POLITECNICO DI TORINO
Repository ISTITUZIONALE

Generative AI and intellectual property: a systematic literature review of possible issues and mitigations

Original
Generative AI and intellectual property: a systematic literature review of possible issues and mitigations / Arnaudo, A., Coppola, R., Morisio, M., Vetro, A., Borghi, M., Raso, R., Khan, B.. - In: PEERJ. COMPUTER SCIENCE. - ISSN 2376-5992. - ELETTRONICO. - 12:(2026). [10.7717/peerj-cs.3928]

Availability:
This version is available at: 11583/3010554 since: 2026-06-30T09:26:11Z

Publisher:
PeerJ / Taylor & Francis

Published
DOI:10.7717/peerj-cs.3928

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

28 November 2026

PeerJ Computer Science

Generative AI and intellectual property: a systematic literature review of possible issues and mitigations

Anna Arnaudo[1], Riccardo Coppola[1], Maurizio Morisio[1], Antonio Vetrò[1], Maurizio Borghi[2], Riccardo Raso[1] and Bryan Khan[2]

[1] Department of Control and Computer Engineering, Polytechnic Institute of Turin, Turin, Italy
[2] Department of Law, University of Turin, Turin, Italy

ABSTRACT

The voracious appetite of Generative Artificial Intelligence (GenAI) for data has raised a number of issues regarding copyright. We conducted a systematic literature review (SLR), focusing on the types of copyright-protected data involved, the main challenges for compliance, and the ongoing mitigation efforts. This study—encompassing the literature published between 2019 and the first half of 2025—seeks to examine the entire Generative AI pipeline under the lights of the European Union (EU) copyright legislation. At the input stage, recurring issues include opacity surrounding the use of training data, misalignment between content licences and their usage, and the lack of mechanisms to regulate web scrapers collecting training material. Concerns also arise from fine-tuning aimed at stylistic emulation, as well as inadvertent memorisation and *verbatim* reproduction of training material. Various mitigation strategies have emerged: watermarking, adversarial perturbation, training data attribution with eventual royalties distribution, synthetic datasets, and machine-readable Text and Data Mining (TDM) opt-out mechanisms. Output-related vulnerabilities are similarly relevant, as adversarial attacks may extract sensitive data or generate style-imitating content. Retrieval-Augmented Generation (RAG) offers improved transparency, but remains prone to misattribution. The review underscores ongoing research dedicated to enhancing transparency in GenAI's output, including provenance-tracking protocols, blockchain-based integrity systems, watermarking techniques, and inference-time defences. Given the heterogeneity of the field, this review prioritised breadth over depth, leaving room for future studies while serving as a foundational guide for individuals with prior exposure or a developing understanding of the field. The findings underscore the necessity for enhanced copyright safeguards—potentially structured as multi-layered protective systems—and cross-sectoral collaboration to align rapid generative AI development with robust intellectual property protection.

Subjects Artificial Intelligence, Data Mining and Machine Learning, Databases, Multimedia
Keywords Generative artificial intelligence, Training datasets, Copyright, Intellectual property, Watermarking, Training data memorisation, Machine learning, Text and data mining, CDSM Article 4

Submitted 3 February 2026
Accepted 20 April 2026
Published 29 June 2026

Corresponding author
Anna Arnaudo,
anna.arnaudo@polito.it

Academic editor
Juan Lara

Additional Information and Declarations can be found on page 33

DOI **10.7717/peerj-cs.3928**

Copyright
2026 Arnaudo et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

How to cite this article Arnaudo A, Coppola R, Morisio M, Vetrò A, Borghi M, Raso R, Khan B. 2026. Generative AI and intellectual property: a systematic literature review of possible issues and mitigations. **PeerJ Comput. Sci. 12:e3928 DOI 10.7717/peerj-cs.3928**

INTRODUCTION AND BACKGROUND

The rise of generative artificial intelligence

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 (*Bommasani et al., 2022*). 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 (https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf). 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 (*Kowala, 2024*).

Legal background: the opt-out reservation

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). Copyright in the Digital Single Market (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 (*Manteghi, 2024*).

The presumption that CDSM Article 4 applies to AI training is often supported by the text of the AI Act itself, whose Article 53(1)(c) states that general-purpose AI model providers should have policies in place to

> *"identify and comply with the reservation of rights expressed by rightsholders pursuant to [CDSM] Article 4(3)".*

However, there is active debate about whether GenAI model training qualifies as 'Text and Data Mining' in the legal and technical sense.

This is highlighted in a recent study commissioned by the Policy Department of the European Parliament, which concludes that

> *"The current EU text-and-data mining (TDM) exception was not designed to accommodate the expressive and synthetic nature of generative AI training, and its application to such systems risks distorting the purpose and limits of EU copyright exceptions."* (*Lucchi, 2025*)

Notwithstanding this debate on interpretation of the scope of TDM, the prevailing state of legal discourse largely assumes that AI training is likely a form of TDM, and that there is a need to develop and interpret technical measures for rights reservations ('opt-out solutions').

Technical background: the quest for largely-accepted solutions

Given the lack of a single proven and broadly accepted solution, various actors have proposed guidelines, standards or protocols. Some examples are the adaptation of the Robots Exclusion Protocol (REP)—also known as robots.txt protocol—as well as TDM ReP—described in 'RQ1.3: What are the Technologies for Input Control, and How Can They Be Mapped to the Identified Issues?'—and other proprietary solutions—such as Google Extended. It should be noted, however, that increasing the cost or difficulty of accessing data may diminish AI technological progress (*Kowala, 2024*; *Manteghi, 2024*). Thus, both legal and technical mechanisms should be appropriately designed to balance copyright compliance and technological innovation.

Moreover, it is not sufficient to focus solely on the ingestion of training data; a comprehensive analysis of potential copyright issues must also consider the internal mechanisms of generative models that enable them to produce meaningful outputs. Several methods have been proposed to attribute the training samples that significantly contributed to a given generation. In addition, particular attention must be directed towards the outputs themselves, which ought to be explicitly tagged as 'synthetic' (as required by the AI Act) in order to ensure transparency and to support copyright-compliant labelling of media content.

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. For example, a known problem is the generation of images representing copyrighted characters (*He et al., 2024*). Even if these generations do not resemble any training data item, they may constitute a copyright infringement due to the represented subjects.

Moreover, copyright-protected content may be used to replicate an artist's style through including representative works in the input prompt 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.

Purpose and scope

This survey aims to explore the vast and fragmented landscape of established and developing techniques to mitigate copyright infringement throughout the entire GenAI pipeline, spanning both the input and output stages. Moreover, it also provides an analysis of the specific reasons why copyright holders are threatened by generative systems—such as the presence of the 'training data memorisation' phenomenon and the possibility to produce new artworks imitating the style of a given artist—with a narrowed attention to the European copyright law. This target was chosen to ensure the complete compliance of the mitigation strategies analysed with the regulatory needs.

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.

The analysis of the input pipeline encompasses the full suite of processes required to enable machine learning, including data collection, cleaning, curation, pre-training, and fine-tuning.

With regard to the output pipeline, several critical dimensions warrant consideration. These include the integration of Retrieval-Augmented Generation (RAG) technologies, the imperative to reliably label generative outputs in order to enhance transparency, and the potential for identifying copyright holders by associating generative outputs with the principal training samples from which they were derived.

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.

METHODOLOGY

Our survey is grounded in an 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 'Online Material'.

Related secondary studies

To the best of our knowledge, at the time of writing a single SLR has been conducted on the topic (*Vig, 2026*). The main information about it is reported in Table 1. However, it is 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 'Goals and Research Questions'.

Ren et al. (2024) present a cognate secondary study that similarly prioritises a 'technical perspective'. However, their published manuscript lacks transparency regarding the methodology utilised for primary study selection. The present research seeks to address this lacuna by adopting a robust, systematically grounded approach.

Finally, the survey by *Buick (2025)* constitutes a notable analogous study. It serves as a complement to *Ren et al. (2024)* by focusing primarily on potential regulatory frameworks. However, *Buick (2025)* similarly omits a detailed account of the underlying search strategy, which limits methodological comparability with the current research. Furthermore, while that study specifically targets transparency obligations concerning the input stages of GenAI systems, this work adopts a more comprehensive scope, investigating the copyright implications across both the input and output pipelines of generative systems.

Table 1 The related secondary study.

Year	2024
Reference	Intersection of generative artificial intelligence and copyright: an Indian perspective (*Vig, 2026*)
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

Table 2 Research questions (RQs).

GenAI input
RQ1.1 What types of copyright-protected input data are involved in each phase of the GenAI pipeline?
RQ1.2 What are the potential copyright compliance issues—with respect to EU legislation—related to the different methods and data input phases?
RQ1.3 What are the mitigation strategies that can be mapped to the issues identified by RQ1.2?
GenAI output
RQ2.1 What are the different methods and phases for generating content with GenAI models that may have copyright implications?
RQ2.2 What are the technologies for output control, and how can they be mapped to the issues identified by RQ2.1?



Figure 1 High-level schema of the GenAI pipeline and the intersection with our research questions (RQs). Full-size DOI: 10.7717/peerj-cs.3928/fig-1

Goals and research questions

The Research Questions (RQs) adopted for guiding the SLR are reported in Table 2, while Fig. 1 illustrates the relationships between the questions and the related stages of the GenAI pipeline from a high-level perspective—*i.e.*, where only the main relevant steps are reported.

Search approach

We followed the steps inspired by *Garousi, Felderer & Mäntylä (2019)*:

- **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,

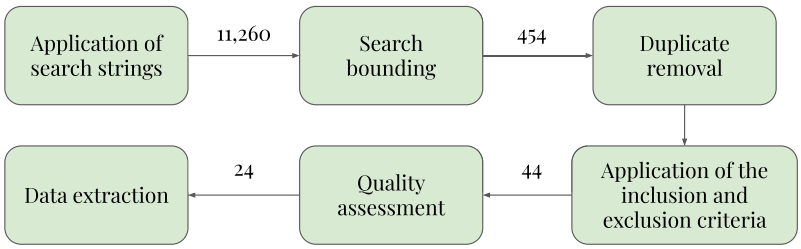


Figure 2 Phases of the SLR process. The arrows are labeled with the number of sources selected after the corresponding phase. Full-size DOI: 10.7717/peerj-cs.3928/fig-2

Springer, ScienceDirect) and Google Scholar. A publication-year filter was applied to select only articles published between 2019 and the early 2025. Since we decided to target EU legislation, this time constraint was chosen in order to ensure a better alignment with the post-CDSM Directive landscape.

- **Search bounding:** Following Garousi, we restricted the scope (*e.g.*, first 100 Google results), yielding 454 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, 44 sources remained.
- **Quality assessment:** Using Kitchenham and Charters' guidelines (*Kitchenham & Charters, 2007*), two authors independently evaluated all articles through tailored questionnaires (See Appendix 'A Online Material'). Sources scoring above the average threshold (3.48/5) were retained, resulting in 24 high-quality studies (the list is available online; See Appendix 'A Online Material').
- **Data extraction:** Relevant information from the final set of studies was extracted in structured form (available online; See Appendix 'A Online Material').

