Copyright Infringement Issues and Mitigations in Data for Training Generative AI

Copyright Infringement Issues and Mitigations in Data for Training Generative AI

1st Anna Arnaudo
Department of Control and Computer Engineering
Politecnico di Torino
Torino, Italy
anna.arnaudo@polito.it

2nd Riccardo Coppola
Department of Control and Computer Engineering
Politecnico di Torino
Torino, Italy

3rd Maurizio Morisio
Department of Control and Computer Engineering
Politecnico di Torino
Torino, Italy

4th Antonio Vetrò
Department of Control and Computer Engineering
Politecnico di Torino
Torino, Italy

5th Maurizio Borghi
Departement of Law
Università di Torino
Torino, Italy

6th Bryan Khan
Departement of Law
Università di Torino
Torino, Italy

7nd Riccardo Raso
Department of Control and Computer Engineering
Università di Torino
Torino, Italy

***Abstract*—This survey provides a synthesis of the practical copyright compliance challenges inherent in the input stages of Generative Artificial Intelligence (GenAI) systems. Specifically, we sought to address three research questions, examining the types of copyright-protected data involved, the corresponding challenges for copyright compliance in input data processing practices, and potential mitigation strategies proposed by researchers. We conducted a Systematic Literature Review (SLR), aimed at establishing a methodical foundation for the research.**

A recurring theme is the opacity of training data usage. This review highlights the frequent misalignement of content licences, together with the absence of mechanisms to govern bot activity, and the risk of copyright infringement through either model fine-tuning aimed at stylistic emulation or inadvertent memorisation of protected training data.

To counter these risks, various mitigation strategies have emerged, including watermarking, adversarial perturbation, training data attribution with eventual distribution of royalties, synthetic datasets, and Text and Data Mining (TDM) opt-out mechanisms in machine-readable formats.

***Index Terms*—Generative AI, Training Datasets, Copyright, Intellectual Property**

I. Introduction

Generative Artificial Intelligence (GenAI) has gained increasing attention in recent years due to its proven versatility and efficacy in supporting humans in performing a wide variety of tasks. This exceptional success is in part due to the scale of training data ingested: both data volume and variety have been shown to be crucial for a successful learning process [1]. However, this voracious need for data has led to various legal and regulatory issues regarding the underlying rights in ingested content.

For example, in December 2023, The New York Times sued OpenAI and Microsoft for unauthorised reproduction of its publications during training of a GPT model [1]. In the same year, in a dispute before the French Competition Authority, French press publishers alleged that they were very often unable to verify whether their press publications had been used by Google's AI Gemini for training purposes. Moreover, they complained of not having access to efficient technical tools to oppose the use of their materials by AI [2].

In the European Union (EU), the regulation of AI at the regional level is guided by the AI Act (2024). It incorporates a recognition of EU copyright principles, including Article 4 of the EU Copyright in the Digital Single Market Directive (Dir. 790/1019; 2019). CDSM Article 4 introduced a general copyright exception for a general Text and Data Mining (TDM), which allows AI dataset developers to use copyright-protected works for training, but also grants copyright holders a right to opt-out of such use. The EU legislator, in Article 4(1), establishes that the reservation of rights (opt-out) must be made "*in an appropriate manner, such as machine-readable means in case of content made publicly available online*". However, there is active debate on how to interpret the "appropriate manner" and "machine-readable means" requirements in both a legal and a technical sense.

Given the lack of a single proven and broadly accepted solution, various actors have proposed guidelines, standards or protocols. It should be noted, however, that increasing the cost or difficulty of accessing data may diminish AI technological progress [2]. Thus, both legal and technical mechanisms should be appropriately designed to balance copyright compliance and technological innovation.

Moreover, due to the probabilistic nature of GenAI models, it is not trivial to trace the training samples from which a specific instance of generative output has been derived.

Finally, it is important not to underestimate the possibility for a GenAI model to produce copyright-infringing outputs independently from the data on which it has been trained. Indeed, copyright-protected content may be incorporated into input prompts and subsequently exploited by end-users, for instance, by replicating an artist's style through inputting

[1] https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

representative works and instructing the model to emulate them. Although copyright law does not extend protection to an author's style per se, direct imitation may nevertheless trigger various legal and regulatory issues.

A. Purpose and Scope

This survey aims to explore the vast and fragmented landscape of established and developing techniques for mitigating copyright infringement risks along the GenAI input pipeline. We prioritised breadth over depth, offering a general overview serving as a foundational guide for individuals with prior exposure or a developing understanding of the field.

On the other hand, the potential copyright concerns associated with the generation procedures—namely, the 'output' pipeline of GenAI systems—are not addressed in this study and are left for future research.

Finally, our focus is solely on copyright infringement and not on other legal or regulatory compliance issues—such as the need to reliably label generative outputs and track their provenance in order to enhance transparency, or the question of the 'copyright-eligibility' of AI-generated works. For simplicity and clarity, we confine our research to the European legislative system and all technologies designed—or potentially capable of—integrating with it.

II. Methodology

Our survey is grounded in a Systematic Literature Review (SLR), aimed at establishing a methodical foundation for the research. All the details about the process—including the Search string, the search hits and the quality assessment of the identified sources—can be found in the replication package, which has been published on Zenodo and referenced in Appendix A.

A. Related Secondary Studies

To the best of our knowledge, at the time of writing a single secondary study has been conducted on the topic [3]. The main information about it is reported in *Table I*. It is, however, targeted for Indian regulation and is based only on a single source of primary studies. Moreover, the review focused on the legal perspective, while this SLR is intended to provide an overview of technical aspects, as better specified by the Research Questions reported in *Section II-B*.

B. Goals and Review Questions

The review questions are intended to investigate (1) the input data, (2) the copyright compliance challenges in training processes and data collection methods, and (3) the potential copyright compliance solutions proposed. They have been formulated as follows:

- **RQ1** Which are the characteristics—e.g., the format, degree of curation, and provenance—of Copyright-protected Input Data Involved in Each Phase of the GenAI Pipeline?
- **RQ2** What are the potential copyright compliance issues—with respect to EU legislation—related to the different methods and data input phases?

