Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. In *IEEE Symposium on Security & Privacy*, IEEE S&P '21, 2021. https://arxiv.org/abs/2101.04535. doi:10.48550/arXiv.2101.04535.

[266] National Cyber Security Centre. Machine learning principles. Technical report, National Cyber Security Centre, United Kingdom, 2024. Accessed: July 18, 2024. URL: https://www.ncsc.gov.uk/collection/machine-learning-principles.

[267] National Security Commission on Artificial Intelligence. Final report. https://www.nscai.gov/2021-final-report/, 2021. doi:10.48550/ARXIV.2006.03463.

[268] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 931–962. PMLR, 16–19 Mar 2021. URL: https://proceedings.mlr.press/v132/neel21a.html, doi:10.48550/arXiv.2007.02923.

[269] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I.P. Rubinstein, Udam Saini, Charles Sutton, and Kai Xia. Exploiting machine learning to subvert your spam filter. In *First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 08)*, San Francisco, CA, April 2008. USENIX Association. URL: https://www.usenix.org/conference/leet-08/exploiting-machine-learning-subvert-your-spam-filter.

[270] J. Newsome, B. Karp, and D. Song. Polygraph: Automatically generating signatures for polymorphic worms. In *2005 IEEE Symposium on Security and Privacy (S&P)*, pages 226–241, 2005. URL: https://ieeexplore.ieee.org/document/1425070, doi:10.1109/SP.2005.15.

[271] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and Thomas Schneider. FLAME: Taming backdoors in federated learning. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1415–1432, Boston, MA, August 2022. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/nguyen.

[272] Nisos. Building Trustworthy AI: Contending with Data Poisoning, retrieved December 2024. URL: https://www.nisos.com/research/building-trustworthy-ai/.

[273] NIST. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf, 2024. NIST AI Publication (NIST AI) NIST AI 600-1 Initial Public Draft, National Institute of Standards and Technology, Gaitherburg, MD. doi:10.6028/NIST.AI.600-1.

[274] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1, 2023. Online.

[275] National Institute of Standards and Technology. Managing misuse risk for dual-use

foundation models. https://doi.org/10.6028/NIST.AI.800-1.ipd, 2024. Online.

[276] Parmy Olson. Faces are the next target for fraudsters, retrieved December 2024. URL: https://www.wsj.com/articles/faces-are-the-next-target-for-fraudsters-11625662828.

[277] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[278] OpenAI. Assistants api overview, 2024. Accessed: 2024-08-18. URL: https://platform.openai.com/docs/assistants/overview.

[279] OpenAI. GPT-4o System Card. Online: https://cdn.openai.com/gpt-4o-system-card.pdf, August 2024.

[280] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4949–4958, 2019. URL: https://ieeexplore.ieee.org/document/8953839, doi:10.1109/CVPR.2019.00509.

[281] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. 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=TG8KACxEON, doi:10.48550/arXiv.2203.02155.

[282] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. https://arxiv.org/abs/1605.07277, 2016. doi:10.48550/ARXIV.1605.07277.

[283] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, page 506–519, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3052973.3053009.

[284] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (S&P)*, pages 582–597, 2016. URL: https://ieeexplore.ieee.org/document/7546524, doi:10.1109/SP.2016.41.

[285] V. Pareto. *Manuale di Economia Politica*. Società Editrice Libraria, Milan, 1906.

[286] V. Pareto. *Manual of Political Economy*. Augustus M. Kelley Publishers, New York, 1971. URL: https://www.loc.gov/item/05022672/.

[287] Dario Pasquini, Martin Strohmeier, and Carmela Troncoso. Neural Exec: Learning (and learning from) execution triggers for prompt injection attacks, 2024. URL: https://arxiv.org/abs/2403.03792, arXiv:2403.03792, doi:10.48550/arXiv.2403.03792.

[288] Arpita Patra, Thomas Schneider, Ajith Suresh, and Hossein Yalame. ABY2.0: Improved Mixed-Protocol secure Two-Party computation. In *30th USENIX Security*

Symposium (USENIX Security 21), pages 2165–2182. USENIX Association, August 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/patra.

[289] Andrea Paudice, Luis Muñoz-González, and Emil C. Lupu. Label sanitization against label flipping poisoning attacks. In Carlos Alzate, Anna Monreale, Haytham Assem, Albert Bifet, Teodora Sandra Buda, Bora Caglayan, Brett Drury, Eva García-Martín, Ricard Gavaldà, Stefan Kramer, Niklas Lavesson, Michael Madden, Ian Molloy, Maria-Irina Nicolae, and Mathieu Sinn, editors, *Nemesis/UrbReas/SoGood/IWAISe/GDM@PKDD/ECML*, volume 11329 of *Lecture Notes in Computer Science*, pages 5–15. Springer, 2018. URL: http://dblp.uni-trier.de/db/conf/pkdd/nemesis2018.html#PaudiceML18.

[290] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. Asleep at the keyboard? assessing the security of github copilot's code contributions, 2021. URL: https://arxiv.org/abs/2108.09293, arXiv:2108.09293, doi:10.48550/arXiv.2108.09293.

[291] R. Perdisci, D. Dagon, Wenke Lee, P. Fogla, and M. Sharif. Misleading worm signature generators using deliberate noise injection. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, Berkeley/Oakland, CA, 2006. IEEE. URL: http://ieeexplore.ieee.org/document/1623998/, doi:10.1109/SP.2006.26.

[292] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022. URL: https://arxiv.org/abs/2202.03286, doi:10.48550/arXiv.2202.03286.

[293] Neehar Peri, Neal Gupta, W. Ronny Huang, Liam Fowl, Chen Zhu, Soheil Feizi, Tom Goldstein, and John P. Dickerson. Deep k-nn defense against clean-label data poisoning attacks. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 55–70, Cham, 2020. Springer International Publishing. URL: https://arxiv.org/abs/1909.13374, doi:10.48550/arXiv.1909.13374.

[294] Neil Perry, Megha Srivastava, Deepak Kumar, and Dan Boneh. Do users write more insecure code with ai assistants? In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23. ACM, November 2023. URL: http://dx.doi.org/10.1145/3576915.3623157, doi:10.1145/3576915.3623157.

