Rethinking open source generative AI: open-washing and the EU AI Act

Andreas Liesenfeld*
Mark Dingemanse*
andreas.liesenfeld@ru.nl
mark.dingemanse@ru.nl
Centre for Language Studies, Radboud University
Nijmegen, The Netherlands

ABSTRACT

The past year has seen a steep rise in generative AI systems that claim to be open. But how open are they really? The question of what counts as open source in generative AI is poised to take on particular importance in light of the upcoming EU AI Act that regulates open source systems differently, creating an urgent need for practical openness assessment. Here we use an evidence-based framework that distinguishes 14 dimensions of openness, from training datasets to scientific and technical documentation and from licensing to access methods. Surveying over 45 generative AI systems (both text and text-to-image), we find that while the term open source is widely used, many models are 'open weight' at best and many providers seek to evade scientific, legal and regulatory scrutiny by withholding information on training and fine-tuning data. We argue that openness in generative AI is necessarily composite (consisting of multiple elements) and gradient (coming in degrees), and point out the risk of relying on single features like access or licensing to declare models open or not. Evidence-based openness assessment can help foster a generative AI landscape in which models can be effectively regulated, model providers can be held accountable, scientists can scrutinise generative AI, and end users can make informed decisions.

CCS CONCEPTS

• **Information systems** → **Open source software**; • **General and reference** → Surveys and overviews.

KEYWORDS

Technology assessment, large language models, text generators, text-to-image generators

ACM Reference Format:
Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative AI: open-washing and the EU AI Act. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, June 03–06, 2024, Rio de Janeiro, Brazil. ACM. doi: 10.1145/ 3630106.3659005

*Both authors contributed equally to this research.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs International 4.0 License.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659005

1 INTRODUCTION

Open generative AI systems are on the rise, with small players and academic initiatives leading the way in open innovation and scientific documentation [20, 32, 61] and several larger corporations joining the fray by releasing models billed as 'open'. But there are three critical challenges to openness in the domain of generative AI systems. The first is that openness is not a binary feature: today's transformer-based system architectures and their training procedures are complex, and they can only be classified into open or closed at the price of severe information loss. Secondly, some systems are open in name only. Ubiquity and free availability are not equal to openness and transparency [3, 9, 21]. For instance, over the past few years, many research teams have relied on OpenAI's popular `davinci` text generation models only to find them deprecated, jeopardising the reproducibility of large amounts of scientific work [43]. The third challenge is a growing number of models that are open in weights only, where model weights are made available under an open licence yet most other aspects of how the system was built are kept under wraps. This practice of open-washing has the effect of compromising professional standards in software and technology development, moving the field away from core tenets of the open source movement like reverse-engineerability and full transparency.

These challenges are compounded by impending legislation. In 2024, the AI landscape will be shaken up by the EU's AI Act, the world's first comprehensive AI law, with a projected impact on science and society comparable to GDPR [48]. Fostering open source driven innovation is one of the aims of this legislation. This means it will be putting legal weight on the term "open source" [48], creating only stronger incentives for lobbying operations driven by corporate interests to water down its definition. The latest version of the EU AI Act features what may turn into an alarming exemption: it allows models released under open licences to forego detailed disclosure of training data and fine-tuning methods, while outsourcing the definition of what is open to a (yet to be established) EU AI Office [27]. This makes clarity about the meaning of open source all the more urgent.

Our goal in this paper is to make a number of critical and constructive contributions. We discuss evidence of open-washing and the deleterious effects it has on the open technology landscape; review current strategies to redefining the notion of open source in

light of the EU AI Act; and propose that the most fruitful conceptions of openness in generative AI must be composite (recognizing that AI systems are made of many moving parts) and gradient (recognizing that openness is not a simple binary property). A survey of 40 large language models and 6 text-to-image models reveals the most important trends in current open generative AI and shows how evidence-based openness assessment can work in practice.

1.1 Open source in the context of the EU AI Act

The EU AI Act [44] will impose on general purpose AI providers a number of potentially burdensome requirements, including going through a conformity assessment, providing human oversight, and providing technical documentation that includes detailed information on system architecture, training datasets, provenance and curation (Annex IV in [44]). This is a major improvement over the current regulatory landscape, where models have been allowed to proliferate under the murkiest of legal conditions and with little to no regulatory oversight [7, 49].

A special feature of the EU AI Act is the importance it accords to open source. In view of the notion that open models can contribute to research and innovation, the EU AI Act provides a number of exemptions for such models, meaning that at least some of the burdensome requirements mentioned above can be escaped by attaining open source status. Key passages in this regard appear in section §60 of the latest iteration:[1]

- "Software and data, including models, released under a free and open-source licence that allows them to be openly shared and where users can freely access, use, modify and redistribute them or modified versions thereof, can contribute to research and innovation in the market and can provide significant growth opportunities for the Union economy." (§60i)
- [models released under a free and open-source licence] "should be subject to exceptions as regards the transparency-related requirements imposed on general purpose AI models." (§60f)

What exactly are these exceptions? Under the latest version of the Act, providers of AI models "under a free and open licence" are exempted from the requirement to "draw up and keep up-to-date the technical documentation of the model, including its training and testing process and the results of its evaluation, which shall contain, at a minimum, the elements set out in Annex IXa" (Article 52c:1a). Instead, they would face a much vaguer requirement to "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model according to a template provided by the AI Office" (Article 52c:1d).

If this exemption or one like it stays in place, it will have two important effects: (i) attaining open source status becomes highly attractive to any generative AI provider, as it provides a way to escape some of the most onerous requirements of technical documentation and the attendant scientific and legal scrutiny; (ii) an as-yet unspecified template (and the AI Office managing it) will become the focus of intense lobbying efforts from multiple stakeholders (e.g., [12]). Figuring out what constitutes a "sufficiently detailed summary" will literally become a million dollar question.

[1]We quote from the February 24, 2024 version at www.europarl.europa.eu

While minor adjustments may still be made, it is clear that any decisions to be made here, both in the EU AI Act and in the requirements and templates for technical documentations surrounding it, will be hugely consequential for the open technology landscape. We are not the first to note this. Legal scholar Kate Downing has described the current version of the EU AI Act as "choose your own adventure" [15] when it comes to openness. The most important gap appears to be that there is very little reference to training datasets. Indeed, sociologist and open policy advocate Alek Tarkowski has warned that the current EU AI Act "fails to set meaningful dataset transparency standards" [56].

In sum, with the enshrinement of open source in the AI Act, the term "open source" comes to carry unprecedented legal weight. Can it bear this weight? In what follows, we briefly review how the open source community has been responding to these developments.

1.2 The moving target of open source AI

Until recently, classifications of software as open source could simply rely on the availability of source code under appropriate licensing: if some software is released under a licence approved by the Open Source Initiative (OSI), it means that it is fully open and minimally restrictive [45]. For software that is relatively portable and user-deployable, this was long sufficient, and it afforded users the rights to make copies, to tinker and, and to make improvements. However, the rise of large language models and text-to-image generators means that a different approach is needed [47].

There are various efforts underway to update and tailor the definition of open source to current generative AI systems. One is a public consultation dubbed the "Open Source AI Deep Dive" that the OSI board may draw on in their efforts to update their definition of open source in the age of generative AI.[2] Another is the "Joint Statement on AI Safety and Openness" by parties including Creative Commons, Mozilla, LAION and Open Future.[3] The challenge that these efforts face is to adopt the notion of open source, which used to be fairly unambiguous, to the increasingly complex world of generative AI systems [34].

The sheer amount of training data, the architectural complexity, and the many moving parts make full openness a tall order for generative AI [53]. The compute needed for training comes with costs that are within reach of none but a handful of larger corporations or government entities [1]. Human labour can be involved at multiple points, from reinforcement learning datasets to crowd-sourced ratings. Comprehensive documentation of such complex systems requires serious effort [20]. And opening up data and training pipelines might pose not only a commercial risk but also lead to legal exposure. Naturally, many actors in the field opt to err on the side of caution and only disclose elements of their system as required [4].