TABLE I
The related secondary study.

Year	2024
Reference	Intersection of generative artificial intelligence and copyright: an Indian perspective [3]
Research Method	SLR + qualitative research
N. of Primary Studies	33
Description	Examines the adequacy of Indian copyright law in addressing AI-generated creations.
Limitations	Sources only from one database (Scopus); focused on Indian law and not on technologies for GenAI's input/output control

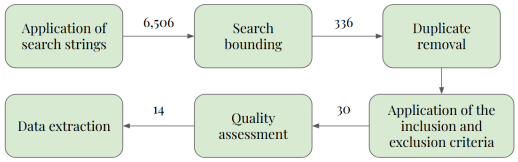

Fig. 1. Phases of the SLR process. The arrows are labeled with the number of sources selected after the corresponding phase.

- **RQ3** What are the mitigation strategies that can be mapped to the issues identified by RQ2?

C. Search Approach

We followed the steps inspired by Garousi [4]:

- **Application of search strings:** Search strings covering the key terms on *GenAI* and *Copyright* were iteratively refined and applied to major digital libraries (ACM, IEEE, Springer, ScienceDirect) and Google Scholar. A publication-year filter (2019–early 2025) ensured alignment with the post-CDSM Directive landscape.[2]
- **Search bounding:** Following Garousi, we restricted the scope (e.g., first 100 Google results), yielding 336 initial sources.
- **Inclusion and exclusion criteria:** Titles, keywords, and abstracts were screened according to IC/EC. Works outside EU copyright law or not accessible linguistically were excluded. After removing duplicates, 30 sources remained.
- **Quality assessment:** Using Kitchenham and Charters' guidelines [5], two authors independently evaluated all papers through tailored questionnaires.[3] Sources scoring above the average threshold (3.41/5) were retained, resulting in 14 high-quality studies (listed in Appendix A).
- **Data extraction:** Relevant information from the final set of studies was extracted in structured form.[3]

Figure 1 summarises the workflow and the number of sources selected at each step; further details are available in the online appendix.[3]

[2]The focus is limited to EU legislation.
[3]See *Appendix A*.

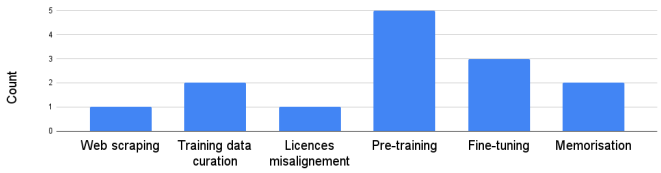

Fig. 2. Distribution of the issues related to the phases of the GenAI input pipeline explicitly mentioned by the selected sources.

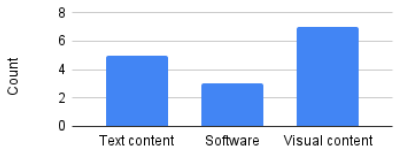

Fig. 3. Distribution of the data formats explicitly mentioned by the selected sources.

III. Results

Figures 2 and *3* report some statistics about the content extracted from the final pool of papers.

A. ***RQ1:*** *Which are the characteristics—e.g., the format, degree of curation, and provenance—of Copyright-protected Input Data Involved in Each Phase of the GenAI Pipeline?*

Summary of the Answer to RQ1 - Types of Input Data

It emerged that there is a marked lack of transparency regarding the specific training data employed in machine learning processes. However, Common Crawl was identified as one of the most common sources. Notably, the press and publishing sector appears to be particularly affected by data harvesting practices, mainly accomplished through web scraping activities. The latter are increasingly generating concern among website managers and copyright owners. This widespread use of web scraping practices can be traced back to the substantial demand for data required to train large foundation models underpinning all GenAI systems. Conversely, while fine-tuning these pre-trained models necessitates a comparatively smaller volume of data, it demands a markedly higher degree of curation.

The lack of access to the details on how commercial models are built—as reported in the European Open Source AI Index[4]—hinders transparency regarding the actual training samples used [6]. Moreover, regulations might be advocated to increase transparency or allow companies to keep model details secret to protect intellectual property [7]. There are instances of models labelled as open-source due to the public availability of their internal parameters; however, this does not necessarily imply that the training datasets have also been disclosed [6], [8]: a notable example is Meta's Llama.

Furthermore, the complexity of dealing with the huge amount of training data ingested certainly reduces transparency. Overall, it is generally valid to say that there are two primary sources of training materials:

- Dataset created by the AI company itself, often collecting data via web scraping;
- Datasets curated by third parties: they can be both licensed or open-source (e.g., the ones hosted on Hugging Face, a large platform allowing the GenAI community to share a variety of tools[5]).

1) Web Scraping: Web scraping differs from crawling in the fact that it is not limited only to indexing pages, but also involves data extraction activities aimed at retrieving relevant information hosted inside the HTML pages.[6] Some of the most known web scrapers employed by AI companies are OpenAI's GPTBot and Microsoft's Bing bot.

Scraping activities particularly involve the vast amount of news articles publicly available online. Indeed, the press sector is specially important for GenAI training since it possesses valuable assets: a large amount of text data and ethical principles for aligning the systems to humans' needs [2].

Another target of scraping, as highlighted by the consulted sources, is open-source code, as exemplified by the 'Doe v. GitHub' case discussed in Section III-B [9].

2) Pre-Training: Foundation models, the large, general-purpose models at the base of each GenAI system, are constructed during a phase called pre-training. This step involves a very large corpus of data in the range of tens or hundreds of terabytes.

A widely mentioned source of textual data for training the LLMs is the Common Crawl dataset, which is gained by scraping subsets of the World Wide Web multiple times a year over the past decade, yielding a vast corpus. Pre-training often uses multiple datasets, Common Crawl being only one of them, albeit typically the largest [10]. However, as declared by the founders of the initiative, the primary objective of Common Crawl was not to feed generative AI models [11]. Consequently, the data gathered by CCBot—the web crawler employed by the non-profit organisation—has not been curated to ensure compliance with copyright requirements for this specific use. As such, AI companies sourcing data from Common Crawl without proper data cleaning and filtering may encounter copyright compliance issues.

