[25] Marieke Bak, Vince Istvan Madai, Marie-Christine Fritzsche, Michaela Th. Mayrhofer, and Stuart McLennan. You can't have ai both ways: Balancing health data privacy and access fairly. *Frontiers in Genetics*, 13, 2022. https://www.frontiersin.org/articles/10.3389/fgene.2022.929453. doi:10.3389/fgene.2022.929453.

[26] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *NeurIPS 2021 Workshop on Privacy in Machine Learning (PRIML)*, 2021. URL: https://openreview.net/forum?id=Yi2DZTbnBl4, doi:10.48550/arXiv.2201.04845.

[27] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017. doi:10.48550/ARXIV.1705.09406.

[28] Anthony M. Barrett, Dan Hendrycks, Jessica Newman, and Brandie Nonnecke. *UC Berkeley AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models*. UC Berkeley Center for Long Term Cybersecurity, 2023. https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/. doi:10.48550/ARXIV.2206.08966.

[29] Lejla Batina, Shivam Bhasin, Dirmanto Jap, and Stjepan Picek. CSI NN: Reverse engineering of neural network architectures through electromagnetic side channel. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 515–532, USA, 2019. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity19/presentation/batina.

[30] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.*, 38(8):2939–2970, August 2022. doi:10.1007/s00371-021-02166-7.

[31] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL: https://arxiv.org/abs/2303.08112, doi:10.48550/arXiv.2303.08112.

[32] Philipp Benz, Chaoning Zhang, Soomin Ham, Gyusang Karjauv, Adil Cho, and In So Kweon. The triangular trade-off between accuracy, robustness, and fairness. *Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV) at CVPR*, 2021. URL: https://dl.acm.org/doi/10.1145/3645088, doi:10.1145/3645088.

[33] Jamie Bernardi, Gabriel Mukobi, Hilary Greaves, Lennart Heim, and Markus Anderljung. Societal adaptation to advanced ai, 2024. URL: https://arxiv.org/abs/2405.10295, arXiv:2405.10295, doi:10.48550/arXiv.2405.10295.

[34] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5050–5060.

Curran Associates, Inc., 2019. URL: http://papers.nips.cc/paper/8749-mixmatch-a-holistic-approach-to-semi-supervised-learning.pdf, doi:10.48550/arXiv.2405.10295.

[35] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Model Poisoning Attacks in Federated Learning. In *NeurIPS SECML*, 2018.

[36] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 634–643. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/bhagoji19a.html, doi:10.48550/arXiv.1811.12470.

[37] Battista Biggio, Igino Corona, Giorgio Fumera, Giorgio Giacinto, and Fabio Roli. Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. In *Proceedings of the 10th International Conference on Multiple Classifier Systems*, MCS'11, page 350–359, Berlin, Heidelberg, 2011. Springer-Verlag. URL: https://api.semanticscholar.org/CorpusID:12680508.

[38] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013. doi:10.1007/978-3-642-40994-3_25.

[39] Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In Chun-Nan Hsu and Wee Sun Lee, editors, *Proceedings of the Asian Conference on Machine Learning*, volume 20 of *Proceedings of Machine Learning Research*, pages 97–112, South Garden Hotels and Resorts, Taoyuan, Taiwain, 14–15 Nov 2011. PMLR. URL: https://proceedings.mlr.press/v20/biggio11.html, doi:10.48550/arXiv.2206.00352.

[40] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning, ICML*, 2012. URL: https://arxiv.org/abs/1206.6389, doi:10.48550/arXiv.1206.6389.

[41] Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, and Fabio Roli. Poisoning behavioral malware clustering. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, AISec '14, page 27–36, New York, NY, USA, 2014. Association for Computing Machinery. doi:10.1145/2666652.2666666.

[42] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, December 2018. URL: https://doi.org/10.1016%2Fj.patcog.2018.07.023, doi:10.1016/j.patcog.2018.07.023.

[43] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *NeurIPS*, 2017.

URL: https://papers.nips.cc/paper_files/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf.

[44] Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. The foundation model transparency index v1.1: May 2024, 2024. URL: https://arxiv.org/abs/2407.12929, arXiv:2407.12929, doi:10.48550/arXiv.2407.12929.

[45] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 141–159. IEEE, 2021. doi:10.1109/SP40001.2021.00019.

[46] Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling laws for data poisoning in llms, 2024. URL: https://arxiv.org/abs/2408.02946, arXiv:2408.02946, doi:10.48550/arXiv.2408.02946.

[47] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=SyZI0GWCZ, doi:10.48550/arXiv.1712.04248.

[48] Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 123–132, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3406325.3451131.

[49] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL: https://arxiv.org/abs/2005.14165, arXiv:2005.14165.

[50] Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, and Michal Irani. Reconstructing training data from multiclass neural networks, 2023. URL: https://arxiv.org/abs/2305.03350, arXiv:2305.03350, doi:10.48550/arXiv.2305.03350.

[51] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FLTrust: Byzantine-robust federated learning via trust bootstrapping. In *NDSS*, 2021. URL: https://arxiv.org/abs/2012.13995, doi:10.48550/arXiv.2012.13995.

[52] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015. URL: https://ieeexplore.ieee.org/document/7163042, doi:10.1109/SP.2015.35.

[53] Nicholas Carlini. Poisoning the unlabeled dataset of Semi-Supervised learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1577–1592. USENIX Association, August 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-poisoning.

[54] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (S&P)*, pages 1519–1519, Los Alamitos, CA, USA, May 2022. IEEE Computer Society. URL: https://doi.ieeecomputersociety.org/10.1109/SP46214.2022.00090, doi:10.1109/SP46214.2022.00090.

