Exploring an Agenda on Memorization-based Copyright Verification

Harry H. Jiang [1] **Aster Plotnik** [2] **Carlee Joe-Wong** [1]

Abstract

The methods and systems through which developers of large language models (LLMs) acquire training data is nebulous and contentious. As a result, many data owners have concerns about whether their data is being used inappropriately. Thus, finding a way for data owners to independently verify whether their data has been used to train LLMs is an increasingly popular area of research. In such situations, a recourse for data owners is civil litigation. In many legal systems, such as that of the United States, a major hurdle for a civil suit is finding the evidence necessary to bring a case from pleadings to discovery; memorization of a text by an LLM would be helpful in this case as evidence a text was present in training data. Currently, there is a disconnect between the legal system, which demands evidence of memorization in text, and the realities of the technology that can rarely produce this. From analysis of both law in various jurisdictions, and a review of memorization techniques and benchmarks, we propose in this paper a set of objectives for researchers to better align work on memorization and other defences for data owners with the practicalities of argumentation in the legal realm, namely: specificity, substantiality, and accessibility.

1. Introduction

Large language models (LLMs) are today lucrative products, built on large amounts of text data used for training; the acquisition of this data is an essential part of LLM creation. However, the training process is often opaque, raising concerns among data rights holders that their intellectual property may be used to train these models without permission (Henderson et al., 2023).

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA [2]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada . Correspondence to: Harry H. Jiang <hhj@andrew.cmu.edu>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

Frameworks such as the "three C's" (Consent, Compensation, and Credit) (Cultural Intellectual Property Rights Initiative, 2017) have been identified as a model for beneficial governance from the perspective of creatives (Kyi et al., 2025), but these objectives are either far away from regulatory implementation through the legislative process in most countries or inaccessible for smaller parties to directly negotiate with the model producers.

As such, creators and data owners have sought recourse through civil litigation and existing concepts of intellectual property (IP), namely copyright, also known as author's rights. These civil suits can not only address specific instances of unauthorized data usage, but may also serve as a deterrent for not following established and expected norms and laws in data collection. Given this turn towards judicial means of protection, technological defences of creators' interests should also be devised with the legal field in mind.

A natural source of evidence to justify civil litigation lies in memorization, which in LLMs refers to the retention and output of information from an identifiable source within the training data. This behaviour thus forms from a causal link to the use of a text in training data. However, as memorization is an active area of research with many competing definitions, not all of which are usable in legal settings, this paper aims to align areas of research in memorization with the practicalities of law in order to foster usable research outcomes for data rights holders. After giving an overview of LLM memorization and copyright (Sections 2 and 3), in Section 4, we contribute a set of desiderata for usability in memorization-based tools: **specificity**, **substantiality**, and **accessibility**. We argue for the necessity of those items in the context of current IP law and the proceedings of existing legal cases regarding IP and generative AI models.

2. LLM Memorization

LLMs are able to generate strings of text as a result of training on existing works of written language. However, LLMs have been known to generate outputs that are identical or provably derived from a portion of the training dataset; this *memorization* phenomenon presents a variety of risks such as privacy and security concerns and violation of copyright law (Hartmann et al., 2023). With models rapidly increasing in size, memorization is expected to become an increasingly

Table 1. Overview of recent language model membership inference methods relevant to copyright verification. "Type" refers to the memorization type (V=verbatim, A=approximate, E=entity-level) used in the work, if applicable. "Evidence" refers to the type of information used as evidence for memorization (M=metric, T=Text). "Access" refers to the access to the model required to evaluate (I=input, L=logits and/or loss, O=output); insertion capability into the training data prior to model training is presumed.

