The Disclosure Delusion: Systemic Challenges in AI Data Transparency Policy

Judy Hanwen Shen [1] **Ken Liu** [1] **Angelina Wang** [1] **Sarah H. Cen** [1] **Andy K. Zhang** [1] **Caroline Meinhardt** [1] **Daniel Zhang** [1] **Kevin Klyman** [1] **Rishi Bommasani** [1] **Daniel E. Ho** [1]

Abstract

Data transparency has emerged as a central rallying cry to address concerns with Generative AI, including data quality, privacy, and copyright. While commentators have called for "Nutrition Facts" for AI, such calls have exhibited limited understanding of the social science and institutional implementation of disclosure. We offer such an institutional perspective and identify three fallacies in calls for data disclosures for AI. First, many proposals exhibit a *disclosure gap* between the stated goals of data transparency and the actual disclosures necessary to achieve such goals. Second, reform attempts exhibit a *remediation gap* between required disclosures on paper and enforcement to ensure compliance in fact. Third, policy proposals manifest an *outcome gap* between disclosure and behavior by AI developers. Informed by the social science on transparency, our analysis identifies affirmative paths for transparency that matters.

1. Introduction

Transparency for Generative AI has recently emerged as a critical goal for policymakers as a mechanism for accountability (Bommasani et al., 2023). Similar to nutrition labels, policymakers want AI data transparency to provide clear, digestible information about AI systems. For example, California's AB 2013 requires developers to publicly post high-level summaries of datasets used in the development of generative AI systems or services (California State Legislature). The EU AI Act similarly requires developers of general purpose AI models to disclose a data summary, including the types of data and whether any data is protected by copyright (European Parliament and Council). These requirements arose in part due to repeated calls from the academic community for data disclosures as nutrition facts for AI (Holland et al., 2018; Chmielinski et al., 2024). However, like nutrition facts for food, these transparency efforts face their own set of policy challenges. Through an institutional perspective, we illustrate the critical misalignment that causes current policies to fall short of delivering their intended goals by identifying three fundamental gaps.

The first fallacy is the *disclosure gap* between the stated transparency goals, if they are even specified, and the necessary disclosures to achieve these goals. Although issues with the broad alignment of AI regulatory standards have been highlighted (Guha et al., 2024), the disconnect in data transparency policy, specifically between ideals and actual implementation, is pronounced. Current transparency requirements hint at several protections, yet the defined disclosures fall short of these aims, a topic seldom explored in the literature. Our work systematically identifies current and prospective transparency objectives and aligns them with minimum disclosure requirements.

The second fallacy is the *remediation gap* that exists between what companies are required to disclose and the mechanisms that ensure that the disclosures are made accurately. Given the invocation of nutrition labeling for AI policy, we illustrate these enforcement challenges in US nutritional labeling itself. The transparency efforts in food regulation illustrate the many challenges that arise when considering the actors, incentives, power, and resources required to ensure meaningful transparency. These challenges are particularly acute for AI, where technical complexity further complicates verification efforts.

The third fallacy is the *outcome gap* between the data transparency compliance behavior and the actions that are actually needed to protect individuals, data creators, other stakeholders. Many works have highlighted the failures of mandatory disclosures (Fung et al., 2007; Ben-Shahar & Schneider, 2017) and we identify how some of these failure modes also translate to AI transparency policy.

In this abridged version of our paper, we briefly discuss the levels of data transparency and the disclosure gap. Please see our full paper for a full discussion of the disclosure gap, remediation gap, outcome gap, AB2013 case study, and technical research agenda in the supplementary materials.

[1]Stanford University, Palo Alto, USA. Correspondence to: Judy Hanwen Shen <jhshen@stanford.edu>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

2. Levels of Data Transparency

Despite data transparency being a frequent goal in AI policy, transparency can refer to many different things. In this section, we taxonomize different forms of data transparency, which we separate into (1) documentation and descriptions and (2) data access. This taxonomy is based on the artifact the disclosure process produces; using this taxonomy allows a deeper analysis of the burden of providing these disclosures and the technical feasibility of validating these disclosures.

2.1. Level 1: Documentation and Descriptions

This level of transparency includes various information *surrounding* the datasets used in the development of generative AI models. We further separate data information into the following two categories: information about the dataset itself (e.g., how data is processed and used) and information about the dataset supply chain (e.g., how data is acquired). This level of disclosure produces an artifact that is a documentation of the data.

Level 1a: Dataset Information Dataset information is metadata on the content and construction of the dataset. Some examples, along with those for which existing regulations mandate them, include

- Summary Statistics (AB 2013): High-level dataset statistics, such as the number of training points.
- Data Sources (AB 2013): Sources (e.g. web, books, video) from which training data was gathered and what purpose each source serves.
- Synthetic Data (AB 2013): Description of whether and how synthetic data was used in the training data, and how the synthetic data was generated.
- Personal Information (AB 2013, EU AI Act): Whether the datasets used contain personally identifiable information or aggregate consumer information.
- Copyrighted Content (AB 2013): Whether the datasets contain copyright, trademarked, or patented information or whether the datasets are in the public domain.
- Data Processing (GDPR, AB 2013, EU AI Act): Cleaning, filtering, removal of PII and other processing steps taken before the data were used for model training.

Level 1b: Data Supply Chain Information We consider the data supply chain broadly to include many different types of data, including public datasets, publicly available data, licensed data, human-generated data and synthetically generated data. Here, data supply chain refers to the process and network by which data used for AI is produced (Hopkins et al., 2025; Lee et al., 2023; Widder & Nafus, 2023; Cobbe et al., 2023). This level of information around the data-generating process of the datasets used involves both the data producers and the model developers. Although existing regulations focus on system developers and model providers, it is likely that these parties know or control aspects of the supply chain when they work with data producers. In this type of disclosure, documentation and descriptions include the following:

- Data Collection (AB 2013, EU AI Act): Information about when, how, and from whom the dataset was collected.
- Consent (GDPR): If training data includes personal data, whether and how consent was collected.
- Licensing (AB 2013): Information about whether the dataset was purchased or licensed; permissions around dataset usage.
- Vendors: Information on which data vendors were used in the dataset collection, labeling, and processing.
- Contracts: Contracts between data providers and model developers that allow access to data for training and evaluation.
- Data workers: Details about who labeled and collected the data and how they were compensated or attributed.
- Data Provenance: Information on how the dataset is derived including sourcing, creation, and licensing.[1]

2.2. Level 2: Data access

This second level of transparency involves direct access to training data in some form. Although not traditionally considered a *disclosure*, access to training data can be required to achieve certain policy goals, despite being costly or in contradiction to the business interests of model developers. For models that disclose training data, if all data used in the training process are publicly available (Groeneveld et al., 2024), releasing sufficient dataset statistics is equivalent to having full access to the training dataset. For example, if a model was trained only on the Pile (Gao et al., 2020), a publicly available dataset, specifying the training data information, is sufficient to access the entire dataset. For proprietary models, datasets are often secret, and dataset descriptions are not equivalent to direct access to datasets. When data contain personal information, access to anonymized data also falls under this category of disclosure,

[1]There is overlap with the prior bullet points, but the focus is on understanding the original sources which training datasets stem from (Longpre et al., 2023).

since anonymized data can be included as a subset or as a full dataset. Access to a dataset itself is not binary; we discuss three forms: membership access, subset access, and full data access.

Level 2a: Membership Access

- Membership query access to training data: Given a document, such as a webpage, a book, or a code repository, a membership query access returns a yes or no answer about whether the piece of data is included in the dataset.

Level 2b: Subset Access

- Sample access to training dataset subset: To illustrate the format of the data used in different phases of training, some samples can be provided on the type of data used for each phase of training.
- Partial access to training data or the generation of training data (e.g., public datasets, prompts) (EU DSA): Portions of the dataset, transformations of public datasets, or instructions for generation of datasets (e.g., prompts) provide partial access to training data.

Level 2c: Full Data Access

- Full data access: The entire dataset used for one or more components of training, validation, and testing of the model before deployment is made available.

Methods of verifying that the data provided in levels 2b and 2c were indeed used in the training set (recall) include comparing the likelihood of the training data vs data not included in training (Shokri et al., 2016).

3. The Disclosure Gap

Our goal is to map each data objective to the requisite levels of data transparency that would enable the objective. In doing so, we can map how well each objective is supported by the actual form of data transparency requested. For example, if the goal is to assess copyright infringement, what is the actual form of transparency that allows this assessment? Table 2 provides a summary of the fine-grained subobjectives and the necessary disclosures and the corresponding verification requirements.[2]

3.1. Protecting Personal Information

When regulation mandates data transparency around personal information, the aim is often to protect individuals and their data. Personal information may include name, image, and government identification numbers. This type of data transparency would empower consumers to make informed decisions around AI products and services that use personal data for training. Data transparency alone is not enough to protect these rights since downstream actions based on privacy law is required. However, data transparency provides the necessary first step of disclosure of when personal information is used.

Necessary Disclosures Although the goal of policy in this area is to protect users if their data are collected for use in training of models, it is unclear what types of disclosure are necessary or sufficient to guarantee these supposed protections. Different stakeholders may want to protect personal information in different scenarios. Although some objectives such as consumer choice of products that do not use PII is possible with level 1 disclosures, other objectives such as testing for leakage of PII are not possible through data disclosures but only possible through model disclosures (Table 1).

3.2. Assurance of Training Data Quality

Existing regulation also has mandates for training data quality which are driven by the belief that data quality would reduce harm or discrimination experienced by downstream users. In the computer science literature, the importance of data quality has been highlighted (Xu et al., 2021; Longpre et al., 2024b; Wettig et al., 2024); sometimes as a factor even more important than the size of the dataset (Zhou et al., 2023; Shen et al., 2024).

Necessary Disclosures For regulators and individual users interested in ensuring data quality and understanding the representativeness of the training dataset, level 1 disclosures are sufficient since the inclusion of various data sources can be reported through these disclosures. If an auditor wanted to evaluate the degree to which the training data is representative of a customer segment or society more generally, full data access (2c) would be necessary. Thus, the validation of training data quality often requires levels of data transparency existing policy does not mandate.

