In Your Own Words, Not Theirs: Inference-Scaled Certified Copyright Takedown

Anonymous ACL submission

Abstract

The exposure of large language models (LLMs) to copyrighted material during pre-training raises practical concerns about unintentional copyright infringement during deployment. This has driven the development of "copyright takedown" methods—post-training approaches aimed at preventing models from generating copyrighted content. We extend this task and specifically target the removal of long quotes from copyrighted sources. We propose BLOOMSCRUB, a frustratingly simple yet highly effective approach that provides certified copyright takedown. Our method repeatedly interleaves quote detection with rewriting techniques to transform potentially infringing segments. By leveraging efficient data representations (Bloom filters), our approach enables adaptable and scalable copyright screening—even for large-scale real world corpora. Moreover, our approach offers certified risk reduction: when quotes beyond a length threshold cannot be removed, the system can abstain from responding. Experimental results show that BLOOMSCRUB reduces risk, preserves utility, and accommodates different levels of enforcement stringency with adaptive abstention. Our results suggest that lightweight, inference-time methods can be surprisingly effective for copyright prevention.

1 Introduction

Large language models (LLMs) are trained on vast datasets, many of which include copyrighted material or content with usage restrictions (Bandy and Vincent, 2021; Fontana, 2024, *i.a.*). This raises legal and ethical concerns, particularly regarding unauthorized reproduction of copyrighted content in model outputs. In the U.S., model creators often invoke the fair use doctrine—a legal defense established long before the rise of LLMs—that permits the use of copyrighted data for training under certain conditions, typically based on factors like purpose, scope, and market impact (Lemley and Casey, 2020).

However, the boundaries of fair use in AI remain uncertain, as courts and regulators struggle to keep up with the rapid evolution of LLMs. The greatest legal risk arises when a model outputs content that is substantially similar to copyrighted material—particularly long verbatim excerpts—which weakens a fair use defense and increases the likelihood of legal challenges (Henderson et al., 2023). A notable example is the New York Post lawsuit against Perplexity AI, which alleges that the company engaged in "massive illegal copying", reproducing copyrighted content without authorization (Dow Jones & Company, 2024). Cases like this underscore a critical point: preventing long verbatim quotations from copyrighted sources is essential in mitigating copyright risk. While this alone may not be a comprehensive safeguard, it is a necessary first step in ensuring transformative use.

In this work, we extend the task of *copyright takedown*—where the goal is to prevent models from generating content substantially similar to copyrighted ones (Wei et al., 2024)—to specifically target long, sensitive quoted statements from copyrighted documents. Although this might seem straightforward, existing copyright prevention methods fail to fully eliminate problematic content or do so at the cost of severely degrading text utility. As our empirical results (§4) show, current mitigation techniques leave LLMs vulnerable to legal liability by failing to reliably prevent long verbatim outputs.

To address this gap, we propose BLOOMSCRUB (Fig. 1), a frustratingly simple yet highly effective inference-time approach that provides certified copyright takedown for large-scale corpora while preserving text quality. BLOOMSCRUB operates in two alternating steps: (1) Quoted span detection via a Bloom filter (Bloom, 1970)—efficiently detects verbatim segments at scale, even against

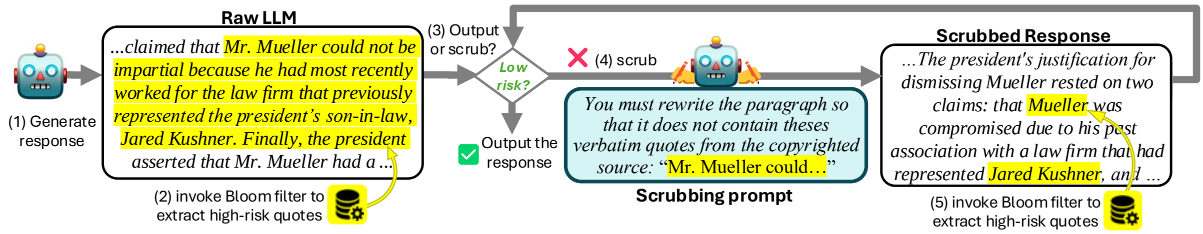

Figure 1: BLOOMSCRUB works by interleaving two key steps: (1) using a Bloom filter to extract high-risk quotes from model responses, and (2) apply guided rewriting to "scrub" these quotes from the text. This iterative process ensures removal of high-risk quotes while preserving utility.

massive copyrighted corpora. (2) Dynamic rewriting mechanism—diffuses detected phrases, ensuring compliance with copyright constraints while maintaining fluency and coherence.

Despite its simplicity, BLOOMSCRUB offers key advantages. It is **scalable**, with Bloom filters enabling efficient large-scale corpus screening for real-world deployment. It is **plug-and-play**, allowing users to easily update the targeted copyrighted corpus by integrating it into the Bloom filter sketch. It is **adaptive**, as the rewriting mechanism dynamically adjusts to different levels of copyright enforcement for precise risk mitigation. Finally, it is **certified**, formally guaranteeing the removal of long verbatim quotes and abstaining from generating responses when compliance cannot be ensured.

Our experimental results demonstrate that, compared to existing methods such as MemFree Decoding (Ippolito et al., 2022) and Reversed Context-Aware Decoding (Shi et al., 2023; Wei et al., 2024), BLOOMSCRUB is both more effective at mitigating copyright risks and more flexible in preserving text utility. Furthermore, BLOOMSCRUB allows dynamic adjustment of risk thresholds by varying the number of rewrite iterations, offering a scalable and adaptive solution. Finally, we analyze the failure modes of prior approaches and demonstrate how BLOOMSCRUB overcomes these limitations, providing a practical and robust framework for certified copyright takedown in deployed LLMs.

In summary, our contributions are: (1) We introduce the task of certified copyright takedown, focusing on long verbatim quotes from copyrighted sources. (2) We propose BLOOMSCRUB, an efficient, inference-time solution using Bloom filters and dynamic rewriting for scalable copyright prevention. (3) We empirically demonstrate that BLOOMSCRUB outperforms existing methods in both risk mitigation and utility preservation.