Figure 2 summarises the workflow and the number of sources selected at each step; further details are available in the online appendix (See Appendix 'A Online Material').

Inclusion and exclusion criteria

Inclusion criteria (IC) and exclusion criteria (EC) were defined to ensure the collection of only the sources relevant to our research goal.

Inclusion criteria:

- The source is directly related to techniques used to facilitate GenAI-related copyright compliance; those may involve techniques used either by copyright owners or by GenAI developers and applied to the input or the output phases of the GenAI pipeline;
- The literature item is written in a language that is directly comprehensible by the authors: English or Italian;
- The source is an item of white literature with the full text available for download and is published in a peer-reviewed journal or conference proceedings.

Exclusion criteria:

- The source is related to AI but not to GenAI;
- The source refers to copyright legislation which diverges from that of the European Union;
- The source refers to legislation older than the DSM copyright directive (2019);

The 454 sources selected by the previous phase have been filtered according to these criteria, resulting in 44 accepted articles.

Quality assessment of sources

To evaluate the quality of the sources in our pool, we followed the guidelines proposed by *Kitchenham & Charters (2007)*. The questions used for assessing the relevance of the identified items, reported in the online Appendix—see 'Online Material'—are inspired by those used in the SLR by *Cursino et al. (2018)*, with an adaptation to the topic of GenAI and Copyright. In particular, since we found two main categories of primary studies, we formulated specific assessment questions:

- **Primary studies presenting a new technology:** These articles showcase innovative solutions. To evaluate the article, it is important to focus on the accuracy of the study validation;
- **Exploratory studies:** Articles that broadly discuss the interaction between GenAI and copyright. In those cases, it is important to focus on the reliability of the data sources and the quality metrics eventually employed to perform the comparisons.

Note that there are some exceptions in which a given source can belong to both categories, *e.g.*, a article presenting an innovative technique also discussing other related approaches.

Two authors independently evaluated all sources, and discussions were conducted when opinions diverged, to finally reach a common agreement.

Since the two quality assessment questionnaires used different scales, we normalised the score for each source on a 0–5 interval. Next, we calculated the average score—which was 3.48/5. Only the sources which achieved a score above this threshold were selected for the next phase, *i.e.*, the data extraction. This selection procedure aims to maximise the relevance of the present work, targeting only works answering our questions 'better than the average'.

Final pool of sources

The SLR procedure led to a final set of 24 sources. The interested reader can find their details in the online appendix (see Appendix 'A Online Material'). As can be seen from the diagram in Fig. 3, the input and output stages are mentioned in a balanced way in the analysed literature.

RESULTS

GenAI input

In Fig. 4, the distribution of the phases within the GenAI input pipeline, as referenced in the selected sources, is presented. Although there may be further steps in an input pipeline,

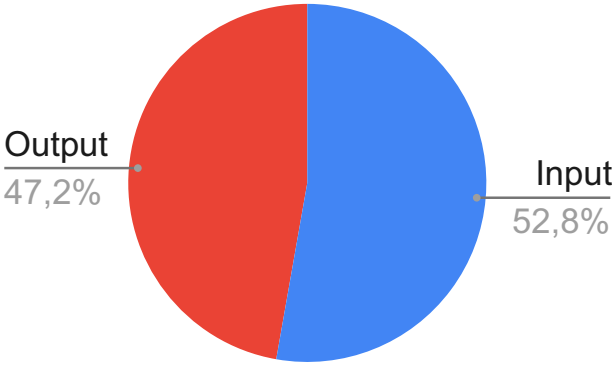


Figure 3 Distribution of the analysed sources between the input and output stages of the Generative AI pipeline. Some articles cover topics related to both. Full-size DOI: 10.7717/peerj-cs.3928/fig-3

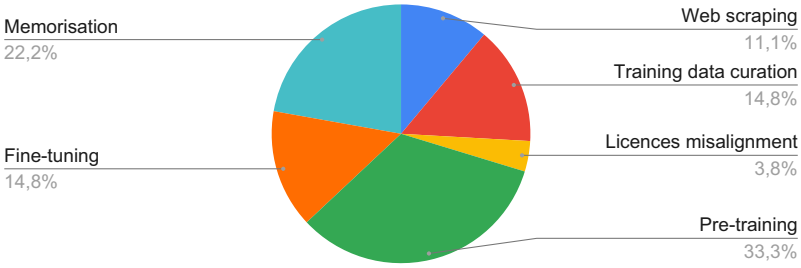


Figure 4 Distribution of the phases within the GenAI input pipeline explicitly mentioned by the selected sources. Full-size DOI: 10.7717/peerj-cs.3928/fig-4

these are the key steps relevant for the copyright issues mentioned in the literature analysed (see 'RQ1.2: What Are the Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?' for more details). Moreover, Figs. 5 and 6 report some statistics about the keywords mentioned in the literature analysed.

RQ1.1: What types of copyright-protected input data are involved in each phase of the GenAI pipeline?

Summary of the Answer to RQ1.1—Types of Input Data.

It emerged that there is a marked lack of transparency regarding the specific training data employed in machine learning processes. This is especially true for commercial models, with AI companies often refraining from disclosing very fragmented composition of large training *corpus*. Common Crawl was identified as one of the most referenced sources. Notably, press contents and open-source code appear 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

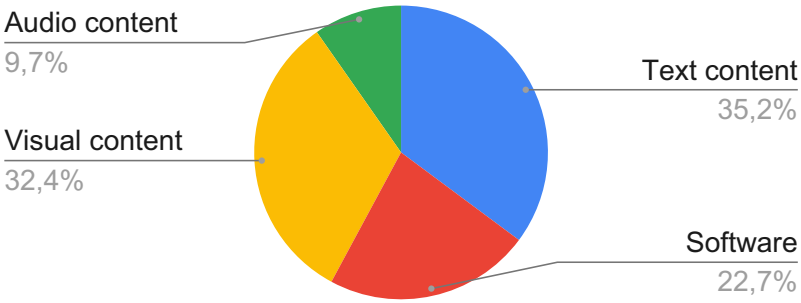


Figure 5 **Distribution of the data formats explicitly mentioned by the selected sources.**
Full-size DOI: 10.7717/peerj-cs.3928/fig-5

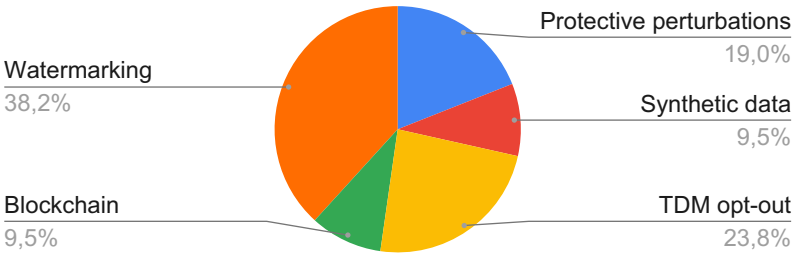


Figure 6 **Distribution of the mitigations to the input copyright issues explicitly mentioned by the selected sources.**
Full-size DOI: 10.7717/peerj-cs.3928/fig-6

models underpinning 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 (https://osai-index.eu/the-index)—hinders transparency regarding the actual training samples used (*Widder, Whittaker & West, 2024*). 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 (*Widder, Whittaker & West, 2024*; *Liesenfeld & Dingemanse, 2024*).

Regulations might be advocated to increase transparency or allow companies to keep model details secret to protect intellectual property or other proprietary interests (*Schneider, 2024*). However, AI Act Article 53(1)(d) (https://artificialintelligenceact.eu/article/53/) prescribes:

> *"In order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and* ***make publicly available a sufficiently detailed summary of the content used for training*** *the general-purpose AI model. While taking into due account the need to protect trade secrets and confidential business information, this summary should be generally comprehensive in*

its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights."

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 (*Sakib et al., 2024*):

- 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 (https://huggingface.co/)).

In the following, the results are grouped around the main topics mentioned in the analysed literature. In most of the real-word scenarios, training datasets are constituted by data from the various sources listed below in different portions. By combining multiple typologies, AI developers can improve the balance and representativeness of the models (*Lukas, 2023*).

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 (https://www.geeksforgeeks.org/difference-between-web-scraping-and-web-crawling/). For example, blog posts and the content of websites like Wikipedia are extracted *via* web scraping. Project Gutenberg offers dozens of thousands freely accessible books available online, a text *corpus* used to train Generative AI (*Lukas, 2023*). Some examples of 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 especially 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 (*Kowala, 2024*). For example, OpenAI has made an agreement with the Associated Press (apnews.com) for the use of its archive for 2 years in order to training ChatGPT (*Lukas, 2023*).

Another target of scraping is open-source code, as exemplified by the 'Doe v. GitHub' case discussed in 'RQ1.2: What Are the Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?' (*Samuelson, 2023*).

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 (*Ludwig et al., 2024*).

However, as declared by the founders of Common Crawl (https://commoncrawl.org/mission), the primary objective of the initiative was not to furnish a dataset specifically intended for training generative AI models, but rather to democratise the access to web data (*Knibbs, 2024*). Consequently, CCBot (Common Crawl's crawler) was not designed to enforce copyright-safe collection, and AI firms that use Common Crawl without proper data cleaning and filtering may encounter copyright compliance issues.

With regard to image datasets, some examples are the ImageNet (*Deng et al., 2009*) and the LAION-5B (*Schuhmann et al., 2022*) datasets. These corpora 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.

Finally, another input format that has been at the centre of copyright disputes in the last years is source code of software programs. The technical sophistication and functionality of software programs make them distinct from text and audio in copyright law. While text and music cases often rely on the 'idea *vs.* expression' distinction, this standard is difficult to apply to functional code. Consequently, the legal assessment of fair use remains a unique and open challenge for code-based models. On online platforms, there is a wide availability of open-source code, which is often subject to strict licences, such as copyleft licences (*Ma et al., 2024*; *Ren et al., 2024*).

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 (*Ludwig et al., 2024*).

These characteristics may lead to a reduced risk of copyright infringement. Indeed, it becomes more plausible that copyright-protected material is identified and excluded during the more meticulous data preprocessing conducted on fine-tuning datasets. Nonetheless, as elucidated in 'RQ1.2: What Are the Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?', emergent fine-tuning paradigms may be leveraged to embed bespoke attributes within model outputs, potentially facilitating stylistic mimicry that contravenes extant copyright frameworks.

RQ1.2: 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 RQ1.2—Copyright Compliance Issues along the Input Pipeline.

Machine learning models are not databases; they don't store training data. However, loading copyrighted content into RAM—*e.g.*, during training—is legally considered reproduction, and is covered by the copyright law.

