: **Overall Scores by Release Strategy.** The overall scores for the 6 open developers (Adept, BigCode/Hugging Face/ServiceNow, Meta, Microsoft, Mistral, Stability AI) and the 8 closed developers (AI21 Labs, Aleph Alpha, Amazon, Anthropic, Google, IBM, OpenAI, Writer).

generally outperform closed developers (Figure 3: the median open developer scores 5.5 points higher than the median closed developer. While making the weights of a model openly available is correlated with greater overall transparency, it does not necessarily imply greater transparency about matters like data, compute, or usage.

The difference in transparency between open and closed developers is attributable to the substantial gap in upstream transparency: within the upstream domain, the median open developer scores 3 additional points on indicators in the upstream subdomain over the median closed developer. Within each subdomain, the median open developer scores as many or more points than the median closed developer on all but 5 of the 23 subdomains (i.e., risks, model mitigations, trustworthiness, distribution, usage policy, and model behavior policy). Open developers outscore closed developers by the widest margin on the following subdomains: data labor, data, and model access. Though open developers may struggle to gauge the downstream use of their models (Klyman, 2024), which closed developers may be able to directly monitor, the median open developer scores the same number of points on the downstream domain as the median closed developer.

Developers that are part of the AI Alliance (Hugging Face, IBM, Meta, ServiceNow, Stability AI), which often advocates for open releases of model weights, outscore developers that are founding members of the Frontier Model Forum (Anthropic, Google, Microsoft, OpenAI) by a median of 12 points, with higher average scores across upstream, model, and downstream indicators.[11]

Still, closed developers outperform open developers in several areas related to policies and enforcement. Closed developers generally share more information about if and how they enforce their policies regarding user and model behavior, outperforming open developers on these subdomains by 2 and 1 points respectively.[12] Closed developers also score higher on the risks and model mitigations subdomains, as several open developers do not

[11]The AI Alliance is a coalition of developers, universities, and researchers "who collaborate to advance safe, responsible AI rooted in open innovation." The Frontier Model Forum is an industry group whose founding members are Anthropic, Google, Microsoft, and OpenAI; the Frontier Model Forum committed to "advancing AI safety research," "identifying best practices," "collaborating across sectors," and "help[ing] AI meet society's greatest challenges." See `https://thealliance.ai/` and `https://www.frontiermodelforum.org/`.

[12]As with some other subdomains, the difference in scores here is driven by just one or two of the 14 developers.

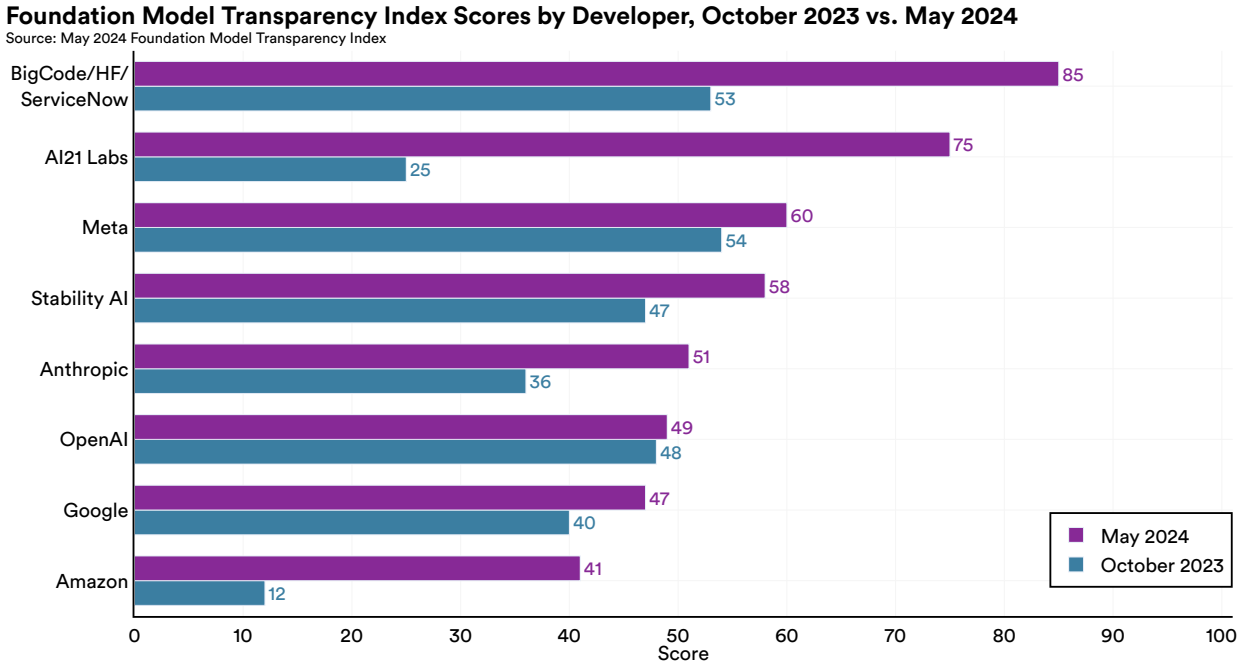

Figure 4: **Change in Overall Scores.** The FMTI v1.0 and v1.1 overall scores for the eight developers assessed in both versions.

describe or demonstrate risks associated with their flagship foundation model and closed developers are more likely to describe and demonstrate any risk mitigations that are taken at the model level.[13] The discrepancy in transparency between open and closed foundation model developers is a reflection of the current state of the ecosystem, not a fundamental reality about the transparency of developers that do or do not make the weights of their foundation models widely available.

4.2 Comparative results between FMTI v1.1 and v1.0

Transparency increased across the board from v1.0 to v1.1. Foundation model developers significantly improved their scores between October 2023 and May 2024, with the average score rising from 37 to 58 out of 100. Scores improved on every domain, with average upstream scores improving by the greatest margin (+7.6 points) followed by downstream (+7.2) and model (+6.1). As a result, there is significantly more information publicly available about the upstream resources developers use to build foundation models, the models themselves and how they are evaluated, and their downstream distribution and use.

Transparency increased in nearly all subdomains from v1.0 to v1.1. Developers improved their scores on every subdomain with the exception of data access. The largest improvements in subdomain scores were in compute (average increase of +2.4 indicators per developer), data labor (+2.3), and risks (+1.6). This broad improvement demonstrates that the overall trend in recent years toward reduced transparency is more nuanced than is commonly understood, though transparency is still lacking with respect to the data used to build foundation models and the impact they have once deployed.

Transparency increased in subdomains that were especially opaque in v1.0. Several of the areas of the index that were least transparent in v1.0 show significant improvement in v1.1, including subdomains such as compute, methods, risks, and usage policy. For example, the total score across companies for the compute subdomain rose from 17% in v1.0 to 51% in v1.1. Compute is potentially one of the most intractable areas for disclosure as it relates to the environmental impact of building foundation models—and therefore could be associated with additional legal liability for developers and deployers—yet we see improvement

[13]Open developers, however, outscore closed developers on data mitigation indicators. They often do not score points on model mitigation indicators because they do not explicitly state that no such mitigations were applied at the model level.

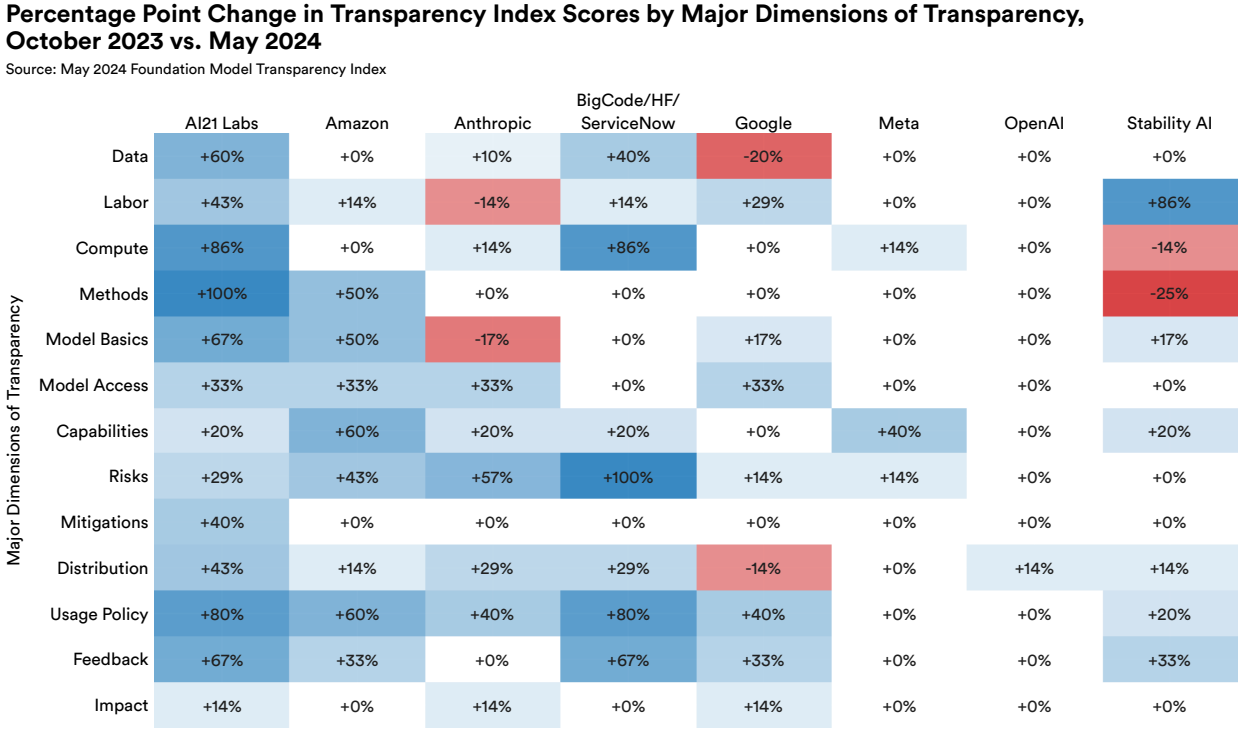

Figure 5: **Change in Subdomain Scores From FMTI v1.0 to FMTI v1.1.** This figure shows the percentage point change in scores for major subdomains for the eight developers that are included in both the October 2023 and May 2024 versions of the Foundation Model Transparency Index.

across compute indicators. This improvement is driven to a significant degree by new information that companies have disclosed, with companies disclosing new information on compute usage (6 companies disclosed new information), development duration (4), energy usage (4), compute hardware (3), hardware owner (3), and carbon emissions (2). In this way, transparency about compute expenditure has spillover effects for transparency about environmental impact, providing a potential model for translating information disclosure about one area into details about the impact of the foundation model supply chain.

Transparency has improved substantially with respect to companies' policies regarding acceptable use and behavior of their models. The total score across companies for the usage policy subdomain rose from to 44% to 57%, which was driven by increased transparency related to how these policies are enforced. Similarly, the total score across companies for the model behavior policy subdomain rose from 23% to 69%. While increased transparency in these domains is a relatively light lift—as companies merely need to state whether they have such policies and, if so, describe how they are enforced—this form of transparency is not costless for companies as it highlights potential failure modes for their models (Klyman, 2024). Transparency about how companies restrict the use and behavior of their models can facilitate independent evaluation and red teaming due to reduced legal uncertainty, promoting research on safety and trustworthiness (Longpre et al., 2024a).