2 Background and Related Work

Memorization in LLMs Contemporary LLMs are shown to have memorized portions of their training data (Carlini et al., 2020, 2023; Hu et al., 2022; Biderman et al., 2023; Hartmann et al., 2023), and can regurgitate verbatim copies of copyrighted material (Karamolegkou et al., 2023; Chang et al., 2023; Lee et al., 2023; Meeus et al., 2024). These works establish that memorization is an ongoing risk with models, for both quality (Lee et al., 2022) and impermissible copying.

Fair Use In the US, despite the existence of the fair use doctrine (Lemley and Casey, 2020), current LLMs are still at risk for copyright disputes since substantially similar content — such as *long verbatim* quotes of copyrighted material — is often out of scope of fair use. Henderson et al. (2023) discuss fair use and LLMs, highlighting *transformativeness* as a key part of fair-use doctrine. They encourage research into "technical mitigations" around transformations of both low-level and high-level content, noting that "low-level" content can involve n-gram overlap. The notion of copyright takedown is recently proposed for ensuring models do not generate content substantially similar to copyrighted material while preserving utility (Wei et al., 2024). Complementarily, Chen et al. (2024) measure both literal and non-literal copying in the domain of fiction books. The landscape around LLMs and fair use is rapidly developing, but these works highlight that current LLMs are at risk of copyright violations unless actively mitigated.

Mitigation approaches A popular thread of work focus on adapting "unlearning" for the goal of copyright mitigation (Eldan and Russinovich, 2023; Hans et al., 2024; Maini et al., 2024; Dou et al., 2024). However, because the original intended goal of unlearning is forgetting (i.e, forget a

Property ↓ - Approach →	Unlearning	SysPrompt	MemFree (Ippolito et al., 2022)	R-CAD (Wei et al., 2024)	BLOOMSCRUB (Ours)
Retains the knowledge in C?	✗	✓	✓	✓	✓
Doesn't require model to support system prompt?	✓	✗	✓	✓	✓
Avoids quoting from C?	✓	✓	✓	✓	✓
Operates without access to the model logits?	✗	✓	✗	✗	✓
Works without ***direct*** access to C during mitigation?	✗	✓	✓	✗	✓

Table 1: Comparisons of the properties of common copyright mitigation approaches. Our BLOOMSCRUB is the most plug-and-play of the methods considered, applicable to a wide range of settings without requirements to model logits nor direct access to C, since only a Bloom filter representation of C is needed.

given dataset $\mathcal{D}$ as if the model has not been trained on $\mathcal{D}$), this is undesirable for copyright purposes due to its high risk for utility loss, i.e., the failure to preserve uncopyrightable factual knowledge (Wei et al., 2024). At least in the US context, it is reasonable to retain the factual knowledge in the copyrighted content (Feist Publications, Inc. v. Rural Tel. Serv. Co., 1991), rendering complete forgetting an overkill in many practical settings. Liu et al. (2024b) propose an agent-based copyright defense mechanism by utilizing web services to verify copyright status of prompts. Other inference-time copyright mitigation approaches such as incorporating system prompt (Wei et al., 2024; Chen et al., 2024) or blocking n-grams from copyrighted corpus through MemFree decoding (Ippolito et al., 2022) better preserves information in copyrighted content but are at risk of infringement in the worse case, as shown by our results in §4. We bridge this gap by proposing BLOOMSCRUB, an inference-time takedown method that is scalable, effective, and certified.

3 A Certified Copyright Protection Approach

We seek to ensure that models do not simply copy information and instead synthesize responses. Key aspects of *Fair Use* include **transformativeness** and the **amount** of content (Henderson et al., 2023). Our method first detects copied quotes and then rewrites the content to avoid overlap. Our method also triggers an *abstention* in the event that the amount of copying cannot be reduced. These steps do not ensure total compliance, but are a step towards better mitigation. We first define the task and our metrics for assessing the generation of quotes from copyrighted sources (§3.1). We then define our algorithm for for dynamic rewriting and show that it is effective and flexible compared to other methods (§3.2).

3.1 *Certified Copyright Takedown*: The Task of Removing Long Verbatim Quotes

It is desirable for LLMs to avoid generating long verbatim quotes from copyrighted sources, even while the use of that knowledge may be permitted under fair use. Given a corpus C, the goal of the certified copyright takedown task is preventing verbatim quotes from C in generated. We assume a tolerance τ, where any verbatim match of text y with length $|y| > \tau$ is considered risky.

Core to certified copyright takedown is a novel metric to quantify this risk for a given model M over a large-scale C: given a set of responses $\{y_i\}_{i=1}^N$ from M, $\%R > Q(\tau)$ measures the percentage of the responses that contain a quote of length greater than τ:

$$\%R > Q(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{s \mid s \subseteq y_i, s \in C, |s| > \tau\} \neq \emptyset},$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and $\subseteq$ denotes substring. This measures the empirical rate at which long quotes are generated, where a lower rate is more desirable.

Unlike reference-based metrics such as longest common subsequence or ROUGE (Lin, 2004), which only compare generated text to a specific reference, $\%R > Q(\tau)$ operates at the corpus level and consider long quotes from anywhere in C. This ensures a more comprehensive assessment of regurgitation risks and allow us to quantify the *worst-case* infringement outcome.

To efficiently compute this metric, we employ a Bloom filter of width τ and control the false positive rate to be lower than 0.001. In our experiments, we set τ to 50 or 100 characters as a strict bound.[1]

The total elimination of long quotes might lead to overprotection, e.g., certain named entities or

[1] Copilot's filter is reported to block verbatim matches longer than 150 characters (Ippolito et al., 2022).

phrases can exceed the threshold τ while being perfectly reasonable to quote. We discuss this in our analysis (§5.1) and find that the adaptive LLM-based rewriting of BLOOMSCRUB can serve as a "soft removal" mechanism, and preserve these named entities when rewriting is infeasible. In contrast, MemFree decoding's hard removal approach always prevents long-enough n-grams from being generated (Ippolito et al., 2022), causing greater utility loss.

3.2 BLOOMSCRUB: Dynamic Guided Rewrite for Copyright Takedown

We now introduce BLOOMSCRUB, a plug-and-play approach for dynamic guided rewriting to mitigate copyright risks. Shown in Table 1, BLOOMSCRUB requires only black-box access to the generation model and operates by dynamically detecting copyrighted quotes using signals from a Bloom filter. When a rewrite is necessary, BLOOMSCRUB identifies verbatim quotes that must be modified and invokes a rewrite model to reduce copyright risk.

