A Common Pool of Privacy Problems: Legal and Technical Lessons from a Large-Scale Web-Scraped Machine Learning Dataset

Rachel Hong[1], Jevan Hutson[2], William Agnew[3], Imaad Huda[2], Tadayoshi Kohno[1], Jamie Morgenstern[1]

ABSTRACT

We investigate the contents of web-scraped data for training AI systems, at sizes where human dataset curators and compilers no longer manually annotate every sample. Building off of prior privacy concerns in machine learning models, we ask: What are the legal privacy implications of web-scraped machine learning datasets? In an empirical study of a popular training dataset, we find significant presence of personally identifiable information despite sanitization efforts. Our audit provides concrete evidence to support the concern that any large-scale web-scraped dataset may contain personal data. We use these findings of a real-world dataset to inform our legal analysis with respect to existing privacy and data protection laws. We surface various privacy risks of current data curation practices that may propagate personal information to downstream models. From our findings, we argue for reorientation of current frameworks of "publicly available" information to meaningfully limit the development of AI built upon indiscriminate scraping of the internet.

KEYWORDS

Empirical Studies, Data Protection, Artificial Intelligence, Ethics, Web Scraping, Dataset Audit

1 INTRODUCTION

With the recent popularity in foundation models like ChatGPT and Midjourney [87, 100], machine learning practitioners often rely on data scraped from the web to train large language or vision models [21, 112, 124]. DataComp CommonPool, for instance, is one of the largest publicly available image-text dataset scraped from the web with over 12.8 billion samples [47]. This dataset has been downloaded over 2 million times at the time of writing with half a million downloads in the month of October 2024 alone [45], and its precursor LAION-5B [116] was used to train well-known image generation models like Midjourney, Stable Diffusion, and Google's Imagen [6, 87, 114]. Since machine learning models are a function of their training data, the downstream models trained on DataComp CommonPool may share problematic behavior [17], including the potential leakage of personally-identifiable information (PII) [25, 92]. Just as prior work highlights the importance of data-centric AI governance [56], we emphasize that regulating a dataset with such wide usage may be more effective than addressing the harms of every model one-by-one – in other words, tackling the "root" rather than the "leaves" as illustrated in Figure 11.

In our work, we use DataComp CommonPool as a case study of web-scraping and conduct an investigation into data privacy concerns. We perform a *legally-grounded audit*, one of the first to our knowledge, in which our audit findings inform our legal analysis on web-scraping, and vice versa, where recent legal literature on data privacy motivates our audit inquiries [68, 124]. Specifically, our audit asks: *What kinds of personally identifiable information are present in DataComp? How do current data cleaning practices address privacy concerns?* To do so, we draw upon prior frameworks on privacy [88], representation [38], and data filtering [64].

Our legal analysis considers how use of DataComp CommonPool for AI development might trigger application of and compliance obligations under existing privacy laws for developers and downstream deployers, including US state comprehensive privacy laws and international data protection laws. We consider and problematize current interpretations of "publicly available data" under existing privacy laws. We also consider how privacy risks and compliance obligations triggered by the production of DataComp CommonPool propagate to downstream models trained on this dataset. Lastly, we consider ongoing privacy risks that are currently not being addressed sufficiently by data filtering and other responsible data curation and hygiene practices, which informs recommendations on how policymakers might address these risks.

We make the following contributions:

(1) We find instances of personal information present in DataComp CommonPool, revealing various privacy concerns in web-scraped image-text datasets. For example, we uncover examples of personal information including credit card numbers and passport numbers, and we estimate at least 142,000 images depict resumes of individuals.
(2) We argue that no automated cleaning of web-scraped data can remove all PII and that ongoing cleaning methods are not sufficient to tackle privacy and must be scrutinized. Specifically, the DataComp CommonPool creators use a face blurring tool to preserve privacy, and we find that this tool fails to catch an estimated 102 million images of real human faces, demonstrating the importance of privacy tool assessments.
(3) We map these audit results to legal concerns to provide a critique of current data curation practices according to existing privacy laws. We also apply our findings from this widely used dataset to demonstrate shortcomings of existing privacy frameworks, such as the implications of the exemption for publicly available information.

We first present the context for web-scraped machine learning dataset development by detailing the history of DataComp CommonPool in Section 2, the stakeholders and artifacts associated in each step of the curation pipeline in Section 3, and related computer science and legal work in Section 4. We then present our audit methodology in Section 5 and the empirical results in Section 6. We use these findings to inform our legal analysis to determine the application of various data protection laws in Section 7. Finally in

1. University of Washington Paul G. Allen School of Computer Science & Engineering.
2. University of Washington School of Law.
3. Carnegie Mellon University Carnegie Bosch Institute.

Section 8, we integrate the concerns revealed by our audit together with the shortcomings of existing privacy frameworks to discuss normative arguments for both policymakers and machine learning practitioners.

2 DATACOMP COMMONPOOL

In April of 2023, Gadre et al. [47] released DataComp CommonPool, a publicly available image-text dataset of 12.8 billion samples collected from the web, as part of the DataComp testbed for assembling datasets to train more effective large image-text models like CLIP [109]. In this section we describe the curation process for the DataComp CommonPool dataset, describe the measures the curators took to protect privacy of individuals in the dataset, highlight the dataset's usage after its release, and situate CommonPool in relation to its predecessor LAION-5B.

2.1 Curation process

The steps to build CommonPool are as follows [47]:

(1) **Gather**: The curators first gather web snapshots from 2014 to 2022, relying on Common Crawl as the data source, which is a nonprofit organization that crawls the entire web to form unformatted web dump archives [32].
(2) **Extract**: The image URLs and accompanying alt-text (alternative text attached to the image for accessibility purposes [23]) are extracted from the web snapshots and deduplicated. The alt-text is referred to as "captions" for the associated images.
(3) **Download**: The images are then downloaded from the URLs, resulting in 16.8 billion successfully downloaded samples at the time of curation.
(4) **Filter**: Next, several toxicity filters are applied in order to discard NSFW-detected images or text. In addition, a face detection algorithm is applied to annotate bounding boxes of faces in the images, as detailed in Section 2.2.1.
(5) **Deduplicate**: Finally, the images are inspected and deduplicated from evaluation sets, resulting in 12.8 billion image-text pairs that comprise DataComp CommonPool.
(6) **Release**: Rather than releasing the image content, the curators release CommonPool as a table where each sample consists of an image URL, the associated text, and additional metadata (image size, image hash, etc). To acquire the dataset, the release is accompanied with a code repository for users to run a script which instantiates a crawler that automatically downloads each image from its URL [34].

2.2 Privacy mitigations

In the CommonPool datasheet [50], the dataset curators disclose that due to its scale and internet sources "it is highly likely that there is sensitive data in the dataset" including identifying information. Therefore, they engage in several mitigations to "prevent making sensitive content more accessible" [47].

2.2.1 Face obfuscation. To address privacy concerns, CommonPool is released with face detection annotations, so that the dataset download script by default hides any detected face in the image via a Gaussian blurring method [139]. To create these annotations, the curators apply the SCRFD algorithm [55] to obtain bounding boxes for detected faces in each image. It is plausible that SCRFD is chosen due to its efficiency and lack of cost — the CommonPool curators compare SCRFD to the commercial system Amazon Rekognition and find that SCRFD has worse precision and recall (75.87% & 90.53%) than Rekognition (86.09% & 93.75%). The curators evaluate image-text CLIP embedding models [109] trained on CommonPool with and without blurred faces and demonstrate similar model performance on their evaluation benchmarks [47]. In the released artifact, the face bounding boxes thus accompany each URL-caption sample as metadata, which also inadvertently allows any user to extract faces in the dataset.

The face blurring is optional, however, as a dataset user downloading CommonPool can easily turn off face blurring through specific parameters [34]. In addition, models trained on CommonPool with blurred faces are able to zero-shot classify race and gender at rates significantly better than random chance [47]. The CommonPool curators speculate that models are still absorbing sociodemographic information outside of faces, or that face blurring does not capture all human faces which we evaluate in our work.

2.2.2 Opt-out mechanisms. Hugging Face, the dataset distribution platform that hosts the CommonPool URL-caption pairs, integrates with Spawning AI, a tool that allows users to search and remove their personal information from the dataset [7]. However, as described in Section 4.3.1, these opt-out policies are often not considered meaningful consent since users must first know of the presence of personal information in the first place and then put in effort to find and remove it [123]. Another mechanism which provides some attempt at privacy protection is Robots Exclusion Protocol, where a website attach a robots.txt file to instruct web crawlers what assets they can access [127]. However, this protocol is not legally enforceable, and AI crawlers have recently been accused of not respecting robots.txt [104]. Since 2023, site hosts have modified their robots.txt files to restrict scraping for AI development, signaling some intention of site content to be kept private from training models (in addition to intellectual property concerns) [81]. While Common Crawl does respect robots.txt [33], CommonPool relies on snapshots from 2014 to 2022 [47] to aggregate URLs, before many sites began these restrictions. According to a developer of the code package used to download CommonPool, the downloading step at the time of writing does not respect up-to-date robots.txt website-level protocols due to practicality reasons, although they do respect image-level robots tags to not crawl [52], which we elaborate further in Section 6.3.2. Certain site hosts may have changed their preferences to prevent crawling across the site (and thus see no reason to update image-by-image tags), yet these preferences are not followed at the site-level.

2.3 Usage

According to Hugging Face, the DataComp datasets (which include CommonPool and smaller subsets) at the time of writing have been downloaded 2 million times, with half a million downloads in the month of October 2024 alone [45]. The CommonPool curators explicitly state that the dataset is intended for academic research and do not condone the dataset being used to train deployed models [47]. However, the URL-text pairs that comprise CommonPool are

released with a CC-BY-4.0 license, which does not prohibit anyone from using the dataset for commercial purposes [30]. The license is also specific to the URL table, rather than the images assets themselves.

Because companies that develop models often do not disclose their training data [57], it is unclear if any popular commercially deployed models have trained on CommonPool. However, we highlight in Section 2.4 that the curators themselves acknowledge that CommonPool has substantial overlap [35] with the LAION-5B dataset, which has been used to train models like Stable Diffusion, Midjourney, and Imagen [6, 87, 114]. We also observe that in several issues published to the dataset Github repository, the posters mention downloading the dataset on behalf of a company [53] — which is unclear if the dataset is being used for research or commercial deployment. DataComp's Github repository has about 700 stars (number of users who have marked it as a favorite) [34], while the original paper has about 400 citations. These numbers, however, do not cover the sheer number of overall dataset downloads.

2.4 LAION-5B

DataComp CommonPool was intended as a follow-up to the LAION-5B dataset released the prior year [116], and follows many of the same data curation steps, originating from Common Crawl and releasing a similar collection of URLs rather than the image assets themselves. LAION-5B is a collection of 5 billion image-text pairs that DataComp authors state have substantial overlap with CommonPool due to reliance on the same data source [35]. In December 2023, the LAION-5B dataset was taken down after Thiel [130] found presence of Child Sexual Abuse Material (CSAM) in the dataset [28]. In August 2024, the dataset developers subsequently removed links to the detected CSAM with the release of Re-LAION 5B [72].

2.4.1 Privacy mitigations. In their paper, the LAION-5B authors highlight issues of PII present in the dataset. They argue their accompanying tool `CLIP retrieval`, which enables text search of LAION-5B images, grants users the ability to find their own personal content potentially present, in order to initiate takedown procedures from the dataset or website hosting provider. While the public availability of this search tool may raise awareness about the content of web-crawled data, the tool can also be used by adversaries to gather personal information; we expand on these concerns in our discussion on profiling Section 7.3.7. Currently the CLIP retrieval website is no longer accessible, but code is available to run the tool locally [11]).

Moreover, the LAION-5B authors place the responsibility on the individuals to find and remove their personal information, yet these opt-out policies again are not very meaningful (Section 4.3.1). For instance, when someone found their medical records leaked on LAION-5B and wished to take them down, a LAION author responded that the hosting website was responsible since the dataset curators "are not hosting any of these images" [43]. As we discuss in Section 8.2, content on the internet propagates and becomes incredibly hard to regulate past the time of upload.

2.4.2 Usage. Originally intended for research, the LAION-5B dataset also has had substantial usage to train commercially deployed models, despite authors explicitly advising "against any applications in deployed systems without carefully investigating behavior and possible biases of models" [116]. Midjourney, Stable Diffusion, and Google's Imagen all disclose training on subsets of LAION-5B, and these models have over millions of users [6, 87, 114]. Recent work has shown that Stable Diffusion and Imagen are subject to training reconstruction attacks, in which supplying captions from the training dataset into these models can generate images almost identical to training examples [25]. In addition, researchers have been able to fine-tune Stable Diffusion models to reconstruct training images without access to the captions [79]. The wide impact of models trained on LAION-5B subsequently demonstrates that publicly-available datasets can be used without regarding the dataset curator's original intent.

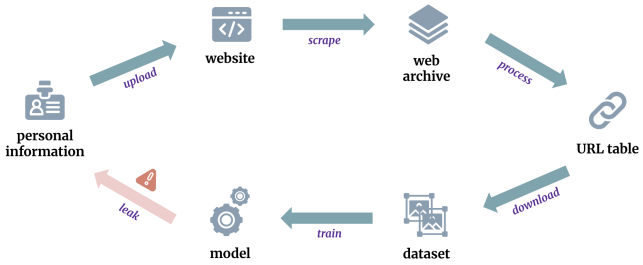


Figure 1: Data lifecycle of how personal information appears in various artifacts in the machine learning pipeline. First, *personal information* may be uploaded to a *website*, which then is scraped and aggregated into a *web archive*. Then, the web archive is cleaned and processed into a *URL table* consisting of text and a link to an image. The table is then downloaded to a *dataset* which then is used to train a machine learning *model*. The model may then be deployed and potentially leak personal data through memorization.

3 STAKEHOLDERS

In this section, we situate web-scraped datasets like CommonPool within the broader machine learning pipeline and provide frameworks to define the various processes and actors that influence or are influenced by large-scale web-scraped training datasets. Within the pipeline, personal information may be encoded in different forms, from a webpage to a machine learning model. In Figure 1, we demonstrate the lifecycle for personal information that may trickle into datasets like CommonPool. First, *personal information* is uploaded to a *website*, which then is scraped and aggregated into a *web archive*, a snapshot of the entire internet obtained through web crawling. Then, the web archive is cleaned and processed into a *URL table* consisting of text and a link to an image. The table is then downloaded to a *dataset* which then is used to train a machine learning *model*. The model may then be deployed and potentially leak personal data through memorization [25].

With the data lifecycle in mind, we now define the various stakeholders that are involved with the content of machine learning datasets, and in our case, focusing specifically on image-text data from the web. We follow the roles from Khan and Hanna [68] but also add additional actors in relation to DataComp CommonPool,

Internet	Dataset	Usage
Data subject	Web archiver	Dataset user
Data owner	Dataset curator	Model user
Data uploader	Dataset annotator	Model subject
Site host	Dataset distributor	

Table 1: Overview of stakeholders by stage of data lifecycle. We demonstrate how stakeholders interact with each other in Figure 11.

of whom we argue are important players that have their own incentives and consequences. As displayed in Table 1, the stakeholders are separated into three stages: the Internet, the Dataset, and the Usage. None of these stakeholders are mutually exclusive from each other, as there may be substantial overlap.

3.1 Internet

In this stage, the actors interact with data on the internet, completely divorced from any expectation the data will be used for downstream applications.

3.1.1 Data subject. The data subject is the individual who the data is about. For instance, this may be the person whose face is in the image, or the person whose address is in the caption of the photo. Privacy laws often center privacy in relation to the data subject due to the presence of their personal information.

3.1.2 Data owner (copyright holder). The data owner is the person who typically creates the image (and thus likely owns the copyright to the image [68]) – this might be a photographer who takes a photo of the data subject. Data ownership may be transferable, so the current owner may not have created the data in the first place.

3.1.3 Data uploader. The data uploader is the person who uploads the data to the website, which, for instance, could be the photographer's company who uploads a picture online for marketing purposes. There can be a distinction here between data uploader and data subject: the data subject may have no knowledge nor given consent to their personal information uploaded onto the web by the data uploader.

3.1.4 Site host. The site host is the maintainer of the website that hosts the data once uploaded. They determine how the data is depicted and accessed, as well as setting the terms of service for how the data can be automatically scraped.

3.2 Dataset

This stage includes actors that are a part of the dataset creation process, to be released as a public artifact explicitly intended to train machine learning models.

3.2.1 Web archiver. The web archiver is the entity that crawls the entire internet to aggregate data in one place as a web archive. In CommonPool's case, this is Common Crawl, the nonprofit organization that provides the data source of the same name. These web archives are typically unformatted and not usable as a training dataset.

3.2.2 Dataset curator. The dataset curator is the actor that compiles the dataset based on the web dump from the web archiver. The curator must establish a process to select and format data from the web archive, in order to output a dataset suitable for training models.

3.2.3 Dataset annotator. The dataset annotator is the person who processes or adds relevant metadata to the data. The annotator may tag information manually, or rely on automated methods built by others. For CommonPool, the data annotator is equivalent to the dataset curator — for instance, they rely on a face detection algorithm to mark the presence of faces. This is not always the case, as dataset compilers may outsource annotation instead [74].

3.2.4 Dataset distributor. The dataset distributor is the entity that is in charge of hosting and distributing the dataset as an artifact for others to download. CommonPool is hosted by Hugging Face, a popular dataset distribution platform.

3.3 Usage

This stage describes the players involved in the downstream usage of the dataset once it is released.

3.3.1 Dataset user (model developer). The dataset user is someone who downloads the machine learning dataset with the intention of processing that data. This might be to develop a model, but could also be for other purposes, such as reformatting to produce another dataset, or searching through the dataset.

3.3.2 Model user. If a machine learning model is developed and released by the dataset user, then the model user is the person who interacts with the model. Given the general-purpose design of foundational models like text-to-image generators, there are many potential use cases for a given model.

3.3.3 Model subject. The model subject, as defined in Khan and Hanna [68], is the person who the model makes decisions about, which may have consequences on the person's life. The model subject may also be equivalent to the data subject of an image in the model's training dataset, but not necessarily. For example, the model subject may be a job applicant who submits an application screened by a machine learning model.

3.4 Stakeholder network

We next map the landscape of web-scraping to build large-scale machine learning datasets in Figure 2. We illustrate how stakeholders interact with each other to pass personal information from the *data subject* to the *model user* (and model subject). The diagram first depicts the various paths between *site hosts* who may scrape and reupload personal information after its original upload by the *data subject* (or data uploader). The *dataset curator* (along with other actors in the Dataset stage) then aggregates data from the site hosts into a funnel, which then is dispersed to *dataset users*, and further dispersed to *model users*. The number of connections between the *dataset curator* and downstream players is enormous, hence demonstrating the "leaves" reliant on the centralized source.

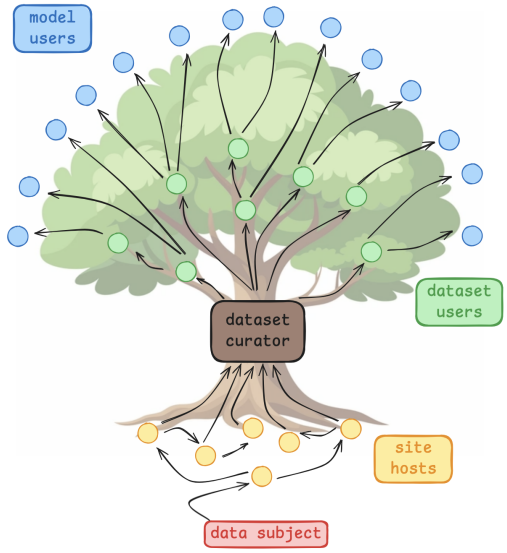


Figure 2: A high-level depiction of how personal information can flow between actors and is aggregated by the *dataset curator* and then dispersed to various *dataset users* and *model users*. A detailed diagram is displayed in Figure 11.

4 RELATED WORK

In this section, we highlight prior work both in the computer science and legal disciplines.

4.1 Data collection

We draw upon many prior audits of datasets to inform our approach. Birhane et al. [16] provide a comprehensive overview of the AI audit ecosystem and analyze the priorities of existing data audits. On the web-scraping side, Dodge et al. [39] use keyword techniques and URL analysis to understand the text and websites within Common Crawl. Recent work has also audited the license and website terms of service restrictions of web-scraped machine learning datasets [80, 81]. Several works examine the LAION-5B and DataComp data collection processes, mostly revolving around sexually explicit content, toxicity, and bias [13, 15, 64, 130]. Díaz et al. [38] provide a general framework to comprehensively assess representation in unstructured data like images; our work further explores the first component of their framework by examining the presence of people in unstructured datasets.

There has also been extensive work examining data curation practices from a sociotechnical lens. Paullada et al. [105] provide a broad survey of dataset collection in machine learning research, while Desai et al. [37] incorporate archival studies to examine datasets — both of which argue the need to analyze the contents of datasets and choices in assembling them. To do so, recent works have traced the values, assumptions, politics, and histories of machine learning datasets [14, 36, 115]. We are inspired by these critiques in our work to better understand and define web-scraping practices.

4.2 Privacy concerns

Researchers have demonstrated that machine learning models are susceptible to leak information of training data due to memorization during training [20]. Various attacks, for instance, can extract potentially personal information from large language models [82, 92], while on the vision side, Carlini et al. [25] generate over a thousand training examples at test time from various diffusion models. Fine-tuning diffusion models can also amplify the leakage of memorized training samples [79].

To understand the privacy risks associated with a model that can output its training data, it is necessary to determine the private information present in a dataset in the first place. Dou et al. [40] develop a taxonomy of various online self-disclosures, where users communicate personal information in text, and Mireshghallah et al. [88] use this taxonomy to find presence of personal information in user interactions with ChatGPT. For vision, most work has focused around image classification tasks like face detection or license plate detection to automatically blur regions to preserve privacy [139, 142]. Other work has focused on manually annotating more nuanced privacy concerns present in images to train models to identify privacy risks [101, 102]. We incorporate several of these personal information detection techniques to understand the risks of DataComp CommonPool.

4.3 Legal literature review

Large-scale web-scraped datasets raise long-standing privacy issues in new forms. Scholars have noted that many privacy risks posed by artificial intelligence (AI) and big data are not fundamentally novel, but rather "remixes" or amplifications of existing problems [122].

4.3.1 Individual control. A key concern is the inadequacy of the prevailing individual control model of privacy, often termed "privacy self-management." [121] Under this model, individuals are expected to read notices and consent to data practices (or opt out), supposedly empowering them to manage their own privacy, like the mechanisms detailed in Section 2.2.2 and Section 2.4.1. In practice, this notice-and-consent regime has been widely critiqued as ineffective and even "farcical" [58, 63, 111, 131, 136]. Solove argues that people cannot meaningfully control personal data in an AI-driven and data-intensive environment: the scale and opacity of data collection and algorithmic processing far exceed individuals' ability to understand or consent [122]. Privacy self-management, as Solove put it, is beset by a "consent dilemma," individuals face too many notices and hidden inferences to practically make informed choices [121]. This critique, echoed by experts, underscores that reliance on individual consent is an inadequate safeguard in the age of web-scraped AI datasets [122–125].

