Towards Data Governance of Frontier AI Models

Jason Hausenloy*
University of California, Berkeley
hausenloy@berkeley.edu

Duncan McClements
University of Cambridge
dm2020@cam.ac.uk

Madhavendra Thakur
Independent
mt3890@columbia.edu

Abstract

Data is essential to train and fine-tune today's frontier artificial intelligence (AI) models and to develop future ones. To date, academic, legal, and regulatory work has primarily addressed how data can directly harm consumers and creators, such as through privacy breaches, copyright infringements, and bias and discrimination. Our work, instead, focuses on the comparatively neglected question of how data can enable new governance capacities for frontier AI models. This approach for "frontier data governance" opens up new avenues for monitoring and mitigating risks from advanced AI models, particularly as they scale and acquire specific dangerous capabilities. Still, frontier data governance faces challenges that stem from the fundamental properties of data itself: data is non-rival, often non-excludable, easily replicable, and increasingly synthesizable. Despite these inherent difficulties, we propose a set of policy mechanisms targeting key actors along the data supply chain, including data producers, aggregators, model developers, and data vendors. We provide a brief overview of 15 governance mechanisms, of which we centrally introduce five, underexplored policy recommendations. These include developing canary tokens to detect unauthorized use for producers; (automated) data filtering to remove malicious content for pre-training and post-training datasets; mandatory dataset reporting requirements for developers and vendors; improved security for datasets and data generation algorithms; and "know-your-customer" requirements for vendors. By considering data not just as a source of potential harm, but as a critical governance lever, this work aims to equip policymakers with a new tool for the governance and regulation of frontier AI models.

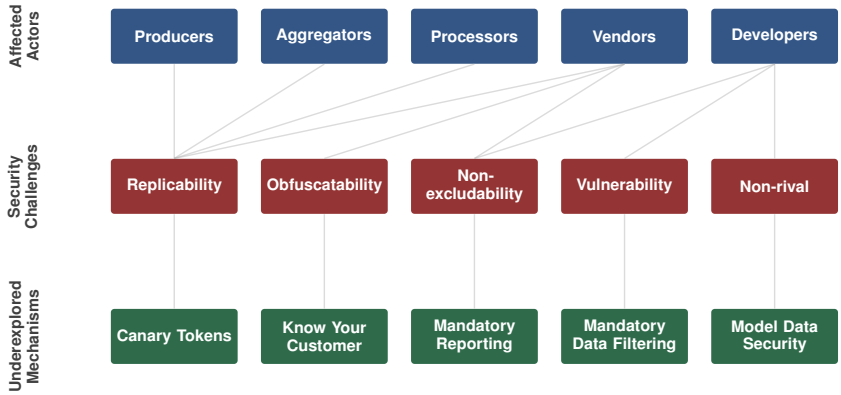

Figure 1: Proposed underexplored mechanisms in relation to challenges of data regulation and the AI Data Supply Chain.

*Corresponding author. Contact: hausenloy@berkeley.edu

2nd Workshop on Regulatable ML at NeurIPS 2024.

1 Introduction

The development of today's frontier artificial intelligence (AI) models [2] , highly capable foundation models, is inextricably linked to data, so much so that the systems are regularly defined by their training on "broad data at scale" [2, 14]. There is a growing scientific consensus that, as well as tremendous benefit, such models may pose risks to public safety [20, 11]. Yet, because of the rapid pace of AI development and the growing secrecy surrounding frontier model training, the production, aggregation, and processing of the datasets used by frontier models has thus far received little regulatory and public attention. As data is a key input to the pre-training and fine-tuning of models, we hope to demonstrate that governing data for frontier models, "frontier data governance", can be a promising approach to monitor and mitigate the risks as these models advance.

In particular, we focus on how policymakers can use the unique mechanisms within this approach to prevent the acquisition of specific dangerous capabilities, whether caused by malicious actors and potential misalignment. Existing "data governance" efforts emphasise detecting and preventing *direct harms* arising from the abuse or misuse of data on consumers and creators. Examples of such efforts include regulation to protect individual privacy [38, 16], data producer's copyright [27, 25], and nondiscrimination in the workplace [47]. [3] Instead, we focus on the comparatively neglected question of how data can enable new governance capacities for frontier AI models.

In this paper, we briefly overview previous attempts to govern data, and popular methods for regulating frontier AI models – namely "compute governance" and model evaluations. We then introduce "frontier data governance" as "the policies, practices and mechanisms that monitor, regulate, and control data throughout the AI development pipeline, to mitigate risks and ensure responsible development of frontier AI systems." We then explain the importance of datasets to the acquisition of specific dangerous capabilities, and to the scaling towards larger, potentially more dangerous advanced AI systems. After, we show why the inherent properties that make data so useful for training AI systems – it is non-rival, non-excludable and easily replicated – pose challenges for using it a governance lever, and that its current supply chain may leave it vulnerable to adversarial attacks, and obfuscatable from regulatory scrutiny. Finally, we propose 15 mechanisms, of which we recommend five exploratory policies:

1. **Canary tokens:** data producers could embed unique identifiers in (particularly dangerous) data to detect and prevent unauthorized use in AI models
2. **Mandatory data filtering:** model developers are required to implement a set of (automated) filtering processes to remove malicious or harmful content from training datasets
3. **Mandatory reporting requirements:** after a certain threshold, model developers and data vendors must disclose their pre-training and fine-tuning datasets to a third-party evaluator.
4. **Model data security:** model developers and data vendors should implement enhanced security measures to protect their datasets, and synethetic generation algorithms.
5. **Know your customer regulations:** data vendors are required to collect, verify and disclose the identity of developers requesting certain datasets to the government.

2 Related work

Governing data predates foundation models. In general, the term "data governance" has been used to refer to various mechanisms and techniques for the organizational handling of data [9, 72], and often in the context of AI [53]. In particular, there is a rich literature on technical methods for privacy-preserving machine learning, including differential privacy, federated learning, and homomorphic encryption [89, 32, 61, 39]. Policymakers have translated some of these techniques into regulatory

[2]Specifically, this paper will focus on mechanisms for regulating text-based frontier artificial intelligence systems, as these are the models with the most training data and the bulk of frontier models

[3]Note: To limit the scope of our paper, we do not focus on the direct harms arising from data, though we recognise these cannot be completely disentangled. For example, if private companies can train on private data (like personal cloud drives or video call transcripts, as some sources have speculated), this could be prevented through enforcing existing privacy regulation – even if this would ultimately have ramifications for the size of datasets that model developers may need for further scaling [42, 93]. Likewise, some outlined mechanisms, such as canary tokens or data filtering, may be employed to prevent privacy or copyright-infringing datasets.

frameworks, with a wide focus on various *direct harms* from data abuse and misuse, including privacy breaches, copyright infringement, algorithmic bias and discrimination [38, 16, 26].