[55] 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, 2023. URL: https://arxiv.org/abs/2301.13188, arXiv:2301.13188, doi:10.48550/arXiv.2301.13188.

[56] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. https://arxiv.org/abs/2202.07646, 2022. doi:10.48550/ARXIV.2202.07646.

[57] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023. URL: https://arxiv.org/abs/2302.10149, doi:10.48550/arXiv.2302.10149.

[58] Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. Cryptanalytic extraction of neural network models. In Daniele Micciancio and Thomas Ristenpart, editors, *Advances in Cryptology – CRYPTO 2020*, pages 189–218, Cham, 2020. Springer International Publishing. URL: https://arxiv.org/abs/2003.04884, doi:10.48550/arXiv.2003.04884.

[59] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The Secret Sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, USENIX '19), pages 267–284, 2019. https://arxiv.org/abs/1802.08232. URL: https://arxiv.org/abs/1802.08232, doi:10.48550/arXiv.1802.08232.

[60] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023. URL: https://arxiv.org/abs/2306.15447, doi:10.48550/arXiv.2306.15447.

[61] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Itay Yona, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model, 2024. URL: https://arxiv.org/abs/2403.06634, arXiv:2403.06634, doi:10.48550/arXiv.2403.06634.

[62] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie

Sun, and J. Zico Kolter. (certified!!) adversarial robustness for free!, 2023. URL: https://arxiv.org/abs/2206.10550, arXiv:2206.10550, doi:10.48550/arXiv.2206.10550.

[63] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[64] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, page 3–14, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3128572.3140444.

[65] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proc. IEEE Security and Privacy Symposium*, 2017. URL: https://arxiv.org/abs/1608.04644, doi:10.48550/arXiv.1608.04644.

[66] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018. URL: https://arxiv.org/abs/1801.01944, doi:10.48550/arXiv.1801.01944.

[67] Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell. Red teaming deep neural networks with feature synthesis tools. *arXiv preprint arXiv:2302.10894*, 2023. URL: https://arxiv.org/abs/2302.10894, doi:10.48550/arXiv.2302.10894.

[68] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. Explore, establish, exploit: Red teaming language models from scratch, 2023. URL: https://arxiv.org/abs/2306.09442, arXiv:2306.09442, doi:10.48550/arXiv.2306.09442.

[69] National Cyber Security Center. Introducing our new machine learning security principles, retrieved February 2023 from https://www.ncsc.gov.uk/blog-post/introducing-our-new-machine-learning-security-principles. URL: https://www.ncsc.gov.uk/blog-post/introducing-our-new-machine-learning-security-principles.

[70] Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. Exploring connections between active learning and model extraction. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. URL: https://arxiv.org/abs/1811.02054, doi:10.48550/arXiv.1811.02054.

[71] Hong Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and R. Shokri. On adversarial bias and the robustness of fair machine learning. https://arxiv.org/abs/2006.08669, 2020.

[72] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J.

Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models, 2024. URL: https://arxiv.org/abs/2404.01318, arXiv:2404.01318, doi:10.48550/arXiv.2404.01318.

[73] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023. URL: https://arxiv.org/abs/2310.08419, doi:10.48550/arXiv.2310.08419.

[74] Harsh Chaudhari, John Abascal, Alina Oprea, Matthew Jagielski, Florian Tramèr, and Jonathan Ullman. SNAP: Efficient extraction of private properties with poisoning. In *2023 IEEE Symposium on Security and Privacy (S&P)*, 2023. URL: https://arxiv.org/abs/2208.12348, doi:10.48550/arXiv.2208.12348.

[75] Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A. Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General trigger attacks on retrieval augmented language generation, 2024. URL: https://arxiv.org/abs/2405.20485, arXiv:2405.20485, doi:10.48550/arXiv.2405.20485.

[76] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. https://arxiv.org/abs/1811.03728, 2018. URL: https://arxiv.org/abs/1811.03728, doi:10.48550/arXiv.1811.03728.

[77] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. https://arxiv.org/abs/1712.02051, 2017. URL: https://arxiv.org/abs/1712.02051, doi:10.48550/ARXIV.1712.02051.

[78] Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. DeepInspect: A black-box trojan detection and mitigation framework for deep neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4658–4664. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi:10.24963/ijcai.2019/647.

[79] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. HopSkipJumpAttack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1277–1294. IEEE, 2020. doi:10.1109/SP40000.2020.00045.

[80] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec ’17, page 15–26, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3128572.3140448.

[81] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. *ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector*, page 52–68. Springer International Publishing, 2019. URL: http://dx.doi.org/10.1007/978-3-030-10925-7_4, doi:10.1007/978-3-030-10925-7_4.

[82] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Annual Computer Security Applications Conference*, ACSAC '21, page 554–569, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3485832.3485837.

[83] Xiaoyi Chen, Siyuan Tang, Rui Zhu, Shijun Yan, Lei Jin, Zihao Wang, Liya Su, Zhikun Zhang, XiaoFeng Wang, and Haixu Tang. The Janus interface: How fine-tuning in large language models amplifies the privacy risks, 2024. URL: https://arxiv.org/abs/2310.15469, arXiv:2310.15469, doi:10.48550/arXiv.2310.15469.

[84] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. URL: https://arxiv.org/abs/1712.05526, doi:10.48550/arXiv.1712.05526.

[85] Heng-Tze Cheng and Romal Thoppilan. LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything. https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html, 2022. Google Brain. URL: https://research.google/blog/lamda-towards-safe-grounded-and-high-quality-dialog-models-for-everything/.