[295] Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. Intriguing properties of adversarial ML attacks in the problem space. In *2020 IEEE Symposium on Security and Privacy (S&P)*, pages 1308–1325. IEEE Computer Society, 2020. URL: https://doi.ieeecomputersociety.org/10.1109/SP40000.2020.00073, doi:10.1109/SP40000.2020.00073.

[296] Julien Piet, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David Wagner. Jatmo: Prompt injection defense by task-specific finetuning, 2024. URL: https://arxiv.org/abs/2312.17673, arXiv:2312.17673, doi:10.48550/arXiv.2312.17673.

[297] Krishna Pillutla, Galen Andrew, Peter Kairouz, H. Brendan McMahan, Alina Oprea,

and Sewoong Oh. Unleashing the power of randomization in auditing differentially private ml. In *Advances in Neural Information Processing Systems*, 2023. URL: https://arxiv.org/abs/2305.18447, doi:10.48550/arXiv.2305.18447.

[298] PromptArmor and Kai Greshake. Data exfiltration from writer.com with indirect prompt injection, 2023. URL: https://promptarmor.substack.com/p/data-exfiltration-from-writercom.

[299] Jonathan Protzenko, Bryan Parno, Aymeric Fromherz, Chris Hawblitzel, Marina Polubelova, Karthikeyan Bhargavan, Benjamin Beurdouche, Joonwon Choi, Antoine Delignat-Lavaud, Cédric Fournet, Natalia Kulatova, Tahina Ramananandro, Aseem Rastogi, Nikhil Swamy, Christoph Wintersteiger, and Santiago Zanella-Beguelin. EverCrypt: A fast, verified, cross-platform cryptographic provider. In *Proceedings of the IEEE Symposium on Security and Privacy (Oakland)*, May 2020. URL: https://eprint.iacr.org/2019/757, doi:10.1109/SP40000.2020.00114.

[300] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to!, 2023. URL: https://arxiv.org/abs/2310.03693, arXiv:2310.03693.

[301] Gauthama Raman M. R., Chuadhry Mujeeb Ahmed, and Aditya Mathur. Machine learning for intrusion detection in industrial control systems: Challenges and lessons from experimental evaluation. *Cybersecurity*, 4(27), 2021. URL: https://arxiv.org/abs/2202.11917, doi:10.48550/arXiv.2202.11917.

[302] Aida Rahmattalabi, Shahin Jabbari, Himabindu Lakkaraju, Phebe Vayanos, Max Izenberg, Ryan Brown, Eric Rice, and Milind Tambe. Fair influence maximization: A welfare optimization approach. In *Proceedings of the AAAI Conference on Artificial Intelligence 35th*, 2021. URL: https://arxiv.org/abs/2006.07906, doi:10.48550/arXiv.2006.07906.

[303] Adnan Siraj Rakin, Md Hafizul Islam Chowdhuryy, Fan Yao, and Deliang Fan. DeepSteal: Advanced model extractions leveraging efficient weight stealing in memories. In *2022 IEEE Symposium on Security and Privacy (S&P)*, pages 1157–1174, 2022. URL: https://ieeexplore.ieee.org/document/9833743, doi:10.1109/SP46214.2022.9833743.

[304] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. URL: https://ieeexplore.ieee.org/document/8103116, doi:10.1109/MSP.2017.2738401.

[305] Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback, 2024. URL: https://arxiv.org/abs/2311.14455, arXiv:2311.14455, doi:10.48550/arXiv.2311.14455.

[306] Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails, 2023. URL: https://arxiv.org/abs/2310.10501, arXiv:2310.10501, doi:10.48550/arXiv.2310.10501.

[307] Johann Rehberger. Data exfiltration via markdown injection - exploiting chatgpt's webpilot plugin, May 16 2023. Embrace The Red, Accessed: 2024-08-18. URL: https://embracethered.com/blog/posts/2023/chatgpt-webpilot-data-exfil-via-markdown-injection/.

[308] SNYK Report. AI Code Security and Trust: Organizations must change their approach. https://go.snyk.io/2023-ai-code-security-report-dwn-typ.html?aliId=eyJpIjoiUDFvdzRSdHI0dm5rVktvSSIsInQiOiJxOElRU2dQdkdqQm03ZjNLSDFFVkxBPT0ifQ%253D%253D, 2023. Human Centered AI, Stanford University.

[309] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Comput. Surv.*, 56(4), November 2023. doi:10.1145/3624010.

[310] Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Comput. Surv.*, 56(4), November 2023. doi:10.1145/3624010.

[311] Rishi Bommasani, et al. On the opportunities and risks of foundation models, 2024. URL: https://arxiv.org/abs/2108.07258, arXiv:2108.07258.

[312] Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *ArXiv*, abs/2310.03684, 2023. URL: https://api.semanticscholar.org/CorpusID:263671542, doi:10.48550/arXiv.2310.03684.

[313] Robust Intelligence. AI Firewall, 2023. URL: https://www.robustintelligence.com/platform/ai-firewall.

[314] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241. PMLR, 2020. URL: https://arxiv.org/abs/1902.02918, doi:10.48550/arXiv.1902.02918.

[315] Benjamin IP Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J Doug Tygar. Antidote: understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 1–14, 2009. doi:10.1145/1644893.1644895.

[316] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. *ArXiv*, abs/2404.01833, 2024. URL: https://api.semanticscholar.org/CorpusID:268856920, doi:10.48550/arXiv.2404.01833.

[317] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5558–5567. PMLR, 2019. URL: https://arxiv.org/abs/1908.11229, doi:10.48550/arXiv.1908.11229.

[318] Carl Sabottke, Octavian Suciu, and Tudor Dumitras. Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 1041–1056, Washington, D.C., August 2015. USENIX Association. URL: https://www.usenix.org/conference/

usenixsecurity15/technical-sessions/presentation/sabottke.

[319] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can ai-generated text be reliably detected?, 2024. URL: https://arxiv.org/abs/2303.11156, arXiv:2303.11156, doi:10.48550/arXiv.2303.11156.