On the other end, the regulation of frontier AI systems has instead converged to two popular paradigms. First, the governance of computing power ("compute"), which has taken the form of export controls, compute thresholds, on-chip verification, among others. [75, 58, 44, 15]. Second, model evaluations, testing trained AI models before deployment, are a key component of the safety plans of leading model developers and proposed and implemented regulatory regimes (licensing, third-party verification, auditing) [4, 71].

For governance, the early popularization of these two approaches is understandable. Compute is quantifiable and requires a highly-centralised supply chain, and evaluations generally provide a comprehensible indication of the lower bounds of a model's capability, to inform deployment decisions and development forecasts. [75, 71]. Yet, there are still limitations. Once compute has been allocated (eg. when the data centers have been constructed), we do not yet have the technology to reliably monitor their usage; algorithmic improvements mean that compute thresholds must be steadily lowered over time; evaluations can overlook emergent capabilities; correcting dangerous capabilities with existing techniques, such as reinforcement learning from human feedback, are still not perfect. [19]

Our paper lies at the intersection of data governance and frontier AI regulation, two fields that have been studied independently. Our contribution fits within the framework of the recent subfield of "technical AI governance." In particular, Reuel et al. [73] first introduces the subfield, and provide a comprehensive taxonomy of open research questions in data, compute, algorithms, and deployment, sorted by governance capacity. On data, they primarily focuses on the data's ability to enable assessment, access, verification and security from a primarily technical lens, providing a comprehensive survey of progress in machine learning [73], while we seek to explicitly introduce the approach of "frontier data governance" and detail specific policy mechanisms to combat properties of data we have identified.

3 Frontier data governance

3.1 Motivation

Data is a foundational input to AI models. Particularly for deep learning, models learn patterns, relationships and representations from the data they are trained on. The quality, quantity and nature of the training and fine-tuning data directly influence the model's capabilities, behaviors and potential risks. We follow the standard definition of frontier AI models as "highly capable foundation models that could exhibit sufficiently dangerous capabilities." [2, 14]. We then define frontier data governance as **"the policies, practices and mechanisms that monitor, regulate, and control data throughout the AI development pipeline, to mitigate risks and ensure responsible development of frontier AI systems."**

We briefly explore the risks from misuse and misalignment from frontier models, and data governance's role in mitigating these:

1. **Misuse:** Frontier AI models have the potential to acquire capabilities that pose societal-scale risks, such as for bioweapon design, cyber attacks or autonomous weapons. Malicious actors could intentionally elicit or fine-tune these capabilities, training AIs on specialized datasets [41, 79]. Compared to compute governance, which directly regulates the resources not the harmful content learned, or model evaluations, which occur after training, data governance can address the root cause by eliminating harmful content that provides these abilities from the training process, and can be enforced at multiple points throughout the supply chain.
2. **Scaling:** AI models exhibit emergent capabilities as they scale in size and complexity, often in unpredictable ways [3]. However, scaling requires not just compute but also vast amounts of high-quality data; optimal performance and capability growth depend on both compute and data. Data requirements increase approximately linearly with compute [46]. The availability of high-quality, diverse public data is finite. Estimates suggest that publicly available data suitable for training could be exhausted by 2026–2032 under current growth trajectories. [84]. Without enough new data, models can overfit, and additional compute yields diminishing returns [64, 90]. By controlling data access, we can prevent

rapid, uncontrolled scaling that may outpace our ability to manage associated risks, thereby aligning AI development with societal readiness.

3. **Misalignment:** More speculatively, frontier data governance may be able to align frontier models that could otherwise be misaligned with human values or intentions, reducing the likelihood of scenarios where control over these systems is compromised. Filtering out harmful, biased, or malicious content from training datasets may reduce the likelihood of models learning undesirable behaviors. If the training distribution reflects the ethical standards and societal values may help align AI behaviors.

Overall, frontier data governance may complement existing strategies, filling the gaps left by compute governance and model evaluations, and help proactive shape AI development, prevent dangerous capabilities, control scaling and even enhance alignment.

The AI data supply chain

Frontier models rely on diverse datasets at various stages of their development and deployment. The AI data supply chain involves multiple stages: **production** (creation of raw data by users, creators, researchers, and organizations), **aggregation** (gathering data through web scraping and purchases by tech giants and aggregators), **processing** (cleaning and structuring data by AI company teams and academic institutions), **pre-training** (using these large datasets to optimize model parameters, done by AI companies and research labs), **fine-tuning** (adapting models for specific tasks with smaller datasets, involving specialized providers), **retrieval** (optional, incorporating external knowledge during inference, often with content partners), and **evaluation** (assessing model performance using curated datasets, conducted by internal teams, auditors, and researchers).

Actor	Description and Examples
Data Producers	• Individual users (social media posters, video uploaders, bloggers) • Content creators (YouTubers, podcast hosts, journalists) • Researchers and academics (paper publishers, dataset sharers) • Businesses and organizations (report producers, website maintainers)
Data Aggregators	• Tech giants (Google, Meta, Microsoft) • Specialized data aggregators (CommonCrawl, WebCorpus)
Data Processors	• In-house teams at AI companies (OpenAI, Google DeepMind, Anthropic) • Data science departments at universities and research institutions • Cloud service providers
Model Developers	• Dedicated AI companies (OpenAI, Google DeepMind, Anthropic, Cohere) • Research divisions of tech giants (Microsoft Research, Meta AI) • Academic labs (Stanford AI Lab, MIT CSAIL) • Open-source communities (Hugging Face, EleutherAI)
Data Vendors	• Content partners (Shutterstock, Getty Images, TIME) • Specialized vendors (Scale AI, Surge AI, Appen) • Data marketplaces (Kaggle Datasets, AWS Data Exchange)
Evaluators	• Internal teams within AI companies • Third-party auditors (METR, Apollo Research) • Academic researchers • Government agencies (British and American AI Safety Institute)

Table 1: Actors in the AI Data Supply Chain

For a more comprehensive breakdown of the AI data supply chain stages and the key actors involved at each step, refer to Table 1.