[86] Minhao Cheng, Thong Le, Pin-Yu Chen, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=rJlk6iRqKX, doi:10.48550/arXiv.1807.04457.

[87] Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=SklTQCNtvS, doi:10.48550/arXiv.1909.10773.

[88] Alesia Chernikova and Alina Oprea. FENCE: Feasible evasion attacks on neural networks in constrained environments. *ACM Transactions on Privacy and Security (TOPS) Journal*, 2022. URL: https://arxiv.org/abs/1909.10480, doi:10.48550/arXiv.1909.10480.

[89] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/choquette-choo21a.html, doi:10.48550/arXiv.2007.14321.

[90] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? https://arxiv.org/abs/2212.05400, 2022. doi:10.48550/ARXIV.2212.05400.

[91] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security

against training data poisoning. *ACM Computing Surveys*, March 2023. URL: https://doi.org/10.1145%2F3585385, doi:10.1145/3585385.

[92] Jack Clark and Raymond Perrault. 2022 AI index report. https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf, 2022. Human Centered AI, Stanford University.

[93] Joseph Clements, Yuzhe Yang, Ankur Sharma, Hongxin Hu, and Yingjie Lao. Rallying adversarial techniques against deep learning for network security, 2019. URL: https://arxiv.org/abs/1903.11688, doi:10.48550/ARXIV.1903.11688.

[94] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/cohen19c.html.

[95] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[96] Gabriela F. Cretu, Angelos Stavrou, Michael E. Locasto, Salvatore J. Stolfo, and Angelos D. Keromytis. Casting out demons: Sanitizing training data for anomaly sensors. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 81–95, 2008. URL: https://ieeexplore.ieee.org/document/4531146, doi:10.1109/SP.2008.11.

[97] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL: https://openreview.net/forum?id=SSKZPJCt7B, doi:10.48550/arXiv.2010.09670.

[98] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’04, page 99–108, New York, NY, USA, 2004. Association for Computing Machinery. doi:10.1145/1014052.1014066.

[99] DARPA. DARPA AI Cyber Challenge Aims to Secure Nation’s Most Critical Software, 2023. Accessed: 2024-08-22. URL: https://www.darpa.mil/news-events/2023-08-09.

[100] Emiliano De Cristofaro. A critical overview of privacy in machine learning. *IEEE Security & Privacy*, 19(4):19–27, 2021. doi:10.1109/MSEC.2021.3076443.

[101] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate attacks and defenses for LLM agents, 2024. URL: https://arxiv.org/abs/2406.13352, arXiv:2406.13352, doi:10.48550/arXiv.2406.13352.

[102] DeepMind. Building safer dialogue agents. https://www.deepmind.com/blog/building-safer-dialogue-agents, 2022. Online.

[103] Luca Demetrio, Battista Biggio, Giovanni Lagorio, Fabio Roli, and Alessandro Armando. Functionality-preserving black-box optimization of adversarial windows malware. *IEEE Transactions on Information Forensics and Security*, 16:3469–3478, 2021. URL: https://ieeexplore.ieee.org/document/9437194, doi:10.1109/TIFS.2021.3082330.

[104] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 321–338. USENIX Association, 2019. URL: https://www.usenix.org/conference/usenixsecurity19/presentation/demontis.

[105] Serguei Denissov, Hugh Brendan McMahan, J Keith Rush, Adam Smith, and Abhradeep Guha Thakurta. Improved differential privacy for SGD via optimal private linear operators on adaptive streams. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL: https://openreview.net/forum?id=i9XrHJoyLqJ, doi:10.48550/arXiv.2202.08312.

[106] Leon Derczynski. Garak: LLM vulnerability scanner. https://github.com/leondz/garak, 2024. Accessed: 2024-08-18.

[107] Leon Derczynski, Erick Galinkin, Jeffrey Martin, Subho Majumdar, and Nanna Inie. garak: A framework for security probing large language models, 2024. URL: https://arxiv.org/abs/2406.11036, arXiv:2406.11036, doi:10.48550/arXiv.2406.11036.

[108] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=OUIFPHEgJU, doi:10.48550/arXiv.2305.14314.

[109] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019. URL: https://arxiv.org/abs/1803.02815, doi:10.48550/arXiv.1803.02815.

[110] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM Symposium on Principles of Database Systems*, PODS ’03, pages 202–210. ACM, 2003. URL: https://crypto.stanford.edu/seclab/sem-03-04/psd.pdf.

[111] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R. Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *Proceedings of the 37th International Conference on Machine Learning*, ICML’20. JMLR.org, 2020. URL: https://arxiv.org/abs/1910.07870, doi:10.48550/arXiv.1910.07870.

[112] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceed-*

*ings, Part II*, pages 1–12, 2006. URL: http://dx.doi.org/10.1007/11787006_1, doi:10.1007/11787006_1.

[113] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Conference on Theory of Cryptography*, TCC '06, pages 265–284, New York, NY, USA, 2006.

[114] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! A survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017. URL: https://privacytools.seas.harvard.edu/publications/exposed-survey-attacks-private-data.

[115] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *IEEE Symposium on Foundations of Computer Science*, FOCS '15, 2015. URL: https://privacytools.seas.harvard.edu/files/privacytools/files/robust.pdf.

[116] Cynthia Dwork and Sergey Yekhanin. New efficient attacks on statistical disclosure control mechanisms. In *Annual International Cryptology Conference*, pages 469–480. Springer, 2008. URL: https://link.springer.com/chapter/10.1007/978-3-540-85174-5_26, doi:0.1007/978-3-540-85174-5_26.