[320] Vinu Sankar Sadasivan, Shoumik Saha, Gaurang Sriramanan, Priyatham Kattakinda, Atoosa Chegini, and Soheil Feizi. Fast adversarial attacks on language models in one gpu minute, 2024. URL: https://arxiv.org/abs/2402.15570, arXiv:2402.15570, doi:10.48550/arXiv.2402.15570.

[321] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. SoK: Let the privacy games begin! A unified treatment of data inference privacy in machine learning. https://arxiv.org/abs/2212.10986, 2022. doi:10.48550/ARXIV.2212.10986.

[322] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. https://arxiv.org/abs/2003.03675, 2020. doi:10.48550/ARXIV.2003.03675.

[323] Roman Samoilenko. New prompt injection attack on ChatGPT web version. markdown images can steal your chat data, 2023. URL: https://systemweakness.com/new-prompt-injection-attack-on-chatgpt-web-version-ef717492c5c2.

[324] Scale AI. Adversarial robustness leaderboard. https://scale.com/leaderboard/adversarial_robustness, 2024. Accessed: 2024-08-22.

[325] Oscar Schwartz. In 2016, Microsoft's racist chatbot revealed the dangers of online conversation: The bot learned language from people on Twitter—but it also learned values. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation, 2019. IEEE Spectrum.

[326] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. https://doi.org/10.6028/NIST.SP.1270, 2022. Special Publication (NIST SP) 800-1270, National Institute of Standards and Technology, Gaitherburg, MD. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf, doi:10.6028/NIST.SP.1270.

[327] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? A unified benchmark for backdoor and data poisoning attacks. https://arxiv.org/abs/2006.12557, 2020. arXiv. doi:10.48550/ARXIV.2006.12557.

[328] Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Gunnemann. Soft prompt threats: Attacking safety alignment and unlearning in open-source LLMs through the embedding space, 2024. URL: https://arxiv.org/abs/2402.09063, arXiv:2402.09063, doi:10.48550/arXiv.2402.09063.

[329] Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. Explanation-guided backdoor poisoning attacks against malware classifiers. In *30th USENIX Security Symposium (USENIX Security 2021)*, 2021. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/severi.

[330] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor

Dumitras, and Tom Goldstein. Poison frogs! Targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018. URL: https://arxiv.org/abs/1804.00792, doi:10.48550/arXiv.1804.00792.

[331] Shawn Shan, Arjun Nitin Bhagoji, Haitao Zheng, and Ben Y. Zhao. Poison forensics: Traceback of data poisoning attacks in neural networks. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3575–3592, Boston, MA, August 2022. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/shan.

[332] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, October 2016. URL: https://www.ece.cmu.edu/~lbauer/papers/2016/ccs2016-face-recognition.pdf, doi:10.1145/2976749.2978392.

[333] Vasu Sharma, Ankita Kalra, Vaibhav, Simral Chaudhary, Labhesh Patel, and LP Morency. Attend and attack: Attention guided adversarial attacks on visual question answering models. https://nips2018vigil.github.io/static/papers/accepted/33.pdf, 2018.

[334] Ryan Sheatsley, Blaine Hoak, Eric Pauley, Yohan Beugin, Michael J. Weisman, and Patrick McDaniel. On the robustness of domain constraints. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 495–515, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3460120.3484570.

[335] Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021. URL: https://www.ndss-symposium.org/wp-content/uploads/ndss2021_6C-3_24498_paper.pdf.

[336] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pages 1354–1371. IEEE, 2022. doi:10.1109/SP46214.2022.9833647.

[337] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023. URL: https://arxiv.org/abs/2308.03825, doi:10.48550/arXiv.2308.03825.

[338] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *CoRR*, abs/2308.03825, 2023. URL: https://doi.org/10.48550/arXiv.2308.03825, arXiv:2308.03825, doi:10.48550/ARXIV.2308.03825.

[339] Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. Prompt stealing attacks against text-to-image generation models. *arXiv preprint arXiv:2302.09923*, 2023.

URL: https://arxiv.org/abs/2302.09923, doi:10.48550/arXiv.2302.09923.

[340] Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Targeted latent adversarial training improves robustness to persistent harmful behaviors in LLMs, 2024. URL: https://arxiv.org/abs/2407.15549, arXiv:2407.15549, doi:10.48550/arXiv.2407.15549.

[341] Cong Shi, Tianfang Zhang, Zhuohang Li, Huy Phan, Tianming Zhao, Yan Wang, Jian Liu, Bo Yuan, and Yingying Chen. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, MobiCom '22, page 583–595, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3495243.3560531.

[342] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. URL: https://arxiv.org/abs/1610.05820, doi:10.48550/arXiv.1610.05820.

[343] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P), Oakland*, 2017. URL: https://arxiv.org/abs/1610.05820, doi:10.48550/arXiv.1610.05820.

[344] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico Kolter. Simple and efficient hard label black-box adversarial attacks in low query budget regimes. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1461–1469, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3447548.3467386.

[345] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. https://arxiv.org/abs/2006.03463, 2020. doi:10.48550/ARXIV.2006.03463.

[346] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3, January 2019. doi:10.1145/3290354.

[347] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. 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/2001.07685, doi:10.48550/arXiv.2001.07685.

[348] Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, Chandana Satya Prakash, Mukund Sridhar, Fabian Triefenbach, Apurv Verma, Gokhan Tur, and Prem Natarajan. AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model. https://www.amazon.science/publications/al

exatm-20b-few-shot-learning-using-a-large-scale-multilingual-seq2seq-model, 2022. Amazon.

[349] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, Baltimore, MD, August 2018. USENIX Association. URL: https://www.usenix.org/conference/woot18/presentation/eykholt, doi:10.48550/arXiv.1807.07769.

[350] Shuang Song and David Marn. Introducing a new privacy testing library in TensorFlow, 2020. URL: https://blog.tensorflow.org/2020/06/introducing-new-privacy-testing-library.html.

