Watershed: A Unified Benchmark for End-to-End Data Provenance Evaluation

John Russell Himawan [*1] **Gregory Kang Ruey Lau** [*1 2] **Bryan Kian Hsiang Low** [1]

Abstract

Data provenance aims to determine whether and how a data source has influenced a downstream LLM. Despite growing interest in data provenance research, current methods tend to specialize on specific settings and suffer from fragmented evaluation standards. To address this, we introduce Watershed, a unified benchmark and toolkit for end-to-end provenance evaluation. Watershed structures data provenance into stage-wise tests spanning data preparation, LLM training, black-box auditing, and downstream applications such as membership audit, multi-owner source attribution, and unlearning verification. We evaluate existing provenance methods such as watermarking and membership inference attacks on Watershed, across a wide range of datasets, model families and attacks. Our results confirm that methods vary in effectiveness across different stages and tasks. By providing a unified framework and exposing these failure modes, Watershed establishes a rigorous basis for evaluating data provenance methods holistically.

1. Introduction

Data owners face the risk of their published content being scraped, copied, or reused without consent. This problem is exacerbated further as LLMs are trained on a mixture of text, code, and other data collected from heterogeneous sources (Gao et al., 2020; Langlais et al., 2026). Such training data may contain unauthorized content from data owners (Bommasani et al., 2023), and downstream uses of the data are difficult to audit (Mökander et al., 2023). This raises issues in data provenance (Bommasani et al., 2022): Once an LLM is released, how can a data owner evaluate ❶ whether their data was used in training (e.g. *membership audit*) (Lau et al., 2024; Carlini et al., 2021), ❷ whether LLM outputs can be attributed back to them (e.g. *source attribution*) (Lu et al., 2025), or ❸ whether a claimed removal procedure has actually reduced the influence of their data (e.g. *machine unlearning*) (Cao & Yang, 2015; Yao et al., 2024)? These questions are central to dataset governance, attribution, and regulatory compliance, while also having implications to broader data-centric AI applications (Xu et al., 2024; Khoo et al., 2026).

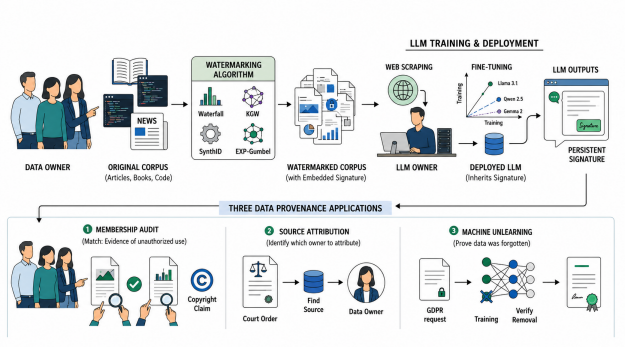

Figure 1. A data owner may release data with data provenance in mind. If a downstream LLM is trained on that data, a provenance signal (e.g., such as watermarking) may persist in model outputs and support black-box audits such as *membership audit*, *source attribution*, or *machine unlearning*.

Existing work has addressed these data provenance questions in LLMs, but they have done so largely in isolation. *Membership audit* is typically approached via membership inference attacks (MIAs) (Shokri et al., 2017; Shi et al., 2024; Zhang et al., 2025) that exploit naturally occurring signals of training exposure, or via data-centric watermarking that actively injects owner-specific signals that persists after model training (Lau et al., 2024). *Source attribution* involves distinguishing between multiple data owners, to know which data owner's data contributed to the LLM's output (Lu et al., 2025). *Machine unlearning* is commonly verified through behavioral probes of the unlearned model over the forget set (e.g., verbatim regurgitation, knowledge questioning, or susceptibility to MIAs (Maini et al., 2024a; Shi et al., 2025), motivating watermark-based unlearning metrics (Lu et al., 2026). These questions are evaluated on different datasets and metrics. It is therefore unclear whether

[1]Department of Computer Science, National University of Singapore, Singapore 117417 [2]CNRS@CREATE, 1 Create Way, #08-01 Create Tower, Singapore 138602. Correspondence to: John Russell Himawan, Gregory Kang Ruey Lau <{johnr.himawan, gregorylau}@u.nus.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

a method that succeeds on one question would succeed on another. However, in practice, a single data provenance method would need to be implemented while satisfying all these scenarios.

For instance, *membership audit* is typically studied on pretraining-data detection benchmarks such as WikiMIA (Shi et al., 2024; Zhang et al., 2025), which score an MIA by how well it separates member from non-member documents of an already-trained model, with no control over how the data entered training. *Machine unlearning* is instead evaluated on dedicated suites such as TOFU (Maini et al., 2024a) and MUSE (Shi et al., 2025), which fine-tune a model on synthetic author profiles or news and book corpora and measure forgetting through question answering and verbatim memorization. Each benchmark fixes its own corpus, training pipeline, and metrics, and neither covers attribution among multiple owners, so evidence obtained under one protocol does not transfer to the others.

To address this, we introduce Watershed, a benchmark and toolkit for evaluating data-centric watermarking on an end-to-end audit pipeline for data provenance. Watershed supports stage-wise evaluation of data-level verifiability, fidelity, robustness, persistence through training, multi-data owner separability, and performance on downstream application tasks such as *membership audit*, *source attribution*, and *machine unlearning verification*.

Our key contributions are the following:

- We formulate data provenance as a systemic audit problem spanning data preparation, model training, black-box querying, scoring, and application-level decisions.
- We present Watershed, an extensible benchmark with stage-wise tests for verifiability, fidelity, robustness, training persistence, multi-owner separability, membership audit, source attribution, and unlearning verification.
- We systematically evaluate existing methods via Watershed, establishing insights such as showing that high data-level verifiability does not guarantee reliable model-level provenance and identifying key failure modes in persistence, robustness, and attribution.

2. Data Provenance as Pipeline Evaluation

We frame data provenance as a single audit pipeline (Fig. 2): a data owner may prepare their data with a provenance method before release, an LLM is trained on the released data, and the owner probes the LLM to answer provenance questions.

2.1. Problem Setting and Access Model

Consider N data owners, where each owner $i \in \{1, \ldots, N\}$ controls a dataset $\mathcal{D}_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and releases a version $\widetilde{\mathcal{D}}_i$ of it. An LLM provider constructs a fine-tuning corpus $\mathcal{T} = \bigcup_{i \in \mathcal{S}} \widetilde{\mathcal{D}}_i$, where $\mathcal{S} \subseteq \{1, \ldots, N\}$ is the set of included owners. Let M_0 denote a base LLM and $M_{\mathcal{T}}$ the model after fine-tuning M_0 on $\mathcal{T}$.

A *data provenance method* $\mathcal{P}$ equips each owner with two functions: (1) a transformation function σ_i that produces owner i's released dataset, $\widetilde{\mathcal{D}}_i = \sigma_i(\mathcal{D}_i)$, and (2) a scoring function assigning any text x with a score $s_i(x) \in \mathbb{R}$, where a higher value indicates stronger evidence that x carries a trace of owner i's data.

This interface covers both families of provenance methods (i.e., *active* and *passive*). *Active* methods transform the data before release. An example is data-centric watermarking, where σ_i embeds an owner-specific signal by paraphrasing with a watermarking LLM M_w under a watermark key k_i, and s_i is the watermark detector keyed by k_i, computed over the tokenizer of M_w. *Passive* methods release the data unchanged (i.e., $\widetilde{\mathcal{D}}_i = \sigma_i(\mathcal{D}_i) = \mathcal{D}_i$). An example is membership inference attacks (MIAs) which instantiate s_i with likelihood, perplexity, or reference-model statistics computed with knowledge of $\mathcal{D}_i$. We evaluate five watermarking schemes and several MIAs (Sec. 3.2).