While we discuss how to achieve assurances of data quality, in some cases, the actual goal might be to ensure the model that the datasets produce does not exhibit undesirable behaviors (e.g., biases, dangerous answers). In these cases, it is important to consider that decisions during model training can be made to mitigate bad behavior even if the training dataset has limitations. For example, even if a dataset does not reflect the distribution of the population using the downstream product, data points representing minority views or preferences can be up-weighted to mitigate the downstream biases of a model. Thus, data transparency is not an effec-

[2]The corresponding data disclosure levels are described in Section C.

SUBOBJECTIVE	STAKEHOLDER	MINIMUM DISCLOSURE	VERIFICATION
Protecting Personal Information			
Understand the presence of PII of citizens in training data	Regulator	(1a) Dataset Information: Personal Information (1b) Data Supply Chain Information: Data Collection	Regulators may need access to the training dataset to verify or rely on complaints of models revealing personal information.
Choose AI products and services that do not use personal data	Individual User	(1a) Dataset Information: Personal Information	Individual users do not have the tools to test comprehensively whether personal data has been used
Discern whether their name and likeness were included in the training data for a specific mode	Individual User	(2a) Membership Access: Membership query access to training data	Users can test whether their information can be revealed by the model, but cannot verify that their personal data was not used at all (Cooper & Grimmelmann, 2024).
Test for the leakage of PII	Auditor	N/A - Leakage must be tested through model access	
Assurance of Training Data Quality			
Assurance of diverse data collection procedures	Regulator	(1a): Dataset Information: Data Sources (1b): Data Supply Chain Information: Data Collection	Regulators can verify that data sources are diverse by higher levels of data access (e.g., 2a membership access)
Assurance of data quality	Regulator	(1a): Dataset Information: Data Processing (1b): Data Supply Chain Information: Data Collection	Regulators can verify that data processing has been done to improve quality by inspecting source code for data.
Interpret whether predictions can be trusted for their specific individual profile	Individual User	(1a): Dataset Information: Data Sources (1b): Data Supply Chain Information: Data Collection	Individual users could verify through higher levels of data access (e.g., 2b subset access)
Assurance of Data Representativeness	Auditor	(2c): Full Data Access	
Copyright and Terms of Use Protections			
Assurance of copyright protection	Data Creator	(2a) Membership Access: Membership query access to training data	Verifying correctness of membership query responses relies on access to the actual training data.
Assurance of copyright and licensing law compliance	Regulator	(1b) Data Supply Chain - Licensing	It is difficult for regulators to verify that all data used is licensed. Regulators may need to rely on complaints from data owners.
Compliance with terms of use	Model Provider	(1b) Data Supply Chain - Data Collection	Model developers may verify that competitors did not train on data generated by their models through watermarking their outputs. (Kirchenbauer et al., 2023)
Evaluation Generalization			
Assurance of no train-test overlap	Regulator	(1a) Dataset Information: Data Processing	While it is impossible to verify the entire data processing pipeline, it might be possible to identify significant omissions through model behavior (Golchin & Surdeanu, 2023; Shi et al., 2023).
Check for the presence of evaluation examples in the training data	Auditor	(2a) Membership Access: Membership query access to training data	Verifying that membership query responses requires access to training data. However, it may be possible to observe behavior on benchmarks to infer potential contamination (Zhang et al., 2024).
Data Laborer Protections			
Choose AI platforms and services that are produced via fair compensation	Consumer	(1b) Data Supply Chain Information: Data workers	Currently consumers cannot verify that AI platforms fairly compensated data workers.
Check that forced or child labor is not a part of generating data	Auditor	(1b) Data Supply Chain Information: Data workers	Regulators may use a complaint system to censure companies that engage in labor practices they are not reporting.

Table 1. Overview of mapping between the objectives of data transparency and the minimum level of *data* disclosure. Different disclosures suffer from different challenges in verification.

tive tool since a representative dataset is neither sufficient nor necessary to achieve an unbiased downstream model.

3.3. Copyright and Terms of Use Protections

As generative AI become increasingly capable of producing high-quality creative and editorial content, concerns have been raised about whether the data used to train these models contain copyrighted materials. Copyright is assigned from 'the moment' a piece is created (United States Congress, 1976; WIPO, 1886), and thus the massive scale of datasets of books and images that are used to train generative models often include copyrighted material (Bandy & Vincent, 2021; Karamolegkou et al., 2023). However, the enforcement of copyright protections often falls in the hands of the creators. As a result, supply chain transparency (Lee et al., 2023), including filtering of training data to reduce the likelihood of copyright infringement (Henderson et al., 2023), has been suggested. Our analysis of data transparency disclosure requirements is again only a necessary but not sufficient step for copyright protections.

Necessary Disclosures While disclosing whether copyright data was used in training data seems low-cost at a high level, these very same disclosures are not useful for creators because they are too vague. Even if companies claim that no copyrighted data was used, it would be difficult to make these claims with certainty for every piece of content with a copyright. The more effective level of disclosure would be to provide access to the data creator in the membership access (2a) (Table 1). At the data collection and curation step, upstream in model training, revealing data sources and their associated licensing may be helpful as an additional incentive to protect copyright.

4. Recommendations

Our analysis highlights that (1) mandated disclosures should provide enough information to be actionable for consumers, data creators, auditors, and regulators, (2) enforcement mechanisms for compliance and verification of these disclosures are crucial, and (3) disclosures should be designed to change the behavior of AI system developers in a way that is aligned with intended policy goals.[3] Our analysis specifically illustrates the importance of the following:

- **Clarity**: Disclosures aimed at addressing too many things may cause confusion. It is more effective to focus on a clear goal and ensure that the transparency measures that correspond to this goal have sufficient enforcement mechanisms.
- **Standardization**: Successful implementation depends on standardized reporting. For example, a common format and a website where all companies that are subject to disclosure requirements can submit information.
- **Information Intermediation**: When public resources are limited, policy should be designed to empower intermediaries, such as private litigants or third parties, capable of conducting informative analysis on disclosed data.

4.1. Research Directions for Computer Scientists for Training Data Disclosure

We repeatedly highlight the importance of verification as a part of the enforcement of data transparency. An essential ingredient for verification is the continued development of empirical techniques for data verification. This section briefly discusses research directions related to both enabling and verifying training data disclosure – this research would be instrumental in understanding which kinds of policy and regulations are actually enforceable in practice.[4]

Granularity of data usage A key challenge in data disclosure is deciding the granularity with which data membership should be reported. Granularity refers to the specificity level at which data usage is tracked and disclosed (Maini et al., 2024). At the coarsest granularity, developers might simply acknowledge that a dataset was used; finer granularity could involve tracking specific documents, paragraphs, or even exact text sequences. More granular disclosure introduces computational overhead but yields more meaningful transparency - for example, sequence-level reporting might require precise but potentially large lookup tables or n gram statistics.

Revisiting definitions for training set inclusion A more fundamental issue with training data disclosure is that data membership is an inherently fuzzy concept. Practitioners typically adopt lossy definitions and tests to operationalize fuzzy concepts, which inevitably leaves room for ambiguity or even malpractice in data transparency.

5. Conclusion

Advanced AI is only made possible through learning from an incredibly large collection of data. If we want true accountability of these AI systems of copyright protections, data quality assurances, and privacy protections, data transparency policies need to be designed to ensure that the necessary disclosures and enforcement mechanisms are well-specified and tractable. In the absence of these properties, data transparency may remain only a fallacy.

[3] See our full analysis of the remediation gap in Section E and enforcement gap in Section F.

[4] We include two other directions: **Canary injection and data watermarks**, and **cryptographic approaches for data usage** in Section H.2.

References

Abbaszadeh, K., Pappas, C., Katz, J., and Papadopoulos, D. Zero-knowledge proofs of training for deep neural networks. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 4316–4330, 2024.

Aleccia, J. New label law has unintended effect: Sesame in more foods — apnews.com. https://apnews.com/article/sesame-allergies-label-b28f8eb3dc846f2a19d87b03440848f1, 2022. [Accessed 05-05-2025].

Ananny, M. and Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20 (3):973–989, 2018.

Bandy, J. and Vincent, N. Addressing" documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Ben-Shahar, O. and Schneider, C. E. The failure of mandated disclosure. *Russian Journal of Economics and Law*, (4 (44)):146–169, 2017.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023.

California State Legislature. Assembly bill no. 2013.

Chmielinski, K., Newman, S., Kranzinger, C. N., Hind, M., Vaughan, J. W., Mitchell, M., Stoyanovich, J., McMillan-Major, A., McReynolds, E., Esfahany, K., et al. The clear documentation framework for ai transparency: Recommendations for practitioners & context for policymakers. *Harvard Kennedy School Shorenstein Center discussion paper*, 2024.

Cobbe, J., Veale, M., and Singh, J. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1186–1197, 2023.

Cooper, A. F. and Grimmelmann, J. The files are in the computer: Copyright, memorization, and generative ai. *arXiv preprint arXiv:2404.12590*, 2024.

Cooper, A. F., Choquette-Choo, C. A., Bogen, M., Jagielski, M., Filippova, K., Liu, K. Z., Chouldechova, A., Hayes, J., Huang, Y., Mireshghallah, N., et al. Machine unlearning doesn't do what you think: Lessons for generative ai policy, research, and practice. *arXiv preprint arXiv:2412.06966*, 2024.

Crosbie, E., Alvarez, M. G. O., Cao, M., Renteria, L. S. V., Rodriguez, E., Flota, A. L., and Carriedo, A. Implementing front-of-pack nutrition warning labels in mexico: important lessons for low-and middle-income countries. *Public Health Nutrition*, 26(10):2149–2161, 2023.

Daily, M. N. Profeco withdraws thousands of products for faulty labeling — mexiconewsdaily.com. https://mexiconewsdaily.com/news/profeco-withdraws-products-faulty-labeling/. [Accessed 02-05-2025].