4 Challenges

While the inherent properties of data—non-rivalry, non-excludability, and replicability—make it incredibly useful for training AI systems, they also pose challenges for using data as a governance tool. Additionally, the current data supply chain may leave data vulnerable to adversarial attacks and obfuscation from regulatory scrutiny. Below, we explain why each of these properties causes a challenge.

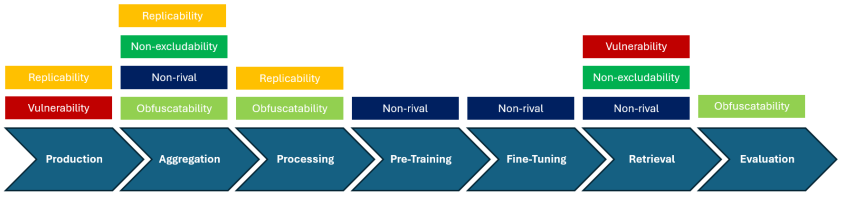

Figure 2: Challenges of regulating data across the data supply chain.

1. **Non-rivalry:** Data is a non-rivalrous good; one party's use doesn't diminish its availability or utility to others. While this has allowed for the widespread use and re-use of data in training AI models, this makes controlling or limits its use difficult [30].
2. **Non-excludability:** Data is often non-excludable; it's difficult to prevent unauthorized access once available. Preventing unauthorized access to data, especially once it is available online, where malicious actors can obtain and use data without authorization, making it hard to control who uses data and for what purposes [30].
3. **Replicability:** Data can be copied and replicated infinitely without degradation or significant cost. The ease of replicating data complicates efforts to control its distribution. Once data is shared or leaked, it can spread uncontrollably, making it nearly impossible to enforce restrictions or track all copies [30]. Data therefore has:
4. **Vulnerability** (to adversarial attacks)**:** Data is susceptible to poisoning and extraction attacks, as adversaries can introduce malicious data into training datasets.[4] This can cause models to learn incorrect or harmful behaviors, leading to unpredictable or dangerous outcomes. Furthermore, using a variety of in-context data-based attacks, information can be inferred or extracted from trained models.
5. **Obfuscatability:** Data can be acquired, transferred, and used in ways difficult to detect or monitor, often due to encryption, anonymization, or the use of covert channels. Unlike compute, which requires significant physical infrastructure (data centers, specialized hardware), data can be stored and transmitted using minimal hardware, such as on hard drives, making it less visible to oversight mechanisms.

For each of these challenges, we highlight how our five proposed mechanisms can uniquely help combat them (in bold), and further classify the remaining 11 similarly.

4.1 Synthetic data

Emerging technologies and industry trends introduce new complexities to frontier data governance, particularly the increasing use of synthetic data—data generated by AI models themselves, designed to mimic real-world data distributions [62, 60].

Furthermore, emergent technologies and industry trends cast doubt on the potential for data governance to act as a sustainable lever for regulation of the frontier. Particularly, as model training dataset sizes approach the limit of human-generated data, model developers are exploring opportunities for pushing the frontier. Although currently not a viable alternative for organic data [60], in the long term, synthetic data could supplant organic data as the primary source of training data in future generations of frontier models. This poses a threat to regulation of the data supply chain, as it relies on the progression of data from data producers to model developers; synthetic data, on the other hand, can be covertly generated and used by model developers, thus potentially allowing for unregulatable and adversarial data production and use. Likewise, the development of more data efficient model architectures, which many industry experts have suggested is the future, could threaten frontier data governance [87], as by reducing models' reliance on data, sufficiently large, and possibly dangerous, datasets could be created covertly, evading many of the detection and visibility mechanisms of frontier data governance.

[4]The barrier to entry for could be low. Public data sources can be manipulated with relatively few resources and technical expertise. Attackers can subtly alter data on websites or platforms that are commonly scraped for training data, embedding harmful content that could be ingested by model.

5 Underexplored mechanisms

Canary tokens

Canary tokens are unique identifiers or markers embedded within data to serve as tripwires, alerting *data producers* and *regulators* to unauthorized access or use, while empowering *data aggregators* and *data vendors* with greater control over the dissemination of their data. Canary tokens can help prevent foundation models from incorporating hazardous or sensitive information into their training datasets. Canary tokens help address the **easily replicable** nature of data by allowing for tracking and blocking of unauthorized replication. Furthermore, they strengthen a key link in the data supply chain, enhancing detectability and mitigating **vulnerabilities to adversarial attacks**.

Implementation: We propose data producers could include random strings or unique codes within webpages or documents containing potentially dangerous information. These tokens would be registered with the appropriate regulatory authority, who would privately inform the model developer about the tokens' presence and meaning. Model developers would be required to scan their training data for these canary tokens and exclude any content containing them.

Existing work: Canary tokens are widely used in cybersecurity to detect unauthorized access, acting as early warning systems against attackers [36, 33, 56]. Grosse et al. [40] demonstrated the use of influence functions to trace model outputs back to training data in large language models. Other methods can detect whether specific data has been used in training models [18, 66, 6, 92]. Canary tokens can be combined with such reverse data attribution to decrease computation expense and increase accuracy. Instead of analyzing influence across the entire training set, canary tokens provide unambiguous markers, eliminating the need for approximate influence calculations in many cases – even if not, models could quickly scan the much smaller space of known canary tokens.

Challenges and mitigation: The interests of data producers, data aggregators and data vendors to protect their own data will help justify the initial technical hurdle of implementing canary tokens across data.

Mandatory data filtering

Mandatory data filtering is a regulatory requirement for *model developers* to implement automated processes that detect and remove malicious, harmful, or unsafe content from training datasets before model training begins. Large language model (LLM) powered filtering of unsafe content in pre-training data could greatly reduce the risk of both the model passively absorbing unsafe content, thus eliminating the model's native potential to be unsafe, as well as targeted data poisoning attacks. This approach addresses the **vulnerable** nature of data to attacks by filtering out these attacks before training, while likely preserving the quality and safety of the information used to train AI models.

Implementation: We propose model developers and other data processors should integrate filtering mechanisms for safety into data preprocessing stages. Government safety agencies would provide the specific safety criteria, and auditors with access can verify their correct implementation and effectiveness (potentially in combination with mandatory reporting requirements below). Although LLM-based methods for data filtering exist, they are not widely used for model safety, but instead to support performance. For example, 25% of Llama-3's pretraining data mix consisted of mathematics and coding tasks [29]. Classifiers can still be used to recognize specific types of unsafe or low-quality content. New techniques such as influence functions to trace back which piece of training data most contributed to a particular dangerous model output, and then require that this subset of data is removed. [40]

Existing Work: Data filtering can be, and is often used to enhance model performance by removing irrelevant or low-quality data. Traditional filtering techniques focus on predefined heuristics for determining unsafe content [29]. Llama-3 uses classifiers like fastText and RoBERTa-based models to filter training data [29]. The use of LLMs to filter data allows for a more dynamic system with minimal prompting [83], and no need to define heuristics for the determination of "unsafe content", as in current safety filtering systems [29].