Against this backdrop of a rapidly-evolving technological and regulatory landscape, the very notion of open source AI is a moving target. While current efforts to arrive at a new open source definition for AI are useful, we observe they are strongly focused on the question of licensing. This makes sense: the OSI has been very successful in shepherding the notion of open source licences. However,

[2]https://opensource.org/deepdive
[3]https://open.mozilla.org/letter/

if everything hinges on licensing, what is to stop model providers from releasing the most inscrutable portion of their system (say, model weights) under an OSI-approved licence and collect open source benefits? This stands to be a major avenue for open-washing.

There is nothing imaginary about this. For one, we are already seeing a strong trend towards selective and self-serving forms of openness, both in release strategies and in empirical surveys of model openness ([32, 53, 60], and see below). Also, other sectors offer well-known precedents of such dynamics. Take fair trade coffee. The international fair trade movement started as a grassroots effort that directly empowered local farmers and waged labourers by giving them fairer terms in global commodity markets. Within a decade, it was adopted by multinationals like Starbucks and Nestlé, who turned it into an efficient marketing tool: green-washing at work. Sociological research has established that this resulted in these multinationals effectively co-opting the notion of fair trade. This work provides an important lesson on the playbook of corporate lobbies: "co-optation ... occurs primarily on the terrain of standards, in the form of weakening or dilution" [26].

We conclude that if open licensing becomes the sole deciding factor for model openness, the open source community faces the risk that community standards will be co-opted and diluted just as Starbucks and Nestlé have co-opted and diluted the notion of fair trade in coffee. Open licensing could become an empty gesture.

1.3 Open-washing and the release-by-blogpost model

Companies operating in the generative AI space currently appear to be converging on a strategy known as open-washing [60]: collect brownie points for openness without disclosing critical information of training and tuning procedures, thereby largely escaping the scientific scrutiny and legal exposure that would come with full openness.

A key sign of open-washing is the growth of what we call a 'release by blogpost'-strategy. The bulk of open generative AI models released in the past year were first made public in a blogpost or press release touting their openness. For instance, TII's Falcon 70B was introduced as the "top-ranked open-source AI model"[4], StabilityAI's Stable Beluga as "open access"[5], Mistral proclaims "we have the best open source models"[6], 01.ai's Yi 34B Chat claims to be the "next generation of open-source and bilingual LLMs"[7], and Alibaba claimed "we opensource our Qwen series"[8]. Probably the strongest claims to the label open source (if not its content) come from Meta and its Llama 2 and Llama 3 models. The corporate blogposts that introduced Llama made the following claims: [9]

- "Today, we're introducing the availability of Llama 2, the next generation of our open source large language model." (Llama 2)
- "Meta has put exploratory research, open source, and collaboration with academic and industry partners at the heart of our AI efforts for over a decade." (Llama 2)
- "Today, we're introducing Meta Llama 3, the next generation of our state-of-the-art open source large language model." (Llama 3)

Arguably, companies can announce and market their products however they want. And when they do, they show their hand. In this case, we can conclude that Meta and other companies in this space see a particular benefit in putting a claim on the term open source. We will see below why this may be.

Alongside claims of openness, the release-by-blogpost model typically features some nicely laid out tables that compare the model to a selection of its competitors on a selection of scientific benchmarks like MMLU, HUmanEval, TruthQA and the like. These evaluation tables, clearly modelled after NLP's coveted SOTA tables [10], allow releases to retain the veneer of scientific work while at the same time avoiding the fine-grained accounting and the scrutiny of peer review that comes with actual scientific publication. These tables also offer ample degrees of freedom for cherry-picking, enabling model providers to present their products in the best light possible. Without technical documentation and peer review, the release-by-blogpost model is little more than pseudoscience.

When generative AI follows the release-by-blogpost model, it is reaping the benefits of mimicking scientific communication —including associations of reproducibility and rigour [21, 23]— without actually doing the work. And when generative AI co-opts the term open source, it is reaping the benefits of *libre* culture —including associations of transparency and associated freedoms [45, 60]— without actually contributing to the commons. This is how open-washing works. There is ample evidence that as a communication strategy, open-washing is highly effective. The launches of Llama 2 and Llama 3 were greeted with considerable excitement in mainstream media outlets, almost without exception uncritically echoing the open source claim as a major selling point. For instance, a Wired headline of April 2024 claims, "Meta's Open Source Llama 3 Is Already Nipping at OpenAI's Heels"[10], and Fortune wrote, "Meta releases its new Llama 3 open-source AI model. Is it enough to keep Meta at the front of the pack?"[11].

The consequences of open-washing are considerable and affect multiple stakeholders. Open-washing stands in the way of innovation, because if large corporations can derive benefits from the trappings of open source without doing the requisite work, this sucks up the oxygen in the open source ecosystem, making it less attractive for smaller entities to find funding for truly open projects [1]. Open-washing is also bad for research, because it means that researchers can no longer count on being able to tinker with models and architectures even if they are advertised as open source [54]. And open-washing is bad for the public understanding of AI, because it creates artefacts designed to impress without providing people with the resources to reach a deeper understanding of the technology [39].

Open-washing is related to the notion of audit-washing [19]: the risks posed by poorly designed auditing procedures. As Goodman and Tréhu note, "Audits without clear standards provide false

[4] falconllm.tii.ae/falcon.html
[5] stability.ai/news/stable-beluga-large-instruction-fine-tuned-models
[6] mistral.ai/technology/
[7] 01.ai
[8] alibabacloud.com/en/solutions/generative-ai/qwen
[9] ai.meta.com/blog/llama-2, ai.meta.com/blog/meta-llama-3
[10] wired.com/story/metas-open-source-llama-3-nipping-at-openais-heels
[11] fortune.com/2024/04/18/meta-ai-llama-3-open-source-ai-increasing-competition

assurance of compliance" (p. 3). As current and impending regulation increasingly puts legal weight on the notion of open source generative AI, the spectre of false assurance looms large.

Our main goal in what follows is to propose a conception of openness that can serve the EU AI Act's goal to foster research and innovation, that can help to specify what makes a sufficiently detailed summary, and that can provide key building blocks for a model openness template.

2 OPENNESS ASSESSMENT IN ACTION: A COMPOSITE AND GRADIENT APPROACH

If open source is given legal weight under the EU AI Act (§1.1), its definition is a moving target (§1.2), and open-washing is a rising challenge (§1.3), what would be the best way to carry out openness assessment? We maintain that given the complexity of generative AI, the most productive approaches will see openness as **composite** and **graded**. Composite, because it is made up of multiple elements, each of which can be assessed. Graded, because each element itself can be realised with different degrees of openness, and it is no longer feasible to maintain a simple open/closed binary.

The composite and gradient nature of openness can be grounded directly in prior work on openness and accountability in AI systems [29, 37, 42, 58]. Practical approaches to implement such ideas typically revolve around the systematic collection and curation of data on various aspects of systems [2, 28, 46]. Important constituent elements are the notions of data sheets [18, 25, 38], model cards [40], and system cards [22]: frameworks that can help structure the systematic presentation of metadata about models and systems and that feed into auditing procedures [37]. The multi-faceted nature of openness is also seen in initiatives that aim to codify transparency from upstream resources to downstream uses [8] and in a proposal like the Linux Foundation's model openness framework [59], which shows many likenesses to the dimensions we use here and introduced in [32]. The necessarily gradient nature of openness is seen in release methods, where prior work has pointed out tradeoffs between full openness and risk mitigation [16, 53].

Here we use and extend a framework for model openness and transparency that has been tried and tested since July 2023 in an openness leaderboard that tracks degrees of openness for a growing number of generative AI models.[12] The auditing procedure is itself fully open to public scrutiny and to community contribution (see Appendix). Because of its openness, the framework doubles as a possible infrastructure for auditing and a public service to foster AI literacy [24, 30]. We use this approach in a systematic sweep of the current generative AI landscape, focusing on 40 text generators and extending the scope to 6 text-to-image generators.

