Data Delineation for Foundation Models

Paymon Haddad **Sheriff Issaka** **Preetham Pangaluru** **Haniyeh Ehsani**

Abstract

Based on the Trust No Bot (Mireshghallah et al., 2024) presentation given earlier in the course, we advocate for a data delineation framework that distinguishes all data used in the context of foundation model training into 3 distinct categories: (1) no training, (2) internalized training, and (3) general training, with category 1 containing the most sensitive data, and category 3 containing the least sensitive data. We advocate that any data that is used to train or fine-tune machine learning systems must first be categorized into 1 of these 3 buckets in accordance with the terms of the regulations (which we define below in detail). In this perspective piece, we will (1) motivate the need for our proposed policy, (2) discuss the existing policy landscape pertaining to this area and how our policy fits into this landscape (3) provide a detailed account of the policy itself (what determines how data is binned into each of these categories), and finally (4) discuss recommendations for enforcing our proposed policy.

1 Motivation

Many of the existing debates and discussions surrounding imminent problems posed by machine learning systems, especially large-language models, have their root in a lack of data protection laws surrounding what data is legally allowed to be used in training these models. We argue that providing a clear set of guidelines surrounding what data is appropriate for machine learning training in certain contexts will help to solve many of these problems. Below we motivate our policy by discussing how it will help to address most of the most pressing concerns regarding the deployment of increasingly sophisticated machine learning systems in our society.

First, our policy will help to mitigate the perpetuation of harmful social and cultural biases that are often accompanied by the deployment of machine learning models in real-world contexts. Information that is deemed to be explicitly and/or implicitly perpetuating sources of societal and/or cultural bias are to be placed in the no training data bucket, which indicates that machine learning models by law cannot use data in this category at training time. (A detailed outline of our policy proposal is provided later in the discussion.) Although there are many existing models that have been previously trained on data that would fall squarely into the no training data bucket outlined by our policy, this policy would eliminate all future legal training on data that has even implicit biases present.

Second, our policy will help governments and organizations enforce existing privacy laws by helping them reconcile data privacy laws that were made before the LLM-era with this new landscape. For example, the GDPR, HIPAA, and CCPA outline a host of regulatory guidelines pertaining to the way personal data needs to be handled. Putting clearly defined guidelines around how this data can be used in machine learning training will enforce culpability upon developers and maintainers of these systems. Take the New York Times' suing of OpenAI over what they claim is unlawful use of their articles used in OpenAI's training corpus. Had our policy structure been imposed prior, there would be no room for ambiguity over this case, and the supposed violation would most likely have never even happened in the first place. This is because OpenAI will be unable to claim ignorance, since they were personally responsible for categorizing data into the buckets based on a clear set of guidelines.

Third, our policy will help protect intellectual privacy. There is an ongoing debate over the legality of model developers scraping the work of artists and other individuals who already exist in a climate which makes it very difficult for them to receive acknowledgment and/or compensation for their work. Our policy provides a clear frame-

work to delineate the extent to which this sort of data could be used for model training by default. By providing strict regulations around the use of copyrighted material in training corpora, this sets a strong framework to incentivize the development of proper compensation structures for rewarding artists and other inventors for their work, or giving them a strong legal basis to deny entry of their creations as part of train corpora.

Fourth, our policy will greatly improve public trust in machine learning systems developed in any corporate or government context. A natural extension of data-use transparency in training machine learning systems is that the public will be more likely to trust and support future development of these systems. This is true for a number of reasons. For example, protecting intellectual privacy can mean a strong AI ecosystem actually benefits artists rather than hurting them. Additionally, having strong guarantees over how data is used to train systems in a way that is known to the public makes it much easier for, for example, a healthcare patient to feel comfortable disclosing their anonymized data for use in training a domain specific system that would help disease diagnosis.

It's clear then that the ratification and implementation of our policy has the potential to provide far-reaching benefits in domains across the public and private sector, while helping to address many existing concerns surrounding machine learning deployment in the wild.

2 Related Work

Large Language Models (LLMs) present a double-edged sword in artificial intelligence. While they offer unprecedented capabilities in generating human-like text, they also pose significant privacy and security risks. These risks stem from several key factors:

Vast and Varied Training Data: LLMs are typically trained on enormous datasets scraped from the internet. These datasets may inadvertently contain sensitive or private information, including: Personally Identifiable Information (PII), confidential business data, private communications, and copyrighted material (Carlini et al., 2019) (Carlini et al., 2021) (Zhang et al., 2020)

Probabilistic Nature of LLMs: The fundamental design of LLMs as probabilistic models makes controlling their outputs extremely challenging. This unpredictability means that: (1) LLMs can unexpectedly reproduce sensitive information from their training data. (2) The model's responses may vary, even for identical inputs. (3) It's difficult to guarantee that private information won't be disclosed (Papernot et al., 2016).