4.3.2 Inference. Compounding the consent problem is the issue of data inference and generation. AI systems not only collect vast amounts of data as input, but also generate new data about individuals via inference [126, 133, 135]. Machine learning algorithms can predict sensitive facts about people that they never directly revealed, blurring the line between data "provided" and data "produced." Solove observes that inference allows companies to end-run

traditional privacy protections: laws typically regulate the collection of personal data from individuals, but if algorithms can create personal data (e.g., a prediction of someone's pregnancy or political views) from other information, those laws offer little direct control [122]. A famous example is Target's analytics identifying a teen's pregnancy from mundane purchasing data. Such inferences "upend the traditional picture" of privacy management, because people cannot anticipate or prevent the creation of sensitive data about them [122]. Scholars like Solow-Niederman [126] have termed this the "inference economy," noting that individuals have no practical way to opt out or correct the myriad inferences drawn about them. Even inaccurate inferences can be harmful, yet privacy frameworks offer little recourse for data generated without one's knowledge. This literature suggests that privacy law must expand its focus beyond the moment of data collection to address downstream inferencing and profiling [126, 133, 135].

4.3.3 Data minimization and purpose limitation. Another body of relevant work concerns core data protection principles, such as data minimization and purpose limitation, and how they clash with the big-data practices behind massive datasets. Data protection laws traditionally require that organizations collect only the personal data that is necessary for a specified purpose, and not repurpose it for incompatible uses. Large-scale AI datasets assembled via indiscriminate web scraping flout these principles dramatically [122, 124]. Solove and Hartzog argue that scraping "ignores" virtually all fair information practice principles: data is collected broadly without notice or consent, without a specified purpose, retained without regard for necessity, and used for new purposes (AI training) never contemplated by the original context [122, 124]. This critique builds on previous scholarship recognizing that "Big Data" often demands maximal collection and open-ended future use, directly at odds with minimization and purpose limitation. For example, Tene and Polonetsky [129] observed that organizations were beginning to effectively "collect everything, just in case" to exploit the potential of data, making it difficult to honor purpose restrictions or deletion obligations. DataComp CommonPool's design — billions of items gathered "just because" they may improve machine learning, exemplifies what Solove and others identify as a profound tension between big-data analytics and the foundational privacy principle of collecting the least data needed. In effect, large web-scraped datasets treat personal data as an unlimited raw resource, whereas privacy scholarship insists on data frugality and contextual integrity.

4.3.4 Publicly available data. Crucially, researchers have challenged the assumption that "publicly available" personal data are free of privacy interests. Many web-scraping efforts defend themselves by noting that the data was already public on the Internet. However, interdisciplinary scholarship has long rejected a simplistic "secrecy paradigm" which equates privacy solely with complete secrecy [59, 97]. Helen Nissenbaum's theory of contextual integrity, for instance, posits that privacy is defined by appropriate information flow within context-specific norms, not by whether information is public or private in an absolute sense [96]. It is well established that people maintain privacy expectations even in public arenas. They may share information on a personal blog or forum for a specific audience or purpose, yet still reasonably object to that data being mined en masse for unrelated uses. As Solove explains, individuals often disclose personal data in limited settings; privacy encompasses the ability to limit the audience and purpose of that disclosure [122].

The notion of privacy in public is supported by legal scholars as essential for freedom and democracy [59, 97, 110]. One key concept is "practical obscurity": even if data is technically accessible, it may be difficult to find, scattered, or fleeting, which gives individuals a measure of obscurity that protects their privacy [60]. When scrapers aggregate and centralize such data, they destroy this practical obscurity – effectively a privacy loss even though the data was public before [122]. Hartzog [59] has termed the naive belief that publicly available data is harmless the "public information fallacy." Indeed, privacy law itself historically recognizes interests in public information (for example, the tort of appropriation protects against misuse of one's public name or likeness). The emerging consensus in scholarship is that the context and method of data use matter: personal data scraped from the open web is not per se exempt from privacy concerns. Solove [122] and Hartzog [59] both argue that privacy frameworks must "safeguard obscurity" and place limits on the unfettered harvesting of personal data from the internet.

4.3.5 Web scraping. Finally, there is growing interdisciplinary work examining web scraping and privacy harm. Solove and Hartzog [124] characterize web-scraping as a direct clash with privacy norms. They document how companies like Clearview AI have scraped social media photos to build facial recognition databases, actions that regulators around the world deemed unlawful and harmful to privacy. The Clearview incident, resulting in multimillion-dollar fines for violating data protection laws, is frequently cited as a cautionary example of treating "public" personal data as fair game. In the AI context, massive text and image datasets have been assembled by scraping platforms like Twitter, Reddit, Flickr, and personal websites without consent. This practice has been denounced in legal and ethics literature for sidelining individual autonomy and data subject rights. Scholars emphasize that scraping undermines nearly every element of the modern privacy toolkit: individuals typically do not receive any notice their data is taken, do not consent, cannot opt out, and often cannot even exercise rights like deletion or correction because the scraper may remain unknown to them [124]. In summary, the relevant literature paints a stark picture: large-scale dataset compilers are operating in a legal and ethical gray zone, relying on outdated notions of public data and consent. Foundational principles (data minimization, purpose limitation) are being overridden by the imperatives of "more data at any cost." Commentators call for a reconceptualization of privacy law to address these challenges, shifting away from exclusive reliance on individual consent, and imposing accountability on data collectors and AI developers to respect privacy constraints even when dealing with public or inferred data.

4.4 Privacy law review

4.4.1 EU General Data Protection Regulation (GDPR). The GDPR provides a comprehensive privacy framework that is highly relevant to web-scraped datasets. The GDPR applies to any "controller" or "processor" who processes personal data in the context of an EU establishment, or who processes data of individuals in the EU

for purposes of offering them goods/services or monitoring their behavior (Article 3). This broad jurisdictional reach means that even non-EU entities can be subject to GDPR if they scrape or use personal data from EU residents in a way that qualifies as monitoring or offering services. In practical terms, if the CommonPool dataset includes information about EU persons (highly likely given its web-scale), any entity using that data in a manner targeting the EU or involving EU operations would need to comply with GDPR requirements. There is no monetary or size threshold for GDPR coverage – it applies irrespective of company size, so long as the activity falls within its scope.

The GDPR defines "personal data" very broadly as "any information relating to an identified or identifiable natural person." This definition easily encompasses CommonPool's contents: for example, an image of a person's face or a snippet of their resume "relates to" an identifiable individual (even if names are not explicitly included, identifiability can be inferred from context or by combining data). Notably, unlike some U.S. laws, the GDPR does not exempt publicly accessible personal information from its scope, if the data relates to an individual, it is protected, regardless of source. Even then, processing must occur on a proper legal basis. The GDPR also recognizes certain categories of "special" (sensitive) personal data that merit heightened protection. Article 9 enumerates sensitive data such as racial or ethnic origin, health information, biometric data processed for identification purposes, sexual orientation, political or religious beliefs, and information about children. Processing these special categories is generally prohibited unless a specific condition is met (such as explicit consent), including the condition of "personal data which are manifestly made public by the data subject," but this is a narrow carve-out applicable mainly to the data subject's own deliberate public disclosures. This condition does not cover all special category data in the public domain. It only covers personal data that the individual themselves has made public. In the context of DataComp, any photos revealing race or health traits, biometric identifiers (faces used for recognition), or data about children would fall under these special categories, requiring rigorous justification.

Another core concept in GDPR is purpose specification and limitation: personal data must be collected for "specified, explicit and legitimate" purposes and not further processed in incompatible ways (Article 5(1)(b)). Similarly, the principle of data minimization mandates that only data which is "adequate, relevant and limited to what is necessary" for the stated purpose should be collected (Article 5(1)(c)). These principles directly speak to the DataComp scenario: repurposing people's information from the web for AI training (a new purpose) would typically require a fresh legal basis, and collecting 12.8 billion data points "just in case" would seem to violate the necessity limitation. However, the GDPR does include some contextual exceptions: for example, purely personal or household use of data is exempt (Article 2), and there are allowances for scientific or statistical research that might relax certain obligations (with strict conditions and safeguards).

Overall, GDPR sets a high bar: it prescribes legal grounds for processing (consent, contractual necessity, legal obligation, vital interests, public interest, or legitimate interests – Article 6), requires transparency to data subjects (Articles 13–14), grants individuals robust rights (access, deletion, objection, etc.), and mandates security and breach notification (Articles 32–34). If a web-scraped dataset contains personal data, a GDPR-regulated entity handling it must navigate all these obligations, unless the data can truly be anonymized such that no individual is identifiable (a standard the law and EU regulators interpret very strictly).

4.4.2 California Consumer Privacy Act (CCPA, as amended by the CPRA). California's privacy law is the first comprehensive state privacy regime in the U.S. The CCPA, as amended by the CPRA, imposes obligations on covered "businesses." A business is defined generally as a for-profit entity doing business in California that meets certain thresholds: (a) annual gross revenues over 25 million; or (b) buys/sells or shares personal information of 100,000 or more California consumers or households; or (c) derives 50 percent or more of annual revenue from selling or sharing personal information. These criteria limit the law's application to larger data handlers. For example, an academic or non-profit entity that compiled CommonPool might not be a "business" under CCPA, but a large tech company downloading and using it likely would be. The CCPA grants California residents rights over their personal information held by businesses, including the right to know what data is collected, to delete data, to opt out of its sale or sharing, and to non-discrimination for exercising rights. The CPRA added a right to correct inaccurate data and to limit use of "sensitive personal information."

The CCPA defines "personal information" in expansive terms but explicitly carves out "publicly available" data. Under the CCPA (2018) and its amendment via the California Privacy Rights Act (effective 2023), personal information means any information that "identifies, relates to, or could reasonably be linked with" a particular consumer or household. This would include typical identifiers (names, emails), internet activity data, biometric information, geolocation, and even inferences drawn about preferences or characteristics. Crucially, however, the CCPA's coverage excludes "publicly available information." Initially, "publicly available" was narrowly defined to mean data lawfully made available from government records. The CPRA expansion broadened this definition: now it also includes information a business has a reasonable basis to believe was lawfully made public by the individual or through widely distributed media. In other words, personal data that the consumer themselves made public , or which is in public news or media, may fall outside CCPA's definition of regulated personal information. This could potentially exempt large swathes of DataComp content; for instance, images and text that people posted on public forums or social media without restrictive privacy settings might be considered "publicly available" under California law. Notably, the law excludes data that the consumer has restricted to a specific audience, and it excludes usage that is not aligned with the data's purpose of publication. Under the law, "publicly available" also does not mean biometric information collected by a business about a consumer without the consumer's knowledge. These nuances mean the exemption is not absolute.

Sensitive personal information under CPRA is a subset of personal data including items like Social Security or driver's license numbers, financial account details, precise geolocation, racial or ethnic origin, union membership, contents of private communications,

genetic data, and biometric identifiers, among others. For such sensitive data, businesses must disclose if they collect it and honor consumers' requests to limit its use to what is necessary to provide the requested services. In the DataComp context, any scraped data like full credit card numbers, government IDs, or precise location coordinates would qualify as sensitive; California consumers could demand that businesses cease using those items beyond core functions. It's important to note that the CCPA's publicly-available exemption also applies to sensitive data: e.g., a person's publicly posted phone number might not be protected as "personal information," and likewise their race or religion if obviously made public by them could be deemed "public." However, the presence of children's data triggers additional rules: the CCPA requires opt-in consent (through a parent for under 13, or the minor's consent if 13–15) before selling personal information of minors. While DataComp researchers are not "selling" data, any downstream commercial use of children's personal data could implicate these protections.

Enforcement of CCPA/CPRA is primarily by the California Privacy Protection Agency and state Attorney General (individuals have a limited private right of action for data breaches). Businesses in possession of DataComp-derived personal information would need to provide notice in their privacy policy about categories of personal info collected (potentially listing data obtained from third-party sources like web scraping), and honor deletion or opt-out requests if a California resident somehow identified their data in the dataset. In practice, exercising rights on scraped data is challenging, but the legal framework puts the onus on the business to comply where possible.

4.4.3 Oregon Consumer Privacy Act (OCPA). Enacted in 2023 and effective July 1, 2024, the OCPA is part of the new wave of U.S. state privacy laws. The OCPA applies to "controllers" and "processors" meeting threshold criteria. Uniquely, Oregon's law has no revenue threshold for applicability. It applies to any entity that conducts business in Oregon or targets products/services to Oregonians, provided that it controls or processes the personal data of at least 100,000 Oregon consumers in a year (excluding purely payment data), or of at least 25,000 consumers if deriving over 25 percent of revenue from selling personal data. This thresholds test means the law mainly catches mid-size and large data handlers. Notably, the OCPA does not exempt non-profits, making it broader in coverage than CCPA. By July 2025, many non-profit organizations will also be subject to Oregon's requirements.

The OCPA defines "personal data" as information that is linked or reasonably linkable to an identified or identifiable individual (a "consumer" who is an Oregon resident). Importantly, the definition excludes de-identified data and "publicly available information." The statute regards data as publicly available if it is lawfully made available from government records or widely distributed media, or if the individual made the information public (in line with the CPRA's broader approach). Thus, similar to California, Oregon's law might exempt certain categories of CommonPool data from regulation on the premise that they were publicly accessible online. That said, OCPA's exact definition hews closely to the individual's intent and the nature of distribution; not everything on the internet would automatically count as "public" under the law's terms. Assuming CommonPool contains typical web content, much of it could be argued to be publicly available (e.g. images from public websites), and thus outside OCPA's scope of "personal data." OCPA's definition of sensitive data includes personal data revealing racial or ethnic origin, religious beliefs, sexual orientation, status as transgender or non-binary, immigration status, health information, genetic or biometric data, precise geolocation (within a 1,750-foot radius), and any personal data of a known child (under 13).

Key consumer rights under OCPA include the right to confirm if a controller is processing one's data, to access a copy, to correct inaccuracies, to delete personal data, and to opt out of targeted advertising, sales of data, or certain profiling decisions. There is also a requirement to honor browser opt-out signals for selling or targeted ads. The law imposes several obligations on controllers that align with GDPR-like principles: data minimization (collect only what is "adequate, relevant, reasonably necessary, and proportionate" to the purposes disclosed), purpose specification (process data only for purposes that are disclosed and reasonable), and reasonable security measures. Notably, if processing "sensitive data," the controller must obtain the consumer's opt-in consent. This means if DataComp CommonPool contains, say, images of children or data about minors, or biometric identifiers like facial scans, an Oregon-covered controller would legally need parental consent (for under 13) or the individual's consent (for other sensitive data) before processing that data.

In addition, OCPA mandates transparent privacy notices detailing categories of data collected, purposes of processing, categories of data shared and with whom, and how consumers can exercise their rights. It also requires controllers to conduct and document Data Protection Assessments for certain high-risk processing, such as processing sensitive data or any processing for targeted advertising, sale, or profiling that presents a significant risk of harm.

The OCPA, enforced by the state Attorney General, thus creates a compliance regime similar to other state laws but with its own nuances (like no exemption for nonprofits and a consent requirement for all sensitive data use). For a company leveraging DataComp CommonPool, if that company has a user base or market in Oregon (or otherwise falls under OCPA), it would need to treat any personal data in the dataset in accordance with these rules – unless it can argue the data is outside the law's scope (e.g. truly de-identified or public information).

In summary, all three frameworks (GDPR, CCPA, OCPA) share a broad view that personal data covers any identifiable information about individuals, which certainly includes much of DataComp CommonPool. The GDPR is the most encompassing, applying to essentially all personal data and imposing strict principles and rights. CCPA and OCPA similarly cover a wide range of personal information but carve out publicly available data and apply only to entities meeting certain thresholds. Each has special provisions for sensitive categories of data (especially data about children, biometric identifiers like faces, and financial or health information) and expects data handlers to practice data minimization, purpose limitation, and data security. As Section 7 will analyze, the presence of personal and sensitive information in CommonPool triggers these legal frameworks — raising questions about whether those compiling or using such datasets can meet the legal obligations, and whether current exceptions (like "public data" loopholes) undermine privacy in practice.

5 AUDIT METHODOLOGY

We now describe the methods we used for our privacy audit, which is inspired by similar audits of web-scraped datasets that inspect images [13], text [88], and the websites that host the samples [39, 64]. Our audit is motivated by various legal definitions of personal and sensitive information [98, 108] under various state and federal privacy laws, like the California Consumer Privacy Act (CCPA) [1] and the General Data Protection Regulation (GDPR) [44]. We consider personal information to be information that is *identifying* — in the case of image and text, samples where a face or name is present. Other privacy concerns exist even if a sample by itself is not identifying; for instance, prior work has demonstrated ways to re-identify data by linking with external sources [91, 128]. With unstructured images, however, linkage attacks can be difficult to do at scale, which we consider out of scope for our work. We explicitly refer to "personal information" rather than the term "personal data," as that term is a legal definition that depends on the relevant privacy law, where personal information that is widely accessible may not be considered "personal data" in some cases.

We downloaded DataComp CommonPool in the month of April 2025, following their code package with parameters set to default in the way it is intended to be downloaded [34]. Due to space constraints, we download the `small` scale version of CommonPool, which consists of 12.8 million randomly selected samples. Because this subset is only 0.1% of the entire dataset (and even at this scale still challenging to examine every sample), we use our observations to estimate quantities of information present in all 12.8 billion samples of CommonPool, accompanied by confidence intervals which quantify the probable estimation error due to sampling. We do not aim to capture *all* possible privacy concerns in our audit, but rather establish a lower bound based on our various approaches to inform our legal analysis. As we describe further in Section 7, the determination of whether privacy laws are triggered is not always based on scale; the presence of sensitive content alone can be enough to inform our legal analysis. Thus, to demonstrate this presence, throughout our audit we surface individual images as case studies for legal implications.

5.1 Audit techniques

We use a variety of tools to understand the privacy concerns of web-scraped datasets. The techniques that are specific to search categories, such as sociodemographic information or children's information, we define in Section 6. In this section, we highlight the study-wide techniques incorporated throughout our analysis of the contents of CommonPool.

5.1.1 Optical character recognition. To examine the contents of scraped images, such as documents or screenshots, we use optical character recognition, or OCR, to extract the text that is contained in every image. Prior OCR comparisons [132] are difficult to extend to this dataset, as images on the web may be of lower quality or depict non-document text and therefore likely follow a different distribution than these benchmarks. As a result, we perform an evaluation of various popular open-source OCR methods on a random subset of CommonPool samples in Appendix B.1 and determine that PaddleOCR [103] is most effective. We defer to Appendix C.1 for an overview of the OCR-extracted text and captions.

PII Entity	Caption	OCR
Name	1.3M	3.3M
Address	370K	1.4M
Date Time	500K	880K
Demographics	86K	240K
Email	16K	3.0K
Medical	8.9K	8.2K
URL	5.2K	2.4K
Government ID	4.1K	2.8K
Business ID	3.0K	2.3K
Financial ID	1.8K	1.2K
IP Address	494	22
Vehicle ID	171	133

Table 2: Sample counts of Presidio-detected PII entities in captions and OCR-extracted text of `small` scale dataset (12.8 million samples, or 0.1% of CommonPool). Upon manual inspection, many of these detections are false positives.

5.1.2 Entity extraction. To surface examples with personal information, we apply Microsoft Presidio's PII detection tool (version 2.2.357) to the captions and OCR-extracted text of CommonPool [86]. Presidio's recognizers incorporate regular expression matching and named entity recognition [90] to find sensitive data like credit card numbers, social security numbers, and individual names, as shown in Table 2. As in prior work [88] we find errors in detected PII entities, so we therefore use Presidio to flag content as *possibly* containing PII. Dependent on our search category (for example, sociodemographic dimensions in Section 6.1 or identity documents in Section 6.2), we then manually inspect the flagged content and only the samples that satisfy our criteria are included in our counts. We additionally flag content using basic keyword searches from prior works [15, 39] and also manually inspect matches for inclusion in our counts.

5.1.3 URL analysis. Because CommonPool is released as an index of URL-caption pairs, we evaluate the URLs storing the images as well as the image and text content of the samples. We follow similar URL analysis from recent audits [39, 64] to assess the website category and earliest recorded timestamp of the URL. Source analysis helps us understand how certain kinds of PII may have been uploaded onto the internet.

5.2 Limitations

In conducting an audit on a dataset of this size, our methodology may have certain limitations, some of which are inherited from tools we use. These limitations also extend to addressing privacy concerns via automated cleaning methods in the curation and usage of web-scraped training datasets more broadly, which we highlight in Section 8.

False Positives and Mitigation. Certain algorithms like OCR or URL categorization may make incorrect predictions, so these predictions cannot be treated as ground-truth. We also observe that Presidio's PII detection tool flags random sequences of numbers

or letters as identification numbers or financial accounts. Due to these errors, we manually inspect the samples flagged by either PII entity recognition or OCR-based keyword search and confirm which samples contain personal information.

False Negatives. We additionally recognize that our audit will miss certain kinds of information. PII detection tools do not capture all personal information, especially nuanced content or text that does not match regular expressions [88, 138]. For annotation of text, our analysis is focused on the English language, and expanding privacy audits to include non-English languages is an important avenue for future work. As a consequence, reliance on PII tools and the use of manual inspection constrains the scale of our audit to the millions rather than billions. As we discuss in Section 8.3, our work speaks to the challenges of building web-scraped datasets more broadly at scales in which every sample can no longer be individually examined.

Missing Samples. At the time of data collection in April 2025, about 21.4% of URLs failed to download, which we investigate in Section 6.3.2. This indicates that certain image assets in the dataset are no longer available, but may have been used to train models in prior dataset downloads. Moreover, these failed-to-download images still have text, URL, and dataset metadata currently available for use and for inspection. We highlight the potential reasons the URLs are missing and implications for downstream uses in Section 8.

5.3 Ethical considerations

Our institution's IRB did not consider this study to involve human subjects research due to the dataset being collected from the web, including the online presence validation approach described in Section 6.2.2. Nevertheless, as IRB approval alone is not sufficient to guarantee that a study is ethical [4, 61], we carefully considered ethics throughout our study, beginning with study conception. Given the sensitive nature of our study that can reveal personal information, we store images and indirect identifiers to these images on a secure server. In our results in Section 6, we aggregate all measurements, carefully anonymize examples to preserve privacy, and ensure that searching our redacted text on the web or the dataset does not return the actual samples.

The ethical tensions of studying public data that may violate individual privacy have long been discussed in social computing and computer security research [22, 70, 143]. We follow best practices from prior works on Internet user perceptions of the use of their data for research [41, 46]. To do so, in presenting our work we obfuscate personal information and rewrite verbatim text in individual case studies, such that the image and caption cannot be directly retrieved. We also follow the level of heavy disguise from Bruckman [22] and deliberately introduce false details so that privacy concerns are demonstrated in spirit without allowing the data subject to be recognized.

We also address various potential ethical implications of our work. (1) Our methods may be easily replicated, yet this set of 12.8 billion URLs has already been crawled over two million times [45]. (2) Our privacy audit may indirectly encourage future machine learning systems to become even less transparent for fear of legal risk, but prior work has already observed the lack of transparency of model training [57]. (3) There may also be second-order effects, as removing personal information may alleviate legal risk, although recent work shows that PII removal may lead to model memorization of leftover PII [19].