Black-box access. To check whether its data was used, a data owner probes the deployed LLM under black-box access (to resemble real-world scenarios). The data owner may query $M_{\mathcal{T}}$ with prompts q and observe responses $y \sim M_{\mathcal{T}}(\cdot \mid q)$, but cannot inspect weights, gradients, optimizer states, or the training corpus. Given a set of audit prompts $\mathcal{Q}_i$, the owner scores the responses with s_i.

2.2. Audit Tasks

Data provenance methods facilitate audit decisions on the three questions raised in Sec. 1. Each task scores LLM outputs for owner-specific evidence, but differs in the decision it must support.

Q1 Membership Audit. We need to decide whether the LLM was trained on owner i's data. To accomplish this, we test $H_0 : i \notin \mathcal{S}$ against $H_1 : i \in \mathcal{S}$. The owner queries the model with their audit prompts $\mathcal{Q}_i$ and computes

$$A_i(M) := \tfrac{1}{|\mathcal{Q}_i|} \textstyle\sum_{q \in \mathcal{Q}_i} s_i(y_q), \qquad y_q \sim M(\cdot \mid q),$$

the mean owner-i score of the model's responses. Membership is predicted by thresholding, $\widehat{m}_i = \mathbf{1}\{A_i(M_{\mathcal{T}}) > \tau_i\}$. We measure how well the statistic separates $M_{\mathcal{T}}$ from a reference model M_{ref} never trained on owner i's data, reporting AUROC and TPR at fixed FPR.

Q2 Source Attribution. Identify *which* of the N owners' data the LLM's outputs reflect. The auditor issues a set of queries $\mathcal{Q}_i$ all relevant to a given content. Every owner's

scoring function is applied to the same responses, and we attribute to the owner whose audit statistic over the query set is highest, $\widehat{\jmath} := \arg\max_{j\in\{1,\dots,N\}} A_j(M_{\mathcal{T}})$; a correct attribution is $\widehat{\jmath} = i$. We report the accuracy as this is an N-way identification among owners.

Q3 Machine Unlearning. A subset of owners $\mathcal{F} \subseteq \mathcal{S}$ requests removal of their data, and an unlearning procedure U produces $M_{-\mathcal{F}} = U(M_{\mathcal{T}}, \mathcal{D}^{\mathrm{F}})$, where $\mathcal{D}^{\mathrm{F}} = \bigcup_{f\in\mathcal{F}} \mathcal{D}_f$. We evaluate $M_{-\mathcal{F}}$ against the retrain-on-retain oracle $M_{\mathcal{R}}$, fine-tuned only on the remaining owners' ($\mathcal{R}$) data. On the forget side, the signal drop $\Delta_f := A_f(M_{\mathcal{T}}) - A_f(M_{-\mathcal{F}})$ should be large for every $f \in \mathcal{F}$; on the retain side, the oracle gap $R_r := A_r(M_{-\mathcal{F}}) - A_r(M_{\mathcal{R}})$ should be near zero for every $r \in \mathcal{R}$. In practice, we report the per-side audit signals of $M_{\mathcal{T}}$, $M_{-\mathcal{F}}$, and $M_{\mathcal{R}}$ directly, from which both quantities can be read off. Reading both sides together distinguishes genuine removal from indiscriminate damage to the model.

2.3. Evaluation Criteria

A provenance method can fail at multiple points before an audit decision is made. For example, the signal may already be weak in the prepared data, degrade data quality, be removed by pre-processing (such as paraphrasing), fail to persist through training, or collide with another owner's signal.

We therefore evaluate provenance methods through three criteria defined at the data level (D1–D3); the rest of the pipeline then tests whether they are *maintained* across stages: training must preserve verifiability (*persistence*), multi-owner deployments must keep owners' signals mutually distinguishable (*separability*), and the audit tasks Q1–Q3 stress these maintained properties in combination.

D1 Verifiability. Owner i's score s_i should distinguish the prepared corpus from unprepared data. For $\tilde{x} \in \widetilde{\mathcal{D}}_i$ and $x \in \mathcal{D}_i$, ideally $\mathbb{P}[\, s_i(\tilde{x}) > s_i(x) \,] \approx 1$, i.e., the AUROC of s_i against unprepared negatives approaches 1. In multi-owner settings, owner i's prepared text should likewise score higher under s_i than under any other owner's function, $\mathbb{P}[\, s_i(\tilde{x}) > s_j(\tilde{x}) \,] \approx 1$ for $j \neq i$, so that owners remain distinguishable.

D2 Fidelity. The prepared corpus $\widetilde{\mathcal{D}}_i$ should preserve the meaning, fluency, and downstream utility of $\mathcal{D}_i$, measured via semantic similarity and task utility: a signal that renders the data unusable may be detectable but impractical. *Passive* methods satisfy D2 trivially while *active* methods may trade off fidelity for stronger owner-specific signals.

D3 Robustness. Verifiability should survive realistic perturbations g of the released data or observed outputs, such as token deletion, synonym substitution, or paraphrasing: $s_i(g(\tilde{x}))$ should still separate owner i's data from unprepared and wrong-owner controls. The most consequential transformation is LLM training itself: *persistence* is D3 with g taken to be the training-and-generation channel, requiring responses of $M_{\mathcal{T}}$ to score higher than those of a reference model M_{ref} never trained on owner i's data, i.e., $\mathbb{P}[\, s_i(y) > s_i(y') \,] \approx 1$ for $y \sim M_{\mathcal{T}}(\cdot \mid q)$, $y' \sim M_{\mathrm{ref}}(\cdot \mid q)$, $q \in \mathcal{Q}_i$.

3. Watershed: A Benchmark for Data-Centric Provenance

The previous section defines data provenance as a systemic, multi-stage evaluation problem. Watershed operationalizes this view in a benchmark and toolkit for evaluating provenance methods under a shared setting. Rather than treating *membership audit*, *source attribution*, and *machine unlearning* as unrelated tasks, Watershed exposes them as different audit decisions (Q1–Q3) over a common pipeline.

Watershed is designed to answer two broad questions. First, *where in the pipeline does a provenance method succeed or fail?* Second, *how do active and passive provenance methods compare when evaluated under matched LLMs, datasets, and controls?* This section describes the benchmark design, components, and evaluation.

3.1. Benchmark Overview

Given a collection of owner corpora $\{\mathcal{D}_i\}_{i=1}^{N}$, Watershed proceeds through five stages, with attacks and evaluations applied across the stages:

1. Owner construction. The benchmark begins by partitioning a dataset into owner-specific corpora. Data owners correspond to sources, such as authors or publishers.

2. Provenance preparation. Each data owner applies their transformation function to their corpus, yielding the prepared corpus $\widetilde{\mathcal{D}}_i = \sigma_i(\mathcal{D}_i)$.

3. LLM training. A base LLM M_0 is trained on the corpus $\mathcal{T} = \bigcup_{i\in\mathcal{S}} \widetilde{\mathcal{D}}_i$ of included owners' prepared data. This stage converts a data-level provenance signal into an LLM-level audit problem.

4. Black-box audit. The trained model is queried using audit prompts derived from the owner's data (e.g., prefixes of released data). For each prompt q, the model produces a response $y \sim M_{\mathcal{T}}(\cdot \mid q)$. The benchmark then scores the response under one or more candidate owners using method-specific scoring functions $s_i(y)$.