European Parliament and Council. Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act). URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

Fang, C., Jia, H., Thudi, A., Yaghini, M., Choquette-Choo, C. A., Dullerud, N., Chandrasekaran, V., and Papernot, N. Proof-of-learning is currently more broken than you think. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 797–816. IEEE, 2023.

Federal Trade Commission. Pom wonderful llc et al. F.T.C. File No. 082-3122 Docket No. 9344, Federal Trade Commission, January 2013.

Fung, A., Graham, M., and Weil, D. *Full disclosure: The perils and promise of transparency*. Cambridge University Press, 2007.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., Daumé III, H., and Crawford, K. Datasheets for datasets. corr abs/1803.09010 (2018). *arXiv preprint arXiv:1803.09010*, 2018.

Golchin, S. and Surdeanu, M. Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*, 2023.

Google. An expanded partnership with reddit, February 2024. URL https://blog.google/inside-google/company-announcements/expanded-reddit-partnership/. [Online; accessed May 4, 2025].

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Guha, N., Lawrence, C. M., Gailmard, L. A., Rodolfa, K. T., Surani, F., Bommasani, R., Raji, I. D., Cuéllar, M.-F., Honigsberg, C., Liang, P., et al. Ai regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *Geo. Wash. L. Rev.*, 92:1473, 2024.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Heinzerling, L. The varieties and limits of transparency in us food law. *Food and Drug Law Journal*, 70(1):11–24, 2015.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023.

Ho, D. E. Fudging the nudge: Information disclosure and restaurant grading. *Yale LJ*, 122:574, 2012.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The dataset nutrition label: A framework to drive higher data quality standards.(2018). arxiv. *arXiv preprint arXiv:1805.03677*, 2018.

Hopkins, A., Cen, S. H., Ilyas, A., Struckman, I., Videgaray, L., and Madry, A. Ai supply chains: An emerging ecosystem of ai actors, products, and services. *arXiv preprint arXiv:2504.20185*, 2025.

Intel. Intel® software guard extensions (intel® sgx) overview. Intel Developer Tools, 2025. URL https://www.intel.com/content/www/us/en/developer/tools/software-guard-extensions/overview.html?utm_source=chatgpt.com. Accessed: April 30, 2025.

Jia, H., Yaghini, M., Choquette-Choo, C. A., Dullerud, N., Thudi, A., Chandrasekaran, V., and Papernot, N. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1039–1056. IEEE, 2021.

Jurowetzki, R., Hain, D. S., Wirtz, K., and Bianchini, S. The private sector is hoarding ai researchers: what implications for science? *AI & SOCIETY*, pp. 1–8, 2025.

Kapoor, S. and Narayanan, A. Leakage and the reproducibility crisis in ml-based science. *arXiv preprint arXiv:2207.07048*, 2022.

Karamolegkou, A., Li, J., Zhou, L., and Søgaard, A. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*, 2023.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In *International Conference on Machine Learning*, pp. 17061–17084. PMLR, 2023.

Lee, K., Cooper, A. F., and Grimmelmann, J. Talkin"bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*, 2023.

Lemley, M. A. and Henderson, P. The mirage of artificial intelligence terms of use restrictions. *Available at SSRN*, 2024.

Liu, K. Z., Choquette-Choo, C. A., Jagielski, M., Kairouz, P., Koyejo, S., Liang, P., and Papernot, N. Language models may verbatim complete text they were not explicitly trained on. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. 2023.

Longpre, S., Mahari, R., Obeng-Marnu, N., Brannon, W., South, T., Kabbara, J., and Pentland, S. Data authenticity, consent, and provenance for ai are all broken: What will it take to fix them? 2024a.

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno, D., et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3245–3276, 2024b.

Luthi, S. Functionally useless: California privacy law's big reveal falls short. *Politico*, August 2021. URL https://www.politico.com/states/california/story/2021/08/05/functionally-useless-california-privacy-laws-big- Accessed May 23, 2025.

Maini, P., Jia, H., Papernot, N., and Dziedzic, A. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092, 2024.

Montti, R. Openai secretly funded benchmarking dataset linked to o3 model, Jan 2025. URL https://www.searchenginejournal.com/openai-secretly-funded-frontiermath-benchmarking-dataset/537760/. [Online; accessed May 4, 2025].

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Partnership on AI. Protecting AI's essential workers: Introducing our vendor engagement guidance & transparency template, August 2024. URL https://partnershiponai.org/protecting-ais-essential-workers-introducing-our-vendor-engagement-guidance-transparency-te [Online; accessed May 4, 2025].

Perrigo, B. Exclusive: The $2 per hour workers who made ChatGPT safer. *TIME*, January 2023. URL https://time.com/6247678/openai-chatgpt-kenya-workers/. [Online; accessed May 4, 2025].

POM Wonderful LLC v. Coca-Cola Co. Pom wonderful llc v. coca-cola co., 2014. S. Ct.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., et al. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*, 2024.

Sander, T., Fernandez, P., Durmus, A., Douze, M., and Furon, T. Watermarking makes language models radioactive. *Advances in Neural Information Processing Systems*, 37:21079–21113, 2024.

Shen, J. H., Raji, I. D., and Chen, I. Y. The data addition dilemma. *arXiv preprint arXiv:2408.04154*, 2024.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. S. Detecting pretraining data from large language models. *ArXiv*, abs/2310.16789, 2023. URL https://api.semanticscholar.org/CorpusID:264451585.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models,(2016). *Available: arXiv*, 1610, 2016.

Stark, D. P., Choplin, J. M., and LeBoeuf, M. A. Ineffective in any form: how confirmation bias and distractions undermine improved home-loan disclosures. *Yale LJF*, 122: 377, 2012.

Taylor, M. How the FDA is picking its food label battles. *The Atlantic*, July 2010. URL https://www.theatlantic.com/health/archive/2010/07/how-the-fda-is-picking-its-food-label-battles/59927/. Accessed May 23, 2025.

Tremblay v. OpenAI 2023. Tremblay et al v. OpenAI, inc. et al. Docket No. 3:23-cv-03223. URL https://www.bloomberglaw.com/public/desktop/document/TremblayetalvOpenAIIncetalDocketNo323cv03223NDCal14?doc_id=X39LOQS37QF8OCO6BDEB1UG0PUE. Bloomberg Law, accessed April 30, 2025.

United States Congress. 17 u.s. code § 102 - subject matter of copyright: In general. U.S. Code, 1976. URL https://www.law.cornell.edu/uscode/text/17/102. As amended through December 2024.

Wang, A., Datta, T., and Dickerson, J. P. Strategies for increasing corporate responsible ai prioritization. *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2024.

Wartella, E., Lichtenstein, A., and Boon, C. History of nutrition labelling. *Front-of-package nutrition rating systems and symbols: Phase 1 report*, 2, 2010.

Wei, J. T.-Z., Wang, R. Y., and Jia, R. Proving membership in llm pretraining data via data watermarks. *arXiv preprint arXiv:2402.10892*, 2024.

Wettig, A., Gupta, A., Malik, S., and Chen, D. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*, 2024.

Widder, D. G. and Nafus, D. Dislocated accountabilities in the "ai supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1): 20539517231177620, 2023.

WIPO. Berne convention for the protection of literary and artistic works. International Treaty, 1886. URL https://www.law.cornell.edu/treaties/berne/5.html. Paris Text 1971, as amended through September 28, 1979.

Worledge, T., Shen, J. H., Meister, N., Winston, C., and Guestrin, C. Unifying corroborative and contributive attributions in large language models. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 665–683. IEEE, 2024.

Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M., and Klein, D. Detoxifying language models risks marginalizing minority voices. *arXiv preprint arXiv:2104.06390*, 2021.

Zama. TFHE-rs: A Pure Rust Implementation of the TFHE Scheme for Boolean and Integer Arithmetics Over Encrypted Data, 2022. https://github.com/zama-ai/tfhe-rs.

Zhang, A. K., Klyman, K., Mai, Y., Levine, Y., Zhang, Y., Bommasani, R., and Liang, P. Language model developers should report train-test overlap. *arXiv preprint arXiv:2410.08385*, 2024.

Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

A. Introduction

Transparency for Generative AI has recently emerged as a critical goal for policymakers as a mechanism for accountability (Bommasani et al., 2023). Similar to nutrition labels, policymakers want AI data transparency to provide clear, digestible information about AI systems. For example, California's AB 2013 requires developers to publicly post high-level summaries of datasets used in the development of generative AI systems or services (California State Legislature). The EU AI Act similarly requires developers of general purpose AI models to disclose a data summary, including the types of data and whether any data is protected by copyright (European Parliament and Council). These requirements arose in part due to repeated calls from the academic community for data disclosures as nutrition facts for AI (Holland et al., 2018; Chmielinski et al., 2024). However, like nutrition facts for food, these transparency efforts face their own set of policy challenges. Through an institutional perspective, we illustrate the critical misalignment that causes current policies to fall short of delivering their intended goals by identifying three fundamental gaps.

The first fallacy is the *disclosure gap* between the stated transparency goals, if they are even specified, and the necessary disclosures to achieve these goals. Although issues with the broad alignment of AI regulatory standards have been highlighted (Guha et al., 2024), the disconnect in data transparency policy, specifically between ideals and actual implementation, is pronounced. Current transparency requirements hint at several protections, yet the defined disclosures fall short of these aims, a topic seldom explored in the literature. Our work systematically identifies current and prospective transparency objectives and aligns them with minimum disclosure requirements.

The second fallacy is the *remediation gap* that exists between what companies are required to disclose and the mechanisms that ensure that the disclosures are made accurately. Given the invocation of nutrition labeling for AI policy, we illustrate these enforcement challenges in US nutritional labeling itself. The transparency efforts in food regulation illustrate the many challenges that arise when considering the actors, incentives, power, and resources required to ensure meaningful transparency. These challenges are particularly acute for AI, where technical complexity further complicates verification efforts.