Challenges and mitigation: Implementing LLM-based filtering at scale can be expensive. This can be mitigated by the training/fine-tuning of specialised-models. Rather than a mandate, governments

could partner with model developers, as they partnered with cloud service providers for transparency measures of cloud services usage[48]. A mandate, in principle, already has some regulatory footing, such as in the European Union's AI Act, which sets minimum quality requirements for pre-training datasets. The potential for data filtering to improve the quality of a model also incentivizes its use by model developers, helping to justify the cost of data filtering.

Mandatory reporting requirements

Mandatory reporting requirements are regulatory policies that compel *model developers* and *data vendors* to disclose their pre-training and post-training dataset to a government auditor, potentially including additional information about their data practices, model training processes, and data transactions. The policy addresses data's **non-excludability** by ensuring datasets used by model developers are acquired through legitimate means, regulating the link between data vendors and their clients, and model developers and their sources. The approach also has the additional benefit of detecting the misuse of data, as highlighted by the recent public attention on NVIDIA's alleged scraping of YouTube data for AI training without proper authorization [76]. Furthermore, comprehensive reporting on training data and compute usage could provide valuable insights into the scaling achievements of various AI companies, offering a clearer picture of the competitive landscape and technological progress.

Implementation: We propose that regulatory bodies would provide clear guidelines on the specific information to be reported, including dataset sizes, sources, composition, data filtering techniques, and significant changes in data practices. A secure digital platform would be established for companies to submit their reports, ensuring data security and confidentiality. Similar to existing regulation, like SB1047, these requirements would only apply to developers training models over a certain threshold [13].

Existing work: To our knowledge, no one has proposed pre-training and post-training dataset reporting. Some AI companies do publish transparency reports or model cards that provide insights into their data practices and model characteristics [69].

Challenges and mitigation: Governments will need to build technical expertise and resources to effectively evaluate the large volumes of complex data reported. To reduce the resource-intensive burden on smaller companies, to start, these regulations will only apply above some monetary threshold for frontier model training. Strict security protocols can allay potential confidentiality and competitive concerns when companies are disclosing this proprietary or sensitive information, which can also help address similar consumer privacy concerns.

Model data security

Model data security involves *model developers* and *aggregators* implementing robust security measures to protect pre-training, post-training datasets and synthetic data generation algorithms from unauthorized access, theft, or tampering. This extends existing security practices used for safeguarding model weights to the data used in training AI models [67, 37, 80]. Currently, model developers take various precautions to protect model weights. We recommend that the same precautions be taken to secure pre-training and post-training data. The theft of pre-training data, for example, could allow for the recreation of the model or a model of similar performance [50]. Recent work has shown that post-training and fine-tuning data could be essential to the improvement and alignment of models [52]; the theft of this data, thus, would not only allow for the performance improvement of potentially adversarial models, but also the potential adversarial misalignment of frontier models. Increasing security addresses the challenge of data's **non-rival** nature by discouraging unauthorized sharing and access.

Implementation: We propose that model developers establish access control, encryption, secure computing environments and security auditing to model data – in practice, many of these recommendations are similar for those to defend model weights [67]. These security measures would have to be mandated by regulators and implemented by model providers. Regulatory precedent exists for such mandates, such as stringent requirements for protection of personal data and bank information[68, 38]. Due to the potential presence of said sensitive data in training data sets, the same regulatory mechanisms could be used to mandate the protection of training data[57]. Furthermore, out of a

desire to protect their models and maintain the competitive edge of superior data, model developers have an incentive to protect their data in such a manner.

Existing work: In addition to encryption and access control, which are already industry with regards to model weights [67], increasingly, developers have begin to watermark model weights to track downstream usage [37, 59]. We recommend the implementation of these novel methods to defend model data as well, both due to its technical precedent [43] and the importance of data to model security.

Challenges and mitigation: More thorough protection of data, however, would be challenging, as it would have to protect against mechanisms which allow for the extraction of pre-training data, even in black box models [65]. There appear to be few, if any, widely recognized mechanisms for effectively addressing this type of attack. The techniques are still not well-developed, though, and more speculative security measures (such as output monitoring and limiting output bandwidth) may help.

Know your customer regulations

Know your customer (KYC) regulations for *data vendors* require these vendors to verify and document the identities of their customers, often *model developers* particularly for transactions involving significant quantities of data, or types of data. Data vendors should consider implementing KYC procedures, and thereby verifying customer identities. Further, customers purchasing large amounts of data are mandated to provide verifiable identification. This policy addresses the challenge of data's **obfuscatability** by reducing the potential for covert data transactions.

Implementation: We propose that data vendors collect identifying information, and verify through independent sources or services, and maintain secure records of customer identities and transcation details. Regulators should set specific transaction thresholds above which KYC procedures based on risk assessments.

Existing work: KYC is well-established in finance to prevent money laundering and fraud [12]. It is gaining traction for regulating compute and cloud infrastructure [55, 48]. At present, model developers have untraceable access to immense datasets through private transactions with data vendors [70], leading to an opaque data supply chain which allows adversaries to develop potentially unsafe frontier models covertly.

Challenges and mitigation: The upfront cost for data vendors in establishing KYC processes can be streamlined by governments providing industry-wide KYC standards.

6 Other mechanisms

Restricting and monitoring fine-tuning access (model developers; vulnerability). For highly sensitive or capable models, directly limiting access to fine-tuning capabilities (eg. from the API) or requiring identification can prevent unauthorized or malicious modifications of AI models. Data poses risks at all stages of the pipeline, including fine-tuning, where technical frameworks exist for the cheap and covert exploitation of these post-training methods to instantiate unsafe agents [79, 41]. Additionally, fine-tuning can be exploited through model poisoning attacks, where adversaries subtly manipulate the model's behavior during the fine-tuning process [5]. The democratization of fine-tuning holds serious AI safety risks: "defenders," or people interested in protecting AI, tend to have more resources, so as costs fall, more additional "attackers" gain fine-tuning access than "defenders" [21]. Thus, monitoring and restricting fine-tuning access, such as through license requirements or background checks, would mitigate these risks by blocking adversaries from accessing such resources. This approach already has some regulatory backing, covered under "unsafe post-training modifications" in California's SB1047 [13], and is also addressed in the proposed European Union's AI Act [24].