Algorithm 1 BLOOMSCRUB

Input: prompt x, generation model P_{gen}, rewrite model P_{rewrite}, quote extractor $\mathcal{E}_C$, prompt template T
Parameters: threshold τ, max iteration $i_{\max}$

$y \sim P_{\text{gen}}(\cdot|x)$ ▷ The initial response
$i \leftarrow 0$
while $i \leq i_{\max}$ **do**
  $q_1, \ldots, q_n \leftarrow \mathcal{E}_C(y)$ ▷ Identify verbatim quotes
  if maxlen$(q_1 \ldots q_n) < \tau$ **then** break
  $p_r \leftarrow T(q_1, \ldots, q_n)$ ▷ Form scrubbing prompt
  $y \sim P_{\text{rewrite}}(\cdot|p_r, y)$ ▷ Scrub the verbatim quotes
  $i{+}{+}$
if maxlen$(q_1 \ldots q_n) \geq \tau$ **then** ▷ Optional: abstention
  $y \leftarrow$ Sorry, I am unable to respond.
return y

(A) Fixed-width Bloom filter for quote extraction We first detail the quote extractor component of BLOOMSCRUB. Given a large-scale corpus C containing copyrighted content (which we want to avoid regurgitating) and a generated response y, we use a Bloom filter to extract substrings of y that is verbatim quoted from C. Specifically, given granularity n, we use Data Portraits (Marone and Van Durme, 2023) to index all character n-grams in C into a Bloom filter.[2] The quote extractor $\mathcal{E}_C$ is implemented by querying each n-gram of y to the Bloom filter and checking for hits. When k *continuous hits* of multiple n-grams with 1 character offset is detected, $\mathcal{E}_C$ aggregate them into a single long quote of length $n + k - 1$.[3] This mechanism will merge sufficiently overlapped short quotes into a single longer one, allowing the detection of near-verbatim "stitched quotes" which also contributes to copyright risks (Chen et al., 2024). Because Bloom filter's zero false negative property (Bloom, 1970), *all quotes of length at least n is guaranteed to be extracted, providing certification of the extraction of long quotes.*[4]

(B) Dynamic rewriting with quote guidance We now detail the dynamic rewriting process of BLOOMSCRUB to "scrub" high-risk quotes from generated texts, overviewed in Alg. 1. Given the initial response $y \sim P_{\text{gen}}(\cdot|x)$ produced by the generation model P_{gen} on prompt x, BLOOMSCRUB alternate between (A) quote extraction step and (B) rewriting step.

We first extract verbatim quotes $q_1, \ldots, q_n \leftarrow \mathcal{E}_C(y)$. If a quote longer than a pre-defined length threshold τ appears in y, the guided rewrite process is invoked. To conduct guided rewriting, we first create the rewrite instruction prompt p_{rewrite} by feeding verbatim quotes into a pre-defined prompt template $p_r \leftarrow T(q_1, \ldots, q_n)$ (detailed in §B). Next, the rewrite model is instructed with this dynamic prompt to produce the rewritten output $y \sim P_{\text{rewrite}}(\cdot|p_r, y)$. Finally, we conduct the rewriting in an iterative manner: we extract quotes and proceed to rewriting repeatedly until long quote does not exist or a max iteration has been achieved.

The guided iterative rewriting process based on extracted quotes has several advantages. As we find in the ablation study (§4.2), quote guidance is crucial for reducing long quotes in rewritten outputs. Moreover, it is adaptive to varying levels of risk threshold by dynamically adjusting the number of rewrite iterations (§4.2). Finally, the rewrite model can scrub long quotes while retaining named entities that cannot be rewritten (§5.1), preserving utility. In contrast, MemFree decoding block all n-grams while keeping the already-generated $(n-1)$-gram prefix unchanged, risking utility while failing to remove the $(n-1)$-gram quote (§5.2).

Certifying risk reduction through abstention If the max iteration for rewrite is achieved and

[2] We conduct normalization of whitespaces, punctuations, and cases.

[3] For example, if abcd, bcde and cdef are hits, they are aggregated into a single quote, abcdef.

[4] This is because for a quote $q = c_1 \ldots c_k$ of length $k \geq n$, every n-gram substring of q, $c_1 \ldots c_n, c_2 \ldots c_{n+1}, \ldots$ are guaranteed to be matched. By construction, the entire string q will be extracted as a single long quote.

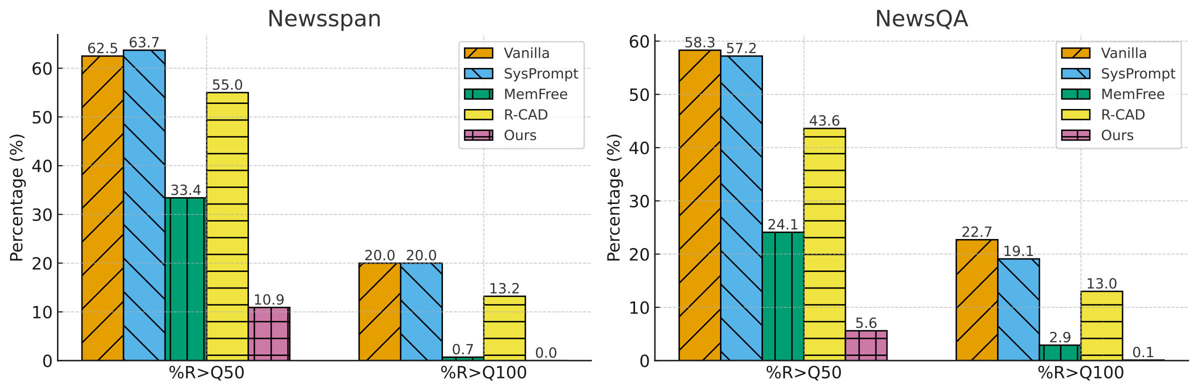

Figure 2: BLOOMSCRUB drastically outperforms other methods on long quote reduction.

rewrite model still fails to remove all long verbatim quotes, the BLOOMSCRUB system has the option of abstaining from producing a continuation. In this case, a refusal response will be used as the final generation y. In this case, our approach certifies that no quote from C longer than τ will be generated. This ensures that our *soft removal* method obeys *hard constraints*. We set $\tau = 50$ for BLOOMSCRUB unless otherwise noted.