Data remains a key area of opacity. Several areas of the Index exhibit sustained and systemic opacity: almost all developers remain opaque on these matters. Developers display a fundamental lack of transparency with respect to data, building on frustrations of data documentation debt (Bandy & Vincent, 2021; Sambasivan et al., 2021). Transparency on the data subdomain rose from 20% to 34% from v1.0 to v1.1, but BigCode/Hugging Face/ServiceNow is the only developer that scores points on indicators relating to data creators, data copyright status, data license status, and personal information in data. Only 3 developers (Aleph Alpha, BigCode/Hugging Face/ServiceNow, IBM) score points on 6 or more of the 10 data indicators, while 6 developers score 2 or fewer points. These low scores reflect the ongoing crisis in data provenance

(Longpre et al., 2024b; 2023a), wherein companies share no information about the license status of their datasets, preventing downstream developers from ensuring they are complying with such licenses.

While scores on data labor improved from 17% to 50% from v1.0 to v1.1, this was driven in large part by an increase in the number of companies that do not use data labor outside of their own organization (BigCode/Hugging Face/ServiceNow, Microsoft, Mistral). Considering only the 11 developers who do not disclose that they do not use external data labor, scores on the data labor subdomain fall to 36%, which would make it the sixth lowest scoring subdomain. Only one developer (Stability AI) that discloses its use of external data labor discloses the wages that it pays data laborers, highlighting how a lack of transparency may limit accountability for worker exploitation (Gray & Suri, 2019b).

Transparency in data access, one of the lowest scoring subdomains across developers in v1.0, declined in v1.1 from 20% to 7%. This reflects the significant legal risks that companies face associated with disclosure of the data they use to build foundation models. In particular, these companies may face liability if the data contains copyrighted, private, or illegal content such as child sexual abuse material (Lee et al., 2024; Solove, 2024; Thorn, 2024).

Developers disclose little information about the real-world impact of their foundation models. Developers still lack transparency about the real-world impact of their models, and any steps they take to mitigate negative impacts pre-deployment. Of the major dimensions of transparency in Figure 2, developers score worst on the impact subdomain (as they did in v1.0). Each of the four indicators where no developer scores points (affected market sectors, affected individuals, usage reports, and geographic statistics) are in the impact subdomain. This means that the public has little to no information about who uses foundation models, where foundation models are used, and for what purpose. The lack of transparency regarding these matters inhibits effective governance of foundation models, as it is difficult for governments or civil society organizations to pressure companies to responsibly deploy their models if there is no information about the impact of deployment. In many cases this opacity stems from a lack of information sharing between developers, deployers, and customers; developers generally do not know how their foundation model is being used unless a deployer monitors use or receives and shares information about use from its customer.

The 8 developers evaluated in both October 2023 and May 2024 showed marked improvement, or 19 points on average. For 3 of these developers we evaluate the same flagship foundation model (Jurassic-2 for AI21 Labs, Llama 2 for Meta, and GPT-4 for OpenAI) as FMTI v1.0 while for the other 5 we evaluate a different flagship foundation model (Titan Text Express for Amazon, Claude 3 for Anthropic, StarCoder for BigCode/Hugging Face/ServiceNow, Gemini 1.0 Ultra API for Google, and Stable Video Diffusion for Stability AI). All 8 companies' scores increased, as shown in (Figure 4): AI21 Labs (+50), BigCode/Hugging Face/ServiceNow (+32) and Amazon (+29) made the largest improvements.

Developers that were assessed only in v1.1 performed slightly worse than those assessed in both v1.0 and v1.1. The 6 developers that were assessed only in v1.1 (Adept, Aleph Alpha, IBM, Microsoft, Mistral, Writer) have slightly lower scores than those assessed in both v1.0 and v1.1. Their mean total score is 57.5, which is 1 point lower than that of the other 8 developers.[14] Developers that were included in FMTI v1.0 had the benefit of having already evaluated their own transparency practices in relation to this initiative (i.e. in FMTI v1.0 they had the opportunity to rebut scores provided by Bommasani et al. (2023a), meaning it may have been relatively easier for them to compile transparency reports. The 6 developers assessed only in FMTI v1.1 include the lowest scoring developer and the two lowest scoring open model developers.

4.3 New information in FMTI v1.1

A key feature of the FMTI v1.1 methodology is that companies were able to disclose *new information*, meaning information that was not public at the onset of the FMTI v1.1 process. In some cases, this information is directly made public for the first time via the FMTI v1.1 transparency reports. In other cases, this information was incorporated into preexisting or new publicly available documentation from companies. For

[14]This is despite the fact that 4 of the 6 developers that are assessed only in FMTI v1.1 are open developers, which score higher on average.

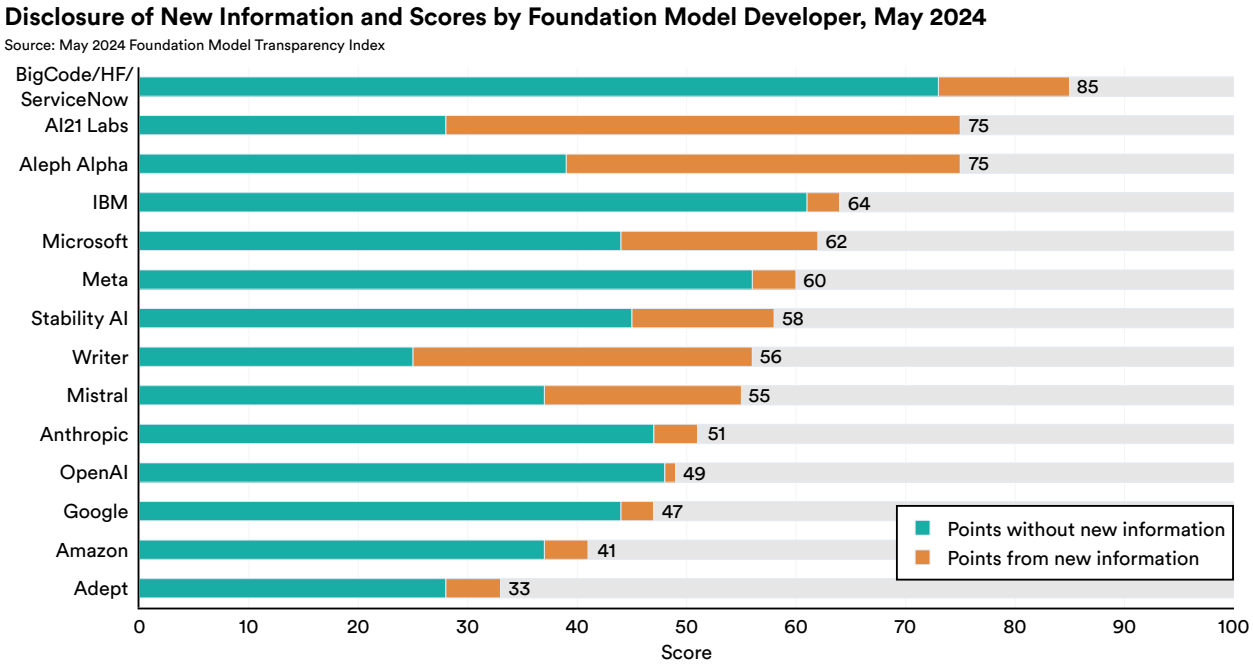

Figure 6: **Scores by New Information Status.** The overall scores disaggregated based on whether the information was newly disclosed.

example, as we noted previously, AI21 Labs released the first model card for Jurassic-2 and Stability AI significantly updated the model card for Stable Video Diffusion.

New information constitutes a large fraction of the score for several companies. Figure 6 breaks down each developer's overall FMTI v1.1 scored based on which indicators were awarded for new information vs. information that was previously publicly available. For three developers (AI21 Labs, Aleph Alpha, Writer), new information constitutes roughly half the points awarded. In the case of Aleph Alpha, several new disclosures are made about data labor: all laborers are employed by Aleph Alpha, work in Germany, and are afforded labor protections as stipulated by German law. For Writer, new information is provided on compute: models are trained on 1024 NVIDIA A100 80GB GPUs for 74 days (910k GPU hours) on the Writer cluster, amounting to 8.2×10^{23} FLOPs in compute, 812 MWh in energy, and 207tCO2eq in emissions.

All developers disclose some new information. In the case of OpenAI, the sole change to its disclosures from FMTI v1.0 is in relation to detecting machine-generated content. Specifically, OpenAI clarify that it "originally released a classifier that was taken down due to lack of accuracy. Our commitments are around audio / visual content for now, so this implies lack of ability to detect GPT-4 generated content". In other cases, new information is disclosed to clarify existing information that was difficult to interpret based on publicly available documentation. For example, Amazon clarified how its model behavior policy and usage policy interoperate: "In the Bedrock user guide, we stated that AWS is committed to the responsible use of AI, and we use an automated abuse detection mechanisms to identify potential violations, we may request information about customers' use of Amazon Bedrock and compliance with our terms of service or a third-party provider's AUP. In the event that a customer is unwilling or unable to comply with these terms or policies, AWS may suspend access to Amazon Bedrock."

New information drives much of the transparency for areas of large improvement. Figure 7 totals the amount of new information across all developers for each major dimension of transparency. The most information is in the area of labor, which we noted previously is largely due to multiple companies clarifying they do not use data labor in building their flagship foundation models. Beyond this, other areas of large improvement from FMTI v1.0 to FMTI v1.1 are precisely those with large amounts of new information. Namely, more than 10% of the 233 pieces of new information are in the areas of compute and usage policy.

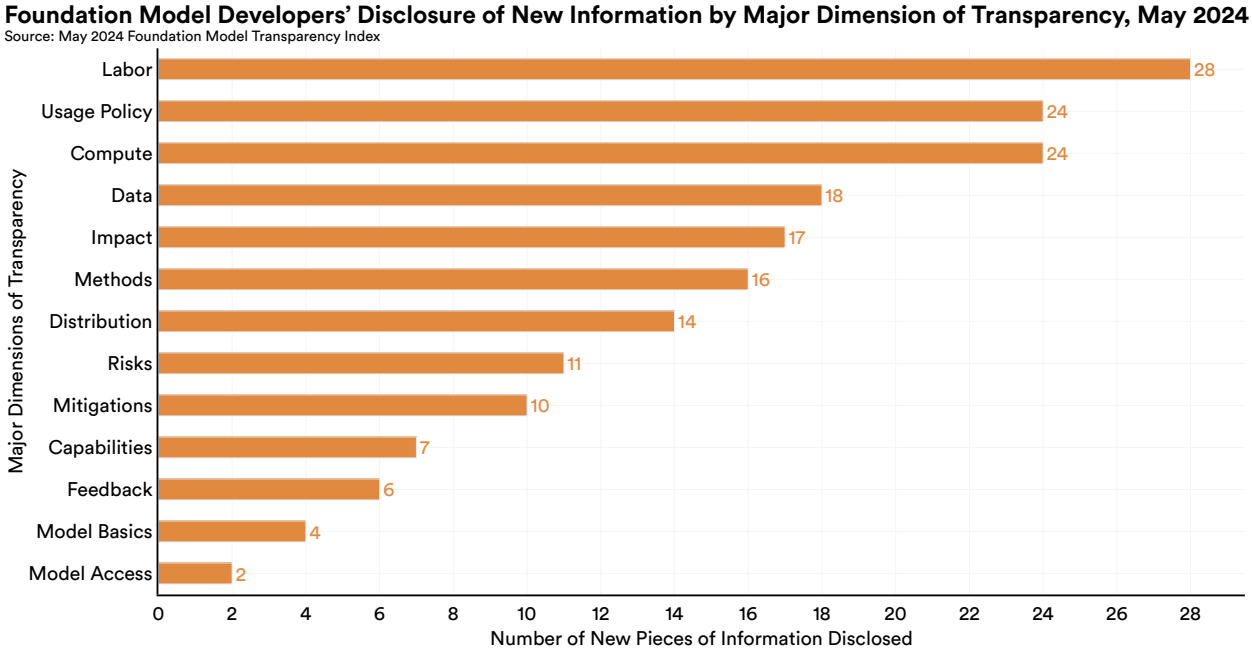

Figure 7: **Aggregate New Information by Major Dimensions of Transparency.** The number of pieces of new information, aggregated across all developers, for each of the 13 major domains of transparency. Note: major dimensions of transparency each have a different numbers of indicators—the largest is data (10 indicators), followed by compute (7), distribution (7), impact (7), labor (7), risks (7), model basics (6), capabilities (5), mitigations (5), usage policy (5), feedback (3), methods (3), and model access (3).