With regard to image datasets, some examples are the ImageNet [12] and the LAION-5B [13] datasets. They do not store the images themselves; rather, they reference the source websites, from which the images may be downloaded, provided that such use is permitted under the terms and conditions governing those sites. This is a common pattern between image datasets.

3) Fine-Tuning: The phase following pre-training—namely fine-tuning—requires task-specific datasets to tailor the foundational model to particular objectives. This stage refines the model's capabilities by introducing targeted modifications to its parameters through an additional machine learning phase. The datasets employed are much smaller and curated [10].

These characteristics may lead to a reduced risk of copyright infringement. Indeed, it becomes more plausible that

[4]https://osai-index.eu/the-index

[5]https://huggingface.co/

[6]https://www.geeksforgeeks.org/difference-between-web-scraping-and-web-crawling/

copyright-protected material is identified and excluded during the more meticulous data preprocessing conducted on fine-tuning datasets.

B. ***RQ2:*** *What are the Potential copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?*

Summary of the Answer to RQ2 - Copyright compliance issues

With regards to web scraping, the identified problems primarily concern the lack of oversight over bot activities—an issue particularly pronounced for copyright owners who lack the technical expertise or direct control over the websites hosting their content.
Moreover, we reported that data collected for training is often deprived of accurate licensing information, leading to possible infringement of the related usage conditions.
Additional identified challenges include the potential of injecting tailored data into models via fine-tuning, which can be employed to emulate an artist's distinctive style.
Finally, another concern is the phenomenon of memorisation, where generative AI models reproduce their training data, thereby unintentionally disseminating copyright-protected material.

1) Web Scraping: Organisations using crawled AI training datasets that are publicly available online typically do not pay compensation for these data sources, but assume that this is legitimate use under some legal doctrine in the jurisdiction in which the crawl takes place or the data source is located [10].

In the context of the dispute with Google before the French Competition Authority mentioned in *Section I*, press publishers in France expressed doubt as to whether the technical and legal mechanisms put in place to reserve their rights were taken into account by AI crawlers. Furthermore, they complained about the lack of a technical solution for differentiating between the instructions for indexing a web page in Google Search and for using the scraped page in Google's GenAI system. This may not be fair, since publishers' reservations should not impede the use of affected publications within other services for which they can be remunerated. [2].

2) Misalignement of Licences: Even if data collection has been carried out in compliance with existing legal frameworks, this does not necessarily imply that these data can be legitimately employed for training generative AI systems. Indeed, the original licensing terms may explicitly restrict such uses, while allowing those for which the crawl was initially conducted. For example, certain open-source licences may be incompatible with the development of commercial generative AI applications.

Moreover, data in training datasets may not be accompanied by the correct associated licence, which may have been lost or wrongly replaced during the collection and curation processes. In November 2023, the Data Provenance Initiative reported that a large portion of training material was linked to a different licence—often more permissive—than the one originally specified by the creator [14].

In 2022[7], GitHub, Microsoft and OpenAI were sued by a group of programmers alleging that GitHub's Copilot and OpenAI's Codex used publicly available open source code posted on GitHub's platform as training data for their Generative AI systems. According to Samuelson [9], the most significant claim is that the companies wrongfully removed copyright-relevant information from open-source programs ingested as training data. The second is that GitHub and OpenAI have breached open-source licence agreements by failing to assign attribution to the respective open-source developers.

3) Fine-Tuning: Recently, fine-tuning methods compatible with generative models belonging to the family of Stable Diffusion enable users to inject personalised concepts into the base model with minimal data and computational resources. These concepts can include specific individuals, objects, and unique styles. Some examples of these methods are Textual Inversion [15], DreamBooth [16], Custom Diffusion [17] and LoRA [18]. Since Stable Diffusion is gaining widespread popularity, concerns have emerged regarding image privacy and copyright infringement. Indeed, fine-tuning on the works of specific artists enables Stable Diffusion to easily replicate their styles.

4) Training Data Memorisation: A training sample is considered to have been memorised by a GenAI system when the latter produces an output that closely resembles, or in some cases replicates, the original data instance. This is an undesirable phenomenon[8], which however, has been observed across various model architectures. For instance, Carlini et al. systematically examined memorisation in both Diffusion Models and Language Models [19], [20]. To discover memorised training samples in LLMs, they repeatedly prompted the models with the first tokens of a string known to be in the training dataset, checking whether the model produced the exact continuation. An analogous methodology was applied to diffusion models, where the researchers prompted the models using the original image captions from the training dataset, observing whether the generated outputs closely replicated the corresponding training images.

Memorisation is a phenomenon related to both GenAI's input and output, even if some authors [21] argue that this issue is caused by how the model is trained and not by how it is prompted. Recent studies have demonstrated that duplication within training datasets is the main factor that can result in both Diffusion Models and LLMs to memorise the samples. Moreover, certain attack methodologies involve crafting prompts specifically designed to extract memorised training data. Inversion attacks exploit this vulnerability by attempting to reconstruct representative examples for specific classes or prompts [19], [22]–[24].

In the context of the earlier mentioned lawsuit against OpenAI and GitHub, the code generated by Copilot and Codex was flagged as potentially infringing [9]. This is due to its supposed substantial similarity to the training data, i.e. the open-source code hosted on GitHub, suggesting that the models may have memorised and reproduced portions of it.

[7]https://www.courtlistener.com/docket/65669506/doe-1-v-github-inc/

[8]Since GenAI models are expected to be able to produce new, unseen data.

TABLE II
MAPPING BETWEEN THE ISSUES IDENTIFIED IN SECTION III-B AND POSSIBLE MITIGATIONS STRATEGIES IDENTIFIED IN SECTION III-C.