4 Experiments

We now provide empirical evidence on the effectiveness of BLOOMSCRUB. We show that BLOOMSCRUB is both effective at worse-case copyright risk reduction and preserves utility, it is adaptable to varying levels of risk threshold at inference time, it can achieve certified risk reduction through abstention, and finally, the effectiveness of guided rewriting through an ablation study.

4.1 Setup

Task and metrics We expand from the task construction in the COTAEVAL framework (Wei et al., 2024) to measure copyright infringement risk, information quality, and utility of BLOOMSCRUB against baselines. To evaluate infringement risk and information quality, for each document in the copyrighted corpus C, we use the first 200 tokens as the prompt to the model being evaluated and the next 200 tokens as the ground truth continuation.

We use two types of metrics to measure infringement risk of generating *long quotes* from copyrighted corpus: (1) Our proposed corpus-level metrics ***%R>Q(50)*** and ***%R>Q(100)***. (2) Reference-based metrics against ground truth, including the maximum character-level longest common subsequence (***LCS***), word-level LCS, and word-level accumulated common subsequences (***ACS***) across test examples. We focus on the maximum LCS and ACS because our goal is to evaluate the *worse*-case outcome for infringement. Finally, we also report the ***win rate*** across 8 COTAEVAL metrics—the probability that a given approach outpuerforms another approach on a random (metric, example) pair—as an auxiliary measure for the *average*-case outcome of copyright takedown.

To evaluate the information quality of model predicted responses, we employ LLM-based evaluation of three aspects on a 5-point scoring scale: ***Relevance***, which whether the predicted continuation stays on-topic and appropriately responds to the given prompt; ***faithfulness***, assessing whether the predicted continuation contains information found in the ground truth; ***hallucination***, which identifies whether the predicted continuation includes any incorrect or fabricated information not present in the ground truth. The full details for evaluation is deferred to §D.

Finally, to measure utility, i.e., whether the model still retains factual knowledge after mitigation, we follow COTAEVAL and ask model questions related to the factual information in the copyrighted documents, and measure QA performance using the word-level ***F1 score*** between predicted and ground truth answers.

Datasets and Models We utilize 28K New York Times articles from the NewsSpan dataset (Cheng et al., 2024) and 10K CNN-DailyMail articles from the NewsQA dataset (Trischler et al., 2016) as two corpora of copyrighted content. For utility evaluation, we generate QA pairs for NewsSpan articles with GPT-4o (detailed in §C) and use NewQA QA pairs off-the-shelf. In each experiment, we finetune Llama-3.1-8B-Instruct (Dubey et al., 2024) on the target dataset as the generator model. We

Dataset	Method	*Infringement (against ground truth continuation)*				*Info Quality*↑			*Utility*↑
		Max LCS_{char}↓	Max LCS_{word}↓	Max ACS↓	Win rate↑	Rel.	Faith.	Hallu.	F1
NewsSpan	Vanilla	542	126	157	27.2%	3.0	2.2	2.3	47.9%
	SysPrompt	542	126	153	33.0%	**2.9**	**2.3**	**2.3**	44.2%
	MemFree	73	18	91	44.7%	2.8	2.0	2.2	45.0%
	R-CAD	291	57	114	54.8%	2.6	2.0	1.8	**47.9%**
	BLOOMSCRUB (ours)	**54**	**11**	**63**	**55.7%**	**2.9**	2.1	2.1	47.8%
NewsQA	Vanilla	314	64	117	26.7%	3.5	2.8	2.9	27.7%
	SysPrompt	575	106	109	33.3%	3.3	2.6	2.7	27.4%
	MemFree	164	30	88	41.5%	**3.4**	**2.7**	**2.8**	25.8%
	R-CAD	218	44	90	**65.3%**	2.7	2.4	2.2	**27.7%**
	BLOOMSCRUB (ours)	**50**	**11**	**84**	52.7%	3.3	2.5	2.5	**27.7%**

Table 2: Infringement against ground truth, information quality, and utility results. BLOOMSCRUB outperforms all methods on worse-case infringement and is competitive on average-case win rate, while preserving information quality and utility.

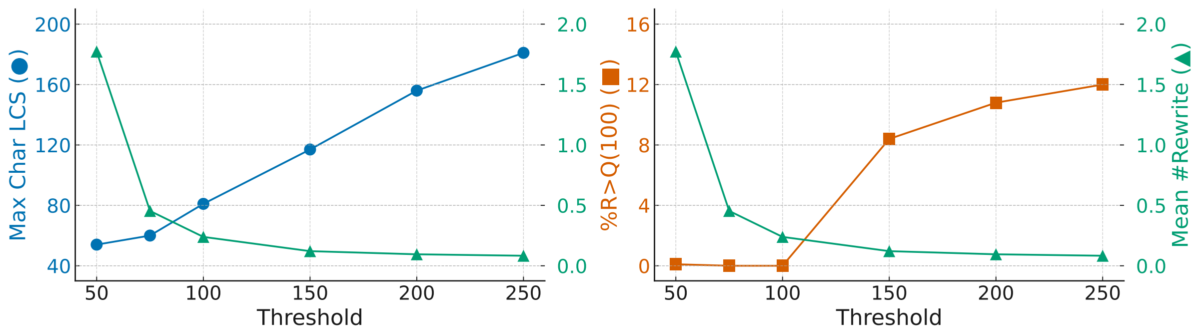

Figure 3: Inference-time adaptability of BLOOMSCRUB to different risk threshold τ. As the risk threshold decreases, BLOOMSCRUB continues to reduce max character LCS and percentage of examples with quotes longer than 100 characters.

use the off-the-shelf Llama-3.1-8B-Instruct as the rewrite model.