Companies disclose new information for which they do not score points. For example, several companies provide additional details about their data labor practices like general information about instructions to annotators and assurances that wages exceed local minimum wage. And on the indicator about data creators, which is awarded to only one of the 14 developers (BigCode/Hugging Face/ServiceNow), AI21 Labs discloses that "Most of the internet-connected population is from industrialized countries, wealthy, younger, and male, and is predominantly based in the United States." While these disclosures are insufficient for the stated standards for the associated indicators, we nonetheless emphasize that it may still be valuable for developers to make such information public. In a few cases, companies provide public justifications for why they do not disclose certain information. Most notably, Aleph Alpha states that "In line with the German copyright act ('Urheberrechtsgesetz'), data used for training is to be deleted after use and therefore can not be made available or distributed to external parties" in relation to the indicators about data access.

5 Discussion

Having characterized how transparency has changed from October 2023 to April 2024, we discuss how we reason about these changes (§5.1), what we recommend going forward (§5.2), and ways in which the Foundation Model Transparency Index could evolve in subsequent iterations (§5.3).

5.1 Interpretation of findings

There is significant room for improvement in the transparency of foundation model developers. Developers on average score just 58 out of 100, with major gaps in multiple subdomains for most developers. Developers' transparency reports are incomplete, lacking disclosures on many important matters.

Nevertheless, our findings provide significant reasons for optimism about the prospects for improved transparency. Foundation model developers shared a significant amount of new information about how they build,

evaluate, and deploy their models via FMTI v1.1. Some of the areas of the index that were least transparent in v1.0 show significant improvement in v1.1, including subdomains such as compute, methods, risks, and usage policy. Several companies have become much more transparent, with some releasing model cards or other documentation for their flagship foundation models for the first time. By engaging directly with foundation model developers, we have shown that many firms are willing to disclose more information about some of their most powerful technologies.

Still, the overall state of transparency in the foundation model ecosystem remains poor. Developers are opaque about the data, labor, and compute used to build their models, they often do not release reproducible evaluations of risks or mitigations, and they do not share information about the impact their models are having on users or market sectors. Transparency in the foundation model ecosystem has been declining for several years, and this trend is unlikely to reverse in the near term.

Where developers do disclose information, they sometimes disclose information for only the least onerous indicators within a subdomain. For example, in the risks subdomain, most developers describe risks (12 of 14 developers) and many demonstrate risks (8 of 14). However, fewer developers evaluate unintentional harm (5 of 14) or intentional harm (4 of 14). We see the same trends for model mitigations and data labor, where developers disclose information regarding less intensive indicators but no others.

There are a variety of different reasons why a developer might not disclose information related to a specific indicator. Developers could face legal exposure if they disclose a substantial amount of information related to some indicators, as is the case with data. Some developers argue that disclosure of certain information could amount to ceding a developer's edge to its competitors. The process of releasing information also presents a potential coordination problem for large developers that need to consult with lawyers, engineers, product managers, and executives before doing so. There were many instances in which it appeared that a developer did not fully understand an indicator and so did not disclose the information needed to satisfy that indicator. In other cases, developers noted that they do not have access to the information in question as it is collected only by deployers or end users. Our results show that these and other factors combine to limit the overall transparency of foundation model developers.

5.2 Recommendations

We present recommendations aimed at different stakeholders based on the findings of FMTI v1.1 as well as changes in the world over the past six months, building on a more extensive set of recommendations made in Bommasani et al. (2023a, §8).

5.2.1 Foundation model developers

As part of FMTI v1.1, we publish transparency reports that consolidate information disclosures from developers and that we release subject to their validation. In light of voluntary codes of conduct promulgated by the White House and the G7 that include commitments for foundation model developers to release transparency reports, we envisage the reports we release as part of FMTI v1.1 as rudimentary forms of such transparency reports (Bommasani et al., 2024). In addition, many of the areas where transparency is lacking in foundation models mirror those in previous waves of digital technology as well as in other industries such as finance and healthcare. For example, shortcomings in downstream transparency by foundation model developers mirror issues faced by social media companies in the last decade (Aspen Institute, 2021). For such indicators, investing in the expertise of trust and safety professionals could help inform how developers develop internal and external policies, including those related to transparency. As one example, social media platforms like Facebook, YouTube, and TikTok detail usage policy violations and government requests for user data in their transparency reports (Narayanan & Kapoor, 2023). Similarly, social media platforms often have well-established processes for communicating with users about account restrictions and handling user appeals (Trust and Safety Professional Association, 2023). Foundation model developers can adopt these policies and processes to improve their downstream transparency (Klyman, 2024).

5.2.2 Customers of foundation models

Purchasers (e.g. downstream developers, enterprise users of chatbots, or government entities) can exert negotiating power in procurement to increase transparency. Notably, some foundation model developers stated that their participation was driven by requests from customers to understand the transparency of their products. Others mentioned that their practices related to transparency with their customers are much better compared to the data they can share publicly—an example of the influence customers can have on business practices. The Foundation Model Transparency Index provides a structured way for customers to advance transparency from foundation model developers—both in terms of the information developers share with their clients and with the broader public.

In addition, governments can play a dual role as customers of foundation models (Quay-de la Vallee et al., 2024). As influential (and lucrative) procurers of technology, this can help them play a standard-setting role. For example, the U.S. government is one of the largest purchasers of various goods and services, which allows it to set the standards for how these goods and services are sold, shaping business practices across industries (Vinsel, 2019), including around transparency. While requirements on transparency by governments may lead to legal concerns around government overreach, such as concerns in the US related to the first amendment implications of compelled speech (Bankston & Hodges, 2024), standard-setting via procurement circumvents these concerns by relying solely on the voluntary commitment to these standards by model developers who want to enter a contract.

5.2.3 Transparency advocates

The transparency reports we release can enable transparency advocates in academic and civil society organizations to better understand developer practices. While for the purposes of the Index scores we set a threshold for constitutes sufficient disclosure to award a point, the underlying disclosures are considerably richer. We encourage researchers and journalists to investigate this information, which includes considerable variation across companies that we do not explore in this paper.

5.2.4 Policymakers

Policymakers can use our results to identify areas of pervasive opacity—including areas with sustained opacity (across FMTI v1.0 and v1.1) as well as areas with systematic opacity (across the developers we score). This can also highlight perverse business incentives that might lead to such a lack of transparency, and in turn, can inform regulation that addresses them. For example, sharing information about the data used to train foundation models might open up companies to liability concerns, such as due to the legal uncertainty around copyright violations. Regulation intended to address such perverse incentives, such as mandatory disclosure of training data, could help address these impediments to transparency.

Various policy efforts in the last two years have focused on addressing transparency in the foundation model ecosystem, including Canada's Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems, the EU AI Act, and the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Still, Bommasani et al. (2024) find that these efforts can lack specificity. Our iterative process with model developers provides an insight into how such specificity might arise in practice—government bodies might consider developing the institutional capacity to coordinate discussions with companies to ensure that they adhere to the standard required by a certain policy.

5.3 Next steps

We plan to conduct future versions of the Foundation Model Transparency Index on a regular basis. As stated by Bommasani et al. (2023c), as part of future iterations we will make changes to the Foundation Model Transparency Index to reflect changes in the foundation model ecosystem and the organizations that build and deploy foundation models.

For example, moving forward, the Foundation Model Transparency Index may change the indicators or their thresholds. For the purposes of clear comparisons, the indicators and scoring thresholds are the same across

FMTI v1.0 and v1.1. Due to ambiguity (as evidenced by developers misinterpreting indicators), one possible change could include splitting existing indicators into multiple different ones in order to more clearly delineate the information required in each indicator. Due to saturation (as evidenced by most developers satisfying many indicators), possible changes could include increasing the scoring thresholds for existing indicators as well as introducing new indicators about information that is not currently assessed by the Index. And due to changes external to FMTI in the foundation model ecosystem, such as growing interest in the encoding of human values (Scherrer et al., 2024) and the implementation of reward models for aligning foundation models (Lambert et al., 2024), possible changes could include adding indicators that specifically reflect transparency in the values encoded in models as well as information about how reward models are trained and operationalized. Most fundamentally, while the initial 100 indicators were decided upon by Bommasani et al. (2023a) with guidance from others in the community, a more open-ended process for community-driven indicator proposals may be implemented.

6 Limitations

In general, transparency as a construct and indices as an approach have well-known limitations (Birchall, 2014; Valdovinos, 2018; Alloa & Thomä, 2018; Alloa, 2018; Sagar & Najam, 1998; Boldin, 1999; Santeramo, 2017; Greco et al., 2019; Schlossarek et al., 2019). There are also a number of well-known limitations of transparency in AI specifically (Ananny & Crawford, 2018b; Phang et al., 2022; Hartzog, 2023). These limitations are discussed at length by Bommasani et al. (2023a) and they apply equally to FMTI v1.1.

In §2.2, Bommasani et al. (2023a) write: "Transparency is far from sufficient on its own and it may not always bring about the desired change (Corbett & Denton, 2023). Salient critiques of transparency include:

- Transparency does not equate to responsibility. Without broad based grassroots movements to exert public pressure or concerted government scrutiny, organizations often do not change bad practices (Boyd, 2016; Ananny & Crawford, 2018a).
- Transparency-washing provides the illusion of progress. Some organizations may misappropriate transparency as a means for subverting further scrutiny. For instance, major technology companies that vocally support transparency have been accused of *transparency-washing*, whereby "a focus on transparency acts as an obfuscation and redirection from more substantive and fundamental questions about the concentration of power, substantial policies and actions of technology behemoths" (Zalnieriute, 2021).
- Transparency can be gamified. Digital platforms have been accused of performative transparency, offering less insightful information in the place of useful and actionable visibility (Ghosh & Faxon, 2023; Mittelstadt, 2019). As with other metrics, improving transparency can be turned into a game, the object of which is not necessarily to share valuable information.[15]
- Transparency can inhibit privacy and promote surveillance. Transparency is not an apolitical concept and is often instrumentalized to increase surveillance and diminish privacy (Han, 2015; Mohamed et al., 2020; Birchall, 2021). For foundation models, this critique underscores a potential tension between adequate transparency with respect to the data used to build foundation models and robust data privacy.
- Transparency may compromise competitive advantage or intellectual property rights. Protections of competitive advantage plays a central role in providing companies to the incentives to innovate, thereby yielding competition in the marketplace that benefits consumers. Consequently, work in economics and management studies have studied the interplay and potential trade-off between competitive advantage and transparency (Bloomfield & O'Hara, 1999; Granados & Gupta, 2013; Liu et al., 2023), especially in the discourse on corporate social responsibility [(Wu et al., 2020; Yu et al., 2018)].

Transparency is not a panacea. In isolation, more information about foundation models will not necessarily produce a more just or equitable digital world. But if transparency is implemented through engagement with

[15]According to Goodhart's Law, "when a measure becomes a target, it ceases to be a good measure" (Goodhart, 1984).

third-party experts, independent auditors, and communities who are directly affected by digital technologies, it can help ensure that foundation models benefit society."