Issue	Mitigations
Web scraping	Text poisoning, solutions for Expressing Text and Data Mining (TDM) Reservation (opt-out);
Misalignement of licences	Text watermarking, visual attribution for generated images, membership inference, olutions for Expressing Text and Data Mining (TDM) Reservation (opt-out);
Fine-tuning for style mimicry	Text poisoning, protective images' perturbations, visual attribution for generated images;
Training data memorisation	Text poisoning, protective images' perturbations, synthetic data, collecting royalties for copyrighted training data, visual attribution for generated images;

C. *RQ3: What Are the Mitigation Strategies That Can Be Mapped to the Issues Identified by RQ2?*

Summary of the Answer to RQ3 - Mitigations

A variety of approaches have been developed to safeguard data from unauthorised use in GenAI training.
Among these, techniques for embedding watermarks into protected contents—thereby enabling the subsequent identification of their presence in a model's training corpus—have gained traction.
Similarly, encoding text using non-standard fonts, which are typically not recognised by AI systems, represents another form of technical deterrence.
In the visual domain, images may be altered by means of adversarial perturbations designed to impede their effective employment in machine learning processes.
Moreover, the use of synthetic data as a substitute for scraped real-world content has also attracted considerable interest.
Finally, more diplomatic strategies—centered on collaboration between copyright owners and AI developers—have emerged. These include methods for tracing AI-Generated Content (AIGC) back to its most influential training samples, thereby facilitating attribution and potential compensation for the respective rightsholders. Additionally, mechanisms have been proposed to enable the latter to express their Text and Data Mining (TDM) opt-out preferences in a standardised, machine-readable format.

In the following, the mitigation strategies proposed in the analysed literature are discussed, while Table II schematises how these techniques can be mapped to the identified issues.

1) Text Watermarking and Poisoning: For copyright-protected texts, typical methods involve embedding watermark information in the text itself. In practice, these involve traditional synonym substitution and recently developed generative model-based techniques. However, watermarking methods, being a passive defense, can respond only after an attack has occurred [25]. Moreover, when applied to text, they are highly vulnerable to manipulations, which can be easily applied using LLMs. Significant research is being carried out to enhance the robustness of text watermarking algorithms. For instance, in 2024 Google open-sourced the reference implementation of SynthID-Text[9], a text watermarking tool developed by the company, with the aim of fostering evolution of the technology.

A new frontier for protecting open-source code against unauthorised use in model training is code dataset watermarking. Some prototypes like CoProtector [26] and CodeMark [27]—which are both available open-source—are designed to trace code usage in training neural code completion models, as outlined by the schema in *Figure 4*. However, as noted by Wu et al. [28], these methods still suffer from vulnerabilities related to code manipulations.

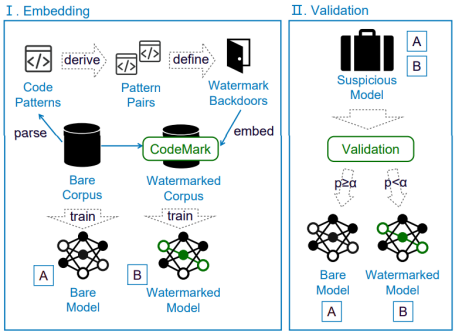

Fig. 4. Schema outlining the working principle of CodeMark, allowing to detect if a code has been used for training a model [27]

A different approach involves using special Unicode characters, making it impossible for AI technologies to process it [25]. An example is the technology proposed by DataDust.ai[10], which protects text content from AI-powered scrapers by using a text font that AI cannot interpret. This strategy proves effective because typically these AI-powered scrapers have not been trained on such rare or non-standard fonts. However, as Zhao et al. pointed out [25], they could be easily bypassed by Optical Character Recognition (OCR).

2) Visual Content - Protective Perturbations: Recent investigations have examined the integration of imperceptible perturbations into images as a means of preventing unauthorised exploitation.

Notably, Glaze [29] is designed to prevent the appropriation of artists' work for stylistic mimicry in Diffusion Models by optimising the distance at the feature level between the original and the protected versions of the images, thereby inducing the model to learn an incorrect artistic style. In their study, they gathered feedback from a cohort of artists, revealing that 93% of respondents consider Glaze to be effective in safeguarding images against generative style mimicry.

Similarly, Anti-DreamBooth [30] explicitly incorporates the DreamBooth[11] fine-tuning process into its framework by formulating a bilevel min-max optimisation procedure to generate protective perturbations.

Nightshade [31] aims to poison images so that, while still looking the same if examined by humans, a model trained on them completely misinterprets their content (e.g., wrongly detects a cat where an image contains a dog). This can lead, during the model functioning, to produce a number of hallucinations proportional to the number of poisoned training samples involved, leaving AI developers with the sole option

[9]Available on GitHub at: https://github.com/google-deepmind/synthid-text

[10]https://www.datadust.ai/

[11]Described in Section III-B in the paragraph dedicated to fine-tuning.

of re-training the model through the exclusion of the poisoned images.

Further research efforts [32], [33] have also explored the generation of protective noise using analogous adversarial techniques.

Zhao et al. [34] pointed out that, while such methods have demonstrated efficacy in controlled environments, their practical applicability remains limited. In real-world contexts, images are frequently subject to various natural transformations—such as cropping, compression, and blurring—during online dissemination. Consequently, any effective image protection technique must account for the robustness of perturbations under such conditions.

Unfortunately, empirical evaluations indicate that even moderate transformations are sufficient to undermine the protective effectiveness of these methods.[12] This suggests that perturbation-based strategies may be inadequate to ensure reliable protection in practice. Moreover, their effectiveness varies considerably depending on the specific fine-tuning approach employed. Because these perturbations typically target the text encoder, they offer limited efficacy in cases where the learning process does not involve modifications to this component. This presents a significant limitation, as image protectors are generally unable to ascertain the fine-tuning methodologies adopted by potential exploiters [34].

In conclusion, while natural transformations may degrade image quality and resolution, they may nonetheless serve as preprocessing tools for circumventing protective measures with acceptable costs. Another critical limitation concerns the proportion of images that are protected: in real-world deployment, it is rarely feasible to apply protection retroactively to content that has already been disseminated publicly [34].