[117] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017. URL: https://arxiv.org/abs/1712.06751, doi:10.48550/arXiv.1712.06751.

[118] Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting LLM Quantization, 2024. URL: https://arxiv.org/abs/2405.18137, arXiv:2405.18137, doi:10.48550/arXiv.2405.18137.

[119] Gemini Team et al. Gemini: A family of highly capable multimodal models. https://arxiv.org/abs/2312.11805, 2023. arXiv:2312.11805.

[120] ETSI Group Report SAI 005. Securing artificial intelligence (SAI); mitigation strategy report, retrieved February 2023 from https://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf. URL: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/005/01.01.01_60/gr_SAI005v010101p.pdf.

[121] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. URL: https://ieeexplore.ieee.org/document/8578273, doi:10.1109/CVPR.2018.00175.

[122] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1625–1634. Computer Vision Foundation / IEEE Computer Society, 2018. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.html, doi:10.1109/CVPR.2018.00175.

[123] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local Model Poi-

soning Attacks to Byzantine-Robust Federated Learning. In *USENIX Security*, 2020.

[124] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. online: https://arxiv.org/abs/2210.14707, 2022. doi:10.48550/ARXIV.2210.14707.

[125] Georgios Fatouros, John Soldatos, Kalliopi Kouroumali, Georgios Makridis, and Dimosthenis Kyriazis. Transforming sentiment analysis in the financial domain with chatgpt. *Machine Learning with Applications*, 14:100508, 2023. URL: https://www.sciencedirect.com/science/article/pii/S2666827023000610, doi:10.1016/j.mlwa.2023.100508.

[126] Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *ACM Symposium on Theory of Computing*, STOC '20, pages 954–959, 2020. https://arxiv.org/abs/1906.05271.

[127] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. URL: https://arxiv.org/abs/2008.03703, doi:10.48550/arXiv.2008.03703.

[128] Ji Feng, Qi-Zhi Cai, and Zhi-Hua Zhou. Learning to confuse: Generating training time adversarial data with auto-encoder. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/1ce83e5d4135b07c0b82afffbe2b3436-Paper.pdf, doi:10.48550/arXiv.1905.09027.

[129] Liam Fowl, Ping-yeh Chiang, Micah Goldblum, Jonas Geiping, Arpit Bansal, Wojtek Czaja, and Tom Goldstein. Preventing unauthorized use of proprietary data: Poisoning for secure dataset release, 2021. URL: https://arxiv.org/abs/2103.02683, doi:10.48550/ARXIV.2103.02683.

[130] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2810103.2813677.

[131] Aymeric Fromherz, Klas Leino, Matt Fredrikson, Bryan Parno, and Corina Pasareanu. Fast geometric projections for local robustness certification. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=zWy1uxjDdZJ, doi:10.48550/arXiv.2002.04742.

[132] Pranav Gade, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Badllama: cheaply removing safety fine-tuning from llama 2-chat 13b, 2024. URL: https://arxiv.org/abs/2311.00117, arXiv:2311.00117, doi:10.48550/arXiv.2311.00117.

[133] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones,

Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL: https://arxiv.org/abs/2209.07858, arXiv:2209.07858, doi:10.48550/arXiv.2209.07858.

[134] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 619–633, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3243734.3243834.

[135] Simson Garfinkel, John Abowd, and Christian Martindale. Understanding database reconstruction attacks on public data. *Communications of the ACM*, 62:46–53, 02 2019. URL: https://dl.acm.org/doi/10.1145/3287287, doi:10.1145/3287287.

[136] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. AI2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (S&P)*, pages 3–18, 2018. URL: https://ieeexplore.ieee.org/document/8418593, doi:10.1109/SP.2018.00058.

[137] Jonas Geiping, Liam H Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=01olnfLIbD, doi:10.48550/arXiv.2009.02276.

[138] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference, 2021. URL: https://arxiv.org/abs/2103.13630, arXiv:2103.13630, doi:10.48550/arXiv.2103.13630.

[139] Antonio A. Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc. URL: https://arxiv.org/abs/1907.05012, doi:10.48550/arXiv.1907.05012.

[140] David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, and Vardan Papyan. LLM censorship: A machine learning challenge or a computer security problem?, 2023. URL: https://arxiv.org/abs/2307.10719, arXiv:2307.10719, doi:10.48550/arXiv.2307.10719.

[141] Micah Goldblum, Avi Schwarzschild, Ankit Patel, and Tom Goldstein. Adversarial attacks on machine learning systems for high-frequency trading. In *Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF '21, New York, NY,

USA, 2021. Association for Computing Machinery. doi:10.1145/3490354.3494367.

[142] Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. https://arxiv.org/abs/2204.06974, 2022. arXiv. doi:10.48550/ARXIV.2204.06974.

[143] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[144] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL: http://arxiv.org/abs/1412.6572, doi:10.48550/arXiv.1412.6572.

[145] Kai Greshake. Prompt injection defenses should suck less. https://kai-greshake.de/posts/approaches-to-pi-defense/, March 2024. Accessed: 2024-08-22.

[146] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *arXiv preprint arXiv:2302.12173*, 2023. URL: https://arxiv.org/abs/2302.12173, doi:10.48550/arXiv.2302.12173.

[147] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26, 2013. URL: https://papers.nips.cc/paper_files/paper/2013/hash/e034fb6b66aacc1d48f445ddfb08da98-Abstract.html.

[148] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. URL: https://ieeexplore.ieee.org/document/8685687, doi:10.1109/ACCESS.2019.2909068.

