Position: Near to Mid-term Risks and Opportunities of Open-Source Generative AI

Francisco Eiras [1] **Aleksandar Petrov** [1] **Bertie Vidgen** [2] **Christian Schroeder de Witt** [1] **Fabio Pizzati** [1] **Katherine Elkins** [3] **Supratik Mukhopadhyay** [4] **Adel Bibi** [1] **Botos Csaba** [1] **Fabro Steibel** [5] **Fazl Barez** [1] **Genevieve Smith** [6] **Gianluca Guadagni** [7] **Jon Chun** [3] **Jordi Cabot** [8 9] **Joseph Marvin Imperial** [10 11] **Juan A. Nolazco-Flores** [12] **Lori Landay** [13] **Matthew Jackson** [1] **Paul Röttger** [14] **Philip H.S. Torr** [1] **Trevor Darrell** [6] **Yong Suk Lee** [15] **Jakob Foerster** [1]

Abstract

In the next few years, applications of Generative AI are expected to revolutionize a number of different areas, ranging from science & medicine to education. The potential for these seismic changes has triggered a lively debate about potential risks and resulted in calls for tighter regulation, in particular from some of the major tech companies who are leading in AI development. While regulation is important, it is key that it does not put at risk the budding field of open-source Generative AI. We argue for the responsible open sourcing of generative AI models in the near and medium term. To set the stage, we first introduce an AI openness taxonomy system and apply it to 40 current large language models. We then outline differential benefits and risks of open versus closed source AI and present potential risk mitigation, ranging from best practices to calls for technical and scientific contributions. We hope that this report will add a much needed missing voice to the current public discourse on near to mid-term AI safety and other societal impact.

1. Introduction

Generative AI (Gen AI), defined as *"artificial intelligence that can generate novel content"* by conditioning its response on an input (Gozalo-Brizuela and Garrido-Merchan, 2023) (e.g., large language or foundation models), is anticipated to profoundly impact a diverse array of domains including science (AI4Science and Quantum, 2023), the economy (Brynjolfsson et al., 2023), education (Alahdab, 2023), the environment (Rillig et al., 2023), among many others. As a result, there has been significant socio-technical work undertaken to evaluate the broader risks and opportunities associated with these models, in a step towards a more nuanced and comprehensive understanding of their impacts (Bommasani et al., 2021), including recent regulatory developments (see Appendix B.1).

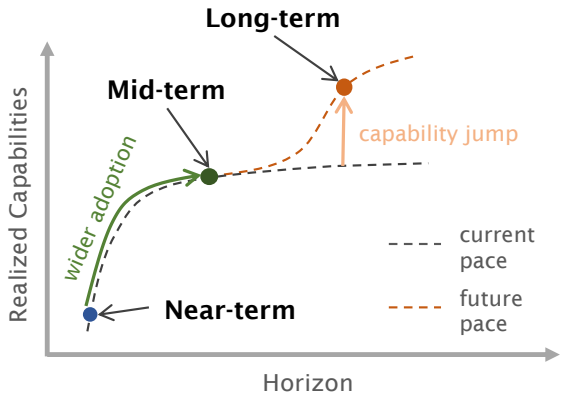

Figure 1: **Three Development Stages for Generative AI Models**: *near-term* is defined by early use and exploration of the technology in much of its current stage; *mid-term* is a result of the widespread adoption of the technology and further scaling at current pace; *long-term* is the result of technological advances that enable greater AI capabilities.

Parallel to these efforts is a debate on the *openness of Gen AI* models. The digital economy heavily relies on open-source software, exemplified by over 60% of global websites using open-source servers like Apache and Nginx (Lifshitz-Assaf and Nagle, 2021). This prevalence is underscored by a 2021 European Union report, which concluded that "overall, the [economic] benefits of open source greatly outweigh the costs associated with it" (Blind et al., 2021). Some developers of Gen AI models have chosen to openly release

[1]University of Oxford [2]MLCommons [3]Kenyon College [4]Center for Computation & Technology, Louisiana State University [5]Institute for Technology & Society (ITS), Rio [6]University of California, Berkeley [7]University of Virginia [8]Luxembourg Institute of Science and Technology [9]University of Luxembourg [10]University of Bath [11]National University Philippines [12]ITESM [13]Berklee College of Music [14]Bocconi University [15]University of Notre Dame. Correspondence to: FE <eiras@robots.ox.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

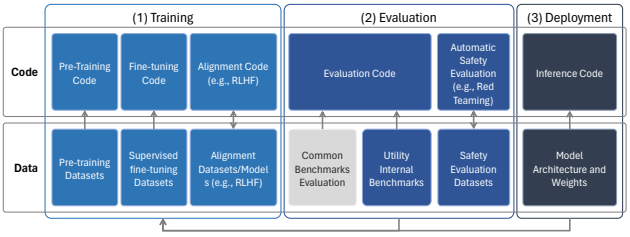

Figure 2: **Model Pipeline**: stages showing (1) training, (2) evaluation, and (3) deployment analyzed in the report. The component Common Benchmarks Evaluation (light gray) is included for completeness yet will not be analyzed in detail as these are standard and commonly available.

trained models (and sometimes data and code too), by leaning on this narrative and claiming that by doing so *"[these models] can benefit everyone"* and that *"it's safer [to release them]"* (Meta, 2023). However, while there has been a flurry of reports and surveys on the impacts of general open-source software in areas such as innovation or research within the last few decades (Paulson et al., 2004; Schryen and Kadura, 2009; Von Krogh and Spaeth, 2007), the discourse surrounding the openness of Gen AI models presents unique complexities due to the distinctive characteristics of this technology, including e.g., potential dual use and run-away technological progress.

This paper argues that the success of open source in traditional software could be replicated in Gen AI with well-defined and followed principles for responsible development and deployment. To this end, we begin by defining different stages of Gen AI development/deployment, followed by an empirical analysis of the openness of existing models through a taxonomy. With this framework, we then focus on evaluating the risks and opportunities presented by open and closed source Gen AI in the near to mid-term. Finally, we make a case for **the responsible open sourcing of generative AI models developed in the near to mid-term stages**, presenting recommendations to developers on how to achieve this safely and efficiently.

2. Preliminaries

To frame our analysis of the impacts of open sourcing generative AI models, we start by defining three-stages of AI development and outline the current pipelines involved in training, evaluating and deploying Large Language Models (LLMs). We focus on LLMs in these definitions and in §3.2 as this is the modality with the most prolific model development and open-sourcing at the moment, but note that it would be easy to extend our analysis to other modalities.

Stages of Development of Gen AI Models Our three-part framework (Figure 1) to describe the evolution of generative AI focuses on adoption rates and technological advancements instead of time elapsed (similar to Anthropic, 2023). The **near-term** stage is defined by the early use and exploration of existing technology, such as deep learning with transformer and diffusion model architectures, utilizing large datasets. This phase is characterized by experimentation, with increasing levels of development, investment and adoption. The **mid-term** is defined by the widespread adoption and scaling of existing technology, and the exploitation of its benefits. We conceptualize this as moving along a predictable 'capability curve', whereby more resources and usage will lead to greater benefits (and risks), but technological capabilities have not radically improved. Increasing use of multimodal models, agentic systems, and retrieval augmented generation are expected at this stage. The **long-term** is defined by a technological advance that will create dramatically greater AI capabilities, and therefore more risks and opportunities. This could manifest as a novel AI paradigm, a departure from traditional deep learning architectures, more efficient data utilization, among others, leading to more powerful AI models. In this paper, we focus primarily on analyzing the risks and opportunities of open-source Gen AI in the near to mid-term stages.

Training, Evaluating, and Deploying LLMs The components typically involved in the (1) training, (2) evaluation, and (3) deployment of models are shown in Figure 2, and they can be divided into two categories: *Code* and *Data*. We briefly describe each of the stages below, and provide a more in-depth component description in Appendix A.

Model training processes can be grouped into three distinct stages: *pre-training*, where a model is exposed to large-scale datasets composed of trillions of tokens of data, with the goal of developing fundamental skills and broad knowledge; *supervised fine-tuning* (SFT), which corrects for data quality issues in pre-training datasets using a smaller amount of high-quality data; and *alignment*, focusing on creating application-specific versions of the model by considering human preferences. Once trained, models are usually evaluated on openly available evaluation datasets (e.g., MMLU by Hendrycks et al., 2020) as well as curated benchmarks (e.g., HELM by Liang et al., 2022). Some models are also evaluated on utility-oriented proprietary datasets held internally by developers, potentially by holding out some of the SFT/alignment data from the training process (Touvron et al., 2023a). On top of utility-based benchmarking, developers sometimes create safety evaluation mechanisms to proactively stress-test the outputs of the model (e.g., red teaming via adversarial prompts). Finally, at the deployment stage, content can be generated by running the inference code with the associated model weights.

3. Openness Taxonomy of LLMs

Model developers decide whether to make each component of the training, evaluation and deployment pipeline (Figure 2) *private* or *public*, with varying levels of restrictions for the latter. For instance, the developers of LLaMA-2 have

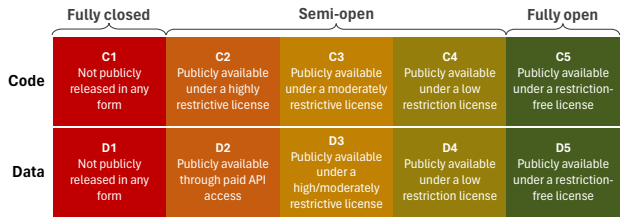

Figure 3: **Openness Scale**: categorization of the levels of openness of the code and data of each model component. See Table 1 (Appendix B) for the restrictions of each license.

publicly released the model architecture and weights, yet they have not shared the code or reward model for Reinforcement Learning from Human Feedback (RLHF) used in the Alignment components (Touvron et al., 2023a). To properly evaluate the openness of each component, we introduce a classification scale for Gen AI models in §3.1, which we then apply to 40 high impact LLMs in §3.2. This will help contextualizing the risks and opportunities discussed in §4, and the responsible open sourcing argument we make in §5. An up-to-date version of the taxonomy of LLMs is also available on this link.