Restricting access to dangerous datasets (data collectors, data processors, data vendors; non-excludability). Unsafe or malicious datasets can be used to create unsafe AI systems at various stages of the data pipeline, from pre-training to fine-tuning [79, 41]. Regulating upstream data sources can significantly reduce these downstream risks. Measures such as mandating licenses, conducting background checks, implementing paywalls for accessing potentially unsafe datasets, or restricting the

public dissemination of particularly dangerous datasets can prevent untraceable adversarial access to these resources. Existing regulatory initiatives, such as the European Union's General Data Protection Regulation (GDPR) [38] and the California Consumer Privacy Act [16], aim to protect sensitive data from unauthorized access. These initiatives should be extended to cover data that could lead to the development of unsafe AI systems.

Output Watermarking (model developers; replicability). Implementing watermarking techniques to attribute data and identify the source model responsible for generating specific outputs allows for better filtering and tracking of AI-generated content. Model output watermarking involves embedding undetectable and unremovable tracing data into the outputs generated by AI models [43]. While traditionally used by model developers to protect intellectual property rights [59, 74], output watermarking holds significant regulatory potential, across different modalities (including text, image, audio, video). Mandating model watermarking would allow regulators to trace unsafe or malicious content back to the source models, facilitating the identification and regulation of unsafe AI systems. Recent advancements have focused on developing robust watermarking methods for generative models to ensure accountability and mitigate misuse [1, 91].

Mandatory attribution for web scrapers (data aggregators; vulnerability). Requiring data aggregators to attribute the sources of their scraped data through standardized meta-reporting ensures transparency and accountability in data collection practices. The opaque processes of data aggregators often make it difficult to trace unsafe or harmful content back to its original source once it is included in a web-scraped dataset. This lack of traceability allows such content to proliferate unchecked into other datasets, potentially leading to unsafe AI models trained on this data [8]. Mandatory and standardized attribution would not only facilitate traceability but also hold web scrapers accountable, enabling regulators to enforce safe scraping practices and prevent the spread of harmful content [63].

Auditing retrieval data (model developers; vulnerability). Utilizing retrieval techniques can significantly enhance AI performance, especially with larger context windows [28]. However, these methods can also be exploited to circumvent restrictions on user prompts and elicit harmful information. By providing a sufficiently large number of examples of undesirable behavior, malicious actors can manipulate the model to produce unsafe outputs [22]. Since this vulnerability relies solely on input data, adversaries can create and distribute packages that are easily replicable to obtain harmful information. To mitigate these risks, it is crucial to audit and monitor retrieval data. User inputs can be analyzed to detect the presence of harmful examples; such inputs can then be entirely refused, portions that fail safety tests can be removed from model responses, or accounts that repeatedly submit such prompts can be tracked and suspended if unsafe behavior continues [86]. Implementing robust content filtering and input validation mechanisms can further enhance security [88]. Additionally, employing techniques like differential privacy during training can help prevent the leakage of sensitive information [31].

Decentralized volunteer classification for safe datasets (data aggregators, data vendors, data processors; obfuscatable). To break out of the chicken-and-egg paradox of relying on AI models to assess the safety of datasets used to train themselves, we propose leveraging volunteer classification efforts. Projects like Galaxy Zoo, launched in 2007, demonstrate the effectiveness of citizen science, where public volunteers classified galaxies and provided valuable training data that eventually enabled machine learning models to automate these tasks [34]. Similarly, volunteers could be enlisted to review random pieces of data and classify their safety according to specified criteria. Once sufficient labeled data is gathered from volunteers, a classifier could be trained to automate verification in the future, with periodic updates from volunteers as the relative safety of data evolves. This approach harnesses the collective intelligence of the public to enhance dataset safety, ensuring that AI models are trained on vetted data [49, 54]. Moreover, involving volunteers in data classification promotes transparency and public trust in AI systems [35]. However, challenges such as ensuring annotation quality and protecting volunteers from exposure to harmful content must be carefully managed [77].

Implementing automated classifiers on input and output (model developers; vulnerability). Fine-tuning a small language model to detect signs of data extraction and other common attacks can provide an additional layer of security on top of existing large language models (LLMs). Data-based input attack vectors are attractive to adversaries due to their accessibility; for instance, attacks on web-scraped datasets can be executed simply by injecting large amounts of data expressing misaligned intent [85]. Similarly, output attacks, such as data extraction, involve malicious prompting that exploits vulnerabilities in the data pipeline, potentially leading to unauthorized access to sensitive information

[18]. Implementing input/output classifiers can detect many of these common attacks, blocking rudimentary threats with relatively simple technologies [51]. These classifiers act as gatekeepers, analyzing inputs and outputs for signs of malicious activity and preventing the exploitation of LLMs.

Differential data spread (data producers, data aggregators, data vendors, model trainers; vulnerability). Implementing techniques to assign greater weighting to positive or beneficial data—or increasing the raw quantity of such data in datasets—can promote more favorable outcomes in AI models. For example, incorporating "good stories" about AGI or emphasizing ethical and aligned content can steer models toward safer behaviors. Meta's Llama 2 model demonstrates the possibility of intentional data weighting by model developers; their pre-training data mix was curated to contain specific proportions of content types, showing that certain types of content can be deliberately emphasized in the training data [81]. Techniques like curriculum learning, where training data is presented in a meaningful order to improve learning efficiency and performance, support the efficacy of intentional data weighting [10]. Beyond this, differential data spread could be executed through the large-scale creation of beneficial data by data producers, an increased sampling of beneficial data by data aggregators, such as through duplication of beneficial data, and the curation of datasets with high amounts of beneficial data by data vendors.

Refine existing regulations to include mechanisms of data generation and collection (regulators, evaluators; synthetic data). Existing regulatory frameworks for data privacy are invaluable but need to evolve as methods of data collection and generation advance. For example, the EU's General Data Protection Regulation (GDPR) defines personal data but does not explicitly address synthetic or AI-generated data that can be linked back to individuals [38]. Expanding the GDPR's definition to explicitly include synthetic and AI-generated data would extend regulatory oversight to these novel forms. Similarly, consent mechanisms mandated by regulations like the California Consumer Privacy Act (CCPA) need to evolve beyond explicit data collection to address AI systems capable of inferring sensitive information from aggregated non-sensitive data and AI models that can infer personal information from seemingly anonymized data [78, 17] or regulate non-sensitive data itself [16]. Updating these privacy regulations and incorporating provisions from recent legislative proposals regarding AI-specific risk, such as the EU's proposed Artificial Intelligence Act [24] is essential for the effective governance of emergent technologies like synthetic data and would enhance protections against potential misuse of AI systems.