5. Application evaluation. The scores help to inform the audit decisions Q1–Q3, evaluated in Sec. 3.3.

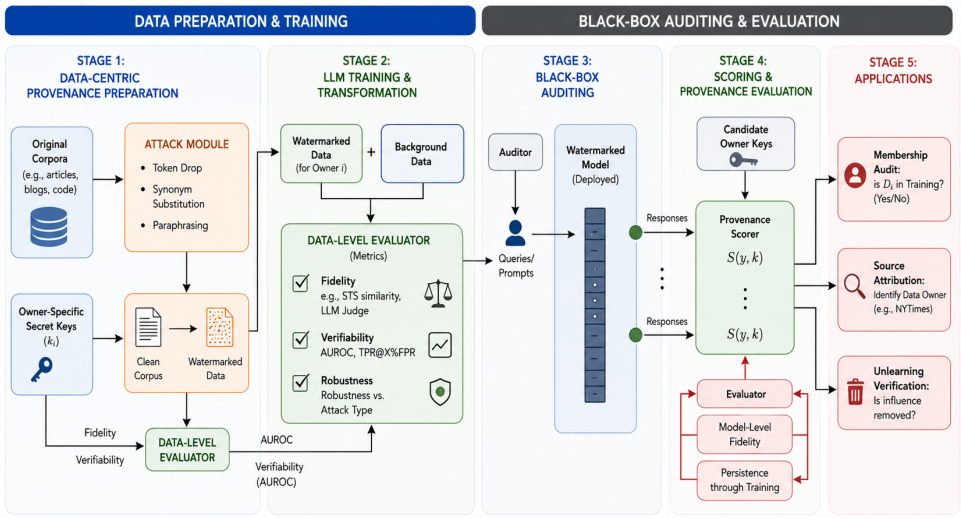

Figure 2. Overview of Watershed. Owners prepare provenance-equipped corpora under owner-specific keys (Stage 1); an LLM is trained on the released data alongside background data (Stage 2); the owner audits the deployed model under black-box access (Stage 3), scoring responses under candidate owner keys (Stage 4) to support membership audit, source attribution, and unlearning verification (Stage 5). Stage-wise evaluators measure fidelity, verifiability, and robustness at the data level, and persistence through training at the model level.

3.2. Components

Watershed consists of modular components for data preparation, provenance methods, model training, auditing, and evaluation. This modularity is important because existing data provenance methods are often evaluated under isolated pipelines, making direct comparison difficult. Prior methods are proposed and evaluated in isolation, while existing watermarking benchmarks focus on model-centric watermarking rather than data provenance tasks.

Datasets. In our experiments, we use text datasets spanning news articles and blogs to resemble real-world scenarios. We then partition these datasets into source-level owners (e.g., publishers or authors). We consider C4-realnewslike for news articles (Raffel et al., 2020) and Blog Authorship Corpus for blogs (Schler et al., 2006), which we have processed into `Watershed-C4` and `Watershed-Blog` respectively (see App. A for more details). This allows us to evaluate data provenance involving multiple owners.

Provenance Methods. We consider both *active* and *passive* provenance methods for the problem setting in Sec. 2.1. For *active* methods, we use the data-centric watermarking framework proposed by Lau et al. (2024), which flexibly enables many watermarking schemes to be used for data provenance. We include Waterfall (Lau et al., 2024), KGW (Kirchenbauer et al., 2023), Unigram (Zhao et al., 2024), EXP-Gumbel (Aaronson, 2022), and SynthID-text (Dathathri et al., 2024). For *passive* methods, we consider MIAs (Shokri et al., 2017) such as Min-K%++ (Zhang et al., 2025). These provenance methods we use are elaborated further in App. B.

Models. The benchmark supports multiple open LLM families. This is important because watermark signals may depend on tokenizer, model distribution, training dynamics, and generation behavior. We therefore distinguish matched settings, where the LLM used to prepare data and the LLM trained on that data come from the same family, from cross-family settings, where they differ. In our implementation, we considered the models Llama 3.1 8B (Grattafiori et al., 2024) and Gemma 2 9B (Team et al., 2024).

Attacks. Watershed supports multiple attacks on the data to evaluate robustness. We include token deletion, synonym substitution, and LLM paraphrasing. These attacks are used to resemble the scenario in which an adversary wants to use the released data while evading a provenance audit. Hence, the adversary perturbs the data with these attacks to weaken the provenance signal while preserving the semantic content. Such perturbations may also arise without an adversary, as part of regular training practices. For example, rephrasing web data with LLMs is an established recipe for data-efficient training (Maini et al., 2024b). These attacks outline real-world scenarios which may challenge the robustness criteria (D3).

3.3. Evaluation Stages

The benchmark reports both stage-wise diagnostics and application-level outcomes. This separation is central to Watershed: when a method fails in an audit task, stage-wise diagnostics help identify whether the failure comes from a weak data-level signal, poor fidelity, lack of robustness, loss during training, or interference from the owner.

Data-stage evaluation. Before training any LLM, Watershed evaluates the prepared corpus directly. This includes data-level verifiability with its wrong-owner separability condition (D1), fidelity with respect to the original data (D2), and robustness to corpus perturbations (D3). These tests answer whether the provenance signal was successfully embedded and whether the prepared data remain useful.

Model-stage evaluation. After training, the benchmark evaluates whether the signal persists in LLM outputs and whether owners remain distinguishable i.e., whether D1 is maintained through training, per owner (persistence), and across owners (separability). The key comparison is between the trained model and suitable controls, such as a base LLM, a clean fine-tuned LLM, or scoring under wrong owners. This stage is the defining test for data provenance: a signal that does not survive training cannot support downstream LLM audit.

Application-stage evaluation. Each audit decision is evaluated with task-appropriate metrics. For *membership audit* (Q1), we report AUROC and TPR at fixed FPR separating member from non-member owners. For *source attribution* (Q2), we report the accuracy of identifying, among the N candidate owners, the owner responsible for the content targeted by a set of related queries. For *unlearning verification* (Q3), we report the audit signal separately on forget-side and retain-side prompts, anchored by the trained model $M_{\mathcal{T}}$ and the retrain-on-retain oracle $M_{\mathcal{R}}$ (Sec. 2.2): the forget-side signal should fall to the oracle's level while the retain-side signal remains at it, separating genuine removal from indiscriminate damage.

4. Empirical Analysis

For the data owner to confidently claim that the LLM has been trained on their data, we must first perform rigorous evaluations on the effectiveness of the provenance methods. We split our analysis on the data-stage (Section 4.1) and model-stage (Section 4.2). This ensures that we can properly verify whether the uploaded data contain the provenance signal and whether the LLM trained on the released data generates outputs that still maintain the signal.

Experimental Settings. For this experiment, we follow the default watermarking-specific hyperparameters provided by MarkLLM (Pan et al., 2024) and we use $K = 20$ for Min-K%++ as suggested by (Shi et al., 2024). We provide a more detailed description of the experimental settings in App. C.

4.1. Data-stage Evaluation

4.1.1. Verifiability and Robustness to Attacks

We first evaluate the verifiability (D1) of the provenance signal before any model training. Since *passive* methods release the data unchanged, we focus more on the *active* methods. Fig. 3 reports the AUROC of each scoring function on prepared vs original text (i.e., watermarked v non-watermarked text). The relatively high AUROC scores before attacks shows that each method embeds a strong signal into the released corpus. Note that while attacks such as token drop and synonym substitution may reduce verifiability slightly, paraphrase attacks are generally much more effective because they can rewrite more tokens while maintaining high semantic similarity to the original. However, Waterfall and KGW remain relatively robust to all attacks.

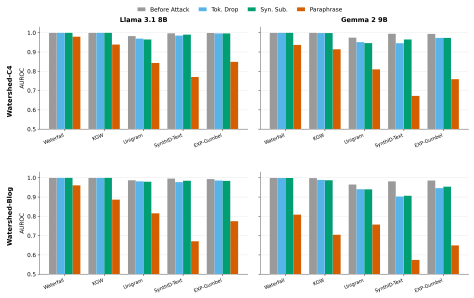

Figure 3. AUROC of the `Watershed-C4` and `Watershed-Blog` datasets watermarked by various watermarking schemes before and after attacks (i.e., Token Drop, Synonym Substitution, and Paraphrase).

As per the multi-owner separability condition of D1, we also expect to be able to distinguish between the N owners' signals. The per-owner signal should not collide across data owners, as owner i's scoring function s_i must distinguish between data from $\widetilde{\mathcal{D}}_i$ vs. $\widetilde{\mathcal{D}}_j$ where $j \neq i$ at the data level, before any LLM is trained on the released corpus.