To further demonstrate the limited robustness of protective perturbations, Zhao et al. [34] proposed a novel purification method called GrIDPure. This technique is designed to eliminate protective perturbations while preserving the structural integrity of the original image to the greatest extent possible. The effectiveness of GrIDPure has been evaluated against several prominent protective approaches, including Glaze (2023 version), AdvDM, and Anti-DreamBooth. The evaluation involves assessing images generated by Stable Diffusion models fine-tuned on datasets purified by GrIDPure. The findings demonstrate that the purified images are effectively learned by the model, enabling the generation of high-quality outputs.

3) Synthetic Data: Synthetic data are a specific corner case of a model's generative output. Indeed, they are specifically designed to reproduce the statistical characteristics of real-data. A generator model is prompted many times to create a dataset of the desired size [10].

Replacing real-world data with synthetic alternatives offers several potential advantages, such as reducing the cost of dataset construction and enabling augmentation in domains where authentic data is scarce. Moreover, relying on synthetic data may lower the risk of copyright infringement. Nevertheless, the copyright safety of the synthetic data itself hinges on the absence of contamination with copyright-protected data in the generator model used.

Overall, while synthetic data offers substantial benefits in terms of scalability, bias reduction, and regulatory compliance, real-world data may remain indispensable for applications that demand authenticity, nuanced complexity, and fidelity to real-world conditions.

4) Collecting Royalties for Copyrighted Training Data: Alongside licensing agreements, one proposed solution for the legal use of copyright-protected works in training is to pay royalties to copyright owners based on the actual use of their content. However, the large volume of samples required for training, combined with the fact that the value of a given sample may vary across different GenAI systems, makes the attribution of royalties a complex task.

A simple approach uses similarity scores between training data and generated content as a valuation metric. For example, Zhang [35] proposes a scoring service for data use based on text similarity. Wang et al. [36] propose WASA, an approach to attribute the sources of generative outputs using watermarks. Sakurai et al. [37] propose a fine-tuned version of BERT tailored for the author attribution task. To address the substantial computational costs noted in prior studies, they integrate the model with distillation techniques aimed at reducing overhead—achieving a sixfold speed-up—and improving the interpretability of the model's outputs.

In contrary, Wang et al. [38] proposes to mitigate the black-box nature of content generation by leveraging the probabilistic nature of generative models: the log-likelihood of generating the user-chosen content is used to measure the utility of the training data. Royalties are subsequently distributed among copyright holders based on their respective contributions, which are analytically determined through the application of Shapley value theory [39].

Although the approach proposed by Wang et al. has proven effective in correctly attributing relevance to data sources associated with generated outputs, it requires retraining the model multiple times to assess the utility of each source. The substantial computational costs involved make it practically feasible only when dealing with a limited number of copyright owners.

5) Visual Attribution for Generated Images: The EKILA framework [23] has been designed to recognise and reward creatives for their contributions to GenAI training. At the core of this theoretical framework, which has not yet been widely implemented, lies a visual attribution technique aimed at identifying the training samples from which a generative output may have originated. Moreover, EKILA leverages Distributed Ledger Technology (DLT), or blockchain, alongside Non-Fungible Tokens (NFTs), to store rights-related data in a decentralized and tamper-proof manner. These are ultimately integrated with the Coalition for Content Provenance and Authenticity (C2PA) standard, enabling the tracing of the provenance of images produced by GenAI systems.

[12]It is worth mentioning that the researchers tested the robustness of the version of Glaze released in 2023 and not the most up-to-date, which has been published in 2025, one year after the study conducted by Zhao et al.

To support attribution, EKILA implements a two-stage visual matching system: (1) patch-level fingerprint extraction and matching, followed by (2) pairwise verification and scoring of the most similar matches. Ultimately, attribution is determined by selecting a given number of training samples that are most similar to the generated content. Visual fingerprints are generated using embeddings produced by a Convolutional Neural Network (CNN) optimised to compute similarity score even with manipulated or degraded images. The experimental results reported by the researchers show that this method outperforms existing similarity metrics such as CLIP [40], LPIPS [41], and SIFID [42] in attribution tasks across millions of image patches.

EKILA extends the concept of Non-Fungible Tokens (NFTs) by proposing a tokenised representation of rights, thereby establishing a triangular relationship among Ownership, Rights, and Attribution (ORA). Unlike traditional NFTs, which often do not confer concrete rights beyond asset ownership, EKILA introduces smart contracts to manage and enforce royalty payments when tokenised rights are exercised.

Inside the C2PA manifest, information about the creators' identity and the relative digital wallet can be included, enabling automated crypto-currency-based royalty payments whenever their content is implicated in AI image generation.

However, this approach faces notable challenges. The high computational cost of visual matching and the complexity of managing attribution across extensive datasets may limit scalability. Furthermore, the correlation-based attribution method, while intuitively plausible, lacks the causal rigor offered by methods such as leave-one-out re-training—approaches which, however, are much more expensive at GenAI-scale. Moreover, EKILA's methodology, being image-specific, is currently not applicable to other data modalities such as text or audio.

6) Membership Inference: Membership Inference Attacks (MIAs) enable the determination of whether a specific data sample was included in the training process of a model. Consequently, they may serve as a means to assess whether the reservation preferences of copyright holders have been upheld.

MIAs often exploit differences in model behaviour when dealing with training data versus unseen data, as models typically perform better when fed with prompts containing training data [43]. However, these attacks exploit vulnerabilities—which are often very peculiar and vary from model to model—that are likely to be progressively mitigated as GenAI technologies advance and become more robust.

Although membership inference attacks are typically complex and expensive, Morone et al. [44] proposed using data portraits, i.e. artifacts that record training data and allow for downstream inspection of training datasets. Their solution is designed to facilitate lightweight membership inference within the training dataset. However, this solution requires direct access to the corpus: as noted in *Section III-A*, GenAI companies often refrain from disclosing the composition of their training material.

7) Solutions for Expressing Text and Data Mining (TDM) Reservation (opt-out): In response to the EU Digital Single Market (CDSM) Directive, previously mentioned in *Section I*, several initiatives were introduced to establish a widely standard way to express Text and Data Mining (TDM) reservation—also referred to as "opt-out".

