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

Francisco Eiras [1] **Aleksandar Petrov** **Bertie Vidgen** **Christian Schroeder de Witt** **Fabio Pizzati** **Katherine Elkins** **Supratik Mukhopadhyay** **Adel Bibi** **Botos Csaba** **Fabro Steibel** **Fazl Barez** **Genevieve Smith** **Gianluca Guadagni** **Jon Chun** **Jordi Cabot** **Joseph Marvin Imperial** **Juan A. Nolazco-Flores** **Lori Landay** **Matthew Thomas Jackson** **Paul Rottger** **Philip Torr** **Trevor Darrell** **Yong Suk Lee** **Jakob Nicolaus Foerster**

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. This regulation is likely to 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.

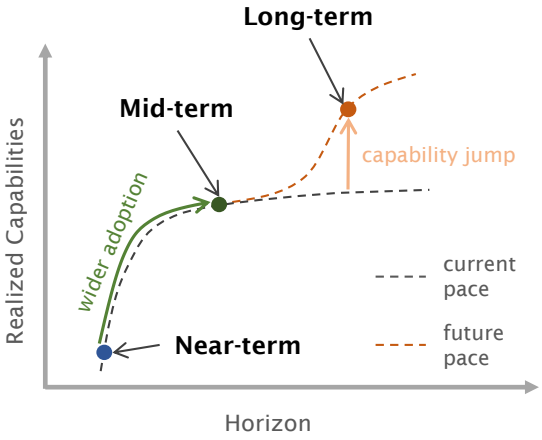

Figure 1: **Three Development Stages for Generative AI Models**: *near-term* is defined by early use and exploration of the technology in much of its current stage; *mid-term* is a result of the widespread adoption of the technology and further scaling at current pace; *long-term* is the result of technological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

1. Introduction

Generative AI (GenAI), 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).

Parallel to these efforts is a debate on the *openness of GenAI* 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 GenAI models have chosen to openly release 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 GenAI 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 generative AI with well-defined and followed principles for responsible development and deployment. To this end, we begin by defining

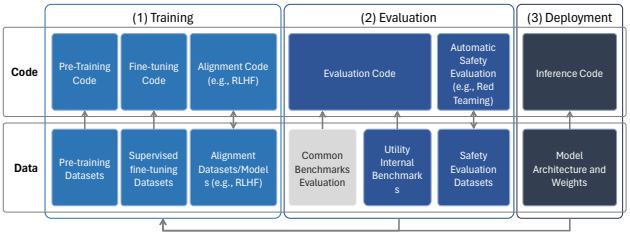

Figure 2: **Model Pipeline**: stages showing (1) training, (2) evaluation, and (3) deployment analyzed in the report. The component Common Benchmarks Evaluation (light gray) is included for completeness yet will not be analyzed in detail as these are standard and commonly available.

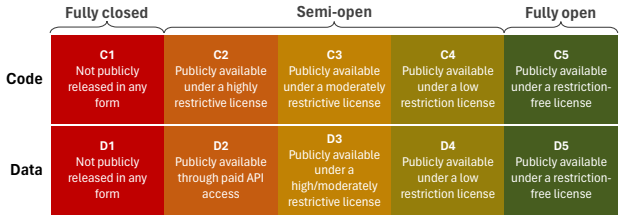

Figure 3: **Openness Scale**: categorization of the levels of openness of the code and data of each model component. See Table 1 (Appendix B) for the restrictions of each license.

different stages of GenAI development/deployment (near-, mid- and long-term), followed by an empirical analysis of the openness of existing models. With this framework, we then focus on evaluating the risks and opportunities presented by open and closed source GenAI 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 on how to achieve this safely and efficiently.

2. Preliminaries

To frame our analysis of the risks and benefits of responsibly 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 these models.

Stages of Development of GenAI 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 GenAI in the near to mid-term stages.

