The following specification is in response to the NIST AI Executive Order,
with particular attention given to the assignments listed below:

1. Developing Guidelines, Standards, and Best Practices
   for AI Safety and Security
   - Forms of transparency and documentation (e.g., model cards, data cards, system cards, benchmarking results, impact assessments, or other kinds of transparency reports)
2. Reducing the Risk of Synthetic Content
   - Approaches that are applicable across different parts of the AI development and deployment lifecycle (including training data curation and filtering)

MILD: Minimal Item-Level Documentation of Training Data

ABSTRACT

MILD is a training data disclosure specification. MILD protects the intellectual property of AI providers, while fulfilling their transparency obligations to rightholders, consumers, and regulators. Instead of full disclosure, an AI provider shares some minimum properties of each item in the training dataset. Borrowing from industry standards in illegal content identification, the hash property uniquely identifies the item's content: eg. text, image, or audio. Stakeholders can search for matching properties to check whether an item was copied into the dataset. Failure to implement would jeopardize citizens' right to property, to privacy, and to freedom from abuse. This is the minimum disclosure required to detect CSAM in training datasets.

1. OVERVIEW

The past few months have revealed to everyone the incredible risks associated with web-scale datasets, from the dissemination of child-abuse imagery to facilitating abuse of women and teenagers — not to mention commercial-scale infringement.

Standard-setting authorities not only have the opportunity, but also a legal responsibility to establish dataset documentation standards that ensure that criminal behavior does not become widespread in the United States. Failure to adopt the minimal specifications needed could be considered to be facilitating these illegal activities.

AI providers are training models on data at a scale that frequently involves large portions of the world wide web. Stakeholders across society are looking for legal answers and balanced policy solutions.

Artists, writers, guilds, record labels, newspapers and privacy-concerned citizens have filed lawsuits against AI corporations for training large commercial models on billions of items of personal or copyrighted data. At the core, plaintiffs are seeking to uphold their right to consent, credit and compensation.

AI providers want to prevent costly lawsuits and regulatory action. Unfortunately, they found themselves locked into a competitive race of scaling model training and productisation. An industry-wide standard would allow AI providers to compete fairly, without having to cut corners on data rights.

A congressional bill has been filed for an AI Foundation Model Transparency Act, aiming to establish "standards specifying information to improve the transparency of foundation models by covered entities with respect to training data".

For a specification to be comprehensive enough to allow parties to enforce their rights, any copyrighted and personal content must be necessarily disclosed *item-by-item*.

Herein lies a conflict:

1. Parties with legitimate interests have a right to access.
   Authors and publishers have a right to know when their copyrighted works are copied into commercialised datasets, as often used to generate competing synthetic content. Consumers have a right to know about any personal information collected under the California Consumer Privacy Act. Finally,

regulators must be able to check for any illegal content collected by AI providers (such as child sexual abuse materials).

2. Commercial AI providers must also protect their intellectual property.
   If an AI provider publishes all of their in-house content online, competitors can scrape that content and train new competing models. Where the AI provider has invested in curating licensed works into a dataset, the overall content and structure of that dataset can be protected under intellectual property law.

The MILD specification strikes a balance between the interests of all stakeholders. Instead of disclosing all items in the training dataset, AI providers can disclose minimum properties about those items.

This minimum disclosure is effective:

- Regulators can ensure that AI providers fulfill their transparency obligations under a broad set of international and country-specific data laws.
- Rightholders and consumers can uphold their data rights.
- Investors can trust an AI provider to not have hidden massive legal violations, ie. that the AI provider will not get hit by costly lawsuits or regulatory actions.
- AI providers can demonstrate best-in-class support of consumers' data rights.

MILD is cheap to administrate. Disclosing these minimum properties is not only not easy to engineer for, but also obviates the need for staff to follow up on access requests. Other practices, such as narrative explanations of the data sourcing process, force rightholders, consumers, and regulators to request further clarifying details from AI providers. This burdens AI providers with the administrative cost of handling those requests. Data and copyright protection agencies in turn lack the staff to investigate whether AI providers handled access requests in compliance with respective laws.

Allowing anyone with legitimate interests to check whether specific items were copied into the dataset, removes all that administrational burden.