[149] Rachid Guerraoui, Arsany Guirguis, Jérémy Plassmann, Anton Ragot, and Sébastien Rouault. Garfield: System support for byzantine machine learning (regular paper). In *DSN*. IEEE, 2021. URL: https://arxiv.org/abs/2010.05888, doi:10.48550/arXiv.2010.05888.

[150] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL: https://aclanthology.org/2021.emnlp-main.464, doi:10.18653/v1/2021.emnlp-main.464.

[151] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024. URL: https://arxiv.org/abs/2402.01680, arXiv:2402.01680, doi:10.48550/arXiv.2402.01680.

[152] Niv Haim, Gal Vardi, Gilad Yehudai, michal Irani, and Ohad Shamir. Reconstructing training data from trained neural networks. In Alice H. Oh, Alekh Agarwal, Danielle

Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL: https://openreview.net/forum?id=Sxk8Bse3RKO, doi:10.48550/arXiv.2206.07758.

[153] Danny Halawi, Alexander Wei, Eric Wallace, Tony T. Wang, Nika Haghtalab, and Jacob Steinhardt. Covert malicious finetuning: Challenges in safeguarding llm adaptation, 2024. URL: https://arxiv.org/abs/2406.20053, arXiv:2406.20053, doi:10.48550/arXiv.2406.20053.

[154] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs, 2024. URL: https://arxiv.org/abs/2406.18495, arXiv:2406.18495, doi:10.48550/arXiv.2406.18495.

[155] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. LLM multi-agent systems: Challenges and open problems, 2024. URL: https://arxiv.org/abs/2402.03578, arXiv:2402.03578, doi:10.48550/arXiv.2402.03578.

[156] Drew Harwell. ID.me gathers lots of data besides face scans, including locations. Scammers still have found a way around it., retrieved December 2024. URL: https://www.washingtonpost.com/technology/2022/02/11/idme-facial-recognition-fraud-scams-irs/.

[157] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. SPECTRE: Defending against backdoor attacks using robust statistics. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4129–4139. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/hayase21a.html, doi:10.48550/arXiv.2104.11315.

[158] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety, 2022. URL: https://arxiv.org/abs/2109.13916, arXiv:2109.13916, doi:10.48550/arXiv.2109.13916.

[159] Isaac Hepworth, Kara Olive, Kingshuk Dasgupta, Michael Le, Mark Lodato, Mihai Maruseac, Sarah Meiklejohn, Shamik Chaudhuri, and Tehila Minkus. Securing the ai software supply chain. Technical report, Google, 2024. URL: https://research.google/pubs/securing-the-ai-software-supply-chain/.

[160] Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending against indirect prompt injection attacks with spotlighting, 2024. URL: https://arxiv.org/abs/2403.14720, arXiv:2403.14720, doi:10.48550/arXiv.2403.14720.

[161] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL: https://arxiv.org/abs/2203.15556, arXiv:2203.15556.

[162] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008. URL: https://pubmed.ncbi.nlm.nih.gov/18769715/, doi:10.1371/journal.pgen.1000167.

[163] Xiaoling Hu, Xiao Lin, Michael Cogswell, Yi Yao, Susmit Jha, and Chao Chen. Trigger hunting with a topological prior for trojan detection. In *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=TXsjU8BaibT, doi:10.48550/arXiv.2110.08335.

[164] Zhengmian Hu, Gang Wu, Saayan Mitra, Ruiyi Zhang, Tong Sun, Heng Huang, and Viswanathan Swaminathan. Token-level adversarial prompt detection based on perplexity measures and contextual information, 2024. URL: https://arxiv.org/abs/2311.11509, arXiv:2311.11509, doi:10.48550/arXiv.2311.11509.

[165] Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*, 2022. URL: https://arxiv.org/abs/2205.12628, doi:10.48550/arXiv.2205.12628.

[166] W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12080–12091. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/8ce6fc704072e351679ac97d4a985574-Paper.pdf, doi:10.48550/arXiv.2004.00225.

[167] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, 2022. URL: https://arxiv.org/abs/2201.07207, arXiv:2201.07207, doi:10.48550/arXiv.2201.07207.

[168] Xijie Huang, Moustafa Alzantot, and Mani Srivastava. NeuronInspect: Detecting backdoors in neural networks via output explanations, 2019. URL: https://arxiv.org/abs/1911.07399, doi:10.48550/ARXIV.1911.07399.

[169] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong

Chen, and Yue Zhao. Position: TrustLLM: Trustworthiness in large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL: https://openreview.net/forum?id=bWUU0LwwMp, doi:10.48550/arXiv.2401.05561.

[170] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. Sleeper agents: Training deceptive llms that persist through safety training, 2024. URL: https://arxiv.org/abs/2401.05566, arXiv:2401.05566, doi:10.48550/arXiv.2201.07207.

[171] W. Nicholson Price II. Risks and remedies for artificial intelligence in health care. https://www.brookings.edu/research/risks-and-remedies-for-artificial-intelligence-in-health-care/, 2019. Brookings Report.

[172] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2142–2151. PMLR, 2018. URL: http://proceedings.mlr.press/v80/ilyas18a.html, doi:10.48550/arXiv.1804.08598.

[173] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=BkMiWhR5K7, doi:10.48550/arXiv.1807.07978.

[174] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf, doi:10.48550/arXiv.1905.02175.

[175] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2008–2016. PMLR, 2021. URL: http://proceedings.mlr.press/v130/izzo21a.html, doi:10.48550/arXiv.2002.10077.