Limitations of Traditional Safeguards: Classic content filtering methods, such as keyword blacklists, prove ineffective in preventing LLMs from revealing sensitive information. This is because LLMs can rephrase or paraphrase information in ways that bypass simple keyword filters (Krishna et al., 2021).

To address the privacy concerns associated with LLMs, several primary approaches have been proposed, including: differential privacy, and output filtering, and machine unlearning. Each method has its own unique challenges and limitations. Here's a brief overview of each:

Differential privacy: Differential privacy adds mathematical noise to the training process to prevent models from learning details about individual data points (Abadi et al., 2016) (Dwork and Roth, 2014). It provides theoretical guarantees against data extraction attacks and restricts the model from memorizing specific details (Tramèr et al., 2022) (Gehrmann et al., 2019). However, it degrades model performance in complex language tasks, needs high computational requirements for LLM-scale implementation, and is impractical for real-world scenarios (Gehrmann et al., 2019).

Output filtering: Output filtering involves inspecting and modifying the model's outputs before delivery to the user. This approach can potentially catch sensitive information before it reaches the end-user but it has difficulty in precisely identifying all occurrences of sensitive information (Henderson et al., 2018). It also results in high false positive rates and struggles with nuances in contextualization (Wallace et al., 2019). Another problem is that LLMs can output semantic variations that bypass filters and experienced users may find ways to circumvent filtering mechanisms (Nguyen et al., 2022).

Machine unlearning: Machine unlearning is a technique designed to selectively remove specific data from trained machine learning models. This post-training data removal approach addresses several key objectives: Enables the deletion of particular training data from a model after it has been trained. Ensures that users' personal information can be removed upon request. Supports compli-

ance with data protection regulations and privacy concerns (Huang et al., 2021). Nevertheless, Machine unlearning can be complex to implement and may lead to decreased model performance due to the removal of specific data points. Additionally, it can require significant computational resources and may not fully eliminate all traces of the unlearned data, raising concerns about residual data leakage.

3 Methods

3.1 Our Detailed Policy

Our data delineation framework divides all data for foundation model training into three distinct categories, each with clear guidelines for usage and implementation. These categories establish a structured approach to data sensitivity and provide organizations with actionable boundaries for responsible AI development.

No Train:. This category encompasses data that should never be used for model training under any circumstances. Such data typically contains highly sensitive personal identifiable information (PII) or content that reinforces or perpetuates harmful societal biases. Organizations must implement robust filtering systems to detect and exclude this data from any training pipeline. The protection of this data category is paramount as its inclusion could lead to severe privacy violations, legal repercussions, and reinforcement of societal inequities through AI systems. Implementation requires multi-layered screening protocols that can identify both explicit and implicit indicators of sensitive content. To ensure compliance, organizations should establish cross-functional oversight committees comprising legal, ethical, and technical experts to periodically audit their data filtering mechanisms. Examples include clinical data with patient identifiers, credit card information and transaction histories, personal health records and medical histories, and biased datasets containing explicit or implicit prejudice based on gender, race, religion, sexual orientation, disability status, or other protected characteristics. Further examples extend to confidential government documents, classified information, private communications without explicit consent such as emails and text messages, biometric data including facial recognition datasets, and children's personal information. Additionally, this category includes data under active litigation, information protected by attorney-client privilege, detailed geolocation histories that could reveal personal routines, trade secrets, proprietary manufacturing processes, and any content that violates existing data protection regulations such as GDPR, HIPAA, or CCPA.

Internalized Train: This category pertains to data that can be utilized in internal and highly controlled environments. Such data is essential for creating models that address complex problems requiring specialized information that cannot be easily determined through other means. For this category to apply, two key conditions must be met: the data must be necessary for solving a specific problem that benefits society, and the insights cannot be reasonably obtained through less sensitive data sources. Organizations employing this data must implement stringent access controls, robust encryption protocols, and comprehensive audit trails to track all interactions with the data throughout the training process. Additionally, they should establish clear documentation practices detailing the deidentification methods applied, risk assessments conducted, and justifications for using this data over less sensitive alternatives. Regular independent verification of these safeguards should be mandated to ensure ongoing compliance. For example, if someone's punishment after being found guilty of a crime can be relatively easily determined by humans through existing legal frameworks, then there is no need to use that data to train AI systems to assign these judgments. However, in an instance where a hospital is using deidentified clinical data for millions of patients to accurately train a model that predicts risk of cancer metastasis, this would fit in our internalized training data segment. Examples include deidentified health information used for medical research, aggregated financial data for fraud detection systems, internal company intellectual property, academic research data with restricted access, anonymized behavioral data for improving user experiences, deidentified legal case histories for legal analysis models, sanitized telecommunications data for network optimization, and proprietary business analytics and market insights.