We intend to investigate the privacy implications of DataComp CommonPool to raise awareness to the degree to which privacy concerns and legal risks may arise in web-scraped data in general. We acknowledge that the datasheet for CommonPool references the presence of sensitive data and clarifies intention as a research artifact [47], yet CommonPool's licensing does not restrict the commercial deployment of models trained on this dataset — speaking to the difficulty of regulating the use of web-scraped data in general. In Section 8, we expand upon these ethical considerations for future dataset use cases as well as the tension between publicly available data and human subjects research.

6 AUDIT RESULTS

When downloaded, each sample in the CommonPool URL-caption artifact contains various components: the caption which contains **text**, the **URL** which upon downloading gives the **image** (and may contain additional text extracted through OCR), and accompanying **metadata** relating to the image. To organize our results, we structure our audit into four sections based on the data modality of our search; each section is narrowed down by search category as motivated by existing privacy laws:

(1) Section 6.1 highlights sociodemographic information and presence of celebrities found in **text** (both captions and OCR-extracted text).
(2) Section 6.2 covers identification documents and resume documents visually presented in the downloaded **images**.
(3) Section 6.3 surfaces platforms relating to children's information as well as samples that are no longer available, based on the **URLs**.
(4) Section 6.4 demonstrates issues relating to the image Exif tags and face bounding boxes in the attached **metadata**.

In our analysis, there is substantial overlap in modality, as we obtain OCR-extracted text from the images, investigate the URLs of the sociodemographic information, or use text keyword search to find visual documents. However, we ascribe each search category to its main data component (for instance, documents are verified upon visual inspection, not via text) and emphasize this groupings is for the purposes of organization rather than a contribution in and of itself. Table 3 gives an overview of our findings split by data modality.

6.1 Text

We search for query keywords in the captions or OCR-extracted **text** of samples to find matches that may contain PII. As described in Section 5, we surface both measurements and individual case studies to inform legal analysis in Section 7. In this section, we describe findings that mention sociodemographic information related to individuals as well as the presence of celebrity names. Section 6.2 later covers personal information that may appear in particular types of documents.

Modality	Search category	Results
Text (6.1)	1. Sociodemographics 2. Celebrities	We find captions that disclose the full name along with sexual orientation, religion, race, or ethnicity. We estimate at least 112 million samples mention names of celebrities mostly from the U.S. and U.K.
Image (6.2)	1. Identity documents 2. Resumes	We find credit cards, drivers licenses, social security numbers, passports, and birth certificates. We estimate at least 142, 000 images depict resumes of individuals with public online presence.
URL (6.3)	1. Children's information 2. Unavailable images	We find children's names, faces, and birth certificates, passports, and health status. Of the 21.4% of links that fail to download, 19.0% of those links fail due to lack of access permissions.
Metadata (6.4)	1. Image Exif tags 2. Face bounding boxes	We find that Exif tags attached to images reveal full names and precise geolocations. We estimate at least 102 million images of human faces are not covered by bounding boxes.

Table 3: Summary of audit findings by data modality and search category based on the April 2025 download of DataComp CommonPool.

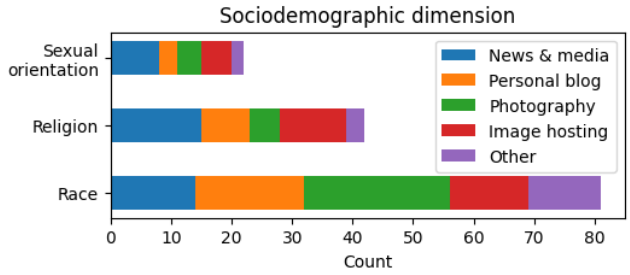


Figure 3: Number of annotated samples that link a name with sociodemographic information. Each bar represents the sociodemographic query described in Section 6.1.1 broken down by website type, which is categorized manually or verified via Cloudflare [27] for the `Image hosting` category.

6.1.1 *Sociodemographic information.*

Approach. Presidio's named entity recognition [86] first flags samples with names mentioned in the OCR-extracted text or caption. To narrow down samples, we manually discard names that do not consist of two words, as well as names of cartoons like "Peter Pan" or historical figures like "George Washington." Among this set, we query the captions and OCR-extracted text for keywords matching regular expressions related to **religion** (following the most popular world religions and religious sects [106]), **race and ethnicity** (such as `African`, `Asian`, `Caucasian`, `Hispanic`, `Latinx`, `Indian` [39]), and **sexual orientation** (such as `queer`, `lgbtq`, `homosexual`, `gay`, `lesbian`, `bisexual`). These sociodemographic dimensions are all considered sensitive data under the CCPA and GDPR [1, 44]. We manually examine these queried samples as an initial exploration and highlight various instances of sociodemographic information linked to names. Of these instances, in Figure 3 we manually categorize the websites (relying on Cloudflare [27] to determine image-hosting sites), while Figure 4 depicts individual case studies.

We find captions that disclose full names paired with sexual orientation, several of which originate from news sites. Keyword search and manual examination surfaces 22 examples depicting the names of certain individuals who identify as LGBTQ+, with some images including the person's face. As depicted in Figure 3, eight (36.4%) of these samples come from news sites. For instance, Figure 4 shows a picture of a couple with the caption describing the name of a high school student attending a queer event — this image is part of an article in which the student was interviewed and submitted the photo. In this case, the individual likely disclosed their name and information for the purposes of the news article, rather than consent to use their sociodemographic information to train a model, which we discuss in Section 7.3.2.

We observe samples that reveal religion, race, or ethnicity terms paired with full names, originating from image hosting or blog sites and some news sites. We flag 42 instances that disclose the religion and full name of an individual, as well as 81 instances that disclose the race or ethnicity of an individual. Keywords like `African` or `Indian` may describe geographic regions or origins, so we only count examples that describe the individual, such as the phrase "Asian artist." A significant portion of samples that disclose religion or race comes from news articles (17.3% for race and 35.7% for religion) similar to our findings on sexual orientation. As an example in Figure 4, it may also be likely that the individual disclosed this information for the purpose of the article. Many examples describe celebrities in which race may be inferred or common knowledge, such as "first African American president," or referring to a religious leader, such as "rabbi." We expand on the prevalence of self-disclosed religion, race, and national origin at a document level in Section 6.2.2.

Of the 142 unique samples that mention full names and sociodemographic keywords relating to sexual orientation, race, or religion, all but three samples depict human faces. We examine the images of these samples and find that 139 samples contain images of human faces. However, only 119 of of the 139 include bounding box annotations that would blur the faces by default at the time of download, which motivates the evaluation of DataComp's face detection algorithm in Section 2.2.1.

6.1.2 *Celebrity names.*

Approach. Presidio's named entity recognition tool [86] extracts the detected names in the captions and OCR text of the samples. We clean any false positives and manually discard names that represent clothing brands, fictional characters, and deceased figures, in order

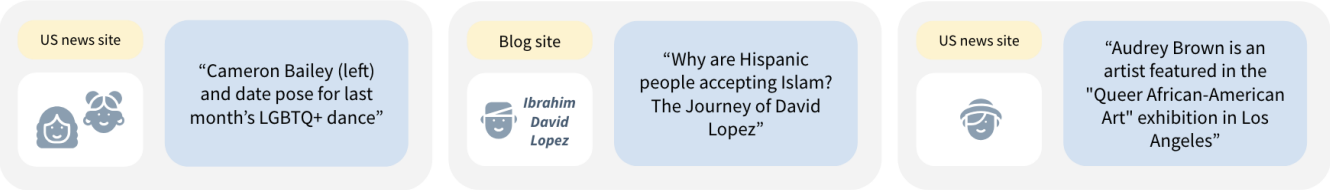


Figure 4: Examples of identifying sociodemographic information found in CommonPool's small scale dataset. For each sample, the type of URL site is shown at the top left, the image in the bottom left, and the caption in quotes on the right. All personal information has been replaced, and text has been paraphrased to avoid direct quotations. Images have been redacted to show the presence of faces without identifying the individuals.

to determine which individuals are described. As an alternative approach, we search the captions and OCR text for celebrity names from Pantheon 2020 [140], a dataset of 48 thousand well-known individuals (alive at the time at the time of collection), based on the criteria that their Wikipedia biographies have been translated into at least 15 languages. We again exclude names shared with designer brands and names that may be used in text outside of describing an individual (such as "50 Cent").

We find about 113 thousand CommonPool samples mention names of celebrities, with Donald Trump being mentioned significantly more than any other name. The Pantheon celebrity search returns mentions of 45,829 unique names (most of the original Pantheon dataset), corresponding to 113 thousand CommonPool samples. Figure 16 plots the sample frequency of the top 50 most common Pantheon celebrity names. We observe that Donald Trump is mentioned more than twice as many times compared to any other celebrity, followed by various other United States politicians, athletes, musicians, and authors. When breaking down by occupation (across all celebrity mentions), we observe in Figure 17b that actors, athletes, musicians, and politicians are the most common. In terms of country of origin, Figure 17a shows that a majority of the samples mention celebrities originate from the United States and United Kingdom. We find similar results with Presidio's named entity recognition in Figure 18.

6.2 Image

In addition to personal information present in text, we search explicitly for specific types of identification or resume documents that may raise privacy concerns. We first incorporate keyword search and the Presidio PII detection tool to surface samples with matching text descriptions of certain documents. Then, we manually verify and discard samples that do not depict documents through visual inspection of the **image** component.

6.2.1 *Identity documents.*

Approach. We use both simple keyword search (relating to drivers license, credit card, etc.) as well as the Presidio PII detection tool to surface examples relating to identity numbers. We then examine images manually to find documents with government identifiers.

We find images that depict credit cards, drivers licenses, social security numbers, passports, and birth certificates. We find pictures or screenshots of credit card numbers with full names and security codes. We also find documents or pictures of U.S. drivers licenses, social security numbers, as well as passports from various countries. We identify birth certificates, mainly of celebrities, although U.S. states often make these documents publicly accessible. A few redacted examples appear in Figure 5.

These identity documents appear on image hosting sites, even as some seem to be uploaded by the data subject themselves. Many of these identity documents are uploaded to various image hosting sites, which makes it difficult to determine if the document comes from a data breach or is uploaded by the data subject themselves. In one specific case, where a U.S. social security number is included in a military document depicted in Figure 5, we trace the document to having been uploaded by a social media account sharing the individual's name. This same image appears on another image hosting site which is crawled by CommonPool, so even if the social media user had taken down the document, the image would still exist elsewhere. We also find a few samples with captions that describe how to generate or purchase credit card numbers and social security numbers, in which the images show examples of fake identity documents.

6.2.2 *Resumes.*

Approach. The sociodemographic keywords from Section 6.1.1 indicated certain kinds of sensitive data present in professional resumes from job seekers. As such, we perform a measurement study of the types of the types of PII disclosed in job application materials and investigate the origins of these resumes. We search for CommonPool images that contain OCR-detected text relating to `resume`, `curriculum vitae`, `cv`, or `cover letter` and exclude samples with filler text in the image or caption, like `lorem ipsum` or `sample text`. For the sake of readability, we refer to "resumes" to describe all resume, cover letter, or curriculum vitae documents.

This initial query surfaces 3,770 samples (out of 12.8M), of which 3,634 images are successfully downloaded. We then engage in several rounds of annotating: (1) **Clean**: We first remove any sample that does not depict a legible resume document or clearly represents a fake individual, resulting in 805 samples. (2) **Validate**: We confirm which resumes and letters describe individuals with online presences. We find public LinkedIn profiles or news media mentioning the same name that have at least three *points of equivalence* to the resume — meaning that both sources share at least three identical attributes like middle name, job title, city, graduation year, or educational institution. (3) **Annotate**: We then manually tag

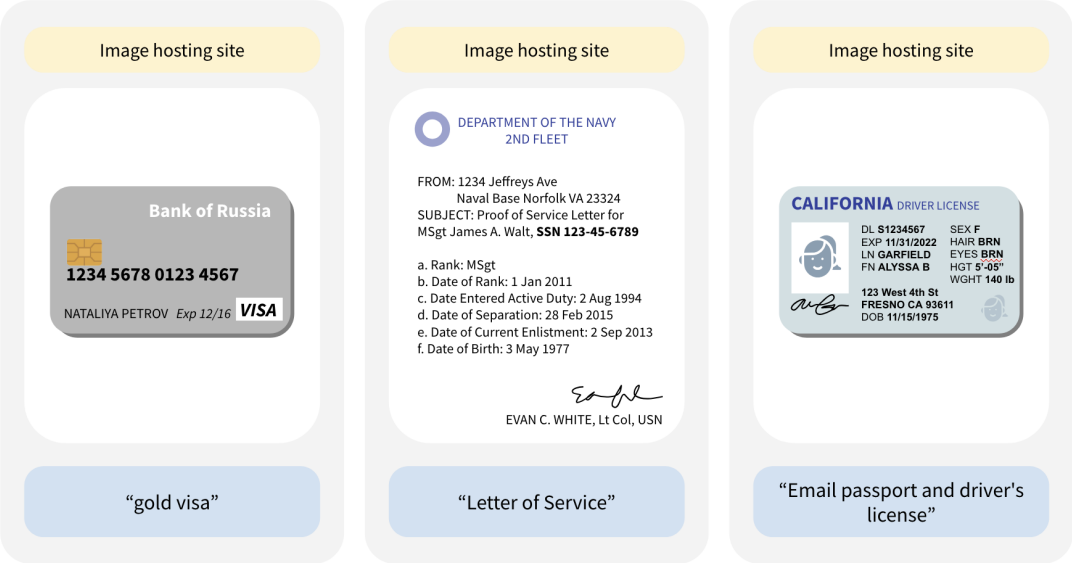


Figure 5: Examples of identity-related documents found in CommonPool's small scale dataset, showing a credit card, social security number, and a driver's license. For each sample, the type of URL site is shown at the top, the image in the middle, and the caption in quotes below. All personal information has been replaced, and text has been paraphrased to avoid direct quotations. Images have been redacted to show the presence of faces without identifying the individuals.

the 168 validated resumes and cover letters for the types of personally identifiable information present. (4) **Automate**: Finally, we automatically analyze the URLs of the validated documents via Cloudflare's URL categorization [27] and the Wayback Machine [9] to understand the origins of these images.

We find specific examples of resumes that disclose background check, disability status, the birth dates and places of dependents, and race. Searching keywords within the captions surfaces additional samples (around 14 thousand) but cannot all be annotated due to scale constraints. To complement our measurement, in Figure 7, we surface several individual examples with captions that contain resume-related words and Presidio-detected names, with additional linking to online profiles.

Overall, we estimate at least 142,000 *images in all 12.8 billion samples of CommonPool depict resume documents linnked to users with public online presence.* As shown in Table 4, out of the 3,634 downloaded images, 805 samples depict resume documents that are not visually fake. Of those, we confirm the public online presence of 168 resume documents, mostly through LinkedIn profiles but some Facebook or news articles sites as well. Given the search is within a random 0.1% subset of CommonPool, at a 95% confidence interval, we estimate between 142 thousand and 194 thousand images that depict resume documents of individuals with public online presence. This number again is a lower bound, as the keyword search does not uncover all resumes; moreover, during the validation step, some resumes may depict individuals but their profiles may be private or non-existent.

Of the validated resumes, we observe careers relating to technology and academia, and many resumes are duplicated on image hosting sites. Of the most recent jobs listed in each sample, there is a high presence of careers relating to information technology, engineering, graphic design, and marketing. We also find six samples of PhD student resumes and five samples of professor resumes. Due to the external nature of certain types of jobs, it may be reasonable to expect that professional experiences are publicly available, especially as some resumes are uploaded by LinkedIn profiles with the same name. At the same time, 20 of these resumes are duplicated across image hosting sites.

Among the resumes with online presence, we find disclosures of contact information, individual faces, government identifiers, and sociodemographic information, as well as the personal information of other individuals. In Figure 6, we manually annotate the types of personal information present in the 168 validated resumes. The majority of these images contain contact information including phone number, email, education, and physical address. Of the 112 samples with physical addresses, 11 explicitly include residential addresses, and 6 include work addresses, while the rest are not clarified. A significant number of resumes also include a photo of the individual, their date and place of birth, personal website URLs, and even government identification numbers like driver's licenses or passports. We also find presence of certain kinds of sociodemographic information like gender, marital status, number of children, religion, race or ethnicity, disability, height and weight, and criminal record. While an individual creating a resume may have disclosed personal information for job-seeking purposes, we observe information relating to other individuals, such as contact information for references, the name of the individual's father, or dates of birth of children as seen in Figure 7. In our legal analysis in Section 7, we comment on the nature of consent of sensitive data

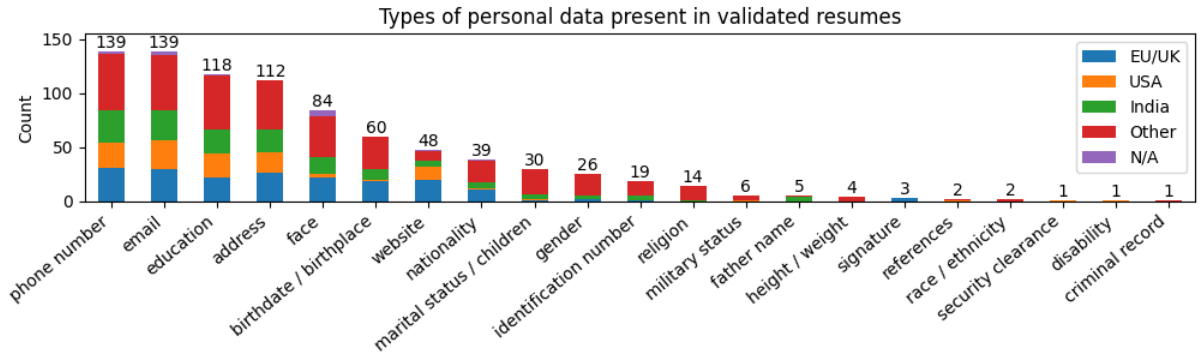


Figure 6: Sample counts of annotated personal information present in the 168 resume documents with validated online presence, broken down by region (if disclosed in resume). We highlight the United States and countries from European Union due the focus of our legal analysis (grouping the United Kingdom with the EU due to their current application of GDPR [66]). Some resumes (10 out of 168) do not include addresses and are labeled "N/A."

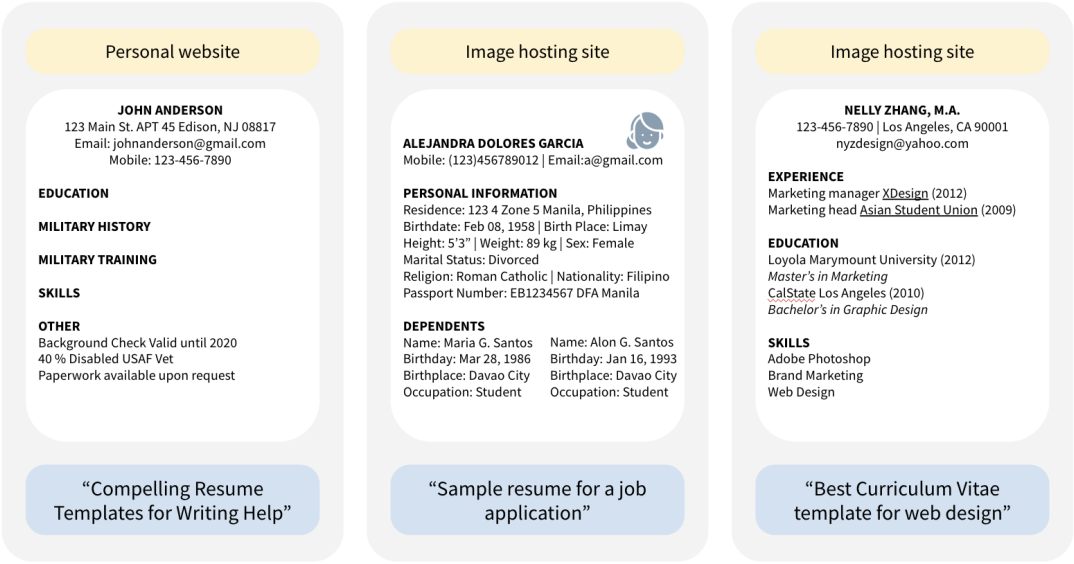


Figure 7: Examples of resume documents and personal disclosures found in CommonPool's small scale dataset. For each sample, the type of URL site is shown at the top, the image in the middle, and the caption in quotes below. All personal information has been replaced, and text has been paraphrased to avoid direct quotations. Images have been redacted to show the presence of faces without identifying the individuals.

like race or religion disclosed in resumes that are later scraped to build datasets to train models.

We find most resumes have addresses associated with India and the United States (and states with consumer privacy laws), with some associated with the European Union. We manually annotate the country associated with validated resumes, in order to inform the possibility of legal attachment based on data subjects in various jurisdictions. Each sample is tagged according to country of address in Figure 19 and national origin or citizenship in Figure 20. We find that India and the United States are the most common countries associated with the 168 validated resumes. A substantial number of resumes come from countries in the European Union. Within the United States, addresses correspond to 15 unique states, notably including California, Texas, Colorado, New Jersey, Massachusetts, Indiana, Oregon, and Illinois.

Most validated resume images come from image hosting or photography sites. The final automation step examines the types of websites that serve the validated resume images. Table 4 shows the most common websites: `bing.net`, `pinimg.net`, `slidesharecdn.net`. These findings again demonstrate the prevalence of personal information appearing on image hosting sites, potentially being propagated or not uploaded to the site by the data subject themselves. In

Resume annotation stage	Count
Overall	3634
→ Cleaned	805
→ Validated	168

Websites (out of 168)	Count
bing.net	54
pinimg.com	53
slidesharecdn.com	27

Table 4: Sample counts of resume annotation process detailed in Section 6.2.2. Left: Funnel of annotation stages, resulting in 168 samples of resumes that have a validated online public presence (e.g. LinkedIn). Right: Breakdown of most common site origins of validated resumes.

Site source	Unique sites	Count
Cloudflare Safe for Kids	493	12698
COPPA Safe Harbor	52	315

COPPA safe harbor PII presence (out of 315)	Count
Adult's face	20
Child's face	14
Name	14

Table 5: Sample counts relating to children's information detailed in Section 6.3.1. Top: Number of unique sites and sample count corresponding to each approach, Cloudflare's categorization [27] or list of COPPA safe harbor programs. Bottom: Breakdown of most common site origins of validated resumes.

Figure 21, we plot the frequency of the earliest timestamp tracked by the Wayback Machine [9] of the resume URLs. The Wayback Machine only found records for 70 of the 168 resumes, noting potential inaccuracies of the earliest recorded timestamps, which also signifies the challenges of tracing the origins of content on the web. Of the sites that were recorded, most images existed before 2022, which aligns with the fact that CommonPool is sourced from Common Crawl snapshots from 2014 to 2022. As a result, most of these resume documents may have been uploaded before the existence of popular generative AI systems [24], yet are now being downloaded over a million times to train models.

6.3 URL

This section presents results on privacy concerns relating to the **URL** component of CommonPool. Section 6.3.1 describes searching for children-related websites to find children's information present in samples, and Section 6.3.2 investigates URLs that fail to download due to DataComp's web crawler.