Training, Evaluating, and Deploying GenAI Models 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*. For practical purposes, we focus on Large Language Models (LLMs), but our definitions are generally applicable to other modalities (e.g., vision, multimodal). In the following, we briefly describe each of the components of the pipeline, with a more in-depth description presented 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 such as MMLU or NaturalQuestions (Hendrycks et al., 2020; Kwiatkowski et al., 2019) as well as curated benchmarks such as HELM or EleutherAI's Evaluation Harness (Liang et al., 2022; Gao et al., 2023). 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., 2023). 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 GenAI Models

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 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

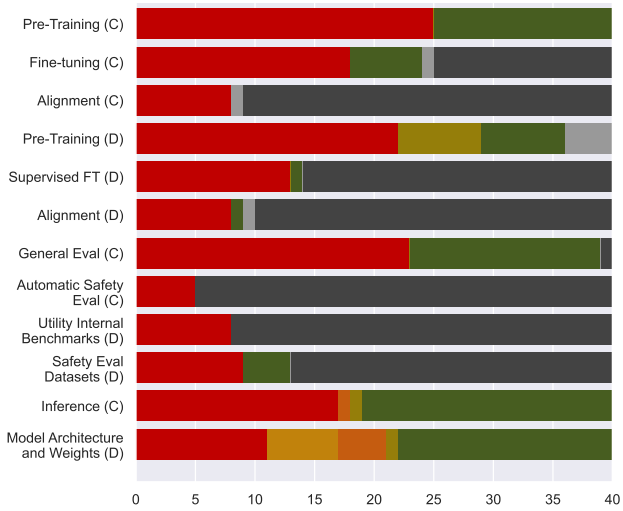

Figure 4: **Distribution of Openness Levels by Pipeline Component**: openness level distribution for each of the pipeline components of the 40 LLMs studied. Color legend: C1/D1, C2/D2, C3/D3, C4/D4, C5/D5, ? (unknown or not publicly available), N/A (not applicable). For conciseness, we use "FT" as a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

Alignment components (Touvron et al., 2023). To properly evaluate the openness of each component, we introduce a classification scale 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.

3.1. Classifying Openness for GenAI Code and Data

We introduce a framework for categorizing the openness of each component of the pipeline in 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 as detailed in 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 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 the 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 evaluation code and data***. As discussed in the next sections, for open source's advantages to be fully leveraged and risks mitigated, a significant shift in this landscape is necessary, achievable only through responsible open source generative AI development and deployment.

4. Near to Mid-term Risks and Opportunities of Open Source 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 opportunities compared to closed source – rather than generative AI in general.

Unless stated explicitly, we refer to all artifacts and compo-

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

nents of AI when using the term "open source".

The Challenges of Assessing Risks and Benefits Generative 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 crowdsourced 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 sociotechnical 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. Considering these aspects, we use a variety of evidence to critically examine and support our arguments. Nonetheless, it is important to recognize the inherent challenges in reaching definitive conclusions, underscoring the need for readers to recognize the evidence's limitations and for the community to improve its relevance and quality.

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. 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 Cheaper 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 greater 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., 2024; 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 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) as well as our recommendations in this paper.

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). In the long-term, efforts should be made to make pre-training datasets more diverse. 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 helps under-resourced actors 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 Democratizes AI With open source, any developer can leverage the investments of larger companies, governments, and research labs. Open source makes more resources available, which encourages reuse, thereby saving time, effort, and money. This is vital given the high costs and complexity of developing AI from scratch, from pre-training models, which can cost tens or hundreds of millions of dollars (Knight, 2023), to creating costly human-labeled datasets. This creates a clear societal benefit by enabling non-elites to access and use AI, which can even include creating economic opportunities (see §4.3).

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

5.1. Addressing Common Concerns when 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. Hard restrictions on open sourcing might not yield large improvements in overall safety, and could even hinder advancements in open safety technologies and auditing procedures.

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). 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. For instance, the White House Executive Order required 15 "leading companies" to "drive safe, secure, and trustworthy development of AI." Regulatory pressure is a lever for society to create incentives for safe model development. However, regulation is not a panacea and several closed source models have been, and could be, 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, regulation 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 stakeholders 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 GenAI Models

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

Enhance Data Transparency and Provenance Responsible open sourcing must be 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 GenAI 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, turn-key APIs, cloud integration, and liability protections. Therefore, comparing the models with their closest commercial closed source alternative is crucial to quantify, clarify, and understand the trade-offs involved in open sourcing decisions.