General Train:. This category consists of general knowledge that contains as little bias as possible. Such data can be freely used for training AI models without additional restrictions. This category forms the foundation of most general-purpose AI systems and represents information that is broadly available, educational in nature, and presents minimal risk of harm when incorporated

into training sets. While this data has fewer restrictions, organizations should still maintain provenance tracking to ensure proper attribution and monitor for any emerging biases that may not have been initially apparent. Regular audits should evaluate whether data continues to meet the criteria for this category as societal norms and understanding of bias evolve. Organizations should implement feedback mechanisms to identify and address any unforeseen consequences resulting from the use of this data. Examples include Wikipedia articles and other encyclopedic content, scientific publications and reports, public domain literature and creative works, open government data and public records, technical documentation and educational resources, news articles from reputable sources with appropriate licensing, open-source code repositories and programming documentation, and publicly available datasets explicitly created for AI training. Additional examples encompass historical archives with appropriate context, standardized benchmark datasets widely accepted in the research community, publicly available language corpora, general knowledge question-answer pairs, open-access medical literature (as distinct from patient records), mathematical formulas and scientific principles, weather and climate data, transportation schedules, product manuals and documentation, and appropriately licensed creative commons media. This category serves as the ethical foundation for general-purpose AI systems, ensuring they can be developed with minimal risk while still accessing the breadth of human knowledge necessary for meaningful functionality.

3.2 Enforcement Protocols

No Train: When dealing with sensitive data that a foundation model should not be trained on, the data source should specify that certain regions of the text have personal identifiers or provide enough information to discern that this information can identify a person or a group of people. This can be done in the metadata of the website, an example of doing so would be to specify this in the robots.txt and/or to create a custom data-pii or data-sens attribute as a HTML data attribute to ensure that a region of the text is explicitly marked as having personal identifiers or being sensitive, respectively. Since this is on the end of the data provider, these are some precautions to be taken in case the provider thinks this would be sensitive data enough for training purposes. On the end of the organization that is training the foundation model there needs to be legal, ethical, and technical experts in a committee that would be overseeing the data filtering to ensure data integrity before the data is used to train the model. This data filtering done by the committee would ensure posthoc filtering of any personal identifying data and any data that would perpetuate harmful societal biases. This committee is to ensure that if individual sources do not specify personal identifying information in the source or mark certain information as being sensitive or perpetuating harmful societal biases this data does not end up as training data for the foundation model. As an example of the above, if there is a politically charged article that has been scraped to be used as training data the author may not necessarily consider this as a harmful societal bias or at least label it as one, however, the committee can then filter this data out before it ends up being trained on. If the organization training the foundation model fails to have an oversight committee that pledges to filter data after scraping from sources or fails to do a proper job of doing so there will be a fine for violation as a way to make sure organizations training foundation models follow through on filtering the data properly.

Internalized Train: For sensitive information that is deidentified and used internally there must be stringent access controls, fail-proof encryption, and auditing done to ensure that the data does not leak beyond the organization and parts of the organization that are allowed access to it. Furthermore, as this includes the case of having identifiers that link the data to a particular person that is internal it still needs to be filtered to ensure that personal identifiers and any implicit identifier that can link a person to the data is discarded. Since this kind of information is still sensitive information there needs to be auditing with clear documentation done to ensure that this information is absolutely necessary and that there is no other form of information that is less sensitive that can be used. Certain data can be deemed sensitive enough that other parts of the organization cannot access (hence the robust encryption and access controls), and therefore, there needs to be a couple of individuals still monitoring that the explicit and implicit personal identifiers are removed. This is especially vital in a hospital setting or in a legal setting where personal identifiers are usually removed but having them by accident could result in a lot of harm. In addition to this, since all the data sources would be internal it is

imperative that they be marked where sensitive information lies, if digital, there should be HTML attributes similar to how data sources for no training (sensitive data) are suggested to mark their data but in this setting it should be enforced internally. As this is internal this is a suggestion to the organization to do so to not leak personal identification information and also to control the spread of data beyond those who should be able to access it. After these filters and oversight this data can then be used to train a model to help an organization or certain parts of it.