[351] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks, 2024. URL: https://arxiv.org/abs/2402.10260, arXiv:2402.10260, doi:10.48550/arXiv.2402.10260.

[352] N. Srndic and P. Laskov. Practical evasion of a learning-based classifier: A case study. In *Proc. IEEE Security and Privacy Symposium*, 2014. URL: https://personal.utdallas.edu/~muratk/courses/dmsec_files/srndic-laskov-sp2014.pdf.

[353] U.S. AI Safety Institute Technical Staff. Strengthening ai agent hijacking evaluations, 2024. URL: https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations.

[354] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/9d7311ba459f9e45ed746755a32dcd11-Paper.pdf, doi:10.48550/arXiv.1706.03691.

[355] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *Advances in Neural Information Processing Systems*, 2023. URL: https://arxiv.org/abs/2305.08846, doi:10.48550/arXiv.2305.08846.

[356] Ellen Su, Anu Vellore, Amy Chang, Raffaele Mura, Blaine Nelson, Paul Kassianik, and Amin Karbasi. Extracting memorized training data via decomposition. *arXiv preprint arXiv:2409.12367*, 2024. URL: https://arxiv.org/abs/2409.12367, doi:10.48550/arXiv.2409.12367.

[357] Octavian Suciu, Scott E Coull, and Jeffrey Johns. Exploring adversarial examples in malware detection. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 8–14. IEEE, 2019. URL: https://arxiv.org/abs/1810.08280, doi:10.48550/arXiv.1810.08280.

[358] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning FAIL? generalized transferability for evasion and poisoning attacks. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1299–1316, 2018. URL: https://arxiv.org/abs/1803.06975, doi:10.48550/arXiv.1803.06975.

[359] Jingwei Sun, Ang Li, Louis DiValentin, Amin Hassanzadeh, Yiran Chen, and Hai Li. FL-WBC: Enhancing robustness against model poisoning attacks in federated learning from a client perspective. In *NeurIPS*, 2021. URL: https://arxiv.org/abs/2110.13864, doi:10.48550/arXiv.2110.13864.

[360] Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H Brendan McMahan. Can you really backdoor federated learning? *arXiv:1911.07963*, 2019. URL: https://arxiv.org/abs/1911.07963, doi:10.48550/arXiv.1911.07963.

[361] Anshuman Suri and David Evans. Formalizing and estimating distribution inference risks. *Proceedings on Privacy Enhancing Technologies*, 2022. URL: https://arxiv.org/abs/2109.06024, doi:10.48550/arXiv.2109.06024.

[362] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL: http://arxiv.org/abs/1312.6199, doi:10.48550/arXiv.1312.6199.

[363] Rahim Taheri, Reza Javidan, Mohammad Shojafar, Zahra Pooranian, Ali Miri, and Mauro Conti. On defending against label flipping attacks on malware detection systems. *CoRR*, abs/1908.04473, 2019. URL: http://arxiv.org/abs/1908.04473, arXiv:1908.04473, doi:10.48550/arXiv.1908.04473.

[364] Azure AI Red Team. Pyrit: The python risk identification tool for generative ai. https://github.com/Azure/PyRIT, 2024. Accessed: 2024-08-18.

[365] The Llama Team. The LLaMA3 Herd of Models, 2024. URL: https://arxiv.org/abs/2407.21783, doi:10.48550/arXiv.2407.21783.

[366] The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, October 2023. The White House.

[367] T. Ben Thompson and Michael Sklar. Breaking circuit breakers. URL: https://confirmlabs.org/posts/circuit_breaking.html.

[368] T. Ben Thompson and Michael Sklar. Fluent student-teacher redteaming, 2024. URL: https://arxiv.org/abs/2407.17447, arXiv:2407.17447, doi:10.48550/arXiv.2407.17447.

[369] Anvith Thudi, Ilia Shumailov, Franziska Boenisch, and Nicolas Papernot. Bounding membership inference. https://arxiv.org/abs/2202.12232, 2022. doi:10.48550/ARXIV.2202.12232.

[370] Lionel Nganyewou Tidjon and Foutse Khomh. Threat assessment in machine learning based systems. *arXiv preprint arXiv:2207.00091*, 2022. URL: https://arxiv.org/abs/2207.00091, doi:10.48550/arXiv.2207.00091.

[371] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302

.13971, arXiv:2302.13971, doi:10.48550/arXiv.2302.13971.

[372] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288, arXiv:2307.09288, doi:10.48550/arXiv.2307.09288.

[373] Florian Tramer. Detecting adversarial examples is (Nearly) as hard as classifying them. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21692–21702. PMLR, 17–23 Jul 2022. URL: https://proceedings.mlr.press/v162/tramer22a.html.

[374] Florian Tramer, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Joern-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9561–9571. PMLR, 13–18 Jul 2020. URL: https://proceedings.mlr.press/v119/tramer20a.html, doi:10.48550/arXiv.2002.04599.

[375] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Mądry. On adaptive attacks to adversarial example defenses. 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/2002.08347, doi:10.48550/arXiv.2002.08347.

[376] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *USENIX Security*, 2016. URL: https://arxiv.org/abs/1609.02943, doi:10.48550/arXiv.1609.02943.

[377] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. https://arxiv.org/abs/1704.03453, 2017. doi:10.48550/ARXIV.1704.03453.

[378] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31.

Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/280cf18baf4311c92aa5a042336587d3-Paper.pdf, doi:10.48550/arXiv.1811.00636.

[379] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=SyxAb30cY7, doi:10.48550/arXiv.1805.12152.

[380] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. In *ICLR*, 2019. URL: https://openreview.net/forum?id=HJg6e2CcK7.

[381] U.K. AI Safety Institute. Advanced ai evaluations: May update, 2024. Accessed: 2024-08-18. URL: https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

[382] Kush R. Varshney. *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA, 2022. URL: https://www.trustworthymachinelearning.com/.

[383] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. URL: https://arxiv.org/abs/1706.03762, doi:10.48550/arXiv.1706.03762.