Indeed, given the significant fragmentation and complexity characterising internet platforms, the absence of a common protocol might impair the rights holders' ability to articulate their TDM reservation preferences in a clear and explicit manner. On the other hand, AI developers and dataset curators would encounter considerable difficulty in automating large-scale data collection while scrupulously adhering to TDM reservations, unless such preferences are conveyed through a standardised and machine-readable format.

The French National Publishing Union (SNE) proposes that a standard clause be included in the general terms and conditions of publishers' websites. In addition, SNE recommends the use of the W3C's TDM Reservation Protocol[13], which has been proposed by the European Digital Reading Lab and enables opting-out through the insertion of a flag (i.e., a boolean value) inside the content's metadata. According to SNE, the use of this metadata, designed to fall into the category of machine-readable means, can be an effective technical response to data harvesting tools [2].

In 2023, Google introduced its own opt-out protocol and launched Google–Extended as an extension to the traditional robots.txt protocol. This allows rightholders to prevent their works from being used to train AI models. Moreover, Google indicated that exercising this opt-out does not affect the indexing on Google Search [2], [45], thus theoretically addressing the issues mentioned in *Section III-B*. However, despite reserving their rights through Google–Extended, some press publishers noticed that their publications still appeared in the responses offered by Gemini [2], [46].

Microsoft and OpenAI also introduced their own standard policies to allow copyright holders to opt-out of having their works used to train the respective AI models.

From the perspective of the copyright owners, this means that they should opt-out for each AI model by following different protocols. This involves a certain amount of complexity, efforts and costs. It also creates uncertainty about whether the reservation is exhaustive—i.e., covering all AI models that can potentially use data for training purposes. Moreover, copyright owners cannot be certain that their rights have been effectively reserved, since there is no mechanism for checking into private AI training datasets. [2], [46]

8) C2PA Training and Data Mining Assertions: The Coalition for Content Provenance and Authenticity (C2PA) standard has attracted increasing attention in recent years. It allows to safely bind provenance information—such as how an image was captured and subsequently manipulated—to digital assets in order to support users in making informed trust decisions. Many software companies—such as Google [47] and Microsoft[14]—and camera producers—like Sony [48]—

[13]https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/

[14]https://www.microsoft.com/en-us/research/project/project-origin/

TABLE III
QUALITY CONCEPTS FOR ASSESSING STUDY'S VALIDITY [5].

Name	Definition
Bias	A tendency to produce results that depart systematically from the 'true' results.
Internal Validity	The extent to which the design and conduct of the study are likely to prevent systematic error.
External Validity	The extent to which the effects observed in the study are applicable outside of the study.

have already adopted this standard.

Although originally intended for image media content, the C2PA standard is designed to be format-agnostic, meaning that any binary asset is potentially capable of bearing a C2PA manifest.

A C2PA manifest is an organised data structure embedded within content's metadata. Manifests contain a customisable list of assertions that describe the history of the asset, including the identity of its creator, the methods employed, and also the constituent ingredient assets involved, i.e., other digital items from which the content has been derived. In turn, each ingredient may be associated with its own manifest. As proposed by Balan et al. in the EKILA framework [23], the list of ingredients can be leveraged to attribute the primary training samples from which an AI Generated Content (AIGC) has been derived (more details can be found in the paragraph on 'Visual Attribution for Generated Images').

The integrity and authenticity of C2PA manifests are ensured through digital signatures. However, although the protocol specifies that stripped manifests may, in principle, be recovered via lookup in a distributed database, the necessary infrastructure to support this mechanism has not yet been realised. As a result, assets remain vulnerable to provenance data loss if metadata is removed.

C2PA has also been proposed as a common standard solution for expressing TDM reservations. Indeed, in addition to the provenance tracking functionality, the protocol also allows embedding Training and Data Mining Assertions. In particular, for any given TDM action between AI training, GenAI training,[15] data mining and AI inference, the syntax allows specifying whether it is allowed, denied or constrained (i.e., allowed only further specified conditions hold). This differentiation provides a certain level of flexibility, which is a desirable feature because of the high level of fragmentation existing between all the sectors affected by the collection of data for AI systems.

IV. THREATS TO THE VALIDITY OF THIS STUDY

We refer to the quality concepts defined by [5] shown in *Table III*.

A. Biases

a) Fragmentation of the Topic: The main challenge in a SLR is ensuring full coverage of relevant studies.

[15]AI training and GenAI training are explicitly distinguished because the first includes categories of systems which do not generate new data, thus having a different probability of infringing copyright.

This review faces limitations due to the fragmented nature of the topic, which spans diverse issues and perspectives. For example, some recent initiatives on rights information management—such as JPEG Trust[16] , Spawning.ai[17] , and Liccium's Trust Engine[18]—fall outside its scope.

The intersection of GenAI and copyright raises multiple issues, each warranting a dedicated review and drawing on strategies from different technological domains. This fragmentation made it harder to design an efficient and comprehensive search string.

B. Threats to Internal Validity

Possible threats to internal validity were mitigated by adhering to established guidelines [4], [5] and by engaging multiple authors in each phase of the review, including reading the articles, thereby ensuring that diverse perspectives were considered.

C. Threats to External Validity

Threats to External Validity arise from potential bias—discussed in Section IV-A—due to incomplete coverage of relevant sources. While some technologies reviewed show adaptability across multiple data formats, this study does not provide balanced coverage of all content types. As highlighted in Figure 3, audio and video remain underexplored. Since sectors handling different data formats often employ distinct copyright management mechanisms, further research is needed to extend and complement the current work.

Another limitation concerns the scope of this study, which focuses exclusively on EU regulation and may therefore lack global generalisability.

The search strategy yielded a limited corpus of 14 systematically analysed sources—fewer than half of those in the comparable secondary study in Section II-A. This narrower result reflects the practice- and computer science-oriented scope of the review, in contrast to the legal focus of the comparative study [3]. The use of legally charged keywords such as "copyright" likely biased ranking algorithms toward law-centric sources, limiting visibility of computer science literature, while bounding strategies in Google Search and Springer Nature further restricted access to technical contributions.