Proof of Training Data Verification (model developers, regulators; non-excludability). Proof-of-Training-Data is an emerging tool for verifying machine learning training datasets [23]. This mechanism enables regulators to audit models for correct dataset attribution, confirming they haven't been trained on unauthorized datasets and acting as an enforcement mechanism. For instance, regulators could use it to detect if a developer used an illicit, sensitive dataset to create a harmful AI agent. Data provenance tracking [45], model watermarking [82], and cryptographic verification [7] enhance this process by ensuring transparency, verifying ownership, and preserving data privacy. This approach addresses the regulatory challenge of black-box models, where outputs can't be traced back to specific training data, supporting accountability for model developers after training.

7 Conclusion

In this paper, we sought to introduce data's role in governing frontier AI models. In particular, as frontier models get more powerful, new vulnerabilities will need to be addressed, and new mechanisms will be required for policymakers to respond. We provided a brief overview of 15 technical mechanisms, which have received varying previous attention, and introduced five, thus far, unexplored central recommendations – canary tokens, data filtering, reporting requirements, data security and know-your-customer regulation – for combating these challenges. Beyond this, there is significant scope for future policy research exploring how existing regulatory regimes (particularly those governing data, which are among the most developed) can be adapted and leveraged for frontier data governance, as well as technical work estimating and formalising the various assumptions of our policy mechanisms.

Acknowledgements

We'd like to thank Samuel Ratnam and Ian Habich-Ramirez for their helpful feedback and comments.

References

[1] Yossi Adi et al. “Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring”. In: *27th USENIX Security Symposium (USENIX Security 18)*. 2018, pp. 1615–1631.

[2] Markus Anderljung et al. *Frontier AI Regulation: Managing Emerging Risks to Public Safety*. 2023. arXiv: 2307.03718 [cs.CY]. URL: https://arxiv.org/abs/2307.03718.

[3] Alexander Andonian. “Emergent Capabilities of Generative Models:“Software 3.0” and Beyond”. PhD thesis. Massachusetts Institute of Technology, 2021.

[4] Anthropic. *Anthropic’s Responsible Scaling Policy*. en. Anthropic is an AI safety and research company that’s working to build reliable, interpretable, and steerable AI systems. 2023. URL: https://www.anthropic.com/news/anthropics-responsible-scaling-policy.

[5] Eugene Bagdasaryan et al. “How to backdoor federated learning”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2938–2948.

[6] Yang Bai et al. “Special characters attack: Toward scalable training data extraction from large language models”. In: *arXiv preprint arXiv:2405.05990* (2024).

[7] Eli Ben-Sasson et al. “Scalable Zero Knowledge via Cycles of Elliptic Curves”. In: *Advances in Cryptology–CRYPTO 2019* (2019), pp. 686–716.

[8] Emily M. Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM. 2021, pp. 610–623.

[9] Olivia Benfeldt, John Stouby Persson, and Sabine Madsen. “Data Governance as a Collective Action Problem”. In: *Information Systems Frontiers* 22.2 (Apr. 2020), pp. 299–313. ISSN: 1572-9419. DOI: 10.1007/s10796-019-09923-z. (Visited on 09/12/2024).

[10] Yoshua Bengio et al. “Curriculum Learning”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. 2009, pp. 41–48.

[11] Yoshua Bengio et al. “Managing extreme AI risks amid rapid progress”. In: *Science* 384.6698 (2024), pp. 842–845. DOI: 10.1126/science.adn0117. eprint: https://www.science.org/doi/pdf/10.1126/science.adn0117. URL: https://www.science.org/doi/abs/10.1126/science.adn0117.

[12] Genci Bilali. “Know Your Customer - Or Not”. In: *University of Toledo Law Review* 43 (2011/2012), p. 319.

[13] *Bill Text - SB-1047 Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.* (Visited on 09/12/2024).

[14] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG]. URL: https://arxiv.org/abs/2108.07258.

[15] Miles Brundage et al. *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. 2020. arXiv: 2004.07213 [cs.CY]. URL: https://arxiv.org/abs/2004.07213.

[16] *California Consumer Privacy Act (CCPA) | State of California - Department of Justice - Office of the Attorney General*. https://oag.ca.gov/privacy/ccpa. (Visited on 09/11/2024).

[17] Nicholas Carlini et al. “Extracting Training Data from Large Language Models”. In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 2633–2650.

[18] Nicholas Carlini et al. “Extracting training data from large language models”. In: *30th USENIX Security Symposium (USENIX Security 21)*. 2021, pp. 2633–2650.

[19] Stephen Casper et al. *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. 2023. arXiv: 2307.15217 [cs.AI]. URL: https://arxiv.org/abs/2307.15217.

[20] Center for AI Safety. *Statement on AI Risk*. en. https://www.safe.ai/work/statement-on-ai-risk. A statement jointly signed by a historic coalition of experts: “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.”. 2023.

[21] Alan Chan et al. *Hazards from Increasingly Accessible Fine-Tuning of Downloadable Foundation Models*. Dec. 2023. arXiv: 2312.14751 [cs]. (Visited on 09/11/2024).

[22] Pengzhou Cheng et al. “TrojanRAG: Retrieval-Augmented Generation Can Be Backdoor Driver in Large Language Models”. In: *arXiv preprint arXiv:2405.13401* (2024).

[23] Dami Choi, Yonadav Shavit, and David Duvenaud. *Tools for Verifying Neural Models' Training Data*. 2023. arXiv: 2307.00682 [cs.LG]. URL: https://arxiv.org/abs/2307.00682.

[24] European Commission. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Accessed: 12/09/24. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[25] *Digital Millenium Copyright Act*. URL: https://www.copyright.gov/dmca/.

[26] *Directive 2019/790*. en. URL: https://eur-lex.europa.eu/eli/dir/2019/790/oj.

[27] *Directive No. 96/9/EC of the European Parliament*. URL: https://www.wipo.int/wipolex/en/text/126788.

[28] Qingxiu Dong et al. "A survey on in-context learning". In: *arXiv preprint arXiv:2301.00234* (2022).