6.3.1 Children's information. While children's information does not fall under the definition of sensitive data in the CCPA and GDPR (but does for the Oregon Consumer Privacy Act [2]), both laws consider special provisions for children's information. Moreover, in the United States, the Children's Online Privacy Protection Rule (COPPA) protects the use of personal information from children under 13 years of age [3].

Approach. We search for personal information relating to children, primarily focusing on online services directed towards children (due to the scope of COPPA requirements). To do so, we initially perform a manual keyword search for samples mentioning `child` and related words in the caption, in order to find individual case studies. We next identify samples that come from children's related websites. We rely on Cloudflare website categorizations [27] from prior work [64] on a random subset of 100,000 domains from CommonPool and isolate samples belonging to sites in the `Safe for Kids` category. While Cloudflare categorizations have been shown to be accurate [113], samples that come from these websites may not necessarily be considered children's information. We thus follow up with an alternative approach to examine samples from sites participating in COPPA safe harbor programs. The safe harbor provision enables industry groups to self-regulate its members to follow COPPA's guidelines [29]. Because members that join these safe harbor programs intend to be in compliance with COPPA, we search for sites from iKeepSafe [67], kidSAFE [69], and PRIVO [107] certification programs, which display their members publicly. Table 5 gives an overview of the sample counts and number of sites by each approach. We examine samples associated with these sites for any personal information.

We locate samples that depict children's birth certificates, passports, and health status, originating from news articles or blogs. With keyword search, we discover various images of children's birth certificates and passport information. As shown in Figure 8, we also find an image of a child's passport and also an image depicting a child unconscious on a hospital bed, with the caption including their full name and health condition. These samples often come from news articles or online blogs, in which it may be plausible that the use of these photographs and full names may have obtained parental consent specifically for inclusion in the article.

Sites categorized as "safe for kids" include child-targeted companies like Hasbro or Disney, as well as platforms in Japan and the United Kingdom, mainly depicting toys or cartoons rather than personal information. We find about 500 websites categorized as `Safe for Kids` by Cloudflare, corresponding to about 13,000 samples, of which 3,000 samples have faces detected by CommonPool annotations. Figure 22a provides a breakdown of the most frequent sites that Cloudflare categorizes as `Safe for Kids`, including various commercial platforms targeted towards children, such as Hasbro or Disney. Some popular sites also have country domain names associated with Japan, Russia, South Africa, and the United Kingdom. We also observe that 34% of samples come from `rcgroups.net`, which is an image hosting site for RCGroups, a radio control forum (not necessarily related to children). Upon further examination of these samples, we find pictures of toys or cartoons rather than personal content, so presence of samples from Cloudflare-categorized sites do not reveal much evidence of private information.

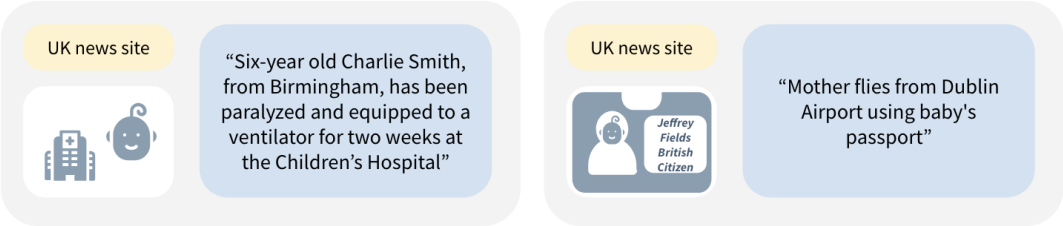


Figure 8: Real examples of children's information found in CommonPool's small scale dataset. For each sample, the type of URL site is shown at the top left, the image in the bottom left, and the caption in quotes on the right. All personal information has been replaced, and text has been paraphrased to avoid direct quotations. Images have been redacted to show the presence of faces without identifying the individuals.

We find child's names and faces revealed in samples from sites that are members of COPPA Safe Harbor platforms. There are 315 samples from COPPA Safe Harbor sites, with most frequent sites are shown in Figure 22b. We manually annotate all of these samples for presence of PII and find that some examples show adults' or children's faces, as well as full or first names (in Table 5). While these numbers are small, not all platforms under COPPA safe harbor programs are captured, and this search only covers a random 1/1000th subset of all of CommonPool — with a 95% confidence interval of this observed sample proportion, we estimate between 280,000 and 350,000 samples from our website set that intends to be COPPA-compliant.

6.3.2 *Unavailable images.*

Approach. At the time we download CommonPool in April 2025, 21.4% or 2.7 million image URLs fail to download, yet during the time of CommonPool's creation in March of 2023, all the images could be successfully scraped from the web, otherwise these samples would have been removed [47]. Even if there are unavailable images in our current version, their corresponding captions, site URLs, and accompanying image annotations still exist with the URL-caption artifact. Because websites frequently change or are no longer maintained, and as some download errors may be a result of our server or connection issues, here we investigate the types of HTTP download errors.

*We find that the most common download error is due to broken or dead links (*35.4%*), while the next most common is due to a lack of permission to access the link (*19.0%*).* In Figure 23, we plot the most common HTTP download errors (after manually merging and renaming similar errors) and find `Not Found`, `Forbidden`, and `Service Not Known` as the most common errors. While many errors relate to a failure to reach the image URL, which perhaps indicates the website or image asset has moved, we find a substantial number errors relating to permissioning with confirmation that the image asset exists. For instance, we observe that some samples with `Forbidden` errors do in fact render manually, which means that the web server may have recognized the download script [34] as a web crawler and subsequently blocked access. The `Forbidden` error is distinct from `Unauthorized`, as it indicates that the web server recognizes the DataComp crawler and verifies a lack of permission. In other words, there is an explicit rejection of consent for users of this dataset to automatically scrape site content, but the images have been scraped in the past, and the captions, URL, and metadata are still available.

We observe that the tool to download DataComp CommonPool by default respects image robots tags when crawling, but not site-level robots.txt protocols. Figure 23 shows that `Robots Disallowed` tag is also a common error for 51 thousand image URLs that fail to download. Upon investigation, we find that the crawler for DataComp by default respects X-robots-tag (unless explicitly modified by the user) [51]. The X-robots-tag is specified by the site host in the image URL HTTP header when the link is crawled [89], and is distinct from the robots.txt protocol which is surfaced at the main webpage (and not each individual image URL). While the DataComp crawler respects site host consent preferences at the image-level, the crawler in its current form requests the *entire* URL content. These requests may increase load on the web server and therefore increase costs for the site host, especially if fetched over two million times! If a site host wishes to prevent crawling at the image-level for server performance or cost reasons, the load increase effectively defeats this purpose. While the tool maintainers are aware of this issue, at the time of writing it has not been resolved [51]. Moreover, if wishing to prevent web-scraping for other purposes, a site host that disallows crawling on its robots.txt file would have to continually attach X-robots-tags to every image URL on the site just to avoid the image content being scraped. The DataComp crawler's current setup to ignore robots.txt runs counter to best practices from the World Wide Web Consortium [134].

Several websites' entire set of samples fail to download in our evaluation set, of which most of these websites no longer load, and one website has a login screen to access these images. Of the 2.7 million image URLs that failed to download in the `small` scale of CommonPool, we find that 1.2 million of these are from website domains that have successfully-downloaded samples, while the leftover 1.5 million image URLs are from websites that are no longer available. As it is plausible that these websites may have available images URLs on larger scales of CommonPool, we examine the error breakdown of these "failed" sites in Figure 9. The most popular failed sites have unavailable image URLs for a variety of reasons, but we note that of the top five, the most common reason is due to a `Forbidden` error. Of the sites listed in Figure 9, we try to load the

main site page and find the majority of these sites fail to load. However, we find that the most common website `specsserver.com`, which composes 0.4% of CommonPool, renders a login screen. If the login screen of the website existed at the time of CommonPool creation, the image URLs that were once available to download may be considered not legally public as described in Section 7.5. If the login screen and authentication to the image URLs were added after CommonPool's release, then the site host may have blocked access to the image assets, although they would have been downloaded in earlier versions.

Of the random subset of failed-to-download images with Wayback Machine records, most of the image URLs had earliest timestamps before 2022. We track the earliest timestamp recorded by the Wayback Machine [9] for a random subset of 1000 image URLs that fail to download. We find records for 21.3% of these URLs in Figure 24a of which most of these existed before 2022 (and now are no longer available). We examine whether image URLs that fail to download have earlier records than image URLs that successfully download — perhaps due to older sites lacking maintenance — and observe that the distribution of existing are roughly similar (shown in Figure 24b). To compare the distributions of successfully-downloaded and failed-to-download groups, we randomly select 1000 samples per group to measure statistical differences in the sample means as shown in Table 8a. Compared to a random subset of successfully-downloaded samples, failed-to-download images on average are larger, have more detected faces, and have higher CLIP-similarity scores (DataComp's measure for image quality), although differences are slight.

We find that captions of samples that mention invoices, social security numbers, and credit cards are associated with higher-than-average download error rates, but by a small amount. Because the captions are still available of image URLs that fail to download, we examine the association between the download error rate and the presence of PII in the caption. We query samples for regular expression matches with personal information, like driver's license, passport, resumes, etc. In Figure 25, we see that captions that mention invoices are associated with higher-than-average error rates, in addition to credit cards and social security numbers, while captions that mention resumes have substantially lower download error rates.

6.4 Metadata

We now focus on the **metadata** component associated with each URL-caption pair. Section 6.4.1 describes the image Exif tags that are extracted when downloading the image assets of CommonPool. Investigation of these tags reveals precise geolocation data accompanied with full names. We then examine the face detection metadata in Section 6.4.2: as described in Section 2.2.1, the released CommonPool artifact comes with bounding box annotations from a face detection algorithm so that when downloaded, the detected faces can blurred in the dataset (unless overridden by the user). We search through the face annotations to evaluate whether this face obfuscation technique effectively anonymizes the presence of faces.

6.4.1 *Exif tags.*

Approach. Each web image is embedded with Exif tags, which can be added manually or automatically by cameras at the time of image creation. However, Henne et al. [62] show that users often are not aware of metadata, which can disclose personal information, being shared when an image is uploaded to the web. DataComp's download tool explicitly extracts the image tags according to the Exif standard for every sample, which means that additional information is also being stored at the time of download. In this section, we investigate these Exif annotations and search for presence of individuals.

We find non-empty Exif tags relating to timestamps, geolocation, and individuals, which upon inspection many of which disclose full names. Figure 10 plots the frequency of non-empty Exif tags that may disclose personal information, where there are hundreds of thousands of samples that are embedded with metadata detailing timestamps, geolocation information, and individual presence. We investigate various Exif tags relating to individuals and find that while some reveal companies or photography studios, a significant amount of metadata text for the `CameraOwnerName` and `Artist` tags include full names.

We re-extract the Exif tags for GPS information and find that 28.6% *of* `GPSInfo` *tags point to precise geolocation, of which* 6.1% *of those also come with full names.* Of the 102 thousand images with `GPSInfo` Exif tags, we observe that the tag information is malformed at the time of download, so we replicate the Exif tag extraction process for a random subsample of 5 thousand geolocation-tagged images. We determine that 28.6% of extracted `GPSInfo` tags contain precise latitude and longitude locations attached to the images, which cameras may include by default [62]. Of those images with precise geolocations, about 5.9% of those images also have Exif tags contain full names.

6.4.2 *Face biometric data.*

Approach. As described in Section 2.2.1, the CommonPool artifact includes annotations of bounding boxes automatically-detected via the SCRFD algorithm. The script to download the images will by default blur those detected face regions, although the dataset user can override the obfuscation step (and can even use these annotations to extract images of detected faces). To evaluate DataComp's method to detect and obfuscate faces, we apply Amazon Rekognition's face detection algorithm applied to 100 thousand randomly selected samples to surface samples uncaught by SCRFD. Among the manually confirmed images of faces, we annotate the presence of a name, location, and whether the face unambiguously depicts a child under 13 years of age. We also investigate the site origins of the images with human faces, as well as any differences in the distributions of human faces caught and uncaught by SCRFD. While prior work has demonstrated that individuals with faces blurred can still be identified [83, 99], we leave this investigation to future work, especially as models trained on CommonPool with faces blurred seem to identify signals to predict gender and race [47].

We estimate at least 102 million images of human faces are not obfuscated when downloading CommonPool with default parameters, although a majority of these facial images are small. Out of the 100

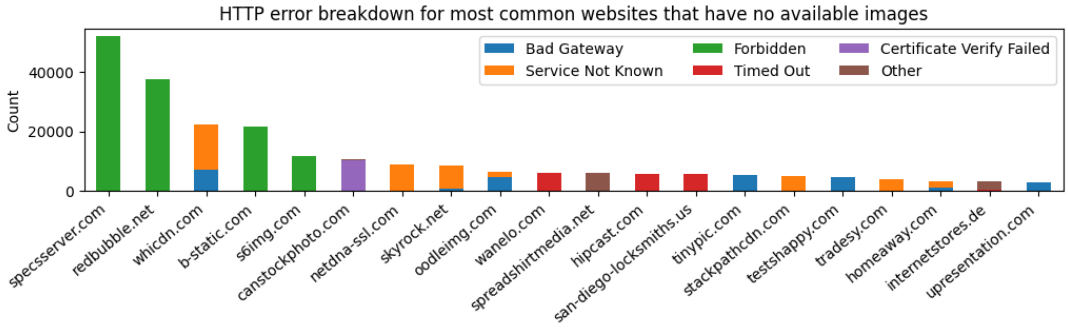


Figure 9: Error breakdown of most common websites of which all samples failed to download for the `small` scale, but at the time of CommonPool curation and for early download versions image URLs on these sites were successfully downloaded.

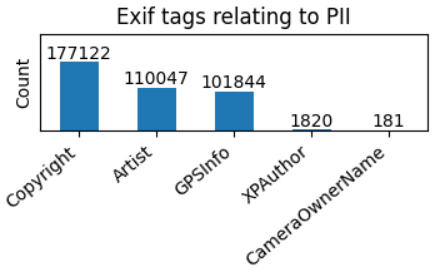


Figure 10: Sample counts of non-empty Exif tags relating to PII for all successfully-downloaded images in small-scale of CommonPool. Upon inspection, many name-related tags, like `CameraOwnerName` or `Artist` contain full names. We re-extract a random set of non-empty `GPSInfo` tags and find that 28.6% contain precise geolocation.

Annotation (out of 1.4K)	**Count**
Human	854
Drawing	357
False	234

PII presence (out of 854)	**Count**
Location	155
Name	52
Child	48

Table 6: Manual annotation counts of samples without facial bounding boxes. Left: Annotation of 1,445 samples with Rekognition positive classifications (at least one face detected) and SCRFD negative classifications (no faces detected). Right: Annotated PII presence of samples with manually verified human faces that were not covered by DataComp's face blurring.

thousand random samples, SCRFD detects 25 thousand samples containing at least one face, while Rekognition detects about 19,000 samples containing at least one space. For about three thousand samples, Rekognition detects more faces than SCRFD, of which 1,445 images SCRFD does not detect a single face. In Table 6, we manually annotate those 1,445 samples and find that 59.1% of these samples depict human faces missed by SCRFD, 16.2% are false positives, and 24.7% are depictions of drawings or cartoons. At this miss rate, for all 12.8 billion samples of CommonPool, we estimate at a 95% confidence interval between 102 million and 117 million samples contain human faces uncaught by DataComp's current face obfuscation mechanism (a lower bound based on the samples surfaced by Rekognition). We observe that a majority of these images contain human faces with bounding boxes of less than 400 square pixels, which means that the depicted faces are low quality — and while this finding implies that identification may be difficult based on the facial image alone, we still find many high-quality images of people's faces.

We find presence of personal information relating to name, location, and depiction of children among the non-obfuscated human faces. We then manually annotate for the types of personal information present among the 854 samples with human faces uncaught by SCRFD. In Table 6, we find mentions of location and name in the caption or image, as well as images that depict children with faces non-obfuscated. We observe various examples of screenshots with un-blurred profile pictures with full names present.

Compared to the images of faces obfuscated by DataComp, the images of faces not obfuscated are on average smaller and have lower pixel brightness. Among the manually confirmed samples of Rekognition predictions that contain a single human face, we randomly draw 400 images each from the subset with faces detected by SCRFD and the subset with faces undetected by SCRFD. We examine differences in the means of these two sampled groups with respect to various image-related variables: Rekognition's bounding box area, average pixel brightness, proportion predicted as `Female` by Rekognition, and the age predicted by Rekognition. Age and gender may be unreliable measures due to biases encoded by Rekognition [117], which make it difficult to make valid inferences. In Table 8b, we find a statistically significant difference (at a 99% confidence level) between the image-related variables of facial images detected and undetected by SCRFD. We accept the alternative hypothesis that unblurred facial images on average have smaller bounding boxes, less bright in average pixel value, and younger Rekognition-imputed ages than facial images that would have been blurred.

Observed differences may be situated against ongoing work of sociodemographic biases in face detection. While these image-related variables represent noisy imputed signals (i.e. Rekognition has its own classification error and biases), these statistical differences in pixel brightness may demonstrate certain instances in which the task to detect faces is not as accurate. These results may possibly relate to other demographic biases, as prior work has established well-known biases in face or person detection along the lines of skin tone [85, 137].

Many images in the set of human faces not covered by bounding boxes originate from image-hosting sites, blogs, and media platforms with earliest records before 2022. Finally, we examine the website URLs of the various images of human faces undetected by SCRFD. In Appendix C.6, we find many images originate from image-hosting sites, blogging sites like Wordpress, and media sites. We also confirm that earliest timestamps recorded by the Wayback Machine for most samples with records are before 2022, before the existence of popular generative AI systems [24].

7 LEGAL ANALYSIS

Our audit results now inform legal analysis of the treatment of personal data in CommonPool, determination of legal attachment and obligations, and the sufficiency of sanitization attempts.

7.1 Does DataComp CommonPool contain "personal data," and if so, how do privacy laws treat it?

Given the empirical findings of our audit, it is clear that CommonPool contains extensive personal information, triggering the definition of "personal data" under GDPR and the equivalent terms under U.S. laws. As highlighted in Table 3, the dataset includes identifiable human faces, full names and contact details on resumes, government ID numbers, financial information (such as credit card numbers with security codes), and even content involving children (e.g. birth certificates). Under the EU's GDPR, virtually any information relating to an identifiable person is personal data. A photograph of a face, for example, is personal data because a person can be identified from it (either directly by recognition, or indirectly via facial recognition technology or matching with other data). Likewise, a resume image clearly "relates to" an identifiable person (the individual named on the resume). Thus, almost all the examples uncovered, faces, resumes, names, emails, credit card details, qualify as personal data under GDPR. The fact that this data was scraped from publicly accessible websites does not remove it from GDPR's ambit. In practice, a photo posted on a personal blog or an image on Flickr is protected personal data in Europe despite being publicly viewable.

Under California's CCPA/CPRA, the scraped information also largely counts as "personal information." The law explicitly includes inferences about a person within the definition of personal information, meaning even any tags or labels inferred in the dataset (for instance, if the dataset or subsequent model infers someone's age or occupation from an image) are considered personal information about the individual. California regulators have affirmed that internally generated profiles or inferences are covered just like collected data. However, California's law has a notable exclusion for "publicly available" information. The DataComp curators might argue that because they scraped data from public internet sources, the data is "publicly available" and thus not subject to CCPA. This argument has some force only if the data squarely fits the statutory definition of publicly available, it was lawfully made available through government records or widely distributed media, or was broadly made public by the individual. Some subset of CommonPool likely does come from widely distributed media (for example, news websites), and some comes from individuals' public postings on social platforms. To that extent, a business could claim those specific portions are exempt. However, it is not a blanket escape hatch. The CPRA version of "publicly available" still requires a case-by-case look at how the data was made public and by whom. For instance, a leaked database posted on a forum would not count as "lawfully made available." A personal photo taken from behind a login-only site (i.e. the example found in Section 6.3.2 if the login screen existed at time of crawling) would not be "publicly available." Even data that was public may lose the exemption if used in a manner different from the purpose for which it was published (the law implies that the context of publication matters). Moreover, any derivative information in the dataset (such as embeddings or metadata added) wouldn't be "publicly available" in origin. In short, while California's law could deem parts of DataComp CommonPool outside its scope, a large portion, especially the more sensitive bits like driver's license images, personal communications, or anything not obviously from a public-facing source, would still be considered personal information subject to the CCPA. And crucially, being "public" does not strip individuals of all protections: if a California resident finds out a business is using their personal photo or essay from the web, they could still exercise rights (like deletion), since the CCPA's public-data exception mainly affects whether the law applies at all, not what happens once data is in play. Publicly available sensitive data can also potentially be regulated if used in certain ways, for example, a business could still be restricted from using a publicly posted race or health detail for targeted advertising without offering an opt-out, because that would be profiling on sensitive grounds.

Under the Oregon Consumer Privacy Act, the definition of personal data likewise covers CommonPool's contents with an exclusion for publicly available data. Oregon's definition of "publicly available" information is similar to California's broadened definition (encompassing information a person has made public or that is available through public sources). If the CommonPool entries are determined to be publicly available under OCPA, they would not be considered "personal data" under that law. For example, an image scraped from a publicly viewable Instagram profile might be deemed public information.

The OCPA mandates opt-in consent for processing sensitive data, which includes "any personal data of a child" and biometric identifiers. One cannot simply scrape a child's personal details from a public website and evade Oregon's consent requirement, because the statute treats all children's data as sensitive regardless of source. Thus, personal data about children in DataComp CommonPool is a particularly problematic category under all frameworks: GDPR accords it special protection (requiring parental consent for young children in many cases), CCPA/CPRA imposes opt-in consent for

selling minors' data and heightened duties to protect children, and OCPA flatly requires consent to process kids' data at all. Our audit unearthed content like birth certificates and photos of children in the dataset– these likely pertain to minors, meaning any entity subject to OCPA or even general consumer protection could face legal risk in using that data.

In sum, the personal data in CommonPool does not cease to be personal data simply because it was scraped from the web. GDPR treats it as fully regulated personal data. CCPA/CPRA and OCPA carve out some public data, but not in a way that would categorically exempt the majority of a massive, wholesale-scraped collection. At minimum, identified or identifiable individuals are present throughout the dataset, and thus privacy laws recognize their personhood in the data. The inclusion of inferred data (e.g. algorithmically generated labels about individuals) also falls under these definitions. The GDPR is clear that profiling data or any information "relating to" an individual is in scope, and California explicitly lists inferences as personal information. Therefore, any notion that CommonPool's billions of samples are completely anonymized or not "personal" cannot be sustained. Despite some sanitization, our findings (such as 102 million images of real, unblurred human faces remaining after filtering in Section 6.4.2) indicate that identifiable data is abundant. Each such face image is biometric data linked to a person; each resume is a dossier of someone's identity. Privacy law is concerned with exactly these kinds of data.

