Evaluating Copyright Takedown Methods for Language Models

Anonymous Authors[1]

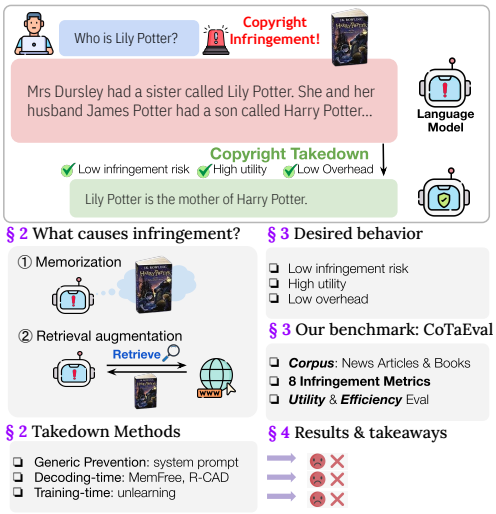

Figure 1. Effective copyright takedown methods should prevent models from generating infringing content (low infringement risk) while preserving uncopyrightable facts and fair use information (high utility).

§1. Overview

Language models (LMs) derive their capabilities from extensive training on diverse data, largely drawn from across the web (Bommasani et al., 2021) and contains copyrighted material. These models can memorize and generate content similar to their training data, potentially risking legal issues like copyright infringement *Tremblay v. OpenAI, Inc.,* (2023); *Kadrey v. Meta Platforms, Inc.* (2023). Generally, there is less legal risk, and a more likely fair use defense, if models do not output content substantially similar to the training data (Henderson et al., 2023; Sag, 2023; Lee et al., 2024). Therefore, model deployers require a strategy to prevent models from outputting infringing content, which they may only notice after training is complete. We refer to this as a **copyright takedown**. Legal scholars suggest that a takedown mechanism may be a necessary and effective part of future policymaking (Henderson et al., 2023; Pasquale & Sun, 2024; Lee et al., 2024). Yet, a key question remains: *Can "takedown" of copyrighted content be operationalized in the context of large language models?*

This paper introduces the first evaluation of the feasibility and side effects of copyright takedowns for LMs. We propose CoTaEval, an evaluation framework to assess the effectiveness of copyright takedown methods, the impact on the model's ability to retain uncopyrightable factual knowledge from the copyrighted content, and how well the model maintains its general utility and efficiency. We examine several strategies, including adding system prompts, decoding-time filtering interventions, and unlearning approaches. Our findings indicate that no method excels across all metrics, showing significant room for research in this unique problem setting and indicating potential unresolved challenges for live policy proposals.

§2. Copyright Infringement in LMs

Causes of infringement. Recent litigation (*Tremblay v. OpenAI, Inc.*,, 2023; *Kadrey v. Meta Platforms, Inc.*, 2023; *Chabon v. OpenAI, Inc.*,, 2023; *DOE 1 v. GitHub, Inc.*, N.D. Cal. 2022) has pointed to two scenarios where a language model deployment might lead to infringement: (1) copyrighted content is memorized within the model's parameters during training (*Memorization*; Carlini et al., 2019; 2021; 2023; Zhang et al., 2023; Nasr et al., 2023), and (2) copyrighted content is incorporated as additional context during retrieval-augmented generation (*RAG*; Lewis et al., 2020). These scenarios necessitate copyright takedown approaches.

Potential copyright takedown methods. Our evaluation considers three types of takedown methods that intervene at different stages of the language model: 1) strategies that generally try to prevent infringement without specifying a blocklist, including System Prompt and Top-k Perturbation (Adding Gaussian noise to the logits in top-k sampling); 2) methods that prevent the generation of blocklisted content during decoding, including MemFree (Ippolito et al., 2023), Reversed Context Aware Decoding (R-CAD; Shi et al., 2023); 3) training-based interventions like unlearning, including *Gradient ascent* ($\text{Unlearning}_{\text{GA}}$; Thudi et al., 2022), *Gradient Difference* ($\text{Unlearning}_{\text{GD}}$; Liu et al., 2022), *KL minimization* ($\text{Unlearning}_{\text{KL}}$; Golatkar et al., 2020), and *Preference Optimization* ($\text{Unlearning}_{\text{PO}}$; Rafailov et al., 2024). R-CAD and unlearning specifically target the memorization scenario, while others can handle both scenarios.