In §9.2, Bommasani et al. (2023a) address how several of these limitations of transparency and indices apply in the context of the Foundation Model Transparency Index. These limitations include equating transparency with responsibility, transparency washing, gaming the index, binary scoring, focusing on language models, and focusing on companies headquartered in the US. Beyond these limitations, we specifically discuss additional considerations that arose in this version.

Considering only publicly available information. At present, the Foundation Model Transparency Index exclusively considers public disclosure of information: such disclosures provide transparency to all stakeholders and are verifiable by the researchers conducting the Index. However, intermediary forms of information disclosure exist between no disclosure to entities beyond the developer and disclosure to everyone. The EU AI Act identifies core types of intermediary information disclosure: disclosure to the government (see Annex IX A) and disclosure to clients downstream in the foundation model supply chain (see Annex IX B). Kolt et al. (2024) provide initial guidance on stratified disclosures approaches for cybersecurity and biosecurity. Assessing companies' release of information via such intermediary forms of disclosure could encourage such disclosures, which can be beneficial even in the absence of public disclosures.

Requiring that disclosure be explicit. In order for a developer to score points on a given indicator, it must make an explicit disclosure related to that indicator. For example, scoring points on the hardware owner indicator requires that a developer explicitly state which organization(s) owns the primary hardware used in building the model; even if the developer considers this to be obvious and not worth stating in its documentation, it must disclose the owner explicitly in its transparency report to score the point. We use this standard to promote reproducibility and remove ambiguity: rather than making assumptions based on what developers imply about their flagship foundation models, explicit disclosures ensure that anyone could re-score the developer in the same way.

Company selection of flagship foundation models. As part of a change to our methodology, for FMTI v1.1 each company selected its flagship foundation model to be assessed. While we provided guidance to companies regarding how to do so—we stated it should be based on a combination of resources expended, capabilities, and societal impact—we do not have sufficient insight into a company's operations (or, for example, what model sits behind an API) to validate that the model a developer designates as its flagship model is in fact its most significant model at present. This flexibility introduces risks, as it is possible that a developer might chose the most transparent of its foundation models in order to increase its score. This limitation stems in large part from the fact that we assess a developer's transparency based on a single flagship foundation model.

Companies submitted transparency reports. In this iteration of the Index, companies submit transparency reports to disclose key information about their foundation models. This comes with several limitations. Where a developer does not disclose information related to a particular indicator or does not affirmatively and explicitly disclose that it satisfies the indicator, we take this at face value and score that indicator as a zero. In FMTI v1.0, however, there were multiple cases that identified information in companies' documentation that they themselves did not believe satisfied a particular indicator. The methodology used in FMTI v1.1 prevents this from happening, instead assuming that companies know their own documentation best.

No distinction between B2B and B2C companies. The developers that we assess include companies with a variety of different business models, including those that primary develop models for enterprise customers (e.g. Aleph Alpha and Amazon) and those with a large number of individuals as customers (e.g. Anthropic and OpenAI). These differences in business models result in different approaches to transparency. B2B companies prioritize disclosing information directly to enterprise customers, who may be less concerned about public facing transparency. We use the same set of 100 transparency indicators for each company, which limits our ability to capture the nuances in how companies conceive of transparency and its value for their customers. A number of companies requested that we assess certain indicators as "Not Applicable" in light of their business model, but our methodology of binary scoring prevented us from contemplating this option.

Low bar for awarding points. Bommasani et al. (2023a) indicate "We were generally quite generous in the scoring process. When we determined that a developer scored some version of a half-point, we usually rounded up. Since we assess transparency, we award developers points if they explicitly disclose that they do not share information about a particular indicator. We also read developers' documents with deference where possible, meaning that we often awarded points where there are grey areas. This means that developers' scores may actually be higher than their documentation warrants in certain cases as we had a low bar for awarding points on many indicators." In future iterations of the index we may raise the threshold for scoring points (e.g. round down scores on indicators that might otherwise be half-points) as doing so would require a greater degree of transparency and potentially foster a more consistent grading standard.

7 Conclusion

The societal impact of foundation models is escalating, attracting the attention of firms, media, academia, government, and the public. The Foundation Model Transparency Index continues to find that transparency ought to be improved in this nascent ecosystem, with some positive developments since October 2023. By dissecting what developers do and do not publicly disclose, and how this has changed, the Index allows different stakeholders (e.g. developers, customers, investors, policymakers) to make more clear-eyed decisions. And, in turn, by establishing the practice of transparency reporting for foundation models, the Index surfaces a new resource that downstream developers, researchers, and journalists should capitalize on to build collective understanding. Moving forward, we hope that headway on transparency will demonstrably translate to better societal outcomes like greater accountability, improved science, increased innovation, and better policy.

References

Emmanuel Alloa. *Transparency: A Magic Concept of Modernity*, pp. 21–55. Springer International Publishing, Cham, 2018. ISBN 978-3-319-77161-8. doi: 10.1007/978-3-319-77161-8_3. URL https://doi.org/10.1007/978-3-319-77161-8_3.

Emmanuel Alloa and Dieter Thomä. *Transparency: Thinking Through an Opaque Concept*, pp. 1–13. 06 2018. ISBN 978-3-319-77160-1. doi: 10.1007/978-3-319-77161-8_1.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018a. doi: 10.1177/1461444816676645. URL https://doi.org/10.1177/1461444816676645.

Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018b. doi: 10.1177/1461444816676645. URL https://doi.org/10.1177/1461444816676645.

Aspen Institute. Commission on information disorder final report. https://www.aspeninstitute.org/wp-content/uploads/2021/11/Aspen-Institute_Commission-on-Information-Disorder_Final-Report.pdf, 11 2021.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Jack Bandy and Nicholas Vincent. Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Kevin Bankston and Jennifer Hodges. Openness and transparency in ai provide significant benefits for society, 2024. URL https://cdt.org/wp-content/uploads/2024/03/Civil-Society-Letter-on-Openness-for-NTIA-Process-March-25-2024.pdf.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6: 587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Clare Birchall. Radical transparency? *Cultural Studies and Critical Methodologies*, 14(1):77–88, 2014. doi: 10.1177/1532708613517442. URL https://doi.org/10.1177/1532708613517442.

Clare Birchall. *Radical secrecy: The ends of transparency in datafied America*, volume 60. U of Minnesota Press, 2021.

Sid Black, Stella Rose Biderman, Eric Hallahan, Quentin G. Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, M. Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, J. Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An open-source autoregressive language model. *arXiv*, 2022.

Robert Bloomfield and Maureen O'Hara. Market transparency: Who wins and who loses? *The Review of Financial Studies*, 12(1):5–35, 1999. ISSN 08939454, 14657368. URL http://www.jstor.org/stable/2645985.

Michael D Boldin. A critique of the traditional composite index methodology. *Journal of economic and social measurement*, 25(3-4):119–140, 1999.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index. *ArXiv*, abs/2310.12941, 2023a. URL https://api.semanticscholar.org/CorpusID:264306385.

Rishi Bommasani, Dilara Soylu, Thomas Liao, Kathleen A. Creel, and Percy Liang. Ecosystem graphs: The social footprint of foundation models. *ArXiv*, abs/2303.15772, 2023b. URL https://api.semanticscholar.org/CorpusID:257771875.

Rishi Bommasani, Daniel Zhang, Tony Lee, and Percy Liang. Improving transparency in ai language models: A holistic evaluation. *Foundation Model Issue Brief Series*, 2023c. URL https://hai.stanford.edu/foundation-model-issue-brief-series.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. Foundation model transparency reports. *ArXiv*, abs/2402.16268, 2024. URL https://api.semanticscholar.org/CorpusID:267938721.

Danah Boyd. Algorithmic accountability and transparency. Open Transcripts, Nov 2016. URL http://opentranscripts.org/transcript/danah-boyd-algorithmic-accountability-transparency/. Presented by danah boyd in Algorithmic Accountability and Transparency in the Digital Economy.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2280–2292, 2022.

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Rosario Cammarota, Matthias Schunter, Anand Rajan, Fabian Boemer, Ágnes Kiss, Amos Treiber, Christian Weinert, Thomas Schneider, Emmanuel Stapf, Ahmad-Reza Sadeghi, et al. Trustworthy ai inference systems: An industry research view. *arXiv preprint arXiv:2008.04449*, 2020.

Eric Corbett and Emily Denton. Interrogating the t in facct. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1624–1634, 2023.

Kate Crawford. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

EU. Official journal of the european union 2016. *Official Journal of the European Union*, L 119/1, Apr 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1552662547490&uri=CELEX%3A32016R0679.

L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. A framework for few-shot language model evaluation, 2021.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Ill, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Ritwick Ghosh and Hilary Oliva Faxon. Smart corruption: Satirical strategies for gaming accountability. *Big Data & Society*, 10(1):20539517231164119, 2023. doi: 10.1177/20539517231164119. URL https://doi.org/10.1177/20539517231164119.

Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice*, pp. 91–121. Springer, 1984.

Nelson Granados and Alok Gupta. Transparency strategy: Competing with information in a digital world. *MIS Quarterly*, 37(2):637–641, 2013. ISSN 02767783. URL http://www.jstor.org/stable/43825928.

Mary L Gray and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019a.

M.L. Gray and S. Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, 2019b. ISBN 978-1-328-56624-9. URL https://books.google.com/books?id=u10-uQEACAAJ.

Salvatore Greco, Alessio Ishizaka, Menelaos Tasiou, and Gianpiero Torrisi. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, 141:1–34, 01 2019. doi: 10.1007/s11205-017-1832-9.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Byung-Chul Han. *The transparency society*. Stanford University Press, 2015.

Karen Hao and Deepa Seetharaman. Cleaning up chatgpt takes heavy toll on human workers. *The Wall Street Journal*, July 2023. URL https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483. Photographs by Natalia Jidovanu.

Woodrow Hartzog. Oversight of a.i.: Legislating on artificial intelligence. Prepared Testimony and Statement for the Record before the U.S. Senate Committee on the Judiciary, Subcommittee on Privacy, Technology, and the Law, Sep 2023. URL https://www.judiciary.senate.gov/imo/media/doc/2023-09-12_pm_-_testimony_-_hartzog.pdf.

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. On the societal impact of open foundation models, 2024.

Stephan Klasen. Human development indices and indicators: A critical evaluation. 2018.

Kevin Klyman. Acceptable use policies for foundation models: Considerations for policymakers and developers. Stanford Center for Research on Foundation Models, April 2024. URL https://crfm.stanford.edu/2024/04/08/aups.html.

Lauren Kogen. From statistics to stories: Indices and indicators as communication tools for social change. *The International Journal of Press/Politics*, 0(0):19401612221094246, 2022. doi: 10.1177/19401612221094246. URL https://doi.org/10.1177/19401612221094246.

Noam Kolt, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin M. Esvelt, Gillian K. Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B. Sandbrink, and Thomas Woodside. Responsible reporting for frontier ai development. 2024. URL https://api.semanticscholar.org/CorpusID:268875838.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. Language generation models can cause harm: So what can we do about it? an actionable survey. *arXiv preprint arXiv:2210.07700*, 2022.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'bout ai generation: Copyright and the generative-ai supply chain, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Yeyi Liu, Martin Heinberg, Xuan Huang, and Andreas B. Eisingerich. Building a competitive advantage based on transparency: When and why does transparency matter for corporate social responsibility? *Business Horizons*, 66(4):517–527, 2023. ISSN 0007-6813. doi: https://doi.org/10.1016/j.bushor.2022.10.004. URL https://www.sciencedirect.com/science/article/pii/S0007681322001306.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023a.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023b.

Shayne Longpre, Sayash Kapoor, Kevin Klyman, Ashwin Ramaswami, Rishi Bommasani, Borhane Blili-Hamelin, Yangsibo Huang, Aviya Skowron, Zheng-Xin Yong, Suhas Kotha, et al. A safe harbor for ai evaluation and red teaming. *arXiv preprint arXiv:2403.04893*, 2024a.

Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Katy Gero, Sandy Pentland, and Jad Kabbara. Data authenticity, consent, & provenance for ai are all broken: what will it take to fix them?, 2024b.

Alexandra Sasha Luccioni and Alex Hernández-García. Counting carbon: A survey of factors influencing the emissions of machine learning. *ArXiv*, abs/2302.08476, 2023.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit, 2023.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. Arctic-embed: Scalable, efficient, and accurate text embedding models. *arXiv preprint arXiv:2405.05374*, 2024.

Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11):501–507, November 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0114-4. URL https://doi.org/10.1038/s42256-019-0114-4.

Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4):659–684, December 2020. ISSN 2210-5441. doi: 10.1007/s13347-020-00405-8. URL https://doi.org/10.1007/s13347-020-00405-8.

Arvind Narayanan and Sayash Kapoor. Generative ai companies must publish transparency reports, 2023. URL https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

Jason Phang, Herbie Bradley, Leo Gao, Louis Castricato, and Stella Biderman. Eleutherai: Going beyond "open science" to "science in the open", 2022.

Hannah Quay-de la Vallee, Ridhi Shetty, and Elizabeth Laird. The Federal Government's Power of the Purse: Enacting Procurement Policies and Practices to Support Responsible AI Use. *Center for Democracy and Technology*, 2024. URL https://cdt.org/wp-content/uploads/2024/04/2024-04-27-AI-Federal-Procurement-report-1.pdf.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 429–435, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314244. URL https://doi.org/10.1145/3306618.3314244.

Ambuj D Sagar and Adil Najam. The human development index: a critical review. *Ecological economics*, 25 (3):249–264, 1998.

SambaNova. Samba-1: 1 trillion parameter composition of experts, 2024. URL https://sambanova.ai/products/samba-1.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *CHI*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445518. URL https://doi.org/10.1145/3411764.3445518.

Fabio Gaetano Santeramo. Methodological challenges in building composite indexes: Linking theory to practice. In *Emerging trends in the development and application of composite indicators*, pp. 127–139. IGI Global, 2017.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Martin Schlossarek, Miroslav Syrovátka, and Ondřej Vencálek. The importance of variables in composite indices: A contribution to the methodology and application to development indices. *Social Indicators Research*, 145(3):1125–1160, 2019.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63 (12):54–63, 2020.

Irene Solaiman. The gradient of generative ai release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 111–122, 2023.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024.

Daniel J. Solove. Artificial intelligence and privacy. *Florida Law Review*, 2024. doi: 10.2139/ssrn.4713111. URL `https://ssrn.com/abstract=4713111`. Forthcoming January 2025.

Elizabeth Stanton. The human development index: A history. *PERI Working Papers*, 2007. URL `https://scholarworks.umass.edu/peri_workingpapers/85/`.

Elham Tabassi. Artificial intelligence risk management framework (ai rmf 1.0), 2023-01-26 05:01:00 2023. URL `https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225`.

Reka Team, Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka core, flash, and edge: A series of powerful multimodal language models. *arXiv preprint arXiv:2404.12387*, 2024.

Thorn. Safety by design for generative ai: Preventing child sexual abuse, 2024. URL `https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf`.

Trust and Safety Professional Association. Transparency reporting. `https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/`, 2023.

Jorge I. Valdovinos. Transparency as ideology, ideology as transparency: Towards a critique of the meta-aesthetics of neoliberal hegemony. *Open Cultural Studies*, 2(1):654–667, 2018. doi: doi:10.1515/culture-2018-0059. URL `https://doi.org/10.1515/culture-2018-0059`.

Lee Vinsel. *Moving violations: automobiles, experts, and regulations in the United States*. JHU Press, 2019.

Jai Vipra and Sarah Myers West. Computational power and ai, Sep 2023. URL `https://ainowinstitute.org/publication/policy/compute-and-ai`.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William S. Isaac. Sociotechnical safety evaluation of generative ai systems. 2023. URL `https://arxiv.org/abs/2310.11986`.

Amy Winograd. Loose-lipped large language modells spill your secrets: The privacy implications of large language models. *Harvard Journal of Law and Technology*, 36(2), 2023.

Yue Wu, Kaifu Zhang, and Jinhong Xie. Bad greenwashing, good greenwashing: Corporate social responsibility and information transparency. *Management Science*, 66(7):3095–3112, 2020.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, Donghyun Kwak, Hanock Kwak, Se Jung Kwon, Bado Lee, Dongsoo Lee, Gichang Lee, Jooho Lee, Baeseong Park, Seongjin Shin, Joonsang Yu, Seolki Baek, Sumin Byeon, Eungsup Cho, Dooseok Choe, Jeesung Han, Youngkyun Jin, Hyein Jun, Jaeseung Jung, Chanwoong Kim, Jinhong Kim, Jinuk Kim, Dokyeong Lee, Dongwook Park, Jeong Min Sohn, Sujung Han, Jiae Heo, Sungju Hong, Mina Jeon, Hyunhoon Jung, Jungeun Jung, Wangkyo Jung, Chungjoon Kim, Hyeri Kim, Jonghyun Kim, Min Young Kim, Soeun Lee, Joonhee Park, Jieun Shin, Sojin Yang, Jungsoon Yoon, Hwaran Lee, Sanghwan Bae, Jeehwan Cha, Karl Gylleus, Donghoon Ham, Mihak Hong, Youngki Hong, Yunki Hong, Dahyun Jang, Hyojun Jeon, Yujin Jeon, Yeji Jeong, Myunggeun Ji, Yeguk Jin, Chansong Jo, Shinyoung Joo, Seunghwan Jung, Adrian Jungmyung Kim, Byoung Hoon Kim, Hyomin Kim, Jungwhan Kim, Minkyoung Kim, Minseung Kim, Sungdong Kim, Yonghee Kim, Youngjun Kim, Youngkwan Kim, Donghyeon Ko, Dughyun Lee, Ha Young Lee, Jaehong Lee, Jieun Lee, Jonghyun Lee, Jongjin Lee, Min Young Lee, Yehbin Lee, Taehong Min, Yuri Min, Kiyoon Moon, Hyangnam Oh, Jaesun Park, Kyuyon Park, Younghun Park, Hanbae Seo, Seunghyun Seo, Mihyun Sim, Gyubin Son, Matt Yeo, Kyung Hoon Yeom, Wonjoon Yoo, Myungin You, Doheon Ahn, Homin Ahn, Joohee Ahn, Seongmin Ahn, Chanwoo An, Hyeryun An, Junho An, Sang-Min An, Boram Byun, Eunbin Byun, Jongho Cha, Minji Chang, Seunggyu Chang, Haesong Cho, Youngdo Cho, Dalnim Choi, Daseul Choi, Hyoseok Choi, Minseong Choi, Sangho Choi, Seongjae Choi, Wooyong Choi, Sewhan Chun, Dong Young Go, Chiheon Ham, Danbi Han, Jaemin Han, Moonyoung Hong, Sung Bum Hong, Dong-Hyun Hwang, Seongchan Hwang, Jinbae Im, Hyuk Jin Jang, Jaehyung Jang, Jaeni Jang, Sihyeon Jang, Sungwon Jang, Joonha Jeon, Daun Jeong, Joonhyun Jeong, Kyeongseok Jeong, Mini Jeong, Sol Jin, Hanbyeol Jo, Hanju Jo, Minjung Jo, Chaeyoon Jung, Hyungsik Jung, Jaeuk Jung, Ju Hwan Jung, Kwangsun Jung, Seungjae Jung, Soonwon Ka, Donghan Kang, Soyoung Kang, Taeho Kil, Areum Kim, Beomyoung Kim, Byeongwook Kim, Daehee Kim, Dong-Gyun Kim, Donggook Kim, Donghyun Kim, Euna Kim, Eunchul Kim, Geewook Kim, Gyu Ri Kim, Hanbyul Kim, Heesu Kim, Isaac Kim, Jeonghoon Kim, Jihye Kim, Joonghoon Kim, Minjae Kim, Minsub Kim, Pil Hwan Kim, Sammy Kim, Seokhun Kim, Seonghyeon Kim, Soojin Kim, Soong Kim, Soyoon Kim, Sunyoung Kim, Taeho Kim, Wonho Kim, Yoonsik Kim, You Jin Kim, Yuri Kim, Beomseok Kwon, Ohsung Kwon, Yoo-Hwan Kwon, Anna Lee, Byungwook Lee, Changho Lee, Daun Lee, Dongjae Lee, Ha-Ram Lee, Hodong Lee, Hwiyeong Lee, Hyunmi Lee, Injae Lee, Jaeung Lee, Jeongsang Lee, Jisoo Lee, Jongsoo Lee, Joongjae Lee, Juhan Lee, Jung Hyun Lee, Junghoon Lee, Junwoo Lee, Se Yun Lee, Sujin Lee, Sungjae Lee, Sungwoo Lee, Wonjae Lee, Zoo Hyun Lee, Jong Kun Lim, Kun Lim, Taemin Lim, Nuri Na, Jeongyeon Nam, Kyeong-Min Nam, Yeonseog Noh, Biro Oh, Jung-Sik Oh, Solgil Oh, Yeontaek Oh, Boyoun Park, Cheonbok Park, Dongju Park, Hyeonjin Park, Hyun Tae Park, Hyunjung Park, Jihye Park, Jooseok Park, Junghwan Park, Jungsoo Park, Miru Park, Sang Hee Park, Seunghyun Park, Soyoung Park, Taerim Park, Wonkyeong Park, Hyunjoon Ryu, Jeonghun Ryu, Nahyeon Ryu, Soonshin Seo, Suk Min Seo, Yoonjeong Shim, Kyuyong Shin, Wonkwang Shin, Hyun Sim, Woongseob Sim, Hyejin Soh, Bokyong Son, Hyunjun Son, Seulah Son, Chi-Yun Song, Chiyoung Song, Ka Yeon Song, Minchul Song, Seungmin Song, Jisung Wang, Yonggoo Yeo, Myeong Yeon Yi, Moon Bin Yim, Taehwan Yoo, Youngjoon Yoo, Sungmin Yoon, Young Jin Yoon, Hangyeol Yu, Ui Seon Yu, Xingdong Zuo, Jeongin Bae, Joungeun Bae, Hyunsoo Cho, Seonghyun Cho, Yongjin Cho, Taekyoon Choi, Yera Choi, Jiwan Chung, Zhenghui Han, Byeongho Heo, Euisuk Hong, Taebaek Hwang, Seonyeol Im, Sumin Jegal, Sumin Jeon, Yelim Jeong, Yonghyun Jeong, Can Jiang, Juyong Jiang, Jiho Jin, Ara Jo, Younghyun Jo, Hoyoun Jung, Juyoung Jung, Seunghyeong Kang, Dae Hee Kim, Ginam Kim, Hangyeol Kim, Heeseung Kim, Hyojin Kim, Hyojun Kim, Hyun-Ah Kim, Jeehye Kim, Jin-Hwa Kim, Jiseon Kim, Jonghak Kim, Jung Yoon Kim, Rak Yeong Kim, Seongjin Kim, Seoyoon Kim, Sewon Kim, Sooyoung Kim, Sukyoung Kim, Taeyong Kim, Naeun Ko, Bonseung Koo, Heeyoung Kwak, Haena Kwon, Youngjin Kwon, Boram Lee, Bruce W. Lee, Dagyeong Lee, Erin Lee, Euijin Lee, Ha Gyeong Lee, Hyojin Lee, Hyunjeong Lee, Jeeyoon Lee, Jeonghyun Lee, Jongheok Lee, Joonhyung Lee, Junhyuk Lee, Mingu Lee, Nayeon Lee, Sangkyu Lee, Se Young Lee, Seulgi Lee, Seung Jin Lee, Suhyeon Lee, Yeonjae Lee, Yesol Lee, Youngbeom Lee, Yujin Lee, Shaodong Li, Tianyu Liu, Seong-Eun Moon, Taehong Moon, Max-Lasse Nihlenramstroem,

Wonseok Oh, Yuri Oh, Hongbeen Park, Hyekyung Park, Jaeho Park, Nohil Park, Sangjin Park, Jiwon Ryu, Miru Ryu, Simo Ryu, Ahreum Seo, Hee Seo, Kangdeok Seo, Jamin Shin, Seungyoun Shin, Heetae Sin, Jiangping Wang, Lei Wang, Ning Xiang, Longxiang Xiao, Jing Xu, Seonyeong Yi, Haanju Yoo, Haneul Yoo, Hwanhee Yoo, Liang Yu, Youngjae Yu, Weijie Yuan, Bo Zeng, Qian Zhou, Kyunghyun Cho, Jung-Woo Ha, Joonsuk Park, Jihyun Hwang, Hyoung Jo Kwon, Soonyong Kwon, Jungyeon Lee, Seungho Lee, Seonghyeon Lim, Hyunkyung Noh, Seungho Choi, Sang-Woo Lee, Jung Hwa Lim, and Nako Sung. Hyperclova x technical report, 2024.