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. Indeed, many contents are easily available online without paywalls, even if they are copyright-protected.
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 'training data memorisation', where generative AI models reproduce their training data, thereby unintentionally disseminating copyright-protected material.

The discussion is divided into paragraphs related to the main keywords encountered during the review.

Web Scraping: Many contents are easily available online without paywalls, even if they are copyright-protected. Public availability does not imply consent for AI training. Even if content is posted consciously, it does not mean the owner has agreed to this use. Social media policies often allow training, but typically only for the platform operator itself (*Lukas, 2023*). 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 (*Ludwig et al., 2024*).

In the context of the dispute with Google before the French Competition Authority mentioned in 'Introduction and Background', 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 (*Kowala, 2024*).

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, saving copyrighted content or loading it into RAM is legally considered reproduction, regardless of whether the copies are permanent or temporary (*Lukas, 2023*). While copyright protects ideas rather than expression, GenAI training technically requires making digital copies to analyse the data, which triggers copyright restrictions (*Manteghi, 2024*).

In addition, 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. Matching the licences for a huge amount of digital data, the ownership of which is usually difficult to reconstruct, is a complex task (*Manteghi, 2024*). As models increase in size, it becomes more difficult to properly ensure that input data is not used without authorisation (*Sakib et al., 2024*). 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 (*Longpre et al., 2023*).

In 2022, 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 (https://www.courtlistener.com/docket/65669506/doe-1-v-github-inc/). According to *Samuelson (2023)*, 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.

Fine-Tuning: Recently, fine-tuning methods compatible with generative models belonging to the family of Stable Diffusion[1] 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 fine-tuning methods are Textual Inversion (*Gal et al., 2022*), DreamBooth (*Ruiz et al., 2023*), and Custom Diffusion (*Kumari et al., 2023*). 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 (*Ren et al., 2024*).

[1] Stable Diffusion (https://stablediffusionweb.com/it) is a generative model designed for text-to-image generation through a diffusion process—*i.e.*, exploiting the capability to add/remove random noise from images.

Training Data Memorisation: Machine learning models are not databases; they do not store training data (*Lukas, 2023*). However, the phenomenon of 'Memorisation' has been widely observed. For example, Midjourney's generative model was accused of producing images strictly resembling commercial films (*Ren et al., 2024*). Moreover, the New York Times has sued OpenAI and Microsoft, alleging they used its articles to train AI without permission. The lawsuit specifically claims ChatGPT reproduces *verbatim* excerpts from paywalled content without providing attribution or referral links (*Ren et al., 2024*).

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, in whole or in part. This is undesirable, as GenAI systems are intended to create novel content distinct from their training data. However, this has been observed across various model architectures (*Ren et al., 2024*). For instance, *Carlini et al. (2023)* systematically examined memorisation in both Diffusion Models and Language Models. 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 (*Cooper & Grimmelmann, 2025*) 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. *Lee et al. (2022)* estimated that LLMs' training datasets contain a number of duplicated and near-duplicated strings that leads to nearly 1% of language model outputs being their *verbatim* copies (*Ren et al., 2024*). Moreover, there exist attack methodologies—namely inversion attacks—crafting prompts specifically designed to extract memorised training data (*Carlini et al., 2023*; *Somepalli et al., 2022*; *Balan et al., 2023*; *Zhang et al., 2020*).

In the context of the earlier mentioned lawsuit against OpenAI and GitHub, the code generated by Copilot and Codex was flagged as potentially infringing (*Samuelson, 2023*). 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 without including a proper attribution (*Ma et al., 2024*).

RQ1.3: What are the technologies for input control, and how can they be mapped to the identified issues?

Summary of the Answer to RQ1.3—Mitigations for the Input Issues

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 traceability and ownership verification—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 invisible 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, while the strategies for erasing the model's knowledge—referred as 'machine unlearning'—remain complex and expensive.

Finally, more diplomatic strategies—centred 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. However, the respect of these preferences is not forced by

Table 3 Mapping between the input issues identified in 'RQ1.2: What are the potential copyright compliance issues—with respect to EU legislation—related to the different methods and data input phases?' and possible mitigations strategies identified in 'RQ1.3: What are the technologies for input control, and how can they be mapped to the identified issues?'.

Issue	Mitigations
Web scraping	Text poisoning, protective images' perturbations, membership inference, synthetic data, solutions for Expressing Text and Data Mining (TDM) Reservation (opt-out), c2pa, watermarking;
Misalignment of licences	Watermarking, fingerprinting for authorship attribution, synthetic data, blockchain, c2pa, solutions for Expressing Text and Data Mining (TDM) Reservation (opt-out);
Fine-tuning for style mimicry	Text poisoning, protective images' perturbations, fingerprinting for authorship attribution, machine unlearning, generation filters;
Training data memorisation	Text poisoning, protective images' perturbations, synthetic data, collecting royalties for copyrighted training data, attribution to originating training samples, fingerprinting for authorship attribution, machine unlearning, knowledge distillation, generation filters, watermarking;

any technical deterrence: for this reason, our review also includes the retrieved information about the effectiveness of Membership Inference Attacks (MIAs), which can be used to reveal whether a given data point was used during training.

In the following, the mitigation strategies proposed in the analysed literature are discussed, while Table 3 schematises how these techniques can be mapped to the identified issues.

Text Poisoning: *Zhao et al. (2024a)* proposed Truncation Protection Examples (TPEs). These are specially manipulated texts designed to make the model to output the special final token, effectively stopping further generation. The technique is intended as a safeguard to prevent users from inserting protected text into LLMs and eliciting potentially infringing continuations. To convert plain text into TPEs, *Zhao et al. (2024a)* introduced an algorithm that aims to minimise semantic alteration. The experiments showed strong effectiveness in some areas, such as mathematics, where truncation efficacy reached 99%, but much weaker results in others, like coding, where effectiveness dropped to 22%. A further limitation is that generating TPEs introduces manipulations into the protected text itself, which may make the text appear noisy or less readable to users.

Ji, Ma & Wang (2022) developed a method to protect open-source code by applying lightweight transformations. By using a pre-trained model like CodeBERT to select specific changes, they turn standard code into 'unlearnable examples' (*Ren et al., 2024*).

A different approach involves encoding the content into special Unicode characters, making it impossible for AI systems to process it (*Zhao et al., 2024a*). An example is the technology proposed by DataDust.ai (https://www.datadust.ai/), which protects text content from AI-powered scrapers by using a text font that AI cannot interpret. This strategy proves effective because typically AI-powered scrapers have not been trained on such rare or non-standard fonts. However, as *Zhao et al. (2024a)* pointed out, it could be easily bypassed by Optical Character Recognition (OCR).

Visual Content—Protective Perturbations: Recent investigations have examined the integration of imperceptible perturbations into images as a means of preventing

unauthorised exploitation, particularly in scenarios where powerful fine-tuning techniques are employed to inject personalised data into Stable Diffusion models[2].

Notably, Glaze (*Shan et al., 2025*) is designed to optimise 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 (*Ren et al., 2024*).

Similarly, Anti-DreamBooth (*Le et al., 2023*) explicitly incorporates the DreamBooth[3] fine-tuning process into its framework by formulating a bilevel min-max optimisation procedure to generate protective perturbations (*Ren et al., 2024*).

Nightshade (*Shan et al., 2024*) 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 of re-training the model through the exclusion of the poisoned images.

Further research efforts (*Ye et al., 2023*; *Zhao et al., 2024b*) have also explored the generation of protective noise using analogous adversarial techniques.

Zhao et al. (2024c) 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[4]. 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 (*Zhao et al., 2024c*; *Ren et al., 2024*).

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 publicly disseminated (*Zhao et al., 2024c*).

To further demonstrate the limited robustness of protective perturbations, *Zhao et al. (2024c)* 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

[2] See 'RQ1.2: What Are the Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?' on Fine-Tuning for more details.

[3] Described in 'RQ1.2: What Are the Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?' in the paragraph dedicated to fine-tuning.

[4] 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, 1 year after the study conducted by *Zhao et al. (2024c)*

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.

Synthetic Data: Synthetic data are a specific corner case of a generative model's output. They can be generated from real-world data by capturing underlying properties, or using numerical models and simulations. 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 (*Ludwig et al., 2024*; *Manteghi, 2024*).

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 are scarce. Moreover, relying on synthetic data may reduce 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 generation process. Indeed, synthetic outputs have often a high probability of resembling the seed data (*Manteghi, 2024*). The quality, diversity, and fidelity of synthetic data are expected to improve rapidly, allowing for more accurate and versatile applications in generative AI training. Beyond technical performance, this approach offers another significant regulatory advantage: developers who build datasets using synthetic information maintain full control over the generation techniques and methods. This level of oversight makes it much easier to identify and disclose data sources, helping providers meet the transparency obligations mandated by the AI Act (What Types of Copyright-protected Input Data are Involved in Each Phase of the GenAI Pipeline?).

Furthermore, it is argued that the quality, reliability and diversity of GenAI models could degrade over generations if training datasets consist only of synthetic data without fresh real-world data (*Manteghi, 2024*). In conclusion, while synthetic data offer 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.

Collecting Royalties for Copyrighted Training Data: Alongside licensing agreements between AI developers and copyright owners, one proposed solution for the legal use of copyright-protected works in training is to pay royalties to copyright owners (*Manteghi, 2024*). Implementing remuneration programs could ensure creators are duly compensated for copyrighted materials used in AI systems. Such models provide a persuasive motivation for content producers to contribute their work as training data. Under a revenue-sharing arrangement, creators would receive a pre-established portion of the AI system's revenue, potentially scaled to the proportion of their contribution to the training set. Alternatively, a royalty-based model could involve a set fee for each use of their work, structured either as a fixed amount per use or as a percentage of the specific revenue generated by the AI system (*Lucchi, 2024*).

However, the large volume of training samples, coupled with the variability in their relative value across different GenAI systems, makes the attribution of royalties a complex task. Local explainability methods—like perturbation, gradients, and Shapley values—identify feature importance for specific outputs. In contrast, global methods analyse the model's overall behaviour by probing its internal states and training data distribution (*Sakib et al., 2024*).

A simple approach uses similarity scores between training data and generated content as a valuation metric. For example, *Zhang (2023)* proposes a scoring service for data use based on text similarity. *Lu et al. (2023)* propose WASA, an approach to attribute the sources of generative outputs using watermarks. *Sakurai et al. (2023)* propose a fine-tuned version of BERT tailored for the author attribution task. To address the substantial computational costs noted in prior studies, they applied distillation techniques (knowledge distillation is discussed later in this section) aimed at reducing overhead—achieving a sixfold speed-up—and improving the interpretability of the model's outputs.

In contrary, *Wang et al. (2024)* proposes to mitigate the black-box nature of content generation by leveraging the probabilistic nature of generative models: the log-likelihood of generating a content is used to measure the utility of the training data from which it derives. Royalties are subsequently distributed among copyright holders based on their respective contributions, which are analytically determined through the application of Shapley value theory (*Shapley, 1953*). Although the approach proposed by *Wang et al. (2024)* 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.