3.1. Classifying Openness for Gen AI Code and Data

We introduce a framework for categorizing the openness of each component of Gen AI pipelines (e.g., Figure 2). At the highest level, a **fully closed** component is not publicly accessible in any form (Rae et al., 2022). In contrast, a **semi-open** component is publicly accessible but with certain limitations on access or use, or it is available in a restricted manner, such as through an Application Programming Interface (API) (Achiam et al., 2023). Finally, a **fully open** component is available to the public without any restrictions on its use (Xu et al., 2022). Further, the semi-open category comprises three subcategories, delineating varied openness levels (see Figure 3). Distinctions are made between Code (C1-C5) and Data (D1-D5) components, where C5/D5 represents unrestricted availability and C1/D1 denotes complete unavailability. For semi-open components, their classification relies on the license of the publicly available code/data.

To evaluate the licenses we introduce a point-based system where each license gets 1 point (for a total maximum of 5) for allowing each of the following: *can use a component for research purposes* (**Research**), *can use a component for any commercial purposes* (**Commercial Purposes**), *can modify a component as desired (with notice)* (**Modify as Desired**), *can copyright derivative* (**Copyright Derivative Work**), *publicly shared derivative work can use another license* (**Other license derivative work**). The total number of points is indicative of a license's restrictiveness. A **Highly restrictive** license scores 0-1 points, aligning with openness levels of code C2 and data D3, imposing significant limitations. A **Moderately restrictive** license, scoring 2-3 points (code C3 and data D3), allows more flexibility

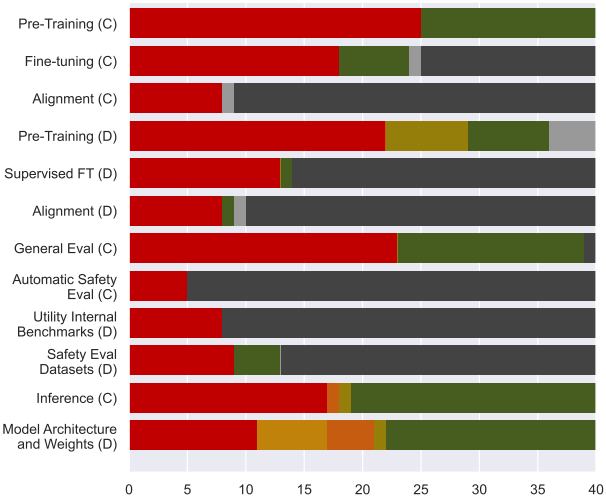

Figure 4: **Distribution of Openness Levels by Pipeline Component**: openness level distribution for each of the pipeline components of the 40 LLMs studied. Color legend: C1/D1, C2/D2, C3/D3, C4/D4, C5/D5, ? (unknown or not publicly available), N/A (not applicable). For conciseness, we use "FT" as a stand in for "Fine-Tuning".

but with some limitations. Licenses scoring 4 points are **Slightly restrictive** (code C4 and data D4), offering broader usage rights with minimal restrictions. Finally, a **Restriction free** license scores 5 points, indicating the highest level of openness (code C5 and data D5), permitting all forms of use, modification, and distribution without constraints.

In Table 1 (Appendix B) we provide a full table with the openness licenses and levels of all models studied in §3.2.

3.2. Openness Taxonomy of Current LLMs

We analyzed the pipeline components of 40 high-impact LLMs released from 2019 to 2023, chosen by optimizing three key impact metrics: *ChatBot Arena Elo Rating*, a crowdsourced benchmark score comparing models[1]; *Google Scholar Citations*, indicating each model's academic impact; and *HuggingFace Downloads Last Month*, reflecting the usage of models openly available on HuggingFace. While we included models that scored high on any of these metrics, we also decided to include other released models for the sake of diversity. Due to space constraints, the full model list is in Table 2 (Appendix B).

A full table with the taxonomy of each of the model components is presented in Table 3 (Appendix B). In Figure 4, we show the distribution of openness levels for each of the pipeline components analyzed. Figure 4 clearly shows a balance between open and closed source deployed components (inference code and weights); however, ***a notable skew exists towards closed source in training data (such as fine-tuning and alignment) and, importantly, in safety***

[1]Introduced in 05/2023; older models may be underrepresented.

evaluation code and data. To fully leverage open source benefits and mitigate risks discussed in the next sections, a significant shift toward responsible development and deployment of open-source generative AI is necessary.

4. Near to Mid-term Risks and Opportunities of Open Source Gen AI Models

We describe the risks and opportunities provided by open-source models in the near and mid-term (as defined in §2). Our focus is how open source catalyses, minimizes or creates risks and benefits compared to closed source – rather than Gen AI in general. Unless stated explicitly, we refer to all artifacts and components of AI when using the term "open source".

The Challenges of Assessing Risks and Benefits Gen AI systems can be evaluated through a variety of methods and frameworks, such as benchmarks like HELM and Big-Bench for task evaluation, Chatbot Arena for crowd-sourced model comparisons, and red teaming for exploratory evaluation (Guo et al., 2023; Liang et al., 2023; Srivastava et al., 2023). However, these approaches face limitations like limited ecological validity and data contamination (Li et al., 2023a; Sainz et al., 2023; Zhou et al., 2023b), and provide only a partial view of how models will perform in real-world settings. In response, some experts suggest socio-technical evaluations that are focused on real-world applications (Weidinger et al., 2023; Solaiman et al., 2023). This is supported by calls for comprehensive pre-release audits of models, datasets, and research artifacts (Derczynski et al., 2023; Mökander et al., 2023; Rastogi et al., 2023). However, even holistic approaches to evaluation face substantial challenges, such as the rapid and unpredictable evolution of AI capabilities, the difficulty of standardizing measurements due to the fast pace of change, and the research community's limited insight into AI's industrial applications. This invariably leads to partial and incomplete evidence. As such, while we use diverse evidence to examine and support our arguments, it is important to recognize the challenges in reaching definitive conclusions as a result of these limitations.

4.1. Quality and Transparency

Open Models are More Flexible and Customizable Having access to open-source models, datasets, and assets significantly aids developers in creating models that are high-performing and specifically tailored to their use-case. Developers have access to far more training approaches, models and datasets. This gives them a powerful starting point when creating a model for a specific application. It also particularly helps cater to less well-resourced languages, domains, and downstream tasks (Bommasani et al., 2023a), as well as enabling personalized models that cater to distinct groups and individuals (Kirk et al., 2023). This has created widespread positive sentiment towards open source, which can be seen in venture capital firm's significant investment in open-sourcing efforts (Bornstein and Radovanovic, 2023; Horowitz, 2023), and the growing adoption of open-source models by companies (Marshall, 2024).

Open Source Improves Public Trust Through More Transparency Nearly three out of five people (61%) are either ambivalent about or unwilling to trust AI, with Gillespie et al. (2023) reporting that cybersecurity risks, harmful use, and job loss are the "potential risks" that people are most concerned about. Closed source models pose challenges for evaluating, benchmarking, and testing them which impede accessibility, replicability, reliability, and trustworthiness (La Malfa et al., 2023). Transparency is a powerful way of improving trust, and addressing this critical problem. Transparency includes providing clear and explicit documentation, such as provenance artefacts like model cards, datasheets, and risk cards (Gebru et al., 2021; Derczynski et al., 2023; Longpre et al., 2023). They can be used to assess and review datasets and models, and are widely-used in the open source community. Open source is itself the best way of creating transparency. It enables widespread community oversight as models and datasets can be interrogated, scrutinised, and evaluated by anyone, without needing to seek approval from a central decision-maker. This empowers developers, researchers and other actors to engage with AI and contribute to discussions, encouraging a culture of contribution and accountability (Sanchez, 2021). At the same time, the highly technical nature of AI research creates substantial barriers to typical citizens. As such, more transparency may not alone drive greater trust – research outputs also need to be *accessible* and *understandable* by non-experts (Mittelstadt et al., 2019).

4.2. Research and Academic Impact

Open Source Advances Research Compared to the machine learning landscape a decade ago, the availability and continuous growth of open source in recent years has enabled the community to do more diverse and innovative research. This includes researchers exploring the inner workings of models through jailbreaking and quality checking for unsafe, harmful, and biased content (see §4.4) as well as probing for misuse of copyrighted data, which can potentially lead to class-action lawsuits (see §4.5). Likewise, the availability of code, data, and proper documentation of open models have allowed researchers to develop novel breakthroughs (e.g., DPO (Rafailov et al., 2023) as a more cost-efficient substitute for RLHF (Ouyang et al., 2022) for capturing human preference), which have been proven to boost open models to gain comparable performances against their closed model counterparts. Closed models, on the other hand, only grant limited access through API calls and restrict access to essential model generation outputs such as logits and token probabilities. Such limitations restrict researchers from forming deeper methodological insights and

limit reproducibility of their research (Rogers, 2023).

4.3. Innovation, Industry and Economic Impact

Open Source Empowers Developers and Fosters Innovation Closed source models accessed via an API make product developers reliant on an external provider for essential components of their product or system. This reliance can limit control and maintainability, especially as models can be updated or removed without warning by their owners. Further, with a closed model developers may not own their data or have full control over their data pipeline, which can make it more difficult to innovate on design, steer model performance, change aspects of their system, or understand their own workflows. In contrast, open models offer significant advantages. Developers can modify the model according to their needs, have complete understanding and transparency of the model, and control the data pipeline, which greatly enhances privacy and auditability (Culotta and Mattei, 2023). One important consideration is whether models are released with permissive licenses that suit commercial usecases (see commercial use in §2). This is increasingly common with more recent releases. Open-source models could be particularly beneficial in the emerging field of generative AI-powered agents (Chan et al., 2024), where outputs involve performing digital or physical actions (for early examples see Adept's blog post (AdeptTeam, 2022), and Amazon's press release (Amazon, 2023)). In this context, product developers are likely to value having more control over models, being able to deploy them on-device, and integrate them in larger, more complex systems.