Despite this, the review provides a valuable contribution as the first secondary study to examine computer science literature on copyright issues in Generative AI and the mitigation strategies proposed in technical research.

V. CONCLUSION

This SLR provided insights into practical issues related to copyright infringement and technical measures related to the GenAI input pipeline.

[16]https://jpeg.org/jpegtrust/
[17]https://spawning.ai/
[18]https://liccium.com

In particular, it underscores significant concerns regarding the opacity of training datasets, where Common Crawl frequently features as a pivotal—yet controversial—data source. This lack of transparency has sparked legal scrutiny, especially in the press and publishing sectors, where web scraping practices have led to increased copyright risks. Moreover, we found that the original licensing terms of web-scraped content are frequently disregarded, raising both legal and ethical concerns. The phenomenon of training data memorisation and the deliberate fine-tuning to imitate identifiable artistic styles further exacerbate these legal and ethical concerns.

In response, technical and strategic countermeasures have emerged. These include watermarking techniques, non-standard font encoding, adversarial image perturbations, and the generation of synthetic datasets as alternatives to real-world data. Diplomatic strategies encompass traceability mechanisms that link AI Generated Content (AIGC) to its training origins, thereby facilitating author attribution and compensation. Additionally, they include machine-readable tools that enable copyright holders to express their opt-out preferences concerning Text and Data Mining (TDM). However, implementing diplomatic protections may prove ineffective if there is no means to inspect training datasets. Consequently, this study examines membership inference attacks, which are closely associated with the phenomenon of memorisation and are then controversial.

The review acknowledged its limitations, noting that specific protections for audio or video content were not covered due to the considerable heterogeneity of the subject. By prioritising breadth over depth, the study leaves room for future work to explore additional copyright issues and mitigations for particular content types.

REFERENCES

[1] B. et al., "On the opportunities and risks of foundation models," 2022. [Online]. Available: http://arxiv.org/abs/2108.07258

[2] M. Kowala, "Protection of press publishers in the age of generative AI – in search of legal remedies to adapt to the pace of technology," vol. 55, no. 10, pp. 1604–1623, 2024. [Online]. Available: https://doi.org/10.1007/s40319-024-01515-y

[3] S. Vig, "Intersection of generative artificial intelligence and copyright: an indian perspective," vol. ahead-of-print, 2024, publisher: Emerald Publishing Limited. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/jstpm-08-2023-0145/full/html

[4] V. Garousi, M. Felderer, and M. V. Mäntylä, "Guidelines for including grey literature and conducting multivocal literature reviews in software engineering," *Information and Software Technology*, vol. 106, pp. 101–121, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584918301939

[5] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007. [Online]. Available: https://docs.edtechhub.org/lib/EDAG684W

[6] D. G. Widder, M. Whittaker, and S. M. West, "Why 'open' AI systems are actually closed, and why this matters," vol. 635, no. 8040, pp. 827–833, 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-024-08141-1

[7] J. Schneider, "Explainable generative AI (GenXAI): a survey, conceptualization, and research agenda," vol. 57, no. 11, pp. 1–38, 2024, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 11 Publisher: Springer Netherlands. [Online]. Available: https://link.springer.com/article/10.1007/s10462-024-10916-x

[8] A. Liesenfeld and M. Dingemanse, "Rethinking open source generative AI: open-washing and the EU AI act," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24. Association for Computing Machinery, 2024, pp. 1774–1787. [Online]. Available: https://dl.acm.org/doi/10.1145/3630106.3659005

[9] P. Samuelson, "Legal challenges to generative AI, part i," vol. 66, no. 7, pp. 20–23, 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3597151

[10] H. Ludwig, Y. Zhou, S. Zawad, Y. Ong, P. Li, E. Butler, and E. Zahid, "Towards collecting royalties for copyrighted data for generative models," in *2024 IEEE International Conference on Web Services (ICWS)*, 2024, pp. 20–26, ISSN: 2836-3868. [Online]. Available: https://ieeexplore.ieee.org/document/10707489

[11] Knibbs. (2024) Publishers target common crawl in fight over AI training data | WIRED. Accessed: 2025-04-20. [Online]. Available: https://www.wired.com/story/the-fight-against-ai-comes-to-a-foundational-data-set/?utm_source=chatgpt.com

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009.

[13] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5b: An open large-scale dataset for training next generation image-text models," 2022. [Online]. Available: http://arxiv.org/abs/2210.08402

[14] S. Longpre, R. Mahari, A. Chen, N. Obeng-Marnu, D. Sileo, W. Brannon, N. Muennighoff, N. Khazam, J. Kabbara, K. Perisetla, X. Wu, E. Shippole, K. Bollacker, T. Wu, L. Villa, S. Pentland, and S. Hooker, "The data provenance initiative: A large scale audit of dataset licensing & attribution in AI," 2023. [Online]. Available: http://arxiv.org/abs/2310.16787

[15] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022. [Online]. Available: http://arxiv.org/abs/2208.01618

[16] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2023. [Online]. Available: http://arxiv.org/abs/2208.12242

[17] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," 2023. [Online]. Available: http://arxiv.org/abs/2212.04488

[18] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021. [Online]. Available: http://arxiv.org/abs/2106.09685

[19] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace, "Extracting training data from diffusion models," 2023. [Online]. Available: http://arxiv.org/abs/2301.13188

[20] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying memorization across neural language models," 2023. [Online]. Available: http://arxiv.org/abs/2202.07646

[21] A. F. Cooper and J. Grimmelmann, "The files are in the computer: Copyright, memorization, and generative AI," 2025. [Online]. Available: http://arxiv.org/abs/2404.12590

[22] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein, "Diffusion art or digital forgery? investigating data replication in diffusion models," 2022. [Online]. Available: http://arxiv.org/abs/2212.03860