7.2 When and how do these privacy laws "attach" to DataComp CommonPool or its use?

The applicability of GDPR, CCPA, and OCPA depends on the circumstances of the entity processing the data. DataComp CommonPool itself is an artifact, a collection of files, and not a legal entity. Thus, the laws apply to the controllers or processors who handle that personal data. Different scenarios illustrate when obligations would kick in:

7.2.1 The dataset creators/distributors. Suppose the team that compiled CommonPool (Gadre et al., per the DataComp paper) is based in the U.S. and released the dataset publicly for research. If they have no business operations in California or Oregon and are an academic/non-profit entity, CCPA and OCPA likely did not apply to their act of compilation (CCPA covers only for-profit businesses, and OCPA only from 2024 with inclusion of non-profits). GDPR might apply if, for example, EU personal data was scraped (such as resumes with disclosed addresses from EU countries in Figure 6) and the act of scraping is considered monitoring behavior of EU residents (web crawling could be seen as a form of monitoring). Clearview AI's scraping of EU citizens' photos led EU regulators to assert GDPR's jurisdiction, even though Clearview was a U.S. company. Under GDPR Article 3(2)(b), monitoring individuals' behavior in the EU (which continuous scraping and analyzing of EU websites could qualify as) brings the activity under GDPR. Additionally, if any EU-based researchers or organizations are involved in hosting or curating CommonPool, GDPR directly binds them. We see that privacy laws can attach even at the dataset creation stage if the compilers meet jurisdictional criteria. However, enforcement at that stage is murky — for instance, if an academic merely scrapes data for research without any commercial purpose, they might invoke exceptions for research or freedom of expression (though GDPR's research exemption still requires safeguards and doesn't nullify data subject rights entirely).

7.2.2 Downstream users (companies or researchers training models on DataComp CommonPool). This is likely the more consequential point of attachment. Any organization that obtains CommonPool and processes it to train an AI model becomes a data controller (determining the purposes and means of processing personal data in the dataset) or a processor for some other controller. If that organization is in the EU, GDPR straightforwardly applies. If it is outside the EU but offering an AI system to EU residents or monitoring EU individuals' behavior through the model, GDPR also applies extraterritorially. For example, a U.S. company training a photo recognition model on CommonPool, which might later identify EU individuals, is arguably processing EU persons' data and could be seen as monitoring them (especially if the model can recognize EU citizens from scraped images, which is precisely what EU regulators objected to in the Clearview case). Under California law, if the user of CommonPool is a for-profit business that does business in California (which includes virtually any larger tech company or any company selling services in CA) and meets a threshold (say they have over 25M revenue or deal in large volumes of data), then any personal information in CommonPool pertaining to California residents falls under the CCPA. It may be hard to know which entries are Californians, but realistically, a significant portion likely are (given California's large online population). The law would require that business to, at minimum, include those categories of data in its privacy disclosures and honor any consumer rights requests related to them. Oregon's law similarly attaches if the user "conducts business in Oregon or targets Oregon residents" and crosses the 100k-resident data processing threshold. The threshold count (100k individuals' data) could easily be met by a dataset of billions (even random sampling would include more than 100k Oregonians). Notably, OCPA has no revenue threshold, so even a smaller company (or a non-profit, starting in 2025) would be covered if they process data about 100k people in Oregon. In essence, any substantial deployment of DataComp CommonPool by a tech company or organization is likely to trigger one or more of these privacy regimes. The only actors who might be outside the laws' reach are, for example, a researcher using the data in a purely non-commercial setting and not sharing the model or outputs in regulated markets. But the moment the data or any model derived from it enters commerce or is made available to individuals in regulated jurisdictions, the privacy laws become relevant.

7.2.3 Thresholds and exemptions. It is worth noting specific threshold quirks: CCPA's threshold of 100k consumers/households for buying/selling data might conceivably rope in the dataset distributor if, for instance, over 100k Californians' data was exchanged (even freely). But since the dataset is openly published (not sold) and the compilers presumably don't have a traditional business relationship with California consumers, CCPA likely wouldn't label the compilers as a "business." Conversely, a big tech company using the data definitely has annual revenue > 25M (threshold a) and will derive value from the data (even if not selling it, simply retaining

it counts as processing). OCPA's inclusion of non-profits means if, say, a non-profit research consortium in Oregon curates or uses CommonPool and it involves >100k individuals, they would have to comply as well (OCPA from 2025 covers non-profits processing large data volumes). GDPR of course has no threshold, even processing data of one EU person can invoke rights and obligations, but enforcement priorities might focus on large-scale systematic processing, which CommonPool certainly is (processing on a "large scale" triggers requirements like Data Protection Impact Assessments under GDPR, per Article 35).

In summary, these laws attach wherever personal data from DataComp CommonPool is processed by an entity within their reach. In practice: a company in California using CommonPool is under CCPA/CPRA; any company of significant size anywhere in the U.S. using it might fall under some state law (if not California's, then perhaps another similar state law, since many states now have comparable statutes). Any company or researcher in Europe using it must comply with GDPR. Even a non-EU company could be subject to GDPR if EU individuals' data in CommonPool is involved in offering a service (for instance, offering a generative image service that might recreate someone's image or personal details). Therefore, the mere presence of regulated personal data in the dataset "anchors" legal obligations to anyone who takes possession of it, unless they undertake robust anonymization (which, as we discuss, was attempted in part but not fully successful).

7.3 What obligations are triggered once these laws apply?

If an entity is subject to GDPR, CCPA, or OCPA while using DataComp CommonPool, a suite of legal duties follow. We outline the most pertinent obligations:

7.3.1 Lawful basis / consent. Under GDPR, every processing of personal data requires a lawful basis (Article 6). For a dataset like CommonPool, it is hard to imagine a lawful basis other than legitimate interests or consent, and consent of the individuals whose data was scraped has not been obtained in any direct way. Legitimate interests (Article 6(1)(f)) might be invoked by an AI developer, arguing that training a model is in their (and perhaps societal) legitimate interest. However, this basis requires a balancing test weighing the impact on individuals' rights. Given the dataset includes sensitive info and people have no expectation of this use, the balance may tip against the controller's interest. Moreover, for special categories of data (GDPR Article 9) like biometric identifiers (faces) or health data that may be present, legitimate interests cannot be used at all – a specific condition like explicit consent or "data manifestly made public by the subject" (Art 9(2)(e)) would be needed. It is highly doubtful that individuals depicted in these images explicitly consented to this use of their data (training an AI). Thus, a GDPR-compliant processor of CommonPool would either have to filter out all special-category data or find an Art 9 exception (scientific research could be one, Art 9(2)(j), but that requires meeting strict necessity and proportionality requirements and providing appropriate safeguards). Under CCPA and OCPA, the concept of lawful basis is less formal, consent is generally not required just to collect or use regular personal data (except for sensitive data under OCPA). However, if the data will be used for certain secondary purposes, consent or opt-outs become relevant: for instance, if a business were to sell any CommonPool personal information (selling in CCPA includes any disclosure for value), it would need to provide an opt-out mechanism. If it engages in "profiling" or automated decision-making that produces legal or similarly significant effects on individuals, some laws (like OCPA and forthcoming CPRA regs) may require consent or at least assessments. Oregon's OCPA explicitly requires opt-in consent for processing sensitive data. So any CommonPool entries that fall under sensitive data (which includes biometric data, specific geolocation, or a child's data) legally mandate obtaining consent from the individual before using. Obviously, in a scraped dataset context, obtaining individual consent post-hoc is nearly impossible (the controller often doesn't even know the identities or have contact with data subjects). This puts the controller in a position of non-compliance by default if they proceed to use sensitive personal data without consent. The only workaround is to exclude or anonymize those pieces – which again raises the question of how effective the dataset's filtering was.

7.3.2 Notice and purpose specification. Privacy laws uniformly require transparency about data practices. A company using DataComp CommonPool to train a model would need to disclose in their privacy notice/policy that they collect and use personal data, potentially from third-party sources (and describe categories such as photos, resumes, etc.). Under GDPR (Articles 13-14), if personal data is collected indirectly (not from the individual), the controller must provide the individual with a privacy notice including the source of the data and the purposes of use. Complying with GDPR's notice obligation for a scraped dataset is logistically daunting, one would have to somehow inform millions of individuals worldwide that their publicly posted content is now being used for machine learning development, giving them details and rights information. In many cases this is practically impossible, and GDPR acknowledges this by allowing exceptions if providing notice is "impossible or would involve disproportionate effort" (Art 14(5)(b)), but then the controller must instead publicly post the information. This means at minimum a public-facing notice should exist. For example, LAION (a German non-profit that created a 5-billion image dataset similar to DataComp) published an online notice listing broad details of the processing and offering an opt-out email for copyright or personal data takedown requests [71]. A business in California would similarly need to include in its CCPA-required privacy notice the categories of personal information it collects (which would include those scraped categories) and the purposes (e.g. "to train and improve AI models"). Oregon's law also mandates clear notices with purpose statements. Furthermore, purpose limitation means the controller should process the data only in ways compatible with those purposes. If the data was originally collected by websites for other purposes, a strict reading of purpose limitation (especially under GDPR) suggests that using it for model training is a new purpose that might not be "compatible" with the original context, absent certain conditions or consent.

7.3.3 Data minimization and scope of collection. All regimes encourage minimizing personal data use. GDPR's Article 5 and OCPA explicitly require that only data which is necessary for the stated

purpose be collected/used. In the context of DataComp, one must ask: is each piece of personal data needed to achieve the aim of training a useful model? Likely not; the collection is opportunistic (grab as much as possible). A privacy officer evaluating this under GDPR would be hard-pressed to justify that, say, 168,000 individuals' resumes (found in Section 6.2.2) are necessary to train a general image-caption model. Similarly, OCPA's requirement that collection be "reasonable in relation to the purposes" might be violated if highly sensitive or irrelevant personal info was included beyond what's needed. This principle might force a controller to actively filter out or minimize the personal data from the dataset (e.g., perhaps hashing faces or excluding text that looks like contact info) to align with the law. DataComp's curators did attempt some minimization by blurring faces and removing some not-safe-for-work content, but our findings in Section 6.4.2 show these measures were incomplete. One hundred two million unblurred faces remain, and of our subsample we find many alongside identifiable context like names or locations (Table 6), which is far from the minimum data necessary for any specific training objective; it's rather an artifact of imperfect filtering. Under privacy law principles, a controller using CommonPool would be expected to proactively weed out unnecessary personal data. Failing to do so could be seen as a breach of the duty to implement Privacy by Design (GDPR Art 25), which requires controllers to integrate data protection principles (like minimization) into the processing activities.

7.3.4 Data subject rights and control. Once personal data is being processed, individuals have rights that the controller must be able to honor. Under GDPR, these include the right to access their data, rectify inaccuracies, erase data (the "right to be forgotten"), restrict or object to processing, and not be subject to certain automated decisions (Art 15–22). For a company holding CommonPool data, responding to such requests is extremely challenging. How would they find one person's data among billions of samples, especially if the person only knows, for example, "there might be a photo of me scraped from my blog?" It's not impossible, the controller could at least attempt to search if provided with identifying details (some projects use perceptual hashes or embeddings to identify specific images). But the scale is prohibitive for tracing all instances, especially as our audit in Section 6.2.1 shows government identifiers like social security numbers can propagate to various image hosting sites. Still, legally, if an EU citizen made a GDPR access or deletion request specifically referencing this dataset, the controller would have to attempt compliance. Failure to comply could result in regulatory action or penalties. Under the CCPA, Californians have the right to request that a business disclose what personal info it has about them and delete it (with some exceptions). A business leveraging CommonPool data would need to have processes for such requests. They might rely on the exception for data collected from a third party that is not maintained in a manner that would be considered personal (for instance, if truly anonymized or if they cannot verify the person in the data). But regulators might not look kindly on "we have your data but can't delete it because we can't find it" as a response. OCPA and similar laws also provide rights to access and delete. Thus, the legal analysis reveals a tension: the very nature of giant scraped datasets makes individual control and data subject rights nearly impossible to operationalize. This is one reason scholars argue that the notice-and-choice or individual control model breaks down at scale. In practice, if individuals start invoking their rights against AI training datasets, controllers might opt to delete broad swathes of data or refrain from use, as compliance on a piecemeal basis could be infeasible.

7.3.5 Special rules for sensitive data. All three regimes impose stricter conditions on sensitive personal data, which DataComp CommonPool unquestionably contains. Under GDPR, as mentioned, processing data like facial images (biometric data) for identification, health information, or data revealing race/ethnicity (which a resume or photo can do, see Figure 4 and Figure 7 for examples) is prohibited unless an exception applies (Art 9). One possible GDPR argument is that the data subjects "manifestly made public" these special-category data themselves – for instance, someone publicly posts their own photo, resume, or medical info. That exception (Art 9(2)(e)) might allow processing, but it's a gray area; arguably they made it public for a certain audience or purpose, not for any use whatsoever. Also, if the person in the photo is not the one who posted it (e.g., a news article about someone's health in Figure 8), the exception doesn't apply. DataComp CommonPool has many images of people taken by others (indicated by mentions of celebrity names in Section 6.1.2), so "manifestly public by the subject" fails. Thus for a GDPR-compliant approach, a controller would need either explicit consent from each person (impossible at scale), or to fit under the research exception (Art 9(2)(j)) which requires that processing be necessary for research in the public interest and subject to EU or member state law providing appropriate safeguards. A commercial company training a product likely cannot claim the research exemption; an academic might, but even then must implement safeguards like de-identification. Under CCPA/CPRA, "sensitive personal information" such as account passwords, financial info (found in Section 6.2.1), precise geo-location (found in Section 6.4.1), or contents of communications can be used by a business only for limited purposes (generally, what is necessary to provide the service, or as permitted with notice) if a consumer directs them to limit it. If a business were, say, using CommonPool and it contained login credentials or credit card numbers (which it does, in some images of documents shown in Figure 5), that's sensitive info that should never be exploited beyond necessary security research. OCPA goes further to require consent for any processing of sensitive data. In context, any use of a child's image or personal details from CommonPool without parental consent (see Section 6.3.1) is a clear violation of OCPA. Also, any biometric data (like using faces to improve a face recognition algorithm) would technically require prior consent of the individual in Oregon. Even if enforcement is unlikely, legally the obligation is there. So a controller would need to filter out all children's data and biometric identifiers or risk non-compliance. Another sensitive category is resumes — these often contain contact info, education, employment history. While not "sensitive" by statutory definition, they are highly personal. If a resume includes something like a Social Security number or driver's license (which some do, as shown in Figure 6), that becomes sensitive (government ID number). CommonPool was found to have images of passports and driver's licenses in Section 6.2.1, which are both sensitive and highly regulated (for example, storing driver's

license numbers triggers breach notification duties if breached, under various U.S. laws). Financial data like credit card numbers in the dataset raise data breach concerns: under all U.S. states' laws, if a company inadvertently exposed those, they'd owe notifications. Thus, even beyond privacy-specific laws, holding such data creates liability if it leaks or is hacked.

7.3.6 Security and breach notification. Privacy laws also require securing personal data. GDPR Article 32 mandates appropriate technical and organizational measures to protect data. CCPA requires "reasonable security" and provides a private right of action (lawsuit) for consumers whose sensitive data (like certain ID numbers) is breached due to lack of reasonable security. OCPA similarly obligates reasonable data security practices. In the context of DataComp, any entity storing the dataset or integrating it into systems must implement strong protections against unauthorized access. This is especially important because the dataset contains some very sensitive elements (e.g., full credit card details, identity documents). If, hypothetically, a company using CommonPool got hacked and the hackers obtained these personal entries, that company could face breach notification duties to potentially millions of individuals (though identifying and contacting them would be almost impossible, which doesn't absolve the duty). The inability to notify affected persons (because the data was scraped without emails or phone numbers perhaps) is a nightmare scenario, it means the company simply cannot fully comply with breach laws if an incident occurred. This is a legal risk of assembling data that you cannot trace back. Regulators might view the initial decision to compile such data as negligent if it could never be properly safeguarded or managed.

7.3.7 Automated decision-making and profiling. One might ask if GDPR's provisions on automated decisions (Article 22) apply. Article 22 gives individuals the right not to be subject to a decision based solely on automated processing (including profiling) that produces legal or similarly significant effects. Training an AI model on DataComp CommonPool doesn't directly make decisions about those individuals, so Article 22 isn't directly triggered by the training process. However, if the resulting model is used in a way that profiles or affects people, then those individuals have rights regarding how their data was used to create that model. This is a cutting-edge area: there's debate about a person's right to opt out of being included in training data that will be used in profiling. GDPR doesn't yet explicitly give a right to opt out of processing for AI model training (unless it's considered processing for a legitimate interest to which they object under Art 21). But some have argued using personal data to materially inform algorithmic decisions about people could trigger obligations of fairness or explanation. For example, if CommonPool were used to build a facial recognition system that is then used by police, EU citizens might challenge the legality of processing under GDPR's law enforcement provisions or human rights law.

7.3.8 Data Protection Impact Assessment (DPIA). Under GDPR Article 35, if processing is likely to result in high risk to individuals (especially using new technologies on a large scale with sensitive data), a DPIA must be conducted. A controller planning to use DataComp CommonPool should perform a DPIA evaluating the risks to rights and freedoms of individuals whose data is in the set. It would almost certainly conclude there are significant risks (e.g., unauthorized disclosure, bias, misuse of personal images). Mitigation measures (like additional filtering, encryption, access controls, or not using certain data) should then be taken. OCPA similarly requires documented risk assessments for processing that presents a heightened risk of harm, such as processing sensitive data or profiling that could lead to unfair outcomes. Training an AI on personal images might qualify as profiling with potential disparate impact (imagine the model reinforces biases or misidentifies certain demographic groups). Thus, these laws demand a proactive, documented examination of the privacy impacts of using CommonPool, something that currently, many AI practitioners might not be doing.

7.4 Were DataComp's own filtering and anonymization efforts legally sufficient?

The dataset creators did implement some privacy filters, notably, automated face blurring to obscure identities in images. However, our audit showed this was far from comprehensive: in Section 6.4.2 approximately 102 million images of real people's faces went unblurred due to the tool's failure to detect them. Legally, if one tries to anonymize personal data but the anonymization is incomplete, the data must still be treated as personal data. GDPR, for instance, considers data "anonymous" only if individuals are no longer identifiable taking into account all means reasonably likely to be used to identify them. A simple blur or pixelation on a face may not meet that standard, especially at scale — advances in AI can reverse blurring [83] or at least identify individuals from unblurred parts (hair, posture) or by correlating with other images [99]. Moreover, blurring the face in an image does nothing if the caption or surrounding text mentions the person's name or other info. In DataComp, even where faces were blurred, in Table 6 we find instances where the accompanying alt-text still states, "Photo of [Name] at [event]." That remains personal data. Thus, from a GDPR perspective, the dataset as released was not effectively anonymized and should be treated as personal data. The legal expectation for anonymization is very high (truly irreversible de-identification). Short of that, one might pursue pseudonymization, replacing identifiers with codes, but here the images are inherently identifying (a face is a unique identifier). The DataComp curators also did not engage in methods to remove obvious PII strings (using tools to detect things like emails or SSNs), and while we found plenty of ID numbers, names, and contacts, prior work shows that PII detection tools are not sufficient [88] and create a "*false sense of privacy*" [138]. This underscores that "no automated cleaning can remove all PII," as we demonstrated in our dataset audit. Privacy law would likely concur: if personal data remains, the controller cannot claim exemption by saying "we tried to filter it." Instead, the controller must continue to handle the data under applicable law or take further steps to mitigate risk.

One could ask: does blurring faces reduce the legal risk at all? It might mitigate it somewhat. For example, a fully blurred face might no longer be "biometric data" because you cannot recognize the person from it. If the blur is strong enough that the person is not reasonably identifiable, that particular image might fall out of definitions of personal data. However, if at least 102 million faces were missed (Section 6.4.2), the effort fails to appreciably lower

the overall risk. Additionally, partial mitigation could demonstrate awareness of privacy issues, which regulators could use to argue the controller knew of the risk yet didn't do enough. In the U.S., attempting to de-identify data can provide some safe harbor (like CCPA says de-identified data is not "personal information" if it meets certain criteria). But de-identified in that context means data that "cannot reasonably be used to infer information about, or otherwise link to, a particular consumer." Given the residual personal info in DataComp CommonPool, it's hard to argue it's de-identified. For example, an image showing a credit card with the numbers visible and a person's name (which our example findings revealed in Figure 5) is clearly identifiable to that cardholder. No amount of general dataset size changes that. Thus, legally, the filtering was not sufficient to escape privacy obligations. At best, it was an attempt at data protection by design, but an underinclusive one.

An interesting legal question is whether releasing the dataset with incomplete blurring could be seen as a form of data processing for research that is privileged. Some laws and courts recognize that publishing personal data for public interest research or journalism can be protected by freedom of expression. But here it's not journalistic, and the personal data belongs to numerous unsuspecting individuals whose interests were not considered individually. The lack of a specific legal basis (no consent, etc.) means that incomplete anonymization doesn't cure the issue; it just demonstrates that a risk was acknowledged. Regulators like the UK's ICO have fined companies even when they attempted anonymization that proved inadequate (e.g., characterizing poor de-identification as essentially an unauthorized disclosure of personal data).

7.5 Is relying on "publicly available" data a defensible legal strategy for AI datasets?

Relying on "publicly available" data may sound like a legal shortcut, but in the context of AI training datasets, it's increasingly a trapdoor. As privacy laws evolve, accessibility is no longer a proxy for permissibility. Laws like the GDPR, CCPA, and OCPA make clear that just because data is online doesn't mean it's free for the taking. All three legal regimes reject the simplistic notion that data is "public" merely because it can be accessed online. Without a reasonable understanding of user intent, context, and consent, sweeping up personal information from the web and calling it "public" is a legally risky and often indefensible strategy.

The DataComp CommonPool dataset does not qualify as "publicly available" data under state consumer privacy laws like the Oregon Consumer Privacy Act (OCPA) and the California Consumer Privacy Act (CCPA), and should not be exempt from legal protections. Though some of the information in CommonPool may have been posted online, the legal definition of "publicly available" is more nuanced than mere accessibility. Both the OCPA and CCPA impose specific conditions to prevent the misuse of personal data that individuals did not affirmatively and knowingly place into the public sphere for unrestricted use.

7.5.1 ***Indiscriminate scraping fails the "reasonable basis" standard.*** Under both the OCPA and the CCPA, information is not considered "publicly available" simply because it can be found on the internet. The laws require that a controller or business have a *reasonable basis* to believe that the data was lawfully made available to the public *by the consumer*. For example, OCPA 646A.570·(13)(b)(B) allows an exemption only where "a controller reasonably has understood [the data] to have been lawfully made available to the public by a consumer." Similarly, CCPA 1798.140(v)(2)(B)(i)(II) requires that the business "has a reasonable basis to believe" that the consumer made the information publicly available.

In the case of DataComp, this standard cannot be met. The dataset is created by automated systems that crawl and scrape data indiscriminately, without human oversight or consumer context. These scrapers cannot discern whether data was posted intentionally for public reuse or under restricted circumstances, such as within a social media profile, a comment section, a classroom forum, or a personal blog with limited viewership. As a result, they cannot reasonably determine consumer intent or consent.

What's more, the volume and automation of this data collection preclude any individualized assessment of context. If a business or controller is scraping billions of data points with no mechanism for filtering out user-restricted or audience-limited disclosures (such as sites from Section 6.3.2), it cannot credibly claim to have a "reasonable basis" for believing the data was lawfully made publicly available. The law contemplates thoughtful, contextual evaluation, not mass extraction based on surface availability.