Membership Inference: Membership Inference Attacks (MIAs) aim to reveal whether a specific data sample was part of a model's training set, making them useful for checking compliance with copyright holders' preferences. They exploit models' tendency to handle training data better than unseen data when included in input prompts at inference time (*Li & Zhang, 2025*). 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 MIAs are typically complex and expensive, *Marone & Durme (2023)* 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 'What Types of Copyright-protected Input Data are Involved in Each Phase of the GenAI Pipeline?', GenAI companies often refrain from disclosing the composition of their training corpora.

Ma et al. (2024) proposed applying Likelihood Radio Test (LRT) to perform membership inference, achieving good results and demonstrating the applicability of this method. They also made available an open-source tool called Codeforensic, which allows to compute the probability that a code snippet has been used to train a given model for

code generation. However, for its proper functioning, this tool needs access to the probabilities computed by the model during inference, which is not feasible in case the generative system is queried through online services.

Solutions for Expressing Text and Data Mining (TDM) Reservation (opt-out): In response to the EU Digital Single Market (CDSM) Directive, previously mentioned in 'Introduction and Background', several initiatives were introduced to establish a standardised 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. Once standard TDM reservation protocols are uniformly adopted, they should be explicitly taken into account in the technical design of web scrapers (*Lukas, 2023*).

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 (https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/). The latter has been designed 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 (*Kowala, 2024*).

OpenAI explicitly pointed out the compatibility of its scrapers with the Robots Exclusion Protocol (also called robots.txt) (*Lukas, 2023*; *Manteghi, 2024*). For three decades, robots.txt has provided a low-cost way to manage search engine access, since it leverages a simple file in the websites' root directory containing the list of the resources the crawlers should not process. Because the file size is strictly limited to prevent memory issues, website owners often cannot include detailed rules for extensive resources. Furthermore, the protocol is entirely voluntary; 'careless' AI trainers can simply ignore these instructions since they do not function as a mandatory digital fence. Even when bots do respect the exclusion rules, they can inadvertently bypass them by accessing protected content through third-party sites that link to the restricted data but do not use their own robots.txt restrictions (*Manteghi, 2024*). To mitigate this issue, the spawning.ai website provides a generator for an ai.txt file, which allows rights holders to set explicit, machine-readable permissions for AI training. Unlike the traditional robots.txt protocol—which is typically checked at the start of a crawl to manage site access—the ai.txt file is designed to be read when content is actually downloaded (*Manteghi, 2024*).

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 (*Manteghi, 2024*). Moreover, Google indicated that exercising this opt-out does not affect the indexing on Google Search (*Kowala, 2024*; *Google, 2023*), thus theoretically addressing the issues mentioned in 'RQ1.2: What Are the

Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?'. However, despite reserving their rights through Google–Extended, some press publishers noticed that their publications still appeared in the responses offered by Gemini (*Kowala, 2024*; *Autorite de la Concurrence, 2024*).

Through a partnership with the startup Spawning, Stability AI integrated an opt-out mechanism that allowed rights holders to manage their presence in AI training sets. Artists used the 'Have I Been Trained?' platform to identify their works within LAION datasets (described in 'What Types of Copyright-protected Input Data are Involved in Each Phase of the GenAI Pipeline?') and request their removal. As a result of this collaboration, approximately 80 million images were excluded from the training data for Stable Diffusion 3 (*Ren et al., 2024*).

Microsoft also introduced its 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 established mechanism for checking into private AI training datasets (*Kowala, 2024*; *Autorite de la Concurrence, 2024*).

A different approach comes from the proposal of C2PA as a common standard solution for expressing TDM reservations. In fact, in addition to the provenance tracking functionality provided by C2PA (described in 'RQ2.2 What are the Technologies for Output Control, and How Can They Be Mapped to the Identified Issues?'), C2PA manifests also allows embedding Training and Data Mining Assertions. These can be used to bind a digital asset with the relative opt-out preferences in a tamper-proof way, having the integrity ensured by a digital signature. In particular, for any given TDM action between AI training, GenAI training[5], 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.

[5] 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.

Unlearning and Knowledge Distillation: A fundamental question is whether an AI model can unlearn material, which would be especially useful if training occurred before rights were reserved (*Ren et al., 2024*). However, current research shows that unlearning remains a highly complex task (*Kowala, 2024*).

Several approaches have been proposed. *Zhang et al. (2024)* categorised them into:

- **Exact Machine Unlearning** requires retraining the model while excluding the targeted samples, thereby ensuring those data points are completely removed from the model's knowledge.

  Methods like SISA (*Bourtoule et al., 2020*) aim to reduce the cost of retraining from

scratch by partitioning the training dataset and process. After each segment, the model's weights are saved, allowing retraining to resume from the last checkpoint before the data to be unlearned was introduced. However, the computational savings vary, as they depend on when the unlearning data entered the partitioned training process.

- **Approximate Machine Unlearning** adjusts model weights through fine-tuning to mimic the removal of specific data. Because of its reduced computational cost, this approach is often preferred (*Ren et al., 2024*). Examples include Stable Sequential Unlearning (SSU) (*Dou et al., 2025*) and Approximate Unlearning with Idiosyncratic Expressions Replacement (*Eldan & Russinovich, 2023*). A key drawback is the risk of over-unlearning, where performance on unrelated tasks is degraded.

Knowledge distillation is an alternative technique; starting from a base model, it produces a distilled version by optimising the use of the internal parameters and thus reducing their number. Larger AI models are more prone to memorising training data (RQ1.2: What Are the Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?). For example, DistilBERT is a distilled version of BERT. It has 40% fewer parameters and retains 97% of the original performance while offering faster inference and better interpretability. These compact architectures should be preferred because their reduced tendency to overfit makes them more secure (*Sakib et al., 2024*).

GenAI output

Figures 7 and 8 report some statistics on the issues and mitigations concerning the output pipeline of GenAI systems.

RQ2.1 What are the different methods and phases for generating content with GenAI models that may have copyright implications?

Summary of the Answer to RQ2.1—Copyright Compliance Issues along the Output Pipeline

Summary of the Answer to RQ2.1

When examining the generation phases of GenAI systems, a number of vulnerabilities have been identified that may lead to potential copyright infringement.

Indeed, various forms of adversarial attacks have proven effective in circumventing system-level output filters, with the intent of extracting information or illicitly generating new content by mimicking the style of a particular artist. Consequently, the proper tagging of AI Generated Content (AIGC) has been emphasised as a necessary prerequisite for fostering transparency.

Moreover, we observed that Retrieval-Augmented Generation (RAG) may be controversial: while enabling a more transparent connection between AIGC and the underlying data sources—significantly more transparent than the opaque relationship between training data and model-derived outputs, due to the non-interpretable nature of model parameters—improper configuration can result in the misattribution of sources or violations of licensing terms.

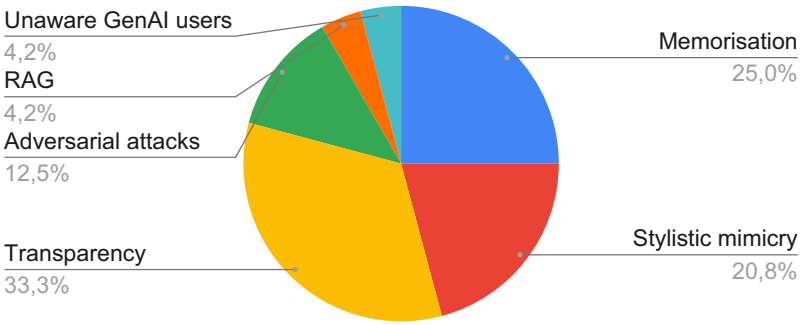


Figure 7 The distribution of issues across the generative AI output phases explicitly identified in the literature under review. Full-size DOI: 10.7717/peerj-cs.3928/fig-7

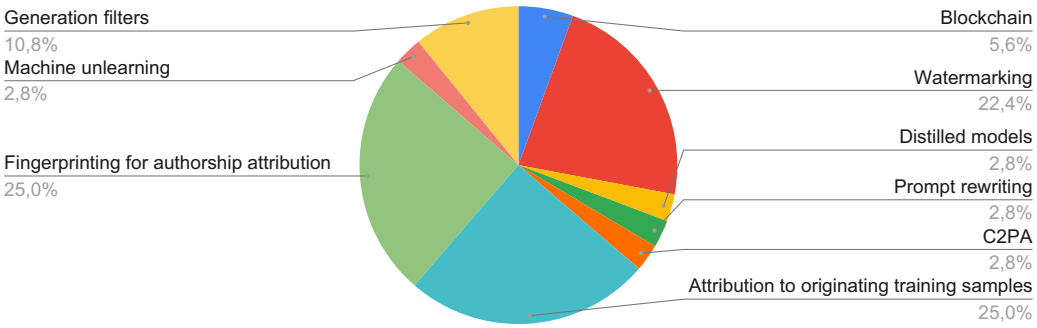


Figure 8 The distribution of mitigations relative to the generative AI output phases explicitly identified in the literature under review. Full-size DOI: 10.7717/peerj-cs.3928/fig-8

Finally, we examined the generation of images depicting copyright-protected characters as a representative case of infringement, even in the absence of direct memorisation.

The discussion has been organised into paragraphs corresponding to the principal themes identified.

Adversarial Attacks *Zhao et al. (2024a)* identified multiple copyright threats arising from adversarial prompting attacks. These fall into two categories: manually constructed prompts requiring greater complexity (*Liu et al., 2024*), and methods to algorithmically generate prompts and automatically optimise their construction (*Shin et al., 2020*; *Guo et al., 2021*). Often, they are able to bypass the security mechanisms of the GenAI system, designed to block the generation of sensitive and/or harmful content.

For example, a writer's style can be emulated through few-shot prompting techniques[6]. Attacks also include the possibility of exploiting the phenomenon of memorisation, in which a model is fed a portion of a target text to extract the remaining part (for more information, see 'RQ1.2: What Are the Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?').

[6] Few-shot prompting refers to the practice of including in the input prompt a varying number of samples to be taken as reference for the model's generation process.

Another example is Textual Inversion: an image modification technique for Stable Diffusion that operates without the need for training or fine-tuning. The process extracts a detailed description from a source image and embeds it into a string. Users can then employ this string to compose new prompts that generate images featuring the original object or style, potentially infringing on the copyright of the source material (*Ren et al., 2024*).

Finally, voice synthesis can now generate realistic audio from only a few seconds of a recording. This makes protection techniques such as watermarking essential to prevent model misuse and unauthorised infringement of a speaker's voice data (*Ren et al., 2024*).

Unaware GenAI Users: Challenges arise not only between AI developers and copyright holders, but also between the latter and the users of AI models. In fact, users may not be aware that a protected work has been reproduced in whole or in part in the AI output generated during their usage. Subsequently, if they distribute such content, they may be unaware of committing an infringement of exclusive rights (*Kowala, 2024*).