General Train: Although the least sensitive data might need less filtering than the other data categories we discussed, it still needs to be audited to ensure that the organization training the foundation model knows where the data is coming from. Tracking the provenance of the data is especially paramount in order to ensure that it comes from a trustworthy data source, since this may lead to downstream factual inaccuracies and misinformation when doing inference if the foundation model is trained on such data. Proper attribution to and crediting data sources should also be done if the data is gathered from external sources because the organization is using another source's data. Furthermore, the organization needs to have a committee of legal, ethical, and technical experts who can determine if certain parts of the scraped data exemplify harmful social biases. This becomes imperative when we consider societal evolution because with it biases evolve and can be a reason to periodically audit the incoming data from data sources. Taking into account that harmful societal biases may not always be filtered out due to training a model for content moderation purposes the committee must also determine when it is necessary to filter such biased data. These factors should all be taken into consideration when determining when or when not to include biased data. This category of less sensitive data also includes harmful societal biases because the other category is excluding this data to be trained on entirely, and since there are instances like content moderation this was worth mentioning. In the case of data being used that is no longer socially acceptable the committee must also monitor this data to ensure that evolving social biases are in check with the data. If this data has already been used for training a foundation model this would be referring to future training and the enforcement of this policy would not apply retroactively.

4 Conclusion

In summary, our data delineation framework provides a set of unequivocal guidelines that can be used to bucket arbitrary data into 1 of 3 buckets that defines its suitability for use in training a machine learning model under a particular jurisdiction. Our policy allows the engineers and researchers in charge of creating these systems to maintain full autonomy over development while establishing an important precedent for responsible government oversight over the industry, which we assert will have far-reaching effects for ensuring (1) fairness (2) data isolation, and (3) increased innovation across many domains. Our advocacy leaves some open questions concerning the nuances of policy implementation. For example, future work may explore what body should serve as the arbiter for enforcement of the framework, or what a new organization to serve as the arbiter would look like. Should arbitration be settled by existing judiciary committees? Should it instead be settled by a new organization all-together? Should this organization be publicly controlled? How can we ensure that control over arbitration does not become a partisan issue? All of these are important real-world considerations to make when implementing our policy.

5 Embedded Ethics Discussion

We would design a module to incentivize students to consider the implications of data control policies that specifically address machine learning training. Our module would be outlined in the following way chronologically: (1) lecture on some of the injustices and breaches in privacy that have resulted from machine learning systems historically due to a lack of such policies, (2) lecture on the existing data privacy landscape and how it fails to reconcile itself with recent developments in machine learning, (3) have an assignment that involves students drafting a clearly defined policy and implementation framework for said privacy that addresses these concerns (much like we did in our methods section for our policy), and lastly (4) model the classroom as a sort of "congress" in which each student advocates for their policy framework in terms of how it addresses the issues taught in the lectures, and ultimately have the class decide on one policy to implement (that final policy may incorporate aspects of multiple suggestions provided by different students' policies).

6 Contribution Statement

Paymon Haddad: Did the sections: **Proposition, Motivation, Conclusion, Embedded Ethics Discussion**

Sheriff Issaka: Did the section: **Methods: Our Policy Detailed**

Preetham Pangaluru: Did the section: **Methods: Enforcement Protocols**

Haniyeh Ehsani: Did the section: **Related Work**

References

M. Abadi, A. Chu, I. Goodfellow, and 1 others. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.

N. Carlini, C. Liu, U. Erlingsson, J. Kos, and D. Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, pages 267–284.

N. Carlini, F. Tramer, E. Wallace, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium*, pages 2633–2650.

C. Dwork and A. Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407.

S. Gehrmann, H. Strobelt, and A. M. Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

P. Henderson and 1 others. 2018. Ethical challenges in data-driven dialogue systems. *arXiv preprint*.

H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang. 2021. Unlearnable examples: Making personal data unexploitable. In *International Conference on Learning Representations*.

K. Krishna, J. Wieting, D. Ippolito, and T. Berg-Kirkpatrick. 2021. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 737–762.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint*, arXiv:2407.11438.

T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. 2022. A survey of machine unlearning. *arXiv preprint*.

N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy*, pages 582–597.

F. Tramèr, N. Carlini, W. Brendel, and A. Madry. 2022. Accuracy first: Selecting a differential privacy level for production machine learning. In *Proceedings on Privacy Enhancing Technologies*, volume 2022, pages 94–110.

E. Wallace, S. Feng, N. Kandpal, S. Singh, and M. Gardner. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2153–2162.

L. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. 2020. Deep mutual learning. *Pattern Recognition*, 100:107173.