§3. The CoTaEval Evaluation Pipeline

To evaluate the effectiveness of copyright takedown methods, we propose a new evaluation pipeline CoTaEval

(**Co**pyright **Ta**kedown **Eval**uation). CoTaEval uses books and news articles as evaluation corpus and considers both the memorization and RAG scenarios. The effectiveness of different takedown methods is quantified based on three desiderata that we propose: **low infringement risk**, **high utility**, and **low overhead**.

Evaluation Corpus. Our evaluation focuses on two prevalent types of text often involved in copyright infringement cases: *news articles* and *books*. For the *news articles* domain, we use the NewsQA dataset (Trischler et al., 2017), which consists of CNN articles paired with questions and answers derived from those articles. For the *books* domain, we use the BookSum dataset (Kryściński et al., 2022), where each example includes a book chapter along with a summary of that chapter's content.

Evaluation metrics. We divide each corpus into blocklisted content, which the model must avoid generating, and in-domain content, which is not subject to takedown requests. We then evaluate takedown methods based on three criteria:

- *Low Infringement Risk*: Given that any insufficient transformation of blocklisted copyrighted content can potentially lead to infringement (Lemley & Casey, 2021; Sag, 2023; Henderson et al., 2023), CoTaEval adopts eight similarity metrics covering three types of potential infringement: exact match, near duplicate, and semantic similarity.
- *High Utility*: Our utility evaluation considers factual knowledge preservation of blocklisted and in-domain content, as well as general utility: For *Blocklisted and in-domain utility*, we ask the model to answer questions related to factual information within the articles in the news articles domain, and ask the model to briefly summarize a book chapter in the books domain. For *General utility*, we use MMLU (Hendrycks et al., 2020) and MT-Bench (Zheng et al., 2024) to evaluate the model's knowledge and reasoning abilities across various range of subjects and tasks.
- *Low Overhead*: We also evaluate the computational efficiency of takedown methods during inference. For a fair comparison, when evaluating the efficiency, we limit the model to generate a fixed number of tokens, and report the average inference speed.

§4. Experiment

We use CoTaEval to evaluate two models: Llama2-7B-chat and Llama2-70B-chat (Touvron et al., 2023). Our observations are listed below:

System Prompt and MemFree offer some mitigation but cannot completely prevent infringement. We observe that System Prompt effectively increases the chances that the model rejects outputting copyrighted content, and it is particularly effective in the RAG scenario within the news domain. However, it still fails occasionally; the model does not correctly reject every instance. MemFree can reduce infringement risk while generally preserving utility, particularly for exact matching infringement, as it employs a Bloom-filter-based detection algorithm, which identifies elements that exactly match those stored in the Bloom filter. However, minor misspellings, extra whitespace, or additional newline characters cannot be captured by the exact match detector and can thus easily bypass detection. Consequently, it struggles to effectively prevent other forms of infringement, such as near-duplicates.

Method	Infringement risk reduction win rate (%, ↑)	Utility (↑) MMLU	MT-Bench	Blocklisted F1	In-Domain F1	Inference speed (↑)
Vanilla	19.2	$35.3_{\pm 3.1}$	$4.7_{\pm 0.5}$	$40.5_{\pm 1.5}$	$40.6_{\pm 1.5}$	1.00×
System Prompt	54.2	$34.0_{\pm 3.1}$	$4.4_{\pm 0.5}$	$33.4_{\pm 2.0}$	$33.0_{\pm 2.0}$	1.00×
Top-k Perturbation	28.7	$14.7_{\pm 1.7}$	$3.0_{\pm 0.4}$	$3.3_{\pm 0.7}$	$1.8_{\pm 0.5}$	0.99×
MemFree	24.3	$35.3_{\pm 3.1}$	$4.7_{\pm 0.5}$	$36.2_{\pm 1.4}$	$37.9_{\pm 1.6}$	0.94×
R-CAD	51.7	$35.3_{\pm 3.1}$	$4.7_{\pm 0.5}$	$40.5_{\pm 1.5}$	$40.6_{\pm 1.5}$	0.53×
Unlearning$_{GA}$	30.1	$27.9_{\pm 3.3}$	$3.3_{\pm 0.5}$	$26.9_{\pm 1.9}$	$25.8_{\pm 1.8}$	1.00×*
Unlearning$_{GD}$	64.2	$15.8_{\pm 3.2}$	$1.5_{\pm 0.3}$	$16.9_{\pm 1.3}$	$16.2_{\pm 1.3}$	1.00×*
Unlearning$_{KL}$	61.6	$17.6_{\pm 3.3}$	$1.5_{\pm 0.3}$	$16.9_{\pm 1.4}$	$15.9_{\pm 1.3}$	1.00×*
Unlearning$_{PO}$	67.5	$33.1_{\pm 3.3}$	$2.4_{\pm 0.4}$	$28.3_{\pm 2.0}$	$24.7_{\pm 2.0}$	1.00×*