[176] Shahin Jabbari, Han-Ching Ou, Himabindu Lakkaraju, and Milind Tambe. An empirical study of the trade-offs between interpretability and fairness. In *ICML Workshop on Human Interpretability in Machine Learning, International Conference on Machine Learning (ICML)*, 2020. URL: https://teamcore.seas.harvard.edu/files/team

core/files/2020_jabbari_paper_32.pdf.

[177] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. URL: https://arxiv.org/abs/1909.01838, doi:10.48550/arXiv.1909.01838.

[178] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 3000–3008. PMLR, 2019. URL: https://arxiv.org/abs/1812.02696, doi:10.48550/arXiv.1812.02696.

[179] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 19–35, 2018. URL: https://arxiv.org/abs/1804.00308, doi:10.48550/arXiv.1804.00308.

[180] Matthew Jagielski, Giorgio Severi, Niklas Pousette Harger, and Alina Oprea. Subpopulation data poisoning attacks. In *Proceedings of the ACM Conference on Computer and Communications Security*, CCS, 2021. URL: https://arxiv.org/abs/2006.14026, doi:10.48550/arXiv.2006.14026.

[181] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private SGD? In *Advances in Neural Information Processing Systems*, volume 33, pages 22205–22216, 2020. URL: https://proceedings.neurips.cc/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf, doi:10.48550/arXiv.2006.07709.

[182] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models, 2023. URL: https://arxiv.org/abs/2309.00614, arXiv:2309.00614, doi:10.48550/arXiv.2309.00614.

[183] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 1895–1912, USA, 2019. USENIX Association. URL: https://arxiv.org/abs/1902.08874, doi:10.48550/arXiv.1902.08874.

[184] Bargav Jayaraman and David Evans. Are attribute inference attacks just imputation? In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 1569–1582, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3548606.3560663.

[185] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September

2017. Association for Computational Linguistics. URL: https://aclanthology.org/D17-1215, doi:10.18653/v1/D17-1215.

[186] Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms, 2024. URL: https://arxiv.org/abs/2402.11753, arXiv:2402.11753, doi:10.48550/arXiv.2402.11753.

[187] Pengfei Jing, Qiyi Tang, Yuefeng Du, Lei Xue, Xiapu Luo, Ting Wang, Sen Nie, and Shi Wu. Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3237–3254. USENIX Association, August 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/jing.

[188] Nikola Jovanovic, Robin Staab, and Martin Vechev. Watermark Stealing in Large Language Models. In *Proceedings of the 41-st International Conference on Machine Learning*, PMLR 235, June 2024. URL: https://files.sri.inf.ethz.ch/website/papers/jovanovic2024watermarkstealing.pdf, doi:10.48550/arXiv.2402.19361.

[189] Peter Kairouz, Brendan Mcmahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5213–5225. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/kairouz21b.html.

[190] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2019. URL: https://arxiv.org/abs/1912.04977, doi:10.48550/ARXIV.1912.04977.

[191] Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In Rupak Majumdar and Viktor Kunčak, editors, *Computer Aided Verification*, pages 97–117, Cham, 2017. Springer International Publishing. URL: https://arxiv.org/abs/1702.01135, doi:10.48550/arXiv.1702.01135.

[192] Michael Kearns and Ming Li. Learning in the presence of malicious errors. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, STOC '88, page 267–280, New York, NY, USA, 1988. Association for Computing Machinery.

doi:10.1145/62212.62238.

[193] Alaa Khaddaj, Guillaume Leclerc, Aleksandar Makelov, Kristian Georgiev, Hadi Salman, Andrew Ilyas, and Aleksander Madry. Rethinking backdoor attacks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16216–16236. PMLR, 23–29 Jul 2023. URL: https://proceedings.mlr.press/v202/khaddaj23a.html, doi:10.48550/arXiv.2307.10163.

[194] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models, 2023. URL: https://arxiv.org/abs/2301.10226, arXiv:2301.10226, doi:10.48550/arXiv.2301.10226.

[195] Marius Kloft and Pavel Laskov. Security analysis of online centroid anomaly detection. *Journal of Machine Learning Research*, 13(118):3681–3724, 2012. URL: http://jmlr.org/papers/v13/kloft12b.html.

[196] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017. URL: https://arxiv.org/abs/1703.04730, doi:10.48550/arXiv.1703.04730.

[197] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L. Buckley, Jason Phang, Samuel R. Bowman, and Ethan Perez. Pretraining language models with human preferences, 2023. URL: https://arxiv.org/abs/2302.08582, arXiv:2302.08582, doi:10.48550/arXiv.2302.08582.

[198] Moshe Kravchik, Battista Biggio, and Asaf Shabtai. Poisoning attacks on cyber attack detectors for industrial control systems. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21, page 116–125, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3412841.3441892.

[199] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning – industry perspectives. https://arxiv.org/abs/2002.05646, 2020. doi:10.48550/ARXIV.2002.05646.

[200] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. https://arxiv.org/abs/1607.02533, 2016. doi:10.48550/ARXIV.1607.02533.

[201] Keita Kurita and Paul Michel amd Graham Neubig. Weight poisoning attacks on pretrained models, 2020. URL: https://arxiv.org/abs/2004.06660, arXiv:2004.06660, doi:10.48550/arXiv.2004.06660.

[202] E. La Malfa and M. Kwiatkowska. The king is naked: On the notion of robustness for natural language processing. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, volume 10, page 11047–57. Association for the Advancement of Artificial Intelligence, 2022. URL: https://arxiv.org/abs/2112.07605, doi:10.48550/arXiv.2112.07605.

[203] Ricky Laishram and Vir Virander Phoha. Curie: A method for protecting SVM classi-

fier from poisoning attack. *CoRR*, abs/1606.01584, 2016. URL: http://arxiv.org/abs/1606.01584, arXiv:1606.01584, doi:10.48550/arXiv.1606.01584.