Ellen Pei-Yi Yu, Christine Qian Guo, and Bac Van Luu. Environmental, social and governance transparency and firm value. *Business Strategy and the Environment*, 27(7):987–1004, 2018.

Monika Zalnieriute. "transparency-washing" in the digital age : A corporate agenda of procedural fetishism. Technical report, 2021. URL `http://hdl.handle.net/11159/468588`.

A Selection decisions

In conducting the index, core structural design decisions are (i) the indicators used to assess developers and (ii) the developers that are assessed. Here, we clarify both matters.

A.1 Indicator selection

We use the same 100 indicators as FMTI v1.0 to facilitate direct comparison. These indicators are listed by domain in Figure 8.

A.2 Developer selection

We contacted 19 foundation model developers to request these developers submit transparency reports for the purpose of conducting FMTI v1.1. Consistent with the principles used by Bommasani et al. (2023a), we only considered developers that are companies and that develop prominent foundation models. Specifically, we contacted leadership via email at 01.ai, Adept, AI21 Labs, Aleph Alpha, Amazon, Anthropic, BigCode (Hugging Face and ServiceNow), Cohere, Databricks, Google, IBM, Inflection, Meta, Microsoft, Mistral, OpenAI, Stability AI, Writer, and xAI. Following this email correspondence, and further clarification of the nature of the request, 14 foundation model developers agreed to provide the requested transparency reports.

Therefore, our selection process deliberately excluded foundation model developers that are not companies, even if they develop prominent foundation models, such as the Allen Institute for AI (developer of models such as OLMo; Groeneveld et al., 2024) and EleutherAI (developer of Pythia; Biderman et al., 2023). While developers such as AI2 and EleutherAI are often leaders in various types of transparency, releasing detailed information about data (Gao et al., 2020; Soldaini et al., 2024), evaluations (Gao et al., 2021; Magnusson et al., 2023), and the model development pipeline (Black et al., 2022; Biderman et al., 2023; Groeneveld et al., 2024), we consider only companies in selecting developers to assess in FMTI v1.1.

Our selection process also did not involve engagement with developers where we lacked connections with their leadership, which often coincides with models developed outside the United States and the Western hemisphere (e.g. the developers of Falcon (Almazrouei et al., 2023), Qwen (Bai et al., 2023), DeepSeek (DeepSeek-AI et al., 2024), and HyperCLOVA (Yoo et al., 2024)).

Finally, we did not engage any developer that released its first prominent foundation model during our execution of FMTI v1.1 (such releases include Apple's MM1 (McKinzie et al., 2024), SambaNova's Samba 1 (SambaNova, 2024), Reka's Reka Core (Team et al., 2024), and Snowflake's Arctic-1 (Merrick et al., 2024)). Moving forward, having successfully demonstrated that prominent foundation model developers have cooperated by submitting transparency reports, subsequent versions of the Foundation Model Transparency Index may engage these companies as well as others.

2023 Foundation Model Transparency Index Indicators

Upstream	Model	Downstream
Data size	Input modality	Release decision-making
Data sources	Output modality	Release process
Data creators	Model components	Distribution channels
Data source selection	Model size	Products and services
Data curation	Model architecture	Detection of machine-generated content
Data augmentation	Centralized model documentation	Model License
Harmful data filtration	External model access protocol	Terms of service
Copyrighted data	Blackbox external model access	Permitted and prohibited users
Data license	Full external model access	Permitted, restricted, and prohibited uses
Personal information in data	Capabilities description	Usage policy enforcement
Use of human labor	Capabilities demonstration	Justification for enforcement action
Employment of data laborers	Evaluation of capabilities	Usage policy violation appeals mechanism
Geographic distribution of data laborers	External reproducibility of capabilities evaluation	Permitted, restricted, and prohibited model behaviors
Wages	Third party capabilities evaluation	Model behavior policy enforcement
Instructions for creating data	Limitations description	Interoperability of usage and model behavior policies
Labor protections	Limitations demonstration	User interaction with AI system
Third party partners	Third party evaluation of limitations	Usage disclaimers
Queryable external data access	Risks description	User data protection policy
Direct external data access	Risks demonstration	Permitted and prohibited use of user data
Compute usage	Unintentional harm evaluation	Usage data access protocol
Development duration	External reproducibility of unintentional harm evaluation	Versioning protocol
Compute hardware	Intentional harm evaluation	Change log
Hardware owner	External reproducibility of intentional harm evaluation	Deprecation policy
Energy usage	Third party risks evaluation	Feedback mechanism
Carbon emissions	Mitigations description	Feedback summary
Broader environmental impact	Mitigations demonstration	Government inquiries
Model stages	Mitigations evaluation	Monitoring mechanism
Model objectives	External reproducibility of mitigations evaluation	Downstream applications
Core frameworks	Third party mitigations evaluation	Affected market sectors
Additional dependencies	Trustworthiness evaluation	Affected individuals
Mitigations for privacy	External reproducibility of trustworthiness evaluation	Usage reports
Mitigations for copyright	Inference duration evaluation	Geographic statistics
	Inference compute evaluation	Redress mechanism
		Centralized documentation for downstream use
		Documentation for responsible downstream use

Figure 8: **Indicators.** The 100 indicators we use across 3 domains (upstream, model, and downstream) that are the same as in the October 2023 Foundation Model Transparency Index.

.

B Extended Results

We present additional results, drawing inspiration from the analyses conducted in the first version of the Foundation Model Transparency Index.

B.1 Developer correlations

Measuring correlations. The 100×14 matrix of scores introduces data-driven structure. In particular, it clarifies relationships that arise in practice between different regions of the index. Here, we consider the *correlations*, in scores, focusing on company-to-company similarity for simplicity. For example, this analysis helps address the following: if two companies receive similar aggregate scores, is this because they satisfy all the same indicators or do they score points on two very different sets of indicators?

In Figure 9, we plot the correlation between every pair of companies. To measure correlation, we report the simple matching coefficient (SMC) or the agreement rate. The SMC is the fraction of the 100 indicators for which both companies receive the same score (i.e. both receive a zero or both receive a 1). As a result, a SMC of 0 indicates there is no indicator such that both companies receive the same score and a SMC of 1 indicates that for all indicators both companies receive the same score. For this reason, the correlation matrix is symmetric and guaranteed to be 1 on the diagonal.

Upstream correlations. In Figure 10, we plot the correlation between every pair of companies when considering only indicators from the upstream domain.

Model correlations. In Figure 11, we plot the correlation between every pair of companies when considering only indicators from the model domain.

Downstream correlations. In Figure 12, we plot the correlation between every pair of companies when considering only indicators from the downstream domain.

B.2 Indicator-level results

Assessing companies at the indicator level. The core of the Foundation Model Transparency Index is the 100 indicators of transparency, which we aggregate into subdomains and domains in order to facilitate discussion and analysis of our results. We score each developer on each indicator (either 0 or 1) based on the information disclosed by the developer in relation to that indicator; each indicator is accompanied by a definition, which describes the indicator, and notes, which provide details about how that indicator is scored (Bommasani et al., 2023a, Appendix B). Below we provide each developers' score on every indicator, broken down by domain (i.e. upstream, model, and downstream).[16]

Upstream Indicators. In Figure 13, we show the scores of every developer on each of the indicators in the upstream domain. We also disaggregate uptream indicators by subdomain (data, data labor, data access, compute, methods, and data mitigations).

Model Indicators. In Figure 14, we show the scores of every developer on each of the indicators in the model domain. We also disaggregate model indicators by subdomain (distribution, usage policy, model behavior policy, user interface, user data protection, model updates, feedback, impact, and downstream documentation).

Downstream Indicators. In Figure 15, we show the scores of every developer on each of the indicators in the downstream domain. We also disaggregate downstream indicators by subdomain (model basics, model access, capabilities, limitations, risks, model mitigations, trustworthiness, and inference).

[16] URL to data is anonymized for submission.

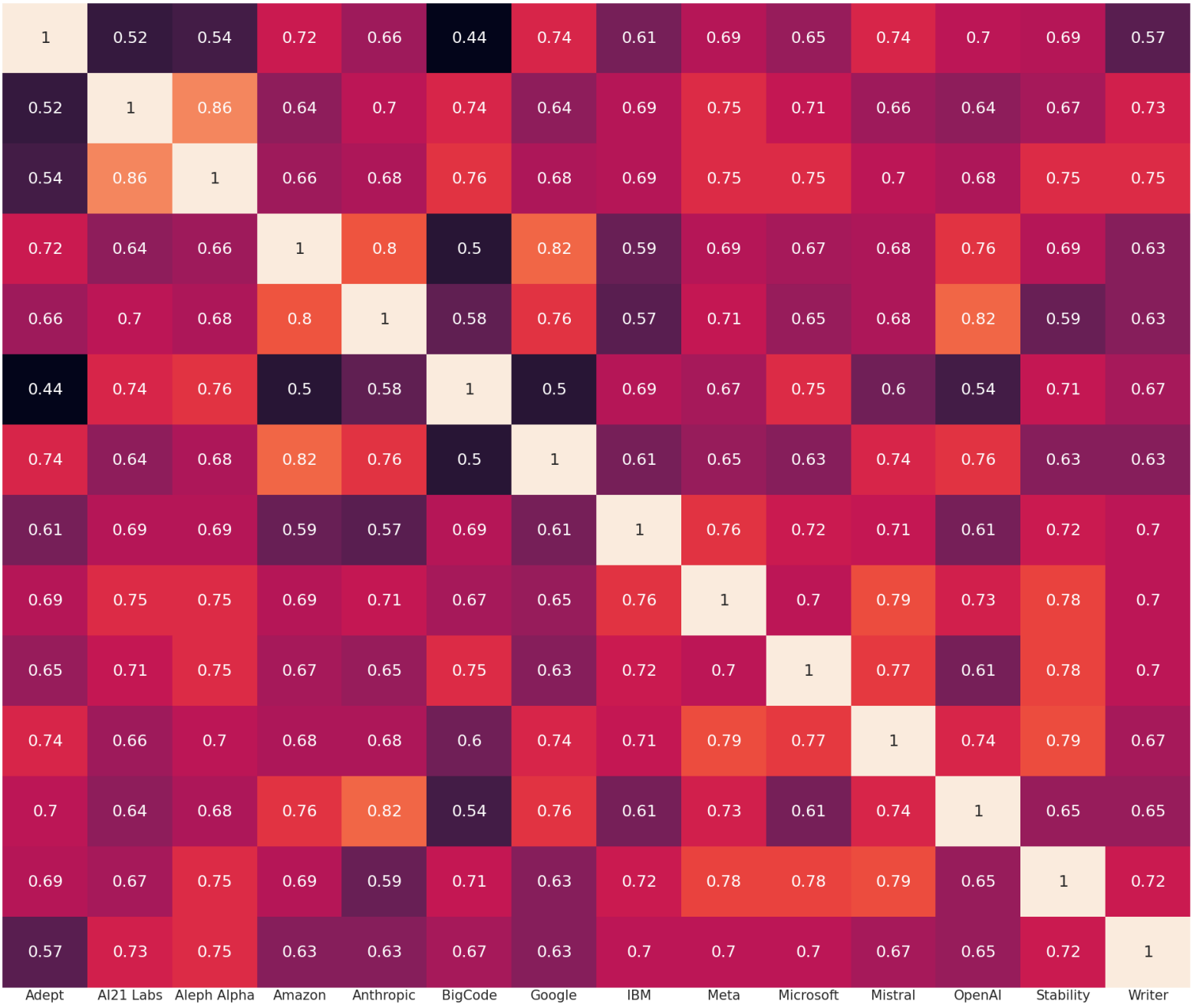

Figure 9: **Correlations between Companies.** The correlation between the scores for pairs of companies across all indicators. Correlation is measured using the simple matching coefficient (i.e. agreement rate), which is the fraction of all indicators for which both companies receive the same score (i.e. both receive the point or both do not receive the point).