2.1 Key elements of an openness matrix: a demonstration using BloomZ and Llama

Ultimately assessment must be evidence-based, and a key question therefore is what the building blocks of openness assessment should be.

We provide a matrix of relevant dimensions of openness in generative AI, each grounded in evidence-based judgements of degrees of openness. The dozen or so dimensions we have identified here along with three levels of openness (open, partial, closed) provide a sufficient level of detail to provide well-informed, high quality, systematic judgements of openness in generative AI.

Here we provide a quick walkthrough of all features by means of a comparison of two systems that are both billed as open source: BloomZ (Bloom for short), introduced by the BigScience Workshop team in May 2023 as an open source chat LLM [61], and Llama 2 (Llama), introduced by Meta as "the next generation of our open source large language model" as we saw above.[13] As we will show, our method provides a way to form a nuanced and evidence-based judgement of the truth value of this claim (Figure 1).

Availability. When it comes to *open code*, we find that BloomZ makes available source code for training, fine-tuning and running the model, while for Llama none of the model's source code is made available, only scripts for running the model are shared. The *LLM data* that was used to train the base model is documented in great detail by Bloom [61], while for Llama only the vaguest details are provided in a corporate preprint: "a new mix of data from publicly available sources, which does not include data from Meta's products or services" [57]. The statement is clearly designed to minimise legal exposure. Both systems make the *LLM weights* of the base model available, though for Llama access is restricted through a consent form. The training data for instruction tuning (*RL data*) is described and documented by Bloom as consisting of xP3 (Crosslingual Public Pool of Prompts); for Llama, the corporate preprint notes that fine-tuning was done based on "a large dataset of over 1 million binary comparisons based on humans applying our specified guidelines, which we refer to as Meta reward modeling data", and which remains undisclosed. (The same preprint mentions that for evaluation, Meta did build on several RLHF datasets openly shared by others.) Model weights for the instruction-tuned version (*RL weights*) are made openly available by BloomZ, while for Llama they require an access request.

Documentation. The BloomZ project *code* is well-documented and actively maintained, while for Llama 2 no documentation of source code is available as the source code itself is not open. The *architecture* is described for BloomZ in multiple scientific papers and supported by a github repository of code and recipes on HuggingFace; for Llama, the architecture is described in less detail and scattered across corporate websites and a preprint.

BloomZ's multiple *preprints* document data curation and fine-tuning [50, 61] in great detail; in contrast, Llama's single preprint offers fewer details and appears strategically vague on crucial details (for instance, training datasets and instruction tuning). The scientific documentation of BloomZ also includes multiple *peer-reviewed papers*, from a scientific description of the multitask fine-tuning procedures [41] to an estimation of the carbon footprint [36] — currently one of the very few scientifically vetted sources of data on the energy footprint of training large language models. No peer-reviewed papers providing scientific documentation or evaluation of Llama are known currently.

[12] https://opening-up-chatgpt.github.io

[13] Llama 3, which was released by blogpost in April 2024, is included in the overview below and is no different from Llama 2 in terms of openness.

Project	Availability						Documentation						Access	
(maker, bases, URL)	Open code	LLM data	LLM weights	RL data	RL weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✓	✓	✓	✗	✓
LLaMA2 Chat	✗	✗	~	✗	~	✗	✗	~	~	✗	~	✗	✗	~

Figure 1: Comparison of BloomZ and Llama 2 on 14 dimensions of openness, illustrating the framework.

BloomZ also released *model cards* that describe the architecture and evaluation results with extensive cross-references to other documentation on training data, training approach, model architecture, fine-tuning and responsible use. In contrast, the Llama model card only provides minimum detail and none whatsoever on training data. A *data sheet* is only available for BloomZ. This means that for Llama, there is no documentation of training datasets whatsoever — a prime example of a strategy described by Birhane et al. as a tactical template of "(non)declaring the training dataset information" [4].

Access and licensing. Neither BloomZ nor Llama distribute models as software *packages* via indexed and version controlled public code repositories such as pypi. Instead, both are primarily intended for local deployment. BloomZ is available through the Petals *API*, while for Llama an API is only available behind a privacy-defying signup form. Finally, the models also differ in terms of *licensing*. BloomZ has two relevant licences. Its source code is Apache 2.0, an OSI-approved open source licence, while the model weights are released under the Responsible AI Licence (RAIL) [13]. Llama 2 is released under Meta's own Community Licence. Both licences aim to restrict harmful use cases, but there is a key difference in how they implement how model outputs are to be represented. RAIL stipulates that a user may not "generate content without expressly and intelligibly disclaiming that the text is machine-generated", while Llama stipulates that a user may not "represent that Llama 2 outputs are human-generated" — a much lower bar, because it leaves open a wide swathe of use cases where there may not be the explicit claim of human-generated output, but merely a strong implication.

This walkthrough shows that drilling into the details of generative AI systems using the dimensions of openness of our framework makes critical differences visible. Only BloomZ can substantially claim open source status, while Meta's Llama is at best open weights, and is closed in almost all other aspects. Llama, in all currently available versions, is a prime example of a model that claims openness benefits by merely providing access to its most inscrutable element: model weights.

2.2 The current open generative AI landscape

With a first view of the framework in hand we can extend our survey to a larger sample of generative AI systems. We focus on models that bill themselves as open, aiming to include well-known players but also small models and work by smaller teams or organisations, some of which make up for their lack in size and model performance with high standards of openness and transparency. Every single openness judgement is directly linked to publicly available evidence, and all data points are available in a versioned data repository.[14]

[14]Supplementary data repository via Open Science Foundation: https://osf.io/f2b7n

Therefore we only focus on describing the most important findings and trends.

2.2.1 Text generators: evading training data disclosure and scientific scrutiny. Our survey yields 40 text generators that are described as "open source" or "open". We examine each system for openness using the assessment framework and rank the systems by openness score. As a reference, we also add ChatGPT. The result is an overview of the current state of openness in text generators (Figure 2).

We observe two broad ways of working. One is the broad open source approach seen in systems like AllenAI's OLMo Instruct [20], BloomZ [61] and LLM360's AmberChat [33], which are approaching full openness status and top the openness leaderboard. The organisations behind these systems have gone to great lengths to make training data, code, training pipelines, and documentation available.

We also find a large number of systems (roughly the bottom third) that make only model weights available but share little to no detail about other parts of their system. These systems are best called open weight rather than open source. Compared to the closed baseline of OpenAI's ChatGPT, some of these systems are barely more open. It is noteworthy that all of the big commercial players —Meta, Google, Cohere, Microsoft and Mistral— are occupying the lower ranks, as are many alternatives that build on them.

We conclude that the current state of openness in text generators is mixed. A few very open systems exist, but the most well-known models are open weights only. Many systems share little information about instruction tuning steps or metaprompting techniques. Datasets and methods for training and fine-tuning are rarely shared or disclosed. System, data and code documentation is often incomplete and lacks academic rigour. Peer-reviewed papers seem to have almost completely fallen out of fashion and are increasingly replaced by blogposts with cherry-picked examples or corporate preprints with minimal detail. If there are technical reports, they tend to focus on performance evaluation at the expense of documenting system architecture and training data.

The lack of openness about training data is particularly worrying. Most models in the bottom half do not provide any details about datasets beyond very generic descriptors obviously designed to evade legal scrutiny.