Finally, MILD supports open-source and open-science collaborations.
The inclusion of access information enables open-source sharing of training data (rather than just 'open-weights'). AI researchers and auditors can test for biases and risks (crucial for dataset reports such as Model Cards). The transparent format enables AI researchers to replicate scientific results, and auditing communities to verify audits.

2. TECHNICAL SPECIFICATION

Minimal item-level documentation (MILD) allows minimal disclosure of training data. In the case of text, each book or article has its own entry in the documentation. In the case of images, each photo or artwork has its own entry in the documentation.

On a computer, if you click on a photo or book you can right-click to see its properties such as date, file size and type. In a similar vein, MILD documents a set of properties for each item in the training data. A content code in the form of a "hash" (a fingerprint for a file that is computed cryptographically) can be used to identify the data.

Compiling the properties of each item in the training data creates the documentation. The documentation can be published alongside regulatory-compliant models when it's released as open source, or made available online alongside a compliant service using a proprietary model. Third-party content must be included in the data documentation, and first-party content is optional at the choice of the developer.

At a high-level, the benefits of MILD:

- Enables true open-source for models and not just open-weights.
- Promotes scientific research and progress through clear data documentation.
- Establishes healthy industry standards that significantly reduce legal risks.
- Improves AI safety due to the transparency of how models are trained.
- Fosters a digital market for legally sourced and high-quality data.
- Increases the standards for data sourcing to eliminate illegal content.

Instead of disclosing all items in the training dataset, AI providers can disclose minimum properties of those items. Two kinds of properties must be disclosed:

1. A content hash, allowing automated identification of the type of content by rightholders, consumers, and regulators.

2. Information about the method through which the item was acquired:
    a. In case of licensing:
    author attribution, as well as any other copyright management information the license requires disclosure of.
    b. In case of collection of unlicensed content:
    specific access information, such as a source url and date of retrieval.

Table 1: A list of disclosable properties for each item in the training data

item_title	The name, title, or caption of this item.
item_size	Number of bytes on disk used by this item.
item_copyright	The author of this item or other CMI
content_type	The mime-type of this item, e.g. image/jpeg.
content_code	A standard content code such as ISCC, encoded in string format (e.g. base58).
content_checksum	A content hash that cryptographically encodes remaining information (e.g. sha256).
source_domain	The domain or company name where this item was licensed.
source_url	The page on which this item was found, *if it is public*.
source_cdn	The domain on which this item was hosted, *if public.*
access_time	A timestamp of the moment when the item was accessed.
access_basis	A shortcode that explains the legal basis for accessing the item.

3. EXAMPLES

These are examples of how higher-quality datasets from the community have features that overlap with MILD in their metadata and documentation. We also show how MILD specifications would help with problematic datasets too.

Positive examples:

- FFHQ
    - Description: A dataset of Faces from Flickr in high-quality (1024x1024).
    - Feature: Item-level metadata for the images with their URL and license.
- Project Gutenberg
    - Description: A repository of public domain books in standard format.
    - Feature: Provides ebook metadata that describes each item.

- Segment Anything
  - Description: Model uses a licensed dataset of 11m images.
  - Feature: Clear licensing from a photo provider provides legal clarity
- NYT Articles
  - Description: Dataset of old articles provided by the New York Times.
  - Feature: Full metadata is available for each article.
- Stability Audio
  - Description: Dataset of licensed music files from audio providers.
  - Feature: Legal clarity for the model licensing and usage.

Negative examples:

- MS Celeb
  - *Problem:* Dataset features personally-identifying information (PII) acquired without consent nor licenses.
  - *Outcome:* Microsoft voluntarily took the dataset down because it violated privacy and publicity rights.
  - *Fix:* MILD would disclose the source URL to allow identification of the people, to ensure they are informed sooner and can take action quicker.
- LAION-5B
  - *Problem:* Scraped dataset that was negligently filtered for illegal content with a substandard technical approach.
  - *Outcome:* Organization was pressured to be taken down because of illegal content (CSAM) being disclosed in the press.
  - *Fix:* The hashes in MILD would allow for easy identification of illegal content, which is a standard approach.
- books3
  - *Problem:* Dataset of books retrieved from pirate websites without license.
  - *Outcome:* Forced to be taken down by takedown request under copyright.
  - *Fix:* The hashes in MILD allow rightholders to detect book content and take action sooner.