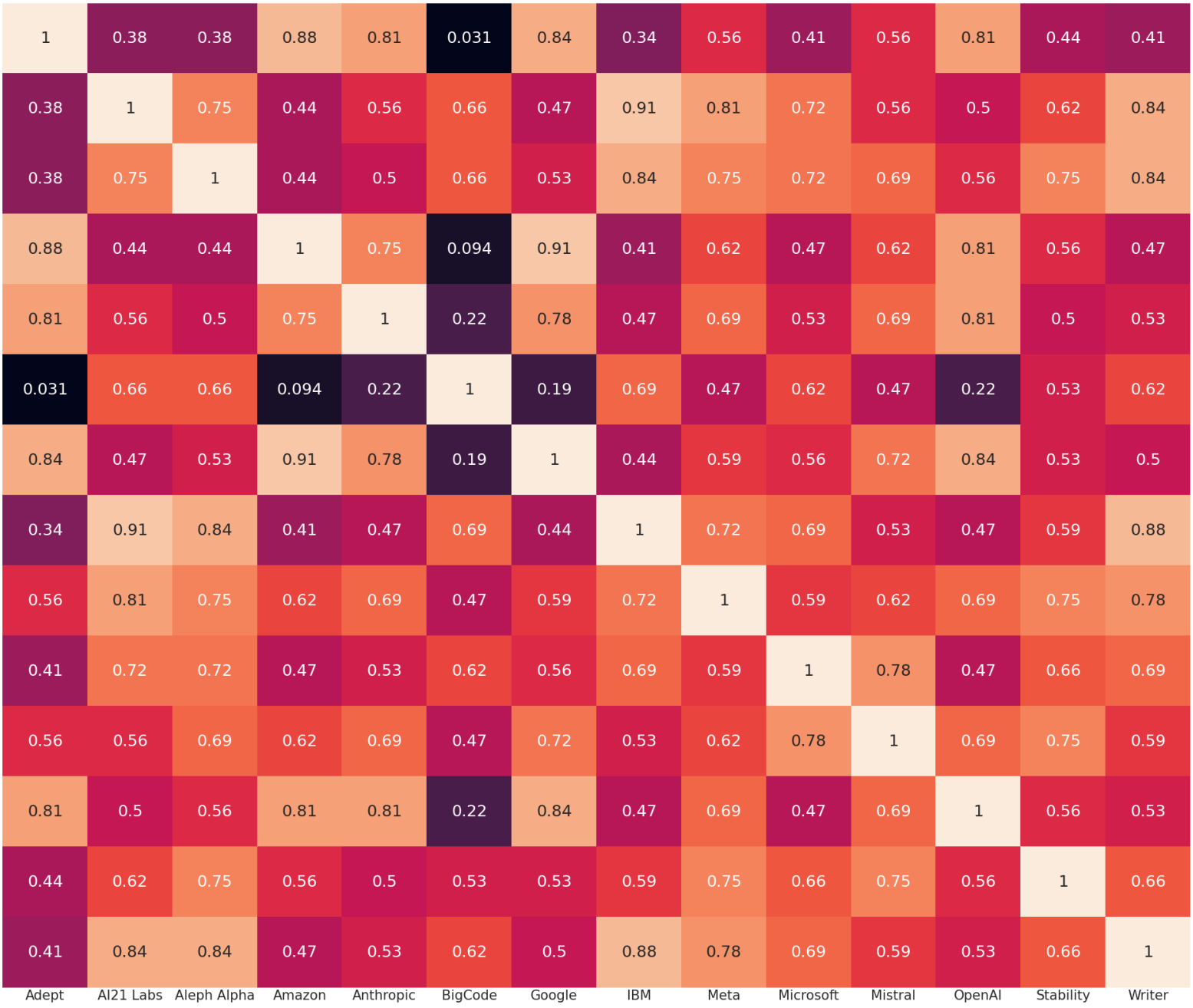

Figure 10: **Correlations between Companies (Upstream Indicators).** The correlation between the scores for pairs of companies across all indicators when only considering upstream indicators. Correlation is measured using the simple matching coefficient (i.e. agreement rate), which is the fraction of all indicators for which both companies receive the same score (i.e. both receive the point or both do not receive the point).

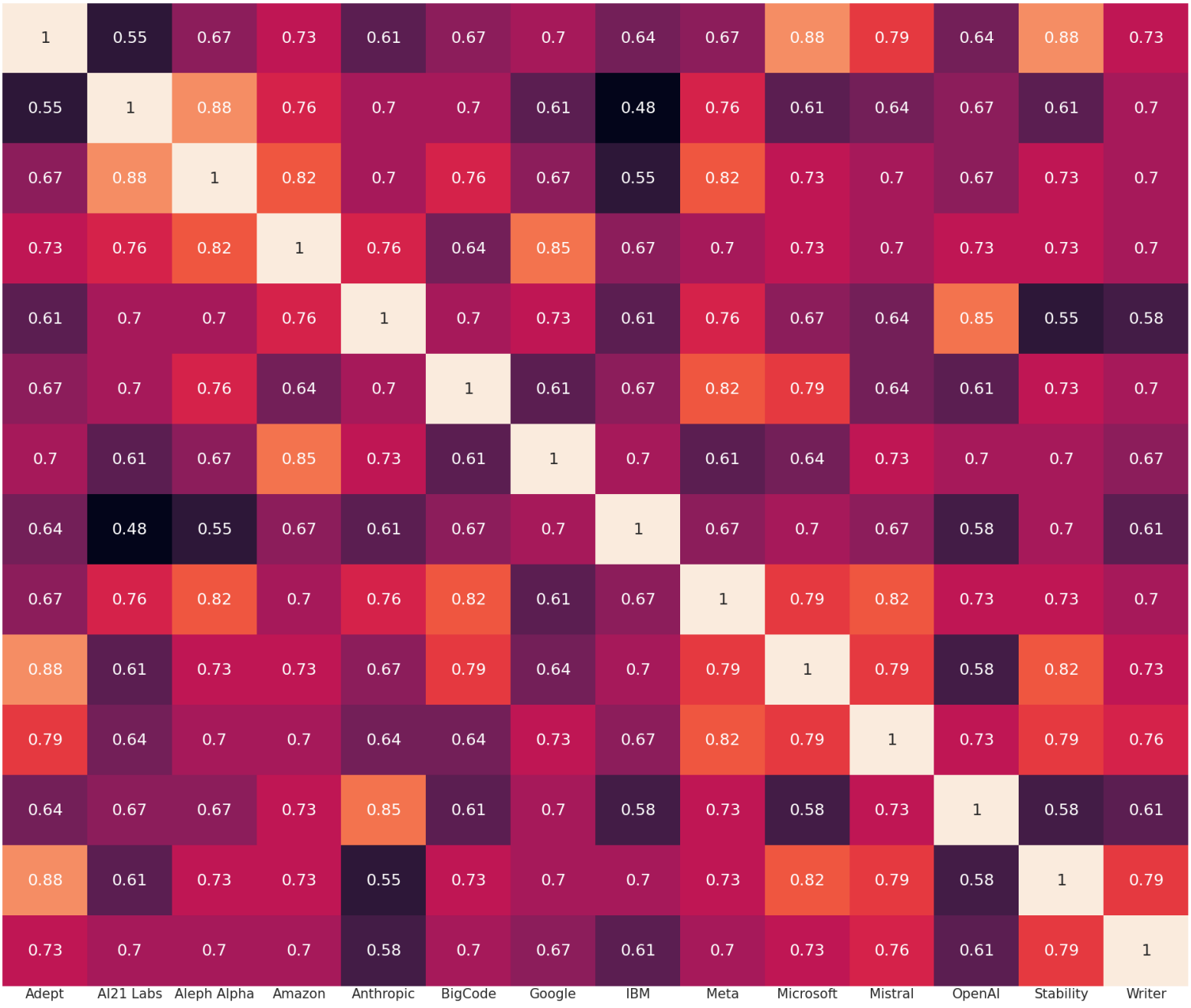

Figure 11: **Correlations between Companies (Model Indicators).** The correlation between the scores for pairs of companies across all indicators when only considering model indicators. Correlation is measured using the simple matching coefficient (i.e. agreement rate), which is the fraction of all indicators for which both companies receive the same score (i.e. both receive the point or both do not receive the point).