The third fallacy is the *outcome gap* between the data transparency compliance behavior and the actions that are actually needed to protect individuals, data creators, other stakeholders. Many works have highlighted the failures of mandatory disclosures (Fung et al., 2007; Ben-Shahar & Schneider, 2017) and we identify how some of these failure modes also translate to AI transparency policy. Our contributions in this work are as follows:

- **Survey motivations** for data transparency and taxonomize specific levels of data disclosures necessary for each objective (Table 2), clarifying the *disclosure gap*.
- **Illustrate the challenges** of enforcing data transparency requirements, drawing analogies from nutrition labels, showing the *remediation gap*.
- **Identify the failure cases** of mandatory disclosures and how the *outcome gap* impacts data transparency policy.
- **Provide a case study** on California's AB2013 to illustrate how policy debates on data transparency play out.
- **Outline a technical research agenda** on evidence needed in order to formulate meaningful policy.

B. Related Work

Many calls have been made for understanding the training data of machine learning models from documentation frameworks for datasets (Guha et al., 2024) to audits of dataset licensing (Longpre et al., 2023) to Dataset Nutrition Labels aimed at avoiding incomplete or biased training data (Holland et al., 2018). These early works have motivated the emergence of disclosure requirements as a lever invoked to regulate AI in the age of foundation models (Guha et al., 2024). Legislation such as California Assembly Bill 2013 (AB2013) on Generative Artificial Intelligence: Training Data Transparency for System Developers (2024) and Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (AI Act) include multiple types of disclosure requirements for system developers and model providers, respectively.

However, the connection between the ultimate objectives of transparency and the narrow, specific transparency requirements in legislation remains unclear. For example, disclosure of training data size will not directly help meet the objective of protecting personal information. Guha et al. highlight the disconnect between AI regulation and its intended purpose, *the regulatory alignment problem*, arguing that disclosure requirements in AI regulation may be feasible and inexpensive to

adopt but may not result in actionable protection of consumers. Furthermore, they note that disclosures in AI often shift the responsibility of designating which information is relevant to system developers or model providers, which may limit the effectiveness of the original goals of policy makers.

For data transparency for foundation models, in particular, the Foundation Model Transparency Index (Bommasani et al., 2023) introduces 10 data-related indicators of model transparency. These indicators include whether developers disclose information about data size, composition, processing, and curation, as well as whether copyright, proprietary, or personal data is included. The Foundation Model Transparency Index also includes transparency indicators with respect to data labor. An index approach allows a comparison of the transparency practices of various model developers, but the downstream impact of improved transparency on each indicator is unclear. Additionally, Ananny & Crawford interrogate transparency as an ideal and how transparency for the sake of transparency may not be an effective tool to achieve algorithmic accountability. The presence of disclosures does not necessarily guarantee adequate auditing to protect users, content creators, and data workers. In a grounded analysis of 18 targeted transparency policies, Fung et al. highlight how public disclosures are frequently ineffective and counterproductive due to information quality, and loopholes found by disclosurers'; targeted disclosures must be user-centric and sustainable to be effective. Ben-Shahar & Schneider point out that most people do not find the flood of information provided through mandated disclosures useful, and policy makers should not lean on disclosures as a crutch.

In this work, we add to this discussion by focusing on data transparency (in comparison to, for instance, model transparency) from an institutional perspective. By focusing on data transparency, we can provide a deep analysis of forms of data transparency alongside various policy objectives, with a particular eye on data-specific enforcement challenges.

C. Levels of Data Transparency

Despite data transparency being a frequent goal in AI policy, transparency can refer to many different things. In this section, we taxonomize different forms of data transparency, which we separate into (1) documentation and descriptions and (2) data access. This taxonomy is based on the artifact the disclosure process produces; using this taxonomy allows a deeper analysis of the burden of providing these disclosures and the technical feasibility of validating these disclosures. This section is intended to provide a common language around which we develop our later discussions. We give an overview of how each form of transparency can be verified, although we do not dive into the details of the remediation gap until a later section. In discussions of verification, model access of different levels is also discussed; however, the scope of this work is on taxonomizing data transparency in particular.

C.1. Level 1: Documentation and Descriptions

This level of transparency includes various information *surrounding* the datasets used in the development of generative AI models. We further separate data information into the following two categories: information about the dataset itself (e.g., how data is processed and used) and information about the dataset supply chain (e.g., how data is acquired). This level of disclosure produces an artifact that is a documentation of the data.

Level 1a: Dataset Information Dataset information is metadata on the content and construction of the dataset. Some examples, along with those for which existing regulations mandate them, include

- Summary Statistics (AB 2013): High-level dataset statistics, such as the number of training points.
- Data Sources (AB 2013): Sources (e.g. web, books, video) from which training data was gathered and what purpose each source serves.
- Synthetic Data (AB 2013): Description of whether and how synthetic data was used in the training data, and how the synthetic data was generated.
- Personal Information (AB 2013, EU AI Act): Whether the datasets used contain personally identifiable information or aggregate consumer information.
- Copyrighted Content (AB 2013): Whether the datasets contain copyright, trademarked, or patented information or whether the datasets are in the public domain.

- Data Processing (GDPR, AB 2013, EU AI Act): Cleaning, filtering, removal of PII and other processing steps taken before the data were used for model training.

Verifying this level of information though observing only a trained model is extremely challenging. Namely, a model can "memorize" a piece of content without "regurgitation" (Cooper & Grimmelmann, 2024). This distinction is important because it means that even when personal information or copyrighted content is a part of training data, the process to generate that content may be hard. Even when the goal is "extraction", when adversarial users intentionally try to make a model produce a certain output verbatim, not every piece of data in the training dataset can be extracted successfully (Nasr et al., 2023). Furthermore, overlap in content between data sources makes extraction and membership inference difficult. For more high-level information, such as the number of data points (required in AB2013) and exact data processing steps, is not possible to verify with current techniques.

Level 1b: Data Supply Chain Information We consider the data supply chain broadly to include many different types of data, including public datasets, publicly available data, licensed data, human-generated data and synthetically generated data. Here, data supply chain refers to the process and network by which data used for AI is produced (Hopkins et al., 2025; Lee et al., 2023; Widder & Nafus, 2023; Cobbe et al., 2023). This level of information around the data-generating process of the datasets used involves both the data producers and the model developers. Although existing regulations focus on system developers and model providers, it is likely that these parties know or control aspects of the supply chain when they work with data producers. In this type of disclosure, documentation and descriptions include the following:

- Data Collection (AB 2013, EU AI Act): Information about when, how, and from whom the dataset was collected.
- Consent (GDPR): If training data includes personal data, whether and how consent was collected.
- Licensing (AB 2013): Information about whether the dataset was purchased or licensed; permissions around dataset usage.
- Vendors: Information on which data vendors were used in the dataset collection, labeling, and processing.
- Contracts: Contracts between data providers and model developers that allow access to data for training and evaluation.
- Data workers: Details about who labeled and collected the data and how they were compensated or attributed.
- Data Provenance: Information on how the dataset is derived including sourcing, creation, and licensing.[5]

Similar to Dataset Information (1a), verifying this category of information broadly is difficult without access to the entire training dataset. For example, from testing the output of a model, it is impossible to discern whether consent was collected from individuals before their data were used for training or whether the reported compensation given to data workers was accurately reported. Even the push to verify these disclosures may need to come from complaints from consumers, data workers, or data vendors who know that the reported information is inaccurate.

C.2. Level 2: Data access

This second level of transparency involves direct access to training data in some form. Although not traditionally considered a *disclosure*, access to training data can be required to achieve certain policy goals, despite being costly or in contradiction to the business interests of model developers. For models that disclose training data, if all data used in the training process are publicly available (Groeneveld et al., 2024), releasing sufficient dataset statistics is equivalent to having full access to the training dataset. For example, if a model was trained only on the Pile (Gao et al., 2020), a publicly available dataset, specifying the training data information, is sufficient to access the entire dataset. For proprietary models, datasets are often secret, and dataset descriptions are not equivalent to direct access to datasets. When data contain personal information, access to anonymized data also falls under this category of disclosure, since anonymized data can be included as a subset or as a full dataset. Access to a dataset itself is not binary; we discuss three forms: membership access, subset access, and full data access.

[5]There is overlap with the prior bullet points, but the focus is on understanding the original sources which training datasets stem from (Longpre et al., 2023).

Level 2a: Membership Access

- Membership query access to training data: Given a document, such as a webpage, a book, or a code repository, a membership query access returns a yes or no answer about whether the piece of data is included in the dataset. The technical feasibility of such an access is contingent on the length and uniqueness of the data segment since overlap itself is a fuzzy concept. Relaxations of this notion include searching for parts of a document through keywords or phrases or approximate matches to the query document.

Level 2b: Subset Access

- Sample access to training dataset subset: To illustrate the format of the data used in different phases of training, some samples can be provided on the type of data used for each phase of training. Although this gives insight into the real data used, it does not reveal all of the data choices.
- Partial access to training data or the generation of training data (e.g., public datasets, prompts) (EU DSA): Portions of the dataset, transformations of public datasets, or instructions for generation of datasets (e.g., prompts) provide partial access to training data.

Level 2c: Full Data Access

- Full data access: The entire dataset used for one or more components of training, validation, and testing of the model before deployment is made available. This includes new data used after substantial changes to the model. Full dataset access would also imply access to the mentioned transparency levels (e.g., 1a, 2a, and 2b).

When verifying the correctness of each of these forms of data disclosure, it is important to consider whether the disclosed samples are actually used (precision) and whether the disclosed data are not exhaustive of all the training data used (recall). For validating the disclosure of training data, there are approximate methods to test precision. For recall, when training data points are available, data attribution methods can be used to test whether training datasets were not reported (Park et al., 2023; Grosse et al., 2023; Worledge et al., 2024). However, without knowledge of what pieces of data to test for, verifying recall (e.g., all training data were reported) is very difficult. Methods of verifying that the data provided in levels 2b and 2c were indeed used in the training set (recall) include comparing the likelihood of the training data vs data not included in training (Shokri et al., 2016).