2.2.2 Text-to-image generators: mostly closed. In the same fashion, we assess 6 text-to-image generators (Figure 3). Again, we add Open AI's DALL-E models as a closed reference. Overall, the survey yields far fewer systems in comparison to text generators. A possible reason for this might be that relatively few image datasets are available. Text-to-image generators also differ in terms of their machine learning architecture. For instance, image generators do

Project	Availability						Documentation						Access	
	Open code	LLM data	LLM weights	RL data	RL weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
OLMo 7B Instruct	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	~
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✓	✓	✓	✗	✓
AmberChat	✓	✓	✓	✓	✓	✓	~	~	✓	✗	~	~	✗	✓
Open Assistant	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✓
OpenChat 3.5 7B	✓	✗	✓	✗	✓	✓	~	✓	✓	✓	~	✗	✓	~
Pythia-Chat-Base-7…	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	~	~	✓	✗
Cerebras GPT 111…	~	✓	✓	✓	✓	~	✗	✓	~	✗	✗	✓	✗	✓
RedPajama-INCITE…	~	✓	✓	✓	✓	~	~	~	✗	✗	✓	✓	✗	~
dolly	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✗
Tulu V2 DPO 70B	✓	✗	~	✓	✓	~	~	~	✓	✗	~	~	✗	✓
MPT-30B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	~	✗	✓	~
MPT-7B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	✓	✗	✓	✗
trlx	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	~	✓
Vicuna 13B v 1.3	✓	~	✓	✗	✗	~	✓	✗	✓	✗	~	✗	✓	~
minChatGPT	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	✗	✓
ChatRWKV	✓	~	✓	✗	✗	✓	~	~	~	✗	✗	✗	✓	~
BELLE	✓	~	~	~	~	✗	~	✓	✓	✗	✗	~	✗	✗
WizardLM 13B v1.2	~	✗	~	✓	✓	~	~	✓	✓	✗	✗	✗	✗	✗
Airoboros L2 70B G…	~	✗	~	✓	✓	~	~	~	✗	✗	~	~	✗	✗
ChatGLM-6B	~	~	✓	✗	✗	✓	~	~	✗	~	✗	✗	✗	✓
Mistral 7B-Instruct	~	✗	✓	✗	~	✓	✗	~	~	✗	✗	✗	~	✓
WizardLM-7B	~	~	✗	✓	~	~	~	✓	✓	✗	✗	✗	✗	✗
Qwen 1.5	~	✗	✓	✗	✓	✗	~	~	✗	✗	✗	✗	~	✓
StableVicuna-13B	~	✗	~	~	~	~	~	~	~	✗	~	✗	✗	~
Falcon-40B-instruct	✗	~	✓	~	✗	✓	✗	~	~	✗	~	✗	✗	✗
UltraLM	✗	✗	~	✓	~	✗	✗	~	✓	✗	~	~	✗	✗
Yi 34B Chat	~	✗	✓	✗	✓	~	✗	✗	✓	✗	✗	✗	✗	~
Koala 13B	✓	~	~	~	✗	~	~	~	✗	✗	✗	✗	✗	✗
Mixtral 8x7B Instruct	✗	✗	✓	✗	~	✓	✗	~	~	✗	✗	✗	~	✗
Stable Beluga 2	✗	✗	~	✗	✓	~	✗	~	~	✗	~	✗	✗	~
Stanford Alpaca	✓	✗	~	~	~	✗	~	✓	✗	✗	✗	✗	✗	✗
Falcon-180B-chat	✗	~	~	~	~	✗	✗	~	~	✗	~	✗	✗	✗
Orca 2	✗	✗	~	✗	✓	✗	✗	~	~	✗	~	✗	✗	~
Command R+	✗	✗	✗	✓	✓	~	✗	✗	✗	✗	~	✗	✗	✗
Gemma 7B Instruct	~	✗	~	✗	~	✗	✗	~	✗	✗	✓	✗	✗	✗
LLaMA2 Chat	✗	✗	~	✗	~	✗	✗	~	~	✗	~	✗	✗	~
Nanbeige2-Chat	✓	✗	✗	✗	✓	~	✗	✗	✗	✗	✗	✗	✗	~
Llama 3 Instruct	✗	✗	~	✗	~	✗	✗	~	✗	✗	~	✗	✗	~
Solar 70B	✗	✗	~	✗	~	✗	✗	✗	✗	✗	~	✗	✗	~
Xwin-LM	✗	✗	~	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	~
ChatGPT	✗	✗	✗	✗	✗	✗	✗	✗	~	✗	✗	✗	✗	✗

Figure 2: Openness of 40 text generators described as open, with OpenAI's ChatGPT (bottom) as closed reference point. Every cell records a three-level openness judgement (✓ open, ~ partial or ✗ closed). The table is sorted by cumulative openness, where ✓ is 1, ~ is 0.5 and ✗ is 0 points. RL may refer to RLHF or other forms of fine-tuning aimed at fostering instruction-following behaviour. For the latest updates see: https://opening-up-chatgpt.github.io

Project	Availability						Documentation						Access	
(maker, bases, URL)	Open code	Training data	Model weights	Watermarking	Prompt mod	Licensing	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
Stable Diffusion	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	~	~
Deep Floyd	~	~	✘	✘	✘	~	~	~	~	✘	✔	✘	✔	✘
Invoke AI	~	✘	✔	✘	~	✔	✘	✘	✘	✘	✘	✘	✘	✘
Craiyon / DALL-E Mini	✘	✘	✔	✘	✘	✔	✘	✘	✘	✘	✔	✘	✘	✘
Dream Shaper	✘	✘	✔	✘	✘	✔	✘	✘	✘	✘	✘	✘	✘	✘
OpenJourney	✘	✘	~	✘	✘	~	✘	✘	✘	✘	✘	✘	✘	✘
DALL-E /2 /3	✘	✘	✘	✘	✘	✘	✘	✘	~	✘	✘	✘	✘	✘

Figure 3: Overview of 6 text-to-image systems described as open, with OpenAI's DALL-E as a reference point. Every cell records a three-level openness judgement (✓ open, ~ partial or ✗ closed). The table is sorted by cumulative openness, where ✓ is 1, ~ is 0.5 and ✗ is 0 points.

not generally implement instruction-tuning, a key component of text generators.

Most relevant for evidence-based openness assessment are the ways in which text-to-image generators implement ways of tracking provenance of synthetic imagery and set up guardrails against creating undesired content. Some systems use *watermarking* to enable a form of provenance tracking. For moderation, text-to-image systems commonly rely on forms of *prompt moderation*, often text filtering or classification. The status of such provenance and safety measures is not always documented, and this is what the openness judgements seek to capture. (Thus, a model may implement prompt moderation but also document it; in that case, it counts as open for that dimension. In contrast, when a model may or may not watermark its output, and does not disclose this either way, it counts against openness.) The respective dimensions of the assessment framework have been adjusted accordingly.

One system stands out when it comes to openness, transparency and documentation: Stable Diffusion by Stability AI, Runway and the Computer Vision and Learning Group at Ludwig Maximilians Universitat Munich, Germany. Some of the other assessed systems build on or fine-tune the various models of Stable Diffusion. Some other systems are open-weight only. Open AI's DALL-E is completely closed.

This means that those interested in text-to-image generators face a relatively clear choice where to look for a very open alternative to proprietary and closed products. It also means that only Stable Diffusion has been open to scrutiny from scientists, regulators and the general public. This has enabled auditing of the underlying datasets, which has revealed legal challenges [49] and deeply problematic content [4, 5].

2.3 Turning evidence-based assessments into openness scores or labels

A key goal of our work is to provide flexible ways for the evidence-based assessment of openness. While the focus is on evidence-based expert judgements of degrees of openness supported by public documentation, the resulting fine-grained data must sometimes be translated into more reductive scores, labels, or classifications.

There are multiple ways to turn fine-grained openness assessments into metrics of openness that support classification or comparison (Figure 4). One is to assign weights to openness classes and to derive, based on this, a cumulative openness score (Figure 4, panel 2). This would be a gradient measure of openness. For simplicity and transparency, here we have picked weights of 1 (open), 0.5 (partial) and 0 (closed) and we have weighted all dimensions equally. Different choices are possible. For instance, in situations where it is important to know exactly what is in the training data or how the instruction-tuning was carried out, these dimensions might be weighted more heavily, penalising models that are less open.

From such a gradient measure, further classifications can be derived. One would be to divide the continuum into separate categories, comparable to the EU energy label system (Figure 4, panel 3). Here too, the simplest approach would be to divide up the space evenly, but different weightings could be used to discretise the space in ways that are more fitting to particular purposes. A third, closely related approach would be to reduce the continuum to a simple dichotomous classification of models into open versus closed (Figure 4, panel 4). The figure makes visible how this cannot be anything else than a gross oversimplification: a dense multidimensional field of openness measures is reduced to a simple binary classification.

Separate from these approaches is a fourth method, increasingly popular but even more reductive (Figure 4, panel 5). This is to single out only a single measure and base an openness classification on that. This is in effect what a focus on open weight models or on open licences accomplishes. If we are satisfied with calling such systems open, we are discarding many relevant dimensions and degrees of openness just to arrive at a single binary classification. What can also be seen is how privileging one such measure may distort the overall picture: there are many open weight models currently, and so focusing on this one dimension makes it seem as if almost half of all text generators are open. By contrast, if we were to focus on the all-important instruction-tuning, by many accounts the secret sauce of chat LLMs, there would be only a handful of models that offer the requisite openness.

We take time to discuss these ways of turning rich openness assessments into reductive metrics because it is important to be aware of the distorting effects of metrics [55]. Metrics can be gamed; indeed important parts of present-day NLP and machine learning could be described as finding clever ways of gaming metrics, whether automated or human [10]. This figure therefore offers both a manual for lobbyists and the means to counter them. It can be

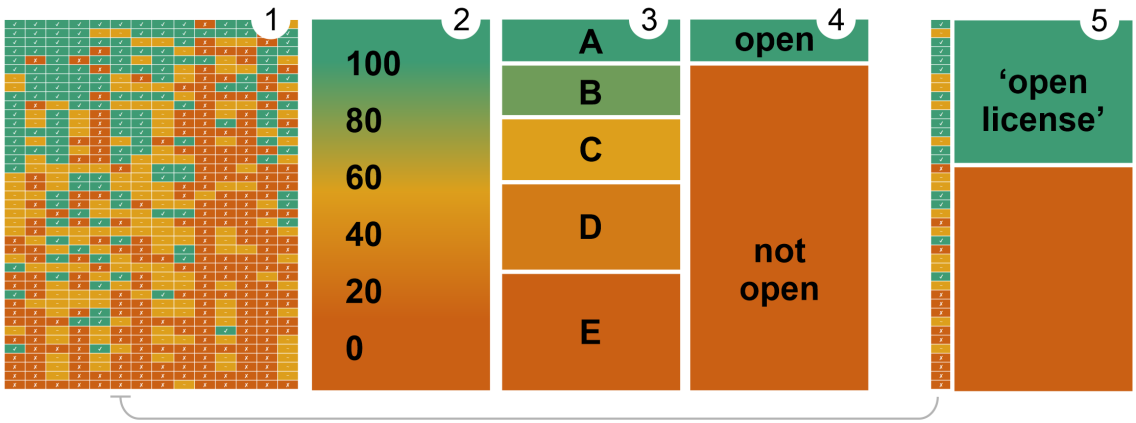

Figure 4: Openness judgements (1) can be turned into actionable metrics in several ways: by assigning weights to openness classes and to specific features, a cumulative openness score can be derived: a gradient measure of openness (2). Any gradient can be discretised into disjunct categories comparable to energy labels (3). Or such data can be turned into a dichotomous measure of openness (4). Each step is increasingly reductive, abstracting away from the full evidence by discretising, categorising and dichotomising it. Yet another method, increasingly popular and maximally reductive, is to base an openness judgement only on a single measure, for instance the mere availability of model weights or an open licence (5). Because this obscures the multidimensional and graded nature of openness, this is one of the most effective methods for open-washing.

predicted that corporate interests will argue for the easiest-to-attain openness dimensions to weigh most heavily — indeed this is likely one of the reasons behind the 'open weight branded as-open source' approach of companies like Meta and Mistral, and behind lobbying attempts to favour simple licence-based decisions. At the same time, knowing this, regulatory entities can make informed decisions on which forms of openness should count.

3 DISCUSSION

Openness is one of the key features that enables transparency and accountability. For present-day generative AI systems, it is not an all-or-nothing feature; rather, it is best conceived as a composite and gradient notion [53]. The place of a system on this gradient is determined by the relative openness of its constituent elements. We have formulated the key elements of a framework that conceptualises AI openness in such a gradient and composite way. When properly implemented, such a framework can safeguard against open-washing by enabling public scrutiny.

Surveying the field of generative AI, we have identified some broad trends. While a first crop of text generators —including BloomZ and OpenAssistant— clearly aimed at meaningful degrees of openness, soon enough large corporate players started releasing systems billed as open source while they were in fact at best open weight, significantly diluting the term [60]. Two influential companies in this domain, Meta and Mistral, have collectively dragged the average level of openness down simply through the ubiquity of their model weights. We find that some smaller players have dropped the development of their own models and now simply plug-and-play Meta or Mistral models, inheriting these systems' lack of transparency about training data and architecture of the model. More worryingly, we might even be seeing a general downturn in efforts to build truly open alternatives to merely open weight systems: well-funded corporate heavyweights are taking oxygen out of the room for smaller organisations that operate with higher professional and ethical standards.

Our survey also brings to light a difference in licensing trends between text generators and text-to-image generators. Whereas the latter often are released under a Responsible AI Licence, restricting harmful uses and adding one more layer of safety and accountability, the most common licences for text generators are classic open source licences like Apache 2.0, presenting no restrictions on use of the technology. It seems makers of text-to-image models have been more aware of harmful uses and legal exposure that may result from use of their models, while makers of text generators seem less concerned about harmful applications. Perhaps this is because for image generators, open datasets like LAION have enabled serious auditing. Such audits have revealed the presence of misogyny, pornography and harmful stereotypes [5, 6] and may have made makers of such models more aware of the harmful uses to which their systems can be put. Another possible reason is that copyright violations in images are easier to pinpoint and prosecute in the legal system [52].

We also want to highlight the meaningful contributions by smaller players and non-corporate entities. In fact, these hidden champions of the generative AI world are where progress towards more open systems is most likely. Sidestepping the toxicity of performance benchmarks and a bigger-is-always-better logic, these small but open models can be just as useful for many end users. Ordinary use cases often do not require the latest gargantuan models [51]. Just like your Ferrari is better left in the garage when you go grocery shopping around town. By publishing our framework as a community platform, we aim to enhance visibility of these small but well-built alternatives. What some systems lack in performance, they make up for in openness and transparency — and this should be rewarded.

Our framework provides a way to surface fine-grained information on openness and transparency. This information can empower regulatory bodies, institutions and the general public to make informed choices for or against deployment of Generative AI. The fully open and community-based nature of our method is one of the things distinguishing it from some other recent initiatives in this space. One is a transparency index released as a preprint [8] and widely publicised Ivy League press release that uses a wide range of indicators and methods, including interviews, but which does not open up individual data points for scrutiny or contestation.[15] Another is a model openness framework that converges on many of the same dimensions as our work [59], and which proposes to embed openness descriptions in model releases themselves.

3.1 Rethinking open source AI risks

Open source AI risks and opportunities are increasingly being studied in the generative AI landscape [16]. Corporate entities in this space have hand-waved at "AI Safety" as a reason to keep system specifications under wraps [43], but this appears mostly a thinly disguised attempt to obscure the clear and present harms such models already pose [17] while minimising the considerable legal and regulatory exposure that would come with disclosing details about training data [35, 49].

Discourse about appropriate levels of openness has been prone to equate 'open source' with two rhetorical extremes: (i) radical openness, which would mean literally sharing every single model component and training dataset, and (ii) homeopathic openness, which is openness diluted beyond recognition, for instance by sharing only model weights. Many corporate players, moving aside the first of these as unrealistic, propose that therefore the second, diluted sense is the only attainable. But this is not the case.

Between radical openness and homeopathic openness lies meaningful openness. Openness comes in degrees, and regulation should be designed to foster meaningful forms of openness. A composite notion of openness can also provide the building blocks that can cater to specific use cases — be it a radically open system for use in research and education, or more privacy-oriented ones with privacy-enhancing techniques in place.

