Dataset Copyright Auditing for Large Models: Fundamentals, Open Problems, and Future Directions

DU Linkang, SU Zhou, YU Xinyi

(Xi'an Jiaotong University, Xi'an 710049, China)

DOI: 10.12142/ZTECOM.202503005

https://kns.cnki.net/kcms/detail/34.1294.TN.20250908.1530.002.html, published online September 8, 2025

Manuscript received: 2025-06-20

Abstract: The unprecedented scale of large models, such as large language models (LLMs) and text-to-image diffusion models, has raised critical concerns about the unauthorized use of copyrighted data during model training. These concerns have spurred a growing demand for dataset copyright auditing techniques, which aim to detect and verify potential infringements in the training data of commercial AI systems. This paper presents a survey of existing auditing solutions, categorizing them across key dimensions: data modality, model training stage, data overlap scenarios, and model access levels. We highlight major trends, including the prevalence of black-box auditing methods and the emphasis on fine-tuning rather than pre-training. Through an in-depth analysis of 12 representative works, we extract four key observations that reveal the limitations of current methods. Furthermore, we identify three open challenges and propose future directions for robust, multimodal, and scalable auditing solutions. Our findings underscore the urgent need to establish standardized benchmarks and develop auditing frameworks that are resilient to low watermark densities and applicable in diverse deployment settings.

Keywords: dataset copyright auditing; large language models; diffusion models; multimodal auditing; membership inference

Citation (Format 1): DU L K, SU Z, YU X Y, et al. Dataset copyright auditing for large models: fundamentals, open problems, and future directions [J]. ZTE Communications, 2025, 23(3): 38 - 47. DOI: 10.12142/ZTECOM.202503005

Citation (Format 2): L. K. Du, Z. Su, X. Y. Yu, et al., "Dataset copyright auditing for large models: fundamentals, open problems, and future directions," *ZTE Communications*, vol. 23, no. 3, pp. 38 - 47, Sept. 2025. doi: 10.12142/ZTECOM.202503005.

1 Introduction

With the rapid advancement of computational power and optimization techniques, deep neural networks (DNNs) with billions or even trillions of parameters, commonly referred to as large models, have become the cornerstone of modern artificial intelligence[1 - 5]. These models are now widely deployed in real-world applications, ranging from text generation and code completion to image synthesis and virtual assistants[6 - 7]. For example, OpenAI's ChatGPT had reportedly reached 500 million weekly active users and 3 million business users by the end of March 2025[8]. To achieve impressive performance, such models rely on massive datasets during pre-training and fine-tuning stages. According to *The Decoder*, ChatGPT-4 was trained on approximately 13 trillion tokens, sourced from a diverse mix of web-scale corpora, including CommonCrawl, Reddit, books, code repositories, and potentially proprietary sources such as educational textbooks[9].

This appetite for data has intensified concerns around the predatory development of training corpora. Public data, often protected by licenses such as the Creative Commons or GPL, are frequently scraped and used at scale, without adequate consent or adherence to usage terms. This practice has led to widespread breaches of data licensing agreements and raised substantial legal, ethical, and economic issues, particularly in domains like publishing, software, and the creative arts. For instance, Getty Images sued Stability AI for allegedly using over 12 million copyrighted and watermarked images without authorization to train its diffusion models[10]. Similarly, Thomson Reuters prevailed in a landmark case against Ross Intelligence, where a US court ruled that using copyrighted legal annotations to train an AI assistant constituted infringement, rejecting claims of fair use[11 - 12]. In the creative domain, authors and artists have brought lawsuits against companies like Meta and OpenAI for training large language models on books and artworks obtained from unauthorized sources, such as pirated eBook repositories[13 - 14]. These cases underscore a growing consensus that large-scale data scraping for AI training, espe-

This work was supported in part by NSFC under Grant Nos. 62402379, U22A2029 and U24A20237.

The corresponding author is SU Zhou.

cially without licensing or compensation, poses serious challenges to existing copyright frameworks and demands clearer regulatory boundaries for responsible AI development.

To address these concerns, researchers have proposed various techniques for dataset copyright auditing. Based on whether the modification of the raw training data is needed, the existing solutions for dataset copyright auditing can be classified into two types, i.e., intrusive auditing[15 - 17] and non-intrusive auditing[18 - 20]. However, these techniques have largely been developed for traditional machine learning models, often for classification tasks and in the image domain, where models are relatively small and datasets are curated manually[21]. The paradigm shift to large models (e. g., large language models and diffusion models) brings unique challenges: Training data is often massive, opaque, and noisy; model behaviors are emergent and stochastic; and auditing is constrained to black-box settings due to proprietary deployment. Consequently, there is an urgent need to reevaluate and redesign dataset copyright auditing techniques in the context of large-scale generative and multimodal models.

Existing surveys have laid valuable groundwork. For instance, HARTMANN et al.[22] introduced a taxonomy of memorization in large language models (LLMs), including verbatim content, factual knowledge, writing styles, and alignment behavior, and examined its implications for privacy, security, and copyright. While memorization is a prerequisite for copyright infringement, the root issue often lies in the unauthorized use of protected datasets during training. Thus, their work is orthogonal to ours. More recently, DU et al.[23] conducted a systematic review of copyright protection techniques and evaluated the existing auditing solutions on classification models in the image domain. However, our focus shifts to dataset auditing for LLMs and diffusion models. Furthermore, given the substantially larger training data scales involved in these models compared with traditional classification models, we evaluate the effectiveness of existing auditing methods under varying injection rates of modified data, with particular emphasis on scenarios involving low injection rates.

In this paper, we provide the first survey of dataset copyright auditing methods specifically for large models. We systematically review and categorize existing techniques, analyze their applicability to large-scale model training pipelines, and identify critical limitations and future challenges. In summary, our contributions are threefold:

• We systematize existing dataset copyright auditing techniques in the context of large models, organizing them across key dimensions including the type of auditing strategy, the specific technique used, the domain of the data, the stage of the model training, the data overlaps, and the model access level.