D. The Disclosure Gap

Our goal is to map each data objective to the requisite levels of data transparency that would enable the objective. In doing so, we can map how well each objective is supported by the actual form of data transparency requested. For example, if the goal is to assess copyright infringement, what is the actual form of transparency that allows this assessment? Table 2 provides a summary of the fine-grained subobjectives and the necessary disclosures and the corresponding verification requirements.[6]

D.1. Protecting Personal Information

When regulation mandates data transparency around personal information, the aim is often to protect individuals and their data. Personal information may include name, image, and government identification numbers. This type of data transparency would empower consumers to make informed decisions around AI products and services that use personal data for training. Data transparency alone is not enough to protect these rights since downstream actions based on privacy law is required. However, data transparency provides the necessary first step of disclosure of when personal information is used.

Existing Policy The European Union's regulatory framework, specifically the General Data Protection Regulation (GDPR) and the EU AI Act, establishes mandatory disclosure requirements for the processing of personal data. Under Article 13 of GDPR, organizations must explicitly inform data subjects about the intended purposes of data collection at the time such

[6]The corresponding data disclosure levels are described in Section C.

SUBOBJECTIVE	STAKEHOLDER	MINIMUM DISCLOSURE	VERIFICATION
Protecting Personal Information			
Understand the presence of PII of citizens in training data	Regulator	(1a) Dataset Information: Personal Information (1b) Data Supply Chain Information: Data Collection	Regulators may need access to the training dataset to verify or rely on complaints of models revealing personal information.
Choose AI products and services that do not use personal data	Individual User	(1a) Dataset Information: Personal Information	Individual users do not have the tools to test comprehensively whether personal data has been used
Discern whether their name and likeness were included in the training data for a specific mode	Individual User	(2a) Membership Access: Membership query access to training data	Users can test whether their information can be revealed by the model, but cannot verify that their personal data was not used at all (Cooper & Grimmelmann, 2024).
Test for the leakage of PII	Auditor	N/A - Leakage must be tested through model access	
Assurance of Training Data Quality			
Assurance of diverse data collection procedures	Regulator	(1a): Dataset Information: Data Sources (1b): Data Supply Chain Information: Data Collection	Regulators can verify that data sources are diverse by higher levels of data access (e.g., 2a membership access)
Assurance of data quality	Regulator	(1a): Dataset Information: Data Processing (1b): Data Supply Chain Information: Data Collection	Regulators can verify that data processing has been done to improve quality by inspecting source code for data.
Interpret whether predictions can be trusted for their specific individual profile	Individual User	(1a): Dataset Information: Data Sources (1b): Data Supply Chain Information: Data Collection	Individual users could verify through higher levels of data access (e.g., 2b subset access)
Assurance of Data Representativeness	Auditor	(2c): Full Data Access	
Copyright and Terms of Use Protections			
Assurance of copyright protection	Data Creator	(2a) Membership Access: Membership query access to training data	Verifying correctness of membership query responses relies on access to the actual training data.
Assurance of copyright and licensing law compliance	Regulator	(1b) Data Supply Chain - Licensing	It is difficult for regulators to verify that all data used is licensed. Regulators may need to rely on complaints from data owners.
Compliance with terms of use	Model Provider	(1b) Data Supply Chain - Data Collection	Model developers may verify that competitors did not train on data generated by their models through watermarking their outputs. (Kirchenbauer et al., 2023)
Evaluation Generalization			
Assurance of no train-test overlap	Regulator	(1a) Dataset Information: Data Processing	While it is impossible to verify the entire data processing pipeline, it might be possible to identify significant omissions through model behavior (Golchin & Surdeanu, 2023; Shi et al., 2023).
Check for the presence of evaluation examples in the training data	Auditor	(2a) Membership Access: Membership query access to training data	Verifying that membership query responses requires access to training data. However, it may be possible to observe behavior on benchmarks to infer potential contamination (Zhang et al., 2024).
Data Laborer Protections			
Choose AI platforms and services that are produced via fair compensation	Consumer	(1b) Data Supply Chain Information: Data workers	Currently consumers cannot verify that AI platforms fairly compensated data workers.
Check that forced or child labor is not a part of generating data	Auditor	(1b) Data Supply Chain Information: Data workers	Regulators may use a complaint system to censure companies that engage in labor practices they are not reporting.

Table 2. Overview of mapping between the objectives of data transparency and the minimum level of *data* disclosure. Different disclosures suffer from different challenges in verification.

data is obtained. Furthermore, Article 10 of the EU AI Act imposes additional obligations on providers of high-risk AI systems to implement data governance practices that protect 'individuals' fundamental rights and freedoms. In California, AB 2013 § 3111(a)(7) specifies that AI system developers must provide information about whether the datasets include personal information or aggregate consumer information.

Necessary Disclosures Although the goal of these provisions is to protect users if their data are collected for use in training of models, it is unclear what types of disclosure are necessary or sufficient to guarantee these supposed protections. Different stakeholders may want to protect personal information in different scenarios. Although some objectives such as consumer choice of products that do not use PII is possible with level 1 disclosures, other objectives such as testing for leakage of PII are not possible through data disclosures but only possible through model disclosures (Table 2).

D.2. Assurance of Training Data Quality

Existing regulation also has mandates for training data quality which are driven by the belief that data quality would reduce harm or discrimination experienced by downstream users. In the computer science literature, the importance of data quality has been highlighted (Xu et al., 2021; Longpre et al., 2024b; Wettig et al., 2024); sometimes as a factor even more important than the size of the dataset (Zhou et al., 2023; Shen et al., 2024).

Existing Policy For applications under high-risk AI systems (e.g., health care, criminal justice), the EU AI Act specifies requirements for training, validation, and testing data. In Article 10(3) of the EU AI Act, datasets are required to be "sufficiently representative, and to the best extent possible, free of errors and complete because of the intended purpose". However, this difficult albeit not impossible because of language like "sufficiently" and "to the best extent possible". Other requirements for high-risk data include disclosure of datasets used to deployers (Article 13(3)(b)(vi)) and documentation of data collection and processing details (Article 17(1)(f)).

Necessary Disclosures For regulators and individual users interested in ensuring data quality and understanding the representativeness of the training dataset, level 1 disclosures are sufficient since the inclusion of various data sources can be reported through these disclosures. If an auditor wanted to evaluate the degree to which the training data is representative of a customer segment or society more generally, full data access (2c) would be necessary. Thus, the validation of training data quality often requires levels of data transparency existing policy does not mandate.

While we discuss how to achieve assurances of data quality, in some cases, the actual goal might be to ensure the model that the datasets produce does not exhibit undesirable behaviors (e.g., biases, dangerous answers). In these cases, it is important to consider that decisions during model training can be made to mitigate bad behavior even if the training dataset has limitations. For example, even if a dataset does not reflect the distribution of the population using the downstream product, data points representing minority views or preferences can be up-weighted to mitigate the downstream biases of a model. Thus, data transparency is not an effective tool since a representative dataset is neither sufficient nor necessary to achieve an unbiased downstream model.

D.3. Copyright and Terms of Use Protections

As generative AI become increasingly capable of producing high-quality creative and editorial content, concerns have been raised about whether the data used to train these models contain copyrighted materials. Copyright is assigned from 'the moment' a piece is created (United States Congress, 1976; WIPO, 1886), and thus the massive scale of datasets of books and images that are used to train generative models often include copyrighted material (Bandy & Vincent, 2021; Karamolegkou et al., 2023). However, the enforcement of copyright protections often falls in the hands of the creators. As a result, supply chain transparency (Lee et al., 2023), including filtering of training data to reduce the likelihood of copyright infringement (Henderson et al., 2023), has been suggested. Our analysis of data transparency disclosure requirements is again only a necessary but not sufficient step for copyright protections. The topic of fair use for generative models is still under debate; if training on copyrighted content is fair use, data transparency would not protect creators.

Furthermore, model creators and developers often require the usage of their services to comply with their terms of use (TOS). For example, OpenAI's terms of use specify that model output cannot be used for the development of models that compete with the company.[7] At the same time, the terms of use grant users ownership of the generated content. Lemley

[7] https://openai.com/policies/row-terms-of-use/

& Henderson highlight that TOS restrictions face legal challenges since it is unclear whether copyright applies to model weights and outputs. Thus, specifying how outputs of generative models can and cannot be used may face legal challenges. Data transparency requirements for disclosing and specifying how data is collected or created may disincentivize companies from generating data using other platforms that explicitly prohibit training on their data.

Existing Policy In the European markets, the EU AI Act specifies that "Providers of general-purpose AI models shall: put in place a policy to comply with Union law on copyright and related rights" (Article 53(1)(c)). Although no transparency requirement is explicitly specified, compliance may require model providers to confirm that copyright data filtering was performed as part of the training data filtering process. For noncommercial purposes, in particular, there are exemptions for training models using copyrighted data. AB2013 specifies that training data transparency should include § 3111(a)(5-6) "Whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain. Whether the datasets were purchased or licensed by the developer." By directly requiring the transparency of training data, creators can theoretically sidestep the tricky legal landscape of eliciting the outputs of generative models.

Necessary Disclosures While disclosing whether copyright data was used in training data seems low-cost at a high level, these very same disclosures are not useful for creators because they are too vague. Even if companies claim that no copyrighted data was used, it would be difficult to make these claims with certainty for every piece of content with a copyright. The more effective level of disclosure would be to provide access to the data creator in the membership access (2a) (Table 2). At the data collection and curation step, upstream in model training, revealing data sources and their associated licensing may be helpful as an additional incentive to protect copyright.

D.4. Forward-Looking Objectives

Next, we look beyond existing regulations to understand the potential future objectives of transparency for generative AI. Although these objectives have been discussed for machine learning in general, translating these objectives to generative AI models has not yet appeared in legislation.[8]

Evaluation Generalization Model evaluations are only valid if there are guarantees evaluation examples are not in the training data or in-context examples. This is often called "train-test overlap" (Kapoor & Narayanan, 2022; Golchin & Surdeanu, 2023; Zhang et al., 2024). Data transparency could theoretically ensure the validity of evaluations against standard benchmarks. The closest existing policy is the EU AI Act Article 10(3) specifying that training data to be free from errors as much as possible. Ensuring that training data do not contain evaluation data could generally fit into this requirement if evaluation data is considered an erroneous inclusion.