Open Source Can be More Affordable AI models can enhance individual productivity by automating repetitive and time-consuming tasks, and augmenting workers when completing more complex and high-value tasks. This can help narrow the productivity gap between workers, improving minimum performance standards (Dell'Acqua et al., 2023). In principle, open-source AI models increase these benefits as they are available for free. However, substantial operational costs are still involved, such as the staff required to run the models, the time of leadership to organise and oversee their use, and the compute costs for inference (Palazzolo, 2023). Some enterprises might also apply additional protections for security and data to ensure compliance when using open-source models, adding further costs. Whether open source is cheaper overall than closed source depends on the maturity and capabilities of the organisation. Generally, larger corporations can bear the overheads involved in open source and overall make substantial savings.

Open Source Can be Easier to Access Open-source models are increasingly easy to use and access, with a range of vendors providing SDKs, APIs and downloadable files, such as Replicate, Together, and HuggingFace. Further, they typically require few approvals to start using models, in comparison with more onerous signup processes from closed source providers. One important area where open source lags behind closed source is in providing user interfaces aimed at non-technical audiences. While ChatGPT is easy to interact with and well-known amongst the general public, few open-source models have widely-used UIs.

Open Source Could Achieve Comparable Performance Today, the preference for closed source models stems from their user-friendly packaging, cost-effectiveness (with lower-income individuals predominantly opting for free versions, see Mollick, 2023), and potentially superior performance across various tasks (Open LLM Leaderboard). However, these dynamics are likely to shift in the near to mid-term. Firstly, with the growth of open source development, the performance gap between open and closed source models is expected to narrow significantly (UK-gov, 2023). Further, open source might be better in specific applications and contexts (see §4.3), driving adoption.

Open Models Could Help Tackle Global Economic Inequalities Knowledge workers in low-income nations, including workers in sectors like call centers and software development, face serious risk of job losses as AI models automate and semi-automate their work. Further, if AI models fail to adapt to local contexts or remain financially inaccessible, the expected economic benefits and new job opportunities may not arise, worsening economic inequalities (Georgieva, 2024). This is a concern as closed source models are often (1) unaffordable for companies in low-income countries and (2) badly-suited to their needs (see §4.5). Local needs are often not met because they lack adequate language support, culturally relevant content, and effective safety measures. This results in higher costs and lower performance, compounding the global inequalities that could be caused by generative AI (Petrov et al., 2023; Ahia et al., 2023). In contrast, open models could significantly change this dynamic. With requisite skill building and support for different communities, open models would enable communities to tailor models to their specific contexts and needs, promoting local innovation, safety, security, and reduced bias. This shift could help bridge the growing global inequality gap, paving the way for a more equitable and inclusive future in generative AI.

4.4. Safety

Generative AI models can create safety risks by increasing the severity and prevalence of harm experienced by individuals and society at large. This can take many forms, including physical, psychological, economic, representational and allocational harms (Shelby et al., 2023; Weidinger et al., 2023). The primary risks from current and near-term generative AI capabilities comprise two distinct pathways. The first is *malevolent use by bad actors*: individuals or organizations might exploit AI to create damaging content or

enable harmful interactions, such as personalized scams, targeted harassment, sexually explicit and suggestive content, and disinformation on a large scale (Vidgen et al., 2023; Ferrara, 2023). The second is *misguidance of vulnerable groups*: inaccurate or harmful advice from AI could lead vulnerable individuals, including those with mental health issues, to engage in self-harm (Mei et al., 2022; 2023; Röttger et al., 2023), radicalise towards supporting extremist groups, or believe in factually inaccurate claims about elections, health, and the environment (Zhou et al., 2023a). In the long-term, AI might develop capabilities that present novel existential threats, creating "catastrophic" consequences for society such as chemical warfare and environmental disaster (Hendrycks et al., 2023; Shevlane et al., 2023; Matteucci et al., 2023). However, these risks are not a substantial concern for existing models given their limited capabilities. Thus, in the near to mid-term, AI safety primarily means preventing models from generating toxic content, giving dangerous advice, and following malicious instructions.

Open Source Enables Technological Innovation for Safety Open source has significantly advanced safety research in the entire model development pipeline. Large open datasets for pre-training, like the Pile (Gao et al., 2020) (released for GPT-Neo, studied in the taxonomy §3.2), Laion (Schuhmann et al., 2022), and RedPajama (Computer, 2023), can be analysed for whether they contain toxic content (Prabhu and Birhane, 2020). Similarly, open research has shown model fine-tuning to be highly efficient in both improving model safety and removing model safeguards (e.g. Bianchi et al., 2023; Qi et al., 2023). Unlike closed APIs, open model analyses permit in-depth exploration of internal mechanisms and behaviors (e.g. Jain et al., 2023; Casper et al., 2024). This transparency enables reproducible and comprehensive evaluations, strengthening our understanding of generative AI safety for models with near and mid-term capabilities. Open source has also driven innovation in developing safeguards and controls for models, such as Meta's LlamaGuard (Inan et al., 2023) and HuggingFace's Safety Evaluation Leaderboard.

Open Models Can Also be Made to Generate Unsafe Content The flexibility of open-source models, as discussed in §4.1, has its drawbacks. Despite their initial alignment, these models can be fine-tuned to produce unsafe content, as exemplified by GPT4Chan and various "uncensored models" on the HuggingFace hub, designed to execute any instruction, irrespective of its safety implications. It is important to recognize, however, that closed models are not impervious to similar risks. Jailbreaks can induce unsafe behaviors in closed models as well (Zou et al., 2023), and recent studies have demonstrated that closed models can easily be fine-tuned to become just as unsafe as open ones (Qi et al., 2023). Nonetheless, ongoing advancements in generative AI safety technology (Dai et al., 2023), particularly through open models, hold the potential for mitigating these risks in the near to mid-term horizon.

Open Models Cannot be Rolled Back or Updated Once a model is made public, anyone can download it and use it indefinitely. In principle, benign users' access (e.g., researchers or rule-abiding corporations) can be regulated through license modifications. However, not all benign users will be aware of license changes and malicious actors will choose to not follow them. This creates a safety risk as any problems that have been identified post-deployment cannot be addressed. In comparison, closed model developers can cut off access to unsafe models if they are gatekept through an API. To reduce these risks, open source developers and the communities that host models (e.g., HuggingFace) must adhere to responsible release and access policies (e.g. Solaiman 2023; Solaiman et al. 2023; Anthropic 2023).

4.5. Societal and Environmental Impact

Open Source Models Can Reduce Energy Use AI model training incurs significant environmental costs from the energy consumption of compute resources. (Strubell et al., 2019; Wu et al., 2022). These impacts, measurable in CO_2 emissions, span the entire AI development process, including training and inference (Verdecchia et al., 2023; Kumar and Davenport, 2023). While accurately quantifying emissions for cloud providers is challenging due to variables like hardware utilization, team practices, geography, and time of day, industry-wide energy consumption can be reduced by sharing of resources that are energy-intensive to create, such as model weights (Saenko, 2023). In addition, open-sourcing can lead to transparent profiling of code to identify energy bottlenecks. This can then be addressed by the community, creating more energy-efficient training methods. For instance, some researchers have put forward small model development paradigms (Schwartz et al., 2019).

Open Models Can Help With Copyright Disputes One of the major legal issues surrounding generative AI is the use of copyrighted data for training without explicit permission (Firm and Butterick; Metz, 2024). This has mostly been identified because models regurgitate memorized data when prompted in specific ways (Karamolegkou et al., 2023; Carlini et al., 2022). The lack of transparency about what data are used in model training for both open and closed source (highlighted in §3.2) can lead to confusion, uncertainty, and misattribution. Open models that release, or describe, their training data can help address these issues of data privacy, memorization and the "fair use" of copyrighted materials. Crowd-sourced data curation also offers a way of minimizing use of proprietary datasets in the future, reducing the risk of copyright disputes (Hartmann et al., 2023).

Open Models Can Serve the Needs and Preferences of Diverse Communities To address global needs effectively, it is crucial that models do not only reflect the values of people who are liberal, culturally Western, and English

speaking (Aroyo et al., 2023; Lahoti et al., 2023). However, models are largely trained on data from the Internet, which is often biased to such people (Joshi et al., 2020). There is a pressing need to make pre-training datasets more diverse, inclusive and representative. In the short-term, models can be *steered* to meet the needs of different contexts, languages, and communities. Open source is a powerful way of achieving this as it enables under-resourced actors to build on top of each other's contributions. For instance, platforms like HuggingFace host a vast array of models, with many designed for specific cultural, geographic, or linguistic needs, e.g., Latxa (Bandarkar et al., 2023) and LeoLM (Plüster), covering diverse domains (e.g. Li et al., 2023b).

Open Source Helps Democratize AI Development Open source empowers developers to utilize resources from major organizations (e.g., companies, governments or research labs), facilitating the reuse of assets and leading to time, effort and money savings. This is crucial for AI development, which is characterized by high costs and complexity, from pre-training models that can cost millions (Knight, 2023) to the creation of expensive human-labeled datasets. This creates a clear societal benefit by enabling non-elites to access and use AI, which can include creating economic opportunities (see §4.3). It is important to acknowledge that, at a higher level, open-source models still contain key decisions, datasets and approaches that influence what is built on top of them. In this sense, they are currently undemocratic. They are informed by the values and market priorities of their largely for-profit driven developers.