This same concern is even more pronounced under the GDPR. Article 6 requires that personal data processing be grounded in a valid legal basis, such as consent, legitimate interest, or performance of a contract. Even publicly accessible data may still require a legal basis for further use. Moreover, Recital 47 of the GDPR states that reliance on "legitimate interest" must be balanced against the reasonable expectations of the data subject. A data subject posting on a message board, publishing a blog post, or uploading a photo cannot reasonably expect their content to be scraped, stored in perpetuity, and used to train AI models — especially if the image was uploaded before these technologies even existed (for instance, some resumes found in Section 6.2.2 or facial images found in Section 6.4.2). The absence of notice, transparency, or opportunity to object violates both the GDPR's fairness principle (Article 5(1)(a)) and the requirement of transparency under Articles 12–14. Without a valid legal basis and fair processing, scraping and reuse of such materials is not lawful, even if the content is technically accessible online.

7.5.2 ***Widely distributed media ≠ automatically public under privacy law.*** Both laws provide a narrow carveout for information made available through "widely distributed media." But that does not give blanket immunity to scraped web content. This provision exists to exclude traditional journalistic content and intentionally public communications, like letters to the editor or public government filings, not to exempt all content accessible via a search engine.

For instance, the CCPA makes clear that *biometric data* collected without the consumer's knowledge is *not* "publicly available," even if it was technically accessible. This reflects an underlying principle: data shared without meaningful understanding or consent is still protected.

Similarly, data scraped from discussion forums or social media platforms may be technically accessible but not "widely distributed"

in the sense intended by the law. Many platforms have shifting or ambiguous privacy settings, and users often do not realize that their content is publicly indexed, especially if they are not sophisticated about data privacy. This ambiguity undermines any argument that consumers clearly and affirmatively made such data available to the general public.

7.5.3 ***Disclosures to a limited audience are not "publicly available".*** California further clarifies that even when consumers disclose information online, it is not "publicly available" if it was disclosed to a specific person or group *with an expectation of audience limitation*. CCPA 1798.140(v)(2)(B)(i)(III) explicitly excludes from the "publicly available" exemption any data that the consumer shared *with audience restrictions*. Again, DataComp's scraping model does not (and cannot) distinguish between content shared globally and content disclosed to a limited group.

As such, data shared in online forums, academic or professional listservs, group chats, or social platforms with customizable privacy settings would often fall outside the CCPA's definition of publicly available information. If the dataset includes these types of data (and preliminary audits suggest that it does, e.g. a social security number originating from LinkedIn Figure 5, resumes from Pinterest in Table 4, and a screenshot of an online forum in Figure 12), they are clearly out of scope for the exemption.

7.5.4 ***Public access does not equal public availability under law.*** Finally, both the CCPA and OCPA rest on the understanding that "publicly available" is a legal term of art, not a synonym for "can be found online." Treating all internet-accessible information as "publicly available" would render the statutory exemptions meaningless and invite systemic abuse by companies that profit from mass scraping. The laws instead require a deeper inquiry into the *source*, *intent*, and *context* of the data shared.

The GDPR, while not using the term "publicly available" as a formal exemption, still requires controllers to consider context and user expectations. The European Data Protection Board has made clear that the publication of data online does not strip individuals of their rights under the GDPR. Even data shared voluntarily does not give downstream actors carte blanche to reuse it for incompatible purposes. Any secondary use, especially for high-impact applications like AI training, requires a fresh legal basis and must be compatible with the original context of collection (per GDPR Article 6(4)).

The act of scraping publicly *accessible* content does not transform it into "publicly available" data under the law. Both the CCPA and OCPA limit this exemption to cases where there is a reasonable, contextual understanding that the data was knowingly placed into the public domain. The GDPR goes further, requiring not just accessibility, but lawful processing grounded in purpose compatibility, transparency, and data subject rights.

The DataComp CommonPool dataset, by design, ignores these safeguards. It amasses personal data without meaningful legal justification, often in direct contradiction to user expectations and platform norms. Its indiscriminate, large-scale scraping practices circumvent not just the spirit but the letter of modern privacy and data protection laws. Policymakers and regulators should be skeptical of claims that internet scraping inherently falls outside privacy regulation. In reality, web-scraped datasets like CommonPool raise urgent legal and ethical questions that warrant scrutiny, not exemption.

7.6 Summary

In conclusion, the legal analysis shows that using a dataset like DataComp CommonPool creates significant compliance challenges under prevailing privacy laws. The personal data in the dataset is squarely within the scope of GDPR, CCPA, OCPA, etc., meaning entities cannot simply ignore those obligations. They must consider jurisdiction (very likely at least one law will apply), then fulfill duties of transparency, lawful basis, and data subject rights – tasks that, given the dataset's nature, are extremely burdensome if not impossible at scale. The attempts at anonymization (face blurring) were not sufficient to remove these obligations because large amounts of identifiable data remain. Indeed, those attempts, while well-intentioned, illustrate the difficulty of truly de-identifying unstructured big data. Relying on the data being "public" is not a silver bullet; privacy laws provide some leeway for public data but also contain important caveats and are backed by a broad consensus that privacy is not forfeited upon disclosure. Ultimately, current notice-and-consent frameworks falter in this scenario — individuals were neither notified nor asked. Thus, any organization leveraging DataComp CommonPool should adopt a very cautious approach: aggressively filter out known personal identifiers, limit the purposes to something justifiable, conduct risk assessments, and be prepared to cease using or delete portions of the data if individuals exercise their rights. They should also monitor legal developments, as regulators are actively grappling with how to apply existing laws to AI datasets and may issue guidance or take enforcement action (for example, enforcement against Clearview AI signals that mass scraping of faces is unacceptable under data protection law). In many ways, DataComp is a test case for the tension between innovation through massive data aggregation and compliance with privacy principles. Our analysis suggests that without changes, either in how datasets are curated or in the legal approach, there is a substantial compliance risk and a broader normative concern that individuals' privacy is being compromised at scale without the tools to effectively protect it.

8 DISCUSSION

We now present the recommendations and implications of our audit results and legal analysis in terms of using datasets like DataComp CommonPool to train models or for other purposes. Our discussion is not specific to the DataComp dataset, as it is likely that other web-scraped large-scale datasets contain similar risks of personal data, despite automated sanitization efforts. Based on the current landscape of scraping the internet [124], we use our findings to inform various recommendations for policymakers as well as for machine learning practitioners.

8.1 Recommendations for law & policy

8.1.1 ***State Attorneys General should enforce aggressively to preserve the integrity of "publicly available" exceptions.*** Attorneys general in states with comprehensive privacy laws, like

California and Oregon (among others), should act decisively to prevent the hollowing out of consumer privacy protections through the misuse of "publicly available" exceptions. The DataComp dataset represents a paradigmatic abuse: personal data scraped indiscriminately, at scale, and without regard for user expectations or context. Permitting companies to sidestep liability merely because data was technically accessible online eviscerates the spirit of these laws.

Both laws impose a "reasonable basis" standard for treating data as publicly available, a deliberately higher bar than mere access. Yet controllers exploiting datasets like DataComp often bypass this safeguard entirely, relying on automation and volume to collect content without any contextual analysis. This practice not only violates the letter of the law but undermines its purpose: to restore agency to individuals over their personal data.

Enforcement authorities should use existing statutory tools to investigate companies that use datasets like DataComp without proper diligence. They should:

- Challenge the assertion that scraped personal data—especially biometric, children's, or resume-related information—was lawfully made public by the data subject.
- Use their investigative powers to examine whether businesses using such data conducted meaningful context assessments.
- Issue interpretive guidance to clarify that mass scraping fails the "reasonable basis" test by default unless extraordinary safeguards are in place.
- Pursue enforcement actions against high-profile users of DataComp as a deterrent, signaling that privacy law will not permit large-scale circumvention through technical loopholes.

Enforcement is not only legally justified; it is normatively essential. The legitimacy of privacy laws depend on the idea that individuals retain rights over their personal data, even when that data is visible online. If regulators do not defend this principle, the web will become a de facto public domain for surveillance, profiling, and commodification.

8.1.2 ***State legislatures should close the web-scraping loophole and modernize the "publicly available" exception.*** Legislatures should act to modernize the "publicly available" exception in consumer privacy statutes by drawing clear lines against the misuse of scraped data. The current definitions were crafted in an era of limited data sharing and do not account for the reality of AI today, in which indiscriminate webscraping is used to vacuum up billions of personal records, often without the knowledge or consent of the data subject.

To address this, state laws should be amended to include the following reforms:

1. ***Express Carveout for Indiscriminately Scraped Data.*** Amend the definition of "publicly available" data to exclude personal data collected through automated or indiscriminate web-scraping, unless the controller can demonstrate that the data was lawfully made available to the public by the data subject with clear intent for unrestricted downstream reuse. This preserves limited, contextual reuse of truly public information (e.g., letters to the editor, government records) but shuts the door on practices like DataComp, where no intent, consent, or meaningful context is established.

2. ***Presumptive Protection for Sensitive or Contextualized Data.*** Automatically disqualify the following from the "publicly available" exemption: (i) sensitive data, including biometric data and children's data, even if accessible online; (ii) any data disclosed on platforms that allow audience restrictions, unless disclosed with a public license or tag; (iii) data disclosed in contexts where reuse is materially incompatible with the original purpose (e.g., training AI on personal essays or support forum posts). This approach aligns with reasonable expectations of privacy and recognizes that technical access does not equate to waiver of privacy interests.

3. ***Mandate Transparency and Attribution.*** Require any controller invoking the "publicly available" exemption to document: (i) the source of the data; (ii) why the data was considered publicly available; (iii) how the controller confirmed the user's intent and awareness; and (iv) whether the platform terms of service allowed for scraping and reuse. This shifts the burden of justification to the party exploiting the data, not the data subject.

Without legislative reform, the "publicly available" exemption becomes a backdoor for pervasive surveillance and privacy harm at scale. It was never meant to immunize AI companies from obligations simply because they used a webscraper. States that claim to lead on consumer privacy cannot permit exceptions that swallow the rule. To that end, we propose the following possible language to modernize the "publicly available" data exception to close the gap that allows AI developers to harvest massive quantities of personal data under the pretense that it is publicly available.

Section [X]: Amendments to the Definition of "Publicly Available" Information

> **(A)** *Revised Definition of Publicly Available Data*
> "Publicly available information" means any personal data that:
> (1) Is lawfully made available from federal, state, or local government records;
> (2) Is lawfully made available to the general public by the consumer with clear and affirmative intent for such information to be broadly accessible without restriction; or
> (3) Is available in widely distributed media intended for general public dissemination (such as news broadcasts or publicly licensed publications).
> **(B)** *Exclusions from Publicly Available Data*
> Notwithstanding subsection (A), the following categories of personal data shall not be considered "publicly available":
> (1) Personal data collected or processed through automated, large-scale, or indiscriminate webscraping methods, unless the controller demonstrates that:
> - (i) The data subject explicitly made the data publicly accessible with no audience restriction;

- (ii) The data subject had actual knowledge and intent to permit unrestricted downstream use; and
- (iii) The platform's terms of service clearly authorized such scraping and reuse.

(2) Personal data disclosed in contexts where audience limitation, expectation of privacy, or contextual sensitivity is apparent, including but not limited to:
- (i) Social media posts with non-public or limited visibility settings;
- (ii) Content from discussion boards, classroom platforms, or professional forums not intended for general indexing;
- (iii) Any disclosure where the data subject did not manifestly intend the information to be used for unrelated commercial or algorithmic training purposes.

(3) Sensitive data, including biometric data and a child's personal data.

(C) *Transparency and Documentation Requirements*
A controller or processor relying on the "publicly available" exception must maintain internal records demonstrating:

(1) The source of the data;
(2) The legal basis for concluding the data was publicly available as defined in this Section;
(3) That reasonable measures were taken to assess the data subject's intent and the original context of disclosure; and
(4) That any scraping or automated collection complied with the originating platform's access terms and community guidelines.

(D) *Purpose Limitation*
Data obtained under the "publicly available" exception may only be processed for purposes that are compatible with the context in which the data was originally disclosed. The use of such data for:
- (i) Training or developing algorithmic models;
- (ii) Profiling; or
- (iii) Commercial repurposing unrelated to the original context

shall not be presumed compatible without specific consumer consent.

By adopting a clarified and modernized exception to the definition of publicly available data, state legislatures can better align privacy law with the realities of contemporary data practices and the technical architectures of AI development. The proposed reform does not prohibit the use of public data outright; rather, it imposes necessary constraints on the indiscriminate scraping and repurposing of personal information in ways that disregard user context, intent, or consent. It operationalizes key privacy principles—purpose limitation, data minimization, and transparency—by ensuring that public accessibility is not conflated with unconditional legal availability. Importantly, it also harmonizes domestic privacy law with emerging international norms, particularly those reflected in the GDPR's emphasis on contextual fairness and lawful reuse. Where AI systems increasingly rely on large-scale ingestion of personal data, these clarifications are essential to preserving the normative foundations of privacy and data protection law. Without such legislative intervention, the exception for publicly available data risks becoming a structural loophole, one that undermines individual rights at scale and erodes the practical enforceability of privacy protections in the age of AI.

8.2 Challenges of existing privacy laws at scale

The findings of our audit illuminate deep structural tensions between the practices of large-scale dataset curation and the enforcement mechanisms of modern privacy law. As the collection, redistribution, and downstream usage of personal information becomes increasingly automated and dispersed, the foundational assumptions of existing privacy frameworks, namely individual control, meaningful consent, and data minimization, are rendered ineffective or outright obsolete. Below, we outline four interlocking challenges that arise when applying current privacy laws to web-scale data practices like those underpinning DataComp CommonPool.

8.2.1 The collapse of individual control. As articulated by privacy scholars, the model of "privacy self-management," which expects individuals to read privacy policies, understand potential downstream uses, and assert their rights has collapsed under the weight of modern data practices [123, 124]. Our audit underscores this failure: the individuals whose resumes, government IDs, or children's medical information appear in CommonPool could not have meaningfully understood or anticipated these downstream uses at the time of upload. Indeed, in many cases, the content was posted years before the rise of large-scale foundation models, making the notion of "informed consent" retroactively implausible. Even where opt-out mechanisms exist, such as the integration of Spawning AI with Hugging Face, these tools presume a level of awareness, technical skill, and effort that is unrealistic at scale. Data subjects are not just unaware that their data has been scraped; they are unaware that it ever could be. And even if they discover their data's inclusion in a training dataset, privacy law does not adequately address revocation of consent post hoc. For example, Section 6.3.2 of our audit shows that a nontrivial portion of CommonPool (over 21 percent of sampled URLs) now fails to download, including roughly 0.4 percent of samples (an estimated 50 million) from a site that currently requires a login. The dataset nonetheless retains the metadata and text of these entries, and models trained on earlier versions may retain the visual content. There is no legal mechanism to retroactively purge these artifacts. The focus on consent at the point of collection falters when the act of collection is hidden from the data subject. Furthermore, it fails to address the dynamic nature of web content: data that was once "public" may later be made private, but this change has no bearing on datasets already scraped. The law thus fails to honor either the data subject's evolving intentions or their right to meaningful withdrawal of consent.

8.2.2 Web-scale data is "too big to privacy". A second, compounding problem is that tracing data provenance at internet scale is functionally infeasible. As web pages and digital images propagate across the internet (shared, reposted, and mirrored) tracing

any given image or caption back to its original context becomes an exercise in futility. This makes it nearly impossible to assess whether scraped data was originally behind a login wall, taken from a compromised database, or uploaded with restricted permissions. Our audit provides a striking example in Figure 5: a social security number uploaded by an individual to a social media site was later scraped from an entirely different image hosting service, raising questions about provenance and the individual's original intent. This demonstrates that even if a user attempts to revoke their disclosure, downstream propagation renders such efforts ineffective. Section 6.3.2 further reveals that many previously accessible images have since become restricted. However, there is no reliable way to determine whether a login requirement was in place at the time of scraping, or if the access was revoked afterward, much less whether the takedown was initiated by the data subject themselves. In either case, the privacy harm remains: prior downloads, including by researchers and commercial entities, retain access to data that is no longer meaningfully public, with no obligation to re-evaluate its lawfulness. This collapse in traceability not only undermines consent but exposes the limits of relying on "public availability" as a legal safe harbor. The practical inability to distinguish between a deliberately public blog post and an inadvertently leaked medical record suggests that more robust frameworks are needed, ones that prioritize data integrity and context, not just accessibility.

8.2.3 Dataset monoculture and the waterfall of harm. A third challenge is the monoculture effect inherent in web-scale datasets: once personal data is included in a widely used dataset like CommonPool, it is replicated and amplified through every model trained on it. This is not merely a matter of one model misusing personal data, it is hundreds or thousands of models, potentially deployed across commercial, academic, and governmental contexts. Unlike previous concerns about model behavior, the risk here stems from the centralization of data sources. As described in Figure 1 and throughout Section 3, the pipeline of AI model development is premised on reusing existing datasets to build ever-larger and more generalizable systems. But the reuse of CommonPool, and its release as a URL index rather than a frozen corpus, means every download triggers a fresh crawl, potentially retrieving newly restricted or outdated content. The same image (if propagated to other image hosting sites) may be downloaded millions of times, long after the data subject has removed or placed the original behind a paywall. Moreover, the burden on data subjects to file data subject access requests (DSARs) for each instantiation of their data is untenable. Even if a subject were to discover their image in CommonPool, they would need to issue requests to every model developer who has used the dataset, a task rendered impractical by the absence of dataset tracking or provenance tools. And even if a request were granted, deleting personal data from a model's training corpus is a deeply unresolved technical and legal problem [65]. The transition from a "web of documents" to a "web of training data" demands a rethinking of data governance. The internet made data widely *available*; web-scraped AI pipelines have made data widely *processed*. This shift changes not just the scale, but the very stakes of privacy.

8.2.4 Incomplete anonymization and the limits of de-identification. Finally, while some developers attempt to mitigate risk through de-identification, our audit demonstrates the limits of those efforts. As detailed in Section 7.4, CommonPool employs face blurring and filters to remove sensitive content, but these methods are incomplete. An estimated 102 million images of real human faces are not blurred despite the default tooling, and OCR-extracted text still reveals credit card numbers, names, and even birth certificates. The result is a system in which purported anonymization is neither comprehensive nor verifiable. Worse, the existence of such flawed de-identification gives dataset developers a false sense of compliance, while leaving data subjects exposed to downstream inference, reidentification, and profiling. Privacy laws allow for certain processing exemptions where data has been sufficiently anonymized. However, as our audit makes clear, the reasonability of these claims collapses at scale. At a dataset of 12.8 billion samples, even a 0.1% failure rate translates into millions of instances of potential privacy harm. Legal frameworks must move beyond the binary of "identified" vs. "anonymous," and instead impose robust standards for anonymization that are tied to dataset size, processing purpose, and downstream risks. Moreover, data minimization principles must be enforced irrespective of data identifiability: collecting vast troves of "possibly anonymous" data still imposes measurable risk. Pending legislation like California's SB 2013 and Colorado's AI Act offer a glimpse of reform. These laws include requirements for dataset documentation and impact assessments. But they must go further, mandating privacy-preserving evaluations of curation practices and requiring public disclosure of filtering tools and their performance metrics.

8.3 Implications for machine learning practice

In this section, we discuss recommendations for *machine learning practitioners and researchers* based on our audit findings and legal analysis, as well as contributions from current work.

8.3.1 Misperceptions of publicly available data. As stated in Section 7.5, legally "publicly available" data is not equivalent to data that is "accessible" via web-scraping. The "publicly available" exception in various data protection laws may not to apply data posted from a breach or data behind a login screen, or in the case of GDPR, data not uploaded by the data subject themselves. Moreover, the "vacuuming" of the internet counters the data protection principle of data minimization to gather only the data "necessary for the stated purpose." The machine learning community should be aware of the distinction between data that is legally public versus data that is available online.

Based on current mechanisms for individual consent, opt-out policies should also be followed, despite the Robots Exclusion Protocol being unenforceable [104]. Websites that disallow web-scraping indicate a revoking of consent of their data being used. Datasets and web-scraping tools therefore should respect these protocols at the *site level* (rather than image-level [51] and *at the time of downloading* (see Section 6.3.2), to align with web-crawling best practices [134]. Sites like spawning.ai [7] that ask for individual opt-out may give some indication of consent, but may not be enough to remedy privacy harms, as it is implausible for individuals to know where their personal data exists on the internet.

8.3.2 General-purpose models. The development of foundation models that are agnostic to a purpose run counter to the data protection principle of purpose specification. Machine learning practitioners should explicitly define narrow use cases of training models [18], especially in cases where consent is required to collect training data. Large vision models can also be considered face recognition tools, as many vision-language models have been shown to generate or classify celebrities in their training datasets [26, 141]. This implies the collection of facial images as biometric data for recognition purposes, even if models are not explicitly trained to do so. Practitioners and researchers should be aware that prior notions of "facial geometry" may change as legal definitions and technologies evolve.

8.3.3 Researching web-scraped data. Specifically for researchers, we repeat the need to release datasets with constricted use licenses that are legally enforceable, such as the RAIL license [31] rather than restriction-free licenses when artifacts are not intended to be deployed for commercial use. While research often builds on top of other work for collaboration and advancements, these dependencies become more difficult to trace as literature proliferates — researchers should be more critical of the practice of training on web-scraped datasets just because prior work has done so. Even as research exemptions are included in existing privacy laws, there are still certain expectations that collected data for research aligns with data protection principles: GDPR safeguards in Article 81 [44], for instance, require data minimization and anonymization to fulfill the research purpose, and that the use of personal data does not affect individuals.

Our work reveals that privacy risks still remain with studying personal data on the internet for research purposes, as we highlight settings where data may not be considered "legally public" even if collected from the web. University institutional review boards should re-evaluate existing web-scraped datasets that previously gained IRB approval and re-examine their exemption protocols, given the amount of personally-identifiable information of real individuals found on the web (whether publicly available or not).

8.3.4 Alternatives to web-scraping. Given the wide usage of datasets like DataComp CommonPool, we ask, is it too late? Where do we go from here? We list several alternatives and concrete remediations, although we acknowledge that these approaches may not completely remove all personal information or downstream privacy harms.

(1) While automated sanitization methods may never guarantee complete removal of personal data [64, 138], proactive evaluation of data cleaning and justification of the use of existing techniques will alleviate privacy risk more than post-hoc audits.
(2) We encourage empirical research on implementing data minimization when training models through techniques like anonymization [54], early stopping [118], and data pruning [49].
(3) Dataset curators should ideally ask for explicit assent for inclusion in datasets or model training [8]. For example, Mozilla Common Voice [10] is a community-driven open speech corpus where contributors volunteer to record. On the text side, the Common Corpus is an open dataset of two trillion tokens that are either uncopyrighted or under permissible licenses [73].
(4) To address the propagation of images across websites, there have been several mechanisms to maintain attribution on the web, such as Adobe's Content Credentials [5]. Not only do Content Credentials label generated media [48], they are also attached to content when shared and can indicate creator's usage preferences, although the extent to which these standards are adopted remains to be seen.
(5) For large models that have already been trained, Lee [75] argues that algorithmic disgorgement may not be necessary. Technical interventions such as machine unlearning [95], data attribution [77] to avoid output attributed to personal data, and fine-tuning [119] may prevent privacy leakages, although we caution that these techniques require careful intervention, as prior work has found that fine-tuning can may increase the rate of training data extraction [79].