Perform Studies of Broader Societal Impact As highlighted in §4.5, properly developed open models can reduce GenAI energy consumption, aid in resolving copyright disputes, cater to diverse communities, and democratize AI. To realize these benefits, it's crucial to undertake, where relevant and feasible, comprehensive broader societal impact studies. These should evaluate corporate practices in model design and management, initiatives for enhancing data diversity and representation, and provide 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. We advocate for the **responsible open sourcing of near to mid-term GenAI models**. 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 GenAI 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 GenAI, and (2) on the debate around open sourcing these models.

The Impact of GenAI 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 GenAI Models A main line of discussion centers on the definition of open sourcing GenAI, 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 GenAI 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.

References

Leila Abboud and Javier Espinoza. 2023. EU's new AI Act risks hampering innovation, warns Emmanuel Macron. *Financial Times*.

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.

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*.

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*.

Connor Benedict. 2023. On openness & copyright, EU AI Act final version appears to include promising changes. *Creative Commons*.

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*.

Brendan Bordelon. 2023. Think tank tied to tech billionaires played key role in Biden's AI order. *POLITICO*.

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

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*.

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*.

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.

EUAIAct.com. 2023. Recitals - EU AI Act.

European Parliament. 2021. Artificial Intelligence Act.

European Parliament. 2023. Artificial Intelligence Act: Council and Parliament strike a deal on the first rules for AI in the world.

European Parliament. Directorate General for Parliamentary Research Services. 2020. The impact of the general data protection regulation on artificial intelligence.

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*.

Senado Federal. PL 2338/2023 - Senado Federal.

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*.

Future of Life Institute. 2023. EU AI Act Compliance Checker.

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*.

Justiça Eleitoral. Eleições 2024: audiência pública debaterá mudanças na prestação de contas.

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

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*.

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*.

Lockton. 2023. Preparing for the EU AI Act: compliance and insurance guidance for AI providers and deployers.

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*.

The Messenger. Think Tank With Tech Billionaire Backing Helped Craft Biden AI Executive Order: Report - The Messenger.

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.

Supantha Mukherjee, Martin Coulter, Foo Yun Chee, Supantha Mukherjee, and Foo Yun Chee. 2023. Explainer: What's next for the EU AI Act? *Reuters*.

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. 2024. Language model tokenizers introduce unfairness between languages. *Neural Information Processing Systems (NeurIPS)*.

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

POLITICO. How a billionaire-backed network of AI advisers took over Washington - POLITICO.

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

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*.

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.

Policy Review. The road to regulation of artificial intelligence: the Brazilian experience.

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.

Bruce Schneier and Nathan Sanders. 2023. Opinion | The A.I. Wars Have Three Factions, and They All Crave Power. *The New York Times*.

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.

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*.

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*.

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, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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.

M Abu Talib. 2017. Open source software in the UAE: Opportunities, challenges and recommendations (a survey research study). *Unpublished*.

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

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.

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*.

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*.

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*.

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 processes can be grouped into three distinct stages:

- The first stage is 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 [PILE, GPT-2]. 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 [PILE, LLaMA, GPT-J].
- The second stage is 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 [LIME, InstructGPT, LLaMA-2, Gemini], 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." [LLaMA-2].
- The third stage is alignment, 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) [InstructGPT, LLaMA-2] 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 [PPO]) to align the LLM. RLHF is not used in models released prior to 2022 [GPT-2, mT5, Megatron-Turing], and it is unclear whether the RLHF is used in models such as PaLM-2 [PaLM-2].

Once trained, models are usually evaluated on openly available evaluation datasets (e.g., MMLU, NaturalQuestions or TruthfulQA) [LLaMA-2, InstructGPT, GPT-4] as well as curated benchmarks (e.g. HELM, BigBench and EleutherAI's Evaluation Harness). 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 [LLaMA-2]. 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 [LLaMA-2, GPT-4]. They are typically the result of applying techniques such as red teaming.

Finally, at the deployment stage the final model is obtained. 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	Microsoft	2.7B	03/2021	N/A	N/A	242,580
GPT-J-6B	Microsoft	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	Microsoft	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	Microsoft	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”.