Fig. 10 shows that the per-owner AUROC ≈ 1 across our various configurations, where the AUROC discriminates rows from $\widetilde{\mathcal{D}}_i$ scored under s_i (positives) from rows of any other owner $\widetilde{\mathcal{D}}_j, j \neq i$ (negatives). This empirically shows there is no cross-owner interference at the data level as each owner's signal uniquely identifies their content.

4.1.2. Fidelity

Data owners may be concerned about a potential utility drop when transforming their data before release, so maintaining fidelity (D2) is an important property of a good provenance method. *Passive* methods trivially satisfy this criterion since the data is unchanged. However, for *active* methods, we evaluate the fidelity of the text using STS scores using an embedding model[1].

In Fig. 11, we notice that the STS scores between the watermarked vs non-watermarked text are very high for `Watershed-C4` ($\approx 0.88 - 0.92$) while it's slightly lower

[1] `sentence-transformers/all-mpnet-base-v2`

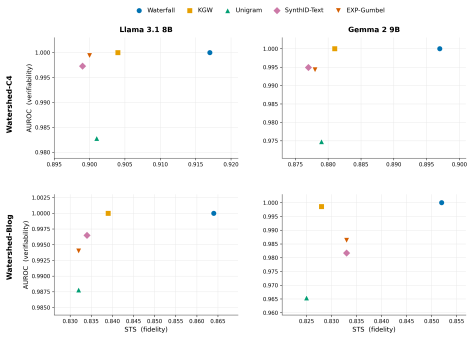

Figure 4. The verifiability-fidelity plot which compares the AUROC (verifiability) and STS scores (fidelity) of each watermarking scheme.

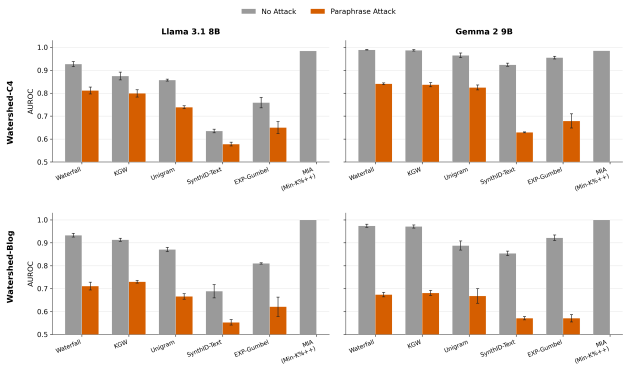

Figure 5. Membership audit AUROC on `Watershed-C4` and `Watershed-Blog`: distinguishing the outputs of an LLM fine-tuned on each owner's prepared corpus from those of a model not trained on it, with the corpus used as released (No Attack) or paraphrased before training (Paraphrase Attack). MIA (Min-K%++) audits an LLM fine-tuned on the original corpus; under the paraphrase attack its AUROC falls below the plotted range.

for `Watershed-Blog` ($\approx 0.83 - 0.86$). This is likely because paraphrasing tends to normalize text toward a neutral style (Krishna et al., 2020). News articles are typically formal and stylistically neutral so paraphrasing does not affect it much, whereas the informal and personal style of blog posts is stripped away.

We also plot the verifiability-fidelity in Fig. 4 to better understand the trade-offs. We see that Waterfall achieves the strongest verifiability and fidelity.

4.2. Model-stage Evaluation

4.2.1. Verifiability and Robustness after Training

For data provenance, it is important to know whether the signal remains verifiable after training, i.e., whether the signal that shows high verifiability on the data-stage will maintain its high verifiability on the model-stage to satisfy D1 across stages. Concretely, if $M_{\mathcal{T}}$ is fine-tuned on a corpus containing owner i's prepared data $\widetilde{\mathcal{D}}_i$, its responses to the audit queries $\mathcal{Q}_i$ should still score highly under s_i.

We fine-tune each base LLM on its own prepared corpus (e.g. if the corpus was watermarked using Waterfall with Llama 3.1 8B as the paraphraser, we fine-tune Llama 3.1 8B on this corpus). We then compute the AUROC between outputs from an LLM trained on watermarked data vs. outputs from an LLM trained on non-watermarked data.

Fig. 5 shows that the provenance signal persists through training (although weaker) for all five watermarking schemes, where the logit-bias-based schemes (Waterfall, KGW, Unigram) have an AUROC of above 0.86, and the non-distortionary schemes (SynthID-Text, EXP-Gumbel) retain a weaker signal. While all watermarking schemes notice a consistently lower AUROC in the model-stage compared to data-stage, we notice that this drop is more prevalent in Llama 3.1 8B than Gemma 2 9B. We hypothesize that this may be a result of how strongly each LLM learns the distribution of the data. Gemma 2 9B shifts its output distribution more closely toward the training text than Llama 3.1 8B, so more of the keyed token bias survives into its generations. Gemma's larger vocabulary (256k vs. 128k tokens) may further sharpen detection, as the watermark bias is expressed over finer-grained tokens.

The *passive* baseline is also strong in this setting. Here, the LLM is fine-tuned on the original corpus, and Min-K%++ computes calibrated token log-probabilities of a candidate document (App. B). We take the AUROC to measure how well this score separates the owner's training documents (member) from held-out documents of the same owner (non-member), and find that the AUROC reaches ≈ 0.98. Without applying any transformations on the dataset, we obtain high verifiability.

However, one must also consider the robustness to attacks. As we have established previously, it is likely that the released data may undergo perturbations before training. Hence, we also consider the robustness of these provenance methods if the data was attacked before training. We paraphrase the released corpora (since paraphrasing is our strongest attack) and the LLM is trained on such paraphrased corpora. For each of the *active* methods (i.e.,watermarking schemes), we paraphrase the watermarked datasets `Watershed-C4` and `Watershed-Blog`. For *passive* methods, we paraphrase the original corpus. In this setting, the trained LLM never encounters the data owners' released corpora and instead only sees paraphrased versions of them.

Under this paraphrase attack, the provenance signal from the watermarking schemes still persists in the outputs of the LLM even though it has been weakened (see Fig. 5). On the other hand, the AUROC of Min-K%++ collapses to

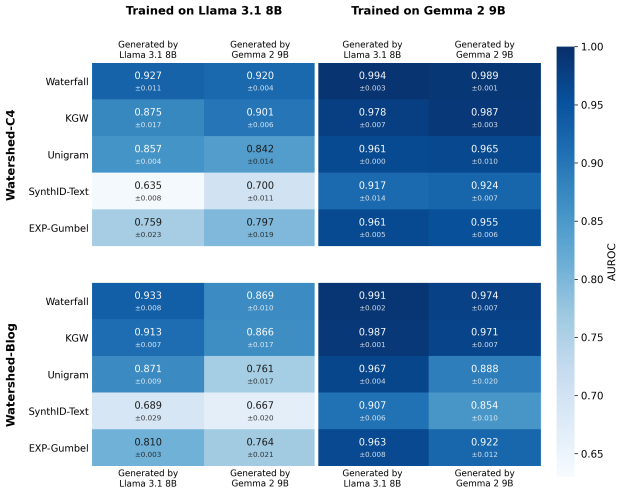

Figure 6. Model-stage detection AUROC when the corpus is watermarked using one model family and fine-tuned on the same or the other family, for each watermarking scheme on `Watershed-C4` and `Watershed-Blog`.

near-random chance.

4.2.2. Cross-Family Watermark Persistence

In practice, the data owner cannot control which LLM is trained on their released data, so the provenance signal must transfer across model families with different tokenizers and output distributions. We therefore evaluate all four combinations of watermarking LLM and trained LLM using Llama 3.1 8B and Gemma 2 9B. Fig. 6 shows that every scheme's signal survives cross-family training; e.g., Waterfall on `Watershed-C4` reaches 0.994 when watermarking with Llama and fine-tuning Gemma, and 0.920 in the reverse direction — comparable to its same-family counterparts (0.989 and 0.927).