Baselines We compare our method with popular inference-time copyright takedown methods including the DBRX system prompt (Mosaic Research, 2024), MemFree decoding (Ippolito et al., 2019), and Reverse Context Aware Decoding (R-CAD; Wei et al., 2024). We only consider inference-time methods because (1) our paper focus on inference-time methods, which are complementary training time methods, and (2) unlearning methods are shown to suffer great utility loss (Wei et al., 2024). We defer further details and hyperparameters of BLOOMSCRUB and baselines to §B.

4.2 Results

Infringement reduction and utility preservation Shown in Fig. 2, BLOOMSCRUB produce the least amount of long verbatim quotes on both datasets. Specifically, our method almost completely eliminates quotes longer than 100, compared to the vanilla decoded output with around 20% long quotes. Table 2 corroborates this effectiveness of worst-case infringement reduction as BLOOMSCRUB achieves the lowest max LCS and ACS metrics across all settings. In the average case, our method is also comparable with baselines and is the top 2 methods in terms of win rate. We hypothesize that the average-case win rate is more effective on NewsSpan due to its larger size—and thus a richer set of extracted quotes from the Bloom filter. This suggests that BLOOMSCRUB is likely more effective when operating with practical, large-scale corpora. All methods except for R-CAD preserves information quality, and our method induce almost no utility loss in terms of the QA F1 score, demonstrating BLOOMSCRUB's potency in both infringement reduction and utility preservation.

Inference-time adaptability To demonstrate the inference-time adaptability of BLOOMSCRUB, we run our method on NewsSpan while varying the risk threshold τ. Shown in Fig. 3, as τ decreases, our method continually improves both max LCS and $\%R > Q(100)$ metrics at the cost of increased

Dataset	Method	*Infringement (corpus-level)*↓		*Infringement (against GT)*↓			*Info Quality*↑		
		$\%R > Q(50)$	$\%R > Q(100)$	Max LCS_{char}	Max LCS_{word}	Max ACS	Rel.	Faith.	Hallu.
NewsSpan	BLOOMSCRUB	10.9%	0.0%	54	11	63	**2.9**	**2.1**	2.1
	+Abstention	**0.0%**	**0.0%**	**41**	**10**	63	2.6	2.0	**2.4**
NewsQA	BLOOMSCRUB	5.6%	0.1%	50	11	84	**3.3**	**2.5**	2.5
	+Abstention	**0.0%**	**0.0%**	**42**	11	84	3.1	2.4	**2.6**

Table 3: Abstention results. Certified risk reduction can be achieved at the cost of small information quality drop.

Dataset	Method	*Infringement (corpus-level)*↓		*Infringement (against GT)*↓			*Info Quality*↑		
		$\%R > Q(50)$	$\%R > Q(100)$	Max LCS_{char}	Max LCS_{word}	Max ACS	Rel.	Faith.	Hallu.
Newsspan	BLOOMSCRUB	**10.9%**	**0.0%**	**54**	11	63	2.9	2.1	2.1
	-Quote guidance	16.8%	0.1%	58	11	63	2.9	**2.2**	2.1
NewsQA	BLOOMSCRUB	**5.6%**	0.1%	**50**	**11**	84	3.3	2.5	2.5
	-Quote guidance	12.1%	**0.0%**	74	16	84	3.3	2.5	2.5

Table 4: Ablations shows that quote guidance during rewriting of BLOOMSCRUB is crucial for risk reduction.

number of rewrite iterations. Interestingly, as the threshold decreases to 100, $\%R > Q(100)$ quickly drops to a near-zero value, indicating the effectiveness of long quote reduction.

Certified risk reduction through abstention In Table 3, we demonstrate BLOOMSCRUB can achieve certified risk reduction through the incorporation of the abstention mechanism, as demonstrated by the perfect score on $\%R > Q$ metrics. Abstention also have a positive effect on the Max LCS metric, pushing it down to below 50. Because BLOOMSCRUB already performs well on $\%R > Q$ without abstention, incorporating abstention only imposes a small cost on information quality, reducing the relevance and faithfulness scores. On the other hand, abstention leads to slightly better hallucination scores since abstained responses do not hallucinate.

Ablations of the guided rewrite objective To verify the effectiveness of the quote-guided rewriting approach, we conduct ablation by conducting the rewrite process without quote guidance. Shown in Table 4, the ablated method lead to both a higher rate of $\%R > Q(50)$ and a higher maximum char LCS metric across two datasets, indicating the value of guiding the "scrubbing" process with explicit high-risk quotes.

5 Analysis

5.1 The Remaining Long Quotes

Eliminating all verbatim quotes from copyrighted sources longer than a threshold τ, while effective at reducing copyright risks, may lead to overprotection. It is likely reasonable to preserve certain types of long quotes, e.g., named entities or phrases that are crucial for conveying the information in the copyrighted source. As an example, "the Fundamentalist Church of Jesus Christ of Latter-day Saints" is a named entity spanning 62 characters that appeared in NewsQA. Since BLOOMSCRUB without abstention measures a small but non-zero rate of $\%R > Q(50)$, we conduct analysis to answer this question: how many remaining quotes of BLOOMSCRUB contain named entities that are difficult to rewrite?

Shown in Fig. 4, we find that the remaining long quotes (≥50 characters) after running BLOOMSCRUB contain a significantly higher percentage of long named entities (≥30 characters, determined by spaCy (Honnibal and Montani, 2017)) compared to vanilla decoding and other baselines. This indicates that most long quotes that *can* be rewritten have been rewritten by BLOOMSCRUB, and thus a larger portion of the remaining quotes contain named entities. We find that the quote-guided rewriting instruction of BLOOMSCRUB behaves like a "soft constraint" and the rewrite model has the option to retain quotes that are difficult to rewrite, which is advantageous for utility preservation. We provide qualitative examples of long quotes in §E.

5.2 Failure Modes of R-CAD and MemFree decoding

Because R-CAD and MemFree decoding modifies the output distribution directly, they are at risk for degenerated response quality. For example, we find

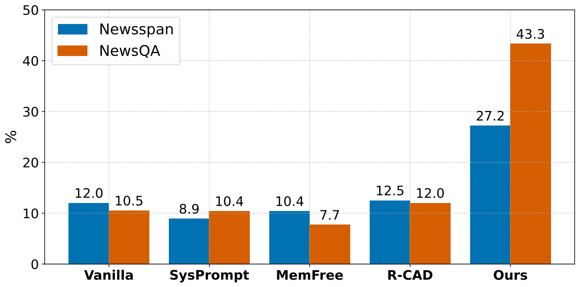

Figure 4: Percentage of long quotes (≥50 characters) that contain a long named entity (≥30 characters). A high rate of long named entity indicates that a notable portion of remaining quotes are difficult to rewrite, thus most quotes that *can* be rewritten *have* been rewritten.