C. Regulatory Landscape, Governance and Geopolitics

C.1. The US Regulatory Landscape

On October 30, 2023 US President Biden signed an executive order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (House, 2023). The order is generally considered to represent a wide range of interests. There has been some criticism, however, about the influence of a handful of tech billionaires, with Politico detailing the extent to which certain parties were confirmed to be present at closed meetings (Bordelon, 2023). This criticism follows on previous concern that the Congressional hearings were unduly influenced by industry given both the nature of certain remarks as well as closed events like a private dinner that was not open to review or feedback.

On the one hand, the US Regulatory landscape evinces a holistic view that accounts for a variety of interested parties and tries to achieve a balance between regulation and innovation through responsible innovation. On the other hand, the US regulatory landscape cannot be seen as divorced from the larger lobbying forces under which some groups, including the leading industry players, may hold outsized influence. As Harvard scholars Schneider and Sanders write, the Biden act represents "a contest about control and power, about how resources should be distributed and who should be held accountable" (Messenger). Brendan Bordelon of Politico goes on to suggest that the act "papers over" the "growing tension between Washington's rival AI factions" (Schneier and Sanders, 2023).

Schneider and Sanders categorize the different factions represented in the executive act as "*doomsayers,*" "*reformers*" and "*warriors.*" Doomsayers highlight the existential risk AI poses. Reformers focus on the already extant series of harms being perpetrated, as well as potential looming economic disruption. Warriors are concerned with the geopolitical sphere amid worries the US will lose its global dominance. Each of these groups highlights the risks and benefits according to their particular concerns and viewpoints. The doomsayers have come under the most criticism for overstating the extent to which fears of existential risk are a dominant viewpoint among AI researchers and developers (POLITICO). Adding to this concern is the fact that highlighting existential risk may benefit leading industry players, for whom regulation could help as an insincere stratagem to cement their market advantage by handicapping smaller competitors.

Reflecting these various positions, the EOpresidential act outlines the various risks of AI and calls for various government agencies to develop plans that include developing critical infrastructure for defense and governmental support of AI research. NIST is tasked with developing practices surrounding their AI Risk Management Framework. To this end, NIST is in the process of creating the US AI Safety Institute Consortium to include members from state and local governments, universities, non-profits, and industry.

The Biden and Harris Executive Order stands apart from the EU regulatory landscape in one key aspect: models with "*freely available weights.*" In the EU, Models below a specific compute budget threshold have been protected from regulation, at least for the present. This is seen as either a positive protection of competition and development or an unfortunate capitulation to lobbying depending on the perspective. The Biden act, on the other hand, specifically calls out open-source models as a locus of potential risk.

C.2. Regulation in the EU: The EU AI Act

The EU AI Act is widely regarded as the world's first attempt at a comprehensive regulation of AI systems. It is important to recognise that this attempt did not happen in a vacuum, but followed and built upon other generalist regulatory initiatives, such as the General Data Protection Regulation (GDPR, 2016), and the Digital Markets Act (DMA, 2012). It is instructive to realise that the EU started examining the compatibility of the GDPR and AI as early as in 2020 (European Parliament. Directorate General for Parliamentary Research Services., 2020). Understanding the Act's future influence on open-source foundation models does necessitate a holistic investigation.

On December 9th, 2023, the Act was provisionally agreed between the European Parliament and the European Council. During the 3-day round of negotiations, the Act's scope was tightened. It was clarified that the Act does not apply outside of European law, does not infringe on the security competences of the member states or so-entrusted entities, nor any military or defence applications. Importantly, it was clarified that the Act would not apply to sole purposes of research and innovation, nor other non-professional use. The EU AI Act act aligns central AI taxonomies with those proposed by the OECD, defining an AI systems as "*a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions that influence physical or virtual environments.*" (European Parliament, 2023).

Roles. The Act distinguishes between the roles of provider, deployer, importer and distributor, and a distinction is made between models provided or deployed from within the EU, or from the outside in a third-country. Providers, i.e. those that develop AI systems with the intention to bring them to the market, whether paid or free, have pre-market obligations including an initial risk assessment, risk-specific compliance, as well as risk-specific post-market obligations. Deployers are users of AI systems that don't fall within a small number of non-professional use cases.