4.2.3. Impact of Watermark on Model Utility

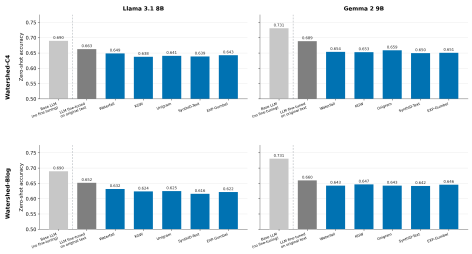

Figure 7. Average zero-shot accuracy across six LM Evaluation Harness tasks to study model utility after fine-tuning. We consider the base LLM, an LLM fine-tuned on the original corpus, and LLMs fine-tuned on each watermarked corpus (using each of the watermarking schemes), on `Watershed-C4` and `Watershed-Blog`.

We also investigate the impact of the watermarks on model utility, given in Fig. 7. We test our LLMs trained on watermarks based on the average zero-shot accuracy across six LM Evaluation Harness (Gao et al., 2024) tasks. The results show that training the LLM on watermarked data does not significantly impact model utility on general tasks. In fact, the degradation of fine-tuning on a watermarked dataset is not too different from the degradation suffered from fine-tuning on a non-watermarked dataset.

4.2.4. Source Attribution

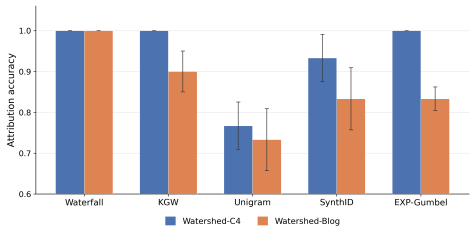

Figure 8. Accuracy which is the N-way identification of the owner responsible for the queried content to perform source attribution, for each watermarking scheme on `Watershed-C4` and `Watershed-Blog`.

Beyond verifying membership for a single owner, an auditor often needs to distinguish between multiple data owners: given some content of interest, which of the N owners' data does the LLM's behavior reflect? The audit naturally involves not just one query but a set of queries all relevant to that content. We therefore score the model's responses to the query set under every candidate owner's key and attribute the content to the owner with the highest aggregate signal. Fig. 8 shows that Waterfall achieves perfect attribution on both datasets, while the other schemes degrade on `Watershed-Blog`; Unigram is the weakest at ≈ 0.75 on both datasets.

4.2.5. Machine Unlearning

Data owners may also request that the influence of their data be removed from the LLM under regulations such as the GDPR, which is known as *machine unlearning* (Cao & Yang, 2015; Yao et al., 2024). Verifying such a claim requires metrics that reveal whether the influence of the forget data is still present in the model (Shi et al., 2025; Maini et al., 2024a; Lu et al., 2026). The gold standard for unlearning is the oracle $M_{\mathcal{R}}$ retrained on only the retain set: a successful unlearning run should be indistinguishable from it on both the forget and retain sides. We therefore compute the AUROC under the same model-side convention as Section 4.2, but separately on prompts from $\mathcal{D}^{\mathrm{F}}$ (forget set) and $\mathcal{D}^{\mathrm{R}}$ (retain set), and report in Fig. 9 the two anchors that bracket any unlearning run which are the trained model $M_{\mathcal{T}}$ (Pre) and the oracle.

On the forget side, the logit-bias watermarking schemes (i.e.,

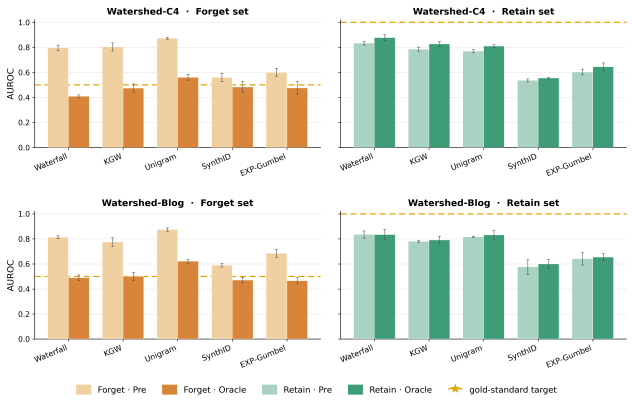

Figure 9. Unlearning-verification anchors on `Watershed-C4` and `Watershed-Blog`: model-side AUROC on forget-set and retain-set prompts for the trained model (Pre) and the retrain-on-retain oracle, per watermarking scheme. Dashed lines mark the ideal targets (chance on the forget set, perfect on the retain set). A successful unlearning run should move the forget-side signal from Pre down to the oracle level while keeping the retain-side signal at it.

Waterfall, KGW, and Unigram) start at a Pre AUROC of 0.78–0.87 on both datasets, while the AUROC for the oracles fall back to near chance. The non-distortionary schemes offer less resolution as EXP-Gumbel retains a moderate window (Pre 0.60–0.68 vs. oracle ≈ 0.47), whereas SynthID's Pre signal (0.56–0.59) is already close to its oracle, leaving little room to distinguish a genuine removal from no removal at all. On the retain side, Pre and oracle nearly coincide for every scheme and both sit below the ideal of perfect detection, so the practical retain-side target for an unlearning run is the oracle level rather than 1.0. Reading both sides together lets an auditor choose the proper scheme to evaluate machine unlearning.

5. Related Works

5.1. Data-centric Watermarking

Data-centric watermarking embeds a watermark directly into a training corpus rather than into model outputs, so that any model trained on the corpus inherits a detectable signal. Existing methods pursue distinct goals such as paraphrase-based rewriting that wraps a model-centric watermark to produce token-biased training text (Lau et al., 2024) and watermark persistence as evidence of incomplete machine unlearning (Lu et al., 2026). Closely related is a line of work which characterizes when the signal of an output watermark persists into a downstream model trained on watermarked text (Sander et al., 2024). These methods have been evaluated under bespoke protocols on different model families and datasets, and to our knowledge have never been directly compared. Watershed addresses this gap with a shared interface and a unified evaluation protocol.

5.2. Benchmarks for Data Provenance Tasks

Each data provenance question has developed its own evaluation ecosystem. Membership audit is studied on pretraining-data detection benchmarks such as WikiMIA (Shi et al., 2024), which score MIAs by separating member from non-member documents of already-trained models (Zhang et al., 2025); the auditor has no control over how the data entered training, and the benchmark carries no multi-owner structure. Machine unlearning is evaluated on dedicated suites such as TOFU (Maini et al., 2024a), which fine-tunes on fictitious author profiles and probes question answering over the forget set, and MUSE (Shi et al., 2025), which measures six criteria including verbatim and knowledge memorization on news and book corpora. Source attribution has no standard benchmark and is evaluated ad hoc by the methods that propose it (Lu et al., 2025). Each of these benchmarks fixes its own corpus, training pipeline, and metrics for a single task, so results do not transfer across questions, and no existing benchmark can evaluate a single provenance method across all three. Watershed provides this missing shared protocol.

6. Conclusion

We presented Watershed, a unified benchmark and toolkit for evaluating data-centric watermarking as an end-to-end provenance audit pipeline. Watershed structures provenance evaluation into stage-wise tests for data-level verifiability, fidelity, robustness, training persistence, and multi-owner separability, together with three downstream audit tasks: membership audit, source attribution, and unlearning verification. Our empirical study across five watermarking schemes, two LLM families, and two data domains shows that data-centric watermarking can support all three audit tasks under a single shared protocol, but a watermark's ranking on one pipeline stage does not predict its ranking on another: SynthID-Text is essentially perfect at the data level but loses substantial signal during training; Unigram is robust through training but yields the weakest attribution; and when a watermark's model-side signal is already weak (e.g., SynthID-Text), the unlearning metric itself loses resolution. These cross-stage gaps justify the stage-wise design — no single number captures whether a watermark is fit-for-purpose across the full audit pipeline.