However, verifying that there is no test set contamination can be difficult. Recent work has shown that specific passages can be generated verbatim even without the passage appearing in the training data (Liu et al., 2025). The key idea is that modern language models demonstrate sufficient generalization capabilities such that they can learn to "piece together" noisy transformations of a passage into the original – these transformations can be hard to detect both for humans and automated checks. Removing test-set contamination is also challenging since since deleting information from models through methods such as unlearning does not provide guarantees about changes in model output (Cooper et al., 2024). Viewed under an adversarial lens, this raises questions as to whether any level of data transparency can guarantee no contamination of evaluation benchmarks.

Data Laborer Protections Many works in AI have shown the importance of data quality in fine-tuning generative models. In fact, for many tasks, data for fine-tuning are human-written or annotated. Existing legislation requires supply chain transparency to protect against labor exploitation. For example, the California Transparency in Supply Chains Act, Cal. Civ. Code § 1714.43 (2010) was designed to help consumers make informed decisions by requiring retail sellers and manufacturers to disclose company standards for trafficking and slavery in their supply chain. As another example, the Uyghur Forced Labor Prevention Act, Public Law No. 117-78, 22 U.S.C. § 6901 prevents the import of goods made wholly or in part by forced labor in the People's Republic of China. Recent investigative reports on data labor practices of large AI companies found that Kenyan workers who provide data annotations are tasked with traumatizing work without ensuring job security (Perrigo, 2023). Protections proposed for data workers have been proposed but rarely adopted (Partnership on AI, 2024).

[8]We also include 3 additional objectives in the supplementary materials that we did not include due to space constraints.

E. The Remediation Gap

While many commentators have invoked nutritional facts as the archetype disclosure for data quality and transparency in AI (Holland et al., 2018; Gebru et al., 2018), such proposals manifest limited understanding of the institutional reality of disclosure, including nutrition labeling. In fact, US nutrition disclosure exemplifies the remediation gap and offers important lessons for naively importing transparency provisions.

E.1. Nutritional Facts: A Motivating Example

Nutrition facts first became widespread in the United States in 1973 when the FDA proposed standardizing nutrition labels on foods (Wartella et al., 2010). This effort was in response to consumers' requests that better information about food be available with the rise of processed foods. By providing nutrition facts, food producers theoretically enable consumers to better choose what they consume. Despite these well-motivated objectives, there remain serious limitations to nutrition labeling Heinzerling. Food regulation in the United States includes compelled disclosures, prohibitions on fraudulent representations, and restrictions on discretionary disclosures. However, this formal focus on disclosure masks structural challenges that undermine the system's commitment to transparency.

Public Enforcement: The first challenge is regulatory fragmentation and capacity. In the United States, the USDA and FDA, the two principal agencies in charge of food labeling, have different processes for reviewing food labels and different definitions of food claims. Labeling emanates from the Food, Drug, and Cosmetics Act, the Federal Meat Inspection Act, the Poultry Products Inspection Act, the Egg Products Inspection Act, the Agricultural Marketing Act, and the Fair Packaging and Labeling Act. If that isn't confusing enough, the Federal Trade Commission has regulatory jurisdiction over deceptive advertising and associated claims. Despite the number of federal authorities, resources are limited. The capacity for public enforcement is dwarfed by hundreds of thousands of claims in the marketplace. From 1998 to 2008, the FDA secured only two court injunctions for misleading labeling (Heinzerling, 2015). Despite the fact that FDA found in random samples in 1994 that 48% of products misrepresented vitamin A and C volumes, both USDA and FDA abandoned attempts at verification via random sampling (id.). Said one FDA official, public enforcement "with the legal and resource restraints we work under is a little like playing Whac-a-Mole, with one hand tied behind your back." (Taylor, 2010)

Private Enforcement: Perhaps private enforcement could make up for the lack of public enforcement. But there is no federal private cause of action in the United States for consumers to sue. The Lanham Act establishes private right of action for *competitors* to seek remedies when harmed by deceptive representations. POM Wonderful, for instance, successfully sued Coca-Cola for misleading claims of using pomegranate juice (POM Wonderful LLC v. Coca-Cola Co., 2014), even whilst their own claims of the disease-preventive properties of pomegranate juice were investigated by the FTC (Federal Trade Commission, 2013). Consumer groups have instead pursued state-level claims, such as under California's Unfair Competition Law. But such suits are resource-intensive and the effect on the reliability of claims overall has been limited. As Heinzerling writes, the "existing legal system for food fails to deliver the transparency it seems to promise." Despite the systems goals of transparency, the effect may be outright confusion. As stated by one former FDA Commissioner, food labels are "so opaque or confusing that only consumers with the hermeneutic abilities of a Talmudic scholar can peel back the encoded layers of meaning. That is because labels spring not from disinterested scientific reasoning but from lobbying, negotiation, and compromise." (Stark et al., 2012)

Put differently, what the underenforcement of nutrition labeling reveals is the remedial gap: the inability to secure remedies for inaccurate, deceptive, or unreliable disclosures. US food labeling is the opposite of what policy advocates desire for AI. It illustrates that transparency is only useful insofar as it can be effectively audited and enforced. That does not have to be the case. Food policies in Mexico, for instance, show signs of stricter enforcement (Crosbie et al., 2023; Daily).

To understand the challenges and suggest best practices for policymakers who wish to see data transparency for AI models, we examine the following two aspects: (1) Compliance with Mandates: Do companies share data information as directed by the data transparency policy? and (2) Verifiability of Disclosed Information: Is the information shared by the companies verifiably correct?

E.2. Component 1: Compliance with Transparency Mandates

As bills like California's AB 2013 and the EU AI Act mandate data information disclosure, the first goal of enforcement is to require model developers to comply with necessary disclosures. However, current enforcement mechanisms for these disclosure requirements are weak and may not promote full compliance. These two bills come into effect in January and

August 2026, respectively. Their scope differs in that AB2013 requires only generative AI models to comply, while the EU AI Act applies to all AI systems and provides different requirements for transparency depending on a system's risk tier.

Public Enforcement: Although the EU AI Act is set to be implemented, supervised, and enforced by the EU AI Office, it is much less clear what bodies could be responsible for enforcing data transparency requirements in the US. First, AB2013 does not assign a specific agency with enforcement and does not discuss auditing and verification. Enforcement will likely take place under California's Unfair Competition Law, which allows the California Attorney General's office to bring public enforcement actions. Such enforcement would be resource intensive, when there is little expertise within state government around AI issues.

Second, AB2013 makes monitoring exceptionally difficult, as there is no standardization for data disclosures are important. AB2013 requires only that data transparency statistics, in some form, be posted on the developer's website, and California's experience with privacy disclosures, which similarly lacked any standardization, were decried as "functionally useless." (Luthi, 2021) The California Privacy Protection Agency (CPPA) has brought actions against data brokers for failing to register and pay an annual fee to fund the California Data Broker Registry, which hosts a data deletion mechanism. However, these legal actions can be costly and may not succeed.

Third, while noncompliance with the transparency provisions Article 50 of the EU AI act subjects model developers to fines up to €15,000,000 or up to 3% of annual worldwide turnover (Article 99), California UCL violations are subject to civil penalties of up to $2500 per violation.

Last, securing compliance with disclosure can also lead to unintended consequences, by siphoning enforcement resources from more nefarious practices (Ho, 2012). To avoid this type of outcome, both disclosures and subsequent compliance mechanisms must be designed taking into account the implementation burden.

Private Enforcement: California law also has a path for private enforcement of AB2013 under California's Unfair Competition Law. However, this path remains limited. First, private parties must suffer injury and financial loss because of an actor's failure to abide by the disclosure requirement of AB2013. A party who owns copyright may not be able to assert harm because a developer trainined on copyrighted data, but must instead point to harm for the failure to *disclose* intellectual property status. Second, private parties may not seek civil penalties, and can seek only injunctive relief or restitution. Private litigants hence have greater incentive to pursue other paths for recovery, such as copyright or privacy law. Abiding by AB2013, however, can increase the risk of litigation, providing a powerful disincentive for developers to fully disclose (Longpre et al., 2024a). Developers may rationally decide not to comply with transparency requirements – given the lower risks of non-compliance – than expose themselves to more serious liability – including fines and criminal sanctions – under copyright or privacy law (Lee et al., 2023).

Takeaways: In its current form, compliance with transparency mandates is difficult to enforce. Instead, we propose that enforcement of transparency mandates is based on three principles. First, legislation should clearly allocate enforcement responsibility to a primary agency. Second, that agency should have adequate resources. Third, penalties should be proportional to the scope and scale of the problem. On each of these counts, in contrast to the EU AI Act, AB2013 commits the same sin of nutrition disclosure: high aspirations coupled with low enforcement.

E.3. Component 2: Verifiability of Data Transparency Disclosures

Once disclosures are made available, the second component of enforcement is to ensure that the disclosures are accurate. In some food label cases, competitor companies took on the burden of verifying or disproving false claims made by some companies (POM Wonderful LLC v. Coca-Cola Co., 2014). Although verifying the accuracy of food labels is laborious, it may not even be possible for external actors to verify data disclosures for AI. However, verifiability is essential because truthful data transparency disclosures enable the achievement of the intended objectives of policymakers. For example, for companies to truthfully specify that they have not used personal data, they would indeed have to avoid using personal data. Verifiable data transparency gives evidence to pursue legal action in areas such as copyright, privacy, and discrimination.

Enforcement Agencies: Although centralized enforcement may allow a specific agency to have adequate authority to ensure that companies make necessary disclosures, verifying that disclosures are accurate requires significant domain expertise. Investigating complaints that data disclosures are misleading is likely beyond the capacity of existing agencies in charge of privacy or consumer protection due to a lack of AI expertise (Jurowetzki et al., 2025). In nutrition, the FDA and USDA are agencies staffed largely by expert scientists, and yet verification of a significant number of food products is already too costly. For data transparency, government agencies have few if any AI experts.