Risk classification systems. The EU AI Act defines two parallel risk classification systems for the regulation of AI models: one for general-purpose AI (GPAI) models and foundation models, and one for AI models of specific purpose. Models of specific-purpose are regulated according to a tiered risk system based on use cases, ranging from prohibited systems, high risk systems, to minimal risk systems. Prohibited systems are those with unacceptable risk to people's safety, security, and human rights. According to Article 28(2a), providers of high risk systems are subject to compliance obligations, including the establishment of risk and quality management systems, data governance, human oversight, cybersecurity measures, postmarket monitoring, and maintenance of the required technical documentation (Future of Life Institute, 2023). It is to be expected that these obligations will be further detailed in later, sector-specific regulation. Providers of minimal risk systems may attract limited obligations if interacting directly with people or if they manipulate or generate visual or audio content, but otherwise are only subject to limited transparency requirements.

General-purpose AI model. According to the December 6th Compromise Proposal (Lockton, 2023), 'General-purpose AI model' means an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is released on the market and that can be integrated into a variety of downstream systems or applications The EU AI Act imposes particular regulation on providers of general-purpose AI models of "*systemic risk*" (Article D), which it defines to be all models for which "the cumulative amount of compute used for its training measured in floating point operations (FLOPs) is greater than 10^{25}". Such providers need to register their model with the European Commission, and need to ensure a wide-ranging catalogue of safety and security criteria. The EU AI Act does appear to leave provisions for adjusting various technical criteria related to "*systemic risk*" classifications in the future (Benedict, 2023).

Exemptions for providers of open source AI systems The latest publicly known draft of the EU AI Act, Commission Compromise Proposal v2, as leaked by Politico (European Parliament, 2021), contains wide-ranging exemptions for providers of certain AI systems provided under free and open source software licences. To be exempt, the models provided may not fall within the "*systemic risk*" category, or otherwise exhibits harmful behaviour of unacceptable risk, such as manipulative or exploitative behaviour, social scoring purposes, and certain biometric identification (Title II), or certain AI systems that leave the user uninformed about their interaction with an AI system, emotion recognition systems or biometric categorisation systems, or AI systems producing "*deep fakes*" (Title IV Article 52 (European Parliament, 2023)).

Are "*open models*" *synonymous to "open source models*"? Interestingly, the leaked draft (European Parliament, 2021) distinguishes the above from providers of "*pre-trained AI models that are made accessible to the public under a licence that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available,*". It is to be noted that the term "*open source model*" is not used explicitly and the degree of legal overlap with open source software is not immediately clear, and hence we will refer to such models as "*open models*". Open models are not exempt from Article C(1)[(c),(d)], as well as Article D and Article 28(2a), the latter governing third-party obligations "along the AI value chain of providers, distributors, importers, deployers or other third parties" (European Parliament. Directorate General for Parliamentary Research Services., 2020) for high risk AI systems.

Are open GPAI models subject to high risk obligations? It is here that the EU AI Act becomes slightly confusing. On the one hand, GPAI models attract their very own risk categorisation into "basic", and "systemic" risk. On the other hand, European Parliament (2021) seemingly reimposes "*high risk*" regulation onto GPAI models if they are open. If Article 28(2a) is indeed meant to apply to open GPAI models, then, in our assessment, this may pose the perhaps biggest legal onus on providers of open GPAI models, insofar as GPAI models, by definition, might be readily used, and trivially adapted, for a wide range of tasks, including those in high risk categories, such as law enforcement, career-critical educational or vocational contexts, or credit scoring. Specifically, Article 28(2a) demands that "*The provider of a high risk AI system and the third party that supplies tools, services, components or processes that are used or integrated in the high risk AI system shall, by written agreement specify the information, capabilities, technical access, and or other assistance, based on the generally acknowledged state of the art, that the third party is required to provide in order to enable the provider of the high risk AI system to fully comply with the obligations under this Regulation.*" It remains unclear to what extent providers of open GPAI models could fulfil obligations under Article 28(2a) by explicitly prohibiting high risk downstream use in their model licences.