[384] Sridhar Venkatesan, Harshvardhan Sikka, Rauf Izmailov, Ritu Chadha, Alina Oprea, and Michael J. De Lucia. Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems. In *MILCOM*, pages 874–879. IEEE, 2021. URL: https://ieeexplore.ieee.org/document/9652916, doi:10.1109/MILCOM52596.2021.9652916.

[385] Sameer Wagh, Shruti Tople, Fabrice Benhamouda, Eyal Kushilevitz, Prateek Mittal, and Tal Rabin. FALCON: honest-majority maliciously secure framework for private deep learning. In *Proceedings on Privacy Enhancing Technologies (PoPETs) 2021, Issue 1*, 20201.

[386] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*, 2019. URL: https://arxiv.org/abs/1908.07125, doi:10.48550/arXiv.1908.07125.

[387] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions, 2024. URL: https://arxiv.org/abs/2404.13208, arXiv:2404.13208, doi:10.48550/arXiv.2404.13208.

[388] Eric Wallace, Tony Z. Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on NLP models. In *NAACL*, 2021. URL: https://arxiv.org/abs/2010.12563, doi:10.48550/arXiv.2010.12563.

[389] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning, 2023. URL: https://arxiv.org/abs/2305.00944, arXiv:2305.00944, doi:10.48550/arXiv.2305.00944.

[390] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng,

and Ben Y. Zhao. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, San Francisco, CA, USA, May 2019. IEEE. URL: https://ieeexplore.ieee.org/document/8835365/, doi:10.1109/SP.2019.00031.

[391] Haotao Wang, Tianlong Chen, Shupeng Gui, Ting-Kuei Hu, Ji Liu, and Zhangyang Wang. Once-for-All Adversarial Training: In-Situ Tradeoff between Robustness and Accuracy for Free. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*, 2020. URL: https://arxiv.org/abs/2010.11828, doi:10.48550/arXiv.2010.11828.

[392] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. In *NeurIPS*, 2020. URL: https://arxiv.org/abs/2007.05084, doi:10.48550/arXiv.2007.05084.

[393] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024. URL: http://dx.doi.org/10.1007/s11704-024-40231-1, doi:10.1007/s11704-024-40231-1.

[394] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Formal security analysis of neural networks using symbolic intervals. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1599–1614, Baltimore, MD, August 2018. USENIX Association. URL: https://www.usenix.org/conference/usenixsecurity18/presentation/wang-shiqi.

[395] Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 22769–22783. PMLR, 2022. URL: https://proceedings.mlr.press/v162/wang22m.html, doi:10.48550/arXiv.2202.02628.

[396] Wenxiao Wang, Alexander J Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (Deterministic) finite aggregation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22769–22783. PMLR, 17–23 Jul 2022. URL: https://proceedings.mlr.press/v162/wang22m.html, doi:10.48550/arXiv.2202.02628.

[397] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1924–1933. Computer Vision Foundation / IEEE, 2021. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Enhancing_the_Transferability_of_Adversarial_Attacks_Through_Va

riance_Tuning_CVPR_2021_paper.html, doi:10.1109/CVPR46437.2021.00196.

[398] Yanting Wang, Wei Zou, and Jinyuan Jia. FCert: Certifiably robust few-shot classification in the era of foundation models. In *Proc. IEEE Security and Privacy Symposium*, 2024. URL: https://arxiv.org/abs/2404.08631, doi:10.48550/arXiv.2404.08631.

[399] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in LLMs, 2023. URL: https://arxiv.org/abs/2308.13387, arXiv:2308.13387, doi:10.48550/arXiv.2308.13387.

[400] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

[401] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. doi:10.1609/aaai.v33i01.33018973.

[402] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12338–12345, 2020. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6918, doi:10.48550/arXiv.1911.09449.

[403] Lilian Weng. Adversarial attacks on latent language models, 2023. URL: https://lilianweng.github.io/posts/2023-10-25-adv-attack-llm/.

[404] Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks against deep learning systems in the physical world. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6202–6211, 2020. URL: https://arxiv.org/abs/2006.14580, doi:10.48550/arXiv.2006.14580.

[405] Simon Willison. The dual LLM pattern for building AI assistants that can resist prompt injection, 2023. Accessed: 2024-08-22. URL: https://simonwillison.net/2023/Apr/25/dual-llm-pattern/.

[406] Yotam Wolf, Noam Wies, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *ArXiv*, abs/2304.11082, 2023. URL: https://api.semanticscholar.org/CorpusID:258291526, doi:10.48550/arXiv.2304.11082.

[407] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16913–16925. Curran Associates, Inc., 2021. URL: https://proceedings.neurips.cc/paper/2021/file/8cbe9ce23f42628c98f80fa0fac8b19a-Paper.pdf, doi:10.48550/arXiv.2110.14430.

[408] Fangzhou Wu, Ning Zhang, Somesh Jha, Patrick McDaniel, and Chaowei Xiao. A new

era in LLM security: Exploring security concerns in real-world LLM-based systems, 2024. URL: https://arxiv.org/abs/2402.18649, arXiv:2402.18649.

[409] Xi Wu, Matthew Fredrikson, Somesh Jha, and Jeffrey F. Naughton. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pages 355–370, 2016. doi:10.1109/CSF.2016.32.

[410] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. SecGPT: An execution isolation architecture for LLM-based systems, 2024. URL: https://arxiv.org/abs/2403.04960, arXiv:2403.04960, doi:10.48550/arXiv.2402.18649.

[411] Zhen Xiang, David J. Miller, and George Kesidis. Post-training detection of backdoor attacks for two-class and multi-attack scenarios. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL: https://openreview.net/forum?id=MSgB8D4Hy51, doi:10.48550/arXiv.2201.08474.

[412] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *International Conference on Machine Learning*, pages 1689–1698, 2015. URL: https://arxiv.org/abs/1804.07933, doi:10.48550/arXiv.1804.07933.

[413] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/44feb0096faa8326192570788b38c1d1-Paper.pdf, doi:10.48550/arXiv.1904.12848.

[414] Weilin Xu, Yanjun Qi, and David Evans. Automatically evading classifiers. In *Proceedings of the 2016 Network and Distributed Systems Symposium*, pages 21–24, 2016. URL: https://www.cs.virginia.edu/~evans/pubs/ndss2016/.