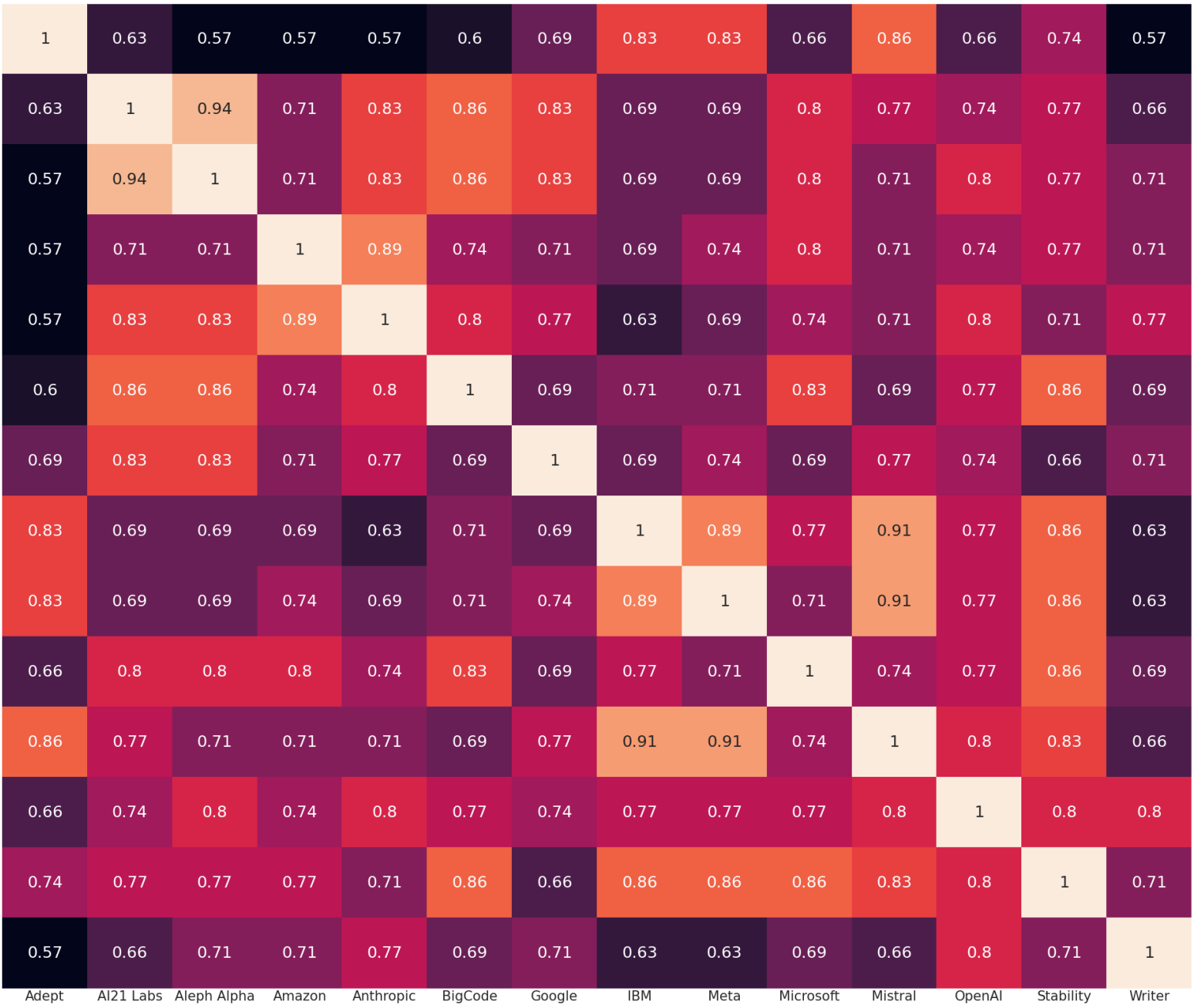

Figure 12: **Correlations between Companies (Downstream Indicators).** The correlation between the scores for pairs of companies across all indicators when considering only downstream indicators. Correlation is measured using the simple matching coefficient (i.e. agreement rate), which is the fraction of all indicators for which both companies receive the same score (i.e. both receive the point or both do not receive the point).

Foundation Model Transparency Index Indicator-Level Scores for Upstream, May 2024

Source: May 2024 Foundation Model Transparency Index

Subdomain	Indicator	Fuyu-8B	Jurassic-2	Luminous	Titan Text Express	Claude 3	StarCoder	Gemini 1.0 Ultra	Granite	Llama 2	Phi-2	Mistral 7B	GPT-4	Stable Video Diffusion	Palmyra-X
Data	Data size	0	1	1	0	0	1	0	1	1	1	0	0	1	1
	Data sources	0	1	0	0	0	1	0	1	0	1	0	0	0	1
	Data creators	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	Data source selection	0	1	0	0	0	1	0	1	0	1	0	0	0	0
	Data curation	0	1	1	0	1	1	0	1	1	0	0	1	1	1
	Data augmentation	0	1	1	0	0	1	0	1	1	1	1	0	1	1
	Harmful data filtration	0	1	1	0	0	1	0	1	1	0	1	1	1	1
	Data copyright status	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	Data license status	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	Personal information in data	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Data Labor	Use of human labor	0	1	1	1	0	1	1	1	1	1	1	0	1	1
	Employment of data laborers	0	0	1	0	0	1	0	1	0	1	1	0	1	1
	Geographic distribution of data laborers	0	0	1	0	0	1	0	0	0	1	1	0	1	1
	Wages	0	0	0	0	0	1	0	0	0	1	1	0	1	0
	Instructions for creating data	0	1	0	0	1	1	0	0	1	1	1	0	1	0
	Labor protections	0	0	1	0	0	1	1	0	0	1	1	1	1	0
	Third party partners	0	1	1	0	0	1	0	1	0	1	1	0	1	0
Data Access	Queryable external data access	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	Direct external data access	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Compute	Compute usage	0	1	1	0	0	1	0	1	1	1	0	0	0	1
	Development duration	0	1	1	0	0	1	0	1	0	1	0	0	0	1
	Compute hardware	0	1	1	0	0	1	0	1	1	1	0	0	1	1
	Hardware owner	1	1	1	0	1	1	1	1	1	1	1	1	0	1
	Energy usage	0	1	1	0	0	1	0	1	1	0	0	0	1	1
	Carbon emissions	0	1	1	0	0	1	0	1	1	0	0	0	1	1
	Broader environmental impact	0	0	1	0	0	1	0	1	0	0	0	0	0	0
Methods	Model stages	0	1	1	1	0	1	1	1	1	1	1	1	1	1
	Model objectives	0	1	1	1	1	1	1	1	1	1	1	1	1	1
	Core frameworks	0	1	1	0	1	1	1	1	0	1	1	0	0	1
	Additional dependencies	0	1	1	0	1	1	0	1	1	1	1	0	1	1
Mitigations	Mitigations for privacy	0	1	1	0	1	1	0	1	1	1	1	1	0	1
	Mitigations for copyright	0	1	0	0	0	1	0	1	0	0	0	0	0	1
	Upstream Subtotal	**3%**	**66%**	**66%**	**9%**	**22%**	**100%**	**19%**	**69%**	**47%**	**62%**	**47%**	**22%**	**53%**	**62%**

Foundation Model Transparency Index Indicator-Level Scores for Model, May 2024

Source: May 2024 Foundation Model Transparency Index

Subdomain	Indicator	Fuyu-8B	Jurassic-2	Luminous	Titan Text Express	Claude 3	StarCoder	Gemini 1.0 Ultra	Granite	Llama 2	Phi-2	Mistral 7B	GPT-4	Stable Video Diffusion	Palmyra-X
Model Basics	Input modality	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Output modality	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Model components	0	1	1	1	0	1	1	1	1	1	1	0	1	1
	Model size	1	1	1	0	0	1	0	1	1	1	1	0	1	1
	Model architecture	1	1	1	1	0	1	1	1	1	1	1	0	1	1
	Centralized model documentation	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Model Access	External model access protocol	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Blackbox external model access	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	Full external model access	1	0	1	0	0	1	0	0	1	1	1	0	1	0
Capabilities	Capabilities description	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Capabilities demonstration	1	1	1	1	1	1	1	0	1	1	1	1	0	1
	Evaluation of capabilities	1	0	1	1	1	1	1	1	1	1	1	1	1	1
	External reproducibility of capabilities evaluation	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Third-party capabilities evaluation	0	1	1	0	1	1	0	0	1	1	1	1	0	1
Limitations	Limitations description	1	1	1	1	1	1	0	1	1	1	0	1	1	1
	Limitations demonstration	0	1	1	1	0	0	0	0	0	0	0	0	0	0
	Third-party evaluation of limitations	1	1	1	1	1	1	1	1	1	1	1	1	1	0
Risks	Risks description	0	1	1	1	1	1	1	1	1	0	1	1	1	1
	Risks demonstration	0	1	1	1	1	1	1	0	1	0	0	1	0	0
	Unintentional harm evaluation	0	0	0	0	1	1	0	1	1	1	0	0	0	0
	External reproducibility of unintentional harm evaluation	0	1	1	1	1	1	0	1	1	1	0	0	0	0
	Intentional harm evaluation	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	External reproducibility of intentional harm evaluation	0	0	0	0	0	1	0	1	0	0	0	1	0	0
	Third-party risks evaluation	0	1	1	0	1	1	0	0	1	0	0	1	0	0
Mitigations	Mitigations description	0	1	1	1	1	0	1	1	1	0	1	1	0	1
	Mitigations demonstration	0	1	0	0	1	0	0	0	1	0	1	1	0	0
	Mitigations evaluation	0	0	0	0	0	0	0	1	1	0	1	1	0	0
	External reproducibility of mitigations evaluation	0	0	0	0	0	0	0	1	0	0	0	0	0	0
	Third-party mitigations evaluation	0	0	0	0	0	0	1	1	0	0	0	0	0	0
Trustworthiness	Trustworthiness evaluation	0	1	1	1	1	0	1	0	0	0	0	1	0	0
	External reproducibility of trustworthiness evaluation	0	1	1	0	0	0	0	0	0	0	0	0	0	0
Inference	Inference duration evaluation	0	1	1	0	0	1	0	0	1	0	0	0	1	1
	Inference compute evaluation	0	1	0	0	0	1	0	0	0	0	0	0	0	1
	Model Subtotal	42%	76%	76%	58%	64%	76%	55%	67%	76%	55%	58%	61%	48%	52%

Foundation Model Transparency Index Indicator-Level Scores for Downstream, May 2024

Source: May 2024 Foundation Model Transparency Index

Subdomain	Indicator	Fuyu-8B	Jurassic-2	Luminous	Titan Text Express	Claude 3	StarCoder	Gemini 1.0 Ultra	Granite	Llama 2	Phi-2	Mistral 7B	GPT-4	Stable Video Diffusion	Palmyra-X
Distribution	Release decision-making protocol	0	1	1	0	1	1	0	0	0	0	0	0	0	0
	Release process	0	1	1	1	1	1	0	1	1	1	1	1	1	0
	Distribution channels	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Products and services	1	1	1	0	1	1	1	1	1	0	1	1	1	1
	Machine-generated content	0	0	1	0	0	1	0	1	0	1	0	1	1	1
	Model License	1	1	1	1	1	1	1	1	1	1	1	0	1	1
	Terms of service	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Usage Policy	Permitted and prohibited users	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Permitted, restricted, and prohibited uses	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Usage policy enforcement	0	1	1	1	1	1	1	0	0	1	0	1	1	1
	Justification for enforcement action	0	1	1	1	1	1	1	0	0	1	0	0	0	0
	Usage policy violation appeals mechanism	0	1	1	0	1	1	1	0	0	1	0	1	0	1
Model Behavior Policy	Permitted, restricted, and prohibited model behaviors	0	1	1	1	1	1	1	0	1	1	1	1	1	1
	Model behavior policy enforcement	0	1	1	1	1	1	0	0	0	1	0	0	1	1
	Interoperability of usage and model behavior policies	0	1	1	1	1	1	0	0	0	1	0	1	1	1
User Interface	Usage disclaimers	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	User interaction with AI system	1	1	1	1	1	1	1	1	1	1	1	1	1	0
User Data Protection	User data protection policy	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Permitted and prohibited use of user data	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Usage data access protocol	1	1	1	0	0	0	1	1	1	1	1	0	1	0
Model Updates	Versioning protocol	1	1	1	1	1	1	1	1	1	1	1	1	1	0
	Change log	1	1	1	0	0	1	0	1	1	1	1	1	1	0
	Deprecation policy	1	1	1	0	0	1	1	1	1	1	1	1	1	1
Feedback	Feedback mechanism	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Feedback summary	1	1	0	0	0	1	0	1	0	1	1	0	1	0
	Government inquiries	0	1	1	0	0	1	1	0	0	0	0	0	0	0
Impact	Monitoring mechanism	1	1	1	0	1	0	1	0	0	0	1	1	0	1
	Downstream applications	1	0	0	0	0	1	0	0	1	0	0	0	1	0
	Affected market sectors	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Affected individuals	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Usage reports	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Geographic statistics	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Redress mechanism	0	1	1	0	0	0	1	0	0	0	0	0	0	0
Downstream Documentation	Centralized documentation for downstream use	0	1	1	1	1	1	1	1	1	1	1	1	1	1
	Documentation for responsible downstream use	0	1	1	1	1	1	1	1	1	0	1	1	1	1
	Downstream Subtotal	**51%**	**83%**	**83%**	**54%**	**66%**	**80%**	**66%**	**57%**	**57%**	**69%**	**60%**	**63%**	**71%**	**54%**