5. Responsible Open Sourcing of Near to Mid-Term Generative AI

5.1. Addressing Common Concerns on Open Sourcing Generative AI

Despite the many benefits of open source, concerns surrounding the increased potential for malicious use, and uncertainty about its societal impact, have prompted calls for keeping generative AI closed source (Seger et al., 2023). There are real risks associated with open-source models. However, we believe these are sometimes exaggerated, possibly motivated by the economic interests of market leaders. Most concerns about open sourcing near to mid-term AI models are also pertinent to closed source models.

CLAIM #1: Closed Models Have Inherently Stronger Safeguards than Open-Source Models Several studies demonstrate that closed models typically demonstrate fewer safety and security risks, compared to open source (Röttger et al., 2023; Chen et al., 2024; Sun et al., 2024). However, closed models still demonstrate weaknesses, and are particularly vulnerable to jailbreaking techniques (Zou et al., 2023; Chao et al., 2023). Closed model safeguards are easily bypassed through simple manipulations like fine-tuning via accessible services (Qi et al., 2023), prompting the model to repeat a word (Nasr et al., 2023), applying a cypher (Yuan et al., 2023), or instructing the model in another language (Deng et al., 2023; Yong et al., 2023). Completely preventing models from exhibiting undesirable behaviors might not even be possible (Wolf et al., 2023; Petrov et al., 2024). Therefore, it is not clear that closed models are definitively "safer" than open-source models. We also anticipate that gaps will narrow over time as open safeguarding methods continue to improve.

CLAIM #2: Access to Closed Models Can Always be Restricted Closed models are often considered more secure because access can be restricted or removed if problems are identified. However, closed models can be compromised via hacking, leaks (Cox, 2023), reverse engineering (AsuharietYgvar, 2021) or duplication (Oliynyk et al., 2023). This perspective also assumes that models are only offered through an API. But some closed models are delivered on premise/device, particularly for sensitive deployments (e.g., government applications). In such cases, access may not be retractable. Finally, closed models can be leaked, e.g., Mistral's 70B parameter was leaked by one of their early customers (Franzen, 2024). Given these factors, developers do not always have the ability to unilaterally revoke access.

CLAIM #3: Closed Source Developers Can be Regulated to be Safer Regulatory pressure is primarily aimed at large companies building closed source models (e.g., see White House Executive Order). While it can create incentives for safe model development, regulation is not a panacea, and several closed source models have been released that are uncensored, poorly safeguarded (Verma, 2023) or deliberately misaligned (Burgess, 2023; Cuthbertson, 2023; Roscoe, 2023). It is also not clear that regulating closed source models is an effective way of stopping malicious actors (Lockie, 2015; Wootson, 2023), who are capable of creating and distributing their own closed source models via illicit sales channels (Sancho and Ciancaglini, 2023). Instead, it might create higher costs for legitimate users who are restricted in what models they can access (Wu et al., 2023).

CLAIM #4: All Safety and Security Problems Must be Addressed By the Model Provider It is becoming increasingly clear that, because of the numerous potential applications of generative models, all safety risks cannot be simply identified (and stopped) by the model provider. First, most model risks depend on the context and actors, and their real-world resources. For instance, real-world constraints significantly hinder activities like acquiring chemicals, equipment, or weapons, thus limiting open source's potential for misuse in such endeavors. Second, models may not have a causal impact on actors if they either (a) have other means of inflicting harm – such as searching on the web for malicious information – or (b) pay little attention to the responses of the model. Third, in practice, other stake-

holders help protect people from risk through established safeguarding practices, such as Internet Service Providers, cloud services, social media, and law enforcement. Given these factors, safety and security issues cannot be seen as solely the responsibility of the model provider.

5.2. Recommendations for Safe and Responsible Open Sourcing of Near to Mid-term Gen AI Models

To safely and responsibly open-source Gen AI models, we outline five important priorities for developers, starting with technical recommendations ahead of broader responsibility and socio-technical considerations.

Enhance Data Transparency and Provenance Responsible open sourcing is linked to greater transparency across the entire the model pipeline. As illustrated by Table 3 (Appendix B), a lack of data transparency is a problem even in relatively open LLMs. Making training and evaluation data publicly available enhances the community's capacity to scrutinize models' capabilities, risks, and limitations, thereby unlocking many of the advantages outlined in §4. It also holds the potential to develop models pre-trained for safety rather than aligned post-hoc. We believe this is an area where more research is needed which requires more parts of the pipeline to be open. Additionally, transparency in dataset composition, including metadata like copyright, is crucial. Maintaining comprehensive audit logs detailing chains of custody, transformations, data augmentation, and synthesis processes is increasingly vital.

Improve Open Evaluation and Benchmarking There has been much progress in open benchmarking of general LLM capabilities (e.g. LMSys, HELM, AlpacaEval), but there is an outstanding need for benchmarks that are specific to particular domains and impact areas, including model safety. This is poignant since, as highlighted in §3.2, most developers do not release their safety training and evaluation data. Generally, new models should be evaluated pre-release, so that their capabilities, risks, and limitations are made clear from day one. Evaluations should include assessments as related to the variety of risks outlined in §4.

Conduct Multilevel Security Audits Open source affords pre- and fine-tuning of models for any downstream tasks. For mission-critical tasks, particularly in areas like mental health, multi-level security audits and procedures must be meticulously designed, documented, implemented, and publicly reported. This should encompass both manual and automated testing, ranging from adversarial jailbreak prompts to expert-led red-teaming for common and edge case exploits, where financially viable. Additionally, incorporating static and dynamic analysis toolchains into developers' IDEs is essential to detect vulnerabilities early in the development process. Establishing and promoting safe design patterns for Gen AI development within the community is also crucial. Once ready for deployment, it is important that developers engage with the wider safety research community to allow for further third-party testing in controlled sandboxes closer to the released model environment.

Compare with Closed Source Models Open-source models offer advantages like enhanced privacy, customization, transparency, efficiency, and cost-effectiveness. In contrast, commercial closed-source models can stand out in performance, usability, and liability protections. Therefore, comparing the models with their closest commercial closed source alternative is important to quantify, clarify, and understand the trade-offs involved in open sourcing decisions.

Conduct Studies of Broader Societal Impact As highlighted in §4.5, properly developed open models can reduce Gen AI energy consumption, aid in resolving copyright disputes, cater to diverse communities, and help democratize AI development. To realize these benefits, it's crucial to undertake comprehensive broader societal impact studies. These should include evaluating corporate practices in model design and management, initiatives for enhancing data diversity and representation, and transparency reports on the environmental impact of the models.

6. Conclusion

The recommendations in §5.2 are a result of combining the openness trends of currently available models in §3.2 with the analysis of §4 on the potential risks and opportunities of open sourcing near to mid-term models. Following this discussion, we advocate for the **responsible open sourcing of near to mid-term Gen AI models**.

Note that our position is a balanced one. We advocate that developers should be allowed and encouraged to responsibly open-source Gen AI models developed in the near to mid-term stages, in as much as it makes economic sense for them to do so. Building Gen AI models is an expensive process, and we are sensitive to the argument that for-profit companies should be able to reap some of the financial benefits of their investments in building the technology. Any other position on this matter (e.g., forcing companies to open source their models/pipelines) would seriously risk investment and progress in this area.

However, often for-profit entities will claim open source Gen AI is fundamentally unsafe, and will publicly use this to argue against the open sourcing of these models altogether. This discourages other developers from open sourcing, and we believe this is one of the main factors that contributes to the current skew in the landscape presented in the taxonomy of §3.2 (Figure 4). We reject this argument, and argue in §4 and §5 that (1) there are many benefits that can only be achieved through open sourcing, and (2) the risks are often exaggerated by these for-profit entities. By making these impacts explicit and laying out recommendations for the responsible open sourcing of these models, our aim is to

encourage developers to improve the notable skew in Figure 4. This does not mean all models will be open-sourced, only that there would be an improved balance. We note that this should always be voluntary rather than imposed, to avoid disrupting the investment in the area.

Our work underscores the importance of mitigating risks and addresses prevalent concerns, thereby paving the way for realizing the vast potential benefits open-source generative AI offers.

7. Related Work

The debate around open sourcing Gen AI differs from the well-studied impacts of open-source software on society (Jaisingh et al., 2008) due to the unique characteristics of the technology. As such, we report related works on two axes: (1) examining the broader impact of Gen AI, and (2) on the debate around open sourcing these models.

The Impact of Gen AI There are many works that focus on the risks and benefits of the technology as it exists today, particularly with respect to areas such as science & medicine (AI4Science and Quantum, 2023; Fecher et al., 2023), education (Alahdab, 2023; Cooper, 2023; Malik et al., 2023), the environment (Rillig et al., 2023), among others. Other research evaluates the potential impacts of a capability shift Seger et al. (2023), emphasizing the critical importance of transparency in analyzing AI failures (Kapoor and Narayanan, 2023a;b).

On Open Sourcing Gen AI Models A main line of discussion centers on the definition of open sourcing Gen AI, highlighting the role of disclosing the training pipeline, weights, and data in achieving the benefits of open source (Bommasani et al., 2023b;a; Liesenfeld et al., 2023; Seger et al., 2023; Shrestha et al., 2023). Notably, AI systems typically encompass more than just code, necessitating custom release pipelines (Liu et al., 2023). Others (LAION.ai, 2023; Hacker et al., 2023; Tumadóttir, 2023) highlight the need to differentiate open-source systems from a regulatory standpoint, to avoid compliance costs unsustainable for open source contributors (Parliament, 2023). Many highlight the risks of centralization in absence of open source (Seger et al., 2023; Horowitz, 2023). On the other hand, open models may exacerbate the risks of misuse (Bommasani et al., 2021; Alaga and Schuett, 2023) unless proper measures are instituted for responsibly open-sourcing them. Interestingly, it has also been shown that open Gen AI tends to be less trustworthy than closed ones (Sun et al., 2024). A relevant paper (Seger et al., 2023) analyzes the risks and benefits of open models, and shapes recommendations for the near future. In our work, we provide a holistic viewpoint centered on near to mid-term models, including a taxonomy of the current landscape and discussion of future impacts.