[415] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. https://arxiv.org/abs/1709.08693, 2017. doi:10.48550/ARXIV.1709.08693.

[416] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A. Gunter, and Bo Li. Detecting AI trojans using meta neural analysis. In *IEEE Symposium on Security and Privacy, S&P 2021*, pages 103–120, United States, May 2021. URL: https://ieeexplore.ieee.org/document/9519467, doi:10.1109/SP40001.2021.00034.

[417] Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Xplore, 2022. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9578130, doi:10.48550/ARXIV.2206.12714.

[418] Limin Yang, Zhi Chen, Jacopo Cortellazzi, Feargus Pendlebury, Kevin Tu, Fabio Pierazzi, Lorenzo Cavallaro, and Gang Wang. Jigsaw puzzle: Selective backdoor attack to subvert malware classifiers. *CoRR*, abs/2202.05470, 2022. URL: https://arxiv.or

g/abs/2202.05470, arXiv:2202.05470, doi:10.48550/arXiv.2202.05470.

[419] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL: https://arxiv.org/abs/2210.03629, arXiv:2210.03629, doi:10.48550/arXiv.2210.03629.

[420] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 2041–2055, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3319535.3354209.

[421] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 3093–3106, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3548606.3560675.

[422] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, CSF '18, pages 268–282, 2018. https://arxiv.org/abs/1709.01604. URL: https://arxiv.org/abs/1709.01604, doi:10.48550/arXiv.1709.01604.

[423] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *ICML*, 2018. URL: https://arxiv.org/abs/1803.01498, doi:10.48550/arXiv.1803.01498.

[424] Youngjoon Yu, Hong Joo Lee, Byeong Cheon Kim, Jung Uk Kim, and Yong Man Ro. Investigating vulnerability to adversarial examples on multimodal data fusion in deep learning. https://arxiv.org/abs/2005.10987, 2020. Online. doi:10.48550/ARXIV.2005.10987.

[425] Andrew Yuan, Alina Oprea, and Cheng Tan. Dropout attacks. In *IEEE Symposium on Security and Privacy (S&P)*, 2024. URL: https://arxiv.org/abs/2309.01614, doi:10.48550/arXiv.2309.01614.

[426] 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 *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, page 363–375, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3372297.3417880.

[427] Santiago Zanella-Beguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, and Daniel Jones. Bayesian estimation of differential privacy. 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 40624–40636. PMLR, 23–29 Jul

2023. URL: https://proceedings.mlr.press/v202/zanella-beguelin23a.html, doi:10.48550/arXiv.2206.05199.

[428] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models, 2021. URL: https://arxiv.org/abs/2106.02636, arXiv:2106.02636, doi:10.48550/arXiv.2106.02636.

[429] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=MeeQkFYVbzW, doi:10.48550/arXiv.2110.03735.

[430] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents, 2024. URL: https://arxiv.org/abs/2403.02691, arXiv:2403.02691, doi:10.48550/arXiv.2403.02691.

[431] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021. doi:10.1145/3446776.

[432] Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *ArXiv*, abs/2311.04378, 2023. URL: https://api.semanticscholar.org/CorpusID:265050535, doi:10.48550/arXiv.2311.04378.

[433] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. 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 7472–7482. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/zhang19p.html, doi:10.48550/arXiv.1901.08573.

[434] Ruisi Zhang, Seira Hidano, and Farinaz Koushanfar. Text revealer: Private text reconstruction via model inversion attacks against transformers. *arXiv preprint arXiv:2209.10505*, 2022. URL: https://arxiv.org/abs/2209.10505, doi:10.48550/arXiv.2209.10505.

[435] Su-Fang Zhang, Jun-Hai Zhai, Bo-Jun Xie, Yan Zhan, and Xin Wang. Multimodal representation learning: Advances, trends and challenges. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–6. IEEE, 2019. URL: https://api.semanticscholar.org/CorpusID:209901378, doi:10.1109/ICMLC48188.2019.8949228.

[436] Susan Zhang, Mona Diab, and Luke Zettlemoyer. Democratizing access to large-scale language models with OPT-175B. https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/, 2022. Meta AI.

[437] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. Leakage of dataset properties in Multi-Party machine learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2687–2704. USENIX Association, August 2021. URL: https://www.us

enix.org/conference/usenixsecurity21/presentation/zhang-wanrong, doi:10.48550/arXiv.2006.07267.

[438] Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(3), apr 2020. doi:10.1145/3374217.

[439] Yiming Zhang and Daphne Ippolito. Prompts should not be seen as secrets: Systematically measuring prompt extraction attack success. *arXiv preprint arXiv:2307.06865*, 2023. URL: https://arxiv.org/abs/2307.06865, doi:10.48550/arXiv.2307.06865.

[440] Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni. Bagflip: A certified defense against data poisoning. 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=ZidkM5b92G, doi:10.48550/arXiv.2205.13634.

[441] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26429–26446. PMLR, 17–23 Jul 2022. URL: https://proceedings.mlr.press/v162/zhang22w.html, doi:10.48550/arXiv.2206.10341.

[442] Zhikun Zhang, Min Chen, Michael Backes, Yun Shen, and Yang Zhang. Inference attacks against graph neural networks. In *31st USENIX Security Symposium (USENIX Security 22)*, 2022. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/zhang-zhikun.

[443] Junhao Zhou, Yufei Chen, Chao Shen, and Yang Zhang. Property inference attacks against GANs. In *Proceedings of Network and Distributed System Security*, NDSS, 2022. URL: https://arxiv.org/abs/2111.07608, doi:10.48550/arXiv.2111.07608.

[444] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. 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 7614–7623. PMLR, 09–15 Jun 2019. URL: https://proceedings.mlr.press/v97/zhu19a.html, doi:10.48550/arXiv.1905.05897.

[445] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL: https://arxiv.org/abs/1909.08593, arXiv:1909.08593, doi:10.48550/arXiv.1909.08593.