Acknowledgement

This project was supported by the Swiss AI Initiative through a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a139 on Alps. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD/2023-01-039J). This research is part of

the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Aaronson, S. My ai safety lecture for UT effective altruism. Shtetl-Optimized (Blog), November 2022. URL https://scottaaronson.blog/?p=6823.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., 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, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., 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, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index, 2023. URL https://arxiv.org/abs/2310.12941.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *Proceedings of the 2015 IEEE Symposium on Security and Privacy*, SP '15, pp. 463–480, USA, 2015. IEEE Computer Society. ISBN 9781467369497. doi: 10.1109/SP.2015.35. URL https://doi.org/10.1109/SP.2015.35.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models, 2021. URL https://arxiv.org/abs/2012.07805.

Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., Matejovicova, T., Hayes, J., Vyas, N., Merey, M. A., Brown-Cohen, J., Bunel, R., Balle, B., Cemgil, T., Ahmed, Z., Stacpoole, K., Shumailov, I., Baetu, C., Gowal, S., Hassabis, D., and Kohli, P. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, October 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-08025-4. URL http://dx.doi.org/10.1038/s41586-024-08025-4.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M.,

Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Khoo, Z.-Y., Ryan, Y., Oo, N. H. Y., Pang, H. E., Orlowski, E. J. W., Norhashim, H., Chew, R. W. T., Choo, D., Sim, R. H. L., Chesterman, S., Hahn, J., and Low, B. K. H. Balance human agency & AI assistance in the tussle for the "right" to choose, own, work, and learn. In *Trustworthy AI for Good (AI4GOOD) Workshop @ ICML 2026*, 2026. URL https://openreview.net/forum?id=Mv8gKVRAqp.

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. A watermark for large language models. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17061–17084. PMLR, 23–29 Jul

2023. URL https://proceedings.mlr.press/v202/kirchenbauer23a.html.

Krishna, K., Wieting, J., and Iyyer, M. Reformulating unsupervised style transfer as paraphrase generation. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 737–762, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.55. URL https://aclanthology.org/2020.emnlp-main.55/.

Langlais, P.-C., Chizhov, P., Arnett, C., Hinostroza, C. R., Nee, M., Jones, E. K., Girard, I., Mach, D., Stasenko, A., and Yamshchikov, I. P. Common corpus: The largest collection of ethical data for LLM pre-training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0wSlFpMsGb.

Lau, G. K. R., Niu, X., Dao, H., Chen, J., Foo, C.-S., and Low, B. K. H. Waterfall: Scalable framework for robust text watermarking and provenance for LLMs. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20432–20466, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1138. URL https://aclanthology.org/2024.emnlp-main.1138/.

Lu, X., Wang, J., Zhao, Z., Dai, Z., Foo, C.-S., Ng, S.-K., and Low, B. K. H. WASA: WAtermark-based source attribution for large language model-generated data. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23791–23824, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1219. URL https://aclanthology.org/2025.findings-acl.1219/.

Lu, X., Niu, X., Lau, G. K. R., Bui, N., Sim, R. H. L., Himawan, J. R., Wen, F., Foo, C.-S., Ng, S.-K., and Low, B. K. H. Waterdrum: Watermark-based data-centric unlearning metric. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5GVfneFvhq.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=B41hNBoWLo.

Maini, P., Seto, S., Bai, H., Grangier, D., Zhang, Y., and Jaitly, N. Rephrasing the web: A recipe for compute and data-efficient language modeling, 2024b. URL https://arxiv.org/abs/2401.16380.

Mökander, J., Schuett, J., Kirk, H. R., and Floridi, L. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115, May 2023. ISSN 2730-5961. doi: 10.1007/s43681-023-00289-2. URL http://dx.doi.org/10.1007/s43681-023-00289-2.

Pan, L., Liu, A., He, Z., Gao, Z., Zhao, X., Lu, Y., Zhou, B., Liu, S., Hu, X., Wen, L., King, I., and Yu, P. S. MarkLLM: An open-source toolkit for LLM watermarking. In Hernandez Farias, D. I., Hope, T., and Li, M. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 61–71, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-demo.7. URL https://aclanthology.org/2024.emnlp-demo.7/.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Sander, T., Fernandez, P., Durmus, A. O., Douze, M., and Furon, T. Watermarking makes language models radioactive. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qGiZQb1Khm.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp. 199–205, 01 2006.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Shi, W., Lee, J., Huang, Y., Malladi, S., Zhao, J., Holtzman, A., Liu, D., Zettlemoyer, L., Smith, N. A., and Zhang, C. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership Inference Attacks Against Machine Learning Models . In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, Los Alamitos, CA, USA, May

2017. IEEE Computer Society. doi: 10.1109/SP.2017.41. URL https://doi.ieeecomputersociety.org/10.1109/SP.2017.41.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Xu, X., Wu, Z., Qiao, R., Verma, A., Shu, Y., Wang, J., Niu, X., He, Z., Chen, J., Zhou, Z., Lau, G. K. R., Dao, H., Agussurja, L., Sim, R. H. L., Lin, X., Hu, W., Dai, Z., Koh, P. W., and Low, B. K. H. Position paper: Data-centric AI in the age of large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11895–11913, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.695. URL https://aclanthology.org/2024.findings-emnlp.695/.

Yao, Y., Xu, X., and YangLiu. Large language model unlearning. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105425–105475. Curran Associates, Inc., 2024. doi: 10.52202/079017-3346.

Zhang, J., Sun, J., Yeats, E., Ouyang, Y., Kuo, M., Zhang, J., Yang, H. F., and Li, H. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZGkfoufDaU.

Zhao, X., Ananth, P. V., Li, L., and Wang, Y.-X. Provable robust watermarking for AI-generated text. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SsmT8aO45L.

A. Datasets

In this section, we provide more details on the curation of our datasets (i.e., `Watershed-C4` and `Watershed-Blog`) used in this project and the reason for the design choices.

A.1. `Watershed-C4`

The original C4 realnewslike dataset has roughly 13 million samples of news articles. Since we are interested in data-centric watermarking, we parse the `url` column of the C4 realnewslike dataset (Raffel et al., 2020) to extract the institution that produced a given news article. Then, to ensure the quality of the news article, we first rank the top 50 institutions by article volume as given in Tab. 1. And then, we randomly sample 10 institutions to be used for our dataset, which resulted in https://aljazeera.com/, https://cnbc.com/, https://nytimes.com/, https://reuters.com/, https://theatlantic.com/, https://npr.org/, https://latimes.com/, https://bbc.com/, https://theguardian.com/, and https://wired.com/. 1,000 news articles are extracted per institution to build our dataset of size 10,000.

Afterwards, we filter the text to only include $300 \leq words \leq 2500$ to ensure that the text is sufficiently long enough to realistically absorb a watermark while not being too long such that it exceeds the context length of open-source LLMs.

Rank	Institution	Count	Rank	Institution	Count
1	reuters.com	184,891	26	usatoday.com	57,751
2	nytimes.com	172,510	27	nypost.com	57,637
3	latimes.com	143,984	28	npr.org	57,337
4	theguardian.com	135,552	29	time.com	57,003
5	businessinsider.com	104,965	30	baltimoresun.com	56,164
6	zdnet.com	92,566	31	stanford.edu	53,903
7	bbc.co.uk	91,266	32	mashable.com	53,886
8	forbes.com	90,558	33	gizmodo.com	53,474
9	foxnews.com	87,269	34	hindustantimes.com	52,547
10	cnet.com	85,451	35	cnbc.com	52,287
11	chicagotribune.com	83,732	36	azcentral.com	50,441
12	washingtonpost.com	82,105	37	adweek.com	47,683
13	aljazeera.com	79,616	38	csmonitor.com	47,671
14	telegraph.co.uk	79,528	39	thesun.co.uk	47,343
15	cnn.com	78,388	40	edweek.org	46,671
16	dailymail.co.uk	78,275	41	nydailynews.com	45,898
17	indiatimes.com	72,990	42	chron.com	45,762
18	express.co.uk	72,761	43	straitstimes.com	45,167
19	theatlantic.com	69,857	44	dailystar.co.uk	44,804
20	sfgate.com	65,764	45	techcrunch.com	43,577
21	foxbusiness.com	63,544	46	pcmag.com	43,146
22	fool.com	63,499	47	inquisitr.com	43,094
23	rt.com	63,085	48	cbsnews.com	42,850
24	ndtv.com	60,731	49	deadline.com	42,742
25	bbc.com	59,581	50	wired.com	42,379

Table 1. Top 50 Institutions by Article Volume in C4 RealNewsLike.