Transparency requirements. Under any circumstances, providers of open GPAI models are responsible for transparency obligations according to Article C(1)[(c),(d)]. These transparency requirements include respecting existing Union copyright law (Article C(1)[(c)]) according to Article 4(3) of the (Digital Single Market) Directive (EU) 2019/790

[14], and the need to make a "*sufficiently detailed summary of the content used for training of the general-purpose AI model*" (Article C(1)[(d)]). The consequences of these transparency requirements for open GPAI model providers have been examined in (Benedict, 2023). Importantly, Directive (EU) 2019/790 expands GDPR regulation to copyright owners who share content online, meaning that key GDPR rights, such as the "*right to opt out*" (Article 7 GDPR) [15] would require that copyright owners could ask for their data to be removed from open GPAI model training data. Another interesting aspect is the question of what constitutes a "sufficiently detailed summary" of training data (Benedict, 2023).

Are providers of open models obliged to implement watermarking provisions? Interestingly, the Politico leak (European Parliament, 2021) shows traces of what may have been last-minute changes to watermarking obligations for GPAI, and hence open GPAI, model providers. Specifically, in Article C(1)[(e)] on watermarking provisions one finds a remark that watermarking provisions have been moved to Title IV Article 52 instead. The now deleted Article C(4) would have explicitly made open GPAI models subject to watermarking provisions.

Draft Forensics. Reading the Commission Compromise v2 closely (European Parliament, 2021), we identify several confusing and seemingly contradicting statements. For example, European Parliament (2023) introduces Article D(4), which states that Article C(4) exemptions for open GPAI models do not apply to GPAI models of systemic risk. However, the preamble states that Article C(4) was bracketed (which it is not) and would be deleted. In addition, we note that Article 28(2a) obligations to open models are bracketed in the preamble without further comment (European Parliament, 2023). This may indicate ongoing debate over whether high risk obligations should indeed apply to providers of open models, including open GPAI models. We cannot help but note that these inconsistencies seem, in fact, not merely clerical, but may indeed reflect a high-level confusion over the very core of the EU AI Act: namely the seemingly conflicting risk categorisations of GPAI models versus non-GPAI models, made blatantly visible in the nexus of open GPAI model regulation.

Measures in support of innovation and small companies. In apparent response to criticism concerning the innovation-friendliness of prior comprehensive EU legislation, such as the GDPR, Commission Compromise v1 introduced several measures in support of innovation, such as e.g. allowing for regulatory sandboxes that provide controlled environments for developing, testing, and validating innovative AI systems to be used for real world tests under suitable conditions. In addition, a number of additional actions are announced that are meant to exempt smaller companies from certain regulation. We find it conceivable that such measures could, in principle, amplify the effect of open GPAI models on innovation in the startup scene.

Obligations for providers of open models in third countries. The EU AI Act applies to both providers and deployers located in a third country where the output produced by the system is used in the EU (European Parliament. Directorate General for Parliamentary Research Services., 2020). Third-country providers of open GPAI models aiming at the EU market therefore will need to fulfill the obligations discussed in this government. While it currently seems unclear how third country providers would choose to fulfill potential obligations surrounding high risk or systemic risk efficiently, it seems that underlying obligations stemming from DSM and GDPR regulation could be fulfilled through existing routes, such as transatlantic data sharing agreements.

Going forward. At the time of writing (January 2024), the EU AI Act is still in the recital phase (EUAIAct.com, 2023), meaning that further changes to the Act are not expected until it is submitted for ratification by the Committee of Permanent Representatives (Coreper) in, supposedly (Mukherjee et al., 2023), early 2024. At the Coreper, a blocking minority of as few as four countries could theoretically derail the law ahead of its publication in the official journal. While Germany, Italy, and France conceded their various calls for exemptions in December, Emmanuel Macron has since been vocal on his stance that the EU AI Act may still hamper innovation (Abboud and Espinoza, 2023).

As to whether the seeming confusion of to what extent non-GPAI risk obligations should apply to open GPAI models will significantly impact the providers of open GPAI models will also be affected by whether usage restrictions and the availability of suitable harmonised standards (see bracketed Article C(3)) will allow providers of open GPAI models to escape from the regulatory maze. As such, with respect to open models, the EU AI Act could still result in workable practices but fail to provide a plausible high-level categorisation of AI risks.