Retrieving Data at Inference Time: Data used at inference time, *i.e.*, for Retrieval Augmented Generation (RAG) include data that changes frequently and cannot be meaningfully included in model training (*Ludwig et al., 2024*).

Typically, there are two types of RAG systems: static and dynamic.

The first consists of a memory source, *e.g.*, a database, connected to a GenAI system that can take into account the information contained during the generation process.

Secondly, dynamic RAG systems leverage web scrapers—which are often powered by AI themselves to be able to interpret webpages' content and retrieve relevant information. An example is the 'Search Online' functionality provided by ChatGPT, which triggers the GPTBot to scrape data to be injected in the subsequent ChatGPT's output.

Both the static and the dynamic mechanisms facilitate proper data attribution, since—differently from what happens when outputs are generated using only the information encoded into the model's parameters—there is a clear link between the source and the generation.

However, the presence of an attribution mechanism may not be enough to completely avoid any copyright infringement, since data may be associated with licences that deny reproduction at all.

Thus, it is important to design RAG scrapers that have the ability to properly extract and discriminate the usage rights associated with the data.

Moreover, there is also the possibility of hallucinations to happen during the attribution task: in that case, the malfunctioning can also result in copyright infringement. More generally, hallucination combined with correct attribution can also lead to copyright infringement claims, on the basis of the integrity of the work to the attributed author.

Finally, the ability to query a GenAI system instead of directly accessing the original websites may result in an economic loss for those websites, particularly when their revenue models rely on visit counts.

Generation of Copyrighted Characters' Representations: Even without reproducing memorised training samples (see 'RQ1.2: What Are the Potential Copyright Compliance

Table 4 Mapping between the issues identified in 'RQ2.1 What are the different methods and phases for generating content with GenAI models that may have copyright implications?' and possible mitigations strategies identified in 'RQ2.2 What are the technologies for output control, and how can they be mapped to the identified issues?'.

Issue	Mitigations
Adversarial attacks	Text/image poisoning, watermarking, distilled models, prompt rewriting, generation filters
Unaware genAI users	Fingerprinting for authorship attribution, attribution to originating training samples, blockchain, C2PA, watermarking
Retrieving data at inference time	Fingerprinting for authorship attribution, blockchain, C2PA, watermarking
Generation of copyrighted characters' representations	Prompt rewriting, generation filters, fingerprinting for authorship attribution, attribution to originating training samples

Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?'), generative models may infringe copyright by creating images of protected characters. In this case, infringement can occur if the model was trained on copyrighted characters and prompted with matching descriptions (*He et al., 2024*).

RQ2.2 What are the technologies for output control, and how can they be mapped to the identified issues?

Summary of the Answer to RQ2.2—Mitigations for the Output Issues

There is a notable effort in adapting text authorship attribution methods to the new challenges posed by LLMs.

The imperative for transparency has also prompted the development of provenance tracking protocols—such as C2PA—supported by robust technologies, including the digital signature and blockchain.

There is also notable research around watermarking algorithms, which could allow the embedding of provenance information directly into the AI Generated Content (AIGC).

Our review also included some proposed methods that intervene at inference time to safely block certain adversarial attacks and the generation of copyright-protected characters' representations.

In the following, the mitigation strategies proposed in the analysed literature are discussed, while Table 4 highlights how these techniques can be mapped to the identified issues.

Text Fingerprinting: Incorporating source citations in language models is a crucial practice for maintaining text accuracy and credibility while curbing the spread of misinformation. Providing transparency regarding data sources helps models transition from 'black boxes' to verifiable tools, a shift that is equally important for AI and human authors (*Lucchi, 2024*).

A recent study (*Ribeiro, Carvalho & Coheur, 2024*) proposed an author classification method based on Fuzzy Fingerprints, which are sets of text features (*e.g.*, word patterns) compared using fuzzy logic. Unlike binary logic, where values are strictly true or false (0 or 1), fuzzy logic allows for degrees of truth between 0 and 1, enabling more flexible and

nuanced comparisons. Although not originally developed for GenAI, this method could be adapted for output filtering to help prevent copyright-infringing content. The approach requires pre-computed fingerprints derived from representative samples of an author's writing, which poses scalability issues given the potentially large number of authors in training datasets. Despite this limitation, the researchers achieved 67.4% attribution accuracy on a benchmark involving 50 authors.

The method introduced by *Ribeiro, Carvalho & Coheur (2024)* stands out from other AI-based approaches to authorship attribution, many of which relied mainly on fine-tuning classifiers on labelled author texts. For instance, *Sakurai et al. (2023)* fine-tuned a model using text from eight Japanese authors, achieving 88.4% accuracy. However, they noted that fine-tuning alone was impractical due to heavy computational demands. Consequently, they decomposed BERT into multiple distilled models running in parallel. On 371 texts from the same eight authors, this setup reduced inference time from 20.33 to 3.38 s but decreased accuracy by about 10%.

In addition, logic-based plagiarism detection methods have gained attention, which convert text into structured representations for similarity computation. While these methods improve the detection of complex content—such as software code—they are often hindered by high computational costs and slow processing speeds at scale. To address this, *Zou et al. (2024)* proposed a new method employing Chain of Thought (CoT)[7] that allows for rapid and accurate analysis by progressively breaking down problems into smaller components. By mapping these logical steps into feature vectors, the CoT approach provides a foundation for precise similarity analysis across large-scale text and code datasets.

[7] Chain of Thought (CoT) is a technique frequently employed in the context of Generative AI. It prescribes to divide a complex problem into multiple intermediate steps that compose the 'chain'.

Li et al. (2023) prioritised the safe release of StarCoder—a model for code generation—by developing a specialised attribution tracing tool to track code origins. Building on the idea of transparency, *Yan & Li (2022)* created WhyGen, which explains how a model generated a specific piece of code by linking it back to its training data (*Ren et al., 2024*). The tool captures a unique 'inference fingerprint' during the generation process. This digital signature is then used to identify the most similar examples from the original training set, effectively showing the user exactly which data points influenced the model's output (*Ren et al., 2024*).

Image Fingerprinting: The EKILA framework (*Balan et al., 2023*) 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)—also called blockchain—alongside Non-Fungible Tokens (NFTs), to store rights-related data in a decentralised and tamper-proof manner. These are ultimately integrated with the Coalition for Content Provenance and Authenticity (C2PA) standard (described later in this section), binding information regarding both the generative model and the specific training data responsible for the generated content.

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 optimised to compute similarity score even with manipulated or degraded images. The experimental results show that this method outperforms existing similarity metrics such as CLIP (*Radford et al., 2021*), LPIPS (*Zhang et al., 2018*), and SIFID (*Shaham, Dekel & Michaeli, 2019*).

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 rigour offered by methods such as leave-one-out re-training—approaches which, however, are much more expensive at GenAI-scale.

Generation of Copyright-Protected Characters: The problem of the generation of copyright-protected characters' representations was described in 'RQ2.1 What are the Different Methods and Phases for Generating Content with GenAI Models that May have Copyright Implications?'. Several studies have been conducted to find effective mitigation strategies. Among those, *He et al. (2024)* proposed COPYCAT, a benchmark based on 50 copyright-protected characters, which has been specifically designed to assess the extent to which text-to-image AI systems are likely to generate their representations.

Building on the work of *He et al. (2024)*, *Chiba-Okabe & Su (2025)* introduced PREGen (Prompt Rewriting-Enhanced Genericisation). It is a model-agnostic approach that works by altering the input prompt and applying a systematic originality reduction mechanism.

PREGen mixes various practices already proven to be highly effective in previous literature (*He et al., 2024*), such as prompt cleaning—*i.e.*, it rewrite input instructions by removing all potentially unsafe content—and negative prompting—*i.e.*, explicitly discourage the model from including certain features in the output.

Specifically, for each generation task, PREGen first rewrites the input prompt using the described techniques. It then creates multiple slightly varied versions of this cleaned prompt, generates an image from each version, and evaluates all outputs for originality.

The 'originality evaluation' algorithm proposed by *Chiba-Okabe & Su (2025)* is based on pairwise similarity computation between all the images: the more similar an image is to the others, the less original it is considered. The final AI system's output is selected as the image that exhibits the lowest originality.

The efficacy of the approach was evaluated using the COPYCAT benchmark (*He et al., 2024*). The authors tested PREGen on three generative models: Playground v2.5, Pixart-α, and SDXL.

In the direct anchoring scenarios—*i.e.*, where prompts explicitly mentioned the names of copyright-protected characters—baseline models generated such characters in 56.0% to 82.6% of the cases. The application of standard prompt rewriting techniques reduced this rate to 2.6–13.4%, while PREGen further reduced it to a range between 0.6% and 6.6%.

In the indirect anchoring scenarios—*i.e.*, where prompts contained only implicit references to copyright-protected entities—baseline generation produced infringing content in 26.6% to 27.4% of the cases; this was reduced to 0–2.0% by using PREGen.

Moreover, although PREGen successfully avoided replicating protected characteristics, it was still able to generate images with general traits that aligned with the original intent of the prompt. This makes the solution an effective mitigation that does not disrupt the model's functionality.

Nonetheless, the study notes a computational trade-off: PREGen requires generating and evaluating multiple internal samples per input prompt, leading to increased resource consumption.

Generation Filters: Generative systems can be equipped with specific modules designed to identify and filter unwanted outputs. In addition to implementing a safety barrier, the same mechanism can also be used to mitigate memorisation (RQ1.2: What Are the Potential Copyright Compliance Issues—with Respect to EU Legislation—Related to the Different Methods and Data Input Phases?). *Ippolito et al. (2023)* developed Memfree. The system checks if any substring in the generated sequence matches the training set; if a match is found, it suppresses the output and triggers a new generation. While efficient, this approach is limited to *verbatim* matches and fails to detect generated content that is similar but not identical to the memorised data (*Ren et al., 2024*). *Chu, Song & Yang (2024)* proposed a copyright regression framework aimed at neutralising the possibility for a model to generate a given content, even if the latter is part of the training dataset. This is achieved through the modification of the objective function used during model training.

Even when models successfully avoid exact memorisation, they remain vulnerable to approximated memorisation, which can be triggered using adversarial 'style-transfer' prompts. To address this, *Kassem, Mahmoud & Saad (2023)* introduced the DeMemorization (DeMem) framework, a model alignment strategy specifically designed to reduce memorisation in LLMs. Unlike standard Reinforcement Learning (RL), DeMem employs negative similarity as a reward signal to discourage the model from reproducing training data (*Ren et al., 2024*).

AI-Generated Content (AIGC) Traceability—RO-SVD & Blockchain: *Ran et al. (2024)* proposed to record the lifecycle of AI-generated content (AIGC) on the blockchain, validating its origin by uniquely identifying the hardware system that produced it. The framework is called Ring Oscillator-Singular Value Decomposition (RO-SVD). It exploits the distinct oscillation patterns of Ring Oscillators (ROs) embedded in hardware systems, which exhibit slight variations due to manufacturing differences.