Legal Action: The success of legal actions for copyright, privacy, and anti-discrimination depends on verified claims of improper data usage. Verifying whether data disclosures are accurate may become the burden of plaintiffs seeking to sue model developers. For example, harm due to misrepresenting dataset content may give competitors incentives to challenge under California's Unfair Competition Law.

Technical Feasibility: As highlighted in the right column of Table 1, the high-level statistics outlined in AB2013, such as the range of the number of training points and whether copyrighted and personal information is used, are too general to verify with only model access. Currently, even verifying the existence of a specific data source or data point used for training in enormous datasets can be difficult (see the Research Directions Section).

Takeaways: Verifying data disclosures is difficult, but necessary to ensure the disclosures are truthful. Data disclosure requirements should (i) create systems, authorities, and infrastructure that enable verification of disclosures, and (ii) not require disclosures that are impossible to verify, without creating adequate supporting processes. However, since significant technical barriers still remain to verified data disclosures through model access, we point to directions of technical research needed towards this effort of verifying disclosures in a later section. The status of the research there will determine what kinds of policy are enforceable in practice.

F. The Outcome Gap

Even when mandated disclosures provide sufficient information and have adequate enforcement mechanisms, a gap remains between their intended goals and their actual impact. For data transparency policies to achieve accountability for AI system developers, including the protection of consumers' personal information and the copyright of creators, the right mechanisms for impact must be in place. The failures of mandated disclosures across domains from privacy regulation to vehicle safety ratings have been well documented (Ben-Shahar & Schneider, 2017). Three key challenges that contribute to this outcome gap include: whether the information released is truly informative for consumers and businesses; whether there are viable alternatives in the market; and whether these factors effectively incentivize companies to change their behavior in a way that aligns with the original policy objectives. These challenges particularly plague AI data transparency disclosures.

First, the impact of disclosures depends on whether the audience of the disclosure can adequately understand them. In federal home loans, for example, the 50 different disclosure forms overwhelm consumers, making it difficult to comprehend the risks that are being disclosed, and therefore these disclosures themselves do not curb predatory lending (Stark et al., 2012). For AI systems, complete disclosures including data sourcing, collection, and processing would create the same information overload for average consumers. Instead, data transparency mandates should focus on enabling information intermediaries such as journalists, academics, and nonprofits to translate these complex disclosures into actionable insights for consumers.

Second, certain data practices are ubiquitous in model training, limiting the effectiveness of disclosures in driving market-based accountability. Modern generative AI models are often pre-trained on as much of the Internet as possible, with companies mostly adopting different data approaches to achieve nuanced goals like safety and preference alignment. This ubiquitous 'throw everything in' approach for pre-training severely limits consumer optionality. For example, it is possible that all models good at news writing need to be trained in copyrighted content. For transparency mandates to successfully generate consumer pressure on AI system developers, meaningful alternatives must be available, a condition that current industry practices make difficult to satisfy.

Third, consumer pressure must incentivize the desired corporate behavior (Wang et al., 2024): better practices around data provenance and quality. Companies might respond to transparency mandates by creating technically compliant but low-quality disclosures, such as listing thousands of data sources without meaningful prioritization or using vague categorical descriptions like 'publicly available online content' instead of specific sources. Returning to the example of food labels, the disclosure of sesame allergens has led increased sesame reports as an ingredient, as there is no penalty for over-reporting (Aleccia, 2022). Similarly, model developers can relabel, reprocess, or cherry pick datasets rather than reducing the use of sensitive data. Fung et al. highlight that effective disclosures give information to consumers that changes their behavior, which then, in turn, changes the behavior of disclosers in a way that serves the original policy goals. As AI advances rapidly, policymakers will likely have to iterate on data transparency policy to address evasive tactics when they arise.

G. The Emergence of Data Transparency Regulation: A Case Study on AB 2013

In practice, AI regulation, like any other legislation, is the result of a compromise between different stakeholders with different interests. Existing data transparency policies are the artifacts of arguments for data transparency and the push against it. To understand how the disclosure and remediation gap arose, we provide a closer analysis of the process that generated the California Bill AB2013.

G.1. Consumer Protection: A Driver for Data Transparency

Introduced in January 2024, California State Bill 2013 initially defined terms related to artificial intelligence systems, developers, and synthetic data generation, along with mandates for disclosing certain aspects of the data supporting model development. Early revisions of the bill extended the bill's scope by widening disclosure to include personal and consumer information, dataset modifications, dataset statistics, and outline exceptions for AI services dedicated solely to security and integrity. The Assembly Committee on Privacy and Consumer Protection emphasized that without knowledge of the data used to train AI products, Californians cannot make informed purchasing decisions. This position was grounded in existing privacy laws. The committee justified the disclosure of synthetic data by pointing out the risks of bias and "model collapse," assessing the proposed requirements as modest. At this stage, supporters of the bill underscored the importance of the bill for public transparency, awareness, and protection for Californians.

Based on an academic report (Chmielinski et al., 2024), the judiciary committee saw data transparency as a way to reducing biases, addressing hallucinations and problematic outputs, and easing privacy and copyright issues. At this stage in the Senate, the California Labor Federation began to support the bill motivated by concerns that AI systems might be used to evaluate employee performance and hiring. The Federation stressed that workers should be made aware of the training data used for these decisions to prevent the non-consensual use of personal data and to tackle both implicit and explicit biases in training data.

G.2. The Push Against Data Transparency

Resistance to the bill emerged early from several key groups expressing doubts about the bill's technical feasibility, vague definitions of key terms, and weak protection of trade secrets and intellectual property. To address these concerns, committee amendments narrowed the definition of artificial intelligence and lowered disclosure requirements from "a description" to "a high-level summary." As the bill reached Assembly Floor Analysis, lobbyists in opposition argued that revealing training data could hinder competition and terms like 'but not limited to' were too vague. As a result, the Senate version included more exceptions for domains where AI systems need not disclose training data (e.g., a GenAI system for operation of aircraft in the national airspace) and relaxed data point disclosures to general ranges.

In the Senate judiciary analysis phase, opposition expanded as a group of organizations raised further criticism mainly focused on two issues: the absence of a risk-level distinction (suggesting only high-risk AI should meet data transparency standards) and the broad interpretation of "artificial intelligence system or service." These concerns shaped subsequent Senate floor amendments, notably narrowing the bill's scope to "generative artificial intelligence" and restricting disclosure obligations to the original developers of a GenAI system, even if modified by a third party.

The forces for and against data transparency illustrates the path from the idea of enabling consumer protection through data transparency to passing a bill with very high-level disclosures. The end product neither ensures that the disclosures are sufficient for the original goal nor provide enforcement provisions to ensure compliance of these disclosures.

H. Recommendations

While AI disclosures share some challenges with the vast landscape of mandated disclosures, there are many unique technical and policy considerations for AI in particular. We conclude our work by describing both recommendations for policymakers creating or revising data transparency policy as well as research directions for computer scientists.

H.1. Towards Better Data Transparency Policy

Our analysis highlights that (1) mandated disclosures should provide enough information to be actionable for consumers, data creators, auditors, and regulators, (2) enforcement mechanisms for compliance and verification of these disclosures are crucial, and (3) disclosures should be designed to change the behavior of AI system developers in a way that is aligned

with intended policy goals. For policymakers considering AI transparency legislation in the future, our analysis specifically illustrates the importance of the following:

- **Clarity**: Disclosures aimed at addressing too many things may cause confusion. It is more effective to focus on a clear goal and ensure that the transparency measures that correspond to this goal have sufficient enforcement mechanisms.
- **Standardization**: Successful implementation depends on standardized reporting. For example, provide a common format and a website where all companies that are subject to disclosure requirements can submit their information.
- **Information Intermediation**: When public resources are limited, policy should be designed to empower intermediaries, such as private litigants or third parties, capable of conducting informative analysis on disclosed data.

While these recommendations are a starting point for today's AI landscape, policymakers will likely have to iteratively update transparency requirements to account for the shifting technical landscape and the actual observed impact of requiring data transparency disclosures. Ultimately, our paper emphasizes that disclosure is not a free lunch for accountability – careful alignment of goals with objectives and robust enforcement mechanisms are necessary.

H.2. Research Directions for Computer Scientists for Training Data Disclosure

We repeatedly highlight the importance of verification as a part of the enforcement of data transparency. An essential ingredient for verification is the continued development of empirical techniques for data verification. Following previous work identifying key technical topics for AI governance (Reuel et al., 2024); we give a deeper analysis of technical research directions for the verifiability of data disclosures, since significant technical bottlenecks prevent efficient auditing of the truthfulness of data transparency disclosures. This section discusses research directions related to both enabling and verifying training data disclosure – this research would be instrumental in understanding which kinds of policy and regulations are actually enforceable in practice.

Granularity of data usage A key challenge in data disclosure is deciding the granularity with which data membership should be reported. Granularity refers to the specificity level at which data usage is tracked and disclosed (Maini et al., 2024). At the coarsest granularity, developers might simply acknowledge that a dataset was used; finer granularity could involve tracking specific documents, paragraphs, or even exact text sequences. More granular disclosure introduces computational overhead but yields more meaningful transparency - for example, sequence-level reporting might require precise but potentially large lookup tables or n gram statistics. Potential research directions include:

- Developing compact representations of training examples that enable efficient lookups;
- Studying trade-offs between data privacy, utility of transparency, and computational feasibility across different levels of granularity.

Revisiting definitions for training set inclusion A more fundamental issue with training data disclosure is that data membership is an inherently fuzzy concept. Practitioners typically adopt lossy definitions and tests to operationalize fuzzy concepts, which inevitably leaves room for ambiguity or even malpractice in data transparency. Consider the recent U.S. district court ruling granting data owners permission to "inspect" developers' training data for unauthorized content use (Tremblay v. OpenAI 2023). A non-technical inspector or auditor might simply perform a substring search—implicitly using n-gram overlap as a membership test—while their true intent is to identify semantically equivalent use (e.g., typos, paraphrasing, or even multilingual translations). However, Liu et al. recently demonstrated that LLMs could synthesize and regurgitate text without including any of its original n-grams from the training data; LLMs effectively "stitch together" fragments due to strong generalization. In such cases, any disclosure mandates based on n-gram overlap may be circumvented by model developers. Potential research directions include:

- Data membership definitions/ tests that do not rely on n-gram overlap, or are robust to text perturbations.
- Probabilistic definitions of membership and as well as probabilistic testing methods.