• We summarize four observations based on the surveyed papers and find that there is a pressing need for more auditing techniques that can handle more comprehensive data types, such as audio and video. In addition, we emphasize that the practical auditing method should be robust across various levels of overlap, especially under partial or sparse inclusion settings.

• We conclude three open problems and corresponding future directions to guide the development of scalable, reliable, and legally sound dataset auditing mechanisms for the governance of large models.

2 Preliminaries

This section introduces the essential definitions and components that are crucial to understanding the context of dataset copyright auditing for large models.

2.1 LLMs

LLMs are deep neural networks designed to process and generate human language. Typically based on the Transformer architecture, these models are trained on vast numbers of textual data using unsupervised learning. The most common objective for training LLMs is language modeling, where the model learns to predict the next word in a sequence given its previous words.

Mathematically, given a sequence of tokens $x = (x_1, x_2, \cdots, x_n)$, the goal is to maximize the probability of predicting the next token x_{i+1} based on the preceding tokens:

$$P(x_{i+1} | x_1, x_2, \cdots, x_i) = \frac{P(x_1, x_2, \cdots, x_{i+1})}{P(x_1, x_2, \cdots, x_i)} \quad (1).$$

The model is trained to optimize the likelihood function over large datasets by minimizing the cross-entropy loss:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log P(x_i | x_1, \cdots, x_{i-1}; \theta) \quad (2),$$

where θ represents the parameters of the model, and N is the total number of tokens in the dataset. LLMs are typically pretrained using massive web-scale corpora (e.g., CommonCrawl and Wikipedia) and fine-tuned for specific tasks (e. g., text generation and summarization).

LLMs can be considered generative models, as they generate plausible text sequences. These models have demonstrated emergent capabilities, including in-context learning, zero-shot classification, and even reasoning, depending on the scale of training and model architecture.

2.2 Diffusion Models

Diffusion models are a class of generative models that learn to create data (e.g., images and audio) by simulating a physical diffusion process, which gradually adds noise to the data until it becomes pure noise. The model then learns the reverse process to transform random noise back into structured data.

Formally, a diffusion model defines a forward noising process that corrupts an image x_0 into a sequence of noisy images x_t over T timesteps, according to a Markov process:

$$q\left(x_t \middle| x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{\alpha_t}\, x_{t-1}, \sigma_t^2 I\right) \tag{3},$$

where α_t controls the variance schedule, and σ_t^2 represents the noise variance at time t. The forward process progressively adds Gaussian noise $\mathcal{N}$ to the image x_0 until it is destroyed by the final timestep T.

The reverse process is learned by the model, which tries to denoise the noisy samples step by step:

$$p_\theta\left(x_{t-1} \middle| x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_t\right) \tag{4},$$

where $\mu_\theta\left(x_t, t\right)$ is the predicted mean of the reverse process and Σ_t is the variance. The model is trained to minimize the denoising score matching loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{q\left(x_0, \cdots, x_T\right)}\left[\left\| z_t - z_\theta \right\|^2\right] \tag{5},$$

where z_t is the noise predicted by the model at each timestep t, and z_θ is the true noise. Popular diffusion models like Stable Diffusion and DALL·E leverage this framework to generate high-fidelity images from text descriptions, where the text serves as a conditioning signal.

2.3 Backdoor Attacks

Backdoor attacks (BA) are a type of data poisoning attacks where an adversary intentionally injects a small subset of poisoned samples into the training set. The poisoned data includes a trigger (a specific pattern or input feature) that induces the model to behave maliciously when the trigger is present during inference. The idea of BAs is usually adopted in intrusive auditing strategies, which embed a hidden signal into training data, making it detectable if unauthorized models exhibit specific responses to trigger patterns.

Formally, let D_{clean} be the original clean dataset and D_{poisoned} the crafted dataset. The goal of a BA is to train the model f_θ such that it correctly classifies the normal data from D_{clean}, but when given a poisoned input x_{trigger} (with the trigger t applied), the model outputs a predefined class x_{target}:

$$f_\theta\left(x_{\text{trigger}}\right) = y_{\text{target}} \text{ when } x_{\text{trigger}} = x + t \tag{6}.$$

The model is trained to minimize the loss of clean data but also ensures that for poisoned data, and the output is as desired (the target class). The loss function is typically augmented to include a trigger-specific objective that steers the model's behavior for the poisoned data:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \lambda \mathcal{L}_{\text{trigger}} \tag{7},$$

where λ is a weighting factor that controls the strength of the trigger influence during training.

2.4 Membership Inference

Membership inference (MI) aims to determine whether a particular data point was used in the training set of a machine learning model. Therefore, MI can inspire the design of non-intrusive methods, which are used to detect whether a model has been trained on a dataset that includes protected content.

Given a model f_θ and a query input x, the MI task is to predict whether x is a member of the training set D_{train}. A MI attack can be formalized as:

$$\hat{y} = \text{MI}(x) = \begin{cases} 1 & \text{if } x \in D_{\text{train}} \\ 0 & \text{if } x \notin D_{\text{train}} \end{cases} \tag{8}.$$

This is typically done by observing the model's confidence levels or output probabilities. If the model outputs high confidence for a given sample, this may indicate that the sample was used during training. A key approach involves using the perplexity of LLMs or the log-likelihood of a token sequence to measure how likely the model is to have used the data.

$$\text{Perplexity}(x) = \exp\left(-\frac{1}{n} \sum_{i=1}^{n} \log P\left(x_i \middle| x_1, \cdots, x_{i-1}; \theta\right)\right) \tag{9},$$

where x_i represents tokens in the sequence, and n is the number of tokens. A low perplexity suggests that the input is likely part of the model's training data.

3 Dataset Copyright Auditing

In this section, we provide the definition of the dataset copyright auditing problem along with a summary of the existing solutions.

3.1 Problem Definition

Considering the auditing scenarios in practice, we first introduce the key roles in the dataset copyright auditing. Then, we explain the different auditing settings based on three pillars: the stages of use, the data overlaps, and the model access levels.

1) Key roles: the data owner, the model trainer, and the auditor.

• Data owner ($\mathcal{P}_{\text{owner}}$): This is the entity that generates and holds the copyright to a dataset D. The data owner may distribute or sell the dataset under specific licensing agreements.