C.3. The Middle East

Saudi Arabia In response to Saudi Arabia's newly announced 2030 Vision, which was introduced by the Crown Prince Mohammed Bin Salman in April 2016, numerous government entities have emerged where advanced technology is their main focus. A key development is the August 2019 establishment of the Saudi Data and AI Authority (SDAIA) by a royal decree to facilitate advancing the 2030 vision with a National Center for AI as one of its arms. 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. The AI ethics reads "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 codes "in a way that opens the field of cooperation and unified standards among government agencies". The general directions with the laid down ahead-of-its-time thorough compliance regulations, stated principles, and open-source government suggests that Saudi Arabia is in the direction of supporting open-source.

United Arab Emirates In October 2017, the UAE Government launched the pioneering "UAE Artificial Intelligence Strategy" (UAE, 2023), spanning sectors from education to space. Shortly after, Omar Al Olama became the world's first AI minister. Studies (Talib, 2017) highlight significant awareness of open-source software in the UAE. This all indicates that similarly to Saudi Arabia, the UAE has had the ground fertile for slowly adopting open-source policies. Following the UAE AI strategy, 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".

C.4. Other countries

Besides the countries addressed above, at least 35 (OECD) other countries have deployed some form of AI policy or regulation (Australian Government, 2024a). Nonetheless, by January 2024, none of those have issued policies specifically about open AI, suggesting the leading geopolitics of regulation lies on the aforementioned forums, mainly the EU and the US.

[3]

It is important to acknowledge that some countries have issued policies specifically on Generative AI, addressing mainly sector-based issues. This is the case for example of Australia (i.e. guidelines for schools (Australian Government, 2024b; **?**) and for the public sector (of New Zealand)), Canada (i.e. general guidelines link and general principles (of New Zealand; **?**)], New Zealand (i.e. judicial system, (Kaldestad, 2023)], Norway (i.e. consumer protection, link), and Singapore (i.e. financial sector, (Monetary Authority of Singapore), and sandboxes, (Infocomm)). We can also point that other few 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).

It is also relevant to point out that AI policies started to appear mainly from 2017 onwards (Federal), while generative AI and open AI attracted attention of policymakers from mid-2023 onwards. As a hypothesis, we can suggest that even countries that had advanced AI regulatory debates for many years have not reached the point to debate generative AI, and even less, open AI. Brazil is a good case to illustrate that.

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). None of those have so far provision on generative AI or open AI. The EU-inspired bill is of particular interest. This proposal has been heavily influenced by the EU AI Act debate. Although heavily altered to address local demands, and enhanced with a human-rights approach that goes even beyond the EU framework (Review), no congressional debate addressed the issues of generative AI or open AI by the end of 2023. In fact, the only policy window in the country, so far, is a public consultation opened by the Supreme Electoral Count on AI-generated electoral materials (Justiça Eleitoral).

C.5. Geopolitics.

On the global stage, the Bletchley Declaration [link] and the G7 Hiroshima AI Process [link] are pioneering efforts to establish standards and guiding rules for AI developers and users. These initiatives aim to engage emerging national AI safety organizations around the world to form a network aligned with shared agreements and policies, which will influence ongoing discussions about open-source AI.

[3] Argentina, Australia, Canada, Chile, Colombia, Costa Rica, Egypt, Iceland, India, Israel, Japan, Kazakhstan, Kenya, Korea, Mauritius, Mexico, Morocco, New Zealand, Nigeria, Norway, Peru, Rwanda, Serbia, Singapore, South Africa, Switzerland, Thailand, Tunisia, Turkey, Uganda, Ukraine, Uruguay, Uzbekistan, and Vietnam.

The EU is advocating for regulation with its proposed EU AI Act, which awaits approval from individual EU countries but is already influencing AI regulatory standards.

With the rise of AI technologies, nations are entering a new kind of arms race. Leading countries have the opportunity to use AI as a means of soft power—even as an open-source tool—by providing, maintaining, and updating AI systems for developing nations, thus gaining strategic influence over AI implementation in those countries.