[23] K. Balan, S. Agarwal, S. Jenni, A. Parsons, A. Gilbert, and J. Collomosse, "EKILA: Synthetic media provenance and attribution for generative art," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 913–922. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023W/WMF/html / Balan_EKILA_Synthetic_Media_Provenance_and_Attribution_for_Generative_Art_CVPRW_2023_paper.html

[24] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," 2020. [Online]. Available: http://arxiv.org/abs/1911.07135

[25] J. Zhao, K. Chen, X. Yuan, Y. Qi, W. Zhang, and N. Yu, "Silent guardian: Protecting text from malicious exploitation by large language models," vol. 19, pp. 8600–8615, 2024, conference Name: IEEE Transactions on Information Forensics and Security. [Online]. Available: https://ieeexplore.ieee.org/document/10669119

[26] Z. Sun, X. Du, F. Song, M. Ni, and L. Li, "CoProtector: Protect open-source code against unauthorized training usage with data poisoning," 2022. [Online]. Available: http://arxiv.org/abs/2110.12925
[27] Z. Sun, X. Du, F. Song, and L. Li, "CodeMark: Imperceptible watermarking for code datasets against neural code completion models," in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2023, pp. 1561–1572. [Online]. Available: http://arxiv.org/abs/2308.14401
[28] B. Wu, K. Chen, Y. He, G. Chen, W. Zhang, and N. Yu, "CodeWMBench: An automated benchmark for code watermarking evaluation," in *Proceedings of the ACM Turing Award Celebration Conference - China 2024*, ser. ACM-TURC '24. Association for Computing Machinery, 2024, pp. 120–125. [Online]. Available: https://dl.acm.org/doi/10.1145/3674399.3674447
[29] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by text-to-image models," 2025. [Online]. Available: http://arxiv.org/abs/2302.04222
[30] T. V. Le, H. Phung, T. H. Nguyen, Q. Dao, N. Tran, and A. Tran, "Anti-DreamBooth: Protecting users from personalized text-to-image synthesis," 2023. [Online]. Available: http://arxiv.org/abs/2303.15433
[31] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models." IEEE Computer Society, 2024, pp. 807–825. [Online]. Available: https://www.computer.org/csdl/proceedings-article/sp/2024/313000a212/1WPcYDRkVX2
[32] X. Ye, H. Huang, J. An, and Y. Wang, "DUAW: Data-free universal adversarial watermark against stable diffusion customization," 2023. [Online]. Available: http://arxiv.org/abs/2308.09889
[33] Z. Zhao, J. Duan, X. Hu, K. Xu, C. Wang, R. Zhang, Z. Du, Q. Guo, and Y. Chen, "Unlearnable examples for diffusion models: Protect data from unauthorized exploitation," 2024. [Online]. Available: http://arxiv.org/abs/2306.01902
[34] Z. Zhao, J. Duan, K. Xu, C. Wang, R. Zhang, Z. Du, Q. Guo, and X. Hu, "Can protective perturbation safeguard personal data from being exploited by stable diffusion?" in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 24 398–24 407, ISSN: 2575-7075. [Online]. Available: https://ieeexplore.ieee.org/document/10656194
[35] D. Zhang, "Should ChatGPT and bard share revenue with their data providers? a new business model for the AI era," 2023. [Online]. Available: http://arxiv.org/abs/2305.02555
[36] X. Lu, J. Wang, Z. Zhao, Z. Dai, C.-S. Foo, S.-K. Ng, and B. K. H. Low, "WASA: WAtermark-based source attribution for large language model-generated data," 2023. [Online]. Available: https://openreview.net/forum?id=FDfq0RRkuz
[37] W. Sakurai, M. Asano, D. Imoto, M. Honma, and K. Kurosawa, "Efficient authorship attribution method using ensemble models built by knowledge distillation," in *2023 9th International Conference on Computer and Communications (ICCC)*, pp. 2357–2362, ISSN: 2837-7109. [Online]. Available: https://ieeexplore.ieee.org/document/10507584
[38] J. T. Wang, Z. Deng, H. Chiba-Okabe, B. Barak, and W. J. Su, "An economic solution to copyright challenges of generative AI," 2024. [Online]. Available: http://arxiv.org/abs/2404.13964
[39] Kuhn, "7. a value for n-person games. contributions to the theory of games II (1953) 307-317." in *Classics in Game Theory*, H. W. Kuhn, Ed. Princeton University Press, 1997, pp. 69–79. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/9781400829156-012/html
[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: http://arxiv.org/abs/2103.00020
[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018. [Online]. Available: http://arxiv.org/abs/1801.03924
[42] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," 2019. [Online]. Available: http://arxiv.org/abs/1905.01164
[43] Z. Li and Y. Zhang, "Advancing membership inference attacks: The present and the future," vol. 4, p. 2024017, 2025, publisher: EDP Sciences and CSPM. [Online]. Available: https://sands.edpsciences.org/articles/sands/abs/2025/01/sands20240020/sands20240020.html
[44] M. Marone and B. Van Durme, "Data portraits: recording foundation model training data," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Curran Associates Inc., 2023, pp. 15 121–15 135.
[45] Google, "An update on web publisher controls," https://blog.google/technology/ai/an-update-on-web-publisher-controls/, 2023, accessed: 2023-09-28.
[46] Autorité de la concurrence. (2024) Décision 24-d-03 du 15 mars 2024. https://www.autoritedelaconcurrence.fr/fr/decision/relative-au-respect-des-engagements-figurant-dans-la-decision-de-lautorite-de-la-0. Accessed: 2024-03-20.
[47] L. Richardson. (2024) How we're increasing transparency for gen AI content with the c2pa. [Online]. Available: https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/
[48] M. Goodman. (2024) Sony delivers highly anticipated firmware updates including c2pa compliancy and ensuring authenticity of images. [Online]. Available: https://www.sony.eu/presscentre/sony-delivers-highly-anticipated-firmware-updates-including-c2pa-compliancy-and-ensuring-authenticity-of-images

APPENDIX

A. Online Appendix

The details regarding the process underlying our Systematic Literature Review are available at: https://zenodo.org/records/15492021.