Work(s)	Type	Evidence	Access
(Ravichander et al., 2025)	—	M	I, O
(Zhao et al., 2025)	—	M	I, O
(Cui et al., 2025)	E	M	I, L, O
(He et al., 2025)	A	M	I, O
(Meeus et al., 2024a;b)	A	M	I, L, O
(Duarte et al., 2024)	—	M	I, O
(Zhao et al., 2024)[1]	V	M, T	I, O
(Shi et al., 2024)	A	M	I, L, O
(Wei et al., 2024)	A	M	I, L, O
(Chang et al., 2023)	E	T	I, O

significant problem as larger models have been shown to be more prone to memorization (Kiyomaru et al., 2024). LLM developers have therefore become interested in mitigating memorization and its potential harms. We follow Satvaty et al. (2025) in defining types of memorization: *verbatim* (outputs string from training data without alteration), *approximate* (significantly similar output to portion of training data), and *entity-level* (output traceable to a set of information from a sample in training data) memorization.

The key properties of verbatim memorization are presented by Carlini et al. (2023), who show that memorization probability of a string increases with model size, duplication within training data, and context length of prompt. A related concept is "membership inference", which refers to a method which aims to verify whether certain data has been used to train a model; for data privacy reasons, membership inference is often conceived as an attack (MIA) on a model.

Memorization is also a highly subjective topic, with many positions taken on the feasibility and validity of research in the area. Given the scale of LLM training data, Duan et al. (2024); Rando et al. (2025); Meeus et al. (2024a) point out the difficulty in achieving and evaluating membership inference attacks on recent and future LLMs. Similarly, Zhang et al. (2025) question the validity of hypothesis testing for membership inference. Conversely, Ippolito et al. (2023) warn against verbatim memorization mitigation methods as incomplete towards membership inference.

2.1. Existing Copyright Verification Work

Membership inference can be used to verify usage of copyrighted works in training LLMs, as summarized in Table 1's references. Some, such as (Wei et al., 2024; Zhao et al., 2025), measure memorization at a data repository level (e.g. arXiv, the *NYT*), but most focus on single documents, such as Cooper et al. (2025)'s work.

Meeus et al. (2024b); Shilov et al. (2024) propose the concept of *copyright traps*, which are sequences within a text deliberately intended for memorization as a method to evaluate whether a model has infringed upon the rights of a given property, though the authors did find significant challenges regarding retrieving verbatim text and required duplication count (Meeus et al., 2025). Zhao et al. (2024)[1] propose a similar method to copyright traps with similar findings on effective repetition requirements.

Outside of proposing verification methods, Xu et al. (2024) observe the memorization of copyright notices, finding that many LLMs indeed fail to respect copyright notices, leading to memorization. Karamolegkou et al. (2023) explore verbatim memorization of code and copyrighted books. Chen et al. (2024) propose a benchmark to measure verbatim, approximate, and entity-level memorization on copyrighted works of fiction. For specific works, the benchmark proposed by Duarte et al. (2024) consists of multiple-choice prompts which aims to determine membership of copyrighted works; notably, Rosenblat et al. (2025) used this benchmark to argue that copyrighted works from the publisher O'Reilly has been trained on by some of OpenAI's models.

3. Memorization and Intellectual Property

Since there exist scenarios in which copyrighted works can be legally reproduced and used as part of other works, and since copyrighted works can be part of the training data used to produce LLMs and contribute to their performance, the question arises of whether using a copyrighted work in LLM training is permissible under the law. In the United States, the doctrine determining this permissibility is known as *fair use*; Henderson et al. (2023) cover considerations of fair use in terms of foundation models. Fair use eligibility is determined through four factors: 1) transformativeness, 2) nature of copyrighted work, 3) amount and substantiality of use, and 4) market effect of use.

3.1. Relevant U.S. Cases and Commentary

In common law jurisdictions, including most Commonwealth nations and the U.S., determinations of legal tests are predicated upon precedent. Thus, even cases outside of generative AI are relevant to the bounds of fair use in LLMs.

Memorization of copyrighted text has already appeared as evidence submitted in a legal complaint of *New York Times v. Microsoft*[2], seeking damages for OpenAI and Microsoft allegedly having "used copyrighted works, without consent or

[1]This work has been withdrawn from submission by its authors.

[2]No. 1:23-cv-11195 (S.D.N.Y.), consolidated thereafter into No. 1:25-md-03143-SHS-OTW (S.D.N.Y.)

compensation", in LLM training. Examples of memorized passages found in NYT articles were featured.[3] Though this case has not yet entered discovery, it demonstrates the legal relevance of memorization, which was invoked by a major rights holder as evidence of alleged harm by an LLM.