[29] Abhimanyu Dubey et al. *The Llama 3 Herd of Models*. Aug. 2024. arXiv: 2407.21783 [cs]. (Visited on 09/10/2024).

[30] Néstor Duch-Brown, Bertin Martens, and Frank Mueller-Langer. *The Economics of Ownership, Access and Trade in Digital Data*. SSRN Scholarly Paper. Rochester, NY, Feb. 2017. DOI: 10.2139/ssrn.2914144. (Visited on 09/10/2024).

[31] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science, 2014.

[32] Cynthia Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis". en. In: *Theory of Cryptography*. Ed. by Shai Halevi and Tal Rabin. Berlin, Heidelberg: Springer, 2006, pp. 265–284. ISBN: 9783540327325. DOI: 10.1007/11681878_14.

[33] Shehroze Farooqi et al. "Canarytrap: Detecting data misuse by third-party apps on online social networks". In: *arXiv preprint arXiv:2006.15794* (2020).

[34] Lucy Fortson et al. "Galaxy zoo". In: *Advances in machine learning and data mining for astronomy* 2012 (2012), pp. 213–236.

[35] Chiara Franzoni and Henry Sauermann. "Crowd Science: The Organization of Scientific Research in Open Collaborative Projects". In: *Research Policy* 43.1 (2014), pp. 1–20.

[36] Daniel Fraunholz et al. "On the detection and handling of security incidents and perimeter breaches-a modular and flexible honeytoken based framework". In: *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE. 2018, pp. 1–4.

[37] *Frontiers | A Systematic Review on Model Watermarking for Neural Networks*. https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2021.729663/full. (Visited on 09/09/2024).

[38] *General Data Protection Regulation (GDPR) – Legal Text*. https://gdpr-info.eu/. (Visited on 09/09/2024).

[39] Craig Gentry. "Fully homomorphic encryption using ideal lattices". In: *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*. STOC '09. Bethesda, MD, USA: Association for Computing Machinery, 2009, pp. 169–178. ISBN: 9781605585062. DOI: 10.1145/1536414.1536440. URL: https://doi.org/10.1145/1536414.1536440.

[40] Roger Grosse et al. "Studying large language model generalization with influence functions". In: *arXiv preprint arXiv:2308.03296* (2023).

[41] Danny Halawi et al. *Covert Malicious Finetuning: Challenges in Safeguarding LLM Adaptation*. June 2024. DOI: 10.48550/arXiv.2406.20053. arXiv: 2406.20053 [cs]. (Visited on 09/11/2024).

[42] Craig Hale. *Gemini AI platform accused of scanning Google Drive files without user permission*. en. July 2024. URL: https://www.techradar.com/pro/security/gemini-ai-platform-caught-scanning-google-drive-files-without-user-permission.

[43] Frank Hartung and Martin Kutter. "Multimedia watermarking techniques". In: *Proceedings of the IEEE* 87.7 (1999), pp. 1079–1107.

[44] Lennart Heim and Leonie Koessler. *Training Compute Thresholds: Features and Functions in AI Regulation*. 2024. arXiv: 2405.10799 [cs.CY]. URL: https://arxiv.org/abs/2405.10799.

[45] Melanie Herschel, Robert Diestelkämper, and Houssem Ben Lahmar. *Provenance: An Introduction to PROV in Data Science*. Morgan & Claypool Publishers, 2017.

[46] Jordan Hoffmann et al. *Training Compute-Optimal Large Language Models*. 2022. arXiv: 2203.15556 [cs.CL]. URL: https://arxiv.org/abs/2203.15556.

[47] The White House. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. Oct. 2023. (Visited on 09/13/2024).

[48] The White House. *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*. https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/. July 2023. (Visited on 09/05/2024).

[49] Jeff Howe. "The Rise of Crowdsourcing". In: *Wired Magazine* 14.6 (2006), pp. 1–4.

[50] Andrew Ilyas et al. *Datamodels: Predicting Predictions from Training Data*. Feb. 2022. DOI: 10.48550/arXiv.2202.00622. arXiv: 2202.00622 [cs, stat]. (Visited on 09/09/2024).

[51] Hakan Inan et al. *Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations*. 2023. arXiv: 2312.06674 [cs.CL]. URL: https://arxiv.org/abs/2312.06674.

[52] *Intuitive Fine-Tuning: Towards Unifying SFT and RLHF into a Single Process*. https://arxiv.org/html/2405.11870v1. (Visited on 09/09/2024).

[53] Marijn Janssen et al. "Data Governance: Organizing Data for Trustworthy Artificial Intelligence". In: *Government Information Quarterly* 37.3 (July 2020), p. 101493. ISSN: 0740-624X. DOI: 10.1016/j.giq.2020.101493. (Visited on 09/12/2024).

[54] Aniket Kittur, Ed H. Chi, and Bongwon Suh. "Crowdsourcing User Studies with Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2008, pp. 453–456.

[55] *Know-Your-Customer Is Coming for the Cloud—The Stakes Are High*. https://www.lawfaremedia.org/article/know-your-customer-is-coming-for-the-cloud-the-stakes-are-high. (Visited on 09/05/2024).

[56] Ioannis Koutsikos. "Improving Infrastructure Security using Deceptive Technologies". In: (2024).

[57] Christopher Kuner et al. "Machine Learning with Personal Data: Is Data Protection Law Smart Enough to Meet the Challenge?" In: *International Data Privacy Law* 7.1 (Feb. 2017), pp. 1–2. ISSN: 2044-3994. DOI: 10.1093/idpl/ipx003. (Visited on 09/09/2024).

[58] Michelle Kurilla. *What Is the CHIPS Act?* en. URL: https://www.cfr.org/in-brief/what-chips-act.

[59] Yue Li, Hongxia Wang, and Mauro Barni. "A Survey of Deep Neural Network Watermarking Techniques". In: *Neurocomputing* 461 (Oct. 2021), pp. 171–193. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2021.07.051. (Visited on 09/12/2024).

[60] Lin Long et al. *On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey*. June 2024. DOI: 10.48550/arXiv.2406.15126. arXiv: 2406.15126 [cs]. (Visited on 09/10/2024).

[61] H. Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: arXiv:1602.05629 (Jan. 2023). arXiv:1602.05629 [cs]. DOI: 10.48550/arXiv.1602.05629. URL: http://arxiv.org/abs/1602.05629.