Table 1. **Performance of takedown methods in the memorization scenario** with fine-tuned Llama2-7b-chat model on news articles. A darker cell indicates better performance. * denotes offline costs. Despite extensive hyperparameter tuning, unlearning methods reduce infringement but also harm utility. R-CAD is effective but compromises both efficiency and utility.

Unlearning and Top-k Perturbation reduce infringement but significantly compromises factual knowledge from the blocklisted content. Although some of the unlearning methods show their capability to reduce the infringement risk (for example, Unlearning$_{PO}$ and Unlearning$_{GD}$), we find they have several downsides. First, most of the unlearning methods are hyperparameter sensitive and require extensive search. Second, existing unlearning methods are not designed to preserve factual knowledge and often inadvertently remove it. Another concern about the unlearning process is that it cannot guarantee the unlearned content will not be generated again (Shi et al., 2023; Patil et al., 2023), necessitating careful audits (Huang et al., 2022). Similarly, for Top-k Perturbation, it will sacrifice a lot of utility when it becomes effective in reducing the infringement risk.

R-CAD is effective for takedown but comes at the cost of efficiency and risk of utility drop. In the news articles domain, R-CAD achieves a 51.7% win rate across all methods. It avoids retrieval when the Faiss distance (Douze et al., 2024) exceeds 0.15, reverting to vanilla decoding, thus maintaining the utility. However, in the worst-case scenario, the retriever might still retrieve the "gold document". If all the context can be retrieved, the blocklisted F1 score is only $5.7_{\pm 1.0}$, indicating a significant risk of utility drop when R-CAD is triggered. Additionally, it introduces an extra inference process during the intervention, reducing the model's inference efficiency by approximately half.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. Quantifying memorization across neural language models. In *ICLR*, 2023.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library. 2024.

Chabon v. OpenAI, Inc.,. 3:23-cv-04625, (N.D. Cal.), 2023.

DOE 1 v. GitHub, Inc. 4:22-cv-06823, N.D. Cal. 2022.

Kadrey v. Meta Platforms, Inc. 3:23-cv-03417, 2023.

Tremblay v. OpenAI, Inc.,. 23-cv-03416-AMO, (N.D. Cal.), 2023.

Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*, 2020.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *ICLR*, 2020.

Huang, Y., Huang, C.-Y., Li, X., and Li, K. A dataset auditing method for collaboratively trained machine learning models. *IEEE Transactions on Medical Imaging*, 2022.

Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette-Choo, C. A., and Carlini, N. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pp. 28–53. Association for Computational Linguistics, 2023.

Kryściński, W., Rajani, N., Agarwal, D., Xiong, C., and Radev, D. Booksum: A collection of datasets for long-form narrative summarization. In *EMNLP-Findings*, 2022.

Lee, K., Cooper, A. F., and Grimmelmann, J. Talkin"bout ai generation: Copyright and the generative-ai supply chain (the short version). In *Proceedings of the Symposium on Computer Science and Law*, pp. 48–63, 2024.

Lemley, M. A. and Casey, B. Fair learning. *Texas Law Review*, 99(4):743–785, 2021.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 2020.

Liu, B., Liu, Q., and Stone, P. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*. PMLR, 2022.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Pasquale, F. and Sun, H. Consent and compensation: Resolving generative ai's copyright crisis. *Cornell Legal Studies Research Paper Forthcoming*, 2024.

Patil, V., Hase, P., and Bansal, M. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. In *ICLR*, 2023.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36, 2024.

Sag, M. Copyright safety for generative ai. *Forthcoming in the Houston Law Review*, 2023.

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 *ICLR*, 2023.

Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2022.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 2017.

Zhang, C., Ippolito, D., Lee, K., Jagielski, M., Tramèr, F., and Carlini, N. Counterfactual memorization in neural language models. *NeurIPS*, 36:39321–39362, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36, 2024.