[204] Lakera. Guard, 2023. URL: https://www.lakera.ai/.

[205] Harry Langford, Ilia Shumailov, Yiren Zhao, Robert D. Mullins, and Nicolas Papernot. Architectural neural backdoors from first principles. *CoRR*, abs/2402.06957, 2024. URL: https://doi.org/10.48550/arXiv.2402.06957, arXiv:2402.06957, doi:10.48550/ARXIV.2402.06957.

[206] Learn Prompting. Defensive measures, 2023. URL: https://learnprompting.org/docs/category/-defensive-measures.

[207] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 656–672. IEEE, 2019. doi:10.1109/SP.2019.00044.

[208] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *Proceedings of the 29th USENIX Conference on Security Symposium*, SEC'20, USA, 2020. USENIX Association. URL: https://arxiv.org/abs/1906.11798, doi:10.48550/arXiv.1906.11798.

[209] Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=YUGG2tFuPM, doi:10.48550/arXiv.2006.14768.

[210] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. arXiv:2005.11401.

[211] Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, 22-26 May 2023*. IEEE, 2023. URL: https://arxiv.org/abs/2009.04131, doi:10.48550/arXiv.2009.04131.

[212] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL: https://arxiv.org/abs/2403.03218, arXiv:2403.03218, doi:

10.48550/arXiv.2403.03218.

[213] Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. Hidden backdoors in human-centric language models. In Yongdae Kim, Jong Kim, Giovanni Vigna, and Elaine Shi, editors, *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 3123–3140. ACM, 2021. doi:10.1145/3460120.3484576.

[214] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18:2088–2105, 2021. URL: https://arxiv.org/abs/1909.02742, doi:10.48550/arXiv.1909.02742.

[215] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. *CoRR*, abs/1807.00458, 2018. URL: http://arxiv.org/abs/1807.00458, arXiv:1807.00458, doi:10.14722/ndss.2019.23202.

[216] Ji Lin, Chuang Gan, and Song Han. Defensive quantization: When efficiency meets robustness. *ArXiv*, abs/1904.08444, 2019. URL: https://api.semanticscholar.org/CorpusID:53502621, doi:10.48550/arXiv.1904.08444.

[217] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In Michael Bailey, Sotiris Ioannidis, Manolis Stamatogiannakis, and Thorsten Holz, editors, *Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Proceedings*, Lecture Notes in Computer Science, pages 273–294. Springer Verlag, 2018. URL: https://link.springer.com/chapter/10.1007/978-3-030-00470-5_13, doi:10.1007/978-3-030-00470-5_13.

[218] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. URL: https://openreview.net/forum?id=Sys6GJqxl, doi:10.48550/arXiv.1611.02770.

[219] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. Prompt injection attack against LLM-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023. URL: https://arxiv.org/abs/2306.05499, doi:10.48550/arXiv.2306.05499.

[220] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking ChatGPT via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023. URL: https://arxiv.org/abs/2305.13860, doi:10.48550/arXiv.2305.13860.

[221] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. ABS: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 1265–1282, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3319535.3363216.

[222] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *NDSS*. The Internet Society, 2018. URL: http://dblp.uni-trier.de/db/conf/ndss/ndss2018.html#LiuMALZW018.

[223] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 182–199, Cham, 2020. Springer International Publishing. URL: https://arxiv.org/abs/2007.02343, doi:10.48550/arXiv.2007.02343.

[224] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, and toxicity, 2023. URL: https://arxiv.org/abs/2305.13169, arXiv:2305.13169, doi:10.48550/arXiv.2305.13169.

[225] Martin Bertran Lopez, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Scalable membership inference attacks via quantile regression. In *NeurIPS 2023*, 2023. URL: https://www.amazon.science/publications/scalable-membership-inference-attacks-via-quantile-regression.

[226] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 641–647, New York, NY, USA, 2005. Association for Computing Machinery. doi:10.1145/1081870.1081950.

[227] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles, 2017. URL: https://arxiv.org/abs/1707.03501, arXiv:1707.03501, doi:10.48550/arXiv.1707.03501.

[228] Yiwei Lu, Gautam Kamath, and Yaoliang Yu. Indiscriminate data poisoning attacks on neural networks. https://arxiv.org/abs/2204.09092, 2022. doi:10.48550/ARXIV.2204.09092.

[229] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE Computer Society, 2023. URL: https://arxiv.org/abs/2302.00539, doi:10.48550/arXiv.2302.00539.

[230] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[231] Pooria Madani and Natalija Vlajic. Robustness of deep autoencoder in intrusion detection under adversarial contamination. In *HoTSoS '18: Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security*, pages 1–8, 04 2018. doi:10.1145/3190619.3190637.

[232] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In

*6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=rJzIBfZAb, doi:10.48550/arXiv.1706.06083.

[233] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. Property inference from poisoning. In *2022 IEEE Symposium on Security and Privacy (S&P)*, pages 1120–1137, 2022. URL: https://ieeexplore.ieee.org/document/9833623, doi:10.1109/SP46214.2022.9833623.

[234] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs, 2024. URL: https://arxiv.org/abs/2401.06121, arXiv:2401.06121, doi:10.48550/arXiv.2401.06121.

[235] Neal Mangaokar, Ashish Hooda, Jihye Choi, Shreyas Chandrashekaran, Kassem Fawaz, Somesh Jha, and Atul Prakash. Prp: Propagating universal perturbations to attack large language model guard-rails, 2024. URL: https://arxiv.org/abs/2402.15911, arXiv:2402.15911, doi:10.48550/arXiv.2402.15911.