Impact Statement

This work presents an attempt at a comprehensive evaluation of the risks and benefits associated with open-sourcing generative AI models as well as a list of prescriptions for responsible open-sourcing. The speculative nature of our work comes naturally with a broad impact potential. From a regulatory viewpoint, this paper could influence policy makers in the decision-making process concerning lawmaking oriented to open-source generative AI. Also, the impact on companies and open-source communities' release processes is potentially significant, considering the recent extremely high interest in developing and releasing open-source models. We stress that although our analysis is thorough, our risk assessment has fundamental assumptions that must be respected, and re-evaluated in case of disruptive unpredictable changes violating our hypotheses.

Disclaimer

This paper represents the collaborative effort of a diverse group of researchers, each bringing their own unique perspectives to the table. We note that not every viewpoint expressed within this work is necessarily unanimously agreed upon by all authors.

Acknowledgments

The authors would like to thank Meta for their generous support, including travel grants and logistical assistance, which enabled this collaboration, as well as for the organization of the first Open Innovation AI Research Community workshop where this work was initiated. Meta had no editorial input in this paper, and the views expressed herein do not reflect those of the company.

FE is supported by EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems [EP/S024050/1] and Five AI Limited. AP is funded by EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems [EP/S024050/1]. FP is funded by KAUST (Grant DFR07910). JMI is funded by National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent AI [EP/S023437/1] of the University of Bath. PR is supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA). PHST is supported by UKRI grant: Turing AI Fellowship EP/W002981/1, and by the Royal Academy of Engineering under the Research Chair and Senior Research Fellowships scheme. JF is partially funded by the UKI grant EP/Y028481/1 (originally selected for funding by the ERC).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo

Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

AdeptTeam. 2022. ACT-1: Transformer for actions.

Agencia de Gobierno. Mesa de diálogo "Inteligencia Artificial: oportunidades y desafíos de una estrategia nacional". *Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento*.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Microsoft Research AI4Science and Microsoft Azure Quantum. 2023. The impact of large language models on scientific discovery: a preliminary study using GPT-4. *arXiv preprint arXiv:2311.07361*.

Jide Alaga and Jonas Schuett. 2023. Coordinated pausing: An evaluation-based coordination scheme for frontier AI developers. *arXiv preprint arXiv:2310.00374*.

Fares Alahdab. 2023. Potential impact of large language models on academic writing. *BMJ Evidence-Based Medicine*.

Amazon. 2023. AWS expands Amazon Bedrock with additional foundation models, new model provider, and advanced capability to help customers build generative AI applications.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthropic. 2023. Anthropic's responsible scaling policy.

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. DICES Dataset: Diversity in conversational AI evaluation for safety. *arXiv preprint arXiv:2306.11247*.

Asia Society. 2024. China's Emerging Approach to Regulating General-Purpose Artificial Intelligence: Balancing Innovation and Control | Asia Society.

AsuharietYgvar. 2021. AppleNeuralHash2ONNX: Reverse-engineered Apple NeuralHash, in ONNX and Python.

Australian Government. 2024a. Australian Framework for Generative Artificial Intelligence (AI) in Schools.

Australian Government. 2024b. Interim guidance on government use of public generative AI tools.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2023. Safety-Tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*.

Knut Blind, Mirko Böhm, Paula Grzegorzewska, Andrew Katz, Sachiko Muto, Sivan Pätsch, and Torben Schubert. 2021. The impact of Open Source Software and Hardware on technological independence, competitiveness and innovation in the EU economy. *European Commission, Brussels*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the Opportunities and Risks of Foundation Modelss. *arXiv preprint arXiv:2108.07258*.

Rishi Bommasani, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake, Daniel E. Ho, Arvind Narayanan, and Percy Liang. 2023a. Considerations for governing open foundation models.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023b. Introducing the foundation model transparency index. *arXiv preprint arXiv:2310.12941*.

Matt Bornstein and Rajko Radovanovic. 2023. Supporting the open source AI community. *Andreessen Horowitz*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. 2023. Generative AI at work. Technical report, National Bureau of Economic Research.

Matt Burgess. 2023. Criminals have created their own ChatGPT clones. *Wired*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models.

In *The Eleventh International Conference on Learning Representations*.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. 2024. Black-box access is insufficient for rigorous AI audits. *arXiv preprint arXiv:2401.14446*.

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. 2024. Visibility into AI agents.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, and Shafiq Joty. 2024. Chatgpt's one-year anniversary: Are open-source large language models catching up?

Together Computer. 2023. RedPajama: an Open Dataset for Training Large Language Models.

Grant Cooper. 2023. Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*.

Norwegian Consumer Council. 2023. Ghost in the machine: Addressing the consumer harms of generative ai. *Norwegian Consumer Council, June*.

Joseph Cox. 2023. Facebook's powerful large language model leaks online. *Vice*.

Aron Culotta and Nicholas Mattei. 2023. Use open source for safer generative AI experiments. *MIT Sloan Management Review*.

Anthony Cuthbertson. 2023. Elon Musk's new AI bot will help you make cocaine which proves it's 'based' and 'rebellious'. *The Independent*.

Cyberspace Administration of China. (translated) interim measures for the management of generative artificial intelligence services office of the central cybersecurity and information technology commission.

Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. 2023. Exploring large language models for multi-modal out-of-distribution detection. *arXiv preprint arXiv:2310.08027*.

SDAIA: Saudi Data and AI Authority. 2023. AI ethics principles.

Fabrizio Dell'Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. 2023. Assessing language model deployment with risk cards. *arXiv preprint arXiv:2303.18190*.

Digital Government Authority. The Digital Government Authority issues free and open-source government software licenses to 6 government agencies.

European Parliament. 2021. Artificial Intelligence Act.

Benedikt Fecher, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. Friend or foe? Exploring the implications of large language models on the science system. *arXiv preprint arXiv:2306.09928*.

Emilio Ferrara. 2023. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *arXiv preprint arXiv:2310.00737*.

Joseph Saveri Law Firm and Matthew Butterick. LLM litigation.

Carl Franzen. 2024. Mistral CEO confirms "leak"' of new open source AI model nearing GPT-4 performance. *VentureBeat*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, et al. 2023. A framework for few-shot language model evaluation.

Saudi Gazette. 2024. SDAIA launches ALLAM AI application for Arabic chat.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Kristalina Georgieva. 2024. AI will transform the global economy. Let's make sure it benefits humanity.

Nicole Gillespie, Steven Lockey, Caitlin Curtis, Javad Pool, and Ali Akbari. 2023. Trust in artificial intelligence: A global study.

Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *arXiv preprint arXiv:2301.04655*.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey.

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. SoK: Memorization in general-purpose Large Language Models. *arXiv preprint arXiv:2310.18362*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*.

Andreseen Horowitz. 2023. House of Lords Communications and Digital Select Committee inquiry: Large language models.

The White House. 2023. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama Guard: LLM-based input-output safeguard for Human-AI conversations. *arXiv preprint arXiv:2312.06674*.

Infocomm. First of its kind Generative AI Evaluation Sandbox for Trusted AI by AI Verify Foundation and IMDA. *Infocomm Media Development Authority*.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*.

Jeevan Jaisingh, Eric WK See-To, and Kar Yan Tam. 2008. The impact of open source software on the strategic choices of firms developing proprietary software. *Journal of Management Information Systems*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*.

Oyvind Kaldestad. 2023. New report: Generative AI threatens. *Forbrukerrådet*.

Rahul Kapoor, Shokoh H Yaghoubi, and Theresa T Kalathil. 2024. Ai regulation in india: Current state and future perspectives.

Sayash Kapoor and Arvind Narayanan. 2023a. Licensing is neither feasible nor effective for addressing AI risks. *AI Snake Oil*.

Sayash Kapoor and Arvind Narayanan. 2023b. Three ideas for regulating generative AI. *AI Snake Oil*.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright Violations and Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback.

Will Knight. 2023. OpenAI's CEO says the age of giant AI models is already over. *Wired*.

Ajay Kumar and Tom Davenport. 2023. How to make generative ai greener. *Harvard Business Review*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.

Emanuele La Malfa, Aleksandar Petrov, Simon Frieder, Christoph Weinhuber, Ryan Burnell, Raza Nazar, Anthony G. Cohn, Nigel Shadbolt, and Michael Wooldridge. 2023. Language Models as a Service: Overview of a new paradigm and its challenges. *arXiv preprint arXiv:2309.16573*.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, et al. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *arXiv preprint arXiv:2310.16523*.

LAION.ai. 2023. A call to protect open source AI in europe. Accessed: 2024-01-29.

Jiatong Li, Rui Li, and Qi Liu. 2023a. Beyond static datasets: A deep interaction approach to LLM evaluation. *arXiv preprint arXiv:2309.04369*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. ChatDoctor: A medical chat model fine-tuned on a large language model meta-ai (LLaMA) using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational user interfaces*.

Hila Lifshitz-Assaf and Frank Nagle. 2021. The digital economy runs on open source. here's how to protect it. *Harvard Business Review*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2023. LLM360: Towards fully transparent open-source LLMs. *arXiv preprint arXiv:2312.06550*.

Alex Lockie. 2015. The wealthiest mafia in the world is undergoing a schism and it could get ugly. *Business Insider*.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. 2023. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. *arXiv preprint arXiv:2310.16787*.

Tegwen Malik, Laurie Hughes, Yogesh K Dwivedi, and Sandra Dettmer. 2023. Exploring the transformative impact of generative AI on higher education. In *Conference on e-Business, e-Services and e-Society*.

Matt Marshall. 2024. How enterprises are using Open Source LLMs: 16 examples. *VentureBeat*.

Kayla Matteucci, Shahar Avin, Fazl Barez, and Sean O hEigeartaigh. 2023. AI systems of concern. *arXiv preprint arXiv:2310.05876*, abs/2310.05876.

Alex Mei, Anisha Kabir, Sharon Levy, Melanie Subbiah, Emily Allaway, John Judge, Desmond Patton, Bruce Bimber, Kathleen McKeown, and William Yang Wang. 2022. Mitigating covertly unsafe text within natural language systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Alex Mei, Sharon Levy, and William Wang. 2023. ASSERT: Automated safety scenario red teaming for evaluating the robustness of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Meta. 2023. Meta and Microsoft Introduce the Next Generation of Llama.

Cade Metz. 2024. Openai says new york times lawsuit against it is "without merit".

MinCiencia. Artículo: Ministerio De Ciencia Abre Consulta Ciudadana Para Actualizar Política Nacional De Inteligencia Artificial.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM.

Ethan Mollick. 2023. An Opinionated Guide to Which AI to Use.

Monetary Authority of Singapore. MAS Partners Industry to Develop Generative AI Risk Framework for the Financial Sector.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

OECD. OECD's live repository of AI strategies & policies.

Courts of New Zealand. Guidelines for use of generative artificial intelligence in Courts and Tribunals — Courts of New Zealand.

Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. 2023. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Comput. Surv.*

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Stephanie Palazzolo. 2023. Meta's free ai isn't cheap to use, companies say.

EU Parliament. 2023. EU AI Act. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Accessed: 2024-01-29.

James W Paulson, Giancarlo Succi, and Armin Eberlein. 2004. An empirical study of open-source and closed-source software products. *IEEE Transactions on Software Engineering*, 30(4):246–256.

Aleksandar Petrov, Emanuele La Malfa, Philip HS Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Neural Information Processing Systems (NeurIPS)*.

Aleksandar Petrov, Philip HS Torr, and Adel Bibi. 2024. Prompting a pretrained transformer can be a universal approximator. *arXiv preprint arXiv:2402.14753*.

Björn Plüster. Laion leolm: Linguistically enhanced open language model.

Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision?

PricewaterhouseCoopers. 2024. Overview of 'The Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence'.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*.

Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 913–926, New York, NY, USA. Association for Computing Machinery.

Reuters. 2023. Abu Dhabi makes its Falcon 40B AI model open source. https://www.reuters.com/technology/abu-dhabi-makes-its-falcon-40b-ai-model-open-source-2023-05-25/.

Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*.

Anna Rogers. 2023. Closed AI Models Make Bad Baselines. Accessed on January 31, 2024.

Jules Roscoe. 2023. Elon Musk's Grok AI is pushing misinformation and legitimizing conspiracies. *Vice*.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2023. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. *arXiv preprint arXiv:2308.01263*.

Kate Saenko. 2023. A computer scientist breaks down generative AI's hefty carbon footprint. *Scientific American. https://www. scientificamerican. com/article/a-computer-scientist-breaks-down-generative-ais-hefty-carbon-footprint*.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark.

C Sanchez. 2021. Civil society can help ensure ai benefits us all. here's how. In *World Economic Forum*.

David Sancho and Vincenzo Ciancaglini. 2023. Hype vs. reality: AI in the cybercriminal underground.

Guido Schryen and Rouven Kadura. 2009. Open source vs. closed source software: towards measuring security. In *Proceedings of the 2009 ACM Symposium on Applied Computing*, pages 2016–2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green ai.

Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, et al. 2023. Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*.

Matt Sheehan. 2023. China's AI Regulations and How They Get Made.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *arXiv preprint arXiv:2210.05791*.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Yash Raj Shrestha, Georg von Krogh, and Stefan Feuerriegel. 2023. Building open-source AI. *Nature Computational Science*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Irene Solaiman. 2023. The Gradient of Generative AI Release: Methods and Considerations.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Anna Gamvros Steven Chong, Edward Yau (HK). 2023. China finalises its Generative AI Regulation.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

The UK Government. 2023. The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023.

TII. 2023. Falcon. `https://falconllm.tii.ae/`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

China Law Translate. 2023. Interim Measures for the Management of Generative Artificial Intelligence Services.

A. Tumadóttir. 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/`. Accessed: 2024-01-29.

UAE. 2023. UAE Strategy for Artificial Intelligence. `https://u.ae/en/about-the-uae/strategies-initiatives-and-awards/strategies-plans-and-visions/government-services-and-digital-transformation/uae-strategy-for-artificial-intelligence`.

UK-gov. 2023. Safety and security risks of generative artificial intelligence to 2025.

Roberto Verdecchia, June Sallou, and Luís Cruz. 2023. A systematic review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1507.

Pranshu Verma. 2023. They thought loved ones were calling for help. It was an AI scam.

Bertie Vidgen, Hannah Rose Kirk, Rebecca Qian, Nino Scherrer, Anand Kannappan, Scott A Hale, and Paul Röttger. 2023. SimpleSafetyTests: a Test Suite for Identifying Critical Safety Risks in Large Language Models. *arXiv preprint arXiv:2311.08370*.

Georg Von Krogh and Sebastian Spaeth. 2007. The open source software phenomenon: Characteristics that promote research. *The Journal of Strategic Information Systems*, 16(3):236–253.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv preprint arXiv:2310.11986*.

Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Cleve R. Wootson. 2023. It's time to stop laughing at Nigerian scammers – because they're stealing billions of dollars. *The Washington Post*.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable AI: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813.

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. LLMDet: A third party large language models generated text detection tool. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2113–2133.

Frank F. Xu, Uri Alon, Graham Neubig, and Vincent J. Hellendoorn. 2022. A systematic evaluation of large language models of code. *arXiv preprint arXiv:2202.13169*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023. GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher. *arXiv preprint arXiv:2308.06463*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023a. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023b. Don't make your LLM an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A. Further details on training, evaluation and deployment

Model training (1) processes can be grouped into three distinct stages:

1. *Pre-training*, where a model is exposed to large-scale datasets composed of trillions of tokens of data, typically scraped from the internet and usually uncurated. The goal is for the model to see a diversity of data, and through that process develop fundamental skills (e.g., grammar, vocabulary, text structure) and broad knowledge (Gao et al., 2020; Radford et al., 2019). An example of a commonly used open source dataset for pre-training LLMs such as LLaMA or GPT-J is The Pile which combines 22 smaller datasets into a diverse 825Gb text dataset (Gao et al., 2020; Touvron et al., 2023a; Wang and Komatsuzaki, 2021).

2. *Supervised fine-tuning (SFT)*, which is intended to correct for data quality issues in pre-training datasets. Usually, a much smaller amount of high quality data is used to improve model performance. Several works observe that at this stage the quality of the data used is essential to the downstream performance of the models (Zhou et al., 2024; Ouyang et al., 2022; Touvron et al., 2023b; Team et al., 2023), with the authors of LLaMA-2 pointing out that *"by setting aside millions of examples from third-party datasets and using fewer but higher-quality examples from our own vendor-based annotation efforts, [their] results notably improved."* (Touvron et al., 2023b).

3. *Alignment*, which is used to create an application-specific version of the foundation model (e.g., a chatbot or translation model). Reinforcement Learning with Human Feedback (RLHF) or Direct Preference Optimisation (DPO) (Ouyang et al., 2022; Touvron et al., 2023b) is used to create a model that follows instructions and is better-aligned with human preferences. With RLHF, a dataset of human preferences over model outputs is used to train a Reward model, which in turn is used with a reinforcement learning algorithm (e.g., PPO; Schulman et al., 2017) to align the LLM. RLHF is not used in models released prior to 2022 (Brown et al., 2020; Xue et al., 2020; Smith et al., 2022), and it is unclear whether the RLHF is used in models such as PaLM-2 (Anil et al., 2023).

Once trained, models are usually evaluated (2) on openly available evaluation datasets such as MMLU or NaturalQuestions (Hendrycks et al., 2020; Kwiatkowski et al., 2019) as well as curated benchmarks such as HELM, BigBench EleutherAI's Evaluation Harness (Liang et al., 2022; Srivastava et al., 2022; Gao et al., 2023). Some models are also evaluated on proprietary datasets held internally by developers, potentially by holding out some of the SFT/RLHF data from the training process (Touvron et al., 2023b). However, there is little publicly available information on how this is implemented, and few details are shared about the composition of such datasets. On top of utility-based benchmarking, developers sometimes create safety evaluation mechanisms to proactively stress-test the outputs of the model. These include human-annotated safety evaluation datasets (e.g., through creating adversarial prompts), as well as automatic safety evaluation algorithms (Touvron et al., 2023b; Yuan et al., 2023). They are typically the result of applying techniques such as red teaming. Finally, at the deployment stage (3), content can be generated by running the inference code with the associated model weights.

B. Full Taxonomy Tables

Important disclaimer: Table 3 focuses on component openness in model pipelines, not reproducibility. GLM-130B and Falcon provide detailed training procedures, unlike GPT-4, yet those are all classified as C1 due to unreleased pre-training code. A full reproducibility assessment falls beyond this report's scope.

License	Research	Commercial Purposes	Modify as Desired	Copyright derivative work	Other license for derivative	Final score	Code Openness	Data Openness
MIT/Mod. MIT	Y	Y	Y	Y	Y	5 (Restriction free)	C5	D5
Apache 2.0	Y	Y	Y	Y	Y	5 (Restriction free)	C5	D5
Common Crawl (ComCrawl)	Y	Y	Y	Y	Y	5 (Restriction free)	C5	D5
BSD-3	Y	Y	Y	Y	Y	5 (Restriction free)	C5	D5
RAIL	Y	Y	Y	Y	N	4 (Slightly restrictive)	C4	D4
LLaMA-2	Y	Y[2]	N	Y	N	3 (Moderately restrictive)	C3	D3
ODC-By	Y	Y	Y	Y	N	4 (Slightly restrictive)	N/A	D4
CodeT5 Data	Y	Y	Y	Y	N	4 (Slightly restrictive)	N/A	D4
RedPajama Data (Full)	Y	Y	Y	Y	N	4 (Slightly restrictive)	N/A	D4
OPT Data	Y	N	N	N	N	1 (Highly restrictive)	N/A	D3
GLM-130B Data	Y	N	N	N	N	1 (Highly restrictive)	N/A	D3
Falcon-180B Data	Y	Y	Y	Y	Y	5 (Restriction free)	N/A	D5

Table 1: **License Openness Taxonomy**: categorization of commonly used licenses in a variety of relevant open source criteria, and resulting code and data openness categories.

Model	Developer	Largest Model Size (params)	Release Date	Impact Metrics		
				ChatBot Arena Elo Rating	Google Scholar Citations	HuggingFace Downloads Last Month
GPT-2	OpenAI	1.5B	02/2019	N/A	8,015	17,984,300
T5	Google	11B	10/2019	873	12,162	3,295,844
GPT-3	OpenAI	175B	05/2020	N/A	18,759	N/A
mT5	Google	13B	10/2020	N/A	1,439	631,429
GPT-Neo	EleutherAI	2.7B	03/2021	N/A	N/A	242,580
GPT-J-6B	EleutherAI	6B	06/2021	N/A	465	95,620
CodeT5	Salesforce	16B	09/2021	N/A	703	23,549
Megatron-Turing	Microsoft, NVIDIA	530B	10/2021	N/A	379	N/A
Anthropic LM	Anthropic	52B	12/2021	N/A	70	N/A
ERNIE 3.0	Baidu	260B	12/2021	N/A	248	728
Gopher	DeepMind	280B	12/2021	N/A	598	N/A
GLaM	Google	1.2T	12/2021	N/A	255	N/A
XGLM	Meta	7.5B	12/2021	N/A	79	12,884
FairSeq Dense	Meta	13B	12/2021	N/A	34	6,129
LaMDA	Google	127B	01/2022	N/A	819	N/A
GPT-NeoX-20B	EleutherAI	20B	02/2022	N/A	364	37,122
PolyCoder	Carnegie Mellon	2.7B	02/2022	N/A	259	554
Chinchilla	DeepMind	70B	03/2022	N/A	245	N/A
PaLM	Google	540B	04/2022	1,004	2,342	N/A
OPT	Meta	175B	05/2022	N/A	1,105	191,115
UL2	Google	20B	05/2022	N/A	99	20,731
BLOOM	Big Science	176B	05/2022	N/A	814	1,172,142
GLM-130B	Tsinghua University	130B	10/2022	N/A	129	345
Pythia	EleutherAI	12B	12/2022	896	195	55,398
Anthropic 175B LM	Anthropic	175B	02/2023	N/A	55	N/A
LLaMA	Meta	13B	02/2023	800	2,793	N/A
GPT-4	OpenAI	N/A	03/2023	1,243	308	N/A
Claude	Anthropic	N/A	03/2023	1,149	N/A	N/A
Cerebras-GPT	Cerebras	13B	03/2023	N/A	23	124,561
Stable LM	Stability AI	7B	04/2023	844	N/A	15,282
PaLM-2	Google	N/A	05/2023	N/A	372	N/A
OpenLLaMA	UC Berkeley	13B	06/2023	N/A	N/A	58,991
Claude-2	Anthropic	N/A	07/2023	1,131	N/A	N/A
LLaMA-2	Meta	70B	07/2023	1,077	1,197	742,238
Falcon	TII	180B	09/2023	1,035	65	1,341,297
GPT-3.5-turbo	OpenAI	N/A	09/2023	1,117	N/A	N/A
Mistral-7B	Mistral AI	7B	10/2023	1,023	15	510,471
Grok-1	xAI	N/A	11/2023	N/A	N/A	N/A
Phi-2	Microsoft	2.7B	11/2023	N/A	N/A	85,200
Gemini	Google DeepMind	N/A	12/2023	1,111	N/A	N/A

Table 2: **Model Information**: table containing the basic information about each of the models classified under the openness taxonomy. Developers highlighted in purple correspond to companies, in pink are non-profit entities, and in light blue are government institutes. All data accessed on 28th of December 2023.

Model	(1) Training						(2) Evaluation				(3) Deployment	
	Code			Data			Code		Data		Code	Data
	Pre-Training	Fine-tuning	Align-ment	Pre-Training	Super-vised FT	Align-ment	General Eval	Automatic Safety Eval	Utility Bench-marks	Safety Eval Datasets	Inference	Model Architecture and Weights
GPT-2	C1	N/A	N/A	D1	N/A	N/A	C1	N/A	D1	N/A	C5 (Mod. MIT)	D5 (Mod. MIT)
T5	C5 (Apache 2.0)	C5 (Apache 2.0)	N/A	D4 (ODC-By)	N/A	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)
GPT-3	C1	C1	N/A	D1	N/A	N/A	C1	N/A	D1	N/A	C1	D2
mT5	C5 (Apache 2.0)	C5 (Apache 2.0)	N/A	D4 (ODC-By)	N/A	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)
GPT-Neo	C5 (MIT)	C5 (MIT)	N/A	D5 (MIT)	N/A	N/A	C5 (MIT)	N/A	N/A	N/A	C5 (MIT)	D5 (MIT)
GPT-J-6B	C5 (Apache 2.0)	C5 (Apache 2.0)	N/A	D5 (MIT)	N/A	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)
CodeT5	C5 (BSD-3)	C5 (BSD-3)	N/A	D4 (CodeT5)	N/A	N/A	C5 (BSD-3)	N/A	N/A	N/A	C5 (BSD-3)	D5 (Apache 2.0)
Megatron-Turing	C1	N/A	N/A	D1	N/A	N/A	C1	N/A	N/A	N/A	C1	D1
Anthropic LM	C1	C1	N/A	D1	N/A	D5 (MIT)	C1	N/A	N/A	D5 (MIT)	C1	D1
ERNIE 3.0	C1	C1	N/A	D1	N/A	N/A	C1	N/A	N/A	N/A	C1	D1
Gopher	C1	C1	N/A	D1	N/A	N/A	C1	N/A	D1	D1	C1	D1
GLaM	C1	N/A	N/A	D1	N/A	N/A	C1	N/A	N/A	N/A	C1	D1
XGLM	C5 (MIT)	N/A	N/A	D5 (ComCra	N/A	N/A	C5 (MIT)	C1	N/A	D5 (Public datasets)	C5 (MIT)	D5 (MIT)
FairSeq Dense	C5 (MIT)	N/A	N/A	D5 (ComCra	N/A	N/A	N/A	N/A	N/A	N/A	C5 (MIT)	D5 (MIT)
LaMDA	C1	C1	N/A	D1	D1	N/A	C1	C1	D1	D1	C1	D1
GPT-NeoX-20B	C5 (Apache 2.0)	N/A	N/A	D5 (MIT)	N/A	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)
Poly-Coder	C5 (MIT)	N/A	N/A	? (D3 or D4)	N/A	N/A	C5 (MIT)	N/A	N/A	N/A	C5 (CC BY-SA-4.0)	D5 (CC BY-SA-4.0)
Chinchilla	C1	C1	N/A	D1	N/A	N/A	C1	N/A	N/A	N/A	C1	D1
PaLM	C1	C1	N/A	D1	D1	N/A	C1	N/A	N/A	N/A	C1	D1
OPT	C5 (MIT)	N/A	N/A	?	N/A	N/A	C1	N/A	N/A	N/A	C5 (MIT)	D3 (OPT Data)
UL2	C5 (Apache 2.0)	C5 (Apache 2.0)	N/A	D4 (ODC-By)	N/A	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)
BLOOM	C5 (Apache 2.0)	?	N/A	? (D3 or D4)	D5 (Apache 2.0)	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D4 (RAIL)
GLM-130B	C1	N/A	N/A	D1	N/A	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D3 (GLM-130B Data)
Pythia	C5 (Apache 2.0)	N/A	N/A	D5 (MIT)	N/A	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)
Anthropic 175B	C1	C1	C1	D1	D1	D1	C1	N/A	N/A	D1	C1	D1
LLaMA	C1	N/A	N/A	? (likely D5)	N/A	N/A	C1	C1	N/A	D5 (Publicly avail-able)	C4 (GNU GPL)	D3 (LLaMA)
GPT-4	C1	C1	C1	D1	D1	D1	C5 (MIT)	N/A	D1	D1	C1	D2
Claude	C1	C1	C1	D1	D1	D1	C1	N/A	N/A	D1	C1	D1
Cerebras-GPT	C5 (Apache 2.0)	N/A	N/A	D5 (MIT)	N/A	N/A	C5 (Publicly avail-able)	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)

Stable LM	C1	C1	N/A	D4 (CC BY-SA-4.0)	D1	N/A	C1	N/A	N/A	N/A	C5 (CC BY-SA-4.0)	D5 (CC BY-SA-4.0)
PaLM-2	C1	N/A	N/A	D1	N/A	N/A	C1	N/A	N/A	D5 (Publicly available)	C1	D1
OpenLLaMA	C5 (Apache 2.0)	N/A	N/A	D4 (RedPaja Data)	N/A	N/A	C5 (Apache 2.0)	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)
Claude-2	C1	C1	C1	D1	D1	D1	C1	C1	D1	D1	C1	D2
LLaMA-2	C1	C1	C1	D1	D1	D1	C1	N/A	N/A	D1	C3 (LLaMA-2)	D3 (LLaMA-2)
Falcon	C1	C1	C1	D4 (ODC-By)	D1	D1	C1	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Falcon-180B Data)
GPT-3.5-turbo	C1	C1	C1	D1	D1	D1	C5 (MIT)	N/A	D1	D1	C1	D2
Mistral-7B	C1	C1	N/A	D1	D1	N/A	C1	N/A	N/A	N/A	C5 (Apache 2.0)	D5 (Apache 2.0)
Grok-1	C1	C1	?	D1	D1	?	C1	N/A	N/A	N/A	C1	D2
Phi-2	C1	N/A	N/A	D1	N/A	N/A	C1	N/A	N/A	N/A	C5 (MIT)	D5 (MIT)
Gemini	C1	C1	C1	D1	D1	D1	C1	C1	D1	D1	C1	D2

Table 3: **Model Pipeline Classification**: openness classification of components of the training, evaluation and deployment pipelines of currently available large language models. “N/A” in this table corresponds to ”Not Applicable”, whereas “?” means the information is not publicly available. If a model has more than one source of code or data source for a given component, the final classification is taken by considering the strictest license. For conciseness, in the table header we use ”FT” as a stand in for ”Fine-Tuning”.