An important relevant case is *Kadrey v. Meta*[4] in which authors sought damages for the use of their works in training the Llama family of models. First, the proceedings of the case shows the importance of moving to discovery: it was uncovered during discovery that Meta knowingly torrented copyrighted works for its training data, a possible violation of California Penal Code § 502. Even though memorization has not featured prominently in complaints[5], they may strengthen complaints and aid in passage to discovery. Comments from United States Copyright Office (2025) align, stating that "[w]hether a model's weights implicate the reproduction or derivative work rights turns on whether the model has retained or memorized substantial protectable expression from the work(s) at issue."

Second, and most crucially, the summary judgment on the direct copyright infringement claims in *Kadrey v. Meta*[6], despite dismissing them, clearly delineated the utility of memorization in LLM litigation: since there is a weak case for LLMs being non-transformative[7], the clearest path to restitution in the U.S. system is through the fourth fair use factor. The judge remarked that a drastic change to the market for a work, such as that wrought upon the literary market by the introduction of LLMs, is difficult to be understood as fair use. This again shows that memorization is most influential as evidence not of an LLM's capacity to disseminate a work in deployment, but of anterior unauthorized copying in the training of an LLM. This could also tie the outcome of the case to the precedent set by *Thomson Reuters v. ROSS*[8] in which summary judgment found that the defendant, ROSS Intelligence, a (non-generative) AI company, infringed on the copyright of news company Thomson Reuters by copying its content to train a model for a competing product.

3.2. Other Jurisdictions

Though fair use is a uniquely American doctrine, most other jurisdictions also have carve-outs to exclusive rights of reproduction. The EU notably has a pre-existing text and data mining (TDM) exception which may apply to LLMs and would give reproduction and extraction rights to research organizations and cultural heritage institutions, predicated on lawful access of the content (Novelli et al., 2024).

In addition, there exist possibly "data havens", jurisdictions in which model producers are given much more latitude in gathering data and training models without IP risks. Japan is one such possible location, where upcoming policy by the Agency of Cultural Affairs states that any unmanaged work in Japan may be free for commercial use in limited respects if no reply is given within 14 days of requesting[9]; coupled with the TDM clause in Japanese copyright[10], this may allow large swaths of works copyrighted elsewhere to be open to commercial model training in Japan. In such situations, though training may be allowed, but as argued in (Dornis & Stober, 2024; Dornis, 2025), a model's deployment in a jurisdiction home to a memorized work's reproduction rights may nonetheless bear IP risk.

4. Desiderata for Memorization-based Tools

The key objective for a copyright verification tool should be to bring actionable evidence of reproduction as to allow a rights infringement case to proceed to discovery. Also, for models trained offshore, memorization can impact their legal deployment and help rights holders keep developers accountable to local rights and regulations.

We identify the main baseline properties for a tool to be practical as an effective copyright verifier:

1) Specificity: A tool should allow LLM outputs to trace back specifically to a singular original text.

2) Substantiality: The output text should be long enough as to carry a meaningful portion of the expression of a work.

3) Accessibility: Evidence should be retrievable with only access to the model available to the general public.

5. Discussion

In this section we discuss the technical specifics of implementing the requirements listed in the previous section as to form sufficient conditions for producing effective evidence.

Firstly, in regards to specificity, it is not sufficient for a copyright holder to prove that an LLM was trained on a set of works. This means that even if an LLM has provably memorized a portion of text that is traceable to a small number of copyrighted works[11], the copyright holder cannot make legal claims about any given work since the evidence is not unique. The copyright holder must be able to map a specific piece of evidence to a specific, singular work.

[3] *supra*, Doc. 170 (First Amended Complaint) pp. 30-46

[4] Nos. 3:23-cv-03417, 3:24-cv-06893 (N.D. Cal.).

[5] *supra*, Doc. 407 (Third Amended Complaint) p. 17

[6] *supra*, Doc. 598 (Partial Summary Judgment)

[7] *cf. supra*, Doc. 598, and *Bartz v. Anthropic PBC*, 3:24-cv-05417, (N.D. Cal.), Doc. 231 (Partial Summary Judgment)

[8] No. 1:20-cv-00613 (D. Del.).

[9] *What is the Unmanaged Work Arbitration System?*

[10] Copyright Act (Act No 48 of 1970), Articles 30-4, 67, 67-3

[11] This would not apply if the duplicates are quotes of a single original work.

Any textual inclusions intended to strengthen legal evidence must be entirely unique to the specific work in question.

Second, substantiality refers to memorization producing a substantial portion of the original expression of the text such that reading the output is a substitute for reading the original text in terms of understanding the intent of the expression. This must also be communicated verbatim or near-verbatim, i.e., minimally transformative. Many works listed in Table 1 rely on statistical evidence of a work's use in training, but this may not be an obvious enough evidence for a non-technical adjudicator, and thus is not as effective of a support to a legal claim. Moreover, many approximate memorization-based tools use similarity metrics that are too lax compared to legal tests such as the transformativeness standard in fair use. It may be the case that paraphrasing non-fiction works eliminates what is considered "protectable expression" in a text under copyright, in which case certain semantic metrics cannot be used and thresholds for other similarity metrics must be kept very low.

Third, the evidence must be accessible by methods which may be reproduced by a general public end-user of the LLM. Methods which require access to features of the model such as logits or a loss function are not practical given that the copyright holder in a real-world setting will not be able to access these. For most models that are relevant today, the general public is only able to access a prompting user interface, and thus defences should be designed to be retrievable through prompting the model.

In addition to the desiderata above, derived from the legal burdens placed on the copyright holder, the design of the tool should consider **essentiality**. Any portion of text that is intended to be used as evidence must be resilient to removal by techniques such as deduplication (Kandpal et al., 2022; Tirumala et al., 2023; Sakarvadia et al., 2025), or allow it such that the removal of this essential portion of the text should hinder the overall usefulness of the training data item, either by causing the text to lose meaning or causing too much of the overall text to be lost. This would prevent dataset creators from obfuscating their inclusion of protected data by means of automated filtering.

One item that we deem not necessary is retrievability. This is because even though specificity is stressed, many existing cases represent plaintiffs holding rights to multiple works (e.g., publisher in *NYT v. Microsoft*, class-action in *Kadrey v. Meta*). As such, it may be sufficient to have a tool with low or probabilistic retrievability (see probabilistically discoverable extraction in (Hayes et al., 2025)) if evidence for a few works allows for a case to enter discovery where defendants may be compelled to provide facts on whether other works owned by the plaintiffs are used in training.

5.1. Further Research

Existing work on memorization often emphasizes the impact of duplication on memorization. Primarily, more research is required to determine characteristics of text that is more prone to memorization beyond duplication, e.g. for a passage with highly unusual structure. Tangentially, Shilov et al. (2024) provides a promising addition to copyright traps where intra-document duplication need not to be exactly identical. Secondarily, as quotations can be fair use and traceable to a single original work, there may be methods to use quotations strategically as to induce memorization if a full text were trained on.

Substantiality is an area where the state of the art vastly underperforms sufficient conditions. An obvious area of work is to convert many of the methods in Table 1 to use textual evidence and near-verbatim memorization. Also, while Zhao et al. (2024) does use both of the above items, they still operate with a very short memorization length, offering another opening for further research.

5.2. Weaknesses

Though we assume access to inputs and outputs of an LLM as a typical access level available to the public, we did not consider post-processing of LLM outputs in applications, which could filter out memorized texts before they reach the user, as in (Ippolito et al., 2023). Similarly, much work in this area presumes greedy sampling and zero temperature, and the influence of modifying those inference parameters on memorization-based methods is not well studied. Another area to monitor is retrieval-augmented generation (RAG), where explicit knowledge retrieval is enabled for an LLM (Guu et al., 2020). If RAG is declared openly, e.g., with sources cited, then the tool may be subject to precedent such as *Authors Guild v. Google*[12] ; if used clandestinely to obfuscate memorization, it could pose additional IP risk.

Memorization can also be countered by further training of models; for example, instruction tuning significantly reduces verbatim memorization as per (Chen et al., 2024). Similarly, machine unlearning is an area which may affect effectiveness of memorization-based verification.

6. Conclusion

In this work, we survey the field of memorization-based copyright verification methods and legal background relevant to LLMs and memorization, and we offer desiderata for usable methods in the context of civil litigation, namely **specificity**, **substantiality**, and **accessibility**. Though the sufficient conditions formed through our findings are difficult to achieve, we identify areas of research to bring these

[12] 804 F.3d 202 (2nd Cir. 2015)

objectives closer to reality. Furthermore, we express optimism for achieving these desiderata, especially as properties favourable to memorization, such as model size, continue to increase.

Acknowledgments

This research has been funded by the CyLab Future Enterprise Security Initiative at Carnegie Mellon University.

References

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models, 2023. URL https://arxiv.org/abs/2202.07646.

Chang, K., Cramer, M., Soni, S., and Bamman, D. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7312–7327, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.453. URL https://aclanthology.org/2023.emnlp-main.453/.

Chen, T., Asai, A., Mireshghallah, N., Min, S., Grimmelmann, J., Choi, Y., Hajishirzi, H., Zettlemoyer, L., and Koh, P. W. Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation, 2024. URL https://arxiv.org/abs/2407.07087.

Cooper, A. F., Gokaslan, A., Cyphert, A. B., De Sa, C., Lemley, M. A., Ho, D. E., and Liang, P. Extracting memorized pieces of (copyrighted) books from open-weight language models. *arXiv preprint arXiv:2505.12546*, 2025.

Cui, X., Wei, J. T.-Z., Swayamdipta, S., and Jia, R. Robust data watermarking in language models by injecting fictitious knowledge, 2025. URL https://arxiv.org/abs/2503.04036.

Cultural Intellectual Property Rights Initiative. Consent Credit Compensation: The Legal Literacy Campaign, 2017. URL https://www.culturalintellectualproperty.com/the-3cs.

Dornis, T. W. Generative ai, reproductions inside the model, and the making available to the public. *International Review of Intellectual Property and Competition Law (IIC)*, 2025. doi: https://dx.doi.org/10.2139/ssrn.5036008. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5036008.

Dornis, T. W. and Stober, S. Copyright law and generative ai training - technological and legal foundations (urheberrecht und training generativer ki-modelle - technologische und juristische grundlagen). *Open Access book in the NOMOS Verlag (Baden-Baden) publisher's series Recht und Digitalisierung/Digitization and the Law.*, 2024. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4946214.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and

Hajishirzi, H. Do membership inference attacks work on large language models?, 2024. URL https://arxiv.org/abs/2402.07841.

Duarte, A. V., Zhao, X., Oliveira, A. L., and Li, L. De-cop: Detecting copyrighted content in language models training data, 2024. URL https://arxiv.org/abs/2402.09910.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training, 2020. URL https://arxiv.org/abs/2002.08909.

Hartmann, V., Suri, A., Bindschaedler, V., Evans, D., Tople, S., and West, R. Sok: Memorization in general-purpose large language models, 2023. URL https://arxiv.org/abs/2310.18362.

Hayes, J., Swanberg, M., Chaudhari, H., Yona, I., Shumailov, I., Nasr, M., Choquette-Choo, C. A., Lee, K., and Cooper, A. F. Measuring memorization in language models via probabilistic extraction, 2025. URL https://arxiv.org/abs/2410.19482.

He, Y., Li, B., Liu, L., Ba, Z., Dong, W., Li, Y., Qin, Z., Ren, K., and Chen, C. Towards label-only membership inference attack against pre-trained large language models, 2025. URL https://arxiv.org/abs/2502.18943.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023. URL http://jmlr.org/papers/v24/23-0569.html.

Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette-Choo, C. A., and Carlini, N. Preventing verbatim memorization in language models gives a false sense of privacy, 2023. URL https://arxiv.org/abs/2210.17546.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models, 2022. URL https://arxiv.org/abs/2202.06539.

Karamolegkou, A., Li, J., Zhou, L., and Søgaard, A. Copyright violations and large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458. URL https://aclanthology.org/2023.emnlp-main.458/.

Kiyomaru, H., Sugiura, I., Kawahara, D., and Kurohashi, S. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pp. 584–596, 2024.

Kyi, L., Mahuli, A., Silberman, M. S., Binns, R., Zhao, J., and Biega, A. J. Governance of generative ai in creative work: Consent, credit, compensation, and beyond. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713799. URL https://doi.org/10.1145/3706598.3713799.

Meeus, M., Jain, S., Rei, M., and de Montjoye, Y.-A. Did the neurons read your book? document-level membership inference for large language models, 2024a. URL https://arxiv.org/abs/2310.15007.

Meeus, M., Shilov, I., Faysse, M., and de Montjoye, Y.-A. Copyright traps for large language models, 2024b. URL https://arxiv.org/abs/2402.09363.

Meeus, M., Shilov, I., Jain, S., Faysse, M., Rei, M., and de Montjoye, Y.-A. Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it), 2025. URL https://arxiv.org/abs/2406.17975.

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., and Floridi, L. Generative ai in eu law: Liability, privacy, intellectual property, and cybersecurity. *Computer Law Security Review*, 55:106066, 2024. ISSN 2212-473X. doi: https://doi.org/10.1016/j.clsr.2024.106066. URL https://www.sciencedirect.com/science/article/pii/S0267364924001328.

Rando, J., Zhang, J., Carlini, N., and Tramèr, F. Adversarial ml problems are getting harder to solve and to evaluate, 2025. URL https://arxiv.org/abs/2502.02260.

Ravichander, A., Fisher, J., Sorensen, T., Lu, X., Lin, Y., Antoniak, M., Mireshghallah, N., Bhagavatula, C., and Choi, Y. Information-guided identification of training data imprint in (proprietary) large language models, 2025. URL https://arxiv.org/abs/2503.12072.

Rosenblat, S., O'Reilly, T., and Strauss, I. *Beyond Public Access in LLM Pre-Training Data: Non-public book content in OpenAI's Models*. April 2025. doi: 10.35650/aidp.4111.d.2025. URL http://dx.doi.org/10.35650/AIDP.4111.d.2025.

Sakarvadia, M., Ajith, A., Khan, A., Hudson, N., Geniesse, C., Chard, K., Yang, Y., Foster, I., and Mahoney, M. W.

Mitigating memorization in language models, 2025. URL https://arxiv.org/abs/2410.02159.

Satvaty, A., Verberne, S., and Turkmen, F. Undesirable memorization in large language models: A survey, 2025. URL https://arxiv.org/abs/2410.02650.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models, 2024. URL https://arxiv.org/abs/2310.16789.

Shilov, I., Meeus, M., and de Montjoye, Y.-A. Mosaic memory: Fuzzy duplication in copyright traps for large language models, 2024. URL https://arxiv.org/abs/2405.15523.

Tirumala, K., Simig, D., Aghajanyan, A., and Morcos, A. S. D4: Improving llm pretraining via document de-duplication and diversification, 2023. URL https://arxiv.org/abs/2308.12284.

United States Copyright Office. Copyright and Artificial Intelligence Part 3: Generative AI Training Pre-publication version, 05 2025. URL https://copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-pdf.

Wei, J. T.-Z., Wang, R. Y., and Jia, R. Proving membership in llm pretraining data via data watermarks, 2024. URL https://arxiv.org/abs/2402.10892.

Xu, J., Li, S., Xu, Z., and Zhang, D. Do llms know to respect copyright notice?, 2024. URL https://arxiv.org/abs/2411.01136.

Zhang, J., Das, D., Kamath, G., and Tramèr, F. Membership inference attacks cannot prove that a model was trained on your data, 2025. URL https://arxiv.org/abs/2409.19798.

Zhao, B., Maini, P., Boenisch, F., and Dziedzic, A. Unlocking post-hoc dataset inference with synthetic data. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. URL https://openreview.net/forum?id=yMvaGZCY8b.

Zhao, S., Zhu, L., Quan, R., and Yang, Y. Protecting copyrighted material with unique identifiers in large language model training, 2024. URL https://openreview.net/forum?id=5xbKFaaqkS.