We include two other directions: **Canary injection and data watermarks**, and **cryptographic approaches for data usage** in the supplementary materials.

A. Other Objectives

Analyses of Environmental Impact In current practice, the size of the dataset used for training provides crucial information on how much compute is used to train the model. Work in model scaling laws associates the predicted performance with the number of total flops, which is derived from the size of the dataset used for training (Hoffmann et al., 2022). Thus, data transparency around the rough size of the dataset used to train a model could give a general estimate of the computational resources, and thus the environmental impacts of training the model. However, the size of the dataset itself does not directly reveal the amount of resources used for the development of the model. The former is because model developers train many iterations of the model and may only release the final version.

Competition between LLM developers Despite the increasing availability of open source model weights (e.g., Llama, DeepSeek, and Qwen models), the actual datasets used to develop these models are not transparent. Little is known about what kind of data allowed DeepSeek's R1 to possess state-of-the-art math and coding skills (Guo et al., 2025). In addition, deals for exclusive access to data (Google, 2024), including test sets (Montti, 2025), are beginning to appear in the data supply chain for AI models. Better transparency in data collection, exclusive data contracts, and ethical data collection practices could actually develop healthy competition between LLM developers and better allow new players to enter the market with competing services.

Evaluation Validity: Test Set Transparency Thus far, our work focuses on datasets themselves as static objects. However, with the growing popularity of inference time computation, the capabilities and behaviors of models are increasingly defined by data generated at test time. For example, given a math question, a model might generate a series of answers and rely on a verifier (e.g., another model or a proof system) to provide feedback in order to arrive at a final answer. Thus, being able to compare the performance of different models in a reliable manner may require transparency about the data that are produced at test time.

B. Research Directions for Computer Scientists for Training Data Disclosure

For many components of compliance and verifiability that we have discussed, technical feasibility comes into play. For instance, given a piece of data, are we able to tell whether a model has actually been trained on it? On a technical level, training data disclosure fundamentally involves the concept of training data membership—or training set inclusion—of a set of examples (Shokri et al., 2016), as well as how to precisely define and verify such membership. This section discusses several topics and potential research directions related to both enabling and verifying training data disclosure. Technical research in these directions would be instrumental for understanding the kinds of policies and regulations are actually enforceable in practice. Our discussions focus primarily on modern AI systems based on large language models (LLMs), though the principles discussed may generalize to other types of AI systems and modalities.

Granularity of data usage A key challenge in training data disclosure is deciding the granularity at which data membership should be reported. Granularity refers to the specificity level at which data usage is tracked and disclosed (Maini et al., 2024). At the coarsest granularity, developers might simply acknowledge that a dataset was used; finer granularity could involve tracking specific documents, paragraphs, or even exact text sequences. More granular disclosure introduces computational overhead but yields more meaningful transparency—for instance, sequence-level reporting might require precise but potentially large lookup tables or n-gram statistics. Potential research directions include:

- Developing compact representations of training examples that enable efficient lookups;
- Studying trade-offs between data privacy, utility of transparency, and computational feasibility across different levels of granularity.

Revisiting definitions for training set inclusion A more fundamental issue with training data disclosure is that data membership is an inherently fuzzy concept. Practitioners typically adopt lossy definitions and tests to operationalize fuzzy concepts, which inevitably leaves room for ambiguity or even malpractice in data transparency.

Consider the recent U.S. district court ruling granting data owners permission to "inspect" developers' training data for unauthorized content use (Tremblay v. OpenAI 2023). A non-technical inspector or auditor might simply perform a substring search—implicitly using n-gram overlap as a membership test—while their true intent is to identify semantically equivalent use (e.g., typos, paraphrasing, or even multilingual translations). However, Liu et al. recently demonstrated that LLMs

could synthesize and regurgitate text without including any of its original n-grams from the training data; LLMs effectively "stitch together" fragments due to strong generalization. In such cases, any disclosure mandates based on n-gram overlap may be circumvented by model developers, allowing data use while evading inspection (see Table 21 of (Liu et al., 2025) for visualization).

For the purposes of data transparency, a key for future research is to develop better membership definitions/tests and data disclosure requirements that emit lower false negatives—that is, when the disclosure reports no data usage, there is likely no data usage. Additionally, future data membership definitions and disclosure requirements should extend beyond simple set membership of text in the raw training dataset to consider data provenance, preprocessing, and other side information accessible during the model training pipeline. Potential research directions include:

- Proposing data membership definitions and tests that do not rely on n-gram overlap, or are robust to text perturbations.
- Investigating probabilistic definitions of membership and as well as probabilistic testing methods.
- Analysis of the advantages and disadvantages of using model-based membership definitions (e.g., using model completions as a test for membership). More broadly, understanding whether data disclosure requirements be based on model behavior, rather than simply on the static training set.

Canary injection and data watermarks Canary injection involves intentionally inserting unique identifiable data points (canaries) into training datasets to verify data use in trained models (Wei et al., 2024). Data watermarking, on the other hand, explores injecting (human-imperceptible) statistical signals into training data that persist in model outputs, which can then aid in data usage detection (Sander et al., 2024). Unlike post hoc verification techniques such as membership inference attacks (Shokri et al., 2016) that depend on (fuzzy) data membership definitions (see above), or self-reported metrics that are exploitable (Zhang et al., 2024; Liu et al., 2025), canaries and watermarks allow for indisputable, side channels for data usage verification.

The key challenges of canaries and watermarks include making them persistent (such that data usage cannot be concealed) and salient (such that data usage is easily detectable for data transparency enforcement). Potential research directions include:

- Developing methods that are robust to adversarial data perturbation aiming to remove the canaries/watermarks;
- Developing methods against intentional or inadvertent post-hoc model sanitization; and
- Evaluating the effectiveness of these methods across diverse LLM architectures and training paradigms.

Cryptographic approaches for data usage. Cryptography also may present opportunities for strong, model-agnostic evidence about data use. Potential research directions include:

- Exploring methods that commit to datasets before training, and subsequently issuing publicly verifiable proofs that every optimization step consumes only data that are consistent with that commitment (Jia et al., 2021; Fang et al., 2023; Abbaszadeh et al., 2024)
- Systems research enabling modern training within trusted execution environments (TEEs)—for instance, sealing datasets within enclaves, with remote attestation certifying enclave identity and code hash before data release (Intel, 2025).
- Efficiency optimization for accelerating existing cryptographic primitives for machine learning workloads, such as fully homomorphic encryption (Zama, 2022).

A key drawback of cryptographic approaches is their dependence on exactness. For instance, hashing training sequences would differ significantly even with minor n-gram variations, despite identical semantics. Therefore, defining appropriate data granularity and training set inclusion tests—as discussed previously—is crucial when considering cryptographic approaches for disclosure. This limitation may restrict cryptographic techniques primarily to the parts of the disclosable data that are otherwise unique in their representation (e.g., metadata, training timestamps).

C. The Emergence of Data Transparency Regulation: A Case Study on AB 2013

In practice, AI regulation, like any other legislation, is the result of a compromise between different stakeholders with different interests. Existing data transparency policies are the artifacts of arguments for data transparency and the push against it. To understand how the disclosure and remediation gap arose, we provide a closer analysis of the process that generated the California Bill AB2013.

C.1. Consumer Protection: A Driver for Data Transparency

Introduced in January 2024, California State Bill 2013 initially defined terms related to artificial intelligence systems, developers, and synthetic data generation, along with mandates for disclosing certain aspects of the data supporting model development. Early revisions of the bill extended the bill's scope by widening disclosure to include personal and consumer information, dataset modifications, dataset statistics, and outline exceptions for AI services dedicated solely to security and integrity. The Assembly Committee on Privacy and Consumer Protection emphasized that without knowledge of the data used to train AI products, Californians cannot make informed purchasing decisions. This position was grounded in existing privacy laws. The committee justified the disclosure of synthetic data by pointing out the risks of bias and "model collapse," assessing the proposed requirements as modest. At this stage, supporters of the bill underscored the importance of the bill for public transparency, awareness, and protection for Californians.

Based on an academic report (Chmielinski et al., 2024), the judiciary committee saw data transparency as a way to reducing biases, addressing hallucinations and problematic outputs, and easing privacy and copyright issues. At this stage in the Senate, the California Labor Federation began to support the bill motivated by concerns that AI systems might be used to evaluate employee performance and hiring. The Federation stressed that workers should be made aware of the training data used for these decisions to prevent the non-consensual use of personal data and to tackle both implicit and explicit biases in training data.

C.2. The Push Against Data Transparency

Resistance to the bill emerged early from several key groups expressing doubts about the bill's technical feasibility, vague definitions of key terms, and weak protection of trade secrets and intellectual property. To address these concerns, committee amendments narrowed the definition of artificial intelligence and lowered disclosure requirements from "a description" to "a high-level summary." As the bill reached Assembly Floor Analysis, lobbyists in opposition argued that revealing training data could hinder competition and terms like 'but not limited to' were too vague. As a result, the Senate version included more exceptions for domains where AI systems need not disclose training data (e.g., a GenAI system for operation of aircraft in the national airspace) and relaxed data point disclosures to general ranges.

In the Senate judiciary analysis phase, opposition expanded as a group of organizations raised further criticism mainly focused on two issues: the absence of a risk-level distinction (suggesting only high-risk AI should meet data transparency standards) and the broad interpretation of "artificial intelligence system or service." These concerns shaped subsequent Senate floor amendments, notably narrowing the bill's scope to "generative artificial intelligence" and restricting disclosure obligations to the original developers of a GenAI system, even if modified by a third party.

The forces for and against data transparency illustrates the path from the idea of enabling consumer protection through data transparency to passing a bill with very high-level disclosures. The end product neither ensures that the disclosures are sufficient for the original goal nor provide enforcement provisions to ensure compliance of these disclosures.