LLMs and Copyright Risks: Benchmarks and Mitigation Approaches

Denghui Zhang
Stevens Institute of Technology
dzhang42@stevens.edu

Zhaozhuo Xu
Stevens Institute of Technology
zxu79@stevens.edu

Weijie Zhao
Rochester Institute of Technology
wjz@cs.rit.edu

Abstract

Large Language Models (LLMs) have revolutionized natural language processing, but their widespread use has raised significant copyright concerns. This tutorial addresses the complex intersection of LLMs and copyright law, providing researchers and practitioners with essential knowledge and tools to navigate this challenging landscape. The tutorial begins with an overview of relevant copyright principles and their application to AI, followed by an examination of specific copyright issues in LLM development and deployment. A key focus will be on technical approaches to copyright risk assessment and mitigation in LLMs. We will introduce benchmarks for evaluating copyright-related risks, including memorization detection and probing techniques. The tutorial will then cover practical mitigation strategies, such as machine unlearning, efficient fine-tuning methods, and alignment approaches to reduce copyright infringement risks. Ethical considerations and future directions in copyright-aware AI development will also be discussed.

1 Introduction

The landscape of artificial intelligence has been dramatically transformed by the advent of Large Language Models (LLMs) such as GPT and its successors (Lewis et al., 2019; Brown et al., 2020a; OpenAI, 2023; Zhang et al., 2022; Touvron et al., 2023). These powerful systems have not only revolutionized natural language processing but have also permeated diverse sectors including healthcare (Peng et al., 2023; Haupt and Marks, 2023), software development (Chen et al., 2021), finance (Yang et al., 2023; Wang et al., 2023b), and education (Firat, 2023; Fuchs, 2023). While LLMs have unlocked unprecedented capabilities in text generation and analysis, they have simultaneously given rise to complex legal and ethical challenges, particularly in the realm of copyright law. The ability of these models to produce human-like text has blurred the boundaries between original creation and potential copyright infringement, as evidenced by recent New York Times legal actions against AI company (nyt, 2023). This tutorial aims to navigate this intricate terrain, providing a comprehensive exploration of the copyright issues surrounding LLMs and equipping participants with the knowledge and tools to address these challenges.

In this tutorial, we will comprehensively review existing paradigms for copyright risk assessment and mitigation in LLMs, focusing on their contributions to responsible AI development and deployment. We categorize the approaches into probing and benchmarking, influence analysis, unlearning techniques, and finetuning-based behavior regulation. For probing and benchmarking, we will explore methodologies for creating quantitative measures to assess the extent of copyrighted content reproduction in LLM outputs. Influence analysis will cover the application of influence functions to detect and quantify the impact of potentially copyrighted material in training data. We will then examine machine unlearning techniques as a means to selectively remove knowledge of copyrighted content from trained LLMs. Finally, we will present efficient finetuning and alignment methodologies designed to modify LLM behavior with respect to copyright considerations. Participants will learn about recent trends and emerging challenges in copyright-aware LLM research, as well as resources and tools to implement these techniques. The tutorial aims to prompt thorough discussions regarding the impact of copyright considerations on LLM development and the broader implications for AI ethics and governance.

2 Tutorial Outline

We will cover four topics on how to measure and mitigate copyright risks of LLMs. **(1) LLMs Copyright Risks Probing and Benchmarking:** We will introduce and analyze state-of-the-art techniques

Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pages 44–50
May 1, 2025 ©2025 Association for Computational Linguistics

for probing LLMs to identify potential copyright infringements. This section will cover methodologies for creating benchmarks that quantitatively assess the extent of copyrighted content reproduction in LLM outputs. **(2) Influence Function for Copyright Detection in Training Data:** We will examine the application of influence functions, a technique from robust statistics, to detect and quantify the impact of potentially copyrighted material in LLM training datasets to model generations. This approach offers a principled way to trace model outputs back to specific training instances. **(3) Mitigating Copyright Risks via Machine Unlearning:** This section will discuss cutting-edge machine unlearning techniques as a means to selectively remove knowledge of copyrighted content from trained LLMs. We will discuss the theoretical foundations of unlearning algorithms and their practical implementation in the context of large-scale language models. **(4) Efficient Finetuning and Alignment to Regulate LLMs Behavior:** We will review efficient finetuning methodologies and alignment techniques designed to modify LLM behavior with respect to copyright considerations. This includes exploring parameter-efficient tuning methods and reinforcement learning approaches for aligning model outputs with copyright regulations.

3 Specification of the Tutorial

3.1 History

This tutorial has not been presented elsewhere. The presented topic has not been covered by previous AAAI/IJCAI/NeurIPS/ACL/EMNLP/NAACL tutorials in the past five years. The most related tutorial is the AAAI 2024 tutorial "Beyond Human Creativity: A Tutorial on Advancements in AI Generated Content" that covers the foundations, recent advancements, applications, and societal implications of large language models and diffusion models in text, image, video, and 3D object generation. They lack in-depth coverage on measuring and mitigating the potential risks of AI-generated content, which this tutorial aims to address.

3.2 Audience

Based on the level of interest in this topic, we expect around 50-100 participants from machine learning (ML) communities, data mining (DM) communities and natural language processing (NLP) communities.

3.3 Prerequisite Knowledge

While no specific background knowledge is assumed of the audience, it would be beneficial for attendees to have familiarity with basic machine learning and deep learning technologies, as well as pre-trained language models (e.g., BERT (Devlin et al., 2018), GPT (Brown et al., 2020b)) and generative AI concepts.

The following reading list could help provide background knowledge to the audience before attending this tutorial:

- Jialiang Xu, Shenglan Li, Zhaozhuo Xu, and Denghui Zhang. Do LLMs Know to Respect Copyright Notice? In EMNLP, 2024.
- Boyi Wei et al. Evaluating Copyright Takedown Methods for Language Models. In NeurIPS, 2024.
- Tong Chen, Akari Asai, et al. CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation. In EMNLP, 2024.
- Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large Language Model Unlearning. In ICLR, 2024
- Ronen Eldan and Mark Russinovich. Who's Harry Potter? Approximate Unlearning in LLMs. arXiv preprint arXiv:2310.02238, 2023.
- Huawei Lin, Jikai Long, Zhaozhuo Xu, and Weijie Zhao. Token-Wise Influential Training Data Retrieval for Large Language Models. In ACL, 2024.
- Wentao Guo, Jikai Long, Yimeng Zeng, Zirui Liu, Xinyu Yang, Yide Ran, Jacob R. Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, et al. Zeroth-Order Fine-Tuning of LLMs with Extreme Sparsity. In ICML, 2024.

4 Tutorial Content

4.1 LLMs Copyright Risks Probing and Benchmarking [50mins]

We will first discuss the concept of copyright infringement in the context of LLMs, highlighting how it differs across regions such as the U.S. and Europe. Then we review quantitative assessment methodologies for copyright infringement risks in LLMs. Research indicates that model memorization is a primary cause of potential infringement behavior (Carlini et al., 2021; Nasr et al., 2023; Karamolegkou et al., 2023). We will present the "memory profile" tool for statistical causal analysis of LLM memorization (Lesci et al., 2024) to factors like data positions in training steps. The CopyBench framework (Chen et al., 2024) will be presented as a means to quantify literal and non-literal overlap between LLM outputs and copyrighted material. Additionally, we present our recent work for evaluating LLM compliance with copyright restrictions on user-provided content and unauthorized commands (Xu et al., 2024). Through these methodologies, we aim to provide a multi-perspective approach and tutorial to assessing copyright risks in LLMs. This part is estimated to be 45 minutes, with 5 minutes for Q&A.

4.2 Influence Function for Copyright Detection in Training Data [50mins]

We will explore how to identify which training data contributed to a specific generation from a Large Language Model (LLM) using the proposed framework, RapidIn (Lin et al., 2024). For instance, when a generation is found to violate copyright, tracing it back to the most influential training data enables developers to filter out infringing data and retrain the model (Ladhak et al., 2023). Additionally, understanding the influence of training data on a specific generation is critical for tasks like machine unlearning (Yao et al., 2023a; Yu et al., 2023), improving explainability (Zhao et al., 2023), detoxification (Welbl et al., 2021; Dale et al., 2021), data cleansing and combating data poisoning (Yan et al., 2023; Wang et al., 2023a; Huang et al., 2023; Ladhak et al., 2023), as well as preserving privacy and security (Brown et al., 2022; Kandpal et al., 2022). RapidIn estimates the influence of training data by compressing gradient vectors over 200,000x, enabling efficient caching and retrieval, offering a practical solution for handling LLMs at scale and enabling crucial tasks like model retraining and machine unlearning. This part will be $\sim$45 minutes, with 5 minutes for Q&A.

4.3 Mitigating Copyright Risks via Machine Unlearning [50mins]

In this part, we will first review the recent research of applying LLM machine unlearning techniques to address copyright infringement issues. Specifically, Yao et al. (2023b) used a gradient ascent-based approach to unlearn copyrighted contents, while Eldan and Russinovich (2023) explored a similar method to unlearn the Harry Potter series. However, Shostack (2024) pointed out that remnants of the Harry Potter books remained in the modified model. More recently, Chen and Yang (2023) proposed adding unlearning layers in transformer blocks for sequential data forgetting, but this approach was only tested on a smaller model focused on movie reviews in a simulated setting. In contrast, our recent work (Dou et al., 2024) formally studies the copyright unlearning problem under the sequential/continual setting, and tries to address the trade-offs between unlearning efficacy and general knowledge retention based on weight saliency. This part will be around 45 minutes, with 5 minutes for Q&A.

4.4 Efficient Finetuning and Alignment to Regulate LLMs Behavior [50mins]

We propose to regulate the LLMs' behavior with memory-efficient fine-tuning strategies. We will introduce a memory-efficient LLM fine-tuning approach, denoted as Winner-Take-All Column-Row Sampling (WTA-CRS) (Liu et al., 2024), which outperforms the existing Low-Rank Adaptation (LoRA) (Hu et al., 2021) techniques with better memory savings. WTA-CRS reduces the memory required for storing activations during training by selectively sampling matrix elements, enabling larger batch sizes and reducing hardware constraints with minimal accuracy loss. Moreover, we will introduce a series of sparse zeroth-order optimization strategies (Guo et al., 2024) that perform backpropagation-free LLM fine-tuning by perturbing 0.1% of LLM parameters. These techniques are particularly effective in aligning LLM behavior to specific tasks, including compliance with copyright regulations, without incurring excessive computational overhead. By leveraging these innovative techniques, we aim to provide a scalable and practical approach to finetuning LLMs, ensuring they align with desired behavioral outcomes while also addressing resource constraints common in large-scale model deployment. This part will be around 45 minutes, with 5 minutes for Q&A.

4.5 Conclusion and Remaining Challenges [30mins]

We will conclude the tutorial by summarizing the key concepts discussed, including techniques for mitigating copyright risks in LLMs. While progress has been made, several challenges remain: (1) Scaling Up Infringement Detection: Existing methods for detecting potential copyright violations struggle to scale when applied to large databases of copyrighted works. Developing more efficient and scalable approaches remains a critical challenge for the future. (2) Internal Concept Decomposition of Copyright Behavior: A deeper understanding of how LLMs internalize and represent copyrighted material is needed. Current methods focus on literal output analysis, but advances in decomposing internal representations could lead to more effective unlearning techniques and better regulation of model behavior.

These open challenges underscore the need for ongoing research to ensure the responsible use of LLMs while balancing innovation and copyright compliance.

5 Tutorial Presenters

Denghui Zhang is An assistant Professor in the School of Business at Stevens Institute of Technology. He studies the interplay between LLMs/GenAI, legal and socioethical issue, business and innovation. His work is published in refereed venues, such as IEEE TKDE, SIGKDD, EMNLP, AAAI, etc., and won Best Student Paper award at 2023 International Conference on Information Systems.

Zhaozhuo Xu is an Assistant Professor at Stevens Institute of Technology, focusing on scalable and sustainable ML. His work, published in venues like NeurIPS, ICML, and ICLR, has been adopted by Huggingface and startups. He is the organizer of Research On Algorithms & Data Structures (ROADS) to Mega-AI Models Workshop at MLSys 2023.

Weijie Zhao is an Assistant Professor in the department of computer science at RIT. He is investigating a broad collection of exciting problems in big data, machine learning systems, AI security, scientific data processing, and database systems.

David Atkinson, J.D., is a lecturer at the McCombs School of Business at The University of Texas at Austin, where he teaches courses on law, ethics, artificial intelligence, and the legal environment of business. He also serves as legal counsel for the Allen Institute for Artificial Intelligence, specializing in privacy, security, export control, and contracts. Previously, he was senior corporate counsel at the autonomous trucking company TuSimple and a technology scout for the U.S. Army. Additionally, Atkinson has served as legal counsel for the education technology company PowerSchool, worked as an InSITE consultant fellow for startups, and co-founded the legal education company Illustrated Law. He also holds master's degrees from Harvard and Kansas State University and a B.S. from Truman State University.

Boyi Wei is a Ph.D. student at Princeton University, advised by Peter Henderson. His research focuses on aligning machine learning systems, especially on understanding the safety alignment of language models and exploring related legal and policy issues. He received his B.Sc from University of Science and Technology of China.

Xiusi Chen is a Postdoctoral Research Fellow at the University of Illinois Urbana-Champaign, working with Prof. Heng Ji. He received his Ph.D. in Computer Science at the University of California, Los Angeles, advised by Prof. Wei Wang. Xiusi's research focuses on enhancing LLM reasoning, alignment, and decision-making. Xiusi has been awarded the SDM Best Poster Award Honorable Mention. His research has generated over 40 publications in top-tier venues in the fields of data mining, natural language processing, machine learning and information retrieval. Xiusi has been invited as a reviewer or a program committee member for conferences including KDD, ICML, NeurIPS, ICLR, ACL Rolling Review, etc.

Qingyun Wang is the incoming assistant professor at William & Mary. He is among the first researchers to develop a virtual scientific research assistant for literature-based discovery by extracting and synthesizing insights from papers. He received the NAACL-HLT 2021 Best Demo Reward. He has experience presenting tutorials at EMNLP 2021, EACL 2024, and LREC-COLING 2024. He organized AI4Research workshop at AAAI 2025 and the Language + Molecules Workshop at ACL 2024.

Jing Gao is an Associate Professor at Purdue University's Elmore Family School of Electrical and Computer Engineering. Her research spans data mining, focusing on information veracity, crowdsourcing, knowledge graphs, anomaly detection, and multi-source data integration, with applications in healthcare, cybersecurity, and education. She has received the NSF CAREER Award, ICDM Tao Li Award, and SDM/IBM Early Career Data Mining Researcher Award. She also serves as an editor for ACM Transactions on Intelligent Systems and Technology and IEEE Transactions on Knowledge and Data Engineering.

Diversity considerations

In this tutorial, we prioritize diversity considerations across multiple dimensions. We will incorporate the use of multilingual data to demonstrate how the described methods for copyright risk assessment and mitigation scale effectively across various languages and domains, highlighting the global relevance of these challenges. The instructional team will include both senior and junior researchers to ensure a range of perspectives and expertise, promoting an inclusive learning environment. Additionally, we emphasize demographic and geographical diversity among instructors to reflect the international nature of AI ethics and copyright issues. To further encourage diverse audience participation, we plan to actively engage underrepresented groups through targeted outreach efforts and provide accessible resources for participants from varying backgrounds and regions. This approach aims to foster a richer, more inclusive dialogue on the ethical considerations of LLMs.

Ethics Statement

The responsible development and deployment of Large Language Models (LLMs) necessitates a careful balance between innovation and ethical considerations, particularly in the context of copyright law. This tutorial acknowledges the profound implications that LLMs have on intellectual property and emphasizes the importance of addressing these challenges proactively.

We are committed to advancing the understanding of LLMs while ensuring that their capabilities are used ethically and in accordance with legal frameworks. Through this tutorial, we aim to foster transparency in AI systems, promote the use of copyright-aware methodologies, and advocate for responsible AI practices. We encourage participants to engage in thoughtful discussions regarding the broader societal impact of LLMs, especially in terms of fairness, accountability, and respect for the rights of content creators.

By equipping attendees with tools for assessing and mitigating copyright risks, we hope to contribute to the development of AI systems that are not only powerful but also aligned with ethical principles and legal norms.

References

2023. New york times company v. microsoft corporation and openai, inc. Filed Dec. 27, 2023.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292, Seoul, Republic of Korea.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

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. *ArXiv preprint*, abs/2407.07087.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 7979–7996, Virtual Event / Punta Cana, Dominican Republic.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. 2024. Avoiding copyright infringement via machine unlearning. *arXiv preprint arXiv:2406.10952*.

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

Mehmet Firat. 2023. How chat gpt can transform autodidactic experiences and open education?

Kevin Fuchs. 2023. Exploring the opportunities and challenges of nlp models in higher education: is chat gpt a blessing or a curse? In *Frontiers in Education*, volume 8, page 1166682. Frontiers Media SA.

Wentao Guo, Jikai Long, Yimeng Zeng, Zirui Liu, Xinyu Yang, Yide Ran, Jacob R Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, et al. 2024. Zeroth-order fine-tuning of llms with extreme sparsity. *arXiv preprint arXiv:2406.02913*.

Claudia E Haupt and Mason Marks. 2023. Ai-generated medical advice—gpt and beyond. *Jama*, 329(16):1349–1350.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. 2023. Composite backdoor attacks against large language models. *CoRR*, abs/2310.07676.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning, ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707, Baltimore, Maryland.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. *ArXiv preprint*, abs/2310.13771.

Faisal Ladhak, Esin Durmus, and Tatsunori Hashimoto. 2023. Contrastive error attribution for finetuned language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 11482–11498, Toronto, Canada.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2024. Causal estimation of memorisation profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15616–15635. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Huawei Lin, Jikai Long, Zhaozhuo Xu, and Weijie Zhao. 2024. Token-wise influential training data retrieval for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 841–860. Association for Computational Linguistics.

Zirui Liu, Guanchu Wang, Shaochen Henry Zhong, Zhaozhuo Xu, Daochen Zha, Ruixiang Ryan Tang, Zhimeng Stephen Jiang, Kaixiong Zhou, Vipin Chaudhary, Shuai Xu, et al. 2024. Winner-take-all column row sampling for memory efficient adaptation of language model. *Advances in Neural Information Processing Systems*, 36.

Milad Nasr, N Carlini, J Hayase, M Jagielski, AF Cooper, D Ippolito, CA Choquette-Choo, E Wallace, F Tramer, and K Lee. 2023. Scalable extraction of training data from (production) language models (2023). *ArXiv preprint*, abs/2311.17035.

OpenAI. 2023. Gpt-4 technical report. In *arXiv*.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Adam Shostack. 2024. The boy who survived: Removing harry potter from an llm is harder than reported. *arXiv preprint arXiv:2403.12082*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. 2023a. Adversarial demonstration attacks on large language models. *CoRR*, abs/2305.14950.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023b. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *NeurIPS Workshop on Instruction Tuning and Instruction Following*.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages

2447–2469, Virtual Event / Punta Cana, Dominican Republic.

Jialiang Xu, Shenglan Li, Zhaozhuo Xu, and Denghui Zhang. 2024. Do llms know to respect copyright notice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2023. Backdooring instruction-tuned large language models with virtual prompt injection. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning-The Good, the Bad, and the Ugly*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023a. Large language model unlearning. *CoRR*, abs/2310.10683.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023b. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL*, pages 6032–6048, Toronto, Canada.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *CoRR*, abs/2309.01029.