A.2. `Watershed-Blog`

The original Blog Authorship Corpus (Schler et al., 2006) contains posts from 19,320 bloggers with a total of 681,288 blog posts. This dataset already has a column `id` as a unique identifier for each blogger which nicely suits our needs. Since realistically there are more bloggers than newspapers on the web, we select randomly select 20 unique id's from the dataset. A blogger would also produce fewer content compared to entire newspapers, so we choose 500 blog posts for each blogger, towards a dataset of size 10,000. Blog posts tend to be much shorter compared to news articles, so we also took this into account during the filtering process. We select blog posts with $100 \leq words \leq 1500$.

B. Provenance Methods

We consider both *active* (i.e., watermarking) and *passive* (i.e., MIA) provenance methods in our evaluation. This section details how each method instantiates the (σ_i, s_i) interface of Sec. 2.1.

B.1. Watermarking

All five watermarking schemes are deployed through the data-centric watermarking framework proposed by Lau et al. (2024) where the transformation function σ_i paraphrases each document of $\mathcal{D}_i$ with a watermarking LLM M_w whose token sampling is biased by the scheme under owner i's key k_i, and s_i scores a text by detecting the keyed bias over M_w's tokenizer. Owner keys are distinct seeds, so a single scheme yields N separable signals. We consider the following watermarking schemes:

Waterfall (Lau et al., 2024) permutes the logit vector into a watermarking token space via a pseudorandom permutation π, re-seeded at each step from k_i and the preceding tokens, adds a keyed orthogonal (Fourier) basis function f there, and maps back before sampling: $\ell_v \leftarrow \ell_v + \kappa\, f(\pi(v))$. Verification maps the tokens of a candidate text into the permuted space and computes the inner product of their empirical (cumulative) token distribution $\bar{C}$ with the normalized basis, $q = \langle \bar{C},\, f/\|f\|_2 \rangle$. The perturbation amplitude κ is the strength parameter.

KGW (Kirchenbauer et al., 2023) partitions the vocabulary V at each generation step t into a green list $G_t \subset V$ of size $\gamma|V|$ and its complement, using a pseudorandom function seeded by k_i and a hash of the preceding context. Generation biases the green logits, $\ell_v \leftarrow \ell_v + \delta$ for $v \in G_t$. The detector counts the green tokens N_G among T observed tokens and reports $z = (N_G - \gamma T)/\sqrt{T\gamma(1-\gamma)}$. The logit bias δ is the strength parameter.

Unigram (Zhao et al., 2024) applies the same logit bias $\ell_v \leftarrow \ell_v + \delta$ for $v \in G$, but the green list G is seeded by k_i alone and fixed across all positions; detection uses the same z-test. The context-independence trades some detectability for provable robustness to edits. The strength parameter is again δ.

EXP-Gumbel (Aaronson, 2022) draws key-seeded uniforms $r_{t,v} \in (0,1)$ from a pseudorandom function of $(k_i, \text{context})$ and emits $y_t = \arg\max_v\, r_{t,v}^{1/p_{t,v}}$, where p_t is the model's temperature-τ next-token distribution; marginalizing over keys, this preserves p_t exactly (non-distortionary). The detector aggregates $\sum_t -\log(1 - r_{t,y_t})$, which is inflated when the observed tokens follow the keyed randomness. The sampling temperature τ acts as the strength parameter: higher entropy leaves more room for the key to steer token choices.

SynthID (Dathathri et al., 2024) runs a depth-d tournament over candidate tokens drawn from p_t — conceptually over 2^d leaves, computed in practice as d successive closed-form updates of p_t: at each layer, paired candidates are compared by key-seeded pseudorandom values $g(v, r_t) \in \{0, 1\}$ and the higher-scoring token advances, biasing outputs toward high-g tokens while remaining non-distortionary on average (we use its non-distortionary configuration). The detector reports the mean of g over the observed text. The tournament depth d is the strength parameter.

B.2. Membership Inference Attack (MIA)

MIAs release the data unchanged (i.e., $\sigma_i(\mathcal{D}_i) = \mathcal{D}_i$) and instantiate s_i with loss-based statistics of the audited model on owner i's records, signed so that higher values indicate stronger membership evidence.

We mainly consider **Min-K%++** (Zhang et al., 2025), a state-of-the-art reference-free MIA. For a candidate text $x = (x_1, \ldots, x_T)$ and audited model M, each token receives a calibrated score

$$\text{Min-K\%++}(x_t) = \frac{\log p_M(x_t \mid x_{<t}) - \mu_{x_{<t}}}{\sigma_{x_{<t}}},$$

where $\mu_{x_{<t}} = \mathbb{E}_{v \sim p_M(\cdot|x_{<t})}[\log p_M(v \mid x_{<t})]$ and $\sigma_{x_{<t}}$ are the mean and standard deviation of the model's next-token log-probabilities at that position. This normalization corrects for tokens that are inherently high- or low-probability regardless of training exposure. The sequence score averages the $k\%$ lowest token scores, $s_i(x) = \frac{1}{|S|}\sum_{t \in S} \text{Min-K\%++}(x_t)$ with S the min-$k\%$ subset. The idea is that a text seen during training should have few abnormally improbable tokens, so a higher score indicates membership.

C. Experimental Settings

Compute. We used a cluster of 4 NVIDIA GH200 GPUs and a cluster of 8 NVIDIA H200 GPUs consuming approximately 2,000 GPU-hours in total for the experiments.

Hyperparameters of provenance methods. We follow the default settings from (Pan et al., 2024) for watermarking schemes included in MarkLLM. We use $\delta = 2.0$ for KGW, $\delta = 2.0$ for Unigram, $\tau = 0.7$ for EXP-Gumbel, $d = 30$ for SynthID, and $\kappa = 6.0$ for Waterfall. For Min-K%++, we consider the standard configuration of K = 20 as recommended by (Shi et al., 2024).

Training. For full fine-tuning, we fine-tune the LLMs for 3 epochs at a learning rate of `2e-5`.

Inference-time. When probing the LLM, we issue $K = 10$ audit queries per data owner, each prompting the LLM with a 50-token prefix and letting it generate a continuation for the response. The response is scored under the audited owner's key. We use $n = 300$ evaluation samples per configuration, and model-side results are averaged over 3 evaluation seeds that resample the prompts and regenerate the outputs.