The researchers successfully isolated the component of the RO signal that remains unaffected by environmental noise, enabling the generation of a unique identifier for each hardware system. This identifier can be recorded on the blockchain along with the AIGC and the associated copyright metadata and rights management information.

This mechanism allows for the unique identification of the hardware that generated the content, rather than the GenAI model itself, which may have many instances. Such a

distinction provides a key advantage in copyright disputes by enabling verifiable attribution of outputs to a specific physical system.

According to *Ran et al. (2024)*, blockchain offers a valuable foundation for AIGC copyright management. It can provide a transparent and immutable record of content provenance, encompassing both the time of creation and associated copyright information, which can be encrypted and securely stored. Furthermore, smart contracts can be employed to automatically execute copyright-related transactions, enhancing transparency and efficiency while reducing the risk of copyright infringement and legal disputes.

Similarly, EKILA (*Balan et al., 2023*)—already described in this section as a framework incorporating visual attribution—leverages blockchain. Specifically, EKILA extends the concept of Non-Fungible Tokens (NFTs) by proposing a tokenised representation of rights. 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.

AI-Generated Content (AIGC) Traceability—C2PA: The Coalition for Content Provenance and Authenticity (C2PA) standard has attracted increasing attention in recent years. It allows to reliably 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. An important functionality is the possibility of tagging AI Generated Content (AIGC). Many software companies—such as Google (*Richardson, 2024*) and Microsoft (https://www.microsoft.com/en-us/research/project/project-origin/)—and camera producers—like Sony (*Goodman, 2024*)—have already adopted this standard. Moreover, it is designed to be format-agnostic, meaning that any binary asset is potentially capable of bearing a C2PA manifest.

The latter 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 and the constituent ingredient assets—*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. (2023)* in the EKILA framework[8], the list of ingredients can be leveraged to attribute the primary training samples from which an AIGC has been derived. According to this framework, information about the creators' identity and the relative digital wallet can be included inside the C2PA manifest, enabling automated crypto-currency-based royalty payments whenever their content is implicated in AI image generation.

C2PA manifests rely on digital signatures to guarantee integrity and authenticity. However, while the protocol envisions recovering stripped manifests through a distributed database, the supporting infrastructure is not yet in place. Consequently, assets remain at risk of losing provenance information if metadata is removed.

Watermarking: To protect source data copyright, researchers are moving from general AI-detection methods—which only identify if an image is AI-generated—to watermarking. This shift addresses the current inability of detection tools to pinpoint the specific origin of the content (*Ren et al., 2024*). Watermarking is a technique used to embed specific

[8] Described in 'RQ1.3: What are the technologies for input control, and how can they be mapped to the identified issues?'.

information in digital files, such as proof of ownership or AIGC tags. Researchers propose using a dedicated deep-learning network to embed watermarks into images immediately after they are generated. This system requires a corresponding external decoder, which is trained specifically to identify and extract the hidden information (*Ren et al., 2024*).

Although these methods are continually evolving to enhance their robustness against forgery and detection, attacks on such systems are also becoming increasingly sophisticated. The findings emphasise the urgent need for generative watermarking strategies that can withstand model-driven attacks while preserving their functionality and integrity (*Zhang et al., 2024*; *Ma et al., 2024*).

Watermarking can be applied to different content types, as detailed in the following paragraphs.

Image Watermarking: *Chen et al. (2024)* conducted a review to study the reliability of the watermarking techniques applied to images, finding that common image processing operations—such as resizing, cropping, flipping, JPEG compression, and frequency filtering—can significantly degrade or destroy the information encoded in the watermark. Moreover, advanced AI-based methods may use Generative Adversarial Networks (GANs) to reconstruct host images, effectively stripping hidden watermarks. On the other hand, the study concluded that GAN-based watermarking is especially robust, even against common image distortions and attacks, provided that the watermarking model has been exposed to the peculiarities of these threats during the training phase.

Cui et al. (2024) introduced FT-Shield to address scenarios where data protectors cannot oversee the model's fine-tuning process. Since traditional watermarking often requires extensive training cycles to take hold, FT-Shield uses subtle, imperceptible perturbations designed to be prioritised by the model. These 'easy-to-learn' features ensure the watermark is absorbed before the model focuses on the actual image content, making the protection much more efficient (*Ren et al., 2024*).

Text Watermarking: Watermarking techniques for LLMs often require partitioning the vocabulary into 'green' and 'red' lists of tokens. Then, the LLM's parameters are perturbed to favour green list selections. This bias allows a detector to identify AI-generated content by calculating if the proportion of green tokens exceeds a specific statistical threshold. While highly effective for direct detection, the signature can be weakened if the text is significantly altered through heavy paraphrasing (*Zhang et al., 2024*). Furthermore, *Zhang et al. (2024)* developed a novel attack that successfully reverse-engineers the 'green list' used in state-of-the-art LLM watermarking schemes. Extensive experiments on models like OPT and LLaMA show that this attack is effective even in extreme scenarios where the adversary has no prior knowledge of the model's parameters, lacks access to a watermark detector API, and has no information about the specific injection scheme. Once the green list is stolen, the watermark can be removed from the text by replacing these identified tokens with 'red list' synonyms, effectively stripping the statistical signature while maintaining text quality.

Besides watermarking applied to LLM-generated text, several techniques cover the case in which an already-existing text needs to be watermarked. These are often based on synonym or token substitution.

SemaMark, introduced by *Cui et al. (2024)*, uses the semantic meaning of a text as the foundation for its watermarking process. While specific words are easy to change through paraphrasing, the underlying message usually remains stable. This approach ensures that even if a malicious user tries to reword the content to strip the watermark, the core meaning stays intact, allowing the system to consistently identify and protect the content's origin (*Ren et al., 2024*).

A completely different approach is adopted in tools belonging to the Easymark family (*Sato et al., 2023*). For example, the Whitemark method—instead of altering how the model chooses words—swaps standard spaces (U+0020) for invisible Unicode alternatives like U+2004. Subsequent identification of the watermark is as simple as counting these specific hidden characters (*Ren et al., 2024*).

Other methods involve collecting AI-generated and human-generated text and then training a classifier to distinguish them. However, these models are often biased toward the specific datasets they were trained on, leading to poor generalisation (*Zhang et al., 2024*).

Software Code Watermarking: Software watermarking techniques can be static or dynamic. The first embeds identifiers into the program's structure, while the second integrates marks into a program's runtime behaviour. Notable examples include SWEET (*Lee et al., 2024*), and SrcMarker (*Yang et al., 2024*), which uses deep learning for subtle, semantics-preserving code modifications.

Wu et al. (2024) conducted a research to evaluate the robustness of code watermarking techniques. In particular, their study highlights the evolution from traditional methods—such as dead code insertion, opaque predicates, and code translation—towards advanced, learning-based schemes. To evaluate watermark robustness, the study explores LLM-based removal techniques using the CodeLlama-7b model. Two methods—rewrite-based and retranslate-based removal—demonstrate how powerful language models can undermine existing watermarking schemes (*Ma et al., 2024*).

A new frontier for protecting open-source code against unauthorised use in model training is code dataset watermarking. Some prototypes like CoProtector (*Sun et al., 2022*) and CodeMark (*Sun et al., 2023*)—which are both available open-source—use backdoor techniques[9] and structural transformations[10], respectively, to trace code usage in training neural code completion models. However, as noted by *Wu et al. (2024)*, these methods still suffer from vulnerabilities related to subsequent code manipulations.

Audio Watermarking: *Cao et al. (2023)* proposed a watermarking technique applied on audio diffusion models, emphasising two critical factors for effective protection: the choice of the watermark's seed and the imperceptibility of the mark. By carefully selecting these parameters, the method ensures that the watermark remains robust enough for tracking and authentication without degrading the acoustic quality of the generated audio.

THREATS TO THE VALIDITY OF THIS STUDY

To better identify the possible limitations of this study, we reference the quality concepts defined by *Kitchenham & Charters (2007)* described in Table 5.

[9] Backdoor refers to the intentional insertion of hidden patterns or triggers into training data such that, when a model learns from this data, the presence of the trigger activates a specific, recognisable behaviour at inference time.

[10] Structural transformations refer to systematic modifications of code—such as altering its syntax, control flow, or organisation—without changing its functional behaviour, in order to embed distinctive patterns that can later be used to trace whether the code was included in a model's training data.

Table 5 Quality concepts for assessing a study's validity (*Kitchenham & Charters, 2007*).

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.

Biases

The main challenge in an SLR is to ensure the full coverage of relevant literature.

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 (https://jpeg.org/jpegtrust/) and Liccium's Trust Engine (https://liccium.com)—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.

Threats to internal validity

Possible threats to internal validity were mitigated by adhering to established guidelines (*Kitchenham & Charters, 2007*; *Garousi, Felderer & Mäntylä, 2019*) and by engaging multiple authors in each phase of the review—including reading the articles—thereby ensuring that diverse perspectives were considered.

Threats to external validity

Threats to external validity arise from potential bias—discussed in 'Biases'—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 Fig. 5, 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 24 systematically analysed sources—fewer than those in the comparable secondary study in 'Related Secondary Studies'. This narrower result reflects the practice- and computer science-oriented scope of the review, in contrast to the legal focus of the comparative study (*Vig, 2026*). The use of legally charged keywords such as 'copyright' likely biased ranking algorithms toward law-centric sources, limiting visibility of computer science literature. At the same time, 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 systematic secondary study to examine computer science literature on copyright issues in Generative AI and the mitigation strategies proposed in technical research.

CONCLUSION

This SLR provided insights into practical issues related to copyright infringement and technical measures at both ends of the GenAI pipeline.

In particular, it underscores significant concerns regarding the opacity of training datasets. This lack of transparency has sparked legal scrutiny, especially in the press and publishing sectors and in online open-source code repositories, 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 standardised, 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 also reported on the literature related to membership inference attacks. Nevertheless, these are closely associated with the phenomenon of memorisation and are then controversial.

At the output stage, vulnerabilities persist: adversarial attacks can bypass safeguards, and models may replicate proprietary styles or generate infringing content, such as visual representations of copyright-protected characters. Additionally, we considered Retrieval-Augmented Generation (RAG) systems, concluding that, even if they enhance transparency by directly linking outputs to data sources, their misconfiguration may result in licence breaches or erroneous attributions.

The efforts to mitigate these issues include adapting authorship attribution techniques to large language models, the employment of provenance-tracking protocols—such as C2PA—underpinned by cryptographic infrastructures, and watermarking algorithms capable of embedding source metadata into content. Furthermore, mitigation strategies at the inference stage—such as prompt-rewriting—aim to prevent adversarial exploitation.

We concluded by acknowledging the limitations of the current review and proposing further investigation targeted at addressing topics not covered by the current work, such as the specific protections for audio or video content. The considerable heterogeneity inherent to the subject of this review has been identified as the main factor contributing to the omission of the aforementioned aspects.