There is a case to be made that open systems that disclose training and fine-tuning data are safer because of the possibility of public scrutiny and professional auditing. It is true openness of this kind, and not just open weight models, that can speed up innovation and afford inclusion and diversity [20, 61]. The EU AI Act and other future regulation should incentivise and regulate data disclosure for generative AI so that it is safer and more auditable, and so that everyone can benefit from a better understanding of how to build and how to do research on this technology.

3.2 Limitations

Generalising the framework. While the framework is in principle applicable to the full range of generative AI, we did not identify systems other than text and text-to-image generators that bill themselves as open source. However, as the field of generative AI is fast growing, this is bound to change in the coming years. We did identify a range of recent hybrid systems that combine text and text-to-image generation in one system, such as DeepFloyd by Stability AI, Open AI's GPT-4 or Google's Gemini model family. As these multimodal models rise in numbers and popularity, their assessment may require additional adjustments to the dimensions of openness laid out so far. Assessing other media types using the framework may require domain-specific dimensions and decisions, just as we found for text versus text-to-image.

Importantly, the overall framework is designed so that the bulk of the features are applicable to any generative AI system. Most generally, the broad areas of *availability*, *documentation*, and *access and licensing* should probably feature in any well-informed assessment of openness and accountability [11]. More specifically, many of the constituent features, from datasets to scientific documentation, are of general relevance to the question of how to define open in the context of AI and machine learning. This means that the framework is flexible enough to serve as a blueprint for the implementation of living guidelines [7] or for the formulation of templates such as those to be developed by the EU AI Office.

Training dataset assessment is complex. Our survey stays relatively superficial when it comes to assessing exactly *how* open the training data of a system is. This is due to three factors. Even the most open models we surveyed only describe what data was used instead of directly sharing it (sometimes due to licensing restrictions). But such descriptions of training data often only provide superficial detail of preprocessing steps and how exactly the data was fed into model training pipelines. This lack of documentation detail is further complicated by the sheer size of the training data that some systems are trained on, which are often combined and edited in ways that make it hard to retrace how data was used (sometimes in the service of privacy enhancing techniques). But probably the most serious complicating factor arises from the complexities of how some larger models are trained. Models can feature a simple training pipeline, but can also employ complex training procedures such as distributed or incremental training techniques (e.g. federated learning, batch learning or online learning), sometimes even using data acquired from user interactions. In sum, such challenges make it hard to assess training data openness well: we are only scraping the surface here.

Some data requires closed-door assessment. Full openness can be harmful, and in some instances assessment should take place behind closed doors. For instance, in the case of dealing with datasets that contain CSAM material, the release of data prior to assessment in the name of openness would pose a clear risk [4, 31]. But this should not mean that organisations that release production-ready systems are relieved from the requirements of full data disclosure by hand-waving to safety concerns. Especially when the systems are advertised as open source.

4 CONCLUSION

The EU AI Act is at risk of tying itself to a moving target: a licence-based definition of 'open source AI' that itself is evolving. A licence and its definition forms a single pressure point that will be targeted by corporate lobbies and big companies. The way to subvert this risk is to use a composite and gradient approach to openness. That makes it possible to cut through the knot of competing stakes and actors trying to influence the definition of open source AI, and to

[15] https://github.com/stanford-crfm/fmti

arrive at meaningful, evidence-based, multidimensional openness judgements. Such judgements can be used for individual models by potential users to make informed decisions for or against deployment of a particular architecture or model. They can also be used cumulatively to derive overall openness scores, and more reductively to classify systems into shades of openness or to define an openness cutoff for regulatory purposes.

Datasets represent the area that is most lagging behind in openness. Despite the challenges of openly sharing all data, we think full disclosure is where a key to meaningful openness lies. Work on AI safety and reproducibility has long pointed out the crucial importance of training data for understanding model performance, ensuring reproducibility, and assessing legal exposure [2, 25, 29, 46, 58].

Full openness is not always the solution: after all, even fully open systems can do harm and may be legally questionable. However, open is better than closed in most cases, and knowing *what* is open and *how* open it is can help everyone make better decisions. Openness is important for risk analysis (the public needs to know); for auditability (assessors need to know); for scientific reproducibility (scientists need to know); and for legal liability (end users need to know).

Our survey has offered a first glimpse at the detrimental effects of open-washing by companies looking to evade scientific, regulatory and legal scrutiny. And our framework hopefully offers the tools to counter it and to contribute to a healthy and transparent technology ecosystem in which the makers of models and systems can be held accountable, and users can make informed decisions.

ACKNOWLEDGMENTS

This research was funded by NWO Vidi grant 016.vidi.185.205 awarded to MD.

REFERENCES

[1] Nur Ahmed, Muntasir Wahed, and Neil C. Thompson. 2023. The growing influence of industry in AI research. *Science* 379, 6635 (March 2023), 884–886. https://doi.org/10.1126/science.ade2420

[2] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. https://doi.org/10.1145/3411764.3445736

[3] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics* 5, 5 (May 2023), 277–280. https://doi.org/10.1038/s42254-023-00581-4 Number: 5 Publisher: Nature Publishing Group.

[4] Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. 2023. Into the LAION's Den: Investigating Hate in Multimodal Datasets. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

[5] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158 ISSN: 2642-9381.

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. https://doi.org/10.48550/arXiv.2110.01963 arXiv:2110.01963 [cs].

[7] Claudi L. Bockting, Eva A. M. van Dis, Robert van Rooij, Willem Zuidema, and Johan Bollen. 2023. Living guidelines for generative AI — why scientists must oversee its use. *Nature* 622, 7984 (Oct. 2023), 693–696. https://doi.org/10.1038/d41586-023-03266-1 Bandiera_abtest: a Cg_type: Comment Number: 7984 Publisher: Nature Publishing Group Subject_term: Machine learning, Technology, Policy, Computer science.

[8] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The Foundation Model Transparency Index. https://doi.org/10.48550/arXiv.2310.12941 arXiv:2310.12941 [cs].

[9] Jean-Claude Burgelman, Corina Pascu, Katarzyna Szkuta, Rene Von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. 2019. Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century. *Frontiers in Big Data* 2 (2019). https://doi.org/10.3389/fdata.2019.00043

[10] Kenneth Ward Church and Valia Kordoni. 2022. Emerging Trends: SOTA-Chasing. *Natural Language Engineering* 28, 2 (March 2022), 249–269. https://doi.org/10.1017/S1351324922000043 Publisher: Cambridge University Press.

[11] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1186–1197. https://doi.org/10.1145/3593013.3594073

[12] Creative Commons, Eleuther AI, GitHub, Hugging Face, LAION, and Open Future. 2023. Supporting Open Source and Open Science in the EU AI Act. https://creativecommons.org/2023/07/26/supporting-open-source-and-open-science-in-the-eu-ai-act/

[13] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022. Behavioral Use Licensing for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 778–788. https://doi.org/10.1145/3531146.3533143

[14] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 427–439. https://doi.org/10.1145/3531146.3533108

[15] Kate Downing. 2024. Choose Your Own Adventure: The EU AI Act and Openish AI. https://katedowninglaw.com/2024/02/06/choose-your-own-adventure-the-eu-ai-act-and-openish-ai-2/

[16] Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder de Witt, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Botos Csaba, Fabro Steibel, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Marvin Imperial, Juan A. Nolazco-Flores, Lori Landay, Matthew Jackson, Paul Röttger, Philip H. S. Torr, Trevor Darrell, Yong Suk Lee, and Jakob Foerster. 2024. Near to Mid-term Risks and Opportunities of Open Source Generative AI. https://doi.org/10.48550/arXiv.2404.17047 arXiv:2404.17047 [cs] version: 1.

[17] Timnit Gebru, Emily M. Bender, Angelina McMillan-Major, and Margaret Mitchell. 2023. Statement from the listed authors of Stochastic Parrots on the "AI pause" letter. https://www.dair-institute.org/blog/letter-statement-March2023

[18] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Nov. 2021), 86–92. https://doi.org/10.1145/3458723

[19] Ellen P. Goodman and Julia Tréhu. 2022. *AI Audit-Washing and Accountability*. Technical Report. German Marshall Fund of the United States. https://www.jstor.org/stable/resrep44893

[20] 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. 2024. OLMo: Accelerating the Science of Language Models. https://doi.org/10.48550/arXiv.2402.00838 arXiv:2402.00838 [cs].

[21] Odd Erik Gundersen, Yolanda Gil, and David W. Aha. 2018. On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. *AI Magazine* 39, 3 (Sept. 2018), 56–68. https://doi.org/10.1609/aimag.v39i3.2816 Number: 3.

[22] Furkan Gursoy and Ioannis A. Kakadiaris. 2022. System Cards for AI-Based Decision-Making for Public Policy. http://arxiv.org/abs/2203.04754 arXiv:2203.04754 [cs].

[23] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush, and Hugo J. W. L. Aerts. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (Oct. 2020), E14–E16. https://doi.org/10.1038/s41586-020-2766-y Number: 7829 Publisher: Nature Publishing Group.

[24] Drew Hemment, Morgan Currie, Sj Bennett, Jake Elwes, Anna Ridler, Caroline Sinders, Matjaz Vidmar, Robin Hill, and Holly Warner. 2023. AI in the Public Eye: Investigating Public AI Literacy Through AI Art. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 931–942. https://doi.org/10.1145/3593013.3594052

[25] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[26] Daniel Jaffee. 2012. Weak Coffee: Certification and Co-Optation in the Fair Trade Movement. *Social Problems* 59, 1 (Feb. 2012), 94–116. https://doi.org/10.1525/sp.2012.59.1.94

[27] Paul Keller. 2023. A Frankenstein-like approach: open source in the AI act. https://openfuture.eu/blog/a-frankenstein-like-approach-open-source-in-the-ai-act

[28] Mehtab Khan and Alex Hanna. 2022. The Subjects and Stages of Al Dataset Development: A Framework for Dataset Accountability. 19, 2 (2022), 171–256.

[29] Florian Königstorfer and Stefan Thalmann. 2022. AI Documentation: A path to accountability. *Journal of Responsible Technology* 11 (Oct. 2022), 100043. https://doi.org/10.1016/j.jrt.2022.100043

[30] Clifford H. Lee and Elisabeth Soep. 2016. None But Ourselves Can Free Our Minds: Critical Computational Literacy as a Pedagogy of Resistance. *Equity & Excellence in Education* 49, 4 (Oct. 2016), 480–492. https://doi.org/10.1080/10665684.2016.1227157

[31] Hee-Eun Lee, Tatiana Ermakova, Vasilis Ververis, and Benjamin Fabian. 2020. Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation* 34 (Sept. 2020), 301022. https://doi.org/10.1016/j.fsidi.2020.301022

[32] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of CUI'23*. Eindhoven. https://opening-up-chatgpt.github.io/

[33] Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. 2024. From LLM to Conversational Agent: A Memory Enhanced Architecture with Fine-Tuning of Large Language Models. https://doi.org/10.48550/arXiv.2401.02777 arXiv:2401.02777 [cs].

[34] Laura Lucaj, Patrick Van Der Smagt, and Djalel Benbouzid. 2023. AI Regulation Is (not) All You Need. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1267–1279. https://doi.org/10.1145/3593013.3594079

[35] Nicola Lucchi. 2023. ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. *European Journal of Risk Regulation* (Aug. 2023), 1–23. https://doi.org/10.1017/err.2023.59

[36] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research* 24, 253 (2023), 1–15. http://jmlr.org/papers/v24/23-0069.html

[37] Jeanna Matthews. 2020. Patterns and Antipatterns, Principles, and Pitfalls: Accountability and Transparency in Artificial Intelligence. *AI Magazine* 41, 1 (2020), 82–89. https://doi.org/10.1609/aimag.v41i1.5204 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1609/aimag.v41i1.5204.

[38] Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2023. Data Statements: From Technical Concept to Community Practice. *ACM Journal on Responsible Computing* (2023). https://doi.org/10.1145/3594737 Just Accepted.

[39] Lisa Messeri and M. J. Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature* 627, 8002 (March 2024), 49–58. https://doi.org/10.1038/s41586-024-07146-0 Publisher: Nature Publishing Group.

[40] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[41] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 15991–16111. https://doi.org/10.18653/v1/2023.acl-long.891

[42] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2023. Accountability in artificial intelligence: what it is and how it works. *AI & SOCIETY* (Feb. 2023). https://doi.org/10.1007/s00146-023-01635-y

[43] OpenAI. 2023. GPT-4 API general availability and deprecation of older models in the Completions API. https://openai.com/blog/gpt-4-api-general-availability

[44] European Parliament. 2024. Artificial Intelligence Act: Provisional Agreement Resulting from Interinstitutional Negotiations. , 245 pages. http://web.archive.org/web/20240310112041/https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/AG/2024/02-13/1296003EN.pdf

[45] Bruce Perens. 1999. The Open Source Definition. In *Open Sources: Voices from the Open Source Revolution*, Chris DiBona, Sam Ockman, and Mark Stone (Eds.). O'Reilly, 171–188.

[46] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 33–44. https://doi.org/10.1145/3351095.3372873

[47] Dirk Riehle. 2023. The Future of the Open Source Definition. *Computer* 56, 12 (Dec. 2023), 95–99. https://doi.org/10.1109/MC.2023.3311648 Conference Name: Computer.

[48] Hannah Ruschemeier. 2023. AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal. *ERA Forum* 23, 3 (Feb. 2023), 361–376. https://doi.org/10.1007/s12027-022-00725-6

[49] Pamela Samuelson. 2023. Generative AI meets copyright. *Science* 381, 6654 (July 2023), 158–161. https://doi.org/10.1126/science.adi0656 Publisher: American Association for the Advancement of Science.

[50] Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M. Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. What Language Model to Train if You Have One Million GPU Hours? https://doi.org/10.48550/arXiv.2210.15424 arXiv:2210.15424 [cs].

[51] Timo Schick and Hinrich Schütze. 2021. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185

[52] Jan Smits and Tijn Borghuis. 2022. Generative AI and Intellectual Property Rights. In *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, Bart Custers and Eduard Fosch-Villaronga (Eds.). T.M.C. Asser Press, The Hague, 323–344. https://doi.org/10.1007/978-94-6265-523-2_17

[53] Irene Solaiman. 2023. The Gradient of Generative AI Release: Methods and Considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 111–122. https://doi.org/10.1145/3593013.3593981

[54] Arthur Spirling. 2023. Why open-source generative AI models are an ethical way forward for science. *Nature* 616, 7957 (April 2023), 413–413. https://doi.org/10.1038/d41586-023-01295-4 Bandiera_abtest: a Cg_type: World View Number: 7957 Publisher: Nature Publishing Group Subject_term: Ethics, Machine learning, Technology, Scientific community.

[55] Marilyn Strathern. 1997. 'Improving ratings': audit in the British University system. *European Review* 5, 3 (July 1997), 305–321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4 Publisher: Cambridge University Press.

[56] Alek Tarkowski. 2024. AI Act fails to set meaningful dataset transparency standards for open source AI. https://openfuture.eu/blog/ai-act-fails-to-set-meaningful-dataset-transparency-standards-for-open-source-ai

[57] Hugo Touvron, Louis Martin, and Kevin Stone. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. (2023). https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/

[58] Joel Walmsley. 2021. Artificial intelligence and the value of transparency. *AI & SOCIETY* 36, 2 (June 2021), 585–595. https://doi.org/10.1007/s00146-020-01066-z

[59] Matt White, Ibrahim Haddad, Cailean Osborne, Xiao-Yang, Liu, Ahmed Abdelmonsef, and Sachin Varghese. 2024. The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency and Usability in AI. https://doi.org/10.48550/arXiv.2403.13784 arXiv:2403.13784 [cs].