Attacks. Data-side perturbations use a token-drop ratio of 0.1, a synonym-substitution ratio of 0.1, and LLM paraphrasing with Llama-3.1-8B-Instruct (Grattafiori et al., 2024) prompted to rewrite each passage while preserving its meaning. For the LLM paraphraser, we use the following prompt:

```
"Paraphrase the user provided text while preserving semantic similarity. Do not
include any other sentences in the response, such as explanations of the
paraphrasing. Do not summarize. Your paraphrase should be approximately
the same length as the original."
```

D. Experimental Results

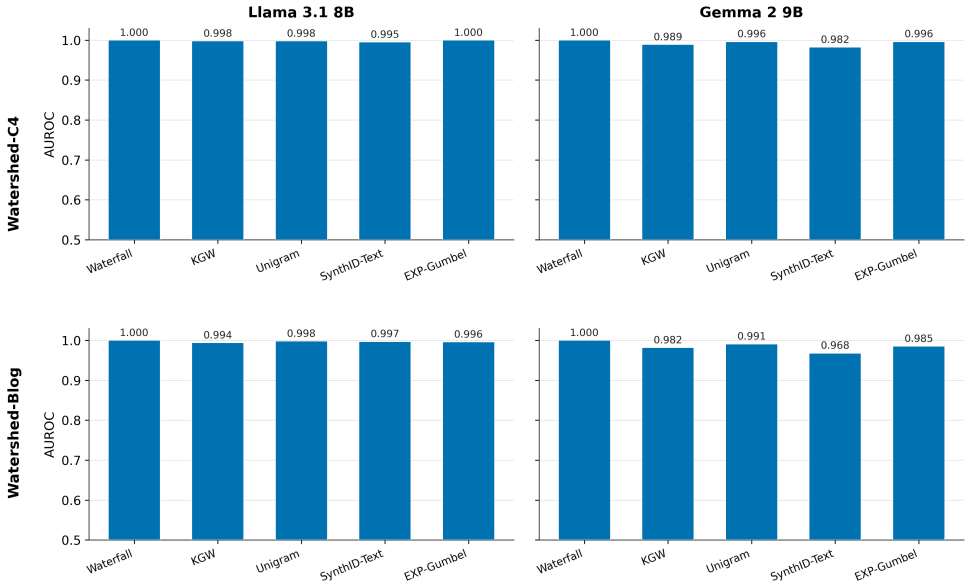

Figure 10. Data-level multi-owner separability: AUROC of each owner's scoring function s_i distinguishing its own prepared data $\widetilde{\mathcal{D}}_i$ (correct key) from other owners' prepared data $\widetilde{\mathcal{D}}_j$, $j \neq i$ (wrong keys), averaged over owners for each watermarking scheme on `Watershed-C4` and `Watershed-Blog`.

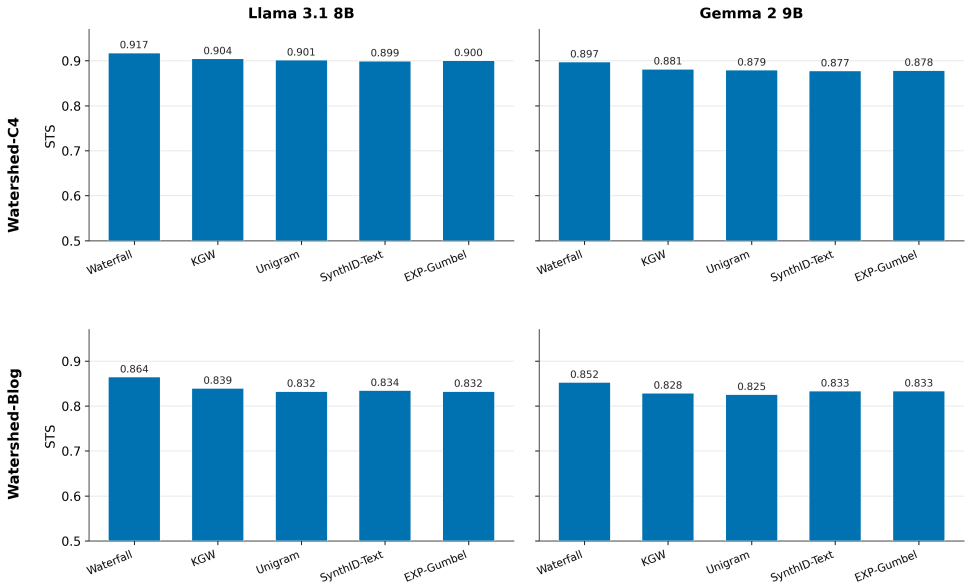

Figure 11. STS scores of the `Watershed-C4` and `Watershed-Blog` datasets between watermarked vs non-watermarked text by various watermarking schemes.

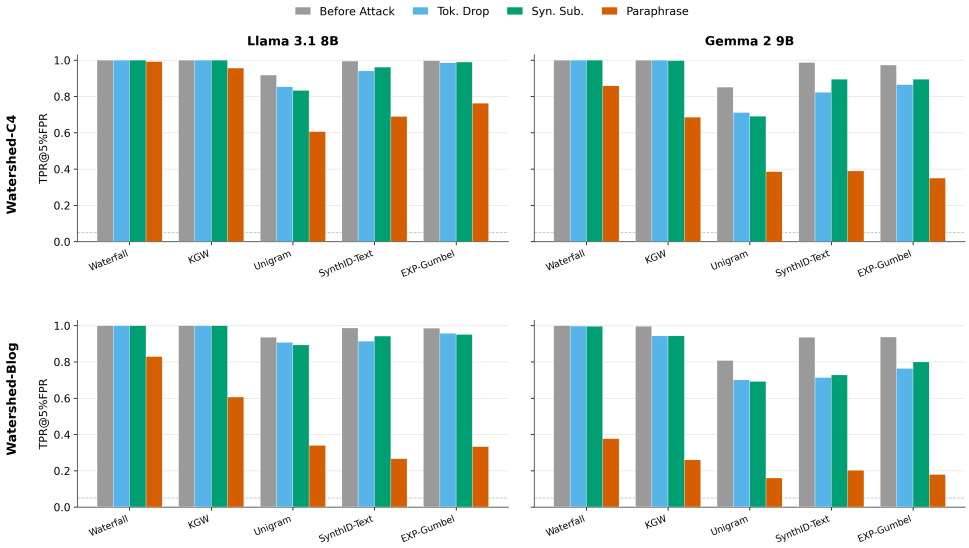

Figure 12. TPR@5%FPR of the `Watershed-C4` and `Watershed-Blog` datasets watermarked by various watermarking schemes before and after attacks (i.e., Token Drop, Synonym Substitution, and Paraphrase).

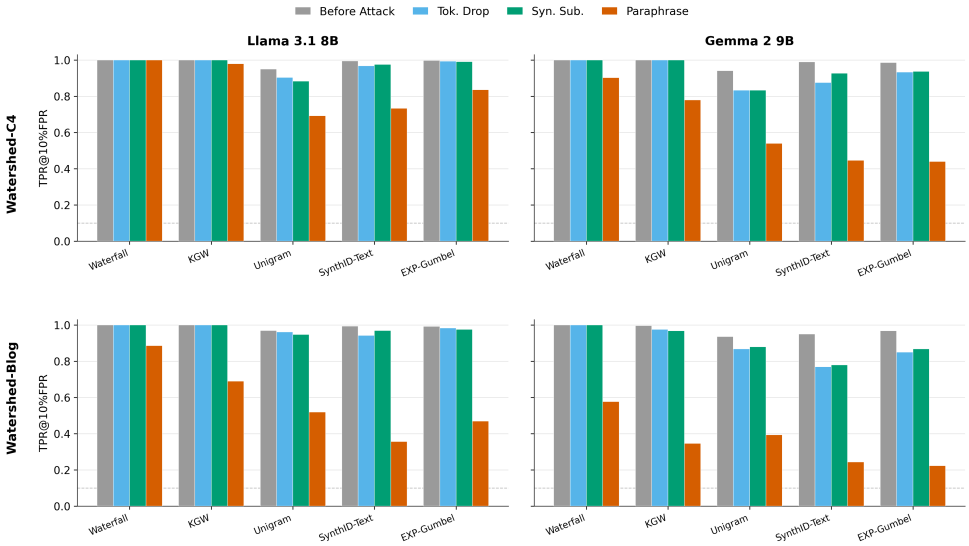

Figure 13. TPR@10%FPR of the `Watershed-C4` and `Watershed-Blog` datasets watermarked by various watermarking schemes before and after attacks (i.e., Token Drop, Synonym Substitution, and Paraphrase).