[446] Giulio Zizzo, Chris Hankin, Sergio Maffeis, and Kevin Jones. Adversarial machine learning beyond the image domain. In *Proceedings of the 56th Annual Design Automation Conference 2019*, DAC '19, New York, NY, USA, 2019. Association for Com-

puting Machinery. doi:10.1145/3316781.3323470.

[447] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers, 2024. URL: https://arxiv.org/abs/2406.04313, arXiv:2406.04313, doi:h10.48550/arXiv.2406.04313.

[448] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. URL: https://arxiv.org/abs/2307.15043, doi:10.48550/arXiv.2307.15043.

[449] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models, 2024. URL: https://arxiv.org/abs/2005.11401, arXiv:2402.07867, doi:10.48550/arXiv.2005.11401.

Appendix A. Glossary

Clicking on the page number at the end of a definition will navigate to the page where the term is used.

A

adversarial example A modified testing sample that induces misclassification or misbehavior of a machine learning model at deployment time. ix, 6

adversarial machine learning Attacks that exploit the statistical, data-based nature of machine learning systems. xii, 1

agent Software programs that can interact with their environment, receive information, and undertake self-directed actions in service of a larger, externally-specified goal. 1, 35, 37, 39, 50, 52, 54

area under the curve A measure of the ability of a classifier to distinguish between classes in machine learning. A higher AUC means that a model performs better when distinguishing between the two classes. AUC measures the entire two-dimensional area under the RECEIVER OPERATING CHARACTERISTIC (ROC) curve. 30

attribute inference attacks An attack against machine learning models that infers sensitive attributes of a training data record, given partial knowledge about the record. 7

availability breakdown In the AML context, a disruption of the ability of other users or processes to obtain timely and reliable access to an AI system's outputs or functionality. 6, 39

B

backdoor pattern A transformation or insertion applied to a data sample that triggers an adversary-specified behaviour in a model that has been subject to a backdoor poisoning attack. For example, in computer vision, an adversary could poison a model such that the insertion of a square of white pixels induces a desired target label. 6, 22, 107

backdoor poisoning attack A poisoning attack that causes a model to perform an adversary-selected behaviour in response to inputs that follow a particular BACKDOOR PATTERN. 6, 42

C

classification The task of predicting which of a set of discrete categories an input belongs to. 5

convolutional neural networks A class of feed-forward neural networks that include at least one convolutional layer, referred to as CNNs. In convolutional layers, feature detectors (known as kernels or filters) detect specific features across the input data. CNNs are primarily used for processing grid-like data, such as images, and are particularly effective for tasks like image classification, object detection, and image segmentation. 5, 31

D

data confidentiality A well-established concept in cybersecurity referring to the protection of sensitive information from unauthorized access and disclosure. 7

data poisoning A POISONING ATTACKS in which an adversary controls part of the training data. 5, 36, 37, 40, 111

data privacy attacks Attacks against machine learning models that extract sensitive information about training data. 7

data reconstruction Privacy attacks that reconstruct sensitive data in a model's training data from aggregate information. 7, 28

deployment stage The stage of the machine learning pipeline in which a model is deployed into a live or real-world environment for use, such as being integrated into an enterprise application or made available to end users through an API. 5, 37, 38

diffusion models A class of latent variable generative models consisting of three major components: a forward process, a reverse process, and a sampling procedure. The goal of the diffusion model is to learn a diffusion process that generates the probability distribution of a given dataset. It is widely used in computer vision on a variety of tasks, including image denoising, inpainting, super-resolution, and image generation. 34

direct prompt injection A DIRECT PROMPTING ATTACK in which the attacker exploits PROMPT INJECTION. 43, 110

direct prompting attack In the generative AI context, an attack conducted by the primary user of the system through QUERY ACCESS (e.g., as opposed to through RESOURCE CONTROL). 34, 43, 108, 110

discriminative A type of machine learning method that learns to discriminate between classes. 5

E

energy-latency attack An attack that exploits the performance dependency on hardware and model optimizations to negate the effects of hardware optimizations, increase computational latency, increase hardware temperature, and massively increase the amount of energy consumed. 6, 8

ensemble learning A type of a meta machine learning approach that combines the predictions of several models to improve performance. 5

expectation over transformation A method for strengthening adversarial examples to remain adversarial under image transformations that occur in the real world, such as angle and viewpoint changes. EOT models these perturbations within the optimization procedure. Rather than optimizing the log-likelihood of a single example, EOT uses a chosen distribution of transformation functions that take an input controlled by the adversary to the "true" input perceived by the classifier. 16

F

federated learning A type of machine learning in which a model is trained in a decentralized fashion using multiple data sources without pooling or combining the data in any centralized location. Federated learning allows entities or devices to collaboratively train a global model by exchanging model updates without directly sharing the data that each entity controls. 5, 31

feedforward neural networks Artificial neural networks in which the connections between nodes is from one layer to the next and do not form a cycle. 31

fine-tuning The process of adapting a pre-trained model to perform specific tasks or specialize in a particular domain. This phase follows the initial pre-training phase and involves further training the model on task-specific data. This is often a supervised learning task. 37

fine-tuning circumvention Fine-tuning to remove model refusal behaviour or other model-level safety interventions. 41

formal methods A mathematically rigorous technique for the specification, development, and verification of software systems. 18

foundation model In generative AI, models trained on broad data using SELF-SUPERVISED LEARNING that can be adapted such as through fine-tuning for a variety of downstream tasks [311]. 111

functional attack An adversarial attack that is optimized for a set of data in a domain rather than per data point. 13, 23

G

generative adversarial networks A machine learning framework in which two neural networks contest with each other in the form of a zero-sum game, where one agent's gain is another agent's loss. A GAN learns to generate new data with the same statistics as the training set. See [143] for further details. 31, 34

generative pre-trained transformer (GPT) A family of machine learning models based on the transformer architecture [383] that are pre-trained through SELF-SUPERVISED LEARNING on large data sets of unlabelled text. This is the current predominant architecture for large language models. 34

graph neural network A neural network designed to process graph-structured data. GNNs perform optimizable transformations on graph attributes (e.g., nodes, edges, global context) while preserving graph symmetries such as permutation invariance. GNNs utilize a "graph-in, graph-out" architecture that takes an input graph with information and progressively transforms it into an output graph with the same connectivity as that of the input graph. 31