These results highlight a critical need to improve the copyright safety of generative models. Research must now keep pace with rapid technical shifts to address the fragmented impacts on intellectual property. Managing this complexity requires closer cooperation between copyright holders and AI developers. While rights holders need a

clearer understanding of how these models operate, developers must fully grasp the legal risks of using protected data. Given the broad range of potential issues, both sides should adopt multiple layers of protection to ensure all risks are covered.

This literature review prioritised breadth over depth. This leaves room for future replications of the study to address further additional copyright issues and mitigations for particular content types.

APPENDICES

A. ONLINE MATERIAL

Details about the research process and the sources analysed when conducting the Systematic Literature Review can be found at: https://zenodo.org/records/19396529.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Aubert Antoine, Drobnic Ziga and Edelbroich Stephan for the constructive dialogue and insights shared during the development of this research.

AI Tools have been used to improve the clarity of the statements in this manuscript. Specifically, Google Gemini and ChatGPT have been prompted to rewrite some of the sentences in this manuscript and refine the formulations from a grammatical and syntaxical perspective. Then, the elaborated strings have been manually checked before being inserted into this manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Anna Arnaudo conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Riccardo Coppola conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Maurizio Morisio analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Antonio Vetrò conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Maurizio Borghi performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

- Riccardo Raso conceived and designed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Bryan Khan conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This is a literature review.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3928#supplemental-information.

REFERENCES

Autorite de la Concurrence. 2024. Decision 24-D-03 of March 15, 2024. Competition Authority. *Available at https://www.autoritedelaconcurrence.fr/fr/decision/relative-au-respect-des-engagements-figurant-dans-la-decision-de-lautorite-de-la-0* [In French].

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

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D, Donahue C, Doumbouya M, Durmus E, Ermon S, Etchemendy J, Ethayarajh K, Fei-Fei L, Finn C, Gale T, Gillespie L, Goel K, Goodman N, Grossman S, Guha N, Hashimoto T, Henderson P, Hewitt J, Ho DE, Hong J, Hsu K, Huang J, Icard T, Jain S, Jurafsky D, Kalluri P, Karamcheti S, Keeling G, Khani F, Khattab O, Koh PW, Krass M, Krishna R, Kuditipudi R, Kumar A, Ladhak F, Lee M, Lee T, Leskovec J, Levent I, Li XL, Li X, Ma T, Malik A, Manning CD, Mirchandani S, Mitchell E, Munyikwa Z, Nair S, Narayan A, Narayanan D, Newman B, Nie A, Niebles JC, Nilforoshan H, Nyarko J, Ogut G, Orr L, Papadimitriou I, Park JS, Piech C, Portelance E, Potts C, Raghunathan A, Reich R, Ren H, Rong F, Roohani Y, Ruiz C, Ryan J, Ré C, Sadigh D, Sagawa S, Santhanam K, Shih A, Srinivasan K, Tamkin A, Taori R, Thomas AW, Tramèr F, Wang RE, Wang W, Wu B, Wu J, Wu Y, Xie SM, Yasunaga M, You J, Zaharia M, Zhang M, Zhang T, Zhang X, Zhang Y, Zheng L, Zhou K, Liang P. 2022. On the opportunities and risks of foundation models. ArXiv DOI 10.48550/arXiv.2108.07258.

Bourtoule L, Chandrasekaran V, Choquette-Choo CA, Jia H, Travers A, Zhang B, Lie D, Papernot N. 2020. Machine unlearning. ArXiv DOI 10.48550/arXiv.1912.03817.

Buick A. 2025. Copyright and AI training data—transparency to the rescue? *Journal of Intellectual Property Law & Practice* **20(3)**:182–192 (visited on 03/28/2026) DOI 10.1093/jiplp/jpae102.

Cao X, Li X, Jadav D, Wu Y, Chen Z, Zeng C, Wei W. 2023. Invisible watermarking for audio generation diffusion models. In: *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. Piscataway: IEEE, 193–202 DOI 10.1109/TPS-ISA58951.2023.00033.

Carlini N, Hayes J, Nasr M, Jagielski M, Sehwag V, Tramèr F, Balle B, Ippolito D, Wallace E. 2023. Extracting training data from diffusion models. ArXiv DOI 10.48550/arXiv.2301.13188.

Carlini N, Ippolito D, Jagielski M, Lee K, Tramer F, Zhang C. 2023. Quantifying memorization across neural language models. ArXiv DOI 10.48550/arXiv.2202.07646.

Chen H, Liu C, Zhu T, Zhou W. 2024. When deep learning meets watermarking: a survey of application, attacks and defenses. *Computer Standards & Interfaces* **89**:103830 DOI 10.1016/j.csi.2023.103830.

Chiba-Okabe H, Su WJ. 2025. Tackling copyright issues in AI image generation through originality estimation and genericization. *Scientific Reports* **15(1)**:10621 DOI 10.1038/s41598-025-90827-1.

Chu T, Song Z, Yang C. 2024. How to protect copyright data in optimization of large language models? *Proceedings of the AAAI Conference on Artificial Intelligence* **38(16)**:17871–17879 DOI 10.1609/aaai.v38i16.29741.

Cooper A, Grimmelmann J. 2025. The files are in the computer: on copyright, memorization, and generative AI. *Chicago-Kent Law Review* **100(1)**:141. *Available at https://scholarship.kentlaw.iit.edu/cklawreview/vol100/iss1/9*.

Cui Y, Ren J, Lin Y, Xu H, He P, Xing Y, Lyu L, Fan W, Liu H, Tang J. 2024. FT-shield: a watermark against unauthorized fine-tuning in text-to-image diffusion models. ArXiv DOI 10.48550/arXiv.2310.02401.

Cursino R, Ferreira D, Lencastre M, Fagundes R, Pimentel J. 2018. Gamification in requirements engineering: a systematic review. In: *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, 119–125 DOI 10.1109/QUATIC.2018.00025.

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. Piscataway: IEEE, 248–255 DOI 10.1109/CVPR.2009.5206848.

Dou G, Liu Z, Lyu Q, Ding K, Wong E. 2025. Avoiding copyright infringement via large language model unlearning. ArXiv DOI 10.48550/arXiv.2406.10952.

Eldan R, Russinovich M. 2023. Who's Harry Potter? Approximate unlearning in LLMs. ArXiv DOI 10.48550/arXiv.2310.02238.

Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano AH, Chechik G, Cohen-Or D. 2022. An image is worth one word: personalizing text-to-image generation using textual inversion. ArXiv DOI 10.48550/arXiv.2208.01618.

Garousi V, Felderer M, Mäntylä MV. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* **106**:101–121 DOI 10.1016/j.infsof.2018.09.006.

Goodman M. 2024. Sony delivers highly anticipated firmware updates including C2PA compliancy and ensuring authenticity of images. *Available at https://www.sony.eu/presscentre/sony-delivers-highly-anticipated-firmware-updates-including-c2pa-compliancy-and-ensuring-authenticity-of-images*.

Google. 2023. An update on web publisher controls. Google. *Available at https://blog.google/technology/ai/an-update-on-web-publisher-controls/* (accessed 28 September 2023).

Guo C, Sablayrolles A, Jégou H, Kiela D. 2021. Gradient-based adversarial attacks against text transformers. ArXiv DOI 10.48550/arXiv.2104.13733.

He L, Huang Y, Shi W, Xie T, Liu H, Wang Y, Zettlemoyer L, Zhang C, Chen D, Henderson P. 2024. Fantastic copyrighted beasts and how (not) to generate them. ArXiv DOI 10.48550/arXiv.2406.14526.

Ippolito D, Tramèr F, Nasr M, Zhang C, Jagielski M, Lee K, Choquette-Choo CA, Carlini N. 2023. Preventing verbatim memorization in language models gives a false sense of privacy. ArXiv DOI 10.48550/arXiv.2210.17546.

Ji Z, Ma P, Wang S. 2022. Unlearnable examples: protecting open-source software from unauthorized neural code learning. In: *Proceedings of the 34th International Conference on Software Engineering and Knowledge Engineering (SEKE 2022)*. 525–530 DOI 10.18293/SEKE2022-066.

Kassem A, Mahmoud O, Saad S. 2023. Preserving privacy through dememorization: an unlearning technique for mitigating memorization risks in language models. In: Bouamor H, Pino J, Bali K, eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 4360–4379 DOI 10.18653/v1/2023.emnlp-main.265.

Kitchenham B, Charters S. 2007. Guidelines for performing systematic literature reviews in software engineering. *Available at https://docs.edtechhub.org/lib/EDAG684W*.

Knibbs K. 2024. Publishers target common crawl in fight over AI training data—WIRED. *Available at https://www.wired.com/story/the-fight-against-ai-comes-to-a-foundational-data-set/?utm_source=chatgpt.com*.

Kowala M. 2024. Protection of press publishers in the age of generative AI in search of legal remedies to adapt to the pace of technology. *IIC—International Review of Intellectual Property and Competition Law* **55(10)**:1604–1623 DOI 10.1007/s40319-024-01515-y.

Kumari N, Zhang B, Zhang R, Shechtman E, Zhu J-Y. 2023. Multi-concept customization of text-to-image diffusion. ArXiv DOI 10.48550/arXiv.2212.04488.

Le TV, Phung H, Nguyen TH, Dao Q, Tran N, Tran A. 2023. Anti-DreamBooth: protecting users from personalized text-to-image synthesis. ArXiv DOI 10.48550/arXiv.2303.15433.

Lee T, Hong S, Ahn J, Hong I, Lee H, Yun S, Shin J, Kim G. 2024. Who wrote this code? Watermarking for code generation. ArXiv DOI 10.48550/arXiv.2305.15060.

Lee K, Ippolito D, Nystrom A, Zhang C, Eck D, Callison-Burch C, Carlini N. 2022. Deduplicating training data makes language models better. ArXiv DOI 10.48550/arXiv.2107.06499.

Li R, Allal LB, Zi Y, Muennighoff N, Kocetkov D, Mou C, Marone M, Akiki C, Li J, Chim J, Liu Q, Zheltonozhskii E, Zhuo TY, Wang T, Dehaene O, Davaadorj M, Lamy-Poirier J, Monteiro J, Shliazhko O, Gontier N, Meade N, Zebaze A, Yee M-H, Umapathi LK, Zhu J, Lipkin B, Oblokulov M, Wang Z, Murthy R, Stillerman J, Patel SS, Abulkhanov D, Zocca M, Dey M, Zhang Z, Fahmy N, Bhattacharyya U, Yu W, Singh S, Luccioni S, Villegas P, Kunakov M, Zhdanov F, Romero M, Lee T, Timor N, Ding J, Schlesinger C, Schoelkopf H, Ebert J, Dao T, Mishra M, Gu A, Robinson J, Anderson CJ, Dolan-Gavitt B, Contractor D, Reddy S, Fried D, Bahdanau D, Jernite Y, Ferrandis CM, Hughes S, Wolf T, Guha A, von Werra L, de Vries H. 2023. StarCoder: may the source be with you! ArXiv DOI 10.48550/arXiv.2305.06161.