B.1. Open-source GenAI Governance

The urgency of assaying the risks and opportunities of open-source GenAI is further underscored by recent regulatory developments around the world. The EU AI Act (European Parliament, 2021) has since matured into the world's first comprehensive and enforcable regulatory framework on AI governance, and is set to introduce specific obligations to providers and deployers (users) of open-source general purpose AI models, and systems built thereon. President Biden's Executive Order on AI (House, 2023) is thought to significantly affect open-source developers also, and, of course, China's approach to AI regulation continuous to be governed by state intervention (Cyberspace Administration of China; Translate, 2023). While these regulations may carve in stone certain aspects of future open-source GenAI governance, fundamental questions surrounding concepts such as *general-purpose models of systemic risk* (EU AI Act) or *dual-use foundation models* (Biden's EO) remain up to debate. Importantly, particularly in the case of the EU AI Act, many regulations have been designed to be adaptable in line with future technological progress. Our debate therefore remains highly relevant to open-source GenAI governance.

Recent years have seen the emergence of regulatory frameworks across the world that are already, or will soon, interact with the real-world governance of open-source Gen AI models. These efforts have been accompanied by increasing efforts at streamlining on the international stage, starting from 2023 G7 Hiroshima Summit and the Bletchley declaration (The UK Government, 2023), and culminating in various national and transnational initiatives forming a network of AI safety institutes in the United Kingdom (UK), United States of America (US), European Union (EU), and elsewhere. Prior to the launch of ChatGPT on November 29th, 2023, such regulations were mostly targeted at (i) containing the spread of *deepfakes* in order to safeguard election integrity – e.g., the EU's 2022 amendments to the Digital Services Act –, or (ii) to exercise wider information control against the spread of "rumors", such as the Chinese government's 2019 *Regulations on the Administration of Online Audio and Video Information Services* (Sheehan, 2023). At the same time, the economic benefits of open-source AI models and systems have been almost unanimously recognized across the world. The launch of ChatGPT, and its rapid adoption among users worldwide, led policymakers to focus on general-purpose AI (GPAI) regulation.

B.1.1. The EU AI Act

The first *comprehensive* regulatory framework governing general-purpose AI – including provisions for open-source Gen AI – may be the EU AI Act, which is expected to come into full force by 2026 (European Parliament, 2021). The legislation will apply to anyone putting AI services, or their outputs, on the EU market for professional purposes, while exempting recreational or academic use, as well as matters relevant to national security. It guards providers of open-source general-purpose models against risks emanating from downstream use by limiting the providers' responsibilities to a number of transparency obligations. These transparency obligations include the high-level documentation of training data provenance, as well a specification of intended use cases. Entities deploying Gen AI *deepfakes* are required to disclose their AI-generated nature. These requirements will apply to small business owners to a lesser degree. While comprehensive, the EU AI Act will not apply to recreational or research use and will be superseded by the EU member states' individual national security interests. Open-source Gen AI providers may face additional procedures and obligations if their models are classified as *general-purpose AI (GPAI) models of systemic risk*, an intentionally vaguely defined criterion that will be adapted as technology progresses. Importantly, the EU AI Act, as perhaps the EU's first transnational legislation, explicitly affirms the economic benefits of open-source AI.

B.1.2. Biden's Executive Order

President Biden's 2023 *Executive Order (EO) on Safe, Secure, and Trustworthy Artificial Intelligence* (House, 2023) continues to follow a *"soft law"* approach of earlier EOs, largely trading enforceable regulation for voluntary industry commitments (PricewaterhouseCoopers, 2024). Safety and security measures surrounding AI technology include requirements for developers to share red-teaming results with the US federal government, and for companies working on "dual-use" foundation models (*i.e.*, systems with civilian and military applications) and/or with large compute clusters to provide regular activity reports. The National Institute of Standards and Technology (NIST) is set up to play a key role in developing standards for secure and safe AI. Instead of placing hard restrictions on the use of certain AI technology (as the EU AI Act explicitly does), Biden's EO focuses on promoting best practices, evaluations, and standard development across a wide variety of aspects including security and risk mitigation. For example, it includes references to biological weapons, AI-generated content watermarking, and labor market impacts, and, additionally, measures for attracting foreign national AI talent through streamlining visa procedures and by providing assistance to small businesses and developers. National security interests are also formulated, including the reporting of foreign users of US Infrastructure as a Service (IaaS) products, as well as promoting the development of AI-driven tools to detect cyber vulnerabilities.

B.1.3. China's Gen AI Legislation

The earliest legal framework specifically targeting Gen AI models and systems, the Chinese government's *Provisional Administrative Measures of Generative Artificial Intelli-*

gence Services (Generative AI Measures) (Cyberspace Administration of China; Translate, 2023), came into force in China in August 2023. These regulations pose strict obligations on providers of Gen AI, ranging from outcome-driven provisions (e.g., requiring generative AI services to not produce illegal or untruthful content) to provenance obligations on training data and model weights, and measures targeted to protect intellectual property and privacy rights (Steven Chong, 2023). From the point of view of open-source model developers, the inability to predict future downstream use of models and systems provided introduces legal risks that require regulatory containment. Although open-source Gen AI plays a significant role in the Chinese economy, however, these regulations do not seem to target open-source (GP)AI models specifically (Asia Society, 2024).

B.1.4. The Middle East

Saudi Arabia. In August 2019, as part of Saudi Arabia's Vision 2030 introduced by Crown Prince Mohammed Bin Salman, the Saudi Data and AI Authority (SDAIA) was established by a royal decree. SDAIA aims to advance this vision, with the National Center for AI serving as a key component. Saudi Arabia, through SDAIA, has adapted and released its first version of AI ethics in September 2023 (Data and Authority, 2023). The document outlines Saudi's stance on AI risks, categorized from minimal to unacceptable risks with a comprehensive risk management plan covering data, algorithms, compliance, operations, legality, and regulatory risks. The AI ethics strongly supports the transparent development and deployment of AI , reflecting that *"transparent and explainable algorithms ensure that stakeholders affected by AI systems [...] are fully informed when an outcome is processed by the AI"*. Moreover, SDAIA has quickly embraced the generative AI wave. In collaboration with NVIDIA, SDAIA developed "Allam" (Gazette, 2024), Saudi Arabia's first national-level LLM model, an Arabic LLM designed to provide summaries and answer questions, drawing information from cross-checked online sources. While Allam was closed source and only a beta version interface is accessible, there are still several pieces of evidences that Saudi Arabia is in favor of open-source. For instance, the Digital Government Authority (Digital Government Authority) issued free and open-source government software licenses to 6 government agencies in 2022. This entails reviewing and publishing the source code "in a way that opens the field of cooperation and unified standards among government agencies". The general directions with the laid down compliance regulations, stated principles, and open source government suggest that Saudi Arabia is in favor of open source.

United Arab Emirates (UAE). In October 2017, the UAE Government launched the *"UAE Artificial Intelligence Strategy"* (UAE, 2023), spanning sectors from education to space. Shortly after, His Excellency, Omar Al Olama became the world's first AI minister. The UAE has been in favor of open-source in their policies, for instance, as stated in the strategy *"Objective 7: Provide the data and the supporting infrastructure essential to become a test bed for AI"* and that *"The UAE has an opportunity to become a leader in available open data for training and developing AI systems"*. Moreover, the strategy states that *"The UAE's ambition is to create a data-sharing program, providing shared open and standardized AI-ready data, collected through a consistent data standard"*. More recently, the UAE through the Technology Innovation Institute (TII) has open-sourced its LLM Falcon (TII, 2023), including its 180B parameter version, for both research and commercial use (Reuters, 2023). This all indicates the UAE's positive take towards open-source models.

B.1.5. AI Regulation Efforts in Other Countries

In 2019, the Organization for Economic Co-operation and Development (OECD) introduced their AI Principles, a recommendation by the council on general-purpose AI. These principles were ratified by the G20 council, and have been adopted by at least 42 of the organization's participating countries (OECD; Australian Government, 2024a).

Some countries have on-going legislation efforts or have issued policies specifically on Gen AI, addressing mainly sector-based issues. These include Australia (Australian Government, 2024b), Canada (of New Zealand), New Zealand (Kaldestad, 2023), Norway (Council, 2023), Singapore (Monetary Authority of Singapore; Infocomm), among others. India has published working papers on the issue of AI and enacted the Digital Personal Data Protection Act in 2023 tackling privacy issues related to Gen AI (Kapoor et al., 2024) – it is yet to regulate on general-purpose Gen AI and the open sourcing of models. Brazil is working on two main legislative proposals to regulate AI, one inspired in the US framework (Bill no. 21, from 2021) and another inspired on the EU framework (Bill No. 2338, from 2023), yet these do not have provisions for open-source Gen AI models. A few other countries are in the process of running public consultations on how to regulate generative AI, such as the case of Chile (MinCiencia) and Uruguay (Agencia de Gobierno).