H

hidden Markov model A Markov model in which the system being modeled is assumed to be a Markov process with unobservable states. The model provides an observable process whose outcomes are influenced by the outcomes of a Markov model in a known way. An HMM can be used to describe the evolution of observable

events that depend on internal factors that are not directly observable. In machine learning, it is assumed that the internal state of a model is hidden but not its hyperparameters. 31

I

indirect prompt injection A type of PROMPT INJECTION executed through RESOURCE CONTROL rather than through user-provided input as in a DIRECT PROMPT INJECTION. 39–41, 50

integrity violation In the AML context, an AI system being forced to misperform against its intended objectives, producing outputs or predictions that align with the attacker's objective. 6, 40

J

jailbreak A DIRECT PROMPTING ATTACK intended to circumvent restrictions placed on model outputs, such as circumventing refusal behaviour to enable misuse. 34, 38, 42, 43, 52

L

label flipping A type of data poisoning attack in which an adversary is restricted to changing the training labels. 20

label limit A capability with which an attacker does not control the labels of training samples in supervised learning. 8

logistic regression A type of linear classifier that predicts the probability of an observation being part of a class. 5

M

machine unlearning A technique that involves selectively removing the influences of specific training data points from a trained machine learning model, such as to remove unwanted capabilities or knowledge in a foundation model, or to enable a user to request the removal of their records from a model. Efficient approximate unlearning techniques may not require retraining the ML model from scratch. 33

membership-inference attack A data privacy attack to determine whether a data sample was part of the training set of a machine learning model. 7, 28

misuse enablement In the AML context, a circumvention of technical restrictions imposed by the AI system's owner on its use, such as restrictions designed to prevent a GenAI system from producing outputs that could cause harm to others. 40

model control A capability with which an attacker can control the machine learning model parameters. 8, 37, 41, 111

model extraction A type of privacy attack that extracts details of the model architecture and/or parameters. 7, 28, 31, 40, 41, 47

model poisoning A POISONING ATTACKS which operates through MODEL CONTROL. 5, 6, 37, 41, 111

model privacy attacks An attack against machine learning models to extract sensitive information about the model. 7

multimodal models A model that processes and relates information from multiple sensory modalities that each represent primary human channels of communication and sensation, such as vision and touch. 58

O

out-of-distribution Data that was collected at a different time and possibly under different conditions or in a different environment than the data collected to train the model. 56

P

poisoning attacks Adversarial attacks in which an adversary interferes with a model during its TRAINING STAGE, such as by inserting malicious training data (DATA POISONING) or modifying the training process itself (MODEL POISONING). 5, 108, 111

pre-training A component of the TRAINING STAGE in which a model learns general patterns, features, and relationships from vast amounts of unlabeled data, such as through SELF-SUPERVISED LEARNING. Pre-training can equip models with knowledge of general features or patterns which may be useful in downstream tasks (see FOUNDATION MODEL), and can be followed with additional training or fine-tuning that specializes the model for a specific downstream task. 37

privacy compromise In the AML context, the unauthorized access of restricted or proprietary information that is part of an AI system, including information about a model's training data, weights or architecture; or sensitive information that the model accesses such as the knowledge base of a GenAI RETRIEVAL-AUGMENTED GENERATION (RAG) application. 7, 40

prompt extraction An attack that tries to divulge the system prompt or other information in the context of a large language model that would normally be hidden from a user. 38, 41

prompt injection An attack which exploits the concatenation of untrusted input with a prompt constructed by a higher-trust party such as the application designer. 38, 41, 108, 110

property inference A data privacy attack that infers a global property about the training data of a machine learning model. 7

Q

query access A capability with which an attacker can issue queries to a trained machine learning model and obtain predictions or generations. 8, 40, 108

R

receiver operating characteristic (ROC) A curve that plots the true positive rate versus the false positive rate for a classifier. 107

red teaming in the AI context, means a structured testing effort, often adopting adversarial methods, to find flaws and vulnerabilities in an AI system, including unforeseen or undesirable system behaviors or potential risks associated with the misuse of the system. [366]. 60

regression A type of supervised machine learning model that is trained on data, including numerical labels (i.e., response variables). Types of regression algorithms include linear regression, polynomial regression, and various non-linear regression methods. 5

reinforcement learning A type of machine learning in which a model learns to optimize its behavior according to a reward function by interacting with and receiving feedback from an environment. 5

resource control A capability in which an attacker controls one or more external resources consumed by a machine learning model at inference time, particularly for GenAI systems such as retrieval-augmented generation applications. 41, 50, 108, 110

retrieval-augmented generation (RAG) A type of GenAI system in which a model is paired with a separate information retrieval system (or ”knowledge base”). Based on a user query, the RAG system identifies relevant information within the knowledge base and provides it to the GenAI model in context for the model to use in formulating its response. RAG systems allow the internal knowledge of a GenAI model to be modified without the need for retraining. 1, 35, 37, 38, 40, 46, 50, 111

rowhammer attack A software-based fault-injection attack that exploits dynamic random-access memory disturbance errors via user-space applications and allows the attacker to infer information about certain victim secrets stored in memory cells. Mounting this attack requires the attacker to control a user-space unprivileged process that runs on the same machine as the victim’s machine learning model. 31

S

self-supervised learning A type of machine learning that relies on generating implicit labels from unstructured data rather than relying on explicit, human-created labels. Self-supervised learning tasks are constructed to allow the true labels to be automatically inferred from the training data (enabling the use of large-scale training data) and to require models to capture essential features or relationships within the data to solve them. For example, a common self-supervised learning task is providing a model with partial data with the task to accurately generate the remainder. 109, 111

semi-supervised learning A type of machine learning in which a small number of training samples are labeled, while the majority are unlabeled. 5

shadow model A model that imitates the behavior of the target model. The training datasets and the truth about membership in these datasets are known for these models.