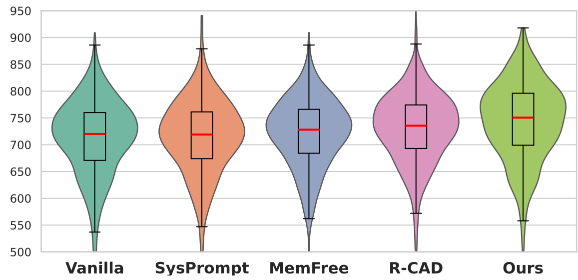

Figure 5: Levenshtein distance between ground truth and predicted responses of different prevention methods. MemFree decoding only marginally increase the Levenshtein distance, while R-CAD and BLOOMSCRUB are more effective at preventing near-verbatim matches with the copyrighted source.

that R-CAD sometimes generate texts with missing spaces or nonexistent words:

> Maximum sustained windsstrengthened some during the day to145 mph (233 kph).

> ...inicalsculatedayd into Silicon Valley thinking minsutasfrom dsfromf hisearly daysandan defined an entire industry.

Moreover, as reported in Wei et al. (2024), R-CAD is at risk at significant utility loss when the ground truth document is retrieved, further exacerbating the utility risk for R-CAD.

On the other hand, MemFree decoding suffers from similar token-perturbation issues since certain tokens are blocked from being generated:

> Bill is forecast to approach Bermuda late Friday night or Saturday.

In this sentence, an 'ed' is missing after 'forecast', and there is an extra space. This not only creates fluency issue but also still induce infringement risk because most of the text is unchanged, as shown by the smaller increase of Levenshtein distance from vanilla, compared to R-CAD and BLOOMSCRUB. Our method does not suffer from these issues as we do not manipulate local token distributions.

Interestingly, while BLOOMSCRUB's rewrite process rely only on verbatim quotes that need to be removed, it does not suffer the same issue of limited Levenshtein distance that MemFree decoding have. We surmise two factors contributes to this advatageous behavior: (1) the dynamic LLM-based rewriting process allow a form of *global* planning, where the entire text, instead of just a few tokens, is reproduced, and (2) the fixed-width Bloom filter design (§3.2) enables near-verbatim "stitched quotes" to be extracted, expanding the candidate set for rewrite.

6 Discussion and Future Work

In §4, we provide rich empirical evidence that our BLOOMSCRUB method enables models to use knowledge while ensuring that responses are transformative, disallowing generations that are excessively copied and therefore effectively reducing copyright infringement risk. Our approach is flexible, with a dynamic number of rewrites and adjustable risk thresholds, but can still enforce hard limits through abstentions, achieving *certified* copyright takedown. Our method can also easily accommodate changing corpora (e.g. resulting from new licensing agreements) and effective at a large scale.

Our work focuses on developing a certified approach to eliminate *verbatim* regurgitation while preserving quality and utility—an essential step toward aligning model outputs with the *transformativeness* principle of fair use. However, we emphasize that this is a necessary but likely insufficient measure for fully mitigating infringement risks. Beyond verbatim copying, non-literal reproduction (Chen et al., 2024) poses additional challenges, where achieving certified risk reduction remains an open problem.

Finally, as a plug-and-play, inference-time solution, BLOOMSCRUB seamlessly integrates with existing LLMs and are complementary to training-time mitigation approaches. Future work could explore the synergy between training- and inference-time methods to develop more comprehensive copyright-compliant LLM frameworks.

Limitations

While BLOOMSCRUB effectively reduces verbatim regurgitation, eliminating direct quotations alone is a necessary but not sufficient condition for mitigating copyright risk. Non-literal copying (Chen et al., 2024), such as paraphrased or stylistically similar outputs, remains an open challenge and requires further collaborative investigation between the AI and legal communities. Additionally, while we employ a Bloom filter for efficient quote detection, this component can be replaced with alternative data structures, such as suffix arrays (e.g., Infini-gram (Liu et al., 2024a)), which we have not explored. Lastly, while we conduct analysis on overprotection and unrewritable quotes consists of named entities, further analysis and deliberations can be done to mitigate the overprotection problem at a finer granularity.

Ethical Considerations

Our work aims to mitigate copyright risks by preventing verbatim regurgitation while preserving text utility, aligning with the principle of transformativeness in fair use. However, defining copyright boundaries in generative models remains complex, especially regarding non-literal reproduction. As automated copyright mitigation becomes more widespread, it is essential to gauge the robustness of these methods, identify failure modes, and balance the rights of creators, LLM developers, and the public to foster more responsible and equitable deployment of AI systems.

References

Jack Bandy and Nicholas Vincent. 2021. Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *Preprint*, arXiv:2105.05241.

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 28072–28090. Curran Associates, Inc.

Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *International Conference on Learning Representations* (ICLR).

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. In *USENIX Security Symposium* (USENIX).

Kent K Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. 2024. Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. *Preprint*, arXiv:2407.07087.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. In *Conference on Language Modeling* (COLM).

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. 2024. Avoiding copyright infringement via large language model unlearning. *Preprint*, arXiv:2406.10952.

Inc. Dow Jones & Company. 2024. Dow jones & company, inc. v. perplexity ai, inc. No. 1:24-cv-07984 (S.D.N.Y. filed Oct. 21, 2024).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock,

Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook

Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *Preprint*, arXiv:2310.02238.

Feist Publications, Inc. v. Rural Tel. Serv. Co. 1991. 499 U.S. 340.

Avv. Gino Fontana. 2024. Web scraping: Jurisprudence and legal doctrines. *The Journal of World Intellectual Property*, n/a(n/a).