[62] Celso M. de Melo et al. "Next-Generation Deep Learning Based on Simulators and Synthetic Data". In: *Trends in Cognitive Sciences* 26.2 (Feb. 2022), pp. 174–187. ISSN: 1364-6613, 1879-307X. DOI: 10.1016/j.tics.2021.11.008. (Visited on 09/10/2024).

[63] Margaret Mitchell et al. "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. 2019, pp. 220–229.

[64] Niklas Muennighoff et al. "Scaling data-constrained language models". In: *Advances in Neural Information Processing Systems* 36 (2024).

[65] Milad Nasr et al. *Scalable Extraction of Training Data from (Production) Language Models*. Nov. 2023. arXiv: 2311.17035 [cs]. (Visited on 09/09/2024).

[66] Milad Nasr et al. "Scalable extraction of training data from (production) language models". In: *arXiv preprint arXiv:2311.17035* (2023).

[67] Sella Nevo et al. *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. Tech. rep. RAND Corporation, May 2024. (Visited on 09/09/2024).

[68] Omolara Patricia Olaiya et al. “Encryption Techniques for Financial Data Security in Fintech Applications”. In: *International Journal of Science and Research Archive* 12.1 (2024), pp. 2942–2949. ISSN: 2582-8185, 2582-8185. DOI: 10.30574/ijsra.2024.12.1.1210. (Visited on 09/09/2024).

[69] OpenAI. “GPT-4 System Card”. In: (Mar. 2023). URL: https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[70] OpenAI et al. *GPT-4 Technical Report*. Mar. 2024. DOI: 10.48550/arXiv.2303.08774. arXiv: 2303.08774 [cs]. (Visited on 09/08/2024).

[71] Mary Phuong et al. *Evaluating Frontier Models for Dangerous Capabilities*. 2024. arXiv: 2403.13793 [cs.LG]. URL: https://arxiv.org/abs/2403.13793.

[72] David Plotkin. *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Academic Press, Oct. 2020. ISBN: 978-0-12-822167-9.

[73] Anka Reuel et al. *Open Problems in Technical AI Governance*. 2024. arXiv: 2407.14981 [cs.CY]. URL: https://arxiv.org/abs/2407.14981.

[74] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. “Deepsigns: A generic watermarking framework for ip protection of deep learning models”. In: *arXiv preprint arXiv:1804.00750* (2018).

[75] Girish Sastry et al. *Computing Power and the Governance of Artificial Intelligence*. Feb. 2024. DOI: 10.48550/arXiv.2402.08797. arXiv: 2402.08797 [cs]. (Visited on 09/12/2024).

[76] Mike Scarcella. “Nvidia, Microsoft Hit with Patent Lawsuit over AI Computing Technology”. In: *Reuters* (Sept. 2024). (Visited on 09/13/2024).

[77] Anna Schmidt and Michael Wiegand. “A Survey on Hate Speech Detection Using Natural Language Processing”. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics. 2017, pp. 1–10.

[78] Reza Shokri et al. “Membership Inference Attacks Against Machine Learning Models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18.

[79] Manli Shu et al. “On the exploitability of instruction tuning”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 61836–61856.

[80] Zhichuang Sun et al. “Mind Your Weight(s): A Large-scale Study on Insufficient Machine Learning Model Protection in Mobile Apps”. In: 30th USENIX Security Symposium (USENIX Security 21). 2021, pp. 1955–1972. ISBN: 978-1-939133-24-3. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/sun-zhichuang (visited on 09/09/2024).

[81] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL]. URL: https://arxiv.org/abs/2307.09288.

[82] Yuji Uchida et al. “Embedding Watermarks into Deep Neural Networks”. In: *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM. 2017, pp. 269–277.

[83] Praneeth Vadlapati. *AutoPureData: Automated Filtering of Web Data for LLM Fine-tuning*. June 2024. DOI: 10.48550/arXiv.2406.19271. arXiv: 2406.19271 [cs]. (Visited on 08/28/2024).

[84] Pablo Villalobos et al. *Will we run out of data? Limits of LLM scaling based on human-generated data*. 2024. arXiv: 2211.04325 [cs.LG]. URL: https://arxiv.org/abs/2211.04325.

[85] Eric Wallace et al. “Concealed Data Poisoning Attacks on NLP Models”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 2021, pp. 139–150.

[86] Eric Wallace et al. “Imitation Attacks and Defenses for Black-box Machine Translation Systems”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. 2020, pp. 4999–5013.

[87] H. James Wilson, Paul R. Daugherty, and Chase Davenport. “The Future of AI Will Be About Less Data, Not More”. In: *Harvard Business Review* (Jan. 2019). ISSN: 0017-8012. (Visited on 09/10/2024).

[88] Jing Xu, Yuan Ju, and Alexander I. Rudnicky. “Recipes for Safety in Open-domain Chatbots”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. 2020, pp. 2957–2970.

[89] Runhua Xu, Nathalie Baracaldo, and James Joshi. “Privacy-Preserving Machine Learning: Methods, Challenges and Directions”. In: *arXiv* arXiv:2108.04417 (Sept. 2021). arXiv:2108.04417 [cs]. DOI: 10.48550/arXiv.2108.04417. URL: http://arxiv.org/abs/2108.04417.

[90] Fuzhao Xue et al. “To repeat or not to repeat: Insights from scaling llm under token-crisis”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[91] Jing Zhang, Zhan Gu, and Licheng Jiao. “Protecting Intellectual Property of Deep Neural Networks with Watermarking”. In: *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*. ACM. 2018, pp. 159–172.

[92] Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. “Text revealer: Private text reconstruction via model inversion attacks against transformers”. In: *arXiv preprint arXiv:2209.10505* (2022).

[93] *Zoom addresses privacy concerns related to AI data colection*. en. Aug. 2023. URL: https://www.nbcnews.com/tech/innovation/zoom-ai-privacy-tos-terms-of-service-data-rcna98665.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's abstract and introduction outline challenges, 15 policy mechanisms and implementation details; we provide these within the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we do so by specifically stating the impact of our work

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There are no theoretical derivations or formal results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA] .

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA] .

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA] .

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: The paper does not include experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Because our work focused primarily on introducing a new paradigm for data governance, our key ethical considerations included to what extent our paper contributed to considering social responsibility, and our process adhered integrity and rigor expected.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Because we focused on safety for societal-scale risks, we hope that by outlining how data can grant policymakers with a new lever, this allows them to better shape the development of frontier AI regulation. In particular, we hope that this form of frontier data governance can prevent the acquisition of dangerous capabilities, including potential malicious or unintended uses.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not use human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We do not use human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.