• Model trainer ($\mathcal{P}_{\text{trainer}}$): This entity acquires datasets either from publicly available online sources or through purchases from authorized markets. Using this data, the trainer builds and optimizes a deep neural network f_θ, where θ denotes the model parameters, typically via loss minimization. The resulting model can be deployed as part of a Machine Learning as a Service (MLaaS) platform to generate commercial revenue.

• Auditor ($\mathcal{P}_{\text{auditor}}$): A neutral third party appointed by the data owner $\mathcal{P}_{\text{owner}}$ to investigate potential unauthorized usage

of dataset D in a suspicious model. If misuse is confirmed, the auditor must provide concrete evidence of copyright infringement. Recent research has enhanced the auditor's capabilities. For example, DONG et al.[24] proposed incorporating an identity registration mechanism to prevent dataset abuse via malicious registration.

2) Stages of use: There are two primary stages in which a dataset can be integrated into the construction of large models.

• Pre-training: The model trainer designs the architecture and optimization strategy of a large model and trains it from scratch using extensive datasets.

• Fine-tuning: As DNNs grow and become more complex, training them from scratch becomes increasingly resource-intensive. Consequently, many model trainers opt to download pre-trained weights and fine-tune the model on task-specific datasets to adapt it for downstream applications.

3) Data overlaps (Fig. 1): The auditing process typically encounters one of five possible scenarios regarding dataset overlaps between the data owner and the model trainer.

• Disjoint (Case 1): The data owner's dataset does not intersect with the model's training dataset $\mathcal{D}_t$, i.e., $\mathcal{D}_a \cap \mathcal{D}_t = \varnothing$.

• Partially overlap (Case 2): The dataset of the data owner partially overlaps with the model's training dataset, that is $\mathcal{D}_a \cap \mathcal{D}_t \neq \varnothing$ and $\mathcal{D}_a \nsubseteq \mathcal{D}_t$.

• The data owner fully covers the model trainer (Case 3): The model's training dataset $\mathcal{D}_t$ is a subset of the data owner, i.e., $\mathcal{D}_t \subseteq \mathcal{D}_a$.

• The model trainer fully covers the data owner (Case 4): The data owner's dataset is a subset of the model's training dataset, represented by $\mathcal{D}_a \subseteq \mathcal{D}_t$.

• Completely overlap (Case 5): The data owner's dataset is the same as the model trainer's training dataset, implying $\mathcal{D}_a = \mathcal{D}_t$.

4) Model access levels: Auditors encounter various levels of access to the suspicious model during auditing.

• Black-box access: The auditor can only query the model with inputs x and observe the corresponding outputs $f_\theta(x)$, without any internal model details.

• Gray-box access: The auditor has partial internal knowledge, such as the model architecture $\mathcal{M}$, alongside inputs x and outputs $f_\theta(x)$.

• White-box access: The auditor has full transparency, including access to model parameters θ, internal training details (e.g., hyperparameters and preprocessing techniques), and all related internal structures.

5) Examples: We present two examples highlighting practical implications in dataset copyright auditing.

• Literary dataset auditing scenario: Consider a scenario in which an author identifies that their publicly shared, yet copyrighted, literary manuscripts have possibly been utilized to train an LLM without consent. The author compiles a small set of specific textual sequences believed to be used by the suspicious LLM. An auditor performs dataset copyright auditing on the suspicious LLM to validate infringement.

• Artwork style piracy scenario: In another scenario, adversaries fine-tune a diffusion model using a small set of online-available artworks by an artist. This enables the model to generate pieces that closely replicate the artworks. Upon discovering artworks resembling their own produced by the suspicious model, the artist suspects unauthorized fine-tuning on their dataset and engages an auditor to check for data infringements.

3.2 Existing Solutions

In this section, we survey recent advances in dataset copyright auditing and categorize the existing works across six key dimensions. Following Ref. [23], we first classify the auditing strategies into intrusive auditing and non-auditing based on whether the auditor needs to modify the original data during the whole auditing process.

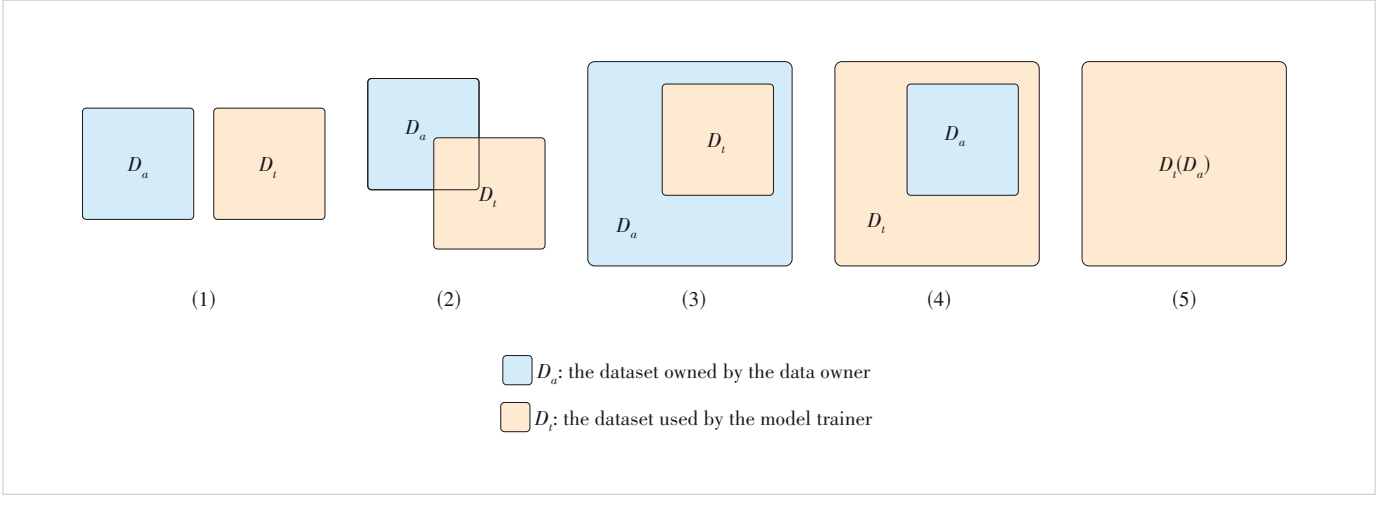

Figure 1. An illustration of data overlaps

3.2.1 Intrusive Auditing

Intrusive auditing techniques embed traceable, imperceptible markers into datasets or models during training to verify data provenance or assert model ownership. These techniques fall under the broader concept of data marking, where artificial signals are intentionally injected during the training process to enable post-hoc verification. In contrast to non-intrusive auditing techniques that infer data usage from model behavior, intrusive auditing techniques actively modify the training data or pipeline. Depending on their embedding strategies, intrusive auditing techniques can be categorized into two types: BA and feature-based watermarks (FW).

In BA methods, researchers insert samples with embedded triggers into the training data so that the model exhibits predefined behaviors when encountering specific inputs during inference. For example, CHEN et al.[25] embed a small number of images with backdoor triggers into the training set, causing the model to misclassify these samples during inference. Similarly, WANG et al.[26] and REN et al.[27] introduce stealthy trigger samples into text-to-image models to detect whether fine-tuning involves a specific dataset. LI et al.[28] further combine personalized triggers to verify the use of authorized data during model fine-tuning.

FW methods avoid explicit triggers and instead modify the model's training objective or representation space to embed watermarks implicitly into the model's features. For instance, HUANG et al.[29] introduce gradient constraints during training to distribute watermark signals within the model parameters, thereby improving the robustness and stealthiness. CUI et al.[30] propose embedding watermarks into the feature space using implicit signals for fine-tuning detection, leveraging shadow models to facilitate watermark learning. Finally, HUANG et al.[31] present a hybrid strategy combining active perturbation and MI, enabling fine-grained auditing of data usage at the image level.

While both BA and FW methods aim to embed verifiable signals into the model, they exhibit distinct trade-offs. BA methods offer strong detectability and relatively low embedding complexity but rely on explicit triggers, making them more vulnerable to trigger removal or data sanitization. In contrast, FW methods enhance robustness and stealth by operating in the representation space or gradient domain, but often incur a higher computational cost and require careful optimization to maintain model performance. These differences highlight a fundamental tension between ease of implementation and resilience against adversarial modifications, which remains an open challenge for intrusive auditing techniques.

3.2.2 Non-Intrusive Auditing

Non-intrusive auditing techniques generate unique identifiers for data or models to trace data provenance, verify legality, or detect potential misuse. The core idea is to leverage the statistical characteristics of the data, model parameter distributions, or training behavior to embed or extract verifiable identifiers without significantly affecting the original performance. Based on detection granularity, non-intrusive auditing techniques can be categorized into MI and dataset inference (DI).

SHI et al.[32] propose an MI approach based on output distribution analysis, detecting anomalies in low-probability tokens to infer whether a text segment is present during pre-training, under the assumption that unseen text exhibits higher uncertainty. This method demonstrates strong performance on LLMs and operates entirely under a black-box setting.

DI determines whether an entire dataset was used during pre-training or fine-tuning. Unlike MI, DI methods typically aggregate multiple statistical signals or leverage distributional properties for more robust inference. MAINI et al.[33] introduce a likelihood ratio-based statistical test that compares the log-likelihood of the target dataset against a reference dataset, followed by hypothesis testing to infer dataset involvement. This approach is particularly suited for auditing large-scale language model pre-training. MA et al.[34] address code generation models by combining likelihood ratio analysis with code-style fingerprinting, leveraging perplexity differences to assess whether code snippets originate from the training set, while incorporating both syntactic and statistical characteristics. DU et al.[35] conceptualize an entire collection of artworks as a unique style fingerprint, extracting multi-granularity style features using CNNs and training a regressor to measure discrepancies between this fingerprint and generated or public images, enabling detection of whether a model has learned specific artistic styles.

In non-intrusive auditing, MI operates at a fine-grained level, aiming to identify whether individual samples were used during training. However, it often faces issues of limited robustness and high false positive rates in large-scale models. In contrast, DI utilizes aggregated statistical characteristics to determine dataset-level inclusion relationships. This achieves higher stability and scalability, but at the cost of requiring more data and computational resources. This trade-off between audit granularity and robustness represents a key design consideration for non-intrusive auditing strategies.

3.2.3 Main Observations

In Table 1, the "Domain" column specifies the modality of the audited dataset, and the "Type" column indicates whether the auditing strategy belongs to intrusive or non-intrusive. The "Technique" column shows the techniques adopted by the auditing strategy. The "Stages of Use" column indicates whether the dataset was used during the pre-training or fine-tuning phase. The "Data Overlaps" column describes the relationship between the data owner's dataset and the model trainer's dataset. The "Model Access Level" refers to the degree of access the auditor has to the model. Finally, the "Used Model" and "Used Dataset" columns list the models and datasets employed in each paper's evaluation. Based on our analysis of Table 1,

Table 1. A summary of existing solutions for dataset copyright auditing in the context of large models, with papers organized by audit domain and type

Reference	Domain	Type	Technique	Stage of Use	Data Overlap	Model Access Level	Used Model	Used Dataset
CHEN et al.[25]	Image	Intrusive	BA	Pre-training	Case 4	Black-box	DeepID, VGG-Face	YouTube Aligned Face Dataset
HUANG et al.[29]			FW	Pre-training	Case 2	Black-box	SimCLR	CIFAR-10, CIFAR-100, and TinyImageNet
HUANG et al.[31]				Pre-training	Case 4	Black-box	ResNet-18, ResNet-34, WideResNet-28-2, VGG-16, ConvNetBN, and SimCLR	CIFAR-100 and TinyImageNet
HUANG et al.[29]	Text	Intrusive	FW	Pre-training	Case 2	Black-box	LLaMA 2	SST2, AG's news, and TweetEval (emoji)
SHI et al.[32]		Non-intrusive	MI	Pre-training	Case 2	Black-box	LLaMA (7 B, 13 B, 30 B, 65 B), GPT-NeoX-20B, OPT-66 B, Pythia-2.8 B, GPT-3 (text-davinci-003), and LLaMA2-7 B-WhoIsHarryPotter	WIKIMIA, Books3 (copyrighted books), RedPajama + downstream tasks (BoolQ, IMDB, TruthfulQA, CommonsenseQA), and Harry Potter series
MAINI et al.[33]			DI	Pre-training	Case 2	Gray-box	Pythia (410 M, 1.4 B, 6.9 B, and 12 B)	PILE (20 subsets including Wiki, Arxiv, OpenWebText, etc.)
MA et al.[34]				Fine-tuning	Case 4	Black-box	CodeGen, GPT-Neo, CodeGPT, InCoder, PolyCoder, and CodeT5	APPS, PY150, MBPP, and MBXP (multi-language versions)
WANG et al.[26]	Text-image	Intrusive	BA	Fine-tuning	Case 4, Case 5	Black-box	Stable Diffusion v2.1	WikiArt and COCO
REN et al.[27]				Fine-tuning	Case 3, Case 5	Black-box	Stable Diffusion v1.4 and Stable Diffusion v2	CC-20k, Sketchyscene, and Cartoon-BLIP-Caption
LI et al.[28]				Fine-tuning	Case 4	Black-box	Stable Diffusion v1.5, Stable Diffusion v2.1, and Kandinsky 2.2	CelebA-HQ, ArtBench, Landscape, MS-COCO, and Pokémon BLIP captions dataset
HUANG et al.[29]			FW	Pre-training	Case 2	Black-box	CLIP	Flickr30k
CUI et al.[30]				Fine-tuning	Case 3, Case 5	Black-box	Stable Diffusion	WikiArt, Pokémon BLIP captions dataset, and CelebA
HUANG et al.[31]				Pre-training	Case 4	Black-box	CLIP	Flickr30k
DU et al.[35]		Non-intrusive	DI	Fine-tuning	Case 1, Case 2, Case 5	Black-box	Diffusion v2.1, Stable Diffusion XL, and Kandinsky	WikiArt and Artist-30

BA: backdoor attack DI: dataset inference FW: feature-based watermark MI: membership inference

we identify four main observations, focusing on the data domain, the training stage, the extent of dataset overlap, and the model access level.

1) Observation 1: The existing auditing methods span a variety of data domains, including texts, images, and text-to-image (multimodal) modalities. Among them, text-to-image models, particularly diffusion-based systems like Stable Diffusion, receive the most attention due to the growing concern over style mimicry and unauthorized use of visual artworks conditioned on textual prompts. Auditing techniques targeting pure text domains typically focus on LLMs trained on datasets such as Books3, Wikipedia, or RedPajama. While these models raise significant copyright concerns, relatively few works address image-only domains, especially in the pre-training stage. Furthermore, although some studies evaluate code datasets as a text subcategory, the structural and legal uniqueness of code suggests it should be treated as a distinct domain. This distribution indicates a pressing need for more auditing techniques that can satisfy the various auditing requirements in real-world applications.

2) Observation 2: In terms of model training stages, most existing methods focus on the fine-tuning process rather than the pre-training phase. This is primarily because fine-tuning often involves smaller, proprietary datasets, such as specific author manuscripts or artist portfolios, which are more easily traceable. These cases align well with real-world scenarios where pre-trained foundation models are adapted to downstream applications, making them an attractive target for auditing. In contrast, only a few methods address auditing at the pre-training stage, which presents a more complex challenge due

to the vast and heterogeneous nature of the training data. Nonetheless, since unauthorized use of copyrighted material is likely to occur during both pre-training and fine-tuning, there is a clear demand for methods capable of handling both phases effectively.

3) Observation 3: The existing works collectively cover the full spectrum of dataset overlap scenarios between the data owner and the model trainer. Many approaches assume that the data owner's content either is completely included in or significantly overlaps with the model's training data, which simplifies the auditing task. However, more comprehensive frameworks, such as those proposed by DU et al.[35], evaluate disjoint, partially overlapping, and completely overlapping conditions, thereby better reflecting the complexities of real-world deployments. Only a limited number of methods, such as CUI et al.[30] explore the case where the model trainer's entire training set is a subset of the data owner's corpus. This observation points to the importance of developing auditing methods that are robust across various levels of overlap, especially under partial or sparse inclusion settings.

4) Observation 4: Most existing solutions are developed under black-box access constraints, where auditors can only query models and observe their outputs without access to internal parameters or training configurations. This is consistent with real-world scenarios where large models, such as commercial LLMs or image generators, are often accessed via closed application programming interfaces (APIs). While this setting reflects deployment reality, it also imposes significant limitations on viable auditing strategies. A few studies, such as the work by MAINI et al.[33], adopt a gray-box approach, assuming partial transparency of model internals such as architecture or intermediate outputs. However, no methods in the reviewed table operate under full white-box access, highlighting a gap in scenarios where such access might be available.

4 Evaluation

4.1 Experimental Setups

We conduct experiments on the methods listed in Table 1, evaluating them under standardized datasets, models, and parameters across text, image, and text-to-image tasks while dynamically aligning with the auditing settings taxonomy in Fig. 2. Specifically, image tasks are evaluated under the pre-training setting with completely overlapping data and black-box access, while text and text-to-image tasks use fine-tuning with the same data-model relationship and access level. We adopt the true positive rate (TPR) @ the false positive rate (FPR)=0.05 as the unified metric, with additional analysis of varying injection rates for text-to-image tasks to assess memorization capabilities, ensuring systematic and fair comparisons within our proposed framework.

4.2 Overall Performance

For text tasks, we employ the industry-standard LLaMA-7B model paired with the comprehensive Wiki dataset; for image tasks, we utilize the well-established ResNet18 architecture with the CIFAR-10 benchmark, maintaining experimental efficiency while ensuring direct comparability with prior research; for text-to-image tasks, we use Stable Diffusion v2.1, the most widely available open-source model, as well as a unique art style and a Pokémon dataset with clear copyright attribution. This standardized experimental framework is designed to yield both statistically robust and practically meaningful results. As shown in Table 2, the performance of different techniques varies significantly across tasks. We observe that methods for text tasks are mostly fingerprint-based and exhibit much lower accuracy compared with other tasks. Additionally, these methods are difficult to generalize in multimodal tasks, and the evaluation metrics are challenging to standardize.

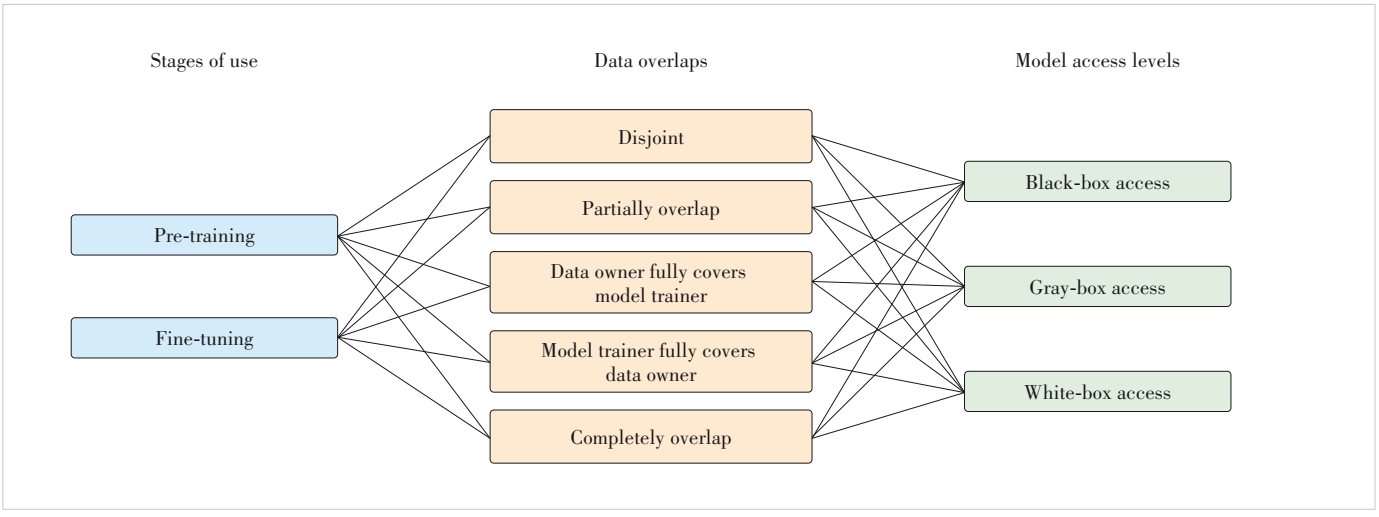

Figure 2. Combinations of auditing settings

Table 2. Performance evaluation of methods, where the injection rate for intrusive methods is fixed at 2%

Reference	Domain	Model	Dataset	Type	TPR@FPR=0.05
CHEN et al.[25]	Image	ResNet18	CIFAR-10	Intrusive	0.155 6
HUANG et al.[29]					0.587 7
SHI et al.[32]	Text	LLaMA-7B	Wiki	Non-intrusive	0.147 9
MAINI et al.[33]					0.068 1
WANG et al.[26]	Text-image	Stable Diffusion v2.1	Pokémon BLIP captions dataset	Intrusive	0.220 0
REN et al.[27]					0.550 0
LI et al.[28]					0.686 7
CUI et al.[30]					0.626 7
DU et al.[35]	Text-image	Stable Diffusion v2.1	Pokémon BLIP captions dataset	Non-intrusive	0.833 0

FPR: false positive rate TPR: true positive rate

4.3 Impact of Injection Rate

We conduct experiments on the intrusive methods listed in Table 1 for text-to-image tasks under different injection rates. In Table 3, evaluations are conducted using the Stable Diffusion v2.1 model and the Pokémon BLIP captions dataset. It can be seen that although these methods achieve nearly 100% accuracy at high injection rates, their performance drops drastically at lower injection rates. For example, when the injection rate is set to 0.005, the model either barely learns most watermarked data or achieves extremely low accuracy.

5 Open Problems and Future Directions

1) Open problem 1: Current copyright protection methods predominantly focus on single-modal data scenarios. From the experimental results in Table 2, most existing techniques are designed for use in unimodal contexts like text-only or image-only datasets. For instance, methods developed for text-to-image models or image-to-text systems often fail to account for the inherent complexity of multimodal correlations. Although there are currently a limited number of multimodal auditing methods available, they primarily offer only a conceptual framework rather than practical solutions. Future direction 1: A promising direction is to construct cross-modal representations that capture the semantic and stylistic alignment between textual prompts and generated images. By leveraging such image-text joint embeddings or alignment scores, auditors can better assess whether unauthorized use of copyrighted material has occurred. For example, if a diffusion model consistently maps a particular artist's textual description to visually similar styles or motifs, this could indicate style piracy and warrant deeper auditing.

2) Open problem 2: In open-world deployment settings with large-scale models and datasets, injection rates are typically low, necessitating robust auditing methods for low watermark densities. As the scales of models grow (e.g., LLaMA-65B and GPT-4), which are trained on massive, heterogeneous datasets, the fraction of modified data becomes increasingly diluted. This low injection rate significantly reduces the signal-to-noise ratio of watermark-based detection methods. From the experimental results in Table 3, existing approaches often suffer from a sharp degradation in auditing accuracy as the injection rate drops, limiting their practical utility. Future direction 2: Future work should emphasize the development of intrusive auditing mechanisms that exhibit low sensitivity to injection rates, perhaps by focusing on instance-level detection, trigger generalization, or aggregating weak signals across multiple inputs. Some promising directions include adaptive watermarking strategies, ensemble detection methods, or leveraging model memorization behavior even for sparsely embedded samples.

3) Open problem 3: There is a lack of consistency in current benchmarking practices, with different methods evaluated on disparate models and datasets. A major limitation in the current literature on dataset copyright auditing is the lack of a standardized experimental protocol. Even though this paper adopts unified performance metrics to evaluate the capabilities of different methods, the accuracy of these methods varies significantly across different tasks. Moreover, the

Table 3. Performance evaluation of intrusive methods under different injection rates, where α denotes the injection rate

Reference	α=0.005	α=0.02	α=0.05	α=0.10	α=0.20	α=0.50	α=1.00
WANG et al.[26]	0.2	0.220 0	0.353 3	0.650 9	0.777 8	0.866 7	0.936 9
REN et al.[27]	0.086 7	0.120 0	0.940 0	0.993 3	1.000 0	1.000 0	1.000 0
LI et al.[28]	0.233 3	0.686 7	0.726 7	0.980 0	0.993 3	1.000 0	1.000 0
CUI et al.[30]	0.006 7	0.626 7	0.640 0	0.653 3	0.746 7	0.913 3	1.000 0

meaning of the metrics also differs slightly between tasks. Many studies introduce their own evaluation datasets, model architectures, and attack settings, which hinders direct comparisons of effectiveness, robustness, and scalability across different methods. Future direction 3: To address this issue, future work should establish uniform benchmarking frameworks, where multiple auditing approaches are assessed under identical experimental settings, including model types (e. g., diffusion and transformer-based LLMs), data domains (e. g., Books3 and WikiArt), and access levels (e. g., black-box vs. gray-box). Such a setup would allow for more comprehensive and fair comparisons, providing deeper insights into each method's strengths, weaknesses, and applicability across scenarios. It would also facilitate the development of standardized metrics for auditing accuracy, robustness to adversarial removal, and computational overhead.

6 Conclusions

This paper systematically reviews the state of dataset copyright auditing in large models, focusing on both methodological advances and practical gaps. We outline a taxonomy based on data domain, usage stages, data overlaps, and model access levels, revealing a landscape largely dominated by black-box methods targeting fine-tuned models. Despite recent progress, significant challenges remain. Notably, current approaches lack cross-modal generalization, perform poorly under low injection rates of modified data, and suffer from inconsistent evaluation practices. To address these issues, we propose advancing toward cross-modal feature auditing, designing low-sensitivity detection techniques, and building standardized benchmark protocols. These future directions are essential to ensure the legal and ethical deployment of large-scale AI systems, especially as they increasingly permeate sensitive and creative domains. By bridging the technical and legal aspects of data provenance, dataset auditing holds promise as a foundational pillar of AI governance.

References

[1] LIANG Z, XU Y, HONG Y, et al. A survey of multimodel large language models [C]//Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering. ACM, 2024: 405 – 409. DOI: 10.1145/3672758.3672824
[2] TRUMMER I. Large language models: principles and practice [C]//Proceedings of IEEE 40th International Conference on Data Engineering (ICDE). IEEE, 2024: 5354 – 5357. DOI: 10.1109/ICDE60146.2024.00404
[3] ZHU X P, YAO H D, LIU J, et al. Review of evolution of large language model algorithms [J]. ZTE technology journal, 2024, 30(2): 9 – 20. DOI: 10.12142/ZTETJ.202402003
[4] TIAN H D, ZHANG M Z, CHANG R, et al. A survey on large model training technologies [J]. ZTE technology journal, 2024, 30(2): 21 – 28. DOI: 10.12142/ZTETJ.202402004
[5] WANG Y T, PAN Y H, SU Z, et al. Large model based agents: state-of-the-art, cooperation paradigms, security and privacy, and future trends [EB/OL]. (2024-09-22) [2025-06-14]. https://arxiv.org/abs/2409.14457
[6] SAMEK W, MONTAVON G, LAPUSCHKIN S, et al. Explaining deep neural networks and beyond: a review of methods and applications [J]. Proceedings of the IEEE, 2021, 109(3): 247 – 278
[7] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models [J]. ACM transactions on intelligent systems and technology, 2024, 15(3): 1 – 45. DOI: 10.1145/3641289
[8] WIGGERS K. OpenAI claims to have hit $10 B in annual revenue [EB/OL]. (2025-06-09) [2025-06-14]. https://techcrunch.com/2025/06/09/openai-claims-to-have-hit-10b-in-annual-revenue/ utm_source=chatgpt.com
[9] SCHREINERM. GPT-4 architecture, datasets, costs and more leaked [EB/OL]. (2023-06-28) [2025-06-14]. https://the-decoder. com/gpt-4-architecture-datasets-costs-and-more-leaked/ utm_source=chatgpt.com
[10] COULTER M. Getty images sues stability AI in UK copyright test case [EB/OL]. (2025-06-09) [2025-06-14]. https://www. reuters. com/business/media-telecom/gettys-landmark-uk-lawsuit-copyright-ai-set-begin-2025-06-09
[11] ENGLUND S, MARINO Z. Client alert: court decides that use of copyrighted works in AI training is not fair use: Thomson Reuters Enterprise Centre GmbH v. s. Ross Intelligence Inc. [EB/OL]. (2025-02-12) [2025-06-14]. https://www.jenner.com/en/news-insights/publications/client-alert-court-decides-that-use-of-copyrighted-works-in-ai-training-is-not-fair-use-thomson-reuters-enterprise-centre-gmbh-v-ross-intelligence-inc
[12] GOODMAN D. Thomson Reuters wins AI copyright 'fair use' ruling against one-time competitor [EB/OL]. (2025-02-11) [2025-06-14]. https://www.reuters.com/legal/thomson-reuters-wins-ai-copyright-fair-use-ruling-against-one-time-competitor-2025-02-11
[13] MORALES J. Meta staff torrented nearly 82 TB of pirated books for AI training [EB/OL]. (2025-02-09) [2025-06-14]. https://arstechnica.com/civis/threads/meta-torrented-over-81-7tb-of-pirated-books-to-train-ai-authors-say.1505523
[14] WEIR K. This is how Meta AI staffers deemed more than 7 million books to have no "economic value" [EB/OL]. (2025-04-15) [2025-06-14]. https://www.vanityfair.com/news/story/meta-ai-lawsuit
[15] GUO J F, LI Y M, CHEN R B, et al. Zeromark: towards dataset ownership verification without disclosing watermark [C]//International Conference on Neural Information Processing Systems. ACM, 2024: 120468 – 120500
[16] LI Y M, BAI Y, JIANG Y, et al. Untargeted backdoor watermark: towards harmless and stealthy dataset copyright protection [C]//International Conference on Neural Information Processing Systems. ACM, 2022: 13238 – 13250
[17] LI Y Z, LI Y M, WU B Y, et al. Invisible backdoor attack with sample-specific triggers [C]//International Conference on Computer Vision (ICCV). IEEE, 2021: 16443 – 16452. DOI: 10.1109/ICCV48922.2021.01615
[18] SZYLLER S, ZHANG R, LIU J, et al. On the robustness of dataset inference [EB/OL]. (2023-06-15) [2025-06-14]. https://openreview.net/pdf id=LKz5SqIXPJ
[19] MAINI P, YAGHINI M, PAPERNOT N. Dataset inference: ownership resolution in machine learning [EB/OL]. [2025-06-14]. https://ppml-workshop.github.io/ppml20/pdfs/Maini_et_al.pdf
[20] LIU G Y, XU T L, MA X Q, et al. Your model trains on my data protecting intellectual property of training data *via* membership fingerprint authentication [J]. IEEE transactions on information forensics and security, 2022, 17: 1024 – 1037. DOI: 10.1109/TIFS.2022.3155921
[21] REN K, YANG Z Q, LU L, et al. SoK: on the role and future of AIGC watermarking in the era of Gen-AI [EB/OL]. [2025-06-14]. https://arxiv.org/html/2411.11478v2#S1
[22] HARTMANN V, SURI A, BINDSCHAEDLER V, et al. Sok: memorization in general-purpose large language models [EB/OL]. (2023-10-24) [2025-06-14]. DOI: 10.48550/arXiv.2310.18362
[23] DU L K, ZHOU X R, CHEN M, et al. SoK: dataset copyright auditing in machine learning systems [EB/OL]. (2024-10-22) [2025-06-14]. DOI: 10.48550/arXiv.2410.16618

[24] DONG T, LI S F, CHEN G X, et al. RAI2: responsible identity audit governing the artificial intelligence [EB/OL]. [2025-06-14]. https://www.ndss-symposium. org/wp-content/uploads/2023/02/ndss2023_f1012_paper. pdf. DOI: 10.14722/ndss.2023.241012
[25] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning [EB/OL]. [2025-06-14]. https://arxiv.org/pdf/1712.05526
[26] WANG S R, ZHU Y B, TONG W, et al. Detecting dataset abuse in fine-tuning stable diffusion models for text-to-image synthesis [EB/OL]. (2024-09-27) [2025-06-14]. https://arxiv.org/abs/2409.18897
[27] REN J, CUI Y Q, CHEN C, et al. EnTruth: enhancing the traceability of unauthorized dataset usage in text-to-image diffusion models with minimal and robust alterations [EB/OL]. (2024-06-20) [2025-06-14]. https://arxiv.org/abs/2406.13933
[28] LI B H, WEI Y H, FU Y K, et al. Towards reliable verification of unauthorized data usage in personalized text-to-image diffusion models [EB/OL]. (2024-10-14) [2025-06-14]. https://arxiv.org/abs/2410.10437
[29] HUANG Z H, GONG N Z, REITER M K. A general framework for data-use auditing of ML models [C]//ACM SIGSAC Conference on Computer and Communications Security. ACM, 2024: 1300 – 1314. DOI: 10.1145/3658644.3690226
[30] CUI Y Q, REN J, LIN Y P, et al. FT-shield: a watermark against unauthorized fine-tuning in text-to-image diffusion models [J]. ACM SIGKDD explorations newsletter, 2025, 26(2): 76 – 88. DOI: 10.1145/3715073.3715080
[31] HUANG Z H, GONG N Z, REITER M K. Instance-level data-use auditing of visual ML models [EB/OL]. (2025-03-28) [2025-06-14]. https://arxiv.org/abs/2503.22413
[32] SHI W J, AJITH A, XIA M Z, et al. Detecting pretraining data from large language models [EB/OL]. [2025-06-14]. https://arxiv. org/html/2310.16789v3
[33] MAINI P, JIA H R, PAPERNOT N, et al. LLM dataset inference: did you train on my dataset [C]//International Conference on Neural Information Processing Systems. ACM, 2024: 124069 – 124092. DOI: 10.48550/arXiv.2406.06443
[34] MA W L, SONG Y L, XUE M H, et al. The "code" of ethics: a holistic audit of AI code generators [J]. IEEE transactions on dependable and secure computing, 2024, 21(5): 4997 – 5013. DOI: 10.1109/TDSC.2024.3367737
[35] DU L K, ZHU Z, CHEN M, et al. ArtistAuditor: auditing artist style pirate in text-to-image generation models [C]//Proceedings of the ACM on Web Conference 2025. ACM, 2025: 2500 – 2513. DOI: 10.1145/3696410.3714602

Biographies

DU Linkang received his BE and PhD degrees from Zhejiang University, China in 2018 and 2023, respectively. He is currently an assistant professor at the School of Cyber Science and Engineering, Xi'an Jiaotong University, China. His research interests include privacy-preserving computing and trustworthy machine learning.

SU Zhou (zhousu@xjtu.edu.cn) is a professor with Xi'an Jiaotong University, China and his research interests include multimedia communication, wireless communication, network security and network traffic. He received the Best Paper Award of International Conference IEEE AIoT 2024, IEEE WCNC 2023, IEEE VTC-Fall 2023, IEEE ICC 2020, etc. He is an associate editor of the *IEEE Internet of Things Journal* and the *IEEE Open Journal of Computer Society*, and the chair of IEEE VTS Xi'an Chapter Section.

YU Xinyi is currently pursuing her master's degree at the School of Cyber Science and Engineering, Xi'an Jiaotong University, China. She received her bachelor's degree in computer science and technology from Hefei University of Technology, China. Her research interests include privacy protection and data traceability within machine learning systems.