[236] James Manyika and Sissie Hsiao. An overview of Bard: an early experiment with generative AI. https://ai.google/static/documents/google-about-bard.pdf, February 2023. Google.

[237] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL: https://arxiv.org/abs/2402.04249, arXiv:2402.04249, doi:10.48550/arXiv.2402.15911.

[238] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science*, FOCS ’07, pages 94–103, Las Vegas, NV, USA, 2007. URL: https://ieeexplore.ieee.org/document/4389483, doi:10.1109/FOCS.2007.66.

[239] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. *arXiv preprint arXiv:2312.02119*, 2023. URL: https://arxiv.org/abs/2312.02119, doi:10.48550/arXiv.2312.02119.

[240] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 691–706. IEEE, 2019. doi:10.1109/SP.2019.00029.

[241] Melissa Heikkilä. This new data poisoning tool lets artists fight back against generative AI. https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/, October 2023. MIT Technology Review.

[242] El Mahdi El Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyên Hoang, and Sébastien Rouault. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). In *NeurIPS*, 2021. URL: https://arxiv.org/abs/2008.00742, doi:10.48550/arXiv.2

008.00742.

[243] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The Hidden Vulnerability of Distributed Learning in Byzantium. In *ICML*, 2018. URL: https://arxiv.org/abs/1802.07927, doi:10.48550/arXiv.1802.07927.

[244] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. Distributed momentum for byzantine-resilient stochastic gradient descent. In *ICLR*, 2021. URL: https://arxiv.org/abs/2003.00010, doi:10.48550/arXiv.2003.00010.

[245] Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. You can't have AI both ways: Balancing health data privacy and access fairly. *Artificial Intelligence Review volume*, 55:3503–3568, 2022. https://doi.org/10.1007/s10462-021-10088-y. doi:10.3389/fgene.2022.929453.

[246] Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019. URL: https://arxiv.org/abs/1908.10530, doi:10.48550/arXiv.1908.10530.

[247] Margaret Mitchell, Giada Pistilli, Yacine Jernite, Ezinwanne Ozoani, Marissa Gerchick, Nazneen Rajani, Sasha Luccioni, Irene Solaiman, Maraim Masoud, Somaieh Nikpoor, Carlos Muñoz Ferrandis, Stas Bekman, Christopher Akiki, Danish Contractor, David Lansky, Angelina McMillan-Major, Tristan Thrush, Suzana Ilić, Gérard Dupont, Shayne Longpre, Manan Dey, Stella Biderman, Douwe Kiela, Emi Baylor, Teven Le Scao, Aaron Gokaslan, Julien Launay, and Niklas Muennighoff. BigScience Large Open-science Open-access Multilingual Language Model. https://huggingface.co/bigscience/bloom, 2022. Hugging Face.

[248] MITRE. ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, retrieved December 2024. URL: https://atlas.mitre.org/.

[249] MITRE ATLAS. AML.M0001: Limit Model Artifact Release. https://atlas.mitre.org/mitigations/AML.M0001, 2023. Last Modified: 12 October 2023.

[250] MITRE ATLAS. AML.M0002: Passive ML Output Obfuscation. https://atlas.mitre.org/mitigations/AML.M0002, 2023. Last Modified: 12 October 2023.

[251] MITRE ATLAS. AML.M0004: Restrict Number of ML Model Queries. https://atlas.mitre.org/mitigations/AML.M0004, 2023. Last Modified: 12 October 2023.

[252] MITRE ATLAS. AML.M0000: Limit Release of Public Information. https://atlas.mitre.org/mitigations/AML.M0000, 2024. Last Modified: 12 January 2024.

[253] Payman Mohassel and Yupeng Zhang. SecureML: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 19–38, 2017. doi:10.1109/SP.2017.12.

[254] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *International Conference on Machine Learning (ICML)*, 2019. URL: https://arxiv.org/abs/1905.06635, doi:10.48550/arXiv.1905.06635.

[255] Olivia Moore. How Are Consumers Using Generative AI? *Andreessen Horowitz (a16z)*, 2023. URL: https://a16z.com/how-are-consumers-using-generative-ai/.

[256] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard.

Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. URL: https://arxiv.org/abs/1610.08401, doi:10.48550/arXiv.1610.08401.

[257] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. https://arxiv.org/abs/1511.04599, 2015. doi:10.48550/ARXIV.1511.04599.

[258] Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H. Falk. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76:355–375, 2021. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001330, doi:10.1016/j.inffus.2021.06.007.

[259] Sasi Kumar Murakonda and Reza Shokri. ML Privacy Meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning, 2020. URL: https://arxiv.org/abs/2007.09339, doi:10.48550/ARXIV.2007.09339.

[260] Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec ’17, 2017. URL: https://arxiv.org/abs/1708.08689, doi:10.48550/arXiv.1708.08689.

[261] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022. URL: https://arxiv.org/abs/2112.09332, arXiv:2112.09332, doi:10.48550/arXiv.2112.09332.

[262] Nina Narodytska and Shiva Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1318, 2017. URL: https://ieeexplore.ieee.org/document/8014906, doi:10.1109/CVPRW.2017.172.

[263] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In Joseph A. Calandrino and Carmela Troncoso, editors, *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 1631–1648. USENIX Association, 2023. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/nasr, doi:10.48550/arXiv.2302.07956.

[264] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy*, pages 739–753. IEEE, 2019. URL: https://arxiv.org/abs/1812.00910, doi:10.1109/SP.2019.00065.

[265] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas