The Space Complexity of Learning-Unlearning Algorithms

Yeshwanth Cherapanamjeri[*] YESH@MIT.EDU
Massachusetts Institute of Technology

Sumegha Garg SG1957@CS.RUTGERS.EDU
Rutgers University

Nived Rajaraman NIVED.RAJARAMAN@BERKELEY.EDU
University of California, Berkeley

Ayush Sekhari AYUSH@SEKHARI.COM
Boston University

Abhishek Shetty SHETTY@MIT.EDU
Massachusetts Institute of Technology

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study the memory complexity of machine unlearning algorithms that provide strong data deletion guarantees to the users. Formally, consider an algorithm for a particular learning *task* that initially receives a training dataset. Then, after learning, it receives data deletion requests from a subset of users (of arbitrary size), and the goal of unlearning is to perform the *task* as if the learner never received the data of deleted users. In this paper, we ask how many bits of storage are needed to be able to delete certain training samples, at a later time. We focus on the task of realizability testing, where the goal is to check whether the remaining training samples are realizable within a given hypothesis class $\mathcal{H}$.

Toward that end, we first provide a negative result showing that the VC dimension, a well-known combinatorial property of $\mathcal{H}$ that characterizes the amount of information needed for learning and representing the ERM hypothesis in the standard PAC learning task—is not a characterization of the space complexity of unlearning. In particular, we provide a hypothesis class with constant VC dimension (and Littlestone dimension), but for which any unlearning algorithm for realizability testing needs to store $\Omega(n)$-bits, where n denotes the size of the initial training dataset. In fact, we provide a stronger separation by showing that for any hypothesis class $\mathcal{H}$, the amount of information that the learner needs to store, so as to perform unlearning later, is lower bounded by the *eluder dimension* of $\mathcal{H}$, a combinatorial notion always larger than the VC dimension. We complement the lower bound with an upper bound in terms of the star number of the underlying hypothesis class, albeit in a stronger ticketed-memory model proposed by Ghazi et al. (2023). We show that for any class $\mathcal{H}$ with bounded star number, there exists a ticketed scheme that uses only $\widetilde{O}(\mathrm{StarNo}(\mathcal{H}))$ many bits of storage and these many sized tickets. Since, the star number for a hypothesis class is never larger than its Eluder dimension, our work highlights a fundamental separation between central and ticketed memory models for machine unlearning.

Lastly, we consider the setting where the number of deletions is *bounded* and show that in contrast to the unbounded setting, there exist unlearning schemes with sublinear (in n) storage for hypothesis classes with bounded *hollow star number*, a notion of complexity that is always smaller than the star number and the eluder dimension.[1]

Keywords: Unlearning, VC/Littlestone dimension, Star Number, space complexity, lower bounds

[*] Authors are listed in alphabetical order of their last names.

[1] Extended abstract. Full version appears as [arXiv:2506.13048, v1].

References

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29, 1996.

Hagit Attiya, Michael A Bender, Martin Farach-Colton, Rotem Oshman, and Noa Schiller. History-independent concurrent objects. *arXiv preprint arXiv:2403.14445*, 2024.

Paul Beame, Shayan Oveis Gharan, and Xin Yang. Time-space tradeoffs for learning finite functions from random evaluations, with applications to polynomials. In *Conference On Learning Theory*, pages 843–856. PMLR, 2018.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36, 2023.

Michael A Bender, Martín Farach-Colton, Michael T Goodrich, and Hanna Komlós. History-independent dynamic partitioning: Operation-order privacy in ordered data structures. *Proceedings of the ACM on Management of Data*, 2(2):1–27, 2024.

Moise Blanchard. Gradient descent is pareto-optimal in the oracle complexity and memory tradeoff for feasibility problems, 2024. URL https://arxiv.org/abs/2404.06720.

Guy E Blelloch and Daniel Golovin. Strongly history-independent hashing with applications. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 272–282. IEEE, 2007.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *proceedings of the 42nd IEEE Symposium on Security and Privacy*, SP '21. IEEE Computer Society, 2021.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal svm bound. In *Conference on Learning Theory*, pages 582–609. PMLR, 2020a.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal SVM bound. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 582–609. PMLR, 2020b. URL http://proceedings.mlr.press/v125/bousquet20a.html.

Mark Braverman, Gillat Kol, Shay Moran, and Raghuvansh R. Saxena. Convex set disjointness, distributed learning of halfspaces, and LP feasibility. *CoRR*, abs/1909.03547, 2019. URL http://arxiv.org/abs/1909.03547.

Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *Proceedings of the 38th International Conference on Machine Learning*, ICML '21, pages 1092–1104. JMLR, Inc., 2021.

Niv Buchbinder and Erez Petrank. Lower and upper bounds on obtaining history independence. In *Advances in Cryptology-CRYPTO 2003: 23rd Annual International Cryptology Conference, Santa Barbara, California, USA, August 17-21, 2003. Proceedings 23*, pages 445–462. Springer, 2003.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *S & P*, pages 1897–1914, 2022.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023a.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *ICLR*, 2023b.

Gert Cauwenberghs and Tomaso Poggio. Incremental and decremental support vector machine learning. *NIPS*, 2000.

CCPA. California consumer privacy act (ccpa). https://oag.ca.gov/privacy/ccpa.

Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. *arXiv preprint arXiv:2401.10371*, 2024.

Rishav Chourasia, Neil Shah, and Reza Shokri. Forget unlearning: Towards true data-deletion in machine learning. In *ICML*, 2023.

Aloni Cohen and Kobbi Nissim. Towards formalizing the gdprs notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, 2020.

Aloni Cohen, Adam Smith, Marika Swanberg, and Prashant Nalini Vasudevan. Control, confidentiality, and the right to be forgotten. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3358–3372, 2023.

R Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980.

T Cover and M Hellman. The two-armed-bandit problem with time-invariant finite memory. *IEEE Transactions on Information Theory*, 16(2):185–195, 1970.

Thomas M Cover. Hypothesis testing with finite statistics. *The Annals of Mathematical Statistics*, 40(3):828–835, 1969.

Thomas M Cover, Michael A Freedman, and Martin E Hellman. Optimal finite memory learning algorithms for the finite sample problem. *Information and Control*, 30(1):49–85, 1976.

Min Du, Zhi Chen, Chang Liu, Rajvardhan Oak, and Dawn Song. Lifelong anomaly detection through unlearning. In *CCS*, pages 1283–1297, 2019.

Yonatan Dukler, Benjamin Bowman, Alessandro Achille, Aditya Golatkar, Ashwin Swaminathan, and Stefano Soatto. SAFE: Machine unlearning with shard graphs. *arXiv preprint arXiv:2304.13169*, 2023.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

Thorsten Eisenhofer, Doreen Riepel, Varun Chandrasekaran, Esha Ghosh, Olga Ohrimenko, and Nicolas Papernot. Verifiable and provably secure machine unlearning. *arXiv preprint arXiv:2210.09126*, 2022.

Ran El-Yaniv and Yair Wiener. Active learning via perfect selective classification. *The Journal of Machine Learning Research*, 13(1):255–279, 2012.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

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

Philippe Flajolet and G Nigel Martin. Probabilistic counting algorithms for data base applications. *Journal of computer and system sciences*, 31(2):182–209, 1985.

Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. Formalizing data deletion in the context of the right to be forgotten. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 373–402. Springer, 2020.

Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. In *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, STOC '18, pages 990–1002, New York, NY, USA, 2018. ACM.

GDPR. Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016. *Official Journal of the European Union*.

Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Ayush Sekhari, and Chiyuan Zhang. Ticketed learning-unlearning schemes, 2023.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 3518–3531. Curran Associates, Inc., 2019.

Jonathan Godin and Philippe Lamontagne. Deletion-compliance in the absence of privacy. In *2021 18th International Conference on Privacy, Security and Trust (PST)*, pages 1–10. IEEE, 2021.

Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022.

Shashwat Goel, Ameya Prabhu, Philip Torr, Ponnurangam Kumaraguru, and Amartya Sanyal. Corrective machine unlearning. *arXiv preprint arXiv:2402.14015*, 2024.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. *arXiv:2003.02960*, 2020b.

Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the 2021 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR '21, pages 792–801. IEEE Computer Society, 2021.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524, 2021.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pages 3832–3842. JMLR, Inc., 2020.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21, pages 16319–16330. Curran Associates, Inc., 2021.

Steve Hanneke. *Theoretical foundations of active learning*. Carnegie Mellon University, 2009.

Steve Hanneke. The star number and eluder dimension: Elementary observations about the dimensions of disagreement. *J. Mach. Learn. Res.*, 247, 2024.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(1): 3487–3602, 2015.

Jason D Hartline, Edwin S Hong, Alexander E Mohr, William R Pentney, and Emily C Rocke. Characterizing history independent data structures. *Algorithmica*, 42:57–74, 2005.

Martin Edward Hellman. *Learning with finite memory*. Stanford University, 1969.

Yiyang Huang and Clément L Canonne. Tight bounds for machine unlearning via differential privacy. *arXiv:2309.00886*, 2023.

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

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

Victor Kac and Pokman Cheung. *q-Binomial Coefficients and Linear Algebra over Finite Fields*, pages 21–26. Springer New York, New York, NY, 2002. ISBN 978-1-4613-0071-7. doi: 10.1007/978-1-4613-0071-7_7. URL https://doi.org/10.1007/978-1-4613-0071-7_7.

Masayuki Karasuyama and Ichiro Takeuchi. Multiple incremental decremental learning of support vector machines. *IEEE Transactions on Neural Networks*, 21(7):1048–1059, 2010.

Gillat Kol, Ran Raz, and Avishay Tal. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1067–1080, 2017.

Satyapriya Krishna, Jiaqi Ma, and Himabindu Lakkaraju. Towards bridging the gaps between the right to explanation and the right to be forgotten. In *ICML*, 2023.

Meghdad Kurmanji, Peter Triantafillou, and Eleni Triantafillou. Towards unbounded machine unlearning. *arXiv preprint arXiv:2302.09880*, 2023.

Gene Li, Pritish Kamath, Dylan J Foster, and Nati Srebro. Understanding the eluder dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750, 2022.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. *Unpublished manuscript*, 1986.

Annie Marsden, Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Efficient convex optimization requires superlinear memory. In *Conference on Learning Theory*, pages 2390–2430. PMLR, 2022.

Andrew McGregor. Graph stream algorithms: a survey. *ACM SIGMOD Record*, 43(1):9–20, 2014.

Jayadev Misra and David Gries. Finding repeated elements. *Science of computer programming*, 2 (2):143–152, 1982.

Shay Moran and Amir Yehudayoff. Sample compression schemes for vc classes. *Journal of the ACM (JACM)*, 63(3):1–10, 2016.

Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Conference on Learning Theory*, pages 1516–1566. PMLR, 2017.

Dana Moshkovitz and Michal Moshkovitz. Entropy samplers and strong generic lower bounds for space bounded learning. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 28:1–28:20, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Wenlong Mou, Zheng Wen, and Xi Chen. On the sample complexity of reinforcement learning with policy space generalization. *arXiv preprint arXiv:2008.07353*, 2020.

J Ian Munro and Mike S Paterson. Selection and sorting with limited storage. *Theoretical computer science*, 12(3):315–323, 1980.

Shanmugavelayutham Muthukrishnan et al. Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2):117–236, 2005.

Moni Naor and Vanessa Teague. Anti-persistence: History independent data structures. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 492–501, 2001.

Moni Naor, Gil Segev, and Udi Wieder. History-independent cuckoo hashing. In *Automata, Languages and Programming: 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7-11, 2008, Proceedings, Part II 35*, pages 631–642. Springer, 2008.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21. JMLR, Inc., 2021.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational Bayesian unlearning. In *NeurIPS*, pages 16025–16036, 2020.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv:2209.02299*, 2022.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*, 2023.

Martin Pawelczyk, Jimmy Z Di, Yiwei Lu, Gautam Kamath, Ayush Sekhari, and Seth Neel. Machine unlearning fails to remove data poisoning attacks. *arXiv preprint arXiv:2406.17216*, 2024.

Binghui Peng and Aviad Rubinstein. Near optimal memory-regret tradeoff for online learning. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1171–1194. IEEE, 2023.

Jeff M Phillips and Jeff M Phillips. Big data and sketching. *Mathematical Foundations for Data Analysis*, pages 261–281, 2021.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR, 2022a.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, 2022b.

Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 266–275. IEEE Computer Society, 2016.

Enrique Romero, Ignacio Barrio, and Lluís Belanche. Incremental and decremental learning for linear support vector machines. In *ICANN*, pages 209–218, 2007.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *NeurIPS*, pages 18075–18086, 2021.

Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. *Advances in Neural Information Processing Systems*, 27, 2014.

Vatsal Sharan, Aaron Sidford, and Gregory Valiant. Memory-sample tradeoffs for linear regression with small error. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 890–901, 2019.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *S & P*, pages 3–18, 2017.

Vaidehi Srinivas, David P Woodruff, Ziyu Xu, and Samson Zhou. Memory bounds for the experts problem. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1158–1171, 2022.

Jacob Steinhardt and John Duchi. Minimax rates for memory-bounded sparse linear regression. In *Conference on Learning Theory*, pages 1564–1587. PMLR, 2015.

Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory*, pages 1490–1516. PMLR, 2016.

Vinith M Suriyakumar and Ashia C Wilson. Algorithms that approximate data removal: New results and limitations. In *NeurIPS*, 2022.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding factors influencing machine unlearning. In *EuroS&P*, pages 303–319, 2022.

Amund Tveit, Magnus Lie Hetland, and Håavard Engum. Incremental and decremental proximal support vector classification using decay coefficients. In *DaWak*, pages 422–429, 2003.

Vladimir N Vapnik. The nature of statistical learning theory, 1995.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*, 2023.

Yair Wiener, Steve Hanneke, and Ran El-Yaniv. A compression technique for analyzing disagreement-based active learning. *J. Mach. Learn. Res.*, 16:713–745, 2015.

Blake Woodworth and Nathan Srebro. Open problem: The oracle complexity of convex optimization with limited memory. In *Conference on Learning Theory*, pages 3202–3210. PMLR, 2019.

Yinjun Wu, Edgar Dobriban, and Susan Davidson. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning (ICML)*, 2020.

Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of updates to natural language models. In *CCS*, pages 363–375, 2020.

Rui Zhang and Shihua Zhang. Rethinking influence functions of neural networks in the over-parameterized regime. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.