9 CONCLUSION

In this work we present an empirical analysis of a popular machine learning dataset and demonstrate that, even with intentions to remove PII, personal data can remain. These results can trigger existing privacy laws for downstream uses of web-scraped data.

ACKNOWLEDGMENTS

We are incredibly grateful for feedback from the 2025 Privacy Law Scholars Conference and especially comments from Inyoung Cheong. The first author is supported by the National Science Foundation Graduate Research Fellowship Program. This work was supported in part by U.S. National Science Foundation awards CCF-2045402 and CNS-2205171, the Carnegie Bosch Postdoctoral Fellowship, the University of Washington Tech Policy Lab, and a grant from the Simons Foundation.

REFERENCES

[1] CA Civ Code § 1798.192. 2018. California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5
[2] OR SB 619. 2018. Oregon Consumer Privacy Act of 2018. https://olis.oregonlegislature.gov/liz/2023R1/Downloads/MeasureDocument/SB619/Enrolled
[3] 15 U.S.C. § 6501. 1998. Children's Online Privacy Protection Act of 1998. https://uscode.house.gov/view.xhtml?req=granuleid%3AUSC-prelim-title15-section6501&edition=prelim
[4] Lura Abbott and Christine Grady. 2011. A systematic review of the empirical literature evaluating IRBs: What we know and what we still need to learn. *Journal of Empirical Research on Human Research Ethics* 6, 1 (2011), 3–19.
[5] Adobe. 2025. Content Credentials overview. https://helpx.adobe.com/creative-cloud/help/content-credentials.html
[6] Stability AI. 2025. https://stability.ai/news/stable-diffusion-public-release
[7] Spawning AI. 2025. Spawning API. https://api.spawning.ai/spawning-api
[8] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. 2023. Ethical considerations for responsible data curation. *Advances in Neural Information Processing Systems* 36 (2023), 55320–55360.
[9] Internet Archive. 2013. Wayback Machine APIs. https://archive.org/help/wayback_api.php
[10] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common Voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670* (2019).

[11] Romain Beaumont. 2022. Clip Retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval.

[12] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* (2001), 1165–1188.

[13] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. 2024. Into the LAION's den: Investigating hate in multimodal datasets. *Advances in Neural Information Processing Systems* 36 (2024).

[14] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[15] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[16] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. SoK: AI Auditing: The Broken Bus on the Road to AI Accountability. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning*.

[17] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems* 35 (2022), 3663–3678.

[18] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[19] Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A Smith, and Christopher A Choquette-Choo. 2025. Privacy Ripple Effects from Adding or Removing Personal Information in Language Model Training. *arXiv preprint arXiv:2502.15680* (2025).

[20] Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. 2021. When is memorization of irrelevant training data necessary for high-accuracy learning?. In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*. 123–132.

[21] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[22] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology* 4 (2002), 217–231.

[23] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)* 290, 1-34 (2008), 5–12.

[24] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226* (2023).

[25] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*. 5253–5270.

[26] Yunzhuo Chen, Nur Al Hasan Haldar, Naveed Akhtar, and Ajmal Mian. 2023. Text-image guided Diffusion Model for generating Deepfake celebrity interactions. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 348–355.

[27] Cloudflare. 2024. Cloudflare API v4 documentation: Get multiple domain details. https://developers.cloudflare.com/api/operations/domain-intelligenceget-multiple-domain-details

[28] Samantha Cole. 2023. Largest dataset powering AI images removed after discovery of child sexual abuse material. *404 Media* 20 (2023).

[29] Federal Trade Commission. 2025. COPPA Safe Harbor Program. https://www.ftc.gov/enforcement/coppa-safe-harbor-program

[30] Creative Commons. 2025. CC BY 4.0. https://creativecommons.org/licenses/by/4.0/deed.en

[31] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022. Behavioral use licensing for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 778–788.

[32] Common Crawl. 2025. Common Crawl. https://commoncrawl.org

[33] Common Crawl. 2025. Frequently asked questions. https://commoncrawl.org/faq

[34] DataComp. 2023. DataComp. https://github.com/mlfoundations/datacomp

[35] DataComp. 2023. Is there overlap between common-pool and laion-5B? https://github.com/mlfoundations/datacomp/issues/19

[36] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2 (2021), 20539517211035955.

[37] Meera A Desai, Irene V Pasquetto, Abigail Z Jacobs, and Dallas Card. 2024. An archival perspective on pretraining data. *Patterns* 5, 4 (2024).

[38] Mark Díaz, Sunipa Dev, Emily Reif, Emily Denton, and Vinodkumar Prabhakaran. 2024. SoUnD Framework: Analyzing (So) cial Representation in (Un) structured (D) ata. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 371–383.

[39] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

[40] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2023. Reducing Privacy Risks in Online Self-Disclosures with Language Models. *arXiv preprint arXiv:2311.09538* (2023).

[41] Brianna Dym and Casey Fiesler. 2020. Ethical and Privacy Considerations for Research Using Online Fandom Data. *Transformative works and cultures* 33 (2020).

[42] EasyOCR. 2025. EasyOCR. https://www.jaided.ai/easyocr/

[43] Benj Edwards. 2022. Artist finds private medical record photos in popular AI training data set. https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/

[44] European Parliament and Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. https://data.europa.eu/eli/reg/2016/679/oj

[45] Hugging Face. 2025. https://huggingface.co/api/datasets/mlfoundations/datacomp_pools?expand%5B%5D=downloads&expand%5B%5D=downloadsAllTime

[46] Casey Fiesler and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society* 4, 1 (2018), 2056305118763366.

[47] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems* 36 (2024).

[48] Dilrukshi Gamage, Dilki Sewwandi, Min Zhang, and Arosha K Bandara. 2025. Labeling Synthetic Content: User Perceptions of Label Designs for AI-Generated Content on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–29.

[49] Prakhar Ganesh, Cuong Tran, Reza Shokri, and Ferdinando Fioretto. 2024. The data minimization principle in machine learning. *arXiv preprint arXiv:2405.19471* (2024).

[50] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[51] Github. 2023. img2dataset ignores X-Robots-Tag. https://github.com/rom1504/img2dataset/issues/298

[52] Github. 2023. Implement Robots.txt support. https://github.com/rom1504/img2dataset/issues/48

[53] Github. 2023. Metadata download error - OSError: Consistency check failed. https://github.com/mlfoundations/datacomp/issues/33

[54] Abigail Goldsteen, Gilad Ezov, Ron Shmelkin, Micha Moffie, and Ariel Farkash. 2022. Data minimization for GDPR compliance in machine learning models. *AI and Ethics* 2, 3 (2022), 477–491.

[55] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2021. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714* (2021).

[56] Ritwik Gupta, Leah Walker, Rodolfo Corona, Stephanie Fu, Suzanne Petryk, Janet Napolitano, Trevor Darrell, and Andrew W Reddie. 2024. Data-Centric AI Governance: Addressing the Limitations of Model-Focused Policies. *arXiv preprint arXiv:2409.17216* (2024).

[57] Jack Hardinges, Elena Simperl, and Nigel Shadbolt. 2024. We must fix the lack of transparency around the data used to train foundation models. *Harvard Data Science Review (Special Issue 5). https://doi. org/10.1162/99608f92. a50ec6e6* (2024).

[58] Woodrow Hartzog. 2018. The case against idealising control. *Eur. Data Prot. L. Rev.* 4 (2018), 423.

[59] Woodrow Hartzog. 2019. The Public Information Fallacy. *BUL Rev.* 99 (2019), 459.

[60] Woodrow Hartzog and Evan Selinger. 2015. Surveillance as loss of obscurity. *Wash. & Lee L. Rev.* 72 (2015), 1343.

[61] Carol A Heimer and JuLeigh Petty. 2010. Bureaucratic ethics: IRBs and the legal regulation of human subjects research. *Annual Review of Law and Social Science* 6, 1 (2010), 601–626.

[62] Benjamin Henne, Maximilian Koch, and Matthew Smith. 2014. On the awareness, control and privacy of shared photo metadata. In *International Conference on Financial Cryptography and Data Security*. Springer, 77–88.

[63] Dennis D Hirsch. 2020. From Individual Control to Social Protection: New Paradigms for Privacy Law in the Age of Predictive Analytics'(2020). *Md L Rev* 79 (2020), 439.

[64] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. 2024. Who's in and who's out? A case study of multimodal CLIP-filtering in DataComp.

In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–17.
[65] Jevan Hutson and Ben Winters. 2024. America's next" Stop Model!" Model Deletion. *Geo. L. Tech. Rev.* 8 (2024), 124.
[66] ICO. 2025. Overview – Data Protection and the EU. https://ico.org.uk/for-organisations/data-protection-and-the-eu/overview-data-protection-and-the-eu/
[67] iKeepSafe. 2025. Certified Products. https://ikeepsafe.org/products/#coppa
[68] Mehtab Khan and Alex Hanna. 2022. The subjects and stages of ai dataset development: A framework for dataset accountability. *Ohio St. Tech. LJ* 19 (2022), 171.
[69] kidSAFE. [n. d.]. kidSAFE Seal Program Member List. https://www.kidsafeseal.com/certifiedproducts.html
[70] Tadayoshi Kohno, Yasemin Acar, and Wulf Loh. 2023. Ethical frameworks and computer security trolley problems: Foundations for conversations. In *32nd USENIX Security Symposium (USENIX Security 23)*. 5145–5162.
[71] LAION. 2025. Privacy Policy. https://laion.ai/privacy-policy/
[72] LAION. 2025. Releasing RE-LAION 5B: Transparent iteration on LAION-5B with additional safety fixes. https://laion.ai/blog/relaion-5b/
[73] Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee, Catherine Arnett, Pavel Chizhov, Eliot Krzystof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P Yamshchikov. 2025. Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training. *arXiv preprint arXiv:2506.01732* (2025).
[74] Clément Le Ludec, Maxime Cornet, and Antonio A Casilli. 2023. The problem with annotation. Human labour and outsourcing between France and Madagascar. *Big Data & Society* 10, 2 (2023), 20539517231188723.
[75] Christina Lee. 2025. Beyond Algorithmic Disgorgement: Remedying Algorithmic Harms. (2025).
[76] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.
[77] Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023. A survey of large language models attribution. *arXiv preprint arXiv:2311.03731* (2023).
[78] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 13094–13102.
[79] Zhangheng Li, Junyuan Hong, Bo Li, and Zhangyang Wang. 2024. Shake to leak: Fine-tuning diffusion models can amplify the generative privacy risk. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 18–32.
[80] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. 2024. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence* 6, 8 (2024), 975–987.
[81] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. 2024. Consent in crisis: The rapid decline of the ai data commons. In *Advances in Neural Information Processing Systems*.
[82] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 346–363.
[83] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. 2016. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408* (2016).
[84] Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE access* 8 (2020), 142642–142668.
[85] Hanna F Menezes, Arthur SC Ferreira, Eanes T Pereira, and Herman M Gomes. 2021. Bias and fairness in face detection. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 247–254.
[86] Microsoft. 2025. Presidio. https://microsoft.github.io/presidio/
[87] Midjourney. 2025. https://www.midjourney.com/home
[88] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438* (2024).
[89] Mozilla. 2025. X-Robots-Tag. https://developer.mozilla.org/en-US/docs/Web/HTTP/Reference/Headers/X-Robots-Tag
[90] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.
[91] Arvind Narayanan and Vitaly Shmatikov. 2006. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105* (2006).
[92] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035* (2023).
[93] Clemens Neudecker, Konstantin Baierer, Maria Federbusch, Matthias Boenig, Kay-Michael Würzner, Volker Hartmann, and Elisa Herrmann. 2019. OCR-D: An end-to-end open source OCR framework for historical printed documents. In *Proceedings of the 3rd international conference on digital access to textual cultural heritage*. 53–58.
[94] Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of OCR evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*. 13–18.
[95] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299* (2022).
[96] Helen Nissenbaum. 2011. A contextual approach to privacy online. *Daedalus* 140, 4 (2011), 32–48.
[97] Helen Nissenbaum. 2020. Protecting privacy in an information age: The problem of privacy in public. In *The ethics of information technologies*. Routledge, 141–178.
[98] Future of Privacy Forum. 2024. An Omnibus Definition of "Sensitive Data" Across Comprehensive State Privacy Laws. https://cdn.sanity.io/files/3tzzh18d/production/eac1440d340a728f1f2c00ab6c27aff446bce67d.pdf
[99] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. 2016. Faceless person recognition: Privacy implications in social media. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 19–35.
[100] OpenAI. 2025. ChatGPT. https://openai.com/chatgpt/overview/
[101] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. 2018. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8466–8475.
[102] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*. 3686–3695.
[103] PaddleOCR. 2025. PaddleOCR. https://paddlepaddle.github.io/PaddleOCR/latest/en/index.html
[104] Katie Paul. 2024. Multiple AI companies bypassing web standard to scrape publisher sites, licensing firm says. https://www.reuters.com/technology/artificial-intelligence/multiple-ai-companies-bypassing-web-standard-scrape-publisher-sites-licensing-2024-06-21/
[105] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021).
[106] Charles Preston. 2020. List of religious populations. britannica.com/topic/List-of-religious-populations
[107] PRIVO. [n. d.]. COPPA Safe Harbor Program. https://www.privo.com/coppa-safe-harbor-program
[108] Paul Quinn and Gianclaudio Malgieri. 2021. The difficulty of defining sensitive data—the concept of sensitive data in the EU data protection framework. *German Law Journal* 22, 8 (2021), 1583–1612.
[109] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[110] Joel R Reidenberg. 2014. Privacy in public. *U. Miami L. Rev.* 69 (2014), 141.
[111] Neil Richards and Woodrow Hartzog. 2018. The pathologies of digital consent. *Wash. UL Rev.* 96 (2018), 1461.
[112] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
[113] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. 2022. Toppling top lists: Evaluating the accuracy of popular website lists. In *Proceedings of the 22nd ACM Internet Measurement Conference*. 374–387.
[114] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.
[115] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.
[116] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.
[117] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. 2020. Diagnosing gender bias in image recognition systems. *Socius* 6 (2020), 2378023120967171.

[118] Divya Shanmugam, Fernando Diaz, Samira Shabanian, Michèle Finck, and Asia Biega. 2022. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 839–849.

[119] Tanmay Singh, Harshvardhan Aditya, Vijay K Madisetti, and Arshdeep Bahga. 2024. Whispered tuning: Data privacy preservation in fine-tuning llms through differential privacy. *Journal of Software Engineering and Applications* 17, 1 (2024), 1–22.

[120] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, Vol. 2. IEEE, 629–633.

[121] Daniel J Solove. 2013. Privacy self-management and the consent dilemma. *Harvard Law Review* 126 (2013), 1880.

[122] Daniel J Solove. 2024. Artificial intelligence and privacy. *Florida L. Rev.* (2024).

[123] Daniel J Solove. 2024. Murky consent: an approach to the fictions of consent in privacy law. *BUL Rev.* 104 (2024), 593.

[124] Daniel J Solove and Woodrow Hartzog. 2024. The Great Scrape: The clash between scraping and privacy.

[125] Daniel J Solove and Woodrow Hartzog. 2024. Kafka in the Age of AI and the Futility of Privacy as Control. *BUL Rev.* 104 (2024), 1021.

[126] Alicia Solow-Niederman. 2022. Information privacy and the inference economy. *Nw. UL Rev.* 117 (2022), 357.

[127] Yang Sun, Ziming Zhuang, and C Lee Giles. 2007. A large-scale study of robots.txt. In *Proceedings of the 16th international conference on World Wide Web*. 1123–1124.

[128] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* 671, 2000 (2000), 1–34.

[129] Omer Tene and Jules Polonetsky. 2012. Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.* 11 (2012), 239.

[130] David Thiel. 2023. *Identifying and eliminating csam in generative ml training data and models.* Technical Report. Technical Report. Stanford University, Palo Alto, CA. https://purl. stanford ....

[131] Charlotte A Tschider. 2020. Meaningful choice: A history of consent and alternatives to the consent myth. *NCJL & Tech.* 22 (2020), 617.

[132] DR Vedhaviyassh, R Sudhan, G Saranya, Mozhgan Safa, and D Arun. 2022. Comparative analysis of easyocr and tesseractocr for automatic license plate recognition using deep learning algorithm. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*. IEEE, 966–971.

[133] Salomé Viljoen. 2021. A relational theory of data governance. *The Yale Law Journal* (2021), 573–654.

[134] W3C. 2011. Write Web Crawler. https://www.w3.org/wiki/Write_Web_Crawler

[135] Sandra Wachter and Brent Mittelstadt. 2019. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.* (2019), 494.

[136] Ari Ezra Waldman. 2014. Privacy as trust: Sharing personal information in a networked world. *U. Miami L. Rev.* 69 (2014), 559.

[137] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* (2019).

[138] Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2025. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. *arXiv preprint arXiv:2504.21035* (2025).

[139] Kaiyu Yang, Jacqueline H Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2022. A study of face obfuscation in imagenet. In *International Conference on Machine Learning*. PMLR, 25313–25330.

[140] Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and César A Hidalgo. 2016. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific data* 3, 1 (2016), 1–16.

[141] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, Wanrong Huang, and Yuhua Tang. 2025. MagicNaming: Consistent Identity Generation by Finding a "Name Space" in T2I Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10439–10447.

[142] Wengang Zhou, Houqiang Li, Yijuan Lu, and Qi Tian. 2012. Principal visual word discovery for automatic license plate detection. *IEEE transactions on image processing* 21, 9 (2012), 4269–4279.

[143] Michael Zimmer. 2018. Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media+ Society* 4, 2 (2018), 2056305118768300.

A STAKEHOLDER NETWORK

Figure 11 provides an overview of the data flow between stakeholders defined in Section 3.

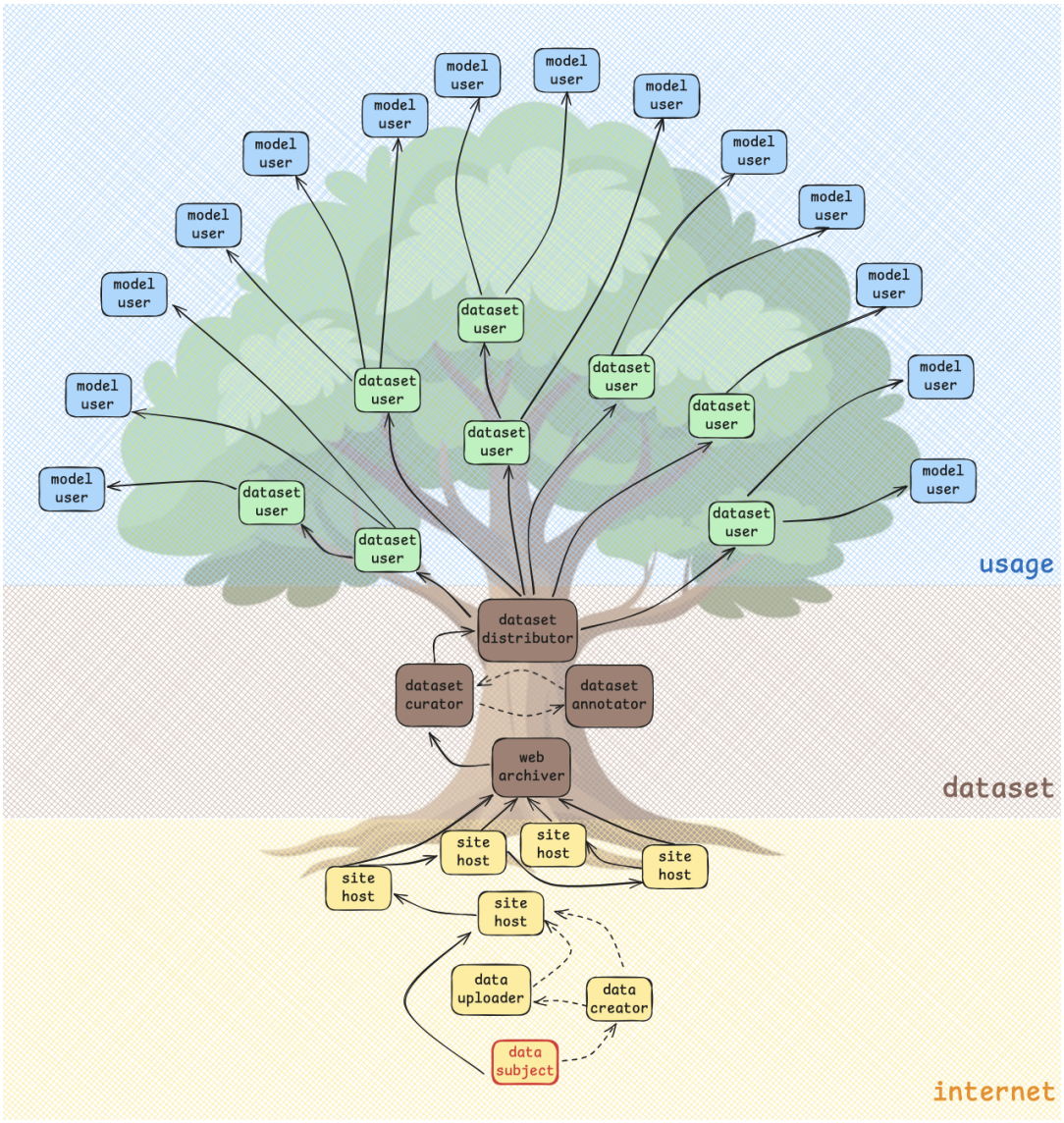
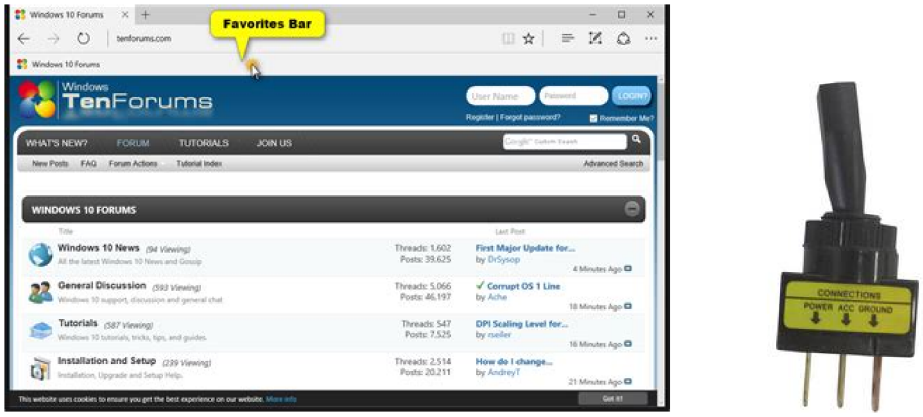


Figure 11: The stakeholder network demonstrates the potential flow of personal information between actors in the Internet (yellow), Dataset (brown), and Usage (blue) stages. Personal information starts from the data subject and is uploaded to the site host, potentially due to the data creator or data uploader without the data subject's knowledge (in dashed lines). Data may pass between site hosts through crawling and uploading, but is all aggregated by the web archiver. The dataset curator gathers from the web archiver and may optionally rely on an eternal dataset annotator (in dashed lines). The curated URL table is given to the dataset distributor who in turn passes data to multiple dataset users, who may create other datasets to pass to other dataset users, or deploy a model for various model users. This diagram demonstrates the trunk as the centralized source before wide dissemination of a popular training dataset like CommonPool with over two million downloads.