[60] David Gray Widder, Sarah West, and Meredith Whittaker. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. https://doi.org/10.2139/ssrn.4543807

[61] BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic

Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. https://doi.org/10.48550/arXiv.2211.05100 arXiv:2211.05100 [cs].

A APPENDIX: ELEMENTS OF A COMMUNITY-DRIVEN ASSESSMENT FRAMEWORK

We propose to assess generative AI systems by collecting evidence-based judgements on relevant dimensions of openness. For each dimension, we distinguish three levels: open (■), partial (■) or closed (■). Here we describe and define all 14 dimensions of openness. Currently, two variations of the framework exist that are tailored to the specific needs to assess text and text-to-image generators respectively.[16]

Method. We conduct a web search to identify text generators that use the term open source to market or advertise their system. For each system, we examine the information provided by the makers regarding information about training data, training pipelines, models, weights, documentation, and user access methods provided by the publisher or maintainer of the system. No third-party information was used.

Assessment procedure. Given all available information, we assess each system on each openness dimension. The dimensions are designed to provide a comprehensive overview of how much detail is provided regarding the training data, model training pipelines and fine-tuning regimes of the system (availability); how well aspects of the architecture, training and evaluation are documented (documentation), and how users can access system either as an end user or as a party interested to learn about the system itself (user access). Using the definitions for each dimension of openness in section 2, we set up a two-step, evidence-based assessment procedure with contributor and reviewer roles using an open data repository. For each dimension, anyone can submit judgements (open/partial/closed) alongside evidence that backs up the claim. Typically evidence is provided in form of a link to the official documentation of the system, a published preprint, or source code. Each contribution is then reviewed by a domain expert before it is published as part of the assessment outcome of the system under scrutiny.

A.1 Availability of training data and weights

The first part of the assessment procedure focuses on the model(s) that are used in currently popular text generators that employ large language models combined with instruction tuning techniques (LLM+RLHF). This includes all training data of all model training, instruction tuning, and/or fine-tuning steps as well as model weights of all components (typically the weights of the base model and the weights of the final tuned model).

A.1.1 Open code. In the classic sense of open source, we ask: is the source code of the model and training pipeline available? Can all source code for training data processing and model training be inspected?

■ System is closed source code.
■ Some source code is open.
■ System source code openly available and fully open available for inspection.

A.1.2 Base LLM data. Are all training datasets of the base model available for inspection?

■ Training data of base large language models (LLM) is not open for inspection.
■ Some of the training data of the large language models (LLM) is open for inspection.

[16] Supplementary data repository via Open Science Foundation: https://osf.io/f2b7n

■ The training data of all large language models (LLM) is fully open for inspection.

A.1.3 Base LLM weights. Are the language model weights (of the production-ready model) openly available?

■ LLM weights are not shared and model training procedure is not open for inspection.

■ LLM weights are not fully shared or model training procedure is not fully open for inspection.

■ LLM weights are shared and model training procedure is fully open for inspection.

A.1.4 Instruction tuning data. Inspect the instruction tuning component (and any additional fine-tuning steps) of the model: Are all datasets used in the instruct tuning component (e.g. reinforcement learning from human feedback) of the system available for inspection?

■ Training data of all instruction-tuning components is not open for inspection.

■ Some of the training data of all instruction-tuning components is open for inspection.

■ Training data of all instruction-tuning components is open for inspection.

A.1.5 Instruction tuning weights. Is the instruction-tuned (or final fine-tuned) model available for inspection (after all training steps have been completed)?

■ The instruction-tuned model weights are not open for inspection.

■ The instruction-tuned model weights are open for limited inspection.

■ The instruction-tuned model weights are fully open for inspection.

A.2 Documentation and transparency

The assessment category assessed the degree to which systems are documented in terms of professional code documentation (of model training, tuning steps), hardware requirements, and model performance and safety evaluation.

A.2.1 Code. This feature considers the level of documentation of the code. Distinct from the mere availability of code, here we ask whether the code base is documented in sufficient detail to allow replication, extension, or modification.

■ Code documentation not available.

■ Some components of the project feature code documentation.

■ All components of the project feature a comprehensive code documentation.

A.2.2 Architecture. Here we look at the documentation of the actual architecture of the system. This includes everything from hardware requirements, to information how the model was trained, tuned and evaluated (e.g. for performance, latency/speed, energy consumption and environmental impact).

■ System architecture and model training setup are not documented.

■ System architecture and model training setup is partially documented.

■ System architecture and model training setup is fully documented.

A.2.3 Preprint. Is an overview of the publication available in the form of a durable publication (including a DOI/ISBN)? Common formats are ArXiv preprints.

■ No archived preprint(s) available.

■ Archived preprint(s) that detail parts of the system are available.

■ Archived preprint(s) are available that cover all parts of the system.

A.2.4 Paper. In addition to a mere preprint, has the publication undergone peer-review in an academic publication venue?

■ No peer-reviewed paper(s) available.

■ Peer-reviewed paper(s) detail parts of the system including base models and tuning components.

■ Peer-reviewed paper(s) are available that cover all parts of the system including data, training, and tuning steps.

A.2.5 Modelcard. Model cards represent the field's standard for disclosing and documenting key facts about the architecture, training and evaluation of the model [14, 40].

■ Model card(s) not available.

■ Model card(s) that provide partial insight on model architecture, training, tuning, and evaluation are available.

■ Model card(s) are available that provide comprehensive insight on architecture, training, tuning, and evaluation.

A.2.6 Datasheet. Data sheets document key aspects of data collection and curation [18, 38]. They ensure that relevant information about training data is made available in systematic and relatively standardised ways.

■ Datasheet(s) are not available.

■ Datasheet(s) that provide partial insight on data collection and curation are available.

■ Datasheet(s) are available that provide comprehensive insight on data collection and curation are available.

A.3 Access and licensing

The third category covers access methods to the system *qua* system, covering features like the availability of software packages for local deployment, APIs, and licensing.

A.3.1 Package. Is an indexed software package available via an open software repository or similar durable web interface?

■ No indexed software package is available.

■ User-oriented code or web-interface is available but not as a versioned, indexed package (e.g. via GitHub).

■ A packaged release of fully open source software (e.g. a Python Package Index, Homebrew) is available.

A.3.2 API. Is the model accessible via an API? How is API access managed?

■ No API available.

■ Commercial or restricted-access user API is available.

■ Open API available that provides unrestricted access to the system (apart from security/CDN restrictions).

A.3.3 Licensing. Licensing relates to the licences that apply to systems or their components. Sometimes systems come with multiple licences, in which case coding is based on the most restrictive licence. Two types of licences are desirable: Open Source Initiate(OSI)-approved licences that allow for maximally unrestricted access/shareability and responsible AI licences that aim to regulate harmful uses of the system. For systems deemed 'minimal risk' (the lowest risk category) by risk-based regulation frameworks such as the EU AI Act, an OSI licence may be more applicable. For any other systems RAIL licences may be more applicable.

- System is not licenced clearly or does not use OSI or RAIL licences.
- System is only partially covered by an OSI or RAIL licence.
- System is fully covered by an OSI or RAIL licence.

B APPENDIX B: ASSESSMENT OF TEXT-TO-IMAGE MODELS

Unlike text generators, most image generators do not feature an instruction-tuning step as part of the model architecture so the two dimensions related to that (RL data and RL model weights) are replaced by two added features with direct relevance to image generators: *watermarking* and *prompt moderation.*

Watermarking. This dimension assessed whether the inclusion of absence of watermarking techniques is specified. We focus on techniques that make images identifiable and trackable as synthetic, often invisible to the human end user. We are not concerned with human-readable watermarks on the image output.

- No information on watermarking available.
- Limited information available.
- Watermarking techniques or the absence thereof are fully documented.

Prompt moderation. Another model output moderation feature are methods to restrict certain types of text prompts that the model can receive, usually using pattern matching techniques and vocabulary lists. To assess this system feature, we look for whether the system specifies whether or not such techniques are employed, and if so, how they work.

- No information available.
- Limited information.
- Sufficient information available.

Received 22 January 2024