Li Z, Zhang Y. 2025. Advancing membership inference attacks: the present and the future. *Security and Safety* **4**:2024017 DOI 10.1051/sands/2024017.

Liesenfeld A, Dingemanse M. 2024. Rethinking open source generative AI: open-washing and the EU AI act. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and*

Transparency (FAccT '24). New York, NY, USA: Association for Computing Machinery, 1774–1787 DOI 10.1145/3630106.3659005.

Liu Y, Deng G, Xu Z, Li Y, Zheng Y, Zhang Y, Zhao L, Zhang T, Wang K, Liu Y. 2024. Jailbreaking ChatGPT via prompt engineering: an empirical study. ArXiv DOI 10.48550/arXiv.2305.13860.

Longpre S, Mahari R, Chen A, Obeng-Marnu N, Sileo D, Brannon W, Muennighoff N, Khazam N, Kabbara J, Perisetla K, Wu X, Shippole E, Bollacker K, Wu T, Villa L, Pentland S, Hooker S. 2023. The data provenance initiative: a large scale audit of dataset licensing & attribution in AI. ArXiv DOI 10.48550/arXiv.2310.16787.

Lu X, Wang J, Zhao Z, Dai Z, Foo C-S, Ng S-K, Low BKH. 2023. WASA: watermark-based source attribution for large language model-generated data. *Available at https://openreview.net/forum?id=FDfq0RRkuz*.

Lucchi N. 2024. ChatGPT: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation* **15(3)**:602–624 DOI 10.1017/err.2023.59.

Lucchi N. 2025. Generative AI and copyright—training, creation, regulation. *Available at https://www.europarl.europa.eu/thinktank/en/document/IUST_STU(2025)774095*.

Ludwig H, Zhou Y, Zawad S, Ong Y, Li P, Butler E, Zahid E. 2024. Towards collecting royalties for copyrighted data for generative models. In: *2024 IEEE International Conference on Web Services (ICWS)*. Piscataway: IEEE, 20–26 DOI 10.1109/ICWS62655.2024.00020.

Lukas A. 2023. Generative artificial intelligence under German copyright law (part 1). *SSRN Electronic Journal* **29**:564 DOI 10.2139/ssrn.4635645.

Ma W, Song Y, Xue M, Wen S, Xiang Y. 2024. The "Code" of ethics: a holistic audit of AI code generators. *IEEE Transactions on Dependable and Secure Computing* **21(5)**:4997–5013 DOI 10.1109/TDSC.2024.3367737.

Manteghi M. 2024. Can text and data mining exceptions and synthetic data training mitigate copyright-related concerns in generative AI? *Law, Innovation and Technology* **16(2)**:663–686 DOI 10.1080/17579961.2024.2392928.

Marone M, Durme BV. 2023. Data portraits: recording foundation model training data. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23*. Red Hook, NY, USA: Curran Associates Inc., 15121–15135.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. 2021. Learning transferable visual models from natural language supervision. ArXiv DOI 10.48550/arXiv.2103.00020.

Ran Z, Abdelgawad MAA, Zhang Z, Cheung RCC, Yan H. 2024. RO-SVD: a reconfigurable hardware copyright protection framework for AIGC applications. In: *2024 IEEE 35th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. Piscataway: IEEE, 135–142 DOI 10.1109/ASAP61560.2024.00037.

Ren J, Xu H, He P, Cui Y, Zeng S, Zhang J, Wen H, Ding J, Huang P, Lyu L, Liu H, Chang Y, Tang J. 2024. Copyright protection in generative AI: a technical perspective. ArXiv DOI 10.48550/arXiv.2402.02333.

Ribeiro R, Carvalho JP, Coheur L. 2024. Leveraging fuzzy fingerprints from large language models for authorship attribution. In: *2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Piscataway: IEEE, 1–7 DOI 10.1109/FUZZ-IEEE60900.2024.10612177.

Richardson L. 2024. How we're increasing transparency for gen AI content with the C2PA. Google. *Available at https://blog.google/technology/ai/google-gen-ai-content-transparency-c2pa/*.

Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K. 2023. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. ArXiv DOI 10.48550/arXiv.2208.12242.

Sakib M, Islam MA, Pathak R, Arifin M. 2024. Risks, causes, and mitigations of widespread deployments of large language models (LLMs): a survey. ArXiv DOI 10.48550/arXiv.2408.04643.

Sakurai W, Asano M, Imoto D, Honma M, Kurosawa K. 2023. Efficient authorship attribution method using ensemble models built by knowledge distillation. In: *2023 9th International Conference on Computer and Communications (ICCC)*, 2357–2362 DOI 10.1109/ICCC59590.2023.10507584.

Samuelson P. 2023. Legal challenges to generative AI, part I. *Communications of the ACM* **66(7)**:20–23 DOI 10.1145/3597151.

Sato R, Takezawa Y, Bao H, Niwa K, Yamada M. 2023. Embarrassingly simple text watermarks. ArXiv DOI 10.48550/arXiv.2310.08920.

Schneider J. 2024. Explainable generative AI (GenXAI): a survey, conceptualization, and research agenda. *Artificial Intelligence Review* **57(11)**:1–38 DOI 10.1007/s10462-024-10916-x.

Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson K, Schmidt L, Kaczmarczyk R, Jitsev J. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. ArXiv DOI 10.48550/arXiv.2210.08402.

Shaham TR, Dekel T, Michaeli T. 2019. SinGAN: learning a generative model from a single natural image. ArXiv DOI 10.48550/arXiv.1905.01164.

Shan S, Cryan J, Wenger E, Zheng H, Hanocka R, Zhao BY. 2025. Glaze: protecting artists from style mimicry by text-to-image models. ArXiv DOI 10.48550/arXiv.2302.04222.

Shan S, Ding W, Passananti J, Wu S, Zheng H, Zhao BY. 2024. Nightshade: prompt-specific poisoning attacks on text-to-image generative models. In: *2024 IEEE Symposium on Security and Privacy (SP)*. Piscataway: IEEE, 807–825 DOI 10.1109/SP54263.2024.00207.

Shapley L. 1953. 7. A value for n-person games. Contributions to the Theory of Games II (1953) 307–317. In: Kuhn HW, ed. *Classics in Game Theory*. Princeton: Princeton University Press, 69–79 DOI 10.1515/9781400829156-012.

Shin T, Razeghi Y, Logan RL IV, Wallace E, Singh S. 2020. AutoPrompt: eliciting knowledge from language models with automatically generated prompts. ArXiv DOI 10.48550/arXiv.2010.15980.

Somepalli G, Singla V, Goldblum M, Geiping J, Goldstein T. 2022. Diffusion art or digital forgery? Investigating data replication in diffusion models. ArXiv DOI 10.48550/arXiv.2212.03860.

Sun Z, Du X, Song F, Li L. 2023. 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*, 1561–1572 DOI 10.1145/3611643.3616297.

Sun Z, Du X, Song F, Ni M, Li L. 2022. CoProtector: protect open-source code against unauthorized training usage with data poisoning. ArXiv DOI 10.48550/arXiv.2110.12925.

Vig S. 2026. Intersection of generative artificial intelligence and copyright: an Indian perspective. *Journal of Science and Technology Policy Management* **17(1)**:118–133 DOI 10.1108/JSTPM-08-2023-0145.

Wang JT, Deng Z, Chiba-Okabe H, Barak B, Su WJ. 2024. An economic solution to copyright challenges of generative AI. ArXiv DOI 10.48550/arXiv.2404.13964.

Widder DG, Whittaker M, West SM. 2024. Why 'open' AI systems are actually closed, and why this matters. *Nature* **635(8040)**:827–833 DOI 10.1038/s41586-024-08141-1.

Wu BL, Chen K, He Y, Chen G, Zhang W, Yu N. 2024. CodeWMBench: an automated benchmark for code watermarking evaluation. In: *Proceedings of the ACM Turing Award Celebration Conference—China 2024. ACM-TURC '24*. New York, NY, USA: Association for Computing Machinery, 120–125 DOI 10.1145/3674399.3674447.

Yan W, Li Y. 2022. WhyGen: explaining ML-powered code generation by referring to training examples. ArXiv DOI 10.48550/arXiv.2204.07940.

Yang B, Li W, Xiang L, Li B. 2024. SrcMarker: dual-channel source code watermarking via scalable code transformations. In: *2024 IEEE Symposium on Security and Privacy (SP)*. Piscataway: IEEE, 4088–4106 DOI 10.1109/SP54263.2024.00097.

Ye X, Huang H, An J, Wang Y. 2023. DUAW: data-free universal adversarial watermark against stable diffusion customization. ArXiv DOI 10.48550/arXiv.2308.09889.

Zhang D. 2023. Should ChatGPT and bard share revenue with their data providers? A new business model for the AI era. ArXiv DOI 10.48550/arXiv.2305.02555.

Zhang D, Finckenberg-Broman P, Hoang T, Pan S, Xing Z, Staples M, Xu X. 2024. Right to be forgotten in the era of large language models: implications, challenges, and solutions. ArXiv DOI 10.48550/arXiv.2307.03941.

Zhang R, Isola P, Efros AA, Shechtman E, Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. ArXiv DOI 10.48550/arXiv.1801.03924.

Zhang Y, Jia R, Pei H, Wang W, Li B, Song D. 2020. The secret revealer: generative model-inversion attacks against deep neural networks. ArXiv DOI 10.48550/arXiv.1911.07135.

Zhang Z, Zhang X, Zhang Y, Zhang LY, Chen C, Hu S, Gill A, Pan S. 2024. Stealing watermarks of large language models via mixed integer programming. In: *2024 Annual Computer Security Applications Conference (ACSAC)*, 46–60 DOI 10.1109/ACSAC63791.2024.00021.

Zhao J, Chen K, Yuan X, Qi Y, Zhang W, Yu N. 2024a. Silent guardian: protecting text from malicious exploitation by large language models. In: *IEEE Transactions on Information Forensics and Security*. Vol. 19. Piscataway: IEEE, 8600–8615 DOI 10.1109/TIFS.2024.3455775.

Zhao Z, Duan J, Hu X, Xu K, Wang C, Zhang R, Du Z, Guo Q, Chen Y. 2024b. Unlearnable examples for diffusion models: protect data from unauthorized exploitation. ArXiv DOI 10.48550/arXiv.2306.01902.

Zhao Z, Duan J, Xu K, Wang C, Zhang R, Du Z, Guo Q, Hu X. 2024c. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE, 24398–24407 DOI 10.1109/CVPR52733.2024.02303.

Zou C, Feng Z, Guan H, Xing C. 2024. Chain-of-thought enhanced content detection in large language models. In: *2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, 610–617 DOI 10.1109/CBASE64041.2024.10824407.