B METHODOLOGY DETAILS

In this section, we provide additional details on our audit methodology for OCR method selection.



(a) Screenshot of webpage **(b) Picture of product**

Figure 12: Example CommonPool images that contain text

OCR method	Accuracy	Precision	Recall
EasyOCR [42]	0.426	0.442	0.469
PaddleOCR [103]	**0.579**	**0.567**	**0.684**
Tesseract [120]	0.168	0.184	0.177
TROCR [78]	0.054	0.049	0.078

Table 7: Evaluation of OCR tools according to bag-of-words accuracy, precision, and recall.

B.1 Optical character recognition evaluation

We consider several open-source state-of-the-art optical character recognition (OCR) methods: EasyOCR [42], PaddleOCR [103], Tesseract [120], and TROCR [78]. Because most OCR methods are intended for handwriting detection or document extraction [84, 132], we perform our own evaluation for web-scraped images in CommonPool. We subsequently manually annotate the visible, legible text contained in 100 randomly-selected images and treat these annotations as ground truth. Images in our evaluation set often include screenshots or products of varying image quality (see Figure 12 for examples), which present text differently from typical OCR use cases.

Metrics: As shown in prior work [94], OCR evaluation metrics are not consistent, as the particular choice of metric may change which OCR tool is considered more accurate. In particular, character error rate (CER) based on Levenshtein distance [76] is often used as a performance metric, which relies on a particular ordering of words in documents. In our analysis, however, ordering does not matter for screenshots or products and not as necessary for keyword queries. We also desire word accuracy rather than close characters, since the words are later searched or fed as input into Presidio's named entity recognition model. As such, we rely on the bag-of-words model for our OCR evaluation, which tracks the number of words that are accurately recognized [93].

Results: Table 7 shows that on our evaluation set, PaddleOCR outperforms the other methods on every bag-of-word metric. Specifically, PaddleOCR has higher recall than the other methods, and thus encompasses the most recognizable words, which we then use to flag samples and subsequently manually verify due to the error in OCR.

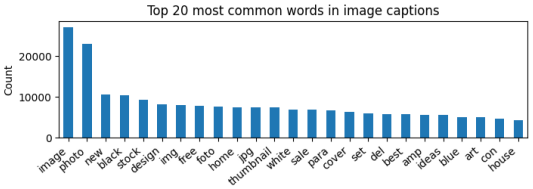
C ADDITIONAL AUDIT RESULTS

In this section we provide additional audit results.

C.1 Text visualizations

In this section we show text visualizations as a cursory analysis of the types of content in CommonPool. Figure 13 shows that image captions are related to stock photos or describing images, which aligns with their usage as alt-text [23]. Figure 14 also shows that detected text in images describes stock photos and images, but in addition we observe that `invoice` is one of the most common words identified in the OCR-extracted text. We find several images of invoices that display business and customer names and addresses, as well as payments issued between them. While these invoices are not considered personal information, they may reveal corporate information that may not be intended to be public.

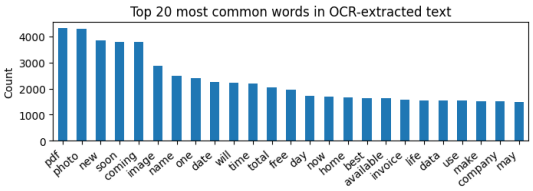
(a) Word cloud where size of the word corresponds to relative size.



(b) Top 20 most common words.

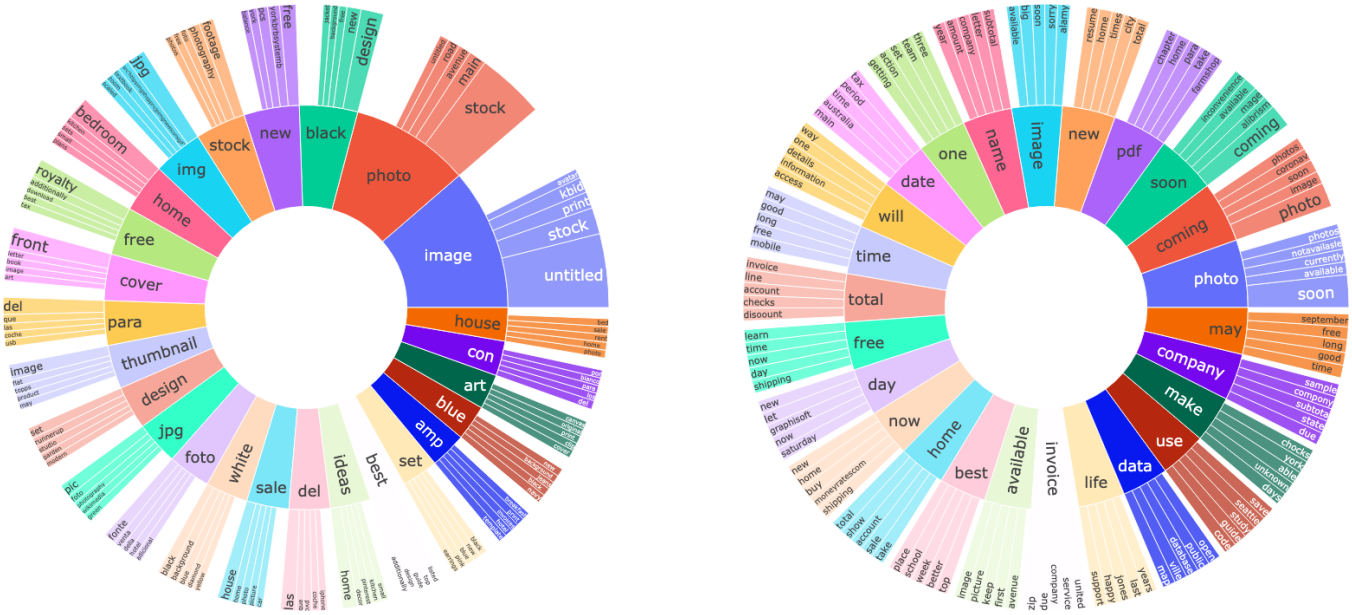
Figure 13: Word visualizations of captions (without stop words) of a 1 million random subsample of CommonPool.

(a) Word cloud where size of the word corresponds to relative size.



(b) Top 20 most common words.

Figure 14: Word visualizations of OCR-extracted text (without stop words) of a 1 million random subsample of CommonPool.



(a) Caption.

(b) OCR-extracted text.

Figure 15: Bigram disk visualizations in the caption and OCR-extracted text of 1 million random subsamples of CommonPool. The words in the first ring have sizes relative to their frequency, but isolated to the top 25 non-stop words for visibility. The outer ring depicts frequent words that appear directly after the corresponding inner word. Within a single segment, the outer words have sizes relative to their frequency, but isolated to the top 5 words within its inner segment and bounded for readability.

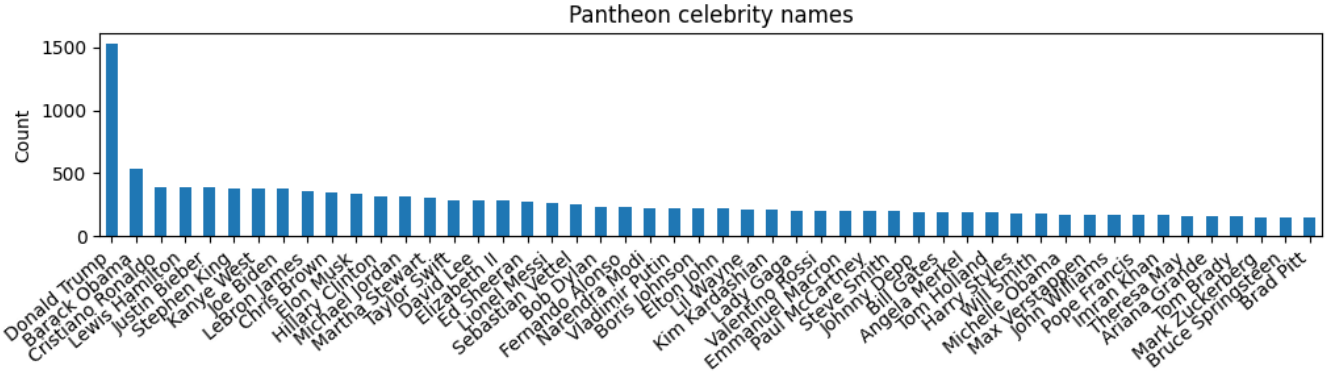
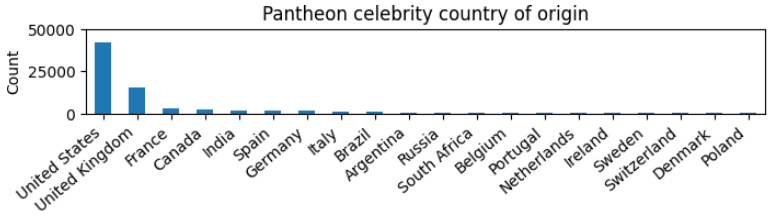


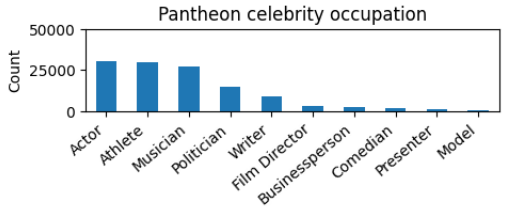
Figure 16: Top 50 most common celebrities from Pantheon 2020 dataset [140] mentioned in CommonPool captions and OCR-extracted text.

C.2 Celebrity name search

This section provides results from searching for celebrity names from the Pantheon dataset (Figure 16 and Figure 17). Figure 18 also shows the most common names using Presidio outside of brands.



(a) Top 20 most common countries of origin.



(b) Top 20 most common occupations.

Figure 17: Additional bar graphs from searching Pantheon 2020 celebrity names [140] from CommonPool captions and OCR-extracted text.

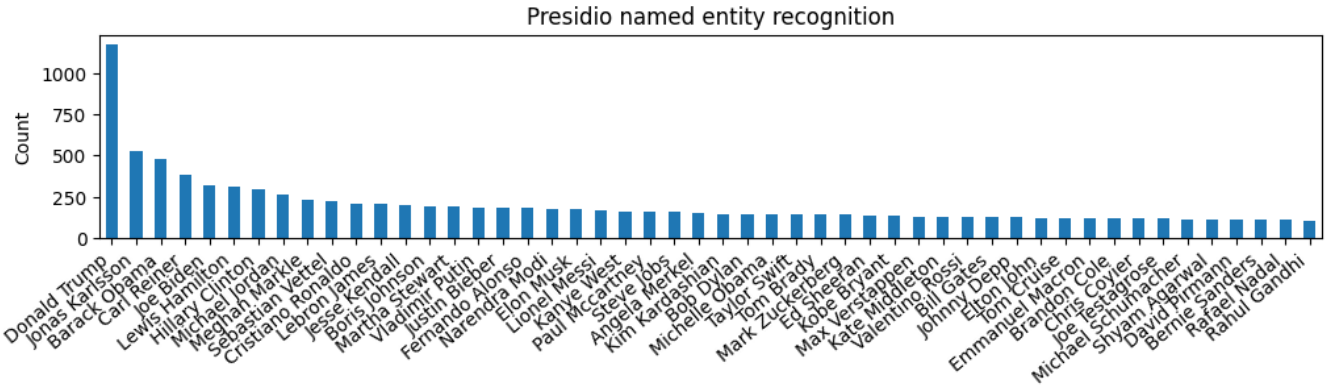


Figure 18: Top 50 most common Presidio-detected names from CommonPool captions and OCR-extracted text.

C.3 Resume documents

In this section, we show additional results of the associated geographic locations of the resume documents associated with online presence of individuals. Figure 19 refers to the disclosed address, while Figure 20 refers to the disclosed national origin or citizenship.

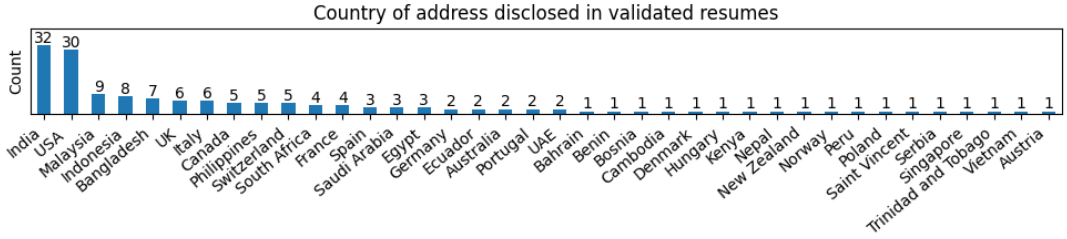


Figure 19: Sample count breakdown of country of address disclosed by validated resumes.

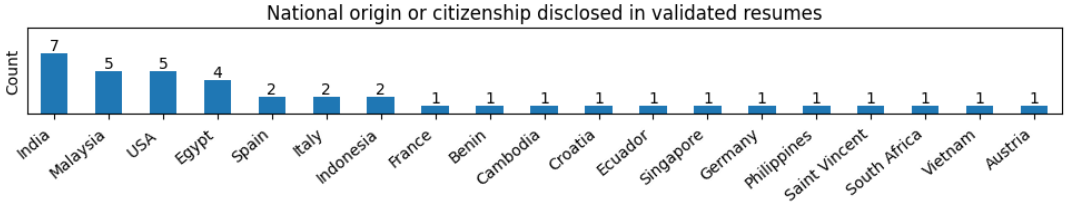


Figure 20: Sample count breakdown of national origin or citizenship disclosed by validated resumes.

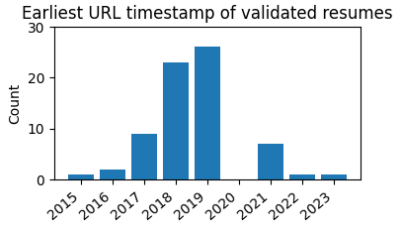
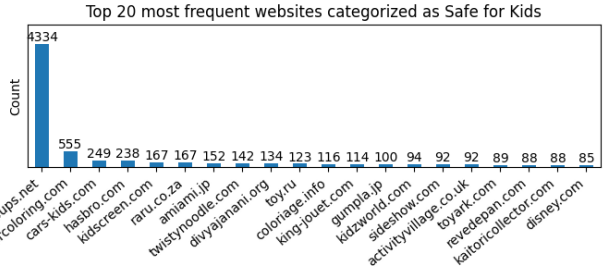


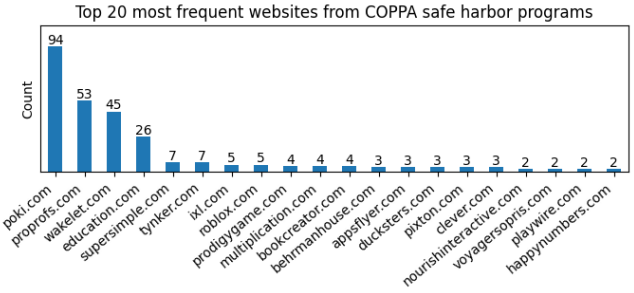
Figure 21: Earliest timestamp of URLs of validated resumes according to Internet Archive's Wayback Machine [9]. Only 70 of 168 validated resume URLs had existing records.

C.4 Children's information

Figure 22 gives additional breakdown of popular websites relating to children based on our two approaches: Cloudflare URL categorization and COPPA safe harbor program membership described in Section 6.3.1.



(a) Sample counts of top 20 websites categorized by Cloudflare as `Safe for Kids`.



(b) Sample counts of top 20 websites that are members of COPPA safe harbor programs [29].

Figure 22: Website frequency of children-related information.

C.5 Download errors of unavailable images

Figure 23 gives a breakdown of most common HTTP errors for images that failed to download, while Figure 24a shows the earliest recorded timestamp according to the Wayback Machine. Table 8a refers to two-sample t-tests to measure statistical differences between samples that failed or succeeded in downloading.

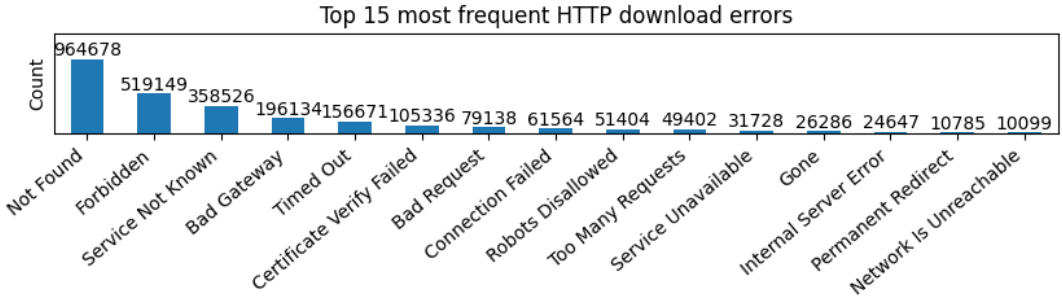
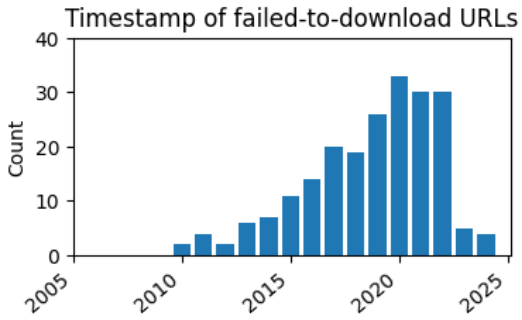
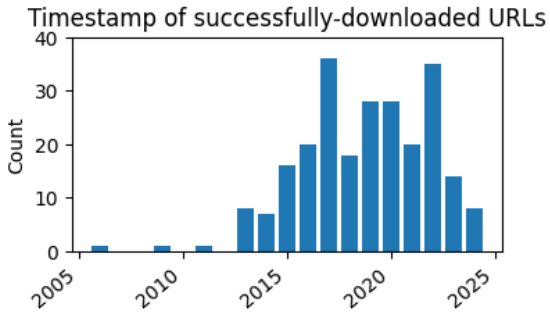


Figure 23: Sample counts of top 15 most common HTTP errors for images that failed to download during a download version run in April 2025.



(a) *Failed-to-download*: 213 out of 1000 image URLs had existing records.



(b) *Successfully-downloaded*: 241 out of 1000 image URLs had existing records.

Figure 24: Sample counts by year of earliest timestamps according to the Wayback Machine records for a random subsample of image URLs during a download version run in April 2025.

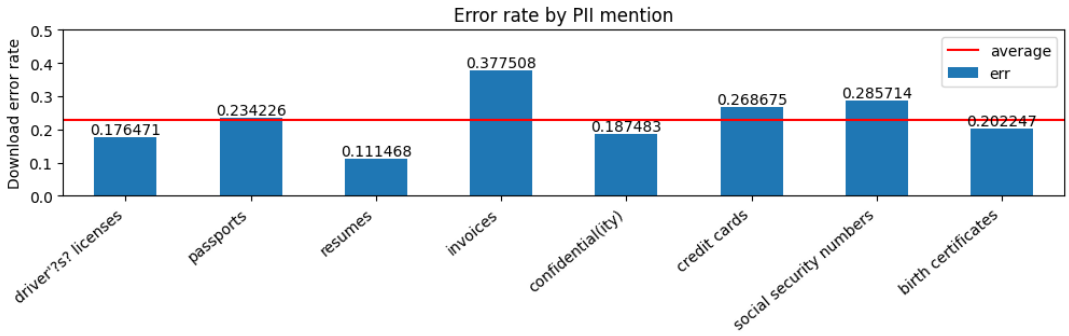


Figure 25: Download error rate for samples grouped by regular expression matches to instances of personal information. The average download error rate is plotted as a red line.

Image-related annotations	μ_1	μ_2	Adjusted p-value
CLIP-similarity score	0.207	0.211	$p < 0.001$
Number of detected faces	0.558	0.587	$p < 0.001$
Image size	390.2	404.0	$p < 0.001$

(a) For each row, μ_1 refers to the sample mean of the variable for 1000 randomly-selected *successfully-downloaded images*, while μ_2 refers to the sample mean for 1000 randomly-selected *failed-to-download images*. The alternative hypothesis is $\mu_1 < \mu_2$. The CLIP-similarity score refers to the cosine similarity in CLIP embeddings of the caption and image [109]. The number of detected faces is according to DataComp's SCRFD algorithm [55].

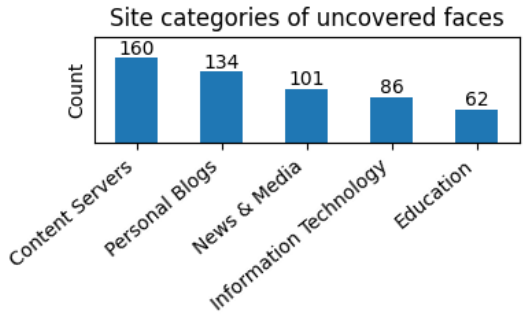
Image-related variable	μ_1	μ_2	Adjusted p-value
Bounding box area	100.4	29.6	$p < 0.001$
Pixel brightness	119.6	98.2	$p < 0.001$
Predicted `Female`	0.42	0.33	$p < 0.01$
Predicted age	34.3	26.2	$p < 0.001$

(b) For each row, μ_1 refers to the sample mean of the variable for images with manually confirmed human faces *detected (and therefore blurred)* by DataComp's face detection algorithm, while μ_2 refers to the sample mean for images *undetected (and therefore not blurred)*. The alternative hypothesis is $\mu_1 > \mu_2$. "Predicted" refers to the Rekognition gender and age annotations, where "Predicted `Female`" represents the the proportion among all images classified as `Female`, and "Predicted age" represents the average age prediction.

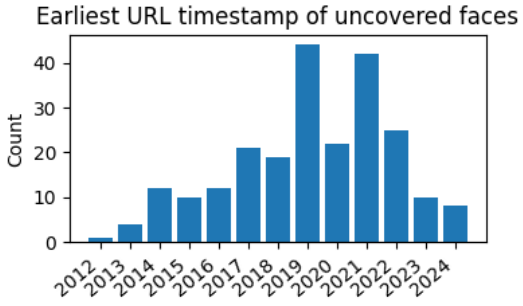
Table 8: Summary of two-sample one-tailed t-tests of various image-related annotations, adjusted by the Benjamini-Yekutieli procedure to control the false discovery rate for dependent tests [12].

C.6 Face detection

Table 8b reports the differences in the sample means of the distributions of manually confirmed human faces that are blurred versus not blurred by DataComp. Figure 26 includes analysis of the URL origins of the real faces uncovered by SCRFD.



(a) Cloudflare categorizations [27] of URLs.



(b) Earliest timestamp of URLs according to Internet Archive's Wayback Machine [9] where 230 out of 854 image URLs had existing records.

Figure 26: Analysis of website URLs of manually confirmed images of faces not caught by SCRFD.

Received 2025