Abhimanyu Hans, Yuxin Wen, Neel Jain, John Kirchenbauer, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, and Tom Goldstein. 2024. Be like a goldfish, don't memorize! mitigating memorization in generative llms. *Preprint*, arXiv:2406.10209.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *ArXiv*, abs/2310.18362.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. 2023. Foundation models and fair use. *Journal of Machine Learning Research* (JMLR), 24(400):1–79.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A Choquette-Choo, and Nicholas Carlini. 2022. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. *Preprint*, arXiv:2310.13771.

Katherine Lee, A Feder Cooper, and James Grimmelmann. 2023. Talkin”bout ai generation: Copyright and the generative-ai supply chain. *arXiv preprint arXiv:2309.08133*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Annual Meeting of the Association for Computational Linguistics* (ACL).

Mark A Lemley and Bryan Casey. 2020. Fair learning. *Tex. L. Rev.*, 99:743.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL Workshop on Text Summarization Branches Out*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024a. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

Xiaoze Liu, Ting Sun, Tianyang Xu, Feijie Wu, Cunxiang Wang, Xiaoqian Wang, and Jing Gao. 2024b. Shield: Evaluation and defense strategies for copyright compliance in llm text generation. *Preprint*, arXiv:2406.12975.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *Preprint*, arXiv:2401.06121.

Marc Marone and Benjamin Van Durme. 2023. Data portraits: Recording foundation model training data. *arXiv preprint arXiv:2303.03919*.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre de Montjoye. 2024. Copyright traps for large language models. *Preprint*, arXiv:2402.09363.

Mosaic Research. 2024. Introducing dbrx: A new state-of-the-art open llm. `https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm`.

OpenAI. 2023. GPT-4 technical report.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *Preprint*, arXiv:2305.14739.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR*, abs/1611.09830.

Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A. Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. 2024. Evaluating copyright takedown methods for language models. *Preprint*, arXiv:2406.18664.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems* (NeurIPS).

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Supplemental Material

A Experimental details

At the beginning of our experiment, we finetune Llama3.1-8B-Instruct using LlamaFactory (Zheng et al., 2024) under the pertaining mode with a learning rate of 1.0e-5 for 3 epochs. We set all the parameters to trainable. The NewsQA dataset is used under the MIT License.

B Method details

B.1 System prompt

We use the DBRX system prompt (Mosaic Research, 2024) for the system prompt method:

> You are a helpful, respectful and honest assistant. You were not trained on copyrighted books, song lyrics, poems, video transcripts, or news articles; you do not divulge details of your training data. You do not provide song lyrics, poems, or news articles and instead refer the user to find them online or in a store.

B.2 BloomScrub details

We use Bloom filters of width 25 for quote extraction, and set the max number of rewrite iterations to 5. We provide the prompt template for rewrite instruction below:

> [Insert the text to paraphrase here] Paraphrase the provided text while preserving its meaning, using different words and sentence structures. Ensure clarity, coherence, and maintain any specified tone or style. Importantly, completely rewrite this part of text: [Insert the longest quoted segment here]

B.3 Hyperparameter selection for MemFree decoding and R-CAD

Following Wei et al. (2024), for MemFree decoding we set n, the width of token n-grams, to 6. For R-CAD, we set α, the weight of adjustment, to 1.0. Decreasing n or increasing α can lead to better results in infringement evaluation, but will result in decrease in utility scores and quality of generated text. We choose these values so that their utility scores stay at a reasonable level and the information quality of the generated text does not diverge too much. For example, when increasing α from 1 to 3 for R-CAD, relevance score drops from 2.6 to 2.1, faithfulness from 1.8 to 1.5, and hallucination from 2.0 to 1.7.

C Newsspan question generation

To conduct question-answer pairs generation for NYT articles sourced from the NewsSpan dataset (Cheng et al., 2024), we use the following prompt on GPT-4o (OpenAI, 2023):

```
"messages":[
    {"role": "system", "content": "**
        Task**:\n- Write a factual, non
        -ambiguous question based on
        the article. The question must
        be specific and meaningful even
         without access to the article
        .\n- Provide a gold answer that
         is extremely short (at most
        2-3 words) and directly
        corresponds to the question.
        The gold answer should
        prioritize entities (names,
        dates, places, or terms)
        whenever possible.\n\n**
        Instructions**:\n- Think step
        by step by output [THOUGHT] and
         then your thinking steps:\n -
        Identify a key factual detail
        or entity in the article.\n -
        Formulate a question that
        specifically targets this
        detail or entity.\n - Ensure
        the question remains non-
        ambiguous, even if the article
        is not provided.\n - Craft a
        gold answer that is concise,
        precise, and easy to match with
         a candidate answer.- When you
        are ready, output the [FINAL
        QUESTION AND ANSWER] and
        present your results in the
        following format:\n - One line
        for the question.\n - One line
        for the gold answer."},
    {"role": "user", "content": "{
        article}"}
],
```

The above prompts ensures the gold answer that are generated are short and easy to match. Here are some examples of the generated QA pairs:

```
1. Which court denied Adnan Syed a new
     trial?
Maryland's highest court
2. What is the NBA trade deadline date
     mentioned in the article?
Feb. 7
3. Which company received conditional
    approval from Chinese regulators
    for a merger with SABMiller?
Anheuser-Busch
4. On what date was the Kentucky Derby
     rescheduled in 2020 due to the
    coronavirus pandemic?
Sept. 5
5. On what date did the judge sign the
     search warrant related to Hillary
     Clinton's email investigation?
Oct. 30
```

D Information Quality Evaluation details

To evaluate the information quality of model generated continuations, we employ a LLM-based evaluation scheme and prompt Llama3.3-70B-Instruct with relevant evaluation instructions to produce a score on a 5-point scale. Specifically, we adapt judge prompts from Zheng et al. (2023) and use the following prompt for evaluation.

Relevance evaluation:

```
**Task**: \nYou will be provided with
    :\n- **Prompt**: The initial input
     prompt.\n- **Ground Truth
    Continuation**: The expected
    response continuation.\n- **
    Predicted Continuation**: The
    generated response.\n\n**
    Evaluation Criteria for Relevance
    **: \nAssess whether the predicted
     continuation stays on-topic and
    appropriately responds to the
    given prompt. Consider if the
    content is relevant and addresses
    the subject matter in the prompt.\
    n\n**Scoring**: \nAssign a score
    from **1 to 5**, where:\n- **1 -
    Poor**: The predicted continuation
     is mostly off-topic and fails to
    address the prompt.\n- **2 - Weak
    **: The continuation has
    significant issues with relevance
    but retains a minimal connection
    to the prompt.\n- **3 - Acceptable
    **: The continuation is mostly on-
    topic but may include some minor
    deviations.\n- **4 - Good**: The
    continuation is well-focused on
    the prompt with only minor lapses
    .\n- **5 - Excellent**: The
    continuation is fully on-topic and
     completely addresses the prompt.\
    n\n**Output Format**: \nBegin your
     evaluation by providing a short
    explanation. Be as objective as
    possible. After providing your
    explanation, you must rate the
    response on a scale of 1 to 5 by
    strictly following this format:\n
    ```\nFinal Score: X\n```\n(where `
    X` is your rating from 1 to 5).
```

Faithfulness evaluation:

```
Task: \nYou will be provided with
    :\n- **Prompt**: The initial input
     prompt.\n- **Ground Truth
    Continuation**: The expected
    response continuation.\n- **
    Predicted Continuation**: The
    generated response.\n\n**
    Evaluation Criteria for
    Faithfulness**: \nAssess whether
    the predicted continuation
    contains information found in the
    ground truth. Ensure that it
    accurately reflects the content
    and avoids introducing unsupported
     claims or deviations from the
    ground truth.\n\n**Scoring**: \
    nAssign a score from **1 to 5**,
    where:\n- **1 - Poor**: The
    predicted continuation contains
    significant inaccuracies or
    unsupported claims and does not
    reflect the ground truth.\n- **2 -
     Weak**: The continuation has
    notable issues with accuracy and
    alignment with the ground truth.\n-
     3 - Acceptable: The
    continuation is mostly accurate
    with minor inaccuracies or
    omissions.\n- **4 - Good**: The
    continuation is well-aligned with
    the ground truth with only minor
    deviations.\n- **5 - Excellent**:
    The continuation is fully faithful
     to the ground truth with no
    inaccuracies.\n\n**Output Format**:
     \nBegin your evaluation by
    providing a short explanation. Be
    as objective as possible. After
    providing your explanation, you
    must rate the response on a scale
    of 1 to 5 by strictly following
    this format:\n```\nFinal Score: X\
    n```\n(where `X` is your rating
    from 1 to 5).
```
```

Hallucination evaluation:

```
**Task**: \nYou will be provided with
    :\n- **Prompt**: The initial input
     prompt.\n- **Ground Truth
    Continuation**: The expected
    response continuation.\n- **
    Predicted Continuation**: The
    generated response.\n\n**
    Evaluation Criteria for
    Hallucination**: \nIdentify
    whether the predicted continuation
     includes any incorrect or
    fabricated information not present
     in the ground truth. Evaluate if
    there are any hallucinated details
     that misrepresent or add
    unwarranted content.\n\n**Scoring
    **: \nAssign a score from **1 to
    5**, where:\n- **1 - Poor**: The
    predicted continuation contains
    numerous hallucinations and
    fabricated details not supported
    by the ground truth.\n- **2 - Weak
    **: The continuation includes
    several instances of hallucination,
     significantly affecting its
    credibility.\n- **3 - Acceptable**:
     The continuation has minor
    hallucinated elements, but these
    do not majorly undermine the
    content.\n- **4 - Good**: The
    continuation contains minimal
    hallucinations with mostly
    accurate representation.\n- **5 -
    Excellent**: The continuation is
    free of hallucinations and
    completely aligns with the ground
    truth.\n\n**Output Format**: \
    nBegin your evaluation by
    providing a short explanation. Be
    as objective as possible. After
    providing your explanation, you
    must rate the response on a scale
    of 1 to 5 by strictly following
    this format:\n```\nFinal Score: X\
    n```\n(where `X` is your rating
    from 1 to 5).
```

E Qualitative examples of long quotes after rewriting

We show qualitative examples of long quotes that are still present in the model generation below. Many of these long quotes contain long named entities that are difficult to rewrite, but are also likely low risk for copyright infringement.

NewsSpan:

```
<quote1>Should healthy people be
    wearing masks when they're outside
     to protect themselves and others?
<quote2> for The Guardian, said he was
     "body slammed" by Greg Gianforte,
     a Republican candidate
<quote3> of communication between the
    incoming administration and the
    Russian government.
<quote4>s. The Federal Reserve and the
     New York State Department of
    Financial Services
<quote5>
...CBS News Magazine "60 Minutes"
    features the story of Beckett
    Brennan, a
<quote6> Dr. Donald Hensrud, director
    of the Mayo Clinic's Healthy
    Living Program.
<quote7> Chris Christie of New Jersey,
     who briefly led the Trump
    transition team,
<quote8> Chris Christie of New Jersey,
     who briefly led the Trump
    transition team,
<quote9> "If I Had a Hammer," "
    Goodnight Irene," and "Kisses
    Sweeter Than Wine,"
<quote10> a billion acres in the
    Arctic, Pacific, Atlantic, and
    Gulf of Mexico. T
```

NewsQA:

```
<quote1>s motivated by a person's
    actual or perceived gender, sexual
     orientation, gender identity, or
    disability.
<quote2> the US Department of Health
    and Human Services and the Centers
     for Disease Control and
    Prevention,
<quote3> David Petraeus, the top US
    commander in Iraq, and Ryan
    Crocker, the US ambassador to
<quote4>s.
The FDA is warning consumers to
    immediately stop using 14
    Hydroxycut products,
<quote5> Rear Admiral Gregory Smith,
    the U.S. military's chief
    spokesman in Iraq,
<quote6>to the Fundamentalist Church
    of Jesus Christ of Latter-day
    Saints (FLDS)
<quote7> the Fundamentalist Church of
    Jesus Christ of Latter-day Saints
    (FLDS).
<quote8> the Fundamentalist Church of
    Jesus Christ of Latter-day Saints
    (FLDS),
<quote9>t:
The Fundamentalist Church of Jesus
    Christ of Latter-day Saints, a
<quote10> Ralph Nicoletti, 18, Michael
     Contreras, 